Analyses of non-coding somatic drivers in 2,658 cancer whole genomes

The discovery of drivers of cancer has traditionally focused on protein-coding genes\(^1\)–\(^4\). Here we present analyses of driver point mutations and structural variants in non-coding regions across 2,658 genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium\(^5\) of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). For point mutations, we developed a statistically rigorous strategy for combining significance levels from multiple methods of driver discovery that overcomes the limitations of individual methods. For structural variants, we present two methods of driver discovery, and identify regions that are significantly affected by recurrent breakpoints and recurrent somatic juxtapositions. Our analyses confirm previously reported drivers\(^6\)–\(^7\), raise doubts about others and identify novel candidates, including point mutations in the 5′ region of \(\text{TP53}\)\(^{12}\), in the 3′ untranslated regions of \(\text{NFKBIZ}\) and \(\text{TOB1}\), focal deletions in \(\text{BRD4}\) and rearrangements in the loci of \(\text{AKR1C}\) genes. We show that although point mutations and structural variants that drive cancer are less frequent in non-coding genes and regulatory sequences than in protein-coding genes, additional examples of these drivers will be found as more cancer genomes become available.

Previous large-scale sequencing projects have identified many putative cancer genes, but most efforts have concentrated on mutations and copy-number alterations in protein-coding genes, mainly using whole-exome sequencing and single-nucleotide polymorphism arrays\(^1\)–\(^4\). Whole-genome sequencing has made it possible to systematically survey non-coding regions for potential driver events, including single-nucleotide variants (SNVs), small insertions and deletions (indels) and larger structural variants. Whole-genome sequencing enables the precise localization of structural variant breakpoints and connections between distinct genomic loci (juxtapositions). Although previous whole-genome sequencing analyses of modestly sized cohorts have revealed candidate non-coding regulatory driver events\(^8\)–\(^13\),

---

The list of affiliations appears at the end of the paper.

---

The discovery of drivers of cancer has traditionally focused on protein-coding genes\(^1\)–\(^4\). Here we present analyses of driver point mutations and structural variants in non-coding regions across 2,658 genomes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium\(^5\) of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). For point mutations, we developed a statistically rigorous strategy for combining significance levels from multiple methods of driver discovery that overcomes the limitations of individual methods. For structural variants, we present two methods of driver discovery, and identify regions that are significantly affected by recurrent breakpoints and recurrent somatic juxtapositions. Our analyses confirm previously reported drivers\(^6\)–\(^7\), raise doubts about others and identify novel candidates, including point mutations in the 5′ region of \(\text{TP53}\)\(^{12}\), in the 3′ untranslated regions of \(\text{NFKBIZ}\) and \(\text{TOB1}\), focal deletions in \(\text{BRD4}\) and rearrangements in the loci of \(\text{AKR1C}\) genes. We show that although point mutations and structural variants that drive cancer are less frequent in non-coding genes and regulatory sequences than in protein-coding genes, additional examples of these drivers will be found as more cancer genomes become available.

Previous large-scale sequencing projects have identified many putative cancer genes, but most efforts have concentrated on mutations and copy-number alterations in protein-coding genes, mainly using whole-exome sequencing and single-nucleotide polymorphism arrays\(^1\)–\(^4\). Whole-genome sequencing has made it possible to systematically survey non-coding regions for potential driver events, including single-nucleotide variants (SNVs), small insertions and deletions (indels) and larger structural variants. Whole-genome sequencing enables the precise localization of structural variant breakpoints and connections between distinct genomic loci (juxtapositions). Although previous whole-genome sequencing analyses of modestly sized cohorts have revealed candidate non-coding regulatory driver events\(^8\)–\(^13\),

---

The list of affiliations appears at the end of the paper.
the frequency and functional implications of these events remain understudied.\textsuperscript{11,13,16,17} Driver identification remains a far greater challenge in non-coding regions than in coding genes, owing to sequencing and mapping artefacts, poorly understood localized hypermutation processes\textsuperscript{14,18,19} incomplete annotation of regulatory regions, inaccurate estimation of the background mutation rate and the unknown functional effect of non-coding mutations. The discovery of drivers from structural variants is further complicated by their sparsity, the lack of obvious neutral events to build background models and their complex functional effects. Adequate statistical methods that address these issues are needed to reliably identify non-coding drivers.

The ICGC and TCGA PCAWG effort, which has collected and systematically analysed cancer genome sequences from 2,658 patients across 38 types of cancer\textsuperscript{1}, offers an opportunity to characterize putative non-coding driver events that cannot be found using data from whole-exome sequencing or single-nucleotide polymorphism arrays. Here we describe a comprehensive search for non-coding somatic drivers. For point mutations (SNVs and indels), we combine results from multiple driver-detection algorithms and, by carefully evaluating the significant hits, reveal that recurrent artefacts and poorly understood mutational processes have led to common false positives among previously reported non-coding drivers. For structural variants, we introduce two new methods for identifying both regions with significantly recurrent breakpoints (SRBs) and with significantly recurrent juxtapositions (SJs), accounting for genomic heterogeneity in the rates of DNA break and repair and the three-dimensional architecture of the genome. Finally, to assess the potential for future non-coding driver discoveries, we quantify our statistical power in the PCAWG dataset and estimate the overall excess of point mutations in non-coding regulatory regions around known cancer genes.

**Hotspot mutations across cancer types**

Many protein-coding driver mutations occur in single-site ‘hotspots’. In the PCAWG dataset, only 12 single-nucleotide positions were mutated in >1%, and 106 in >0.5%, of patients (Extended Data Fig. 1a, Methods). Although protein-coding regions span only about 1% of the genome, 15 out of 50 (30%) of the most frequently mutated sites were well-studied hotspots in cancer genes (KRAS, BRAF, PIK3CA, TP53 and IDH1) (Fig. 1a, Extended Data Fig. 1b), along with the two canonical TERT promoter hotspots\textsuperscript{2,3}.

The remaining non-coding hotspots could be attributed to the following localized mutational processes associated with passenger events: (i) damage from ultraviolet (UV) light and impaired DNA repair in melanoma at sites occupied by transcription factors\textsuperscript{5,18–20}; (ii) somatic hypermutation by activation-induced cytosine deaminase (AID) in B-cell non-Hodgkin lymphoma (Lymph–B NHL) and chronic lymphocytic leukaemia (Lymph–CLL); (iii) palindromic sequence contexts believed to form hairpin DNA structures targeted by APOBEC enzymes (in an intron of ADGRG6) and the PLEKHS1 promoter\textsuperscript{15}; and (iv) presumed technical artefacts (Fig. 1a, Supplementary Note 1). These findings suggest that—besides TERT promoter events—non-coding single-site hotspot drivers are infrequent or fall in regions with low sensitivity to detect mutations.

**Discovery of point-mutation drivers**

To identify recurrently mutated genomic elements, we first analysed somatic SNVs and indels in protein-coding regions, RNA genes (long and short non-coding RNAs and microRNAs (miRNAs)), and regulatory regions (promoters, 5’ untranslated regions (UTRs), 3’ UTRs and enhancers), totalling about 4% of the genome (Extended Data Fig. 2a–c, Methods, Supplementary Table 1). We analysed 2,583 tumours from 27 individual tumour types, and 15 meta-cohorts that grouped cancers by tissue of origin or organ system (Extended Data Fig. 2d, Methods). We identified candidate drivers—that is, cohort–element combinations with \( Q < 0.1 \) (10% false discovery rate (FDR))—by integrating 13 discovery algorithms, circumventing biases introduced by any one method (Extended Data Figs. 2e, 11, Supplementary Tables 2, 3, Supplementary Note 2). We benchmarked this approach by evaluating its ability to detect 603 known cancer genes (from the Cancer Gene Census (CGC))\textsuperscript{21}, v.80, and found that combining methods improved performance compared to single algorithms (Extended Data Fig. 3a, b, Methods). Overall, we identified 1,294 significant hits that involved 520 unique candidates (Supplementary Tables 4, 5).

**Filtering the significant hits**

Even after conservative FDR control, false-positive ‘driver’ loci can remain, owing to inaccurate background models, sequencing and mapping artefacts, or local increases in mutations due to unaccounted-for mutational processes. We therefore systematically filtered the candidate driver elements on the basis of technical and biological criteria, followed by careful review (Extended Data Fig. 3c, Methods, Supplementary Note 3). Examples of filtered elements include the promoters of PIM1 (lymphoid tumours) and ARPL134 (melanoma) because of associations with localized AID and UV-light mutational processes, respectively; PLEKHS1, GPR126, TBC1D12 and LEPROTL1 because of palindromic APOBEC target sequences\textsuperscript{15,18,20}; and the WDR74 5’ UTR and promoter\textsuperscript{15,18,20}, owing to mapping problems detected in downstream manual review (Supplementary Table 5, Supplementary Note 4). In combination, filtering and reapplying FDR control discarded 589 out of 1,294 (46%) of the original cohort–element hits and 341 out of 520 (66%) unique elements (Extended Data Fig. 3c, Supplementary Tables 4, 5).

**Candidate coding and non-coding drivers**

Our stringent combination and filtering strategy yielded 705 hits in 179 genomic elements: 602 hits in 143 protein-coding genes and 103 hits in non-coding elements. We observed wide variability across different types of cancer, from one hit in clear-cell renal cancer to 80 in the pan-cancer meta-cohort (Fig. 1b, Supplementary Tables 4, 5). Although most candidate drivers gained significance in larger meta-cohorts, some genes—such as DAXX (pancreatic endocrine tumour), NRAS (melanoma), SPOP (prostate adenocarcinoma), FGFR1 (pilocytic astrocytoma) and MIR142 (Lymph–B NHL)—scored higher in individual tumour types (Extended Data Fig. 3d). These results emphasize the trade-off between limiting driver discovery analyses to particular types of tumour and maximizing cohort size.

The candidate coding drivers we identified agreed with previous results: of the 143 genes that were significant in at least 1 cohort, 69% are in the CGC and nearly all have previously been implicated in cancer. In contrast to large whole-exome sequencing datasets, the fewer patients per cancer type in this dataset provided power sufficient only to detect genes with the strongest signal. We found 116 additional hits in 84 unique elements that were ‘near significance’ (\( 0.1 < Q < 0.25 \)). Fifty-one per cent of the 63 unique protein-coding genes in this set are in the CGC, which suggests that they would have been discovered in larger cohorts (Supplementary Table 4).

To nominate a significant non-coding element as a candidate driver, we reviewed the supporting evidence from the mutation calls, additional genomic data (chromosomal breakpoints, copy number, loss-of-heterozygosity and expression data), cancer gene databases and the literature (Methods, Supplementary Tables 6, 10). We describe the key candidates below, and in Supplementary Note 4.

The TERT promoter was the most frequently mutated non-coding driver in this dataset (14 cohorts) (Fig. 1b), and these mutations were strongly associated with higher TERT expression, as has previously been reported (Extended Data Fig. 4a, Supplementary Table 10).
Mutations in the promoter and/or 5' UTR of MTG2 (which encodes a GTPase involved in the mitochondrial ribosome) were associated with an expression of $MTG2$ that was marginally significantly lower, in both the pan-cancer ($P = 0.036$, fold difference $= 0.8$) and carcinoma ($P = 0.029$, fold difference $= 0.8$) meta-cohorts (Extended Data Figs. 4a, 5a). Mutations in the 5' UTR have previously been shown to decrease $MTG2$ expression in vitro. 

Recurrent somatic events were identified in the 3' UTRs of TOBI (carcinoma and pan-cancer meta-cohorts), NFKBIZ (lymphomas) and ALB (liver cancer) (Fig. 1b). TOBI encodes an anti-proliferation regulator
that associates with ERBB2, and also affects migration and invasion in gastric cancer. TOB1 regulates other miRNAs through binding to their 3′ UTR and promoting deadenylation. Tumours with 3′ UTR mutations showed a trend towards decreased expression (P = 0.053, fold difference = 0.7). The mutations did not concentrate in known miRNA-binding sites; however, the region is extremely conserved and thus probably functional (Fig. 2a). TOB1 and its neighbouring gene WFIKK2 are focally amplified in breast cancer and pan-cancer, suggesting a complex role in cancer (Extended Data Fig. 4b). \( \text{NFKBIZ} \) is a transcription factor that is mutated in diffuse large B cell lymphoma and amplified in primary lymphomas. Mutations in the 3′ UTR accumulated in a hotspot proximal to the stop codon and upstream of conserved miRNA-binding sites (Extended Data Fig. 5b). The enrichment of indels next to the stop codon suggests that this hotspot is not due to AID off-target activity. Previous functional experiments have shown these mutations with increased \( \text{NFKBIZ} \) expression, which we observed in our lymphoma cohort (P = 0.035, fold difference = 3.2; after correction for copy number, P = 0.03) (Extended Data Fig. 5b).

Both the exon and promoter of the non-coding RNA \( \text{RMRP} \) were significantly mutated in multiple types of cancer (Fig. 1b, Extended Data Fig. 5c). Germline \( \text{RMRP} \) mutations cause cartilage–hair hypoplasia, and previous in vitro studies have shown that some somatic promoter mutations are functional. The \( \text{RMRP} \) locus is also focally amplified in several types of tumour (Extended Data Fig. 4b). The enrichment of mutations in sites that can affect secondary structure suggests that these mutations are functional (P = 0.011, permutation test) (Extended Data Fig. 5c), although caution is required because this locus also appears to be affected by mapping artefacts or increased mutation rates (Supplementary Note 4).

The miR-142 precursor miRNA was significant in Lymph–BNHL and the lymphatic and haematopoietic cohorts (Fig. 1b; Extended Data Fig. 5d). The locus is a known AID off-target region in lymphoma, but 7 out of 8 mutations in the mature miRNA \( \text{mir-142-p3} \) — for which the largest functional effect is expected — were not assigned to AID, which suggests that these mutations are under selection.

**Unbiased genome-wide driver screen**

To test whether we missed drivers by focusing on functionally annotated regions, we applied an unbiased genome-wide survey to all non-overlapping 2-kb windows for excess point mutations. Twenty-two of the resulting 67 significant windows overlap with known protein-coding drivers, and 28 overlap highly transcribed regions with an excess of 2–5-bp indels (described in the ‘Transcription-associated indel signature’ section below) (Extended Data Fig. 5e, Supplementary Table 9, Supplementary Note 5). The remaining 17 windows have no obvious link to cancer, and several appear to be affected by mapping artefacts. A separate analysis of 4,351 ultra-conserved non-coding regions did not yield new candidate drivers (Extended Data Fig. 5e, Supplementary Note 5). Both screens suggest that the paucity of non-coding point-mutation drivers found in this study is not due to the annotation of functional elements.
**Increasing power for known cancer genes**

Finally, we performed restricted hypothesis testing to boost the statistical power to detect cis-regulatory driver mutations near cancer genes from the CGC\(^{39}\) (Supplementary Table 7). Restricted hypothesis testing of cancer gene promoters revealed a significant recurrence of TP53 promoter mutations (11 patients in pan-cancer, \(Q < 0.044\), mostly comprising SNVs and deletions that affect the transcription start site or donor splice site of the first non-coding exon. In 10 out of 11 cases, the mutation occurred in combination with loss-of-heterozygosity, and all samples with expression data showed decreased mRNA levels (Fig. 2b). None of these patients contained additional coding mutations that could instead be responsible for the downregulation of TP53. To our knowledge, this is the first report of a relatively infrequent—but impactful—form of TP53 inactivation by non-coding mutations.

Focal gains or losses in cancer are selected for modulating expression levels of their target genes. Restricting the hypothesis testing to the non-coding elements of such genes (\(n = 216,986\) cohort—element combinations, representing 5,201 unique elements) (Methods) yielded only one new hit, the 3′ UTR of the oncogene FOXA1 in prostate cancer (Supplementary Table 11).

**Transcription-associated indel signature**

Several significant non-coding elements (the ALB 3′ UTR, NEAT1, MALAT1 and MIR1222) were hit by many indels; all have previously been reported to be mutated in cancer\(^{27,33,37,38}\) (Figs. 1b, 2c). To explore whether ALB 3′ UTR events are under selection, we calculated indel rates across the functional regions of this gene. The indel rate is notably high throughout the UTRs, introns and exons, and even downstream of the polyadenylation site—a pattern inconsistent with selection (Fig. 2c, d). Similarly, FOXA1 has high indel rates throughout its locus, whereas the indels in NFKBIZ and TOBI are in their 3′ UTRs, suggesting that these are driver events (Fig. 2d). ALB, NEAT1 and MALAT1 mutations were not associated with changes in gene expression (Extended Data Fig. 4a) and were not associated with high cancer cell fractions or biallelic loss (Extended Data Fig. 6a, b). Likewise, indels in MIR1222 were downstream of the mature miRNA, and were not associated with altered expression of the targets of this miRNA (Supplementary Note 5).

If the indels in these genes were due to a mutational process rather than selection, they might exhibit distinct features. Indeed, indels in NEAT1, MALAT1, MIR1222 and ALB were strongly enriched in 2–5 bp-long events (Fisher’s \(P = 6.8 \times 10^{-5}\), for all) (Fig. 2e). A systematic search of coding and non-coding genes with significantly (\(Q < 0.1\)) increased indel rates of 2–5 bp indels revealed that this mutational process affects at least 18 additional genes in different types of tumour, most of which are highly expressed and tissue-specific (as has previously been reported for some of these genes\(^3\)) (Extended Data Fig. 6e, f). Although less enriched, SNVs also occur at high frequencies in these regions (Fig. 2f).

Overall, our findings suggest that the indels in MALAT1, NEAT1, ALB and MIR1222 are not driver events and are the result of a transcription-associated mutational process. The previously reported oncogenic effect of altered MALAT1 and NEAT1 expression\(^{27,38}\) may thus be unrelated to these mutations. Our findings also suggest that although FOXA1 protein-coding indels are drivers, 3′ UTR indels might be passengers\(^{30}\).

**Breakpoints at driver and fragile sites**

Driver structural variants may act by disrupting one or both of their breakpoint loci (for example, deactivate a tumour suppressor), or by generating a novel juxtaposition between loci. We thus searched both for genomic regions with SRBs and for pairs of regions with SRJs (Extended Data Fig. 7).

For SRBs, we first defined a background model to predict breakpoint density, using eight explanatory variables (Methods, Supplementary Table 13) and accounting for unexplained sources of variation\(^{14}\) (Supplementary Note 6). We identified 53 disjoint regions with SRBs (\(Q < 0.1\)) (Fig. 3a, Supplementary Table 14), which cleanly divided into two groups on the basis of the variability of the breakpoints at the other side of the rearrangements. Eight SRBs had partner breakpoints that were tightly clustered (had low rearrangement dispersion scores; Methods) and represented known oncogenic fusions. The remaining 45 SRBs had dispersed partner breakpoints (had high rearrangement dispersion scores), and were largely associated with previously identified somatic copy-number alterations (SCNAs) (Fig. 3b).

It has been difficult to distinguish recurrent driver SCNAs from passenger events at fragile sites\(^{32}\). At the resolution afforded by whole-genome sequencing, late replication timing predicted fragility-associated SRBs better than existing fragile site annotations (Supplementary Note 7), identifying 12 fragile-like SRBs (Fig. 3b). The remaining 33 SCNA-like SRBs comprised 14 amplifications, 8 deletions and 11 copy-neutral events (Supplementary Table 14).

The different classes of SRB were associated with different effects on neighbouring genes. Five of the eight deletion-associated SRBs were associated with biallelic inactivation of nearby known tumour suppressors, compared to none of the 12 fragile-like SRBs (\(P = 0.039\)) (Extended Data Fig. 8a). The fragile-like SRBs were furthest from tissue-matched enhancers and caused the weakest expression changes, consistent with them being passenger events\(^{31}\). By contrast, fusion-like SRBs were closer to tissue-matched enhancers than the other SRBs (\(P < 0.01\)) (Extended Data Fig. 8b) and were associated with greater changes in expression than all other SRBs except amplifications (\(P < 0.05\) for all types) (Extended Data Fig. 8c, Methods). Our analyses indicate that SRB driver events can be classified using rearrangement dispersion scores, replication timing and gene expression. Notably, neither rearrangement dispersion scores nor association with replication time can be accurately determined from microarrays or whole-exome sequencing, which highlights the importance of whole-genome sequencing. Altogether, we identified SRBs at 34 sites of known oncogenic fusions and recurrent SCNAs, 5 additional sites that are probably due to DNA fragility and 14 novel driver candidates (Supplementary Note 8).

**Novel structural-variant driver candidates**

Although most SCNA-like SRBs act by altering gene copy numbers, several appeared to target regulatory elements. We identified three that were significantly (\(Q < 0.05\)) associated with expression changes of nearby genes after controlling for copy number (Methods), two of which we discuss here. The first comprised structural variants at 10p15, which were associated with a greater than twofold upregulation of AKR1C1, AKR1C2 and AKR1C3 in seven cases of lung squamous cell carcinoma and two cases of liver hepatocellular carcinoma (Extended Data Fig. 8d). AKR1C proteins are aldo-keto reductases involved in steroid homeostasis. Ecotopic expression transforms cell lines, and germline mutations have previously been linked to an increased risk of developing lung cancer\(^{11,13}\). Three-quarters of the breakpoints are near (<10 kb) lineage-specific enhancers, potentially altering promoter–enhancer interactions (and hence gene expression). However, because the highest density of breakpoints lies between two long inverted repeats, the structural variants may have been induced by DNA secondary structure.

The second SRB contains recurrent microdeletions (<50 kb) involving the S′ end of BRD4 in ovarian (eight cases, \(P < 10^{-5}\)) and breast tumours (six cases, \(P = 0.04\)) (Fig. 3c, Extended Data Fig. 8e). These deletions were highly enriched in cancers that amplified a segment that includes BRD4 and NOTCH3 (\(P < 0.004\)) (Fig. 3d, Extended Data Fig. 8f) but were not a direct consequence of these amplifications (Supplementary Note 9). BRD4 is a chromatin regulator and a therapeutic target in several types of cancer\(^{25,38}\) including ovarian and triple-negative breast cancer\(^{25,38}\). Given the increased copy number of the full BRD4 gene, we would expect increased gene expression. However, the microdeletions
**Fig. 3** Significantly recurrent breakpoints and juxtapositions. **a**, Relative enrichment (Fisher’s exact test) for events per tumour type for the 20 most-significant SRBs (circle size). Loci are labelled by the likely driver gene from the CGC3. For gene symbols separated by a solidus, both or either of the genes are intended. B, Rearrangement dispersion score versus mean replication timing. Expression at the primary locus in samples with the rearrangement relative to rearranged samples; the secondary locus (right) is the other breakpoint. The tissue-specific expression at the secondary locus in wild-type (WT) samples, relative to samples of different tissue types, is greater for SRJs than other rearrangements (right). P values represent comparisons to ‘not in SRJ’. **d**, Gene expression per absolute copy number for SRJs (n = 200) and patients (n = 165). **e**, Greater change in expression. **f**, Greater change in expression. **g**, Greater change in expression. **h**, Greater change in expression. **i**, Greater change in expression. **j**, Greater change in expression. **k**, Greater change in expression. **l**, Greater change in expression. **m**, Greater change in expression. **n**, Greater change in expression. **o**, Greater change in expression. **p**, Greater change in expression. **q**, Greater change in expression. **r**, Greater change in expression. **s**, Greater change in expression. **t**, Greater change in expression. **u**, Greater change in expression. **v**, Greater change in expression. **w**, Greater change in expression. **x**, Greater change in expression. **y**, Greater change in expression. **z**, Greater change in expression. **aa**, Greater change in expression. **bb**, Greater change in expression. **cc**, Greater change in expression. **dd**, Greater change in expression. **ee**, Greater change in expression. **ff**, Greater change in expression. **gg**, Greater change in expression. **hh**, Greater change in expression. **ii**, Greater change in expression. **jj**, Greater change in expression. **kk**, Greater change in expression. **ll**, Greater change in expression. **mm**, Greater change in expression. **nn**, Greater change in expression. **oo**, Greater change in expression. **pp**, Greater change in expression. **qq**, Greater change in expression. **rr**, Greater change in expression. **ss**, Greater change in expression. **tt**, Greater change in expression. **uu**, Greater change in expression. **vv**, Greater change in expression. **ww**, Greater change in expression. **xx**, Greater change in expression. **yy**, Greater change in expression. **zz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Greater change in expression. **ooo**, Greater change in expression. **ppp**, Greater change in expression. **qqq**, Greater change in expression. **rrr**, Greater change in expression. **sss**, Greater change in expression. **ttt**, Greater change in expression. **uuu**, Greater change in expression. **vvv**, Greater change in expression. **www**, Greater change in expression. **xxx**, Greater change in expression. **yyy**, Greater change in expression. **zzz**, Greater change in expression. **aaa**, Greater change in expression. **bbb**, Greater change in expression. **ccc**, Greater change in expression. **ddd**, Greater change in expression. **eee**, Greater change in expression. **fff**, Greater change in expression. **ggg**, Greater change in expression. **hhh**, Greater change in expression. **iii**, Greater change in expression. **jjj**, Greater change in expression. **kkk**, Greater change in expression. **lll**, Greater change in expression. **mmm**, Greater change in expression. **nnn**, Great
To our knowledge, this is the first evidence of a recurrent microdeletion limiting expression of an amplified gene.

**Recurrent fusions target gene regulation**

Motivated by the detection of fusion-like SRBs, we specifically looked for genomic loci that were juxtaposed more often than expected by chance, after controlling for both the rate of breakpoints at each locus and the distance between them (Methods). We identified 90 such SRJs (Fig. 3f, Supplementary Table 15), including 13 known oncogenic fusions (including all 5 fusion-like SRBs) and 77 novel hits—18 of which linked to at least one known cancer gene (Supplementary Note 8). Previously reported oncogenic SRJs were observed more frequently (average 24 patients per fusion, range 2–98) than novel ones (most often 2 patients per fusion, range 2–4). As juxtapositions are unlikely to occur by chance, observing even two becomes highly significant. However, it is possible that some SRJs reflect inaccuracies in our background model rather than true drivers. We therefore further evaluated the SRJs on the basis of (i) a ‘robustness factor’ that indicates how much the background rate could increase before the SRJ would become insignificant, and (ii) the ratio between the observed and expected numbers of events under the current background model (‘effect size’) (Extended Data Fig. 9a). Twenty-six SRJs, including 11 of the 13 known drivers and 15 newly identified SRJs, are robust to tripling the expected background rate, and 22 others would remain significant with a doubled rate.

Most canonical driver rearrangements have previously been found in single tumour types, often associated with tissue-specific expression. We found that 9 of our top 10 SRJs are tissue-specific, despite searching across 30 different types of tumour. Such tissue specificity is not observed for cancer genes affected by SCNAS, for which the top 10 are altered in 11.9 cancer types (on average), or by point mutations (for which the top 10 are altered in 6.7 cancer types, on average) (Supplementary Table 16).

The tissue specificity of SRJs suggests that they are strongly shaped by epigenetic state, either owing to mechanistic reasons (for example, tissue-specific three-dimensional proximity of the two DNA breakpoints) or to selection that connects tissue-specific regulatory elements with oncogenes. The latter seems to be more likely because: (i) SRJs are associated with significant overexpression of only one of the rearrangement partners (the ‘primary locus’) relative to randomly selected rearrangements (primary locus, \( P < 10^{-8} \)) (Fig. 3g left); secondary locus, \( P > 0.05 \) (Extended Data Fig. 9b left)); (ii) the rearrangement partner, in the secondary locus, tends to be highly expressed in that tissue type relative to others (Fig. 3g right); and (iii) the distance to the nearest tissue-specific enhancer is smaller for SRJs than for rearrangements overall (Extended Data Fig. 9b). These observations suggest that SRJs act in general by bringing regulatory elements to an oncogene that is otherwise expressed at a low level.

In many cases, SRJs generate truncated or chimeric proteins, and breakpoints within introns or exons were indeed overrepresented (68% versus 56% expected, \( P < 10^{-3} \)). However, only 11 of the 30 (37%) most significant SRJs generated novel proteins in all samples, and 6 others sometimes generated novel proteins; the rest were either non-disruptive or contained breakpoints within the first two introns of the disrupted gene, leaving most of the protein intact (Fig. 3f). Moreover, SRJs that generate novel proteins exhibited expression changes similar to those that do not \( (P = 0.4) \) (Extended Data Fig. 9c). We conclude that altering gene expression is a key function of both classes of SRJs, and that SRJs are akin to non-coding driver point mutations that act on regulatory elements.

We found several SRJs that involve amplified oncogenes, including MDM2, EGRF and TERT (Fig. 3f, h, i, Extended Data Fig. 9d–f, Supplementary Table 15). The TERT promoter region was juxtaposed in four melanomas \( (P < 10^{-3}) \) to a region in the BASP1 gene (both on chromosome 5), and to a region near NDUFC2 \( (t(5,11)) \) in two melanomas and one medulloblastoma \( (P < 10^{-5}) \). Both juxtaposed regions were marked with melanocyte enhancers, which suggests that they could drive TERT expression. Among melanomas, these rearrangements are mutually exclusive with the C22ST and C250T mutations of the TERT promoter \( (P < 10^{-3}) \) (Fig. 3h). Because the juxtapositions were always part of complex events that also amplified TERT, increased TERT expression may be due to amplification, the juxtapositions or both.

**Paucity of non-coding drivers in cancer**

Our analyses of genomic hotspots, functional elements, genomic windows and SRJs all suggest that non-coding drivers are rare compared to protein-coding drivers. This might, in part, be due to a lack of discovery power. We therefore evaluated the discovery power of mutational-burden tests for recurrent events across the different types of element in our tumour cohorts, focusing first on point mutations.

We found that the fraction of mutated patients required for a driver to reach 90% discovery power ranged from <1% in large cohorts with low background-mutation densities to 25% in small cohorts with high background-mutation densities (Fig. 4a). Different types of element were similarly powered, suggesting that the paucity of drivers in non-coding versus coding elements is not due to a lack of power. Similarly, our power to detect SRJs was higher in large cohorts with low rearrangement rates, and for long and interchromosomal rearrangements owing to their lower overall rates (Extended Data Fig. 10a): we were only powered to detect events that recur in 5–20% of samples in most types of cancer (Fig. 4b). Moreover, beginning with about 2,500 tumours, we expect to find a new SRJ with every 25 additional genomes.

Low sequencing coverage (for example, in GC-rich regions) also limits driver discovery. To measure this effect in the PCAWG data, we quantified our ability to detect mutations (detection sensitivity) in cancer gene promoters. Although the mean detection sensitivity in promoters is high (41.9% of genomic positions have mean detection sensitivity >80% across tumours), only 4.1% of the promoters had detection sensitivity >90% in >90% of bases. In particular, the two canonical TERT promoter hotspots had highly variable detection sensitivity: among patients and cohorts, from only 3% of patients in the central-nervous-system pilocytic astrocytoma cohort to 100% in the thyroid adenocarcinoma cohort (Extended Data Fig. 10b). From these data, we inferred the expected number of TERT events in each tumour type (Extended Data Fig. 10c) and found that about 263 (95% confidence interval 232–295) TERT hotspot mutations were probably missed owing to a lack of detection sensitivity. Moreover, on average 9.9% (1.3–13.0% interquartile range) of the cancer gene promoter territory in the tumour of each patient was severely underpowered (an average detection sensitivity of <10%). Therefore, the lack of coverage in promoters may contribute to the paucity of non-coding drivers.

To determine whether the paucity of non-coding drivers discovered thus far could be due to the limited statistical power of current datasets, we estimated the overall excess of point mutations above background (that is, the expected number of driver events) in coding and cis-regulatory non-coding sequences in 603 cancer genes (Methods, Supplementary Table 7, Supplementary Note 11). To minimize the effect of samples with low detection sensitivity, we included only 936 samples with >90% detection sensitivity at the two TERT promoter hotspots (Extended Data Fig. 10c, d, Supplementary Note 11). Overall, this approach predicted more than 1,475 driver mutations (95% confidence interval 1,410–1,687; 1,069 SNVs and 406 indels) in the protein-coding sequences of these cancer genes (Fig. 4d), compared to only 96 (95% confidence interval 30–190) estimated driver mutations in promoters (73 attributed to TERT, 22 (95% confidence interval 0–88) in 5′UTRs, and 68 (95% confidence interval 0–178) in 3′UTRs. Non-coding mutations in cancer-gene promoters were also not generally associated with loss-of-heterozygosity or altered expression, as one would expect if they were enriched with drivers (Supplementary Note 12).
MALAT1 coding RNAs, which contain a high density of indels, found that several previously reported and frequently altered non-coding RNAs, NEAT1 and MALAT1, have critical implications for patients with cancer. Our findings and methods introduced here for the discovery of point-mutation and structural-variant drivers, method integration, vetting of candidates and non-coding mutations and (ii) the target size of functional elements, the paucity of non-coding drivers contrasts with the distribution of germ-line polymorphisms associated with heritability of complex traits, which are most frequently located outside of protein-coding genes. These results collectively indicate that, independently of statistical power, non-coding cis-regulatory driver mutations in known cancer genes besides TERT are much less frequent than protein-coding drivers.

**Discussion**

The accurate and reliable discovery of genomic drivers in tumours may have critical implications for patients with cancer. Our findings and the methods introduced here for the discovery of point-mutation and structural-variant drivers, method integration, vetting of candidates and identification of local hypermutation and fragile sites represent an important contribution to the collective effort towards charting all malignant changes that drive the cancer of each patient.

Among the most interesting candidate non-coding driver elements we uncovered are the 5’-end mutations in TPS3; 3’ UTR mutations in NFKBIZ and TOB1; and rearrangements involving AKRIC genes and BRD4. By careful analysis of the whole-genome sequencing data, we found that several previously reported and frequently altered non-coding elements may not be genuine drivers, including (i) the non-coding RNAs, NEAT1 and MALAT1 (which contain a high density of indels, seemingly owing to a transcription-associated mutational process) and (ii) recurrent structural variants in regions of late replication, indicating DNA fragility.

This study yielded unexpectedly few non-coding driver point mutations and structural variants. SRJs, which appear to act largely through the rearrangement of regulatory elements, are less frequent than SCNA-like SRBs, which directly amplify or delete coding sequences. The results from five analyses—hotspot recurrence, driver-element discovery, structural variants, discovery power and aggregated mutational excess—suggest that this paucity is not caused by a particular analysis strategy, but that regulatory elements truly contribute a much smaller number of recurrent cancer-driving events than protein-coding sequences. This paucity of non-coding drivers contrasts with the distribution of germ-line polymorphisms associated with heritability of complex traits, which are most frequently located outside of protein-coding genes.

At least two factors contribute to the relative paucity of non-coding driver mutations in cancer: (i) the differential fitness effects of coding and non-coding mutations and (ii) the target size of functional elements. The paucity of promoter driver mutations in well-established
cancer genes suggests that point mutations markedly affect the function of non-coding regulatory elements only rarely. This highlights TERT as a notable exception, perhaps because even a modest increase in TERT expression may suffice to circumvent normal telomere shortening. For other cancer genes, directly mutating protein-coding sequences or altering expression levels by copy-number change may provide larger phenotypic effects. For example, complete loss-of-function by nonsense mutations or deletions may be easier to achieve than by disrupting or translocating regulatory regions.

Technical shortcomings (such as coverage ‘blind spots’ in GC-rich promoters and different filtering strategies) may cause genuine drivers to be missed. Therefore, the discovery of non-coding drivers will benefit from technical improvements, including even sequence coverage, longer and accurate reads, and improved variant-calling methods. Moreover, better annotation of functional non-coding elements will increase both the power to discover infrequently mutated driver elements and their interpretability. As datasets grow, yet-identified mutational mechanisms targeting particular genomic regions will emerge and require improved background models, including additional covariates and more-sophisticated statistical models. The analysis of structural variants has greater challenges because (i) accurately modeling their background density is complicated by their lower frequency and larger fraction of drivers (Supplementary Note 6); (ii) their target genes may be far from the breakpoints, as in SCNAs; (iii) the space for modeling SRJs is much larger (the genome squared); and (iv) many structural variants are part of complex events that often involve multiple chromosomes, so that the resultant topology cannot be deduced without technologies such as long- or linked-read sequencing. For these reasons, experimental validation remains important for all—and especially for non-coding—candidate drivers.

Our work suggests that larger datasets and technological advances will continue to identify new non-coding drivers, albeit at considerably lower frequencies than protein-coding drivers. We anticipate that the approaches developed here will provide a solid foundation for the incipient era of driver discovery from ever-larger numbers of cancer whole genomes.

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/s41586-019-1913-9.
50. Bishara, A. et al. Read clouds uncover variation in complex regions of the human genome. Genome Res. 25, 1570–1580 (2015).
51. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).

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, corrected publication 2023
**Methods**

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Detailed methods are provided as Supplementary Methods.

**Dataset generation**

Out of 2,955 samples, we selected 2,583 unique donor samples for SNV and indel driver-discovery analysis on the basis of SNV quality control (Supplementary Methods). We found that 110 additional myeloid–AML samples had robust structural variant calls despite SNV artefacts; we included these in structural variant analyses, for a total of 2,693 samples. For tumour-type cohort analyses, we used only cohorts with at least 20 patients. Tumour meta-cohorts were defined by cell type of origin or by organ system (for example, lung for lung adenocarcinoma and lung squamous cell carcinoma). A pan-cancer meta-cohort was created by combining all tumour cohorts except for Skin–Melanoma and lymphoid tumours (Supplementary Methods).

**Hotspot SNV analysis**

We selected the 50 most-frequent SNV hotspots. These were analysed to identify known driver events; mutational signature biases related to sequence palindromes, immunoglobulin loci and so on; and potential artefacts, including regional mapping problems (Supplementary Methods).

**Mutational signatures**

We performed de novo global-signature discovery and signature attributions with SignatureAnalyzer’s Bayesian non-negative matrix factorization method, based on 1,697 channels—including 1,536 pentanucleotide sequence contexts for single-base substitutions, 83 indel features, and 78 doublet-nucleotide substitution classes (Supplementary Methods).

**Definition of genomic elements**

GENCODE v.19 (ref.53) and other genomic resources were used to define functional genomic elements, including protein-coding genes (CDS, splice sites, 5’ UTR, 3’ UTR and promoters), long non-coding RNAs (gene body, splice site and promoters), short RNAs, miRNAs and enhancers (Supplementary Methods).

**Candidate-driver-mutation identification methods and combination of results**

We obtained results (P values) from 13 methods of driver discovery, including ActiveDriverWGS54, CompositeDriver, DriverPower55, dndscv56, ExhInitor57, LARVA57, MutSig tools57, NBR58, ncdDetect58, ncDriver59, OncodriveFML60 and regDriver61. We integrated the results of all these methods using a custom framework based on a previously published method for combining P values. Results from individual methods that showed large deviations from the expected uniform null distribution of P values were excluded. This approach was evaluated on real and simulated data. We controlled the FDR within each of the sets of tested genomic elements by concatenating all combined Brown’s P values from across all tumour-type cohorts and applying the Benjamini–Hochberg procedure. Cohort–element combinations with Q values < 0.1 were designated as significant hits, and combinations with 0.1 ≤ Q < 0.25 as ‘near significance’. Extensive details are provided in the Supplementary Methods. In addition, we tested for element-independent recurrence with the NBR method on 2-kb bins spanning the entire genome, and non-coding ultraconserved regions.

**Post-filtering of driver mutation candidates**

We applied stringent filters to discern positive selection from technical artefacts and mutational processes. We required at least three mutations to be present in candidate elements, in at least three patients of the tested cohort; more than 50% of mutations in mappable regions; less than 50% of mutations in palindromic DNA; and less than 50% of mutations attributed to APOBEC activity. For lymphoid tumours and skin melanoma, we required that <35% and <50% of mutations were attributed to the AID and UV-light mutational signatures, respectively. The FDR was recalculated after post-filtering.

**Candidate driver structural-variant analyses**

We applied separate analyses to detect recurrent structural variant breakpoints and recurrent juxtapositions. For each analysis, we first binned breakpoints, accepting only one breakpoint per sample per bin. We then determined which bins had more breakpoints than expected by chance (the SRB analysis), and which pairs of bins (or ‘tiles’) were joined by more rearrangements than expected by chance (the SRJ analysis).

**Candidate driver breakpoints**

We calculated the background rate of breakpoints per bin based on a Gamma–Poisson model that took into account genomic covariates, breakpoint counts normalized by the number of bases within each bin that had sufficient mappability to be eligible for breakpoint detection and accounted for an observed overdispersion of breakpoint counts that probably reflects unaccounted-for covariates (Supplementary Methods). We used the Gamma–Poisson model to calculate the P value for each bin (that is, the probability that each bin would exhibit the observed number of breakpoints (or greater) by chance alone), applying the Benjamini–Hochberg procedure to correct for multiple hypotheses.

**Post-filtering of driver breakpoint candidates**

We scored each recurrent breakpoint locus on the basis of the average replication timing of its breakpoints, and filtered those loci with scores >0.5 as probable fragile sites.

**Candidate driver juxtapositions**

We developed a background model to indicate the probability that two loci would be joined, taking into account the observed rate at which each locus underwent DNA breaks (from the breakpoint analysis), the distance between them and the propensity for these rearrangements to reflect a break followed by invasion versus two breaks that were then joined. We determined the probability that each tile would contain the observed number of rearrangements using a binomial test, followed by controlling for multiple hypothesis testing using the Benjamini–Hochberg procedure.

**Gene-expression analyses**

Gene-expression data were provided by the PCAWG Transcriptome Core Group, and also generated using the same approach for an extended set of non-coding transcripts (Supplementary Methods).

**Additional evidence for selection**

In addition to associations between mutations or structural variants and expression, we looked for signals of copy-number-alteration recurrence using the GISTIC2 algorithm. We also tested whether driver candidates showed significantly higher frequency of loss-of-heterozygosity in mutated samples using Fisher’s exact test. We calculated cancer allelic fractions using ploidy and tumour purity predictions from a previous publication.

**Mutational process and indel enrichment**

For every gene, we calculated the proportion of indels of length 2–5 bp out of the total number of indels. This proportion was compared to the genome background proportion using a binomial test. We also compared the indel rate per gene (not distinguishing by length) to the background. Both sets of P values were corrected with the FDR method.
Power calculations
We estimated our power to discover driver elements mutated at a particular frequency in the population as previously described\textsuperscript{1,2,3}, but solving for the lowest frequency for a driver element in the patient population that is powered (≥90%) for discovery. The calculation of this lowest frequency takes into account (i) the average background mutation frequencies for each cohort–element combination; (ii) the median length and average detection sensitivity for each element type and patient cohort size; and (iii) a global desired false-positive rate of 10%.

The effect of element length is discussed in Supplementary Note 10, and details are provided in Supplementary Methods. Power calculations for detection of recurrent juxtapositions was performed similarly, except over a two-dimensional genomic fusion map divided into 100 × 100-kb tiles (Supplementary Methods). We performed this analysis first as a function of the distance between breakpoints (Extended Data Fig. 10a) and second as a function of the median number of rearrangements per sample, spanning values represented by histologies with more than 15 samples (Fig. 4b).

Estimation of the number of mutations in non-coding regions of known cancer genes
NBR was used to estimate the background mutation rate expected across cancer genes, using a conservative list of 19,082 putative passenger genes as background and including as covariates the local mutation rate, gene expression and averaged copy-number states. The resulting model predicted the number of passenger SVNs and indels expected by chance. By aggregating the expected numbers over 603 known cancer genes from the CGC\textsuperscript{64} (CGC v.80) (Supplementary Table 7), we compared the observed and expected numbers of mutations. For this analysis, we excluded samples with problems of low detection sensitivity (Supplementary Methods).

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

Data availability
Data associated with this Article are available at https://dcc.icgc.org/releases/PCAWG/drivers. SRBs and SRJs are available at www.svscape.org. A list of data files used for analyses in this paper is provided in Supplementary Table 20. Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC and TCGA PCWG Consortium are described in an accompanying Article\textsuperscript{6}, and 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 that does not require access approval. To access information that could potentially identify participants, such as germline alleles and the underlying sequencing data, researchers will need to apply to the TCGA data access committee 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 (http://icgc.org/daco) for the ICGC portion of the dataset. In addition, to access somatic single-nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

Code availability
The core computational pipelines used by the PCWG Consortium for alignment, quality control and variant calling are available to the public at https://dockstore.org/search?search=pcawg under the GNU General Public License v.3.0, which allows for reuse and distribution. Code for P-value combination from multiple driver methods is available from https://github.com/broadinstitute/getzlab-PCAWG-pvalue_combination/. Power calculation methods are available from https://github.com/broadinstitute/getzlab-PCAWG-power_calculations/. Structural variant methods are located at https://github.com/msklab/fishhook, https://github.com/walaj/gingseong and https://github.com/walaj/VSSig. Links to individual driver discovery methods are provided in the corresponding section of the Supplementary Methods.

52. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytokine deaminase signatures during indolent chronic lymphocytic leukemia evolution. Nat. Commun. 6, 8866 (2015).
53. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE Genome Project. Genome Res. 22, 1760–1774 (2012).
54. Wadi, L., Uusula-Reimand, L., Iseve, K. & Shuai, S. Candidate cancer driver mutations in super-enhancers and long-range chromatin interaction networks. Preprint at https://www.biorxiv.org/content/10.1101/236802v1 (2017).
55. Shuai, S., Gallinger, S. & Stein, L. DriverPower: combined burden and functional impact tests for cancer driver discovery. Preprint at https://www.biorxiv.org/content/10.1101/215244v1 (2017).
56. Lanos, A. et al. Discovery of cancer driver long noncoding RNAs across 112 tumour genomes: new candidates and distinguishing features. Sci. Rep. 7, 41544 (2017).
57. Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. & Gerstein, M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res. 43, 8123–8134 (2015).
58. Juel, M. et al. Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. eLife 6, e21778 (2017).
59. Hornshøj, H. et al. Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. NPJ Genom. Med. 3, 1 (2018).
60. Mularoni, L., Sabarianathan, R., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. 17, 128 (2016).
61. Umer, H. M. et al. A significant regulatory mutation burden at a high-affinity position of the CTCF motif in gastrointestinal cancers. Hum. Mutat. 37, 904–913 (2016).
62. Brown, M. B. 400: a method for combining non-independent, one-sided tests of significance. Biometrics 31, 987 (1975).
63. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 57, 289–300 (1995).
64. Dmitrieva, S. & Bucher, P. UCNEbase—a database of ultra-conserved non-coding elements and genomic regulatory blocks. Nucleic Acids Res. 41, D101–D109 (2013).
65. Mrasek, K. et al. Global screening and extended nomenclature for 230 aphidicolin-inducible fragile sites, including 61 yet unreported ones. Int. J. Oncol. 36, 929–940 (2010).
66. PCWG Transcriptome Core Group et al. Genomic basis for RNA alterations in cancer. Nature https://doi.org/10.1038/s41586-019-1907-0 (2020).
67. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 12, R41 (2011).
68. Gerstung, M. et al. The evolutionary history of 2,656 cancers. Nature https://doi.org/10.1038/s41586-019-0947-5 (2020).
69. Futreal, P. A. et al. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183 (2004).

Acknowledgements We thank the ICGC and TCGA PCWG Network and the PCWG steering committee, M. Meyerson and E. S. Lander for helpful feedback, and M. Miller for editing this manuscript. Work in the Getz laboratory was partially funded by the GDAC grants (NIH U24CA143845 and NIH U24CA210999), G.G.’s funds at the Broad Institute and MGH. G.G. is also partially supported by the Paul C. Zamecnik Chair in Oncology in MGH.

Author contributions This work was carried out by the PCWG Drivers and Functional Interpretation Group based on data from the ICGC and TCGA PCWG Network. All authors have had access to read and comment on the manuscript. E.R., M.M.N., F.A., J.A.W. and O.S. contributed equally. G.T., H.H., M.J.H. and R.L.J. contributed equally. J. Weischenfeldt, R.B., I.M., J.S.P. and G.G. jointly supervised this work. The full list of PCWG Consortium members and their affiliations appears in the Supplementary Information. For the discovery of...
point-mutation drivers, the following authors contributed: A.L., C. Hermann, C.W., D.A.W., E.K., E.M.L., E.R., G.G., G.S., G.T., H.M.U., I. Kim, J.R., J.S.P., K.A.B., K.D., K.I., L.M., L.U.-R., M.M.N., M.P.H., N.S.-A., N.J.H., P.D., R.J., S.B.A., T.A.J. and T.T. contributed and curated genomic annotations; C. Hermann, C.W.Y.C., I.M., M.K., S.B.A. and Y.E.M. contributed randomized mutational datasets for driver discovery; A.L., A.G.-P., A.H., D. Tamborero, E.K., E.M.L., E.R., H.H., I.M., J. Bertl, J.C.-F., J.M.H., J. Komorowski, J.R., J. Zhang, K.D., K.I., L.M., L.D.S., L.U.-R., L.W., M.B.G., M.J., N.L.-B., O.P., P.D., Q.G., R. Sabarinathan, S.K., S.S. and T.M. contributed driver methods and results. E.R., G.G., Z.L. and G.T. implemented results integration; A. Kahraman, C.W.Y.C., I.M., J.R., L.F. and M.M.N. contributed driver results integration; and C. Hermann, C.W.Y.C., E.K., E.R., G.G., J. Kim, J.M.H., J.S.P., M.M.N. and R.J. contributed single-site recurrence analysis. For the discovery of structural-variant drivers, the following authors contributed: J.A.W., O.S., Y.L., N.D.R., S.S.E., M.I. and J. Weischenfeldt contributed and curated genomic annotations; J.A.W., J.E.H., J.T., O.S., D. Craft, K.K., S.E.S., C. Stewart, C.-Z.Z., M.I., P.J.C., J. Weischenfeldt, X.Y. and R.B. contributed to the development of the structural variant recurrence analysis methods; J.A.W., O.S., K.K., J. Weischenfeldt and R.B. implemented structural variant recurrence analyses; and J.A.W., O.S., J. Busanovich, N.S., P.B., J. Weischenfeldt and R.B. integrated structural variant recurrence results with expression, chromatin organization and functional data. For point mutations candidate vetting and filtering, the following authors contributed: E.R., F.A., H.H., J. Kim, L.F., M.M.N. and T.M. contributed individual candidate filters; and A.L., A. Hong, C.W., E.K., E.M.L., E.R., F.A., G.G., G.T., H.H., H.M.U., I. Kim, J.M.H., J.S.P., K.D., L.F., L. Sieverling, M.M.N., M.S.L., N.S.-A., R.J., R.R. and T.M. performed candidate vetting. For candidate-based analysis, the following authors contributed: E.R., F.A., G.G., H.H., H.N., I.M., J.M.H., J.R., J.S.P., Keunchil Park, M.M.N. and M.P.H. contributed candidate-based analyses; A.G.-P., A.H., A.L., C. Hermann, D. Chakravarty, D. Tamborero, E.K., E.R., F.A., G.G., G.T., H.H., H.M.U., I.M., J.C.-F., J.R., J.S.P., K.I., Keunchil Park, L.M., L.D.S., L.U.-R., L.W., A. Rubin, M.B.G., M.M.N., M.S.L., N.S.-A., N.L.-B., O.P., R.I.J., R.Sabarinathan, S.K. and Y. Kim contributed results interpretation; A. Kahlan, J.S.P., K.A.B., K.V.L., M.M.N., N.A.F., S.B.A., T.A.J. and T.T. contributed expression profiling (extended GENCODE set); A. Kahraman, D. Chakravarty, J.R., J.S.P., K.I., M.A. Rubin, M.M.N., M.S.L., S.B.A. and T.M. contributed mutation-to-expression correlation analysis. A. Kahraman, D. Chakravarty, J.R., L.W., M.A. Rubin and N.S.-A. contributed network or pathway analysis; and R. Sabarinathan, C. Shen, C. Sander and J.S.P. contributed structural RNA analysis. For power analysis and driver mutations at known cancer genes, the following authors contributed: E.R. analysed SNV detection and driver discovery power; Z.L. evaluated sensitivity of methods; F.A. and I.M. contributed mutational excess analysis; M.M.N. integrated additional evidence; and O.S. analysed structural variant detection and driver discovery power. The following authors contributed leadership and organizational work: for point mutations, E.R., G.G. and J.S.P. contributed working group leadership; A.G.-P., B.J.R., D.A.W., E.K., E.R., G.G., G.T., I.M., J.R., J.M.S., J.S.P., L.F., L.M., M.B.G., N.L.-B., O.P., P.J.C., R. Sabarinathan and S.K. contributed organization; and E.R., M.M.N., F.A., G.G., H.H., H.M.U., I.M., J.S.P. and G.G. wrote the manuscript. For structural variants, P.J.C. and R.B. contributed working group leadership; J.A.W., Y.L., P.J.C., J. Weischenfeldt and R.B. contributed organization; and J.A.W., O.S., P.J.C., J. Weischenfeldt and R.B. wrote the manuscript.

Competing interests The following authors declare that they have competing interests. P.B. receives grant funding from Novartis from an unrelated project; R.B. owns equity in Ampressa Therapeutics and receives grant funding from Novartis; G.G. receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSUmutect, MSUmutSig and POLYSOLVER; B.J.R. is a consultant at and has ownership interest (including stock, patents and so on) in Medley Genomics; O.S. is currently an employee of Cedilla Therapeutics; and Y.L. is currently an employee of Seven Bridges Genomics.

Additional information Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-1965-x.

Correspondence and requests for materials should be addressed to J.Weischenfeldt, R.B., I.M., J.S.P. or G.G.

Peer review information Nature thanks Don Conrad, Fran Supek and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | Mutational hotspots in additional tumour types.

**a** Bar plot of number of positions (y axis) mutated in n patients (x axis). The stacked bar charts under the bar plot show the proportion of protein-coding (dark grey) and non-coding (light grey) positions.

**b** Distribution of SNVs in top 50 single-site hotspots across all analysed individual cohorts. Hotspots are grouped as known drivers or induced by mutational processes. The table (middle) shows the frequency of mutations across the PCAWG cohorts. Stacked bar chart (right) shows the contribution of mutational processes to the hotspot mutations (Methods). Gene names are given when hotspots overlap with functional elements (colour-coded), with amino acid alterations for protein-coding genes.
Extended Data Fig. 2 | See next page for caption.
**Extended Data Fig. 2 | Element-based driver discovery and combination of P values.**

**a**, Schematic describing definition of types of functional element (Methods). Functional elements (black) are defined on the basis of transcript annotations from various databases. Elements arising from multiple transcripts with the same gene identity are collapsed, as seen here for the protein-coding isoforms. Promoter elements are defined as 200 bases upstream and downstream of the transcription start sites of the transcripts of a gene (green). Splice site elements extend 6 and 20 bases from the 3′ and 5′ exonic ends into intronic regions, respectively (light blue). Regions overlapping protein-coding bases and protein-coding splice sites are subtracted from other regions. **b**, Percentage of genomic coverage for each element type. **c**, Distribution of element lengths for each element type. Thick lines indicate interquartile ranges and short horizontal bars indicate the medians. **d**, Organization of meta-cohorts defined by tissue of origin and organ system. Pan-cancer contains all cancers, excluding Skin–Melanoma and lymphoid malignancies. **e**, Combination workflow: overview of methods of driver discovery and their lines of evidence to evaluate candidate gene drivers. Methods using each feature are marked with a box in the appropriate track. Heat map displaying Spearman’s correlation of P values across the different driver-discovery algorithms based on simulated (null model) mutational data. Dendrogram illustrates the relatedness of method P values, and algorithm approaches are marked by coloured boxes on dendrogram leaves. Next, P values are combined with Brown’s method on the basis of the calculated correlation structure. Individual method (left) and integrated (right) log-transformed P values are shown in a heat map (grey, missing data). Post-filtering used several criteria to identify likely suspicious candidates. Significant driver candidates were identified after controlling for multiple hypothesis testing based on an FDR Q value threshold of 0.1 (blue asterisk). Candidates with Q values below 0.25 (blue dash) were also considered of interest.
Extended Data Fig. 3 | Sensitivity of driver-discovery methods and filter statistics. **a**, Percentage of coding-driver discovery runs (with stable F score, $n = 33$), across all cohorts, in which the method had the highest F score (Methods). **b**, F score of different methods of driver discovery, and different combinations evaluated in the four largest cohorts (pan-cancer ($n = 2,278$), carcinoma ($n = 1,856$), adenocarcinoma ($n = 1,631$) and digestive tract ($n = 797$)). Only methods that used the same algorithm to call coding and non-coding drivers were evaluated. Vertical lines indicate 95% confidence intervals. Horizontal black lines mark the median in each group. $P$ values were calculated with the two-sided non-parametric Mann–Whitney $U$ test. **c**, On top, the initial number of hits identified as recurrently mutated for each element type. The element types mature miRNA ($n = 2$ before filtering) and miRNA promoters ($n = 16$ before filtering) were omitted from the table. The heat map shows the number of hits filtered at each step in the sequential application of filters and post-filtering re-application of the FDR correction. Background colours indicate the corresponding percentage of input element removed. The final numbers of hits (including those that were later filtered by the comprehensive vetting procedures) are indicated below the heat map. **d**, Sensitivity versus specificity in individual cohorts versus meta-cohorts for candidate drivers: $Q$ values for the most significant individual cohort (x-axis) versus meta cohort (y-axis) are shown. Driver elements are coloured by their element type. $Q$ values derived from combination of $P$ values from individual driver discovery methods (Methods).
Extended Data Fig. 4 | Mutation-to-expression correlation and focal copy-number alterations. a, Expression is compared between mutated and non-mutated samples. For each element, the z score of the expression values for mutated and wild type in the significant cohort is plotted. For copy number, CNA amplification indicates CNA > 10; CNA gain indicates CNA ≥ 3; CNA loss indicates CNA ≤ 1; and no events indicates CNA < 3 and CNA > 1. If a patient is mutated with multiple types of point mutation, indels are indicated over SNVs. For TERT, only samples powered to call mutation status were used. P values are based on a two-sided Wilcoxon rank-sum test. Bars indicate means. b, Copy-number profiles of 55 of 441 stomach adenocarcinomas from TCGA show copy-number gains around HES1, TOB1 and its gene neighbour WIFKKN2 are focally amplified in cancer (172 of 10,844 total samples from 33 cancer types are shown). RMRP focal amplifications in TCGA cancers (160 of 10,844 total tumours shown).
Extended Data Fig. 5 | Non-coding driver candidates. a, MTG2 promoter locus (left) and associated gene-expression changes in carcinoma tumours (right). Expression of MTG2 in mutated (n = 3) versus the carcinoma meta-cohort wild-type cases (n = 896). Two-sided Wilcoxon rank-sum test. Bars represent means. b, Genomic locus of NFKBIZ 3′ UTR (left) and associated gene-expression changes in Lymph–BNHL (right). Expression of NFKBIZ in mutated (n = 6) versus wild-type cases (n = 98). Test and bars as in b. c, Genomic locus of the RMRP transcript and promoter region (left). RMRP is an RNA component of the endoribonuclease RNase MRP, the function of which depends on its RNA secondary and tertiary structure. The RNA secondary structure, tertiary structure interactions, protein and substrate interactions, and mutations with their predicted structural effect (right) of RMRP; lymphoma and melanoma mutations are excluded. d, MIR142 locus and mutations in patients with lymphoma with the AID signature annotation. e, Manhattan-style plot showing significance of mutation recurrence enrichment for genomic bins (top) and ultraconserved elements (bottom) across cohorts (Methods; Supplementary Table 9).
Extended Data Fig. 6 | A transcriptional process creates passenger mutations in highly expressed, tissue-specific genes. 

**a**, Relative rate of loss-of-heterozygosity (LOH) compared between mutated and wild-type samples for all significant elements, coloured by element type and highlighting significant LOH enrichments with an outside black circle (Fisher’s exact test, one-sided; $Q < 0.1$). 

**b**, Average cancer allelic fraction (CAF) compared between each significant genomic element and the corresponding flanking regions ($\pm 2\, \text{kb and introns}$; overlapping coding exons were excluded). The size of the points represents the number of mutated samples for each particular element. Genes with significantly higher CAFs ($t$-test, one-sided; $Q < 0.1$) are highlighted with an outside black circle. 

**c**, mRNA expression of genes enriched in 2–5-bp indels in their respective tissues. Boxes show the interquartile range and median. The first box contains background gene-expression levels. Red and grey dots correspond to samples with $(m)$ and without $(n-m)$ indels in the corresponding gene. 

**d**, Heat map showing the levels of expression across types of cancer for the genes enriched in 2–5-bp indels.
Extended Data Fig. 7 | Overview of structural-variant analysis. Schematic indicating analysis approach. Left, rearrangements and rearrangement junctions in three hypothetical genomes (top) and the two analysis approaches (bottom): the 1D analysis for recurrent breakpoints and the 2D analysis for recurrent juxtapositions between pairs of loci. Right, the 1D density of breakpoints genome-wide (top) and 2D density of juxtapositions (bottom) across 2,693 cancer genomes (Methods).
Extended Data Fig. 8 | Gene-expression effects of SRBs. **a**, Fraction of recurrent breakpoint loci associated with biallelic inactivation of a known tumour suppressor gene (frag-SCNA, 0/12; neutral-SCNA, 0/14; del-SCNA, 5/8; Fisher’s exact test). **b**, Distance in bp to the nearest tissue-specific enhancer for each breakpoint class. Dashed grey line represents 1,000 randomly selected breakpoints from the same tumour samples. All box plots show the interquartile range, median and 95% confidence interval. **c**, Expression fold change for the gene with the most-altered expression within 1 Mb of the cluster centroid in samples with, compared to samples without, a breakpoint at the cluster locus. Random controls (in dashed boxes) represent 1,000 randomly selected breakpoints. P values are from two-sided t-tests (Methods). **d**, Breakpoint density near AKR1C genes (top), locations of enhancers (middle) and expression of local genes (bottom; n = 7 SV+ tumours, n = 41 SV− lung squamous cell tumours; two-sided t-test) in samples with and without local rearrangements. **e**, Ratio of tumour-to-normal read coverage across six breast tumours and eight ovarian tumours with focal BRD4 exon 1 and intron 1 deletions. Red lines indicate rearrangements. **f**, Amplification structure (absolute copy number, y axis) of the BRD4 and NOTCH3 locus in breast and ovarian tumours with a BRD4 focal deletion. In most cases, the copy-number caller identified the focal deletion. However, in some cases, the deletions were too small to be identified only using read depth. When combining read depth and rearrangement signals in a, there is clear evidence for focal deletions. Deletion locations are marked by an asterisk.
Extended Data Fig. 9 | See next page for caption.
Extended Data Fig. 9 | Gene-expression effects of SRJs. a, Assessment of SRJ robustness against unaccounted for mechanistic and technical confounders. Left, a robustness factor, defined as the ratio between the background probability value that would lower the $P$ value of an SRJ below the genome-wide $P$-value threshold and the estimator for the background probability from our 2D model. Higher robustness values represent lower susceptibility to unaccounted variations in the background model. The top 48 SRJs have a robustness factor greater than 2, which suggests that these SRJs would remain significant even if the true background rate was twice as high as our model estimates. Right, the effect size is calculated as the difference in observed and estimated number of SRJs in units of standard deviation (assuming binomial distribution of structural variant count per 2D genomic region). Most SRJs are well above ten standard deviations of the predicted value. b, Characteristics of SRJ secondary loci. Left, fold expression enrichment of the most highly overexpressed gene in the secondary locus in cancer samples with these fusions relative to cancers of the same histology without the fusion. Right, the distance from the SRJ secondary locus (green) to the nearest enhancer is significantly smaller ($P < 0.05$; two-sided $t$-test) compared to randomly selected breakpoints (grey). c, Fold expression enrichment of the most highly overexpressed gene in the primary locus, for fusions that disrupt protein-coding sequences and fusions that do not. All box plots show the interquartile range, median and 95% confidence interval. d, Rearrangements between the TERT promoter and the BASP1 and MYO10 locus result in focal amplification of TERT and relocation of distal enhancers to TERT. e, TERT-NDUFC2 fusion in two melanoma samples connecting TERT with an enhancer-rich region next to NDUFC2. Both samples also have focal amplifications of TERT. f, Recurrent translocation between EGFR in chromosome 7 and the KL and STARD13 locus on chromosome 13. In all three samples, the rearrangement contributed to the amplification of EGFR.
Extended Data Fig. 10 | A lack of detection power in specific elements.

**a**, Number of tumour–normal pairs needed to detect fusions with 90% power as a function of the span of the fusion and the rate above background at which it recurs. The red asterisks indicate the numbers of samples required to detect 100-kb and 100-Mb fusions that recur at 0.5% above their background rates.

**b**, Distribution of TERT promoter hotspot (top, chromosome 5: 1,295,228; bottom, chromosome 5: 1,295,250; hg19) detection sensitivity for each patient, by cohort. Grey dots indicate values for individual patients inside estimated distribution (areas coloured by cohort). Horizontal black bars mark the medians. Numbers above distributions indicate the percentage of patients powered (detection sensitivity ≥ 90%) in each cohort. Cohort sizes as in Fig. 4a.

**c**, Percentage of patients with observed (blue) and inferred missed (red) mutations at the chromosome 5: 1,295,228 and chromosome 5: 1,295,250 TERT promoter hotspot sites. Error bars indicate 95% Poisson confidence interval. Numbers above bars show the total inferred number of TERT promoter mutations for each site in this cohort. Red numbers indicate the absolute number of inferred missed mutations (owing to a lack of read coverage). Cohort sizes as in Fig. 4a.

**d**, Detection sensitivity for the two TERT promoter hotspots across all samples showing the variation in powered samples. Red vertical line (x = 0.9) indicates cutoff for ‘sufficiently powered samples’.

**e**, Mean detection sensitivity in 1,000 randomly selected putative passengers (pass) and 603 cancer genes (driv) across element types: promoters, 5′ UTRs, CDS and 3′ UTRs. The left panel shows the results for all samples and the right panel corresponds to the set of samples with high sensitivity at TERT hotspots. Boxes show the interquartile range and median; outliers are shown as circles. Weighted sensitivity means are shown at the top of the box plot.
Extended Data Fig. 11 | See next page for caption.
Extended Data Fig. 11 | P-value combination details.

**a**, Quantile–quantile plots of P-values reported by various driver-detection algorithms on the three simulated datasets (Broad, DKFZ and Sanger; shown for coding regions \( n = 20,172 \)) in the meta-carcinoma cohort (see Methods for details for the statistical background model or test of each algorithm) showed no major enrichment of mutations above the background rate. Results generally followed the expected null (uniform) distribution, and the P-values reported on simulated data were subsequently used to assess the covariance of method results.

**b**, Quantile–quantile plots of integrated P-values using the Brown and Fisher methods for combining P-values across the results from different driver-detection algorithms were generated for a few representative tumour cohorts (shown here for coding regions). Brown combined P-values (light blue) generally followed the null distribution as expected, whereas Fisher combined P-values were significantly inflated (dark blue), confirming that dependencies existed between the results reported by the various driver-detection algorithms. To simplify the integration procedure, we calculated covariances using P-values from the observed data instead of simulated data and found that the integrated results based on the observed covariances (first column of plots) were essentially the same as the results obtained using the simulated covariances (second, third, and fourth columns of plots).

**c**, Triangular heat maps showing the Spearman correlations of P-values among the various driver-detection methods in observed versus simulated data (coding regions \( n = 20,172 \), colorectal adenocarcinoma cohort) are highly similar. Differences in the observed and simulated correlation values (shown in the heat maps on the far right) were minimal, and thus the final integration of P-values across methods was performed using covariances estimated on observed data.

**d**, Brown combined P-values based on observed and simulated covariance estimations (shown on the right, top heat map, for coding regions in glioblastoma) did not differ noticeably. In cases in which individual methods reported results that yielded substantially fewer hits than the median across all methods (bottom heat map, methods in light grey with results in dashed box), removing the methods from the integration did not affect the number of significant genes identified (right column of results in bottom heat map, shown for coding regions in lung adenocarcinoma). Number of coding regions as in c.
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.

- 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.
- 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 possible.
- 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.

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 v0.7.8-r455; DELLY v0.6.6; ACISeq v1.0.399; DKFZ somatic SNV workflow v1.0.193-1; Phlatypus v0.4.4; ascatNgs v1.5.2; BRASS v4.0.12; grass v1.1.6; CalVeMan v1.50; Pindel v1.5.7; ABSOLUTE/ABBA v1.5; SwaBA 2015-06-20; diRanger 2016-03-13; BreakPointer 2015-12-22; MuTect v1.1.4; MuSE v1.0rc; SmuFIN 2014-10-26; CossG 2016-4-28; VAFGET v2.1.2; ANNOVAR v2014Nov12; VariantBAM v2017Dec12; SV-Merge v2017May26; SV-MERGE v2017Dec12; DKFZ v2018Dec15

The method for combining p-values is available from https://github.com/broadinstitute/getzlab-PCAWG-pvalue_combination/. Power calculations are available from https://github.com/broadinstitute/getzlab-PCAWG-power_calculations. The method for identifying significantly recurrent breakpoints by controlling for covariates is available at: https://github.com/muklab/fish-hook. The method for permuting rearrangement breakpoint pairs to identify covariates affecting rearrangement partner selection is available as the "swap" module of: https://github.com/walajaginseng. The method for identify significantly recurrent rearrangements is available as the "2D" method at: https://github.com/ffershapiro/SVrag. The method for identifying cis-expression consequences of SVs, CESAM, is available at: https://bitbucket.org/weischer/cesam

For manuscripts utilizing custom algorithms or software that are not 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/ICGCWG. 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/hr-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-naive, 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.

For analyses specific to this study, we generated sample subsets as described in the methods.

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% (95% CI: 88-98%) and 95% (95% CI: 71-95%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (79-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 99% 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 |
|-----|-----------------------|
| ☒ | Antibodies |
| ☒ | Eukaryotic cell lines |
| ☒ | Paleontology |
| ☒ | Animals and other organisms |
| ☐ | Human research participants |
| ☒ | Clinical data |

Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒ | Chip-seq |
| ☐ | Flow cytometry |
| ☒ | 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-O-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.