Article

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

https://doi.org/10.1038/s41586-020-1965-x

Received: 19 January 2018

Accepted: 2 December 2019

Published online: 5 February 2020

Open access

Esther Rheinbay[1,2,3,73], Morten Muhlig Nielsen[4,73], Federico Abascal[5,73], Jeremiah A. Wala[1,6,73], Ofer Shapira[1,7,73], Grace Tiao[1], Henrik Hornshøj[4], Julian M. Hess[1], Randi Istrup Juul[4], Ziao Lin[1,8], Lars Feuerbach[9], Radhakrishnan Sabarinathan[10,11], Tobias Madsen[4], Jaegil Kim[1], Loris Mularoni[10,11], Shimin Shuai[12,13], Andrés Lanzós[14,15,16], Carl Herrmann[17,18], Yosef E. Maruvka[1,2], Ciyue Shen[19,20], Samirkumar B. Amin[21,22], Pratiti Bandopadhayay[1,7], Johanna Bertl[4], Keith A. Boroevich[23], John Busanovich[1,7], Joana Carlevaro-Fita[14,15,16], Dimple Chakravarty[24,25], Calvin Wing Yiu Chan[17,26], David Craft[27], Priyanka Dhingra[28,29], Klev Diamanti[30], Nuno A. Fonseca[31], Abel Gonzalez-Perez[10,11], Qianyun Guo[32], Mark P. Hamilton[33], Nicholas J. Haradhvala[1,2], Chen Hong[9,26], Keren Isaev[12,34], Todd A. Johnson[23], Malene Juul[4], Andre Kahles[35], Abdullah Kahraman[36], Youngwook Kim[37], Jan Komorowski[30,38], Kiran Kumar[1,7], Sushant Kumar[39], Donghoon Lee[39], Kjong-Van Lehmann[35], Yilong Li[40,41], Eric Minwei Liu[28,29], Lucas Lochovsky[42], Keunchil Park[37], Oriol Pich[10,11], Nicola D. Roberts[41], Gordon Saksena[1], Steven E. Schumacher[1,7], Nikos Sidiropoulos[43], Lina Sieverling[9,26], Nasa Sinnott-Armstrong[44], Chip Stewart[1], David Tamborero[10,11], Jose M. C. Tubio[45,46,47], Husen M. Umer[30], Liis Uusküla-Reimand[48,49], Claes Wadelius[50], Lina Wadi[12], Xiaotong Yao[51], Cheng-Zhong Zhang[52,53], Jing Zhang[39], James E. Haber[54], Asger Hobolth[32], Marcin Imielinski[51,55], Manolis Kellis[1,56], Michael S. Lawrence[1,2], Christian von Mering[36], Hidewaki Nakagawa[57], Benjamin J. Raphael[58], Mark A. Rubin[59,60,61], Chris Sander[19,20], Lincoln D. Stein[12,13], Joshua M. Stuart[62], Tatsuhiko Tsunoda[23,63,64], David A. Wheeler[65], Rory Johnson[14,16], Jüri Reimand[12,34], Mark Gerstein[39,42,66], Ekta Khurana[28,29,60,61], Peter J. Campbell[5,41], Núria López-Bigas[10,11,67], PCAWG Drivers and Functional Interpretation Working Group[68], PCAWG Structural Variation Working Group[68], Joachim Weischenfeldt[43,69,74*], Rameen Beroukhim[1,6,70,74*], Iñigo Martincorena[5,74*], Jakob Skou Pedersen[4,32,74*], Gad Getz[1,2,3,71,74*] & PCAWG Consortium[72]

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 *TP53*, in the 3′ untranslated regions of *NFKBIZ* and *TOB1*, focal deletions in *BRD4* and rearrangements in the loci of 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[8–15],

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

the frequency and functional implications of these events remain understudied[6,7,13,16,17].

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[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[5], 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-discovery 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 (SRJs), 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[6,7].

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 nucleotide excision repair in melanoma at sites occupied by transcription factors[5,18–20]; (ii) somatic hypermutation by activation-induced cytosine deaminase (AID) in B-cell non-Hodgkin lymphoma (Lymph–BNHL) and chronic lymphocytic leukaemia (Lymph–CLL); (iii) palindromic sequence contexts believed to form hairpin DNA structures targeted by APOBEC enzymes (in an intron of *GPR126* (also known as *ADGRG6*) and the *PLEKHS1* promoter)[10]; 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)[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 *RPL13A* (melanoma) because of associations with localized AID and UV-light mutational processes, respectively; *PLEKHS1*, *GPR126*, *TBC1D12* and *LEPROTL1* because of palindromic APOBEC target sequences[9,10]; and the *WDR74* 5′ UTR and promoter[8,10,14], 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–BNHL)—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[9] (Extended Data Fig. 4a, Supplementary Table 10).

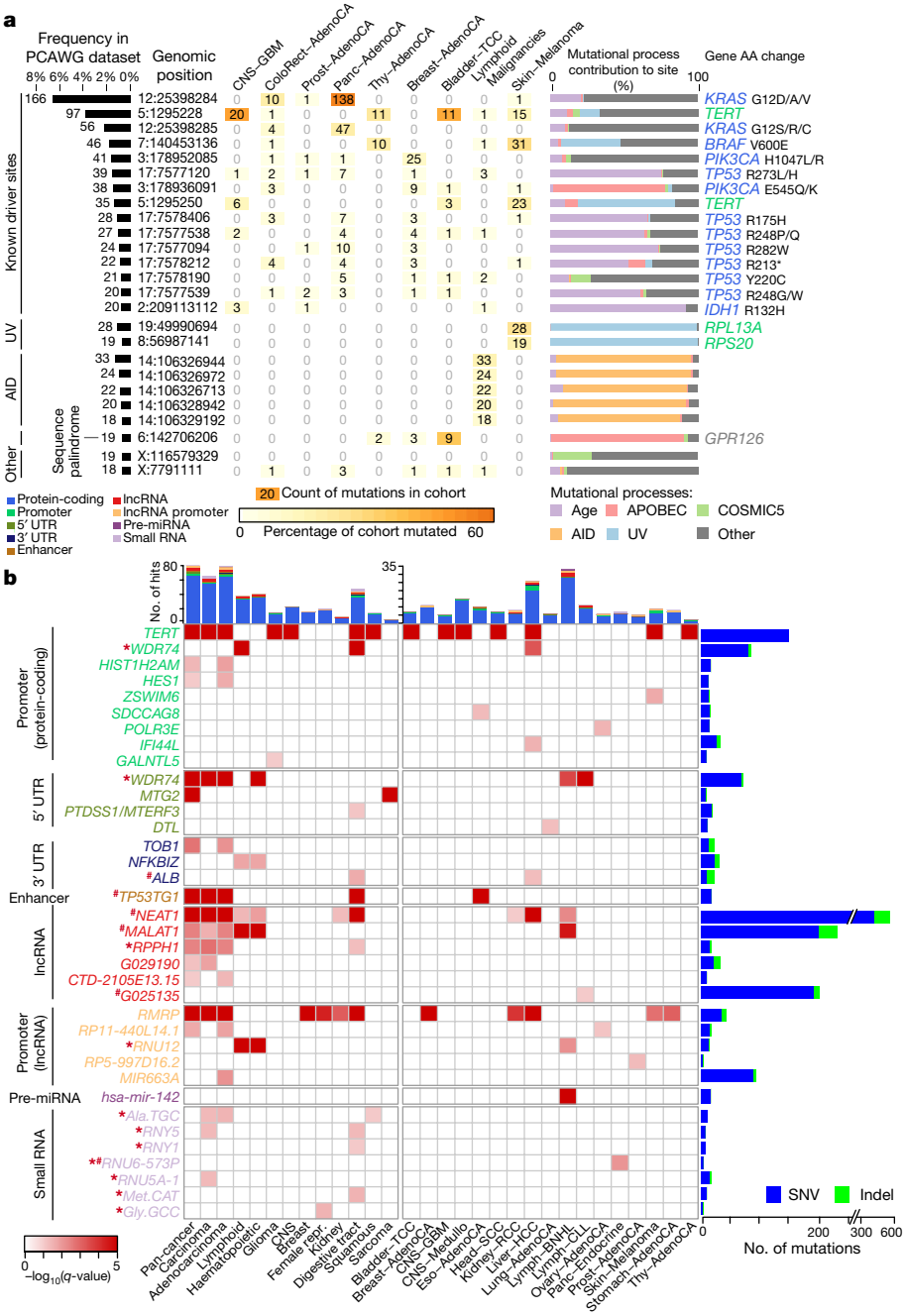


Fig. 1 | Non-coding point mutations in PCAWG. a, The bar chart (left) shows the total number of patients across PCAWG with mutations at a particular genomic hotspot (chromosome:position). The top 25 hotspots are grouped as known drivers or induced by mutational processes. The table (middle) shows the frequency of mutations across a subset of PCAWG cohorts. Lymphoid malignancies comprise Lymph–BNHL and Lymph–CLL. The stacked bar chart (right) shows the contribution of mutational processes to the hotspot mutations (Methods). Gene names are given when hotspots overlap functional elements (colour-coded), with amino acid (AA) alterations for protein-coding genes (solidus denotes substitution with any one of the indicated amino acids). Extended Data Fig. 1b shows the top 50 hotspots, and all cohorts. **b**, Significant non-coding elements ($Q < 0.1$ of Brown's combined P values of up to 13 driver discovery methods; Methods) identified before manual review in cohorts with at least one hit. Colour represents significance levels. Details are provided in Supplementary Table 5. *Potential technical artefact; #targets affected by mutational processes. AdenoCA, adenocarcinoma; CNS, central nervous system; Eso, oesophageal; GBM, glioblastoma; HCC, hepatocellular carcinoma; Medullo, medulloblastoma; Panc, pancreatic; Prost, prostate; RCC, renal cell carcinoma; Repr., reproductive organs; SCC, squamous cell carcinoma; TCC, transitional cell carcinoma; Thy, thyroid. *HIST1H2AM* is also known as *H2AC17*; *Ala.TGC* as *TRA-TGC3-1*; *Met.CAT* as *TRM-CAT1-1*; and *Gly.GCC* as *TRG-GCC2-3*. *PTDSS1/MTERF3* denotes that 5′ UTR mutations in *PTDSS1* also overlap the *MTERF3* promoter.

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[22].

Recurrent somatic events were identified in the 3′ UTRs of *TOB1* (carcinoma and pan-cancer meta-cohorts), *NFKBIZ* (lymphomas) and *ALB* (liver cancer) (Fig. 1b). *TOB1* encodes an anti-proliferation regulator

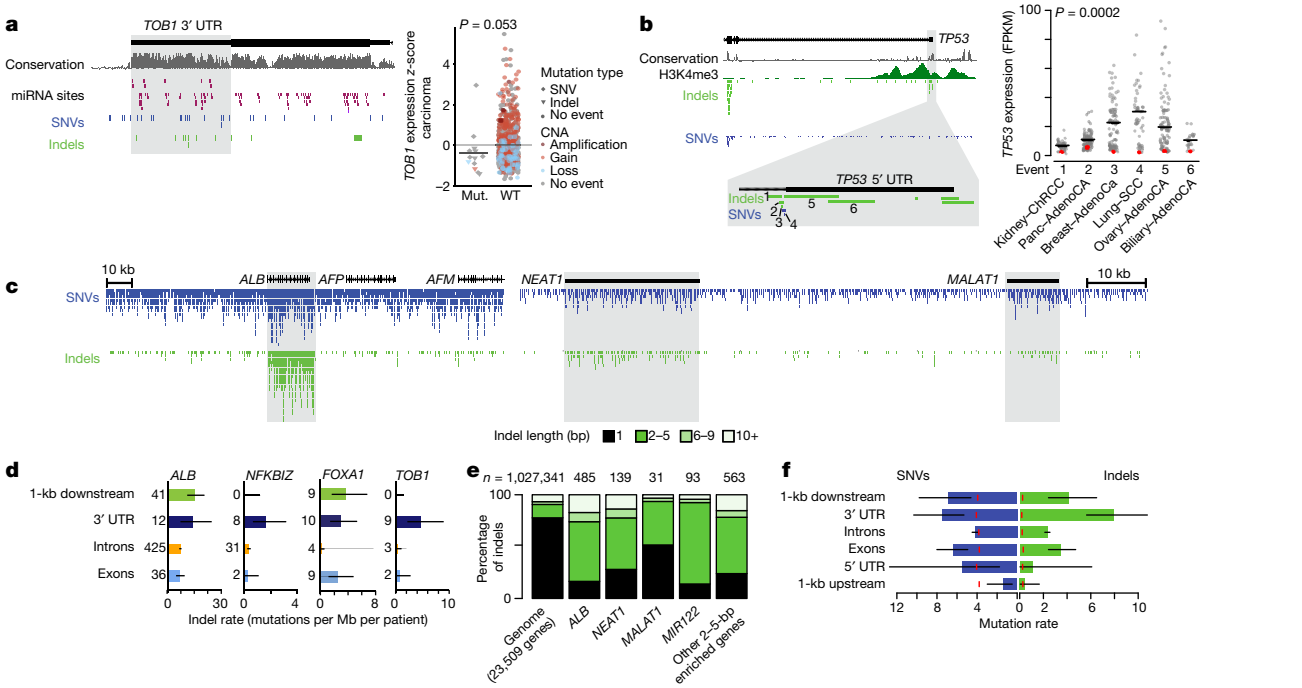


Fig. 2 | Newly identified non-coding driver candidates and localized transcription-associated mutational process. a, Recurrent mutations and associated gene expression in the highly conserved *TOB1* 3′ UTR. Tracks showing conservation score (PhyloP, grey), miRNA-binding sites (TargetScan (top track) and Ago-Clip (bottom track)), and observed SNVs (blue) and indels (green). Expression of *TOB1* in mutated ($n = 13$) and wild-type ($n = 886$) cases (right). *P* value based on two-sided Wilcoxon rank-sum test. Bars represent means. CNA, copy-number alteration. **b**, Indels and SNVs overlapping the *TP53* 5′ region and their effect on gene expression. H3K4me3 from the GM12878 cell line (ENCODE). Event numbers match with gene expression in the right panel (red dot, mutated sample; black bar, median). *P* value represents Fisher's combination of permutation tests within each tumour type. ChRCC, chromophobe renal cell carcinoma; FPKM, fragments per kilobase of transcript per million mapped reads. **c**, Overall pan-cancer distribution of indels and SNVs in *ALB*, *NEAT1* and *MALAT1* genomic loci (lymphoid tumour samples were excluded owing to AID). **d**, Quantification of average indel rates for genes with significantly mutated 3′ UTRs. Error bars represent 95% binomial confidence intervals. **e**, Contribution of indels of different sizes in: all protein-coding and long non-coding RNA genes; *ALB*; *NEAT1*; *MALAT1*; *MIR122*; and the remaining genes enriched in 2–5-bp indels. **f**, SNV and indel rates (total events per Mb per patient) in different functional regions of 18 protein-coding genes enriched in 2–5-bp indels (without *ALB*, which contributed 47% of indels). Red lines indicate background indel and SNV rates estimated from all protein-coding genes. Error bars as in **d**; raw counts provided in Supplementary Table 18. **c**–**f**, Mutations analysed in all unique cases ($n = 2,583$).

that associates with ERBB2, and also affects migration and invasion in gastric cancer[23]. TOB1 regulates other mRNAs through binding to their 3′ UTR and promoting deadenylation[24]. Tumours with 3′ UTR mutations in *TOB1* 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 *WFIKKN2* are focally amplified in breast cancer and pan-cancer, suggesting a complex role in cancer (Extended Data Fig. 4b). *NFKBIZ* is a transcription factor that is mutated in diffuse large B cell lymphoma and amplified in primary lymphomas[25]. 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 associated these mutations with increased *NFKBIZ* expression[25], 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 *RMRP* were significantly mutated in multiple types of cancer (Fig. 1b, Extended Data Fig. 5c). Germline *RMRP* mutations cause cartilage–hair hypoplasia, and previous in vitro studies have shown that some somatic promoter mutations are functional[16]. The *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[12,26], but 7 out of 8 mutations in the mature miRNA *mir-142-p3*––for which the largest functional effect is expected––were not assigned to AID, which suggests that these mutations are under selection[12].

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[21] (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 *MIR122*) were hit by many indels; all have previously been reported to be mutated in cancer[10,15,27] (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 *TOB1* 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 *MIR122* 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*, *MIR122* 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 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[15]) (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 *MIR122* 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[27–29] 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, deactivating 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[31] (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[32]. 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. Ectopic expression transforms cell lines, and germline mutations have previously been linked to an increased risk of developing lung cancer[33,34]. 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 5′ end of *BRD4* in ovarian (eight cases, $P < 10^{-7}$) 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[35,36], including ovarian and triple-negative breast cancer[37,38]. Given the increased copy number of the full *BRD4* gene, we would expect increased gene expression. However, the microdeletions

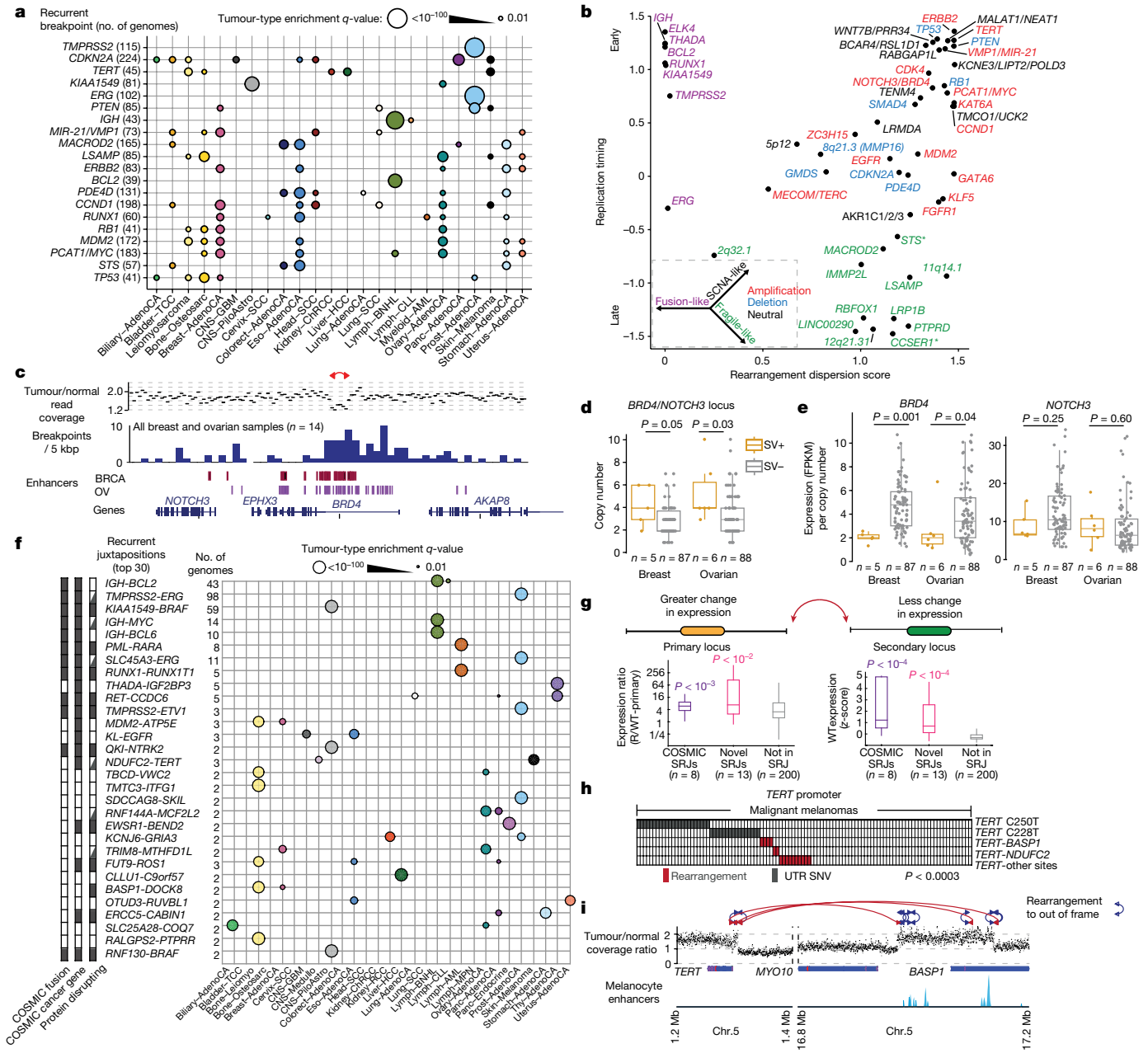


Fig. 3 | Significantly recurrent breakpoints and juxtapositions. a, Relative enrichment (Fisher's exact test) for events per tumour type for the 20 most-significant SRBs (circle size). Loci are labelled by the likely driver gene from the CGC[21]. For gene symbols separated by a solidus, both or either of the genes are intended. **b**, Rearrangement dispersion score versus mean replication timing of the 53 SRBs. Colours indicate fusion (purple), fragile-like (green), deletion (blue), amplification (red) or copy-neutral (black) events. **c**, Tumour-to-normal read coverage ratio in an ovarian tumour with a *BRD4* microdeletion; red arrow indicates the rearrangement (top). Breakpoint density across PCAWG breast and ovarian cancers (middle). Enhancer locations from breast (BRCA) and ovarian (OV) tissue[51] (bottom). **d**, Somatic copy number at the *BRD4* and *NOTCH3* locus in breast and ovarian cancers with (SV+) and without (SV−) rearrangements. **e**, Gene expression per absolute copy number for *BRD4* and *NOTCH3*. **f**, The 30 most-significant SRJs, with their relative enrichment (circle size) per tumour type, annotated with oncogenic fusions from the Catalogue of Somatic Mutations in Cancer (COSMIC) (left), CGC gene (centre) and protein disruption (right) (Methods). *ATP5E* is also known as *ATP5F1E*. **g**, Expression correlates of rearrangements in SRJs from COSMIC (purple), other SRJs (pink) or not in any SRJ (grey). For each rearrangement (R), the primary locus (left) is defined as the breakpoint within 100 kb of the gene that is most overexpressed in rearranged samples; the secondary locus (right) is the other breakpoint. Expression at the primary locus in samples with the rearrangement relative to samples without the rearrangement is greater for SRJs than for other rearrangements (left). 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**, **e**, **g**, Box plots show the interquartile range, median and 95% confidence interval; two-sided *t*-test. **h**, *TERT* promoter mutations and rearrangements across PCAWG melanomas. **i**, Rearrangements between *TERT* promoter and *BASP1* and *MYO10* locus result in focal amplification and relocation of distal enhancers to *TERT*. AML, acute myeloid leukaemia; Colorect, colorectal; Leiomyo, leiomyosarcoma; MPN, myeloproliferative neoplasm; Osteosarc, osteosarcoma; PiloAstro, pilocytic astrocytoma.

are associated with a lower expression of *BRD4* in breast ($P = 0.001$) and ovarian tumours ($P = 0.04$), but not of the neighbouring gene *NOTCH3* (Fig. 3e). The focal deletions in *BRD4* overlap a prominent exon-1 H3K4me3 peak and intron-1 enhancer elements in HMEC (normal breast) and MCF-7 (breast tumour) cells (Extended Data Fig. 8e), which suggests that these deletions disrupt regulatory elements.

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 8 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[39,40]. 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[13,41–43]. 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, $P < 10^{-4}$ (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, $P < 10^{-7}$). 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[44] (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*, *EGFR* and *TERT* (Fig. 3f, h, i, Extended Data Fig. 9d–f, Supplementary Table 15). The *TERT* promoter region was juxtaposed in four melanomas ($P < 10^{-7}$) 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 ($P < 10^{-8}$). 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 C228T 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[3]. 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[3,16]. 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 (Fig. 4c).

Low sequencing coverage (for example, in GC-rich regions[45]) also limits driver discovery. To measure this effect in the PCAWG data, we quantified our ability to detect mutations (detection sensitivity)[16] 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[46] (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).

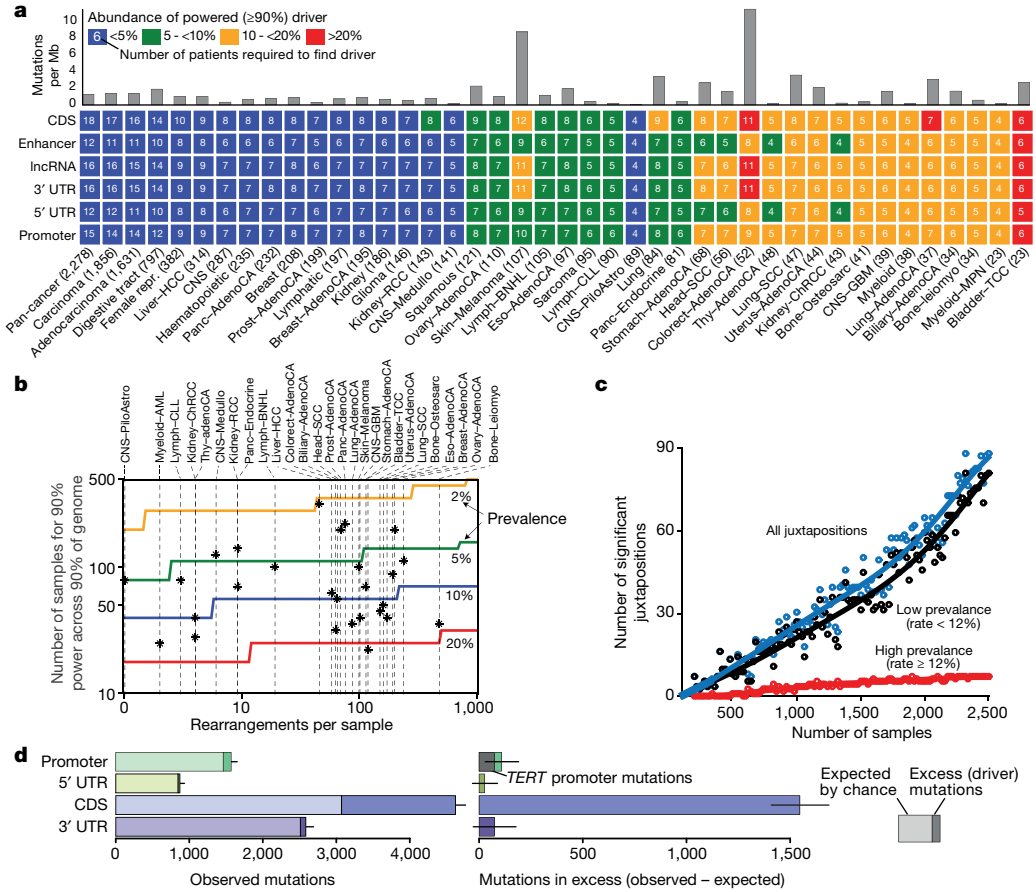


Fig. 4 | Power considerations and paucity of non-coding drivers. a, Heat map shows the minimal frequency of a driver element with ≥90% discovery power. Power is dependent on the background mutation frequency (above the heat map), the element length (median length depicted in Extended Data Fig. 2c) and the number of patients with mutations (cell numbers). For example, the pan-cancer cohort is powered to discover a protein-coding driver gene (coding sequence (CDS)) present in <1% (18 patients), whereas the Bladder–TCC cohort is only powered to discover drivers present in at least 27% (6 patients). **b**, Number of samples required to detect 90% of recurrent juxtapositions across 90% of pairs of loci, as a function of the median number of rearrangements per sample and the rate above background at which the fusion recurs (solid lines). The vertical dashed lines represent the median rearrangement rates of each cancer type, and the stars on these lines indicate the numbers of whole genomes analysed for that cancer type. **c**, Number of SRJs detected after downsampling the data to various sample sizes, separately indicating rearrangements that recur at high (≥12%; red) and low (<12%; black) rates above background; their sum (blue). **d**, Number of observed mutations (SNVs and indels) in *cis*-regulatory and coding regions of 603 protein-coding cancer genes with the expected numbers shown in lighter colours (left). Right, the number of excess mutations (that is, the estimated number of driver mutations) (right). The grey fraction of promoter mutations indicates *TERT* events. Error bars show 95% binomial confidence intervals. Only samples with high detection sensitivity were included ($n = 936$).

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[5].

Among the most interesting candidate non-coding driver elements we uncovered are the 5′-end mutations in *TP53*; 3′ UTR mutations in *NFKBIZ* and *TOB1*; and rearrangements involving AKR1C 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 germline polymorphisms associated with heritability of complex traits, which are most frequently located outside of protein-coding genes[47].

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[48]. 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-unidentified 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 modelling 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 modelling SRJs is much larger (the genome squared); and (iv) many structural variants are part of complex events that often involve multiple chromosomes[31], so that the resultant topology cannot be deduced without technologies such as long- or linked-read sequencing[49,50]. 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/s41586-020-1965-x.

1. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **174**, 1034–1035 (2018).
2. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
3. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
4. Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
5. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network. Pan-cancer analysis of whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1969-6 (2020).
6. Horn, S. et al. *TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
7. Huang, F. W. et al. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
8. Khurana, E. et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
9. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
10. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
11. Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47**, 710–716 (2015).
12. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
13. Northcott, P. A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
14. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).
15. Imielinski, M., Guo, G. & Meyerson, M. Insertions and deletions target lineage-defining genes in human cancers. *Cell* **168**, 460–472.e14 (2017).
16. Rheinbay, E. et al. Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55–60 (2017).
17. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in *IDH* mutant gliomas. *Nature* **529**, 110–114 (2016).
18. Perera, D. et al. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).
19. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
20. Mao, P. et al. ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat. Commun.* **9**, 2626 (2018).
21. Forbes, S. A. et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.* **38**, D652–D657 (2010).
22. Zhang, W. et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* **50**, 613–620 (2018).
23. Li, B.-S. et al. MicroRNA-25 promotes gastric cancer migration, invasion and proliferation by directly targeting transducer of ERBB2, 1 and correlates with poor survival. *Oncogene* **34**, 2556–2565 (2015).
24. Hosoda, N. et al. Anti-proliferative protein Tob negatively regulates CPEB3 target by recruiting Caf1 deadenylase. *EMBO J.* **30**, 1311–1323 (2011).
25. Arthur, S. E. et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* **9**, 4001 (2018).
26. Robbiani, D. F. et al. AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Mol. Cell* **36**, 631–641 (2009).
27. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
28. Ke, H. et al. NEAT1 is required for survival of breast cancer cells through FUS and miR-548. *Gene Regul. Syst. Bio.* **10** (Suppl 1), 11–17 (2016).
29. Han, Y., Liu, Y., Nie, L., Gui, Y. & Cai, Z. Inducing cell proliferation inhibition, apoptosis, and motility reduction by silencing long noncoding ribonucleic acid metastasis-associated lung adenocarcinoma transcript 1 in urothelial carcinoma of the bladder. *Urology* **81**, 209. e1–209.e7 (2013).
30. Annala, M. et al. Frequent mutation of the *FOXA1* untranslated region in prostate cancer. *Commun. Biol.* **1**, 122 (2018).
31. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* https://doi.org/10.1038/s41586-019-1913-9 (2020).
32. Bignell, G. R. et al. Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
33. Chien, C.-W., Ho, I.-C. & Lee, T.-C. Induction of neoplastic transformation by ectopic expression of human aldo-keto reductase 1C isoforms in NIH3T3 cells. *Carcinogenesis* **30**, 1813–1820 (2009).
34. Lan, Q. et al. Oxidative damage-related genes *AKR1C3* and *OGG1* modulate risks for lung cancer due to exposure to PAH-rich coal combustion emissions. *Carcinogenesis* **25**, 2177–2181 (2004).
35. Filippakopoulos, P. et al. Selective inhibition of BET bromodomains. *Nature* **468**, 1067–1073 (2010).
36. Dawson, M. A., Kouzarides, T. & Huntly, B. J. P. Targeting epigenetic readers in cancer. *N. Engl. J. Med.* **367**, 647–657 (2012).
37. Shu, S. et al. Response and resistance to BET bromodomain inhibitors in triple-negative breast cancer. *Nature* **529**, 413–417 (2016).
38. Baratta, M. G. et al. An in-tumor genetic screen reveals that the BET bromodomain protein, BRD4, is a potential therapeutic target in ovarian carcinoma. *Proc. Natl Acad. Sci. USA* **112**, 232–237 (2015).
39. Tomlins, S. A. et al. Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
40. May, W. A. et al. The Ewing's sarcoma EWS/FLI-1 fusion gene encodes a more potent transcriptional activator and is a more powerful transforming gene than FLI-1. *Mol. Cell. Biol.* **13**, 7393–7398 (1993).
41. Weischenfeldt, J. et al. Pan-cancer analysis of somatic copy-number alterations implicates *IRS4* and *IGF2* in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
42. Mani, R.-S. & Chinnaiyan, A. M. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. *Nat. Rev. Genet.* **11**, 819–829 (2010).
43. Schneider, G., Schmidt-Supprian, M., Rad, R. & Saur, D. Tissue-specific tumorigenesis: context matters. *Nat. Rev. Cancer* **17**, 239–253 (2017).
44. St John, J., Powell, K., Conley-Lacomb, M. K. & Chinni, S. R. TMPRSS2-ERG fusion gene expression in prostate tumor cells and its clinical and biological significance in prostate cancer progression. *J. Cancer Sci. Ther.* **4**, 94–101 (2012).
45. Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52**, 87–94 (2012).
46. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
47. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
48. Shuai, S. et al. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. *Nature* **574**, 712–716 (2019).
49. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).

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).

[1]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Aarhus, Denmark. [5]Wellcome Trust Sanger Institute, Hinxton, UK. [6]Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA, USA. [7]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [8]Harvard University, Cambridge, MA, USA. [9]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [10]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [11]Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain. [12]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [13]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [14]Department for BioMedical Research, University of Bern, Bern, Switzerland. [15]Graduate School of Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland. [16]Department of Medical Oncology, Bern University Hospital, University of Bern, Bern, Switzerland. [17]Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [18]Bioquant Center, Institute of Pharmacy and Molecular Biotechnology, University of Heidelberg, Heidelberg, Germany. [19]Department of Cell Biology, Harvard Medical School, Boston, MA, USA. [20]cBio Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [21]Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, USA. [22]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX, USA. [23]Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [24]Department of Genitourinary Medical Oncology - Research, Division of Cancer Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, USA. [25]Department of Urology, Icahn school of Medicine at Mount Sinai, New York, NY, USA. [26]Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. [27]Department of Radiation Oncology, Harvard Medical School, Massachusetts General Hospital, Boston, MA, USA. [28]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [29]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [30]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden. [31]European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, UK. [32]Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, Denmark. [33]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA. [34]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [35]Division of Computational Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [36]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. [37]Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea. [38]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland. [39]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [40]SBGD Inc, Cambridge, MA, USA. [41]Department of Haematology, University of Cambridge, Cambridge, UK. [42]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [43]Biotech Research & Innovation Centre (BRIC), The Finsen Laboratory, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark. [44]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [45]Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [46]Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [47]The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo, Spain. [48]Genetics and Genome Biology Program, SickKids Research Institute, Toronto, Ontario, Canada. [49]Department of Gene Technology, Tallinn University of Technology, Tallinn, Estonia. [50]Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden. [51]New York Genome Center, New York, NY, USA. [52]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [53]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [54]Department of Biology and Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA, USA. [55]Department of Pathology and Laboratory Medicine, and Englander Institute for Precision Medicine, and Institute for Computational Biomedicine, and Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [56]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA. [57]Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan. [58]Department of Computer Science, Princeton University, Princeton, NJ, USA. [59]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. [60]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. [61]Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [62]Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, CA, USA. [63]Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan. [64]Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan. [65]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [66]Department of Computer Science, Yale University, New Haven, CT, USA. [67]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. [68]A list of members and their affiliations appears in the online version of the paper. [69]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [70]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [71]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [72]A list of members and their affiliations appears in the Supplementary Information. [73]These authors contributed equally: Esther Rheinbay, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A. Wala, Ofer Shapira. [74]These authors jointly supervised this work: Joachim Weischenfeldt, Rameen Beroukhim, Iñigo Martincorena, Jakob Skou Pedersen, Gad Getz. *e-mail: joachim.weischenfeldt@bric.ku.dk; rameen@broadinstitute.org; im3@sanger.ac.uk; jakob.skou@clin.au.dk; gadgetz@broadinstitute.org

PCAWG Drivers and Functional Interpretation Working Group

Federico Abascal[5], Samirkumar B. Amin[21,22], Gary D. Bader[75], Pratiti Bandopadhayay[1,7], Jonathan Barenboim[76], Rameen Beroukhim[1,6,70], Johanna Bertl[4], Keith A. Boroevich[23], Søren Brunak[77,78], Peter J. Campbell[5,41], Joana Carlevaro-Fita[14,15,16], Dimple Chakravarty[24,25], Calvin Wing Yiu Chan[17,26], Ken Chen[79], Jung Kyoon Choi[80], Jordi Deu-Pons[81,82], Priyanka Dhingra[28,29], Klev Diamanti[30], Lars Feuerbach[9], J. Lynn Fink[83,84], Nuno A. Fonseca[31], Joan Frigola[81], Carlo Gambacorti-Passerini[85], Dale W. Garsed[86,87], Mark Gerstein[39,42,66], Gad Getz[1,2,3,71], Qianyun Guo[32], Ivo G. Gut[88,89], David Haan[90], Mark P. Hamilton[33], Nicholas J. Haradhvala[1,2], Arif O. Harmanci[91,92], Mohamed Helmy[93], Carl Herrmann[17,18], Julian M. Hess[1], Asger Hobolth[32], Ermin Hodzic[94], Chen Hong[9,26], Henrik Hornshøj[4], Keren Isaev[12,34], Jose M. G. Izarzugaza[77], Rory Johnson[14,16], Todd A. Johnson[23], Malene Juul[4], Randi Istrup Juul[4], Andre Kahles[35], Abdullah Kahraman[36], Manolis Kellis[1,56], Ekta Khurana[28,29,60,61], Jaegil Kim[1], Jong K. Kim[95], Youngwook Kim[37], Jan Komorowski[30,38], Jan O. Korbel[96,97], Sushant Kumar[39], Andrés Lanzós[14,15,16], Erik Larsson[98], Michael S. Lawrence[1,2], Donghoon Lee[39], Kjong-Van Lehmann[35], Shantao Li[91], Xiaotong Li[91], Ziao Lin[1,8], Eric Minwei Liu[28,29], Lucas Lochovsky[42], Shaoke Lou[91,99], Tobias Madsen[4], Kathleen Marchal[100,101], Iñigo Martincorena[5], Alexander Martinez-Fundichely[102,103,104], Yosef E. Maruvka[1,2], Patrick D. McGillivray[99], William Meyerson[91,105], Ferran Muiños[82,106], Loris Mularoni[10,11], Hidewaki Nakagawa[57], Morten Muhlig Nielsen[4], Marta Paczkowska[76], Keunchil Park[37], Kiejung Park[107], Jakob Skou Pedersen[4,32], Tirso Pons[108], Sergio Pulido-Tamayo[100,101], Benjamin J. Raphael[58], Jüri Reimand[12,34], Iker Reyes-Salazar[106], Matthew A. Reyna[109], Esther Rheinbay[1,2,3], Mark A. Rubin[59,60,61], Carlota Rubio-Perez[82,106,110], S. Cenk Sahinalp[94,111,112], Gordon Saksena[1], Leonidas Salichos[91,99], Chris Sander[19,20], Steven E. Schumacher[1,7], Mark Shackleton[87,113], Ofer Shapira[1,7], Ciyue Shen[19,20], Raunak Shrestha[111], Shimin Shuai[12,13], Nikos Sidiropoulos[43], Lina Sieverling[9,26], Nasa Sinnott-Armstrong[44], Lincoln D. Stein[12,13], Joshua M. Stuart[62], David Tamborero[10,11], Grace Tiao[1], Tatsuhiko Tsunoda[23,63,64], Husen M. Umer[30], Liis Uusküla-Reimand[48,49], Alfonso Valencia[83,114], Miguel Vazquez[83,115], Lieven P. C. Verbeke[101,116], Claes Wadelius[50], Lina Wadi[12], Jiayin Wang[117,118,119], Jonathan Warrell[91,99], Sebastian M. Waszak[97], Joachim Weischenfeldt[43,69], David A. Wheeler[65], Guanming Wu[120], Jun Yu[121], Jing Zhang[39], Xuanping Zhang[118,122], Yan Zhang[91,123,124], Zhongming Zhao[125], Lihua Zou[126] & Christian von Mering[36]

PCAWG Structural Variation Working Group

Kadir C. Akdemir[79], Eva G. Alvarez[127,128,129], Adrian Baez-Ortega[130], Rameen Beroukhim[1,6,70], Paul C. Boutros[76,131,132,133], David D. L. Bowtell[86,134], Benedikt Brors[135,136,137], Kathleen H. Burns[138], John Busanovich[1,7], Peter J. Campbell[5,41], Kin Chan[139], Ken Chen[79], Isidro Cortés-Ciriano[140,141,142], Ana Dueso-Barroso[83], Andrew J. Dunford[143], Paul A. Edwards[144,145], Xavier Estivill[146,147], Dariush Etemadmoghadam[86,87], Lars Feuerbach[9], J. Lynn Fink[83,84], Milana Frenkel-Morgenstern[148], Dale W. Garsed[86,87], Mark Gerstein[39,42,66], Dmitry A. Gordenin[149], David Haan[90], James E. Haber[54], Julian M. Hess[1], Barbara Hutter[135,150,151], Marcin Imielinski[51,55], David T. W. Jones[152,153], Young Seok Ju[80,154], Marat D. Kazanov[155,156,157], Leszek J. Klimczak[158], Youngil Koh[159,160], Jan O. Korbel[96,97], Kiran Kumar[1,7], Eunjung Alice Lee[161], Jake June-Koo Lee[141,162], Yilong Li[40,41], Andy G. Lynch[144,145,163], Geoff Macintyre[144], Florian Markowetz[144,145], Iñigo Martincorena[5], Alexander Martinez-Fundichely[102,103,104], Matthew Meyerson[143,164,165,166,167], Satoru Miyano[168], Hidewaki Nakagawa[57], Fabio C. P. Navarro[99], Stephan Ossowski[89,146,169], Peter J. Park[141,162], John V. Pearson[170,171], Montserrat Puiggròs[83], Karsten Rippe[172], Nicola D. Roberts[41], Steven A. Roberts[173], Bernardo Rodriguez-Martin[127,128,129], Steven E. Schumacher[1,7], Ralph Scully[174], Mark Shackleton[87,113], Nikos Sidiropoulos[43], Lina Sieverling[9,26], Chip Stewart[1], David Torrents[83,114], Jose M. C. Tubio[45,46,47], Izar Villasante[83], Nicola Waddell[170,171], Jeremiah A. Wala[1,6], Joachim Weischenfeldt[43,69], Lixing Yang[175], Xiaotong Yao[51], Sung-Soo Yoon[160], Jorge Zamora[127,128,129,154] & Cheng-Zhong Zhang[52,53]

[75]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [76]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [77]Technical University of Denmark, Lyngby, Denmark. [78]University of Copenhagen, Copenhagen, Denmark. [79]University of Texas MD Anderson Cancer Center, Houston, TX, USA. [80]Korea Advanced Institute of Science and Technology, Daejeon, South Korea. [81]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [82]Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain. [83]Barcelona Supercomputing Center (BSC), Barcelona, Spain. [84]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, St Lucia, Queensland, Australia. [85]University of Milano Bicocca, Monza, Italy. [86]Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [87]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Victoria, Australia. [88]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [89]Universitat Pompeu Fabra, Barcelona, Spain. [90]Biomolecular Engineering Department, University of California, Santa Cruz, Santa Cruz, CA, USA. [91]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [92]Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA. [93]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [94]Simon Fraser University, Burnaby, British Columbia, Canada. [95]Research Core Center, National Cancer Centre Korea, Goyang-si, South Korea. [96]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK. [97]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [98]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [99]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [100]Department of Information Technology, Ghent University, Ghent, Belgium. [101]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [102]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [103]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [104]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. [105]Yale School of Medicine, Yale University, New Haven, CT, USA. [106]Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain. [107]Cheonan Industry-Academic Collaboration Foundation, Sangmyung University, Cheonan, South Korea. [108]Spanish National Cancer Research Centre, Madrid, Spain. [109]Department of Computer Science, Princeton University, Princeton, NJ, USA. [110]Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain. [111]Vancouver Prostate Centre, Vancouver, British Columbia, Canada. [112]Indiana University, Bloomington, IN, USA. [113]Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Victoria, Australia. [114]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [115]Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway. [116]Department of Information Technology, Interuniversitair Micro-Electronica Centrum (IMEC), Ghent University, Ghent, Belgium. [117]The McDonnell Genome Institute, Washington University, St Louis, MO, USA. [118]School of Computer Science and Technology, Xian Jiaotong University, Xian, China. [119]School of Electronic and Information Engineering, Xian Jiaotong University, Xian, China. [120]Oregon Health & Sciences University, Portland, OR, USA. [121]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, Hong Kong, China. [122]The University of Texas Health Science Center at Houston, Houston, TX, USA. [123]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA. [124]The Ohio State University Comprehensive Cancer Center (OSUCCC – James), Columbus, OH, USA. [125]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA. [126]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. [127]Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [128]Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [129]The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo, Spain. [130]Transmissible Cancer Group, Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. [131]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [132]Department of Pharmacology, University of Toronto, Toronto, Ontario, Canada. [133]University of California Los Angeles, Los Angeles, CA, USA. [134]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia. [135]National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany. [136]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [137]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. [138]Johns Hopkins School of Medicine, Baltimore, MD, USA. [139]Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada. [140]Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK. [141]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [142]Ludwig Center, Harvard Medical School, Boston, MA, USA. [143]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [144]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [145]University of Cambridge, Cambridge, UK. [146]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [147]Quantitative Genomics Laboratories (qGenomics), Barcelona, Spain. [148]The Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel. [149]Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. [150]German Cancer Consortium (DKTK), Heidelberg, Germany. [151]Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center (DKFZ), Heidelberg, Germany. [152]Hopp Childrens Cancer Center (KiTZ), Heidelberg, Germany. [153]Pediatric Glioma Research Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. [154]Wellcome Sanger Institute, Hinxton, UK. [155]Skolkovo Institute of Science and Technology, Moscow, Russia. [156]A. A. Kharkevich Institute of Information Transmission Problems, Moscow, Russia. [157]Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia. [158]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences (NIEHS), Durham, NC, USA. [159]Center For Medical Innovation, Seoul National University Hospital, Seoul, South Korea. [160]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. [161]Division of Genetics and Genomics, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA. [162]Ludwig Center at Harvard, Boston, MA, USA. [163]School of Medicine and School of Mathematics and Statistics, University of St Andrews, St Andrews, UK. [164]Harvard Medical School, Boston, MA, USA. [165]Dana-Farber Cancer Institute, Boston, MA, USA. [166]Department of Medical Oncology, University Hospital, University of Bern, Bern, Switzerland. [167]Department of Pathology, The University of Melbourne, Melbourne, Victoria, Australia. [168]The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [169]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. [170]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [171]Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, Queensland, Australia. [172]German Cancer Research Center (DKFZ), Heidelberg, Germany. [173]Center for Reproductive Biology, School of Molecular Biosciences, Washington State University, Pullman, WA, USA. [174]Cancer Research Institute, Beth Israel Deaconess Medical Center, Boston, MA, USA. [175]Ben May Department for Cancer Research, Department of Human Genetics, The University of Chicago, Chicago, IL, USA.

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[52], 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 ActiveDriverWGS[54], CompositeDriver, DriverPower[55], dndscv[46], ExInAtor[56], LARVA[57], MutSig tools[3], NBR[10], ncdDetect[58], ncDriver[59], OncodriveFML[60] and regDriver[61]. We integrated the results of all these methods using a custom framework based on a previously published method[62] 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[63]. Cohort–element combinations with *Q* values < 0.1 were designated as significant hits, and combinations with $0.1 \leq Q < 0.25$ 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[64].

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[15] 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[63] 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[65].

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[63].

Gene-expression analyses

Gene-expression data were provided by the PCAWG Transcriptome Core Group[66], 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[67]. 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[68].

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[3,16], 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 SNVs and indels expected by chance. By aggregating the expected numbers over 603 known cancer genes from the CGC[69] (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 PCAWG Consortium are described in an accompanying Article[5], 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 PCAWG 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/mskilab/fishHook, https://github.com/walaj/ginseng and https://github.com/walaj/SVsig. 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 cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
53. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE pProject. *Genome Res.* **22**, 1760–1774 (2012).
54. Wadi, L., Uuskula-Reimand, L., Isaev, 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. Lanzos, A. et al. Discovery of cancer driver long noncoding RNAs across 1112 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. Juul, 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., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: 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. Dimitrieva, S. & Bucher, P. UCNEbase—a database of ultraconserved 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. PCAWG Transcriptome Core Group et al. Genomic basis for RNA alterations in cancer. *Nature* https://doi.org/10.1038/s41586-020-1970-0 (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 2,658 cancers. *Nature* https://doi.org/10.1038/s41586-019-1907-0 (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 PCAWG Network and the PCAWG steering committee for enabling this work, and for guidance throughout the study. We thank K. Kübler for assistance with meta-cohort generation and R. Heller for discussion on FDR. We are grateful to the PCAWG 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. J.S.P. was partially funded by Independent Research Fund Denmark (12-126439 and 7016-00379) and The Danish Cancer Society (R124-A7869). R.B. received funds from the National Institutes of Health (U54CA143798, R01CA188228, R35GM127029, and R01CA215489), the DFCI-Novartis Drug Discovery Program, the Pediatric Low Grade Astrocytoma Foundation, the Cure Starts Now Foundation and The Fund for Innovation in Cancer Informatics. J.W. was partly funded by Independent Research Fund Denmark (4183-00233B and 8020-00282B) and Danish Cancer Society (R147-Rp12977). N.L.-B. acknowledges funding from the European Research Council consolidator grant 682398) and Spanish Ministry of Economy and Competitiveness (SAF2015-66084-R, MINECO/FEDER, UE). We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for collation, realignment and harmonized variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

Author contributions This work was carried out by the PCAWG Drivers and Functional Interpretation Group based on data from the ICGC and TCGA PCAWG 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., J.M.H. and R.I.J. contributed equally. J. Weischenfeldt, R.B., I.M., J.S.P. and G.G. jointly supervised this work. The full list of PCAWG 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.M., J. 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., P.J.C., 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.L., 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.L., 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.v.M., G.T., H.H., 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.I.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.E.S., 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., C. Hong, C.W., E.K., E.M.L., E.R., F.A., G.G., G.T., H.H., H.M.U., J. Kim, J.M.H., J.S.P., K.D., L.F., L. Sieverling, M.M.N., M.S.L., N.S.-A., R.I.J., R.J. 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 analysis; 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., I.M., J.C.-F., J.R., J.S.P., K.I., Keunchil Park, L.M., L.D.S., L.U.-R., L.W., M. 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. Kahles., 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., L.W., 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.T., H.H., J.M.H., R.I.J., 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.

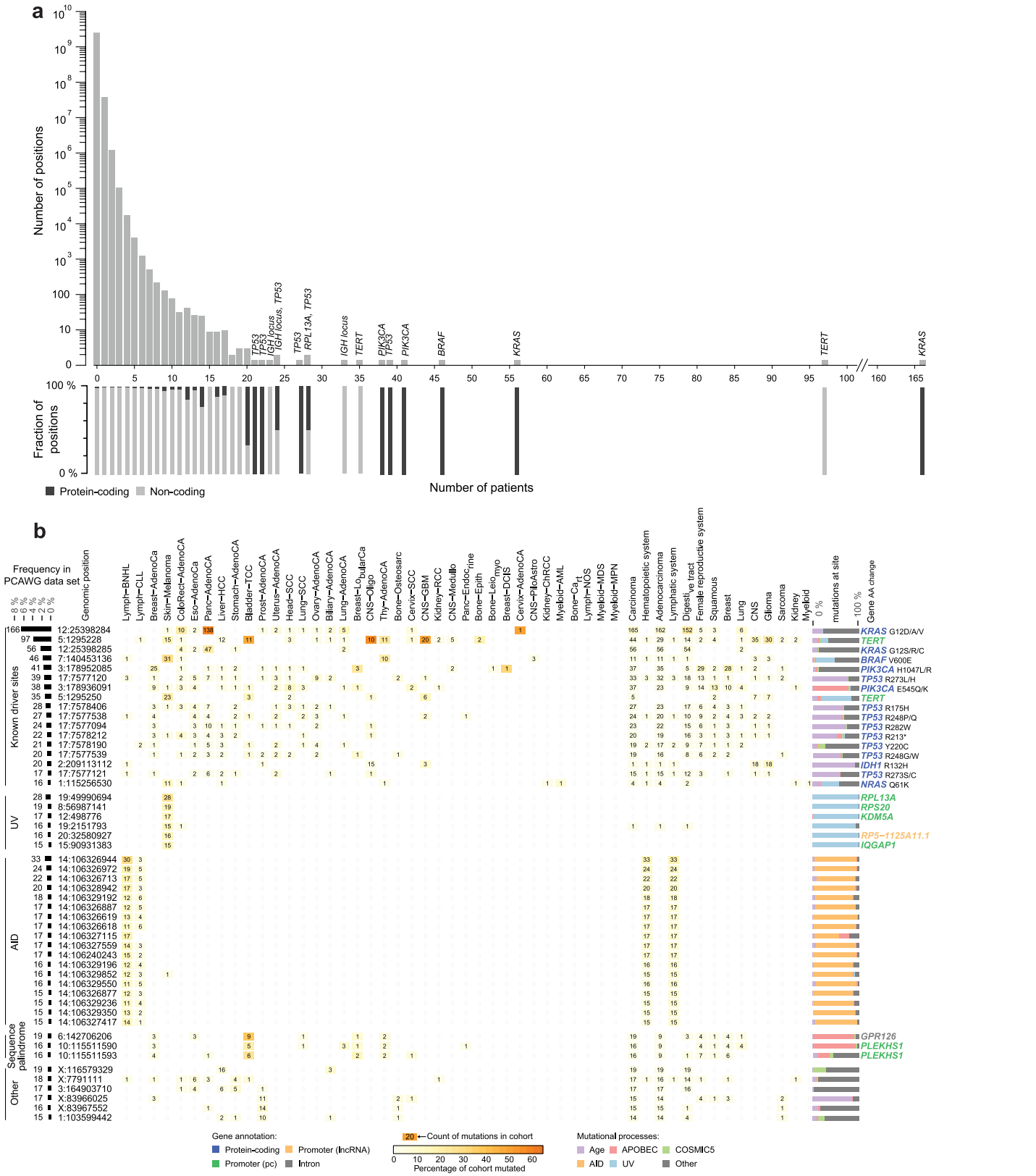
Competing interests The following authors declare that they have competing interests. P.B. receives grant funding from Novartis from an unrelated project; R.B. owns equity in Ampressa Therapeutics and receives grant funding from Novartis; G.G. receives research funds from IBM and Pharmacyclics and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect, MSMutSig 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

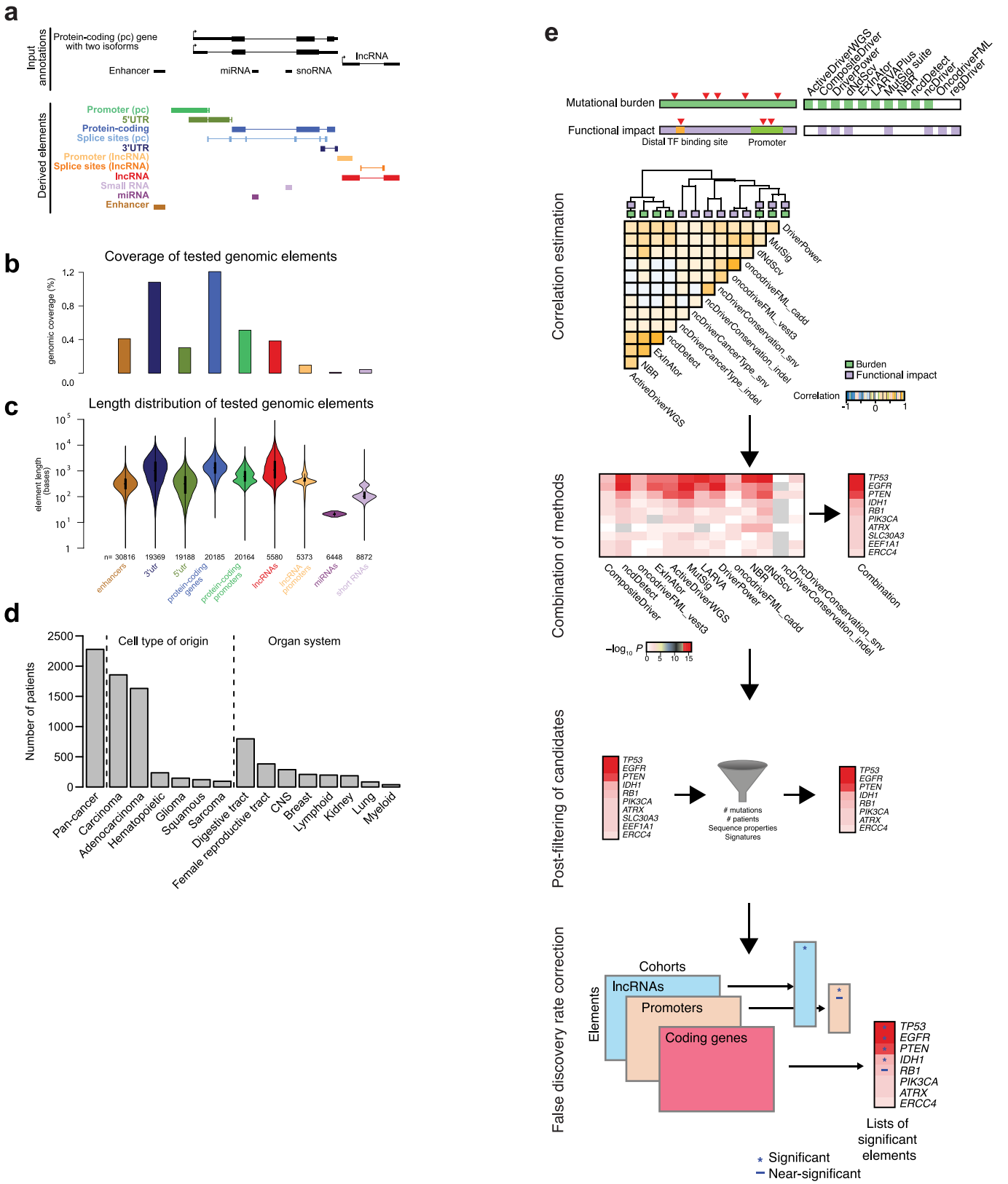
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-1965-x.

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.

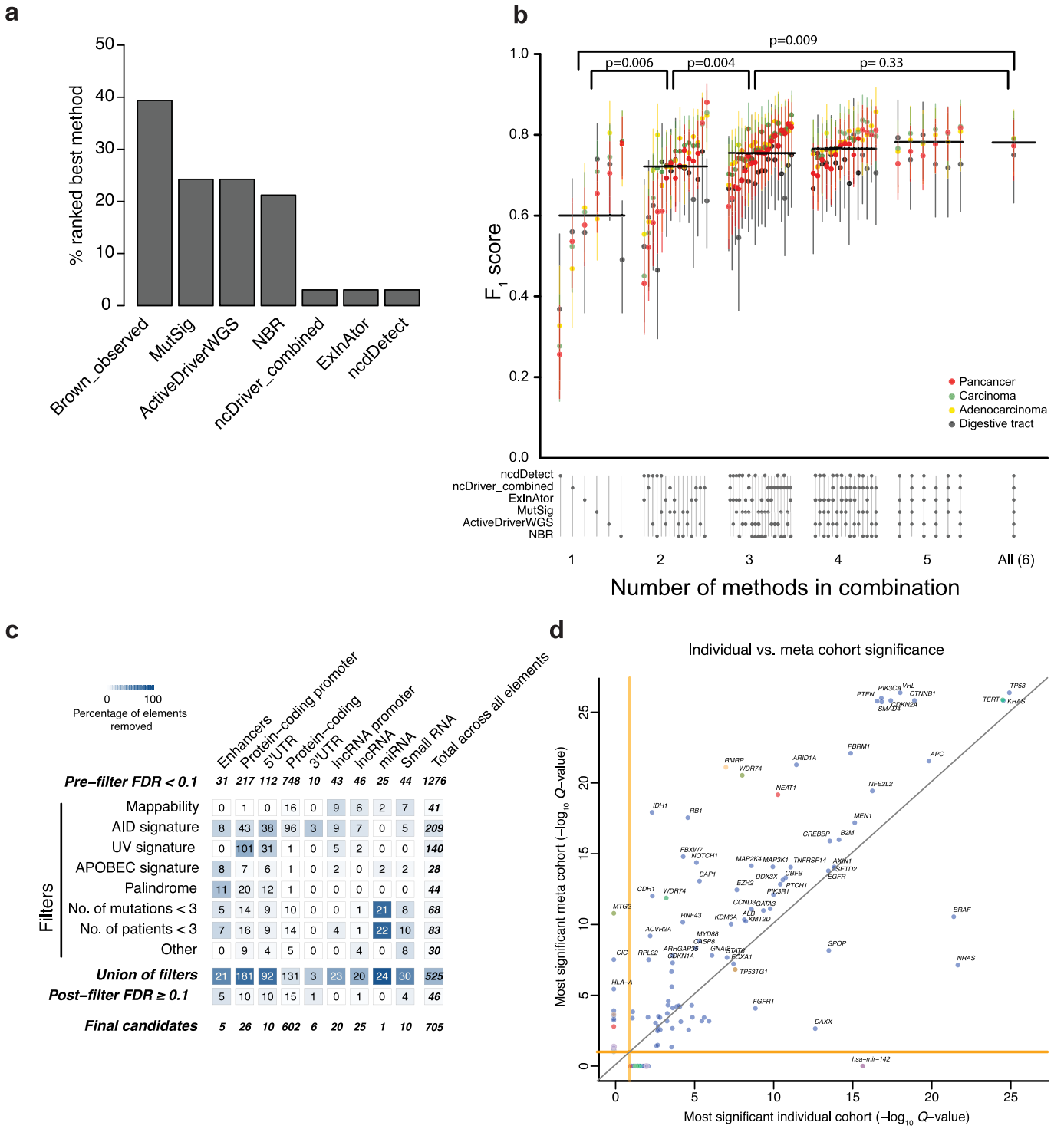


Extended Data Fig. 1 | Mutational hotspots in additional tumour types. a, Bar plot of number of positions (*y* axis) mutated in *n* patients (*x* axis). The stacked bar charts under the bar plot show the proportion of protein-coding (dark grey) and non-coding (light grey) positions. **b**, Distribution of SNVs in top 50 single-site hotspots across all analysed individual cohorts and meta-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.

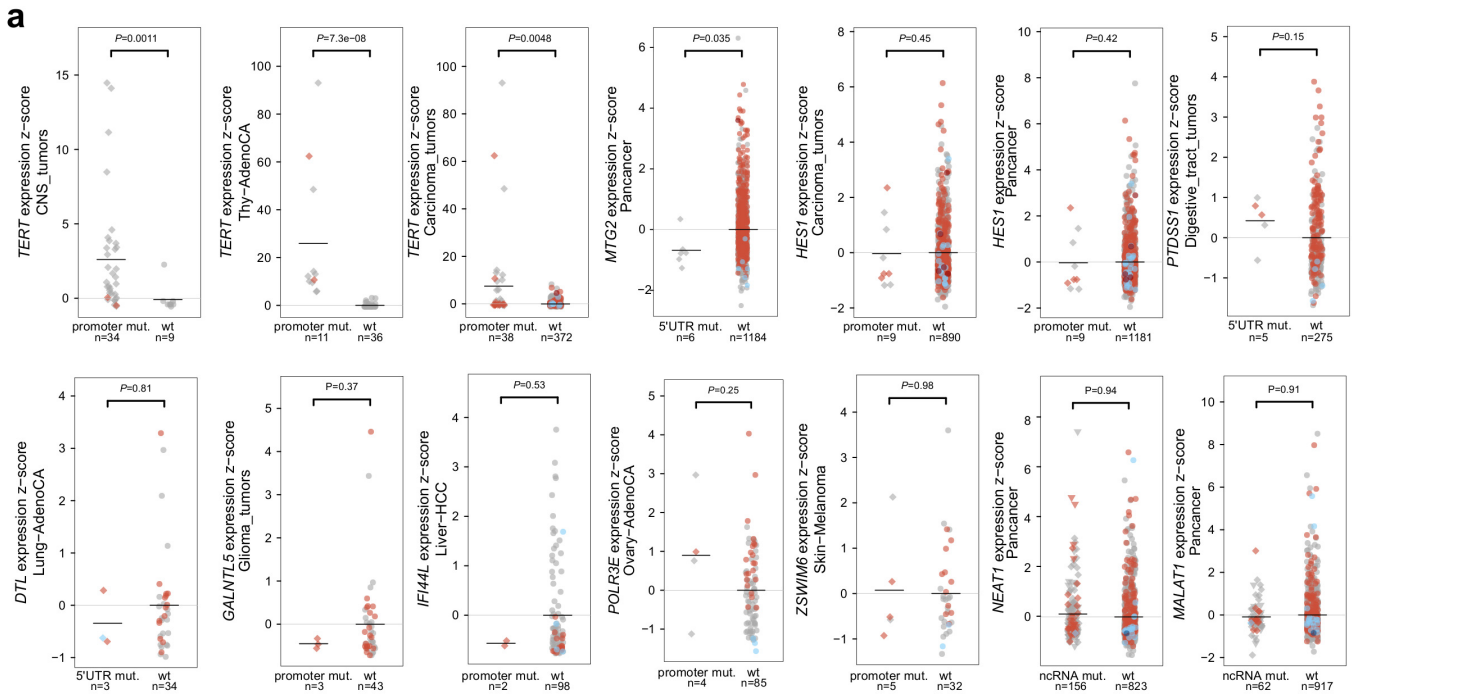


Extended Data Fig. 2 | See next page for caption.

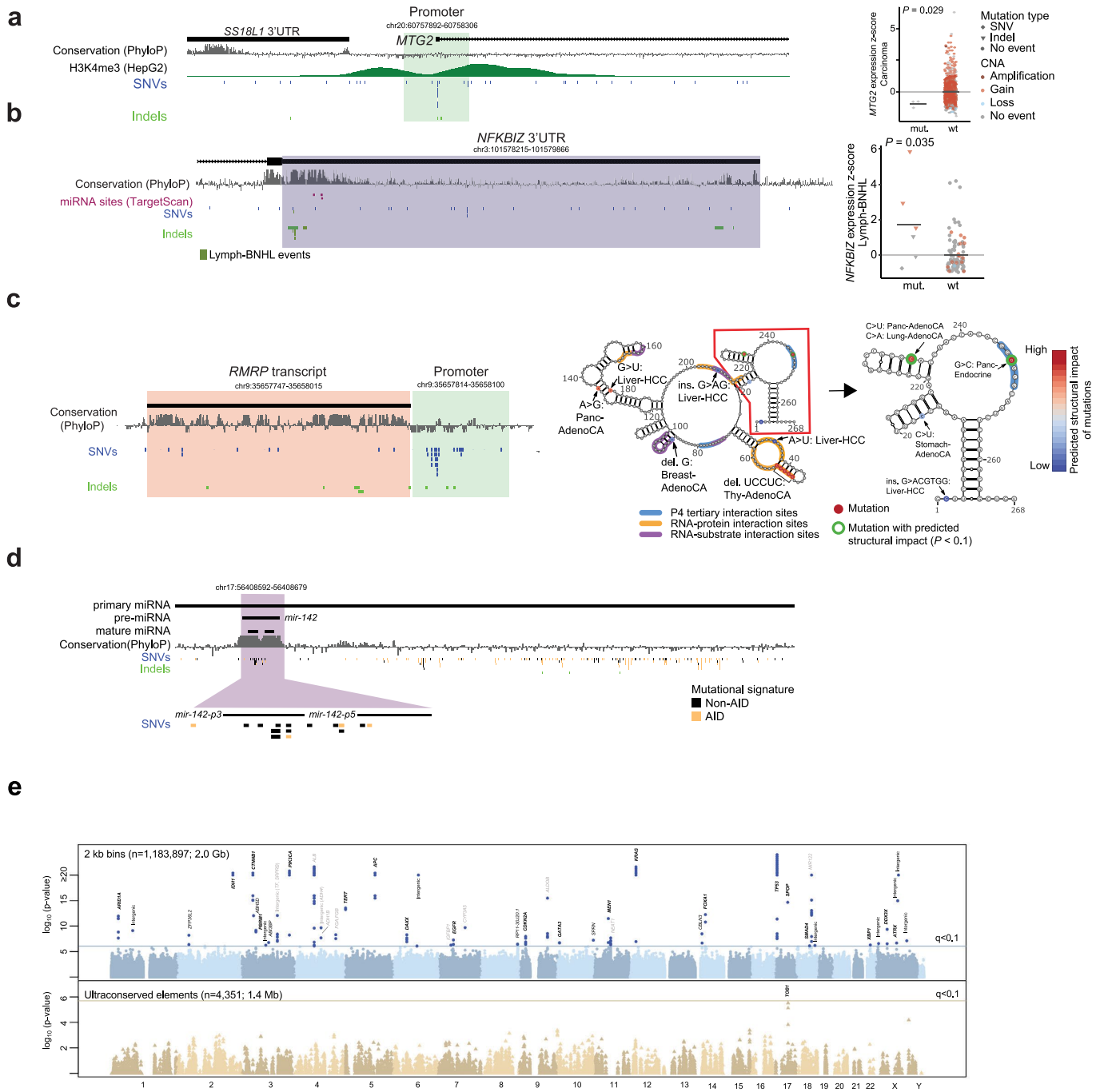
Extended Data Fig. 2 | Element-based driver discovery and combination of *P* values. a, Schematic describing definition of types of functional element (Methods). Functional elements (black) are defined on the basis of transcript annotations from various databases. Elements arising from multiple transcripts 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.



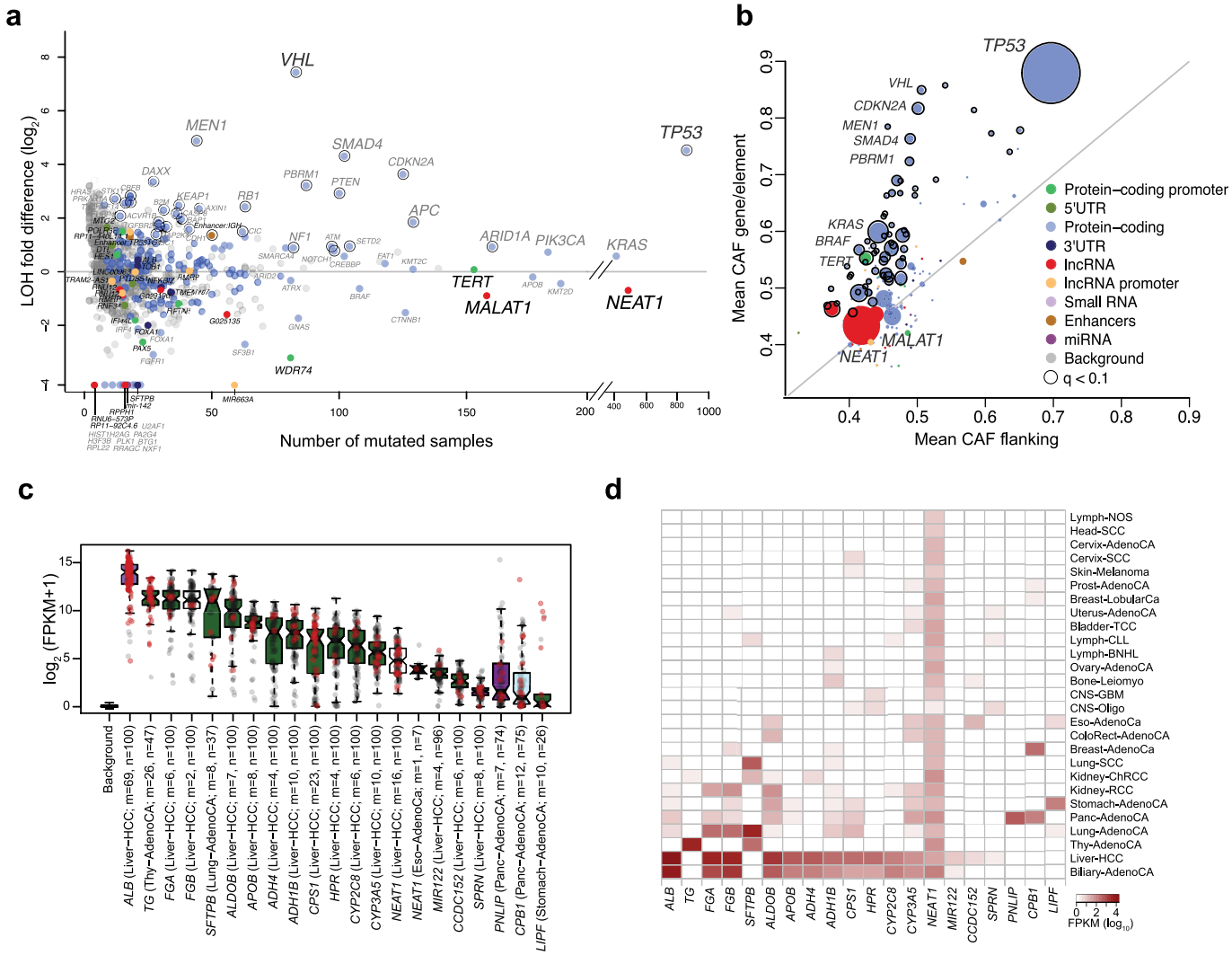
Extended Data Fig. 3 | Sensitivity of driver-discovery methods and filter statistics. a, Percentage of coding-driver discovery runs (with stable F_1 score, $n = 33$), across all cohorts, in which the method had the highest F_1 score (Methods). **b**, F_1 score of different methods of driver discovery, 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).



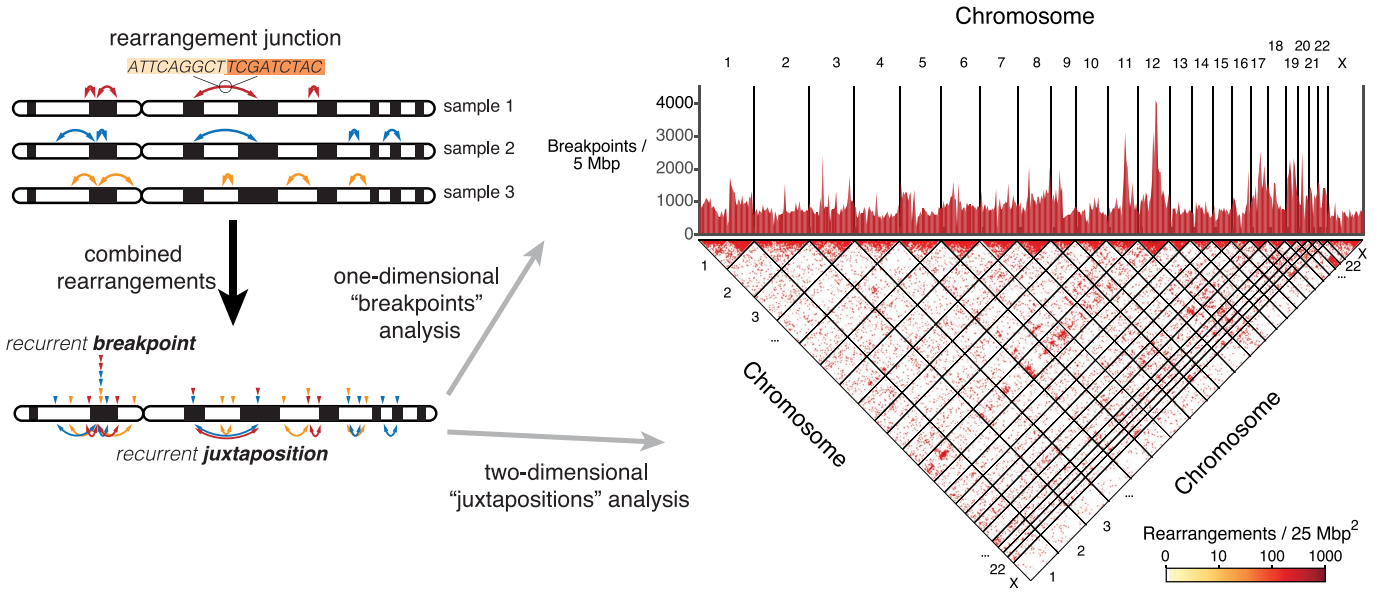
Extended Data Fig. 4 | Mutation-to-expression correlation and focal copy-number alterations. a, Expression is compared between mutated and non-mutated samples. For each element, the *z* score of the expression values for mutated and wild type in the significant cohort is plotted. For copy number, CNA 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 *WFIKKN2* 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).



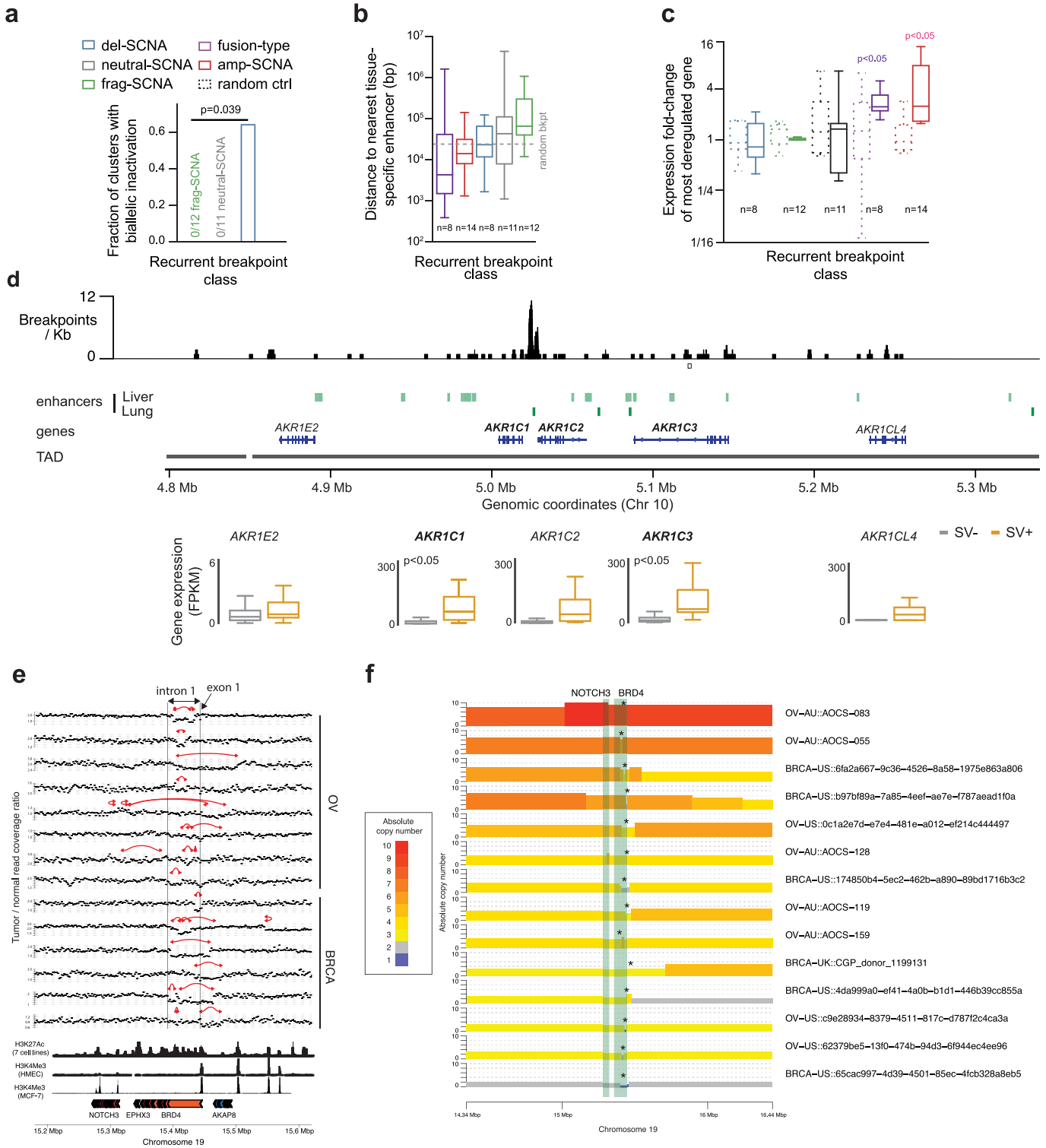
Extended Data Fig. 5 | Non-coding driver candidates. a, *MTG2* promoter locus (left) and associated gene-expression changes in carcinoma tumours (right). Expression of *MTG2* in mutated ($n = 3$) versus the carcinoma meta-cohort wild-type cases ($n = 896$). Two-sided Wilcoxon rank-sum test. Bars represent 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).



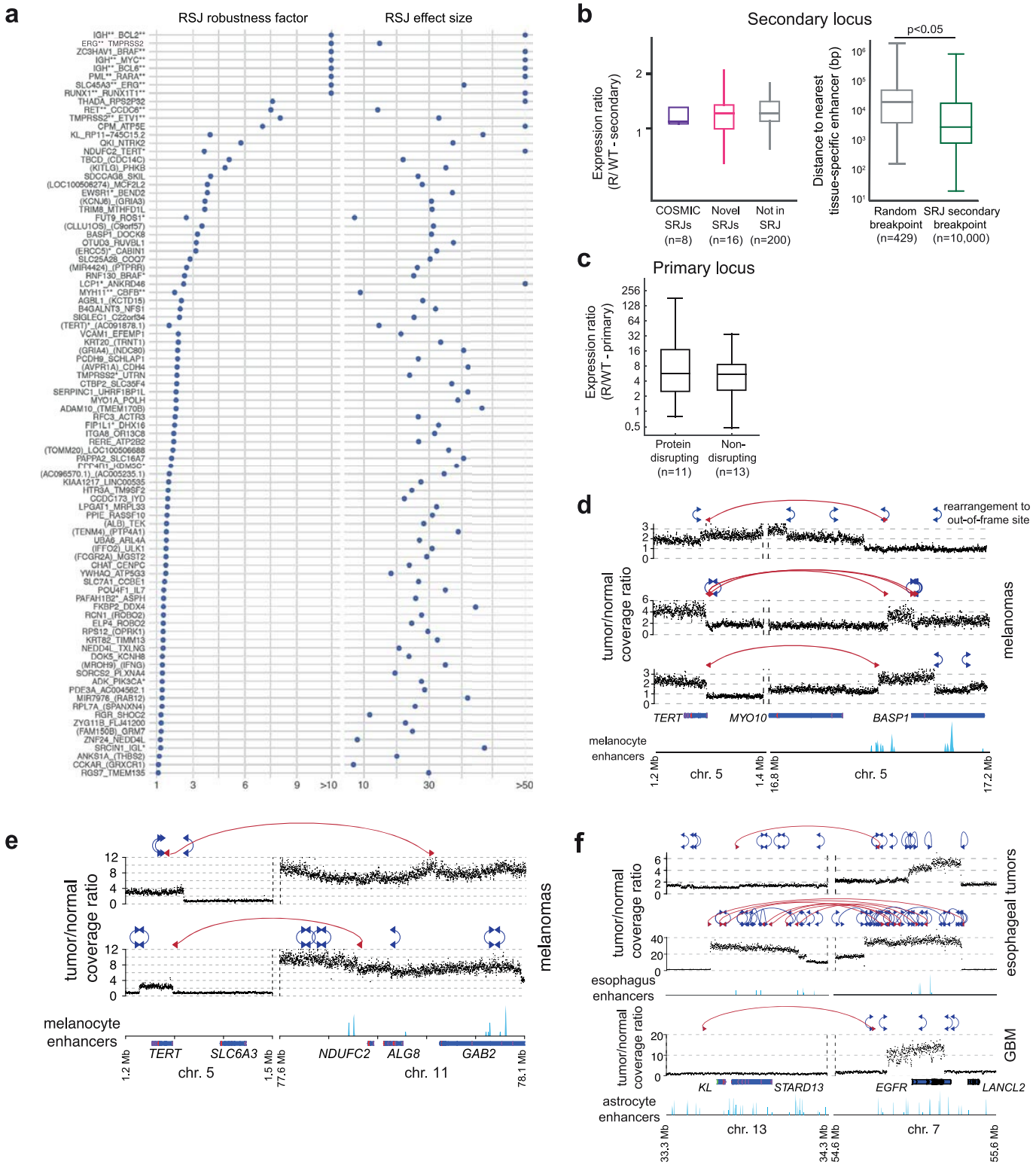
Extended Data Fig. 6 | A transcriptional process creates passenger mutations in highly expressed, tissue-specific genes. a, Relative rate of loss-of-heterozygosity (LOH) compared between mutated and wild-type samples for all significant elements, coloured by element type and highlighting significant 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 (±2 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).

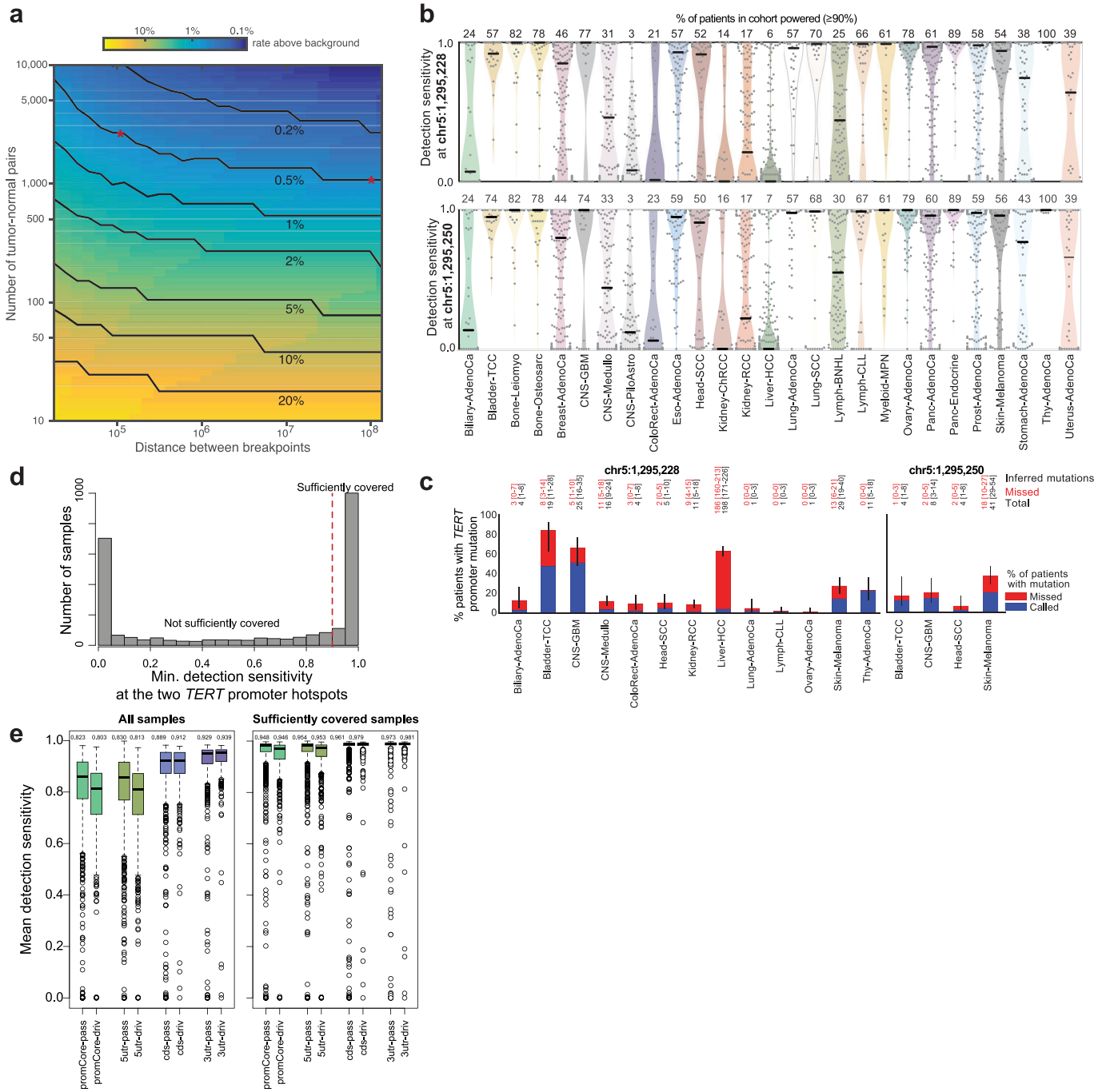


Extended Data Fig. 8 | Gene-expression effects of SRBs. a, Fraction of recurrent breakpoint loci associated with biallelic inactivation of a known tumour suppressor gene (frag-SCNA, 0/12; neutral-SCNA, 0/14; del-SCNA, 5/8; Fisher's exact test). **b**, Distance in bp to the nearest tissue-specific enhancer 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.

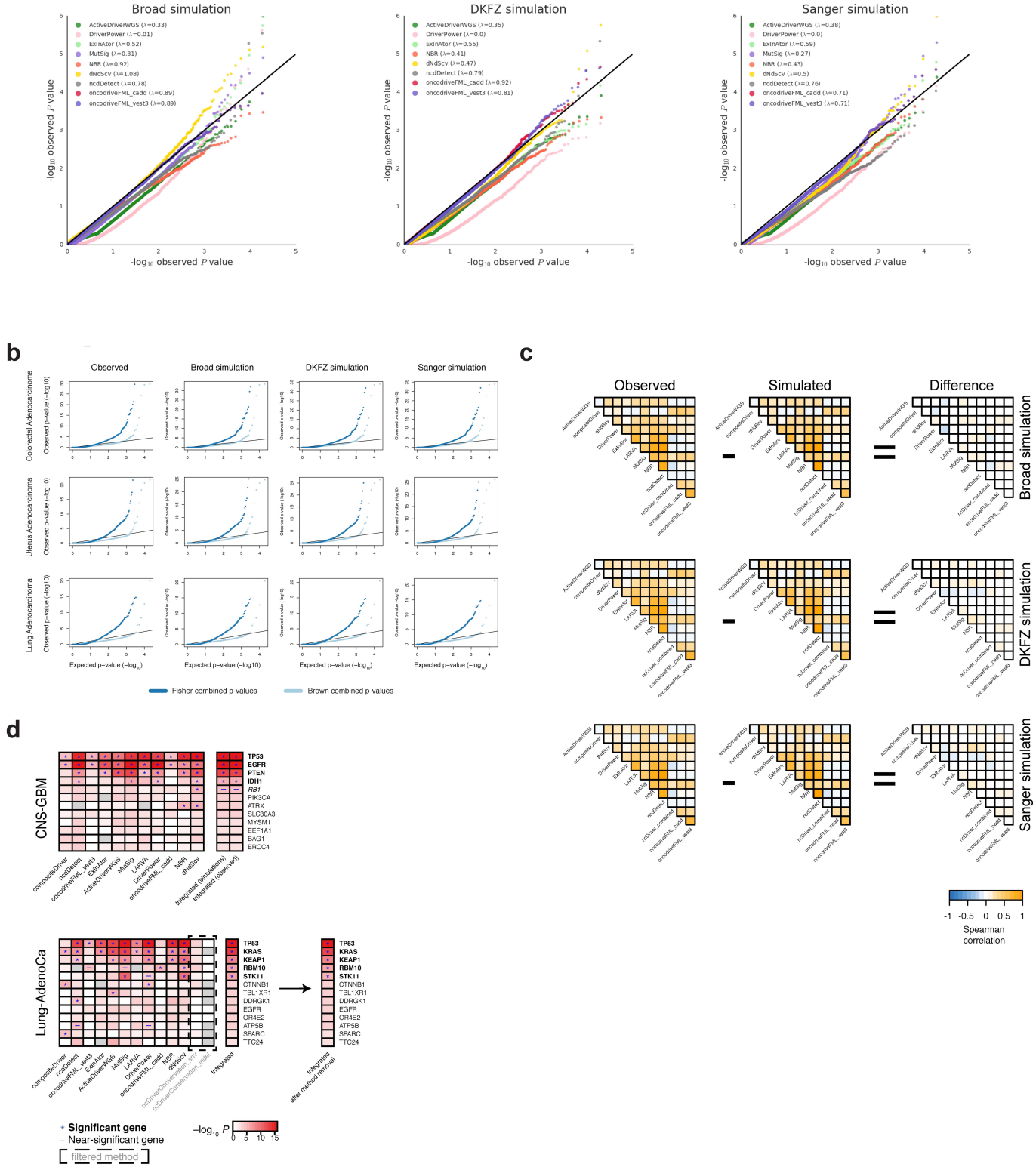


Extended Data Fig. 9 | See next page for caption.

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**.

nature research

Corresponding author(s): Gad Getz

Last updated by author(s): Oct 18, 2019

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Data and metadata were collected from International Cancer Genome Consortium (ICGC) consortium members using custom software packages designed by the ICGC Data Coordinating Centre. The general-purpose core libraries and utilities underlying this software have been released under the GPLv3 open source license as the "Overture" package and are available at https://www.overture.bio. Other data collection software used in this effort, such as ICGC-specific portal user interfaces, are available upon request to contact@overture.bio.
Data analysis	The workflows executing core WGS alignment, QC and variant-calling software are packaged as executable Dockstore images and available at: https://dockstore.org/search?labels.value.keyword=pcawg&searchMode=files. Individual software components are as follows: BWA-MEM v0.78.8-r455; DELLY v0.6.6; ACEseq v1.0.189; DKFZ somatic SNV workflow v1.0.132-1; Platypus v0.7.4; ascatNgs v1.5.2; BRASS v4.012; grass v1.1.6; CaVEMan v1.50; Pindel v1.5.7; ABSOLUTE/JaBbA v1.5; SvABA 2015-05-20; dRanger 2016-03-13; BreakPointer 2015-12-22; MuTect v1.1.4; MuSE v1.0rc; SMuFIN 2014-10-26; OxoG 2016-4-28; VAGrENT v2.1.2; ANNOVAR v2014Nov12; VariantBAM v2017Dec12; SNV-Merge v2017May26; SV-MERGE v2017Dec12; DKFZ v2016Dec15 The 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/mskilab/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/walaj/ginseng. The method for identify significantly recurrent rearrangements is available as the "2D" method at: https://github.com/ofershapira/SVsig. The method for identifying cis-expression consequences of SVs, CESAM, is available at: https://bitbucket.org/weischen/cesam

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

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

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	We compiled an inventory of matched tumour/normal whole cancer genomes in the ICGC Data Coordinating Centre. Most samples came from treatment-naïve, primary cancers, but there were a small number of donors with multiple samples of primary, metastatic and/or recurrent tumours. Our inclusion criteria were: (i) matched tumour and normal specimen pair; (ii) a minimal set of clinical fields; and (iii) characterisation of tumour and normal whole genomes using Illumina HiSeq paired-end sequencing reads. 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% (CI90%: 88-98%) and 95% (CI90%: 71-99%) respectively for SNVs. For somatic indels, sensitivity and precision were 60% (34-72%) and 91% (73-96%) respectively. Regarding SVs, we estimate the sensitivity of the merging algorithm to be 90% for true calls generated by any one caller; precision was estimated as 97.5% - that is, 97.5% of SVs in the merged SV call-set have an associated copy number change or balanced partner rearrangement.
Randomization	No randomisation was performed.
Blinding	No blinding was undertaken.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ 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-0-3 tumour site controlled vocabulary. Overall, the PCAWG data set comprises 38 distinct tumour types. While the most common tumour types are included in the dataset, their distribution does not match the relative population incidences, largely due to differences among contributing ICGC/TCGA groups in numbers sequenced.
Recruitment	Patients were recruited by the participating centres following local protocols.
Ethics oversight	The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local arrangements for ethics oversight and regulatory alignment.

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