The Origins and Consequences of Localized and Global Somatic Hypermutation

Fouad Yousif[1*], Stephenie D. Prokopec[1*], Ren X. Sun[1,2*], Fan Fan[1], Christopher M. Lalansingh[1], David H. Park[1], Lesia Szyca[1], PCAWG Network, Paul C. Boutros[1,2,3,x]

[1] Ontario Institute for Cancer Research, Toronto, Canada

[2] Department of Pharmacology & Toxicology, University of Toronto, Toronto, Canada

[3] Department of Medical Biophysics, University of Toronto, Toronto, Canada

* Joint first authors

x Corresponding author

Address for correspondence:
Dr. Paul C. Boutros
Ontario Institute for Cancer Research
661 University Avenue, Suite 510
Toronto, Ontario, Canada,
M5G 0A3
Email: Paul.Boutros@oicr.on.ca
Phone: 416-673-8564

Abstract

Cancer is a disease of the genome, but the dramatic inter-patient variability in mutation number is poorly understood. Tumours of the same type can differ by orders of magnitude in their mutation rate. To understand potential drivers and consequences of the underlying heterogeneity in mutation rate across tumours, we evaluated both local and global measures of mutation density: both single-stranded and double-stranded DNA breaks in 2,460 tumours of 38 cancer types. We find that SCNAs in thousands of genes are associated with elevated rates of point-mutations, while similarly point-mutation patterns in dozens of genes are associated with specific patterns of DNA double-stranded breaks. These candidate drivers of mutation density are enriched for known cancer drivers, and preferentially occur early in tumour evolution, appearing clonally in all cells of a tumour. To supplement this understanding of global mutation density, we developed and validated a tool called SeqKat to identify localized “rainstorms” of point-mutations (kataegis). We show that rates of kataegis differ by four orders of magnitude across tumour types, with malignant lymphomas showing the highest. Tumours with *TP53* mutations were 2.6-times more likely to harbour a kataegic event than those without, and 239 SCNAs were associated with elevated rates of kataegis, including loss of the tumour-suppressor *CDKN2A*. We identify novel subtypes of kataegic events not associated with aberrant APOBEC activity, and find that these are localized to specific cellular regions, enriched for MYC-target genes. Kataegic events were associated with patient survival in some, but not all tumour types, highlighting a combination of global and tumour-type specific effects. Taken together, we reveal a landscape of genes driving localized and tumour-specific hyper-mutation, and reveal novel mutational processes at play in specific tumour types.

Introduction

Genome instability is one of the hallmarks of cancer, and cancer is often referred to as a "disease of the genome"[1]. Just as cancers are heterogeneous over time and space[2-4], they are also heterogeneous in the number of mutations they harbour -- their "mutational density". Some tumour types, like melanomas and lung cancers, harbour hundreds of thousands or even millions of single nucleotide variants (SNVs). These large mutational burdens are thought to reflect the effects of environmental carcinogens, like UV radiation and the by-products of cigarette smoke. Other tumour types, like prostate cancers, can harbour only a few hundreds SNVs[5-11]. This variability in SNV mutational density across tumour types has been well-demonstrated in previous pan-cancer exome sequencing studies[12,13].

But it is less clear what drives some tumours to harbour more DNA damage than others. Even within an individual cancer type, individual tumours of similar clinical grade and stage can vary by orders of magnitude in their number of SNVs. Theoretical modeling studies have attributed large fractions of the divergence across tumour types to differences in the number and rate of replication-induced errors, rather than the effects of heredity or environmental influences[14-16]. What drives these differences in mutational density at the SNV level? And are these differences in mutational density in point-mutations reflected by similar trends in somatic copy number aberrations (SCNAs), translocations and other copy-neutral structural variants (SVs) and in localized hypermutation events like kataegis?

To address these questions, we evaluated both local and global measures of mutation density: both single-stranded and double-stranded DNA breaks in 2,460 tumours of 38 cancer types. We then identify individual genes whose mutation status is associated with changes in the mutation density of different types of aberrations. For example, we identify SNVs in dozens of genes as associated with changes in the burden of copy number aberrations and the number of copy-neutral SVs identified. These candidate drivers of mutation density are enriched for known cancer drivers, and preferentially occur early in tumour evolution, appearing clonally in all cells of a tumour. To supplement this understanding of global mutation density, we developed and validated a tool called SeqKat to identify localized "rainstorms" of point-mutations (kataegis). We show that rates of kataegis differ by four orders of magnitude across tumour types, with malignant lymphomas showing the highest. Tumours with *TP53* mutations were 2.6-times more likely to harbour a kataegic event than those without, and 239 SCNAs were associated with elevated rates of kataegis, including loss of the tumour-suppressor *CDKN2A*. We identify novel subtypes of kataegic events not associated with aberrant APOBEC activity, and find that these are localized to specific cellular regions, enriched for MYC-target genes. Kataegic events were associated with patient survival in some, but not all tumour types, highlighting a combination of global and tumour-type specific effects. Taken together, we reveal a landscape of genes driving localized and tumour-specific hyper-mutation, and reveal novel mutational processes at play in specific tumour types.

Results

Experimental Design

The PCAWG project evaluated 2,703 tumours from 37 histological subtypes[17] (https://dcc.icgc.org/pcawg#!%2Fmutations). Samples were excluded if they were not flagged for removal, if they represented additional specimens from a single donor, if donor sex was unknown or if donor age was unknown. This left 2,460 tumours for which we evaluated the density of single-stranded breaks, double-stranded breaks, hypermutation and kataegic events (**Supplementary Figure 1; Supplementary Table 1**). We note that the selection procedures for tumours in PCAWG have biased the cohort towards larger tumours that were surgically managed and yielded sufficient DNA for sequencing, and that the PCAWG marker paper outlines variant calling, coverage and other technical aspects[17-19].

Drivers of Single Nucleotide Variation

We first examined the density of SNVs within the PCAWG cohort, and observed a broadly consistent number of SNVs across the genome (chromosomes 1-22 and X), with a median of 12,950 SNVs within each 1 Mbp bin (n = 2,450; **Figure 1A**). After controlling for differential coverage across samples, hypermutated bins did not encompass known driver genes, but rather reflected other sources of mutational burden like replication-timing effects and environmental exposures (**Figure 1B, Supplementary Figure 2**)[8,20,21]. We confirm the relatively small intra-tumoural heterogeneity seen in previous studies[6,22-25], with 99.3% of variance in SNV mutational burden between tumour types and only 0.7% of variance within tumour types. This overall variance masks significant differences between tumour types. For example at the extremes, in melanoma the median SNV mutation density of a tumour was 26.4 SNVs/Mbp, but this varied dramatically (SD = 54.25, IQR = 45.59). By contrast, in pilocytic astrocytomas mutation density was both lower (0.06 SNVs/Mbp) and much more consistent across tumours (SD = 0.08, IQR = 0.076; **Supplementary Table 2**). Thus not only does mutation density vary significantly between tumour types, but so too does its consistency within tumour types.

To understand the drivers of increased point mutation density, we assessed the relationship between somatic copy number aberrations (SCNAs) and SNV mutational density (SNVs/Mbp) in each tumour that had both data types available. Linear mixed effects modeling was used to quantify the effects of individual SCNAs on SNV mutational density, while controlling for confounding variables including tumour-type, patient sex and patient age (**Supplementary Table 3**). Because of our robust statistical power, we identified almost half of all coding genes as significantly associated with changes in SNV mutational density (FDR < 0.01). Gains in 7,580 different genes, including *ATM*, *ATR*, *PIK3CA* and *MYC*, were associated with increased SNV mutational density. By contrast, copy number gain of only a single gene (*MUC3A*) was associated with decreased SNV mutational density. Similarly, copy number loss of 1,494 genes were associated with increased SNV mutational density, including classic cancer driver genes like *BRCA2*, *NF1*, *APC* and *PBRM1* (**Figure 1C (top), Supplementary Figure 3**). Interestingly, an additional 216 genomic regions, comprising 1,083 genes (including *NFE2L2*) were positively associated with SNV mutational density, regardless of CN type. For these genes, either a gain

or a deletion was associated with elevated rates of SNV mutational density (**Supplementary Table 4; Supplementary Figure 4**). These associations may represent variant-calling artifacts, genes with divergent effects across tumour types, or other phenomena.

To focus on the best candidate drivers of changes in SNV mutation density, we focused on statistically-significant genes with large effect sizes: some individual genes were associated with very large changes in SNV mutational density. For example, gain of the nuclear-encoded mitochondrial gene *BCKDHB* was associated with an increase of 1.62 SNVs/Mbp in a pan-cancer way (**Figure 1D, left panel**). This corresponds to ~5,000 additional SNVs genome-wide in a tumour with amplification of *BCKDHB*. To confirm our multivariable modeling identified trends that occur in individual tumour types as well as pan-cancer, we performed subgroup analyses on the five tumour types with the highest individual sample number: medulloblastoma, renal cell carcinoma (RCC), hepatocellular carcinoma (HCC), pancreatic and prostate adenocarcinomas. *BCKDHB* was associated with a 1.5 SNVs/Mbp elevation in mutation rates in medulloblastomas, despite the smaller sample size and concomitantly reduced statistical power of this sub-group analyses. Similarly, deletion of the mitochondrial creatine kinase (*CKMT2*) was associated with an increase in SNV mutational density of 1.59 SNVs/Mbp, again corresponding to ~5,000 additional SNVs genome-wide. Interestingly, while CN losses in *CKMT2* were associated with an increase of 1.9 SNVs/Mbp in prostate adenocarcinoma (FDR = 0.095) as observed pan-cancer, *CKMT2* amplification led to increased SNV mutational density to 1.64 SNVs/Mbp in medulloblastoma (FDR = 0.028; **Figure 1E, right panel**). Thus individual gene-wise SCNAs are associated with large-scale changes in SNV mutational burden, both pan-cancer and within individual tumour types.

Globally, genes with SCNAs associated with increased SNV mutational density were enriched for nucleosome and chromatin assembly, and peptide cross-linking (CN gains) or protein catabolic processes (CN deletions; **Supplementary Table 5**). They did not preferentially locate to specific chromosomal regions (**Supplementary Table 6**) and typically originated in subclonal tumour populations[26]. Of genes with SCNAs associated with SNV mutational density, 206 contained deletions identified as clonal mutations in at least 50% of patients (**Supplementary Table 7**).

Finally, to better understand the tumour-type-specificity of candidate drivers of SNV mutational density, we performed genome-wide subgroup analyses on the five tumour types with the most samples: medulloblastoma, RCC, HCC, pancreatic and prostate adenocarcinomas. We replicated our mixed linear modeling strategy for each tumour-type independently, controlling for age and sex (where appropriate). Despite small sample sizes of 107-241, we detected 7,774 genes with SCNAs associated with changes to SNV mutational density in at least one tumour types (FDR < 0.1; **Figure 1C, Supplementary Table 8**). As before, effect-sizes could be very large, ranging from 0.24-8.4 SNVs/Mbp. Many genes (2,121) were associated with increased SNV mutational density in multiple tumour types, and these overlapped pan-cancer candidates. Taken together, these results uncover a landscape of pan-cancer and tumour-type-specific effects in driving changes in SNV mutational density. Indeed even the large PCAWG dataset employed here likely results in significant false-negative rates for smaller effect-size associations.

Drivers of Copy Number Changes

We next sought to reciprocally examine patterns of copy number gains and losses across tumour types. We considered a panel of copy-number summary features, including total SCNA count, proportion of the genome altered (PGA)[27] and average SCNA length. These were further sub-categorized by the direction of change (gain, loss or overall; **Supplementary Table 1**).

We find that the well-known variability in SNV mutational density (**Figure 1A**) are paired to even larger intra- and inter-tumour type variability in their SCNA alteration patterns. For example, individual tumours and tumour-types differed remarkably in their ratio of SCNA gains to losses for each tumour type (**Figure 2A**). Pilocytic astrocytomas harboured orders of magnitudes more gains than losses (median = 1.94×10^6 PGA gain:loss (autosomes only); SD = 2.90×10^8, IQR = 1.96×10^6), while prostate adenocarcinomas had ~7-fold more losses than gains (median = 0.01 PGA gain:loss (autosomes only); SD = 2.82×10^4, IQR = 0.3; **Supplementary Figure 5, Supplementary Table 2**). Even within individual tumour types, these metrics ranged dramatically, with some having relatively balanced gains and losses and others showing large bias. For example, HCC shows a balance of gains to losses (median ratio = 1), however with a large degree of variability around this (SD = 1.14×10^7, IQR = 2.58); similarly pancreatic adenocarcinomas show a median ratio of 0.55, with a SD of 3.80×10^5 due primarily to a single outlier (IQR = 0.79).

This picture of large inter- and intra-tumour type heterogeneity was not restricted to any single feature of the CN landscape, but held true for overall burden, for gains and losses separately and for many other features, with particularly large variability seen in the length distributions of SCNAs across tumour types (**Supplementary Figure 6**). In general, tumour types dominated by single base mutations as opposed to SCNAs were consistent with the M-class (those tumour types presenting predominantly with point mutations) and C-class (predominantly copy-number aberrations) tumours described previously[25].

To identify candidate drivers of this heterogeneity in CN mutational density, we again employed mixed effects modeling. In this case, we modeled specific features of the SCNA profile (**Supplementary Table 3**) using functional SNVs and again controlling for tumour type as well as patient sex and age. Given the relative paucity of recurrent SNVs, we focused on consensus driver genes identified by PCAWG[28]. We identified a diverse landscape of SNV-SCNA interactions (**Supplementary Figure 7**). In particular, point mutations in *TP53* were associated with nearly every metric tested, including an increased rate of losses (**Figure 2B**), increased total number of SCNAs (**Figure 2C**) and increased total PGA (**Figure 2D**). But *TP53* was not the only SNV associated with SCNA phenotypes. Point mutations in *VHL* were strongly associated with an increased number of losses relative to gains (**Figure 2B**), while those in *EZH2* and *DDX3X* were associated with reduced numbers of SCNAs (**Figure 2C**). To understand the tumour-type-specificity of these results, we again stratified our analysis and focused on the five highest-powered tumour types. The same linear mixed-modeling strategy with control for age and sex identified a large number of genes associated with multiple metrics in multiple tumour types. These results confirmed the pan-cancer findings, but also highlight a landscape of tumour-type specific effects. For example, SNVs in *ERBB4* were significantly associated with 7 metrics of copy number in all 5 tumour types, including elevated ploidy,

elevated total number of losses and an elevated PGA (**Supplementary Figure 8**). The tumour suppressor *CDKN2A* was, as expected, associated with an increased rate of SCNA loss in each of the five tumour types and in pan-cancer analyses. Some tumour types appeared to show stronger interactions between their SNV and SCNA landscapes: 14 genes were associated with PGA in only RCC, including *VHL* and *BRCA1* (**Supplementary Table 8**). These results suggest that in addition to shared biological pathways that increase SCNA mutational density, distinct molecular processes are also at play in different tumour types.

Drivers of Copy-Neutral Structural Variation

Next, we repeated the above process to assess copy-number neutral structural variants (SVs) of different types (**Supplementary Table 1**). In general, the mutational density of different types of SVs such as tail-to-tail inversions and translocations were tightly correlated in each tumour type (ρ = 0.81-1.00, p-value < 2.6×10^{-22}). BRCA-driven tumours (breast, ovarian, and uterine cancers) having high levels of both translocations (TRAs) and inversions (INVs). Alternatively, tumours often thought to be driven by fusion proteins, such as thyroid carcinoma or AML, had fewer total SVs (**Figure 3A, Supplementary Figure 9**).

To understand the spatial structure of these trends, we collapsed SV events into 1 Mbp bins across the genome and visualized those occurring in at least 10 patients (~0.4% recurrence; **Figure 3B**). While 149 small translocation hotspots were detected, there were two major translocation clusters: one on the q-arm of chromosome 12 (within a region containing *MDM2*) and a highly recurrent translocation between chromosomes 14 and 18 (within regions that include *YY1* on chromosome 14 and *BCL2*, *SMAD4* and *SMAD7* on chromosome 18).

Because of the high-correlations amongst different characteristics of the SV profile of tumours, we focused on identifying point mutations associated with increased translocation density (**Figure 3C**). As expected, *TP53* point mutations were associated with an increase in all SV types (**Figure 3C, Supplementary Figure 10**). By contrast, point mutations in multiple driver genes including *ARID1A*, *CTNNB1*, *KRAS* and *PIK3CA* were associated with reduced translocation burden. Point mutations within these genes most often appear clonal in these tumours (i.e. appear in its trunk); this suggests that these are strong SNV driver events that reduce the likelihood of translocations. These trends are highly tumour-type specific, even more so than were associations for SNV and SCNA mutational density. For example, *KRAS* point mutations are associated with decreased translocation burden in RCC, but with increased translocation burden in many other tumour types (**Supplementary Table 8; Supplementary Figure 11**), however in tumour types that present with a *KRAS* SNV, it is usually an early (clonal) event (**Supplementary Table 7**).

Kataegis

To determine if these broad trends in global mutation burden are mirrored in local mutational hotspots, we focused on kataegis: localized hypermutation of somatic SNVs[29]. Kataegis can be caused by APOBEC-mediated events[30], although this does not account for all occurrences. Kataegis is an important signature of genomic instability, independent of other somatic variants[31,32]. In order to reproducibly and reliability identify kataegic events and quantify its'

extent in individual tumours, we created an open-source tool called **SeqKat**. SeqKat predicts kataegis from paired tumour and normal human genome samples by using a sliding window approach to test deviation of observed SNV trinucleotide content and inter-mutational distance from that expected by chance, after adjusting for the effects of trinucleotide signature and mutation rate (**Supplementary Figures 12-13**). The resulting kataegic score estimates the magnitude of the event, and accounts for features like the deviation from expected C/T base change frequency and the expected inter-SNV distance within each window.

We applied SeqKat to all PCAWG tumours with consensus SNV calls and detected 83,489 events. Overall 56.8% of patients had at least one kataegic event (1,467/2,583 patients; **Supplementary Table 1**), but different tumour types varied dramatically in their number and magnitude. Bladder carcinomas had the highest rate of kataegis, with a median 22 events per tumour, while 14 different tumour types had a median zero events per tumour (**Figure 4A**). These events also differed dramatically in their length and enrichment for APOBEC-activity, with non-hodgkins lymphomas and squamous carcinomas of the lung showing the strongest events and adenocarcinomas of the stomach and glioblastomas being amongst the weakest (**Figure 4B**). Again, these differences were very large in magnitude, with median event size and score differing by orders of magnitude across tumour types. This inter-tumour-type variability was reflected within tumour types as well, with non-hodgkins B-Cell lymphomas showing the highest variance among in kataegis. Individual events could be remarkably strong, with massive enrichment of the classic APOBEC-associated TCX mutations (**Figure 4C; Supplementary Figure 14**). This heterogeneity confirms previous trends in kataegic incidence[11,12,29,32-35] and expands them to dozens of additional tumour types, while describing surprising differences in the strength and extent of individual events.

To better understand why kataegic events afflict some tumours and regions of the genome more than others, we integrated the SeqKat calls with translocations and RNA-seq data. Chromosomes 4, 5, 8, 18 were enriched for kataegic events relative to chance expectations (q-values 5.73×10^{-10}, 0.0019, 2.34×10^{-8}, and 3.63×10^{-5} respectively; **Figure 4D**). There were 3,479 genes affected by kataegic events in at least 2 patients, ranging from 1 to 11 tumour types (**Figure 4E**). *CNTNAP2* harbored the highest number of total kataegic events (n = 304) across 34 patients in eight different tumour types (Skin-Melanoma, ColoRect-AdenoCA, Lung-SCC, Bladder-TCC, Uterus-AdenoCA, Panc-AdenoCA, Ovary-AdenoCA, and Bone-Osteosarcoma). This gene along with three other kataegis-enriched genes *LRRC4C*, *CNTN5*, and *CSMD1* showed significant associations with mutational signatures 7a-d (**Supplementary Table 9**), characteristic of kataegic events (FDR < 0.01). When looking at the number of samples with kataegic genes, we observed seven kataegic hotspots within the genes *IGH*, *IGK*, *BCL2*, *BCL6* and *MYC*. These were exclusively present in lymphomas and are enriched with mutation signature 7a-d (FDR = 3.09×10^{-6}). These genes are either targets for somatic hypermutation (SMG) or aberrant somatic hypermutation (aSHM) in B-Cells[36]. AID and APOBEC editing deaminases play an important role in the initiation of hypermutation and recombination of immunoglobulin genes in B-Cells, which are essential processes for the recognition and disposal of pathogens[37], explaining the high kataegic rate of *IGK* and *IGH* genes. During that process however, AID has been shown to aberrantly target oncogenes and tumour suppressors such as *BCL6* and *MYC*[38]. We assessed the kataegic rate for 56,827

genes, normalizing for gene length, and identified 451 kataegic enriched genes that are potential targets of aSHM in lymphoma (**Supplementary Table 9**).

Aberrant processing by AID also leads to the introduction of translocations via double-stranded breaks required by the repair process[39]. We examined the translocation breakpoints in PCAWG lymphoma patients and found that many of these breakpoints overlapped with kataegic hotspots (**Figure 5A**). One example of such a translocation is *MYC-IGH* (chr8-chr14) that classically identifies Burkitt's lymphoma: patients with this translocation had an enrichment in kataegis events around the translocation breakpoints (10 of 14 patients). Translocated *MYC* has a consistently smaller kataegic score and mutation frequency compared to the translocated *IGH*, confirming that *MYC* kataegic events occurred after the translocation, while under regulation by Ig regulatory elements[40]. We also assessed the effect of kataegic events in *MYC* on its transcriptional patterns. Kataegic *MYC* had significantly higher ($p = 0.005$) FPKM compared to wild-type *MYC* (**Figure 5B**). This form of MYC deregulation has been observed before in B-cell lymphoma cell-lines and can be caused by the complex insertional rearrangements, three way recombinations of *MYC-IGH-BCL2*, and *IGH-MYC* fusions[41,42]. The presence of IG genes transcriptional enhancers in such translocations lead to deregulation of MYC[43]. Concordantly, mRNA abundance of most members of the AID/APOBEC family of deaminases was significantly higher in kataegic samples than non-kataegic ones (**Supplementary Figure 15**), including the transcription factor PAX5, which is involved in AID upregulation[44].

We then assessed the relationship between individual gene-specific mutations and different measures of kataegis in each tumour to help us evaluate its causes and consequences. We again used linear mixed effects modelling to associate the mutation status of each gene with its effect on kataegic mutational density measures adjusting for sex and age, and fitting tumour type as a random effect. This allowed us to assign to each mutated gene, the increase or decrease in the kataegic density of tumours with that mutation, relative to those without (**Supplementary Table 3**). Point mutations in *TP53* were remarkably predictive of increased kataegic burden (FDR = 1.81×10^{-10}, effect size = 0.97; **Figure 5C; Supplementary Figure 16A-B**). Tumours harbouring a *TP53* mutation are 2.6x more likely to have at least one kataegic event then those without (**Supplementary Figure 17**). While *TP53* was the only SNV associated with increased risk of kataegis, a number of CN deletions (but not amplifications) showed a strong association (FDR < 0.01; **Supplementary Table 3, Supplementary Figure 16C-D**). Patients with a deletion in *IFNA5*, for example, are 2.45x more likely to have at least one kataegic event (FDR = 3.23×10^{-4}, effect size = 0.90). We confirmed 103 kataegis-associated SCNAs in at least one of the five tumour types with the most samples (FDR < 0.01 pan-cancer, FDR < 0.15 per tumour; **Supplementary Table 8**).

Kataegis status, identified through an expression signature, has recently been shown to be associated with late onset, better prognosis and higher HER2 levels in breast cancer[45]. To further investigate the prognostic role of kataegis, a Cox regression was fit for overall survival between kataegic and kataegis-free patients, adjusting for age and sex for each cancer type (**Supplementary Figure 18, Supplementary Table 11**). Interestingly, the prognostic direction of kataegis varied significantly across different tumour types. Patients with kataegic events showed significantly better outcomes to those without in both CLL ($p = 9.2 \times 10^{-3}$, HR = 0.35,

Figure 5E) and GBM ($p = 6.5 \times 10^{-2}$, HR = 0.34). By contrast, kataegic events were associated with significantly poorer prognosis in both adenocarcinomas of the pancreas ($p = 2.4 \times 10^{-2}$, HR = 1.94, **Figure 5F**) and prostate ($p = 3.3 \times 10^{-3}$, HR = 20.63). Thus kataegis not only appears to be associated with specific driver architectures of individual tumour types, but these manifest in a diverse and complex tumour-specific effect on the clinical landscape of outcome and treatment response.

Different Mutational Processes Share Common and Distinct Drivers

Throughout our analyses of different types of mutations, *TP53* was recurrently associated with elevated density of almost all mutational features considered. This is consistent with its central role in cancer biology, and a long literature associating *TP53* with DNA damage. These results led us to consider whether there were molecular determinants or correlates of other mutational processes. For each variant type, associations with 27 mutational signatures derived from trinucleotide events were evaluated. SNVs, SCNAs and kataegis events were assessed in a gene-wise fashion (n = 1,722, 19,364 and 27 respectively; **Supplementary Table 9**). Two signatures (8 and R1) showed no statistical associations with any event tested (FDR < 0.01), while 16 signatures were associated with at least a single gene in at least 2 mutational features (SNV mutational density, SCNA/SV metrics or kataegis); the remaining 9 signatures demonstrated significant associations with only a single mutation metric. In particular, signature 5 was negatively associated with *CSMD1* when considering point mutations, kataegis events and CN loss (**Supplementary Figure 19**); in fact, signature 5 showed only negative associations with events. A small set of genes were associated with multiple mutational signatures: signatures 1, 5, and 38 were negatively associated with SNVs and kataegic events in *CNTN5*, *CNTNAP2* and *LRRC4C* (genes involved with neuron development) while signatures 7a and 28 were positively associated with these same events.

Next, we identified common patterns across the overall mutational profile of a tumour (**Figure 6A**). Broadly, the mutational density of one aspect of a tumour was well-correlated with almost every other (**Figure 6B**): tumours that show elevated amounts of one type of mutation tend to show elevated rates of every other type. The one exception was average SCNA length, where tumours with longer SCNAs tend to have fewer other types of mutations. This is surprising, and may to some extent reflect reduced accuracy of mutation detection in tetraploid tumours, as noted in other PCAWG studies[26]. To determine whether or not each mutational process was driven by similar mechanisms, we compared all pairs of genes found to be significantly associated with any type of mutational density (**Supplementary Figure 20A**). Minimal overlap was detected across models for different variant types (SNVs/Mbp, SCNA, SV, and hypermutation/kataegis metrics), with increasing overlap among metrics of a similar variant type (within metrics of hypermutation/kataegis).

We then considered the top 10 genes for each mutation density metric (based on FDR-adjusted p-value and magnitude of coefficient) and limited overlap analyses to models based on large scale events (gene-wise ternary SCNA; SNVs/Mbp or kataegis metrics; **Supplementary Figure 20B**) or functional point mutations (metrics of SCNA, SV burden and kataegis; **Figure 6C**). As expected, *TP53* was amongst the most recurrent associated gene in both lists, showing a positive association of both CN loss with SNVs/Mbp and kataegis and point mutations

associated with many SCNA, SV and kataegis metrics. Point mutations in *BRCA2* were negatively associated with metrics of SCNA density, but positively associated with the number of kataegis events while CN loss of the same gene was positively associated with SNVs/Mbp. *KDR*, *POLE*, *BCOR* and *DDX3X* however were exclusively significant when evaluating functional point mutations (**Figure 6C**). Interestingly, a clear difference between the two mutation types (SCNAs and SNVs) emerged when examining clonality estimates - SCNAs arose predominantly in tumour subclones (with a few exceptions of amplifications arising in the clonal tumour; **Supplementary Figure 20B**), while SNVs were more likely to occur early in tumour development (**Figure 6C**). Furthermore, these genes were assessed for overlap with predictors of specific trinucleotide mutation signatures. In addition to being positively associated with many mutation density metrics, similar associations were detected between CN loss of *TP53* with mutation signature 4 and 7c with point mutations similarly associated with signature 7c but are negatively associated with signature 5. These results outline both the uniqueness of *TP53* as the only gene associated with almost all types of DNA damage, and the complex, tumour-type specific landscape of mutational drivers revealed by pan-cancer analysis.

Discussion

One of the most striking results of cancer sequencing studies has been the establishment of dramatic inter-patient variability in the number and nature of somatic SNVs[17]. Some of this variability has been attributed to differences in “mutational signatures”, which reflect the fidelity of a cell’s DNA repair processes and the features of the specific mutagenic insults to which a tumour has been exposed. Here, we focused on identifying the genomic changes associated with these trends in SNV mutation density, and more broadly with multiple classes of mutation density. In particular, we develop new approaches for identifying the associations between mutation-density and candidate driver events, and for detecting kataegic events from whole-genome sequencing data.

These analyses have confirmed a number of previously observed trends, while providing insight into previously unknown oncogenic mechanisms. As expected, tumour types primarily driven by environmental factors, including skin and lung cancers, demonstrated high rates of single-stranded break events while those previously identified as C-class, including ovarian, uterine and breast cancers, demonstrated increased genomic instability, with above average rates of SCNAs and SVs. We quantify how multiple well-characterized driver genes, like *TP53*, *KRAS*, *BRCA2*, *CTNNB1*, *PIK3CA* and *ARID1A* as associated with specific features of the mutational landscape of individual tumours. Further, we confirm the associations of kataegic events with translocations and other complex structural variants.

However these general observations obscure the remarkable divergence of different tumour types. *TP53* is truly an outlier gene, being not only associated with multiple mutational features, but also doing so in almost every tumour type. By contrast, many other driver events show Janus-like character. Kataegic events can be associated with either good (e.g. GBM) or poor outcome (e.g. prostate cancer); *KRAS* point mutations can be associated with decreased translocation rates (e.g. ccRCC) or increased rates (many other tumour types). These divergences highlight the critical importance of appropriate statistical modeling in pan-cancer studies. And, the given the distinctive landscapes and candidate drivers in each tumour-type, these data highlight the ongoing need for large, clinically-homogeneous cohorts with deep WGS to improving our understanding of the mutational hallmarks of individual tumours.

Methods

Data assembly and formatting

Clinical annotations were downloaded from the PCAWG data portal on 2016-08 (syn7772065), with survival information obtained from the ICGC data portal on 2017-06-14 (release 25). Samples flagged for removal were excluded and a single sample from each multi-sample donor was selected according to PCAWG recommendations, resulting in 2,583 specimens carried forward for downstream analyses. Consensus SNV calls were obtained on 2016-10-12 (syn7357330) while consensus CNA calls and information on sample purity and ploidy were obtained on 2017-01-25, following the latest PCAWG data release on 2017-01-19 (syn8042905). SV calls were obtained from the 2016-11 release (syn7596712). Callable base files were downloaded on 2017-03-20 (syn8492850). Consensus clonality estimates for SNVs were provided in the 2017-03-25 release (syn8532425) while the latest PCAWG ABSOLUTE calls (2016-11-01) were used to annotate clonal and subclonal CNAs (downloaded from Jamboree on 2017-03-29: /pancan/pcawg11/subclonal_architecture/broad/broad_absolute_on_2660_concensus_bp_11_1_2016.tar.gz). Candidate driver genes were identified by PCAWG-2,5,9,14 and obtained on 2017-04-22 (syn9757986). Mutation signatures for each patient were downloaded on 2017-03-20 (syn8366024). The RNA-seq expression data was obtained from the 2016-02-12 release (syn5553991). The expression matrix contained upper quantile normalized expression values for 57,821 genes and 2,011 patients.

Consensus SNV calls were filtered such that only functional variants (those predicted to result in either missense or nonsense mutations by Oncotator, as performed by PCAWG-2,5,9,14) were carried forward. Variants were then collapsed to the gene level (n = 18,571 genes), with mutation status further reduced to either present (1) or absent (0) for each patient. These were then filtered to contain only those genes determined to be among driver gene candidates (n = 152), identified by PCAWG-2,5,9,14. Similarly, consensus SCNA calls were first filtered based on confidence level (classified by a “star” system, where one star represents poor caller concordance and poor confidence and three stars represent a consensus and high confidence) per patient prior to downstream analysis (more details described here: syn8042880). Only SCNA calls with a star level of two or three were kept. The remaining SCNA calls were adjusted for the estimated ploidy of that sample (based on ploidy data from the latest PCAWG release, obtained on 2017-01-25, data release 2017-01-19, syn8042905), in the event that the sample was predicted to have a whole genome duplication (WGD) with a status of “certain”, and rounded to the nearest whole number. These adjusted calls were referred to as ploidy-adjusted copy number changes and were used in subsequent analyses. SCNAs were annotated to genes using the gencode database (v19).

Further, a collapsed SCNA matrix was generated from the consensus SCNA calls obtained from PCAWG (syn8042880) for statistical modelling. Since the boundaries of SCNAs are not restricted to a single gene or locus, copy number calls with identical or nearly-identical status were aggregated into segments in the effort to reduce the burden of multiple testing during model fitting. More specifically, an iterative process was used per chromosome, where a new

segment was defined when all SCNA calls within a start and end chromosomal coordinate had at least 99% concordance in terms of equivalent SCNA status (i.e. mostly gains, mostly losses or mostly neutral). This 99% concordance threshold in SCNA status also had to be consistently met across all samples. The new segment was then annotated with the SCNA status that the majority of the copy number calls within that segment had carried. The process starts at the beginning of a chromosome and the segment is expanded until the 99% concordance threshold is no longer met. This process is then repeated using the next chromosomal coordinate as a segment start point, and subsequently repeated until the end of the chromosome is reached. This aggregation procedure reduced 20,229 gene-level SCNA features to 3,200 SCNA segments.

Clonality timing estimates for SCNAs and SNVs were collapsed to form gene by sample matrices, such that for each sample every gene was classified as either clonal, subclonal or not available (indicating either that no variant was present or that timing could not be estimated). A consensus classification was generated across patients (pan-cancer or for the top-powered tumour types) for each gene and event type (CN gain/loss and SNV) using the proportion of patients with a variant that was deemed subclonal in origin (where 0 indicates a clonal event in 100% of patients and 1 a subclonal event; **Supplementary Table 3**).

Sample summary, along with mutation metrics used, are available in **Supplementary Table 1**. All visualizations were made using the BPG package (v5.6.19) for R, with lattice (v0.20-34) and latticeExtra (v0.6-28) packages.

Data Processing

Associating overall single-stranded break events with variant status

Statistical modeling across the cohort

Uncondensed SNV matrices were first loaded into the R statistical environment (v3.3.1) and overall SNV density was determined for each sample as the total number of single nucleotide variants (SNVs) per callable megabase, ranging from less than 1 to greater than 850 SNVs/Mbp per sample. As this distribution is highly skewed, we defined mutation density for downstream analyses as $\log_{10}$ SNVs/Mbp (determined using the total number of bases with a minimum coverage of 14 or 8 reads in the tumour and normal BAMs respectively; **Supplementary Figure 2A-B**). A total of 2,450 patients had the available callable base data to calculate this metric. The distribution of patient sex and age were also evaluated (**Supplementary Figure 2C-D**).

In order to associate specific genomic events with SNVs/Mbp, ternary SCNA status (n = 20,229 genes; n = 1,778/2,450 patients) was used. Genes were first assessed for recurrence to remove those with a SCNA present in less than 1% of the cohort (19,361 genes passed this threshold). For each gene, a mixed effects linear model was applied using the lme4 (v1.1-12) and lmerTest (v2.0-33) packages for R, to explain mutation density across all samples using SCNA status (VS; gain/loss/neutral), sex and age, with tumour type included as the random effect variable (**Equation 1**). SCNA status, sex and tumour type were treated as factors and age as a continuous variable. False-discovery rate (FDR) adjustment was applied to correct for multiple

testing (**Supplementary Table 3**). This analysis was repeated using collapsed SCNA segments with binary SCNA status (described above; n = 3,200 segments) as further validation (**Supplementary Table 4**).

(**Eq 1**) $MD_t = VS_{SCNA,t} + age + sex + (1|\, tumour\ type)$

Pathway Analysis

For each model term, query lists containing genes significantly associated with SNVs/Mbp (FDR < 0.01) were generated and separate pathway analyses was performed using gProfileR[46] (v0.6.1) with default parameters, however with FDR correction for multiple testing and using only the gene ontology (GO) database. Significantly enriched pathways were identified as those with FDR < 0.01 and ordered according to precision (the proportion of term genes present in the query list; **Supplementary Table 5**).

Chromosome enrichment

Chromosome enrichment of genes statistically associated with SNVs/Mbp was assessed using genes classified as either protein_coding or processed_transcript in the gencode (v19) database. For each chromosome, a hypergeometric test was used to assess the overlap of significantly associated genes (FDR < 0.01) and all genes present on that chromosome, from a total pool of all genes (**Supplementary Table 6**).

Statistical modeling per tumour type

To increase the power of our analyses, models were run as above on each of the 5 most powered tumour types (those with ≥ 100 samples after filtering). For each tumour type (t), a linear model was applied using variant status (as above), patient age and sex (where applicable; **Equations 2 and 3**) to identify tumour type specific associations with mutation density. FDR adjustment was applied for each tumour type independently. Analyses were run using ternary SCNA status (**Supplementary Table 8**) as the predictor variable.

(**Eq 2**) $MD_t = VS_{SCNA,t} + age + sex$

(**Eq 3**) $MD_t = VS_{SCNA,t} + age$

Associating double-stranded break events with mutation metrics

Statistical modeling

Analysis was conducted in the R statistical environment (v3.3.1) for explanation of double-stranded break event metrics (i.e., total SCNA count, PGA, total SV count; full list available in **Supplementary Table 1**). A total of 2,410 samples had the necessary data types for this analysis. Here, driver genes (n = 152) with a functional SNV (described above) in at least 1% of patients were evaluated for associations with these metrics using **Eq 1** described above, however with binary, gene-wise SNV status to replace SCNA status. For each metric, p-values were again adjusted for multiple testing using FDR (**Supplementary Table 3**). Models were

again repeated for individual tumour types, as described above (**Eq 2 and 3; Supplementary Table 8**).

SeqKat: a tool for assessing hypermutation and kataegis

Overview

Genome instability is one of the hallmarks of Cancer[47]. A relatively new measure of genome instability is kataegis, which is a pattern of localized substitution hypermutations. Kataegis was first identified in breast cancer[29] where clusters of C>T and/or C>G mutations were observed in TpCpN trinucleotides on the same strand. Despite the frequent number of studies that have examined kataegis in cancer, its causes and association with other genomic features, there is currently no publicly available bioinformatic tool that can detect and visualize kataegis events per patient using a probabilistic approach. Since kataegis is observed at different rates in different cancer types, a tool that can dynamically optimize kataegis detection per cancer type and assess significance of kataegis events would be of a great use to the scientific community. Here we present **SeqKat**, a novel tool that automates the detection and visualisation of kataegic regions (**Supplementary Figure 12**).

The input to SeqKat is a standard VCF file containing a list of recurrent somatic single nucleotide variants per patient. SeqKat uses a sliding window (of fixed width) approach to test deviation of observed SNV trinucleotide content and inter-mutational distance from expected by chance alone. Additionally, an exact binomial test is performed to test that the proportion of each of the 32 tri-nucleotides within each window is higher than expected. The resulting p-values are then adjusted for multiple hypothesis testing using FDR. Hypermutation and kataegic scores are calculated for each window as follows:

(**Eq 4**) $$hypermutation\ score = -log_{10}(binomial\ p_{adj}) * \frac{N\ observed\ mutations}{N\ expected\ mutations}$$

(**Eq 5**) $$kataegis\ score = hypermutation\ score * \frac{N\ TCX\ bases}{N\ expected\ TCX\ bases}$$

Finally, any statistically significant windows, within an optimized maximum inter-mutation distance, are then combined to obtain regions of hypermutation. Output consists of a text file indicating all potential hypermutated and kataegic regions, their genomic position, and their corresponding hypermutation and kataegic scores.

Visualization

The rainfall plot is a comprehensive way of visualizing hypermutation and kataegis events that incorporates both inter-mutational distance and genomic position for each mutation. SeqKat can automatically generate these plots both at the whole genome level and for individual chromosomes (such as only those with a statistically significant event). Hypermutation clusters can be easily recognized using such plots and can be further classified as kataegic events by showing the specific base change composition of the cluster (**Supplementary Figure 18**).

Parameter Optimization

Mutation data for 149 samples across 7 different cancer types (ALL, breast cancer, CLL, liver cancer, lung adenocarcinoma, B-cell lymphoma and pancreatic cancer) was downloaded from Alexandrov's "Signature of mutational processes in human cancer" paper (ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/somatic_mutation_data/)[12]. Data was available as tab delimited files grouped by cancer type. Files were downloaded and converted to BED format per patient. To validate SeqKat, we performed cross validation using these data and tuned the tool's parameters to maximize prediction performance. Parameters tuned include: 1) Hypermutation score cutoff, used to classify each sliding window as significant 2) Maximum inter-mutation distance cutoff, used to classify significant windows as separate hypermutated events and 3) Cutoff for minimum number of SNVs within a single window for it to be classified as hypermutated/kataegic. A 5-fold cross validation was performed and various parameter combinations were run. The combination that maximized the F score across cancer types was selected and used to set defaults (**Supplementary Figure 12D**).

Application

SeqKat was applied using the optimized parameters on a cohort of 251 primary whole genome pancreatic cancer samples that are part of the International Cancer Genome Consortium (ICGC). At least one kataegic event was detected in 80% of the cohort. Kataegic samples had an average of four events per sample. Clinical information such as overall survival status (OS), time to OS, grade, age, and sex were obtained for 238 samples. To further investigate the consequence of kataegis on patient overall survival, a Cox regression was fit and overall survival was compared between kataegic and kataegis-free patients adjusting for age and sex. Kataegic patients have significantly poorer prognosis compared to non-kataegis patients (**Supplementary Figure 13**).

Download

SeqKat (v0.0.4) is an R pacakge that is currently available in CRAN and can be downloaded from the following link: (https://cran.r-project.org/web/packages/SeqKat/index.html).

Assessing Prognostic Role of Kataegis

Clinical information such as overall survival status (OS), time to OS, grade, age, and sex were obtained for 1,704 PCAWG samples. To further investigate the consequence of kataegis on patient overall survival, a Cox regression was fit and overall survival was compared between kataegic and kataegis-free patients adjusting for age and sex. The analysis was conducted on cancer types that had at least 25 patients with survival and kataegis information available. The test p-values along with the hazard ratios are reported for each cancer type in the Kaplan Meier plots (**Supplementary Figure 18, Supplementary Table 11**).

Associating hypermutation and kataegis with mutation metrics

Statistical modeling

SeqKat (v0.0.4) was used to identify hypermutation events, classified as either APOBEC-mediated kataegis or not, using the PCAWG consensus SNV calls. SeqKat was run using the

default, globally optimized parameters (hypermutation score cutoff = 5, maximum inter-mutation distance cutoff = 3.2 and minimum SNV count cutoff = 4) and scores were generated as described above. APOBEC-mediated kataegis events were identified as those having a kataegic score > 0. This identified between 1 and 19,951 kataegic events per sample (mean = 30.9, median = 1). Where multiple events were called per sample, the event with the highest kataegic score was used to represent the kataegic status of that sample for downstream analyses.

To assess chromosomal enrichment, the genome was split into 3,113 1Mbp bins. The expected kataegic rate was calculated by dividing the number of kataegic bins over the total bins in the genome (2,791/3,113). For each chromosome, the fraction of kataegic bins was calculated and a binomial test was used to test the deviation of observed chromosomal kataegic rate from the expected rate.

For each hypermutation metric (**Supplementary Table 1**), a mixed effects model was applied as above (**Eq 1**) using either SCNA status (VS; gain/loss/neutral), **Supplementary Table 3** [ternary], **Supplementary Table 4** [collapsed]) or driver genes containing functionally relevant SNVs in at least 1% of the dataset (148 genes in 2,563 patients, **Supplementary Table 3**). Finally, analyses were repeated for each powered tumour type independently (**Eq 2 and 3** above**, Supplementary Table 8**).

Associating specific events with trinucleotide mutation signatures

Analysis was conducted in the R statistical environment (v3.4.0). For each patient, the number of mutations contributing to each trinucleotide profile were obtained and converted to proportions to standardize across patients. Each mutation signature was then modeled as described above (**Eq 1**) using a linear mixed effects model, with the presence of SNVs or kataegis events in each gene, or SCNA status (gain/loss/neutral), as the independent variable, and controlling for patient sex, age and tumour type. A total of 27 mutation signatures were present to any degree in at least 1% of the cohort. For the independent variables, 1,722, 19,364 and 27 genes had a recurrent event (>1% of samples) related to SNVs, SCNAs or kataegis respectively. Models were run separately for each data type with FDR adjustment applied to correct for multiple testing (**Supplementary Table 9**).

Integration across mutation metrics

Mutation enrichment of 1 Mbp bins along the genome was assessed using the rank product of a subset of individual mutation metrics (**Supplementary Figure 20A**). The large collection of mutation metrics used were compared using pairwise Spearman's correlations across all available patients (the number of patients differed for each comparison as not all metrics were available for all patients; **Supplementary Figure 20B**). Furthermore, the top associated SNV or SCNA containing genes from each analysis were compared. Gene-wise functional SNVs were used for associations with DSBs and kataegis metrics, while SCNAs were used to find associations with SSBs and kataegis metrics. For each metric, the top 10 associated genes were selected based on FDR-value and magnitude of the coefficient and compared across the different metrics (**Figure 6**).

Acknowledgment

The authors thank all members of the Boutros lab for technical support and insight commentary. This study was conducted with the support of the Ontario Institute for Cancer Research to PCB through funding provided by the Government of Ontario. Dr. Boutros was supported by a Terry Fox Research Institute New Investigator Award, a CIHR New Investigator Award, by the Canadian Institutes of Health Research grant # SVB-145586, and by Prostate Cancer Canada proudly funded by the Movember Foundation - Grant #RS2014-01.

References

1. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
2. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-357 (2015).
3. Liu, B. et al. Spatio-Temporal Genomic Heterogeneity, Phylogeny, and Metastatic Evolution in Salivary Adenoid Cystic Carcinoma. *J Natl Cancer Inst* **109**(2017).
4. Dong, L.Q. et al. Spatial and temporal clonal evolution of intrahepatic cholangiocarcinoma. *J Hepatol* (2018).
5. Drier, Y. et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res* **23**, 228-35 (2013).
6. Lawrence, M.S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
7. Roberts, S.A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**, 970-6 (2013).
8. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-50 (2014).
9. Haradhvala, N.J. et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538-49 (2016).
10. Behjati, S. et al. Mutational signatures of ionizing radiation in second malignancies. *Nat Commun* **7**, 12605 (2016).
11. Fraser, M. et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359-364 (2017).
12. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
13. Hoadley, K.A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944 (2014).
14. Tomasetti, C. et al. Role of stem-cell divisions in cancer risk. *Nature* **548**, E13-E14 (2017).
15. Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78-81 (2015).
16. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330-1334 (2017).
17. Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O. & Stein, L.D. Pan-cancer analysis of whole genomes. *bioRxiv* (2017).
18. Kozarewa, I. et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**, 291-5 (2009).
19. Aird, D. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**, R18 (2011).
20. Greenman, C. et al. Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-8 (2007).
21. Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* **150**, 251-63 (2012).
22. Weinstein, J.N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-20 (2013).
23. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202-9 (2014).
24. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892 (2012).

25. Ciriello, G. et al. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-33 (2013).
26. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *bioRxiv* (2017).
27. Lalonde, E. et al. Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *Lancet Oncol* **15**, 1521-32 (2014).
28. Rheinbay, E. et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. *bioRxiv* (2017).
29. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
30. Lada, A.G. et al. AID/APOBEC cytosine deaminase induces genome-wide kataegis. *Biol Direct* **7**, 47; discussion 47 (2012).
31. Muino, J.M., Kuruoglu, E.E. & Arndt, P.F. Evidence of a cancer type-specific distribution for consecutive somatic mutation distances. *Comput Biol Chem* **53 Pt A**, 79-83 (2014).
32. Chen, X. et al. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep* **7**, 104-12 (2014).
33. Bolli, N. et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**, 2997 (2014).
34. Hoogstraat, M. et al. Genomic and transcriptomic plasticity in treatment-naive ovarian cancer. *Genome Res* **24**, 200-11 (2014).
35. Davis, C.F. et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319-330 (2014).
36. Khodabakhshi, A.H. et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3**, 1308-19 (2012).
37. Keim, C., Kazadi, D., Rothschild, G. & Basu, U. Regulation of AID, the B-cell genome mutator. *Genes Dev* **27**, 1-17 (2013).
38. Mechtcheriakova, D., Svoboda, M., Meshcheryakova, A. & Jensen-Jarolim, E. Activation-induced cytidine deaminase (AID) linking immunity, chronic inflammation, and cancer. *Cancer Immunol Immunother* **61**, 1591-8 (2012).
39. Nambiar, M. & Raghavan, S.C. How does DNA break during chromosomal translocations? *Nucleic Acids Res* **39**, 5813-25 (2011).
40. Dorsett, Y. et al. A role for AID in chromosome translocations between c-myc and the IgH variable region. *J Exp Med* **204**, 2225-32 (2007).
41. Knezevich, S. et al. Concurrent translocation of BCL2 and MYC with a single immunoglobulin locus in high-grade B-cell lymphomas. *Leukemia* **19**, 659-63 (2005).
42. Dyer, M.J. et al. Concurrent activation of MYC and BCL2 in B cell non-Hodgkin lymphoma cell lines by translocation of both oncogenes to the same immunoglobulin heavy chain locus. *Leukemia* **10**, 1198-208 (1996).
43. Bertrand, P. et al. Mapping of MYC breakpoints in 8q24 rearrangements involving non-immunoglobulin partners in B-cell lymphomas. *Leukemia* **21**, 515-23 (2007).
44. Gonda, H. et al. The balance between Pax5 and Id2 activities is the key to AID gene expression. *J Exp Med* **198**, 1427-37 (2003).
45. D'Antonio, M., Tamayo, P., Mesirov, J.P. & Frazer, K.A. Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Higher HER2 Levels. *Cell Rep* **16**, 672-83 (2016).
46. Reimand, J. et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res* **44**, W83-9 (2016).
47. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-74 (2011).

Figure Legends

Figure 1: Specific SCNAs are associated with an increased SNV density. A) Total SNV count across all available patients within each 1Mbp bin along the genome; red points indicate bins containing known driver genes. B) Mutation rate (SNVs/Mbp), stratified by tumour type; red bars indicate median SNVs/Mbp for each type. Volcano plots showing the coefficient and FDR-adjusted p-value for all C) CN gains or losses in the pan-cancer or individual tumour type models. D) Tumours with CN-amplification of *BCKDH8* show elevated SNV mutation rates in a pan-cancer multivariate analysis (left boxplot) and in multiple individual tumour types (violin plots stratified by CN type (deletion = -1, neutral = 0, amplification = 1); coloured violins reflect statistical mixed effects model, FDR < 0.1). The tumour types selected for subgroup analysis are those with the largest sample number (CNS-Medullo, RCC, HCC, pancreatic adenocarcinomas and prostate adenocarcinomas), in boxplots as indicated by the covariates with colours corresponding to part B. E) *CKMT2* loss is associated with elevated mutation density. Figure structure is similar to 1D).

Figure 2: Distinct SNVs are associated with a number of SCNA metrics. A) Most strikingly across the 19 DSB metrics, the ratio of percent genome altered (PGA) by gains *vs*. losses differed by several folds across tumour types. Here, PGA gain-to-loss ratio (autosomes only) is stratified across tumours and ordered along the x-axis by the median decreasing ratio (median ratio for kidney chRCC = 0). Several metrics had associations with specific point mutations that were statistically significant following adjustments for multiple testing, including B) PGA gain-to-loss ratio (without sex chromosomes), C) total SCNA count and D) total PGA.

Figure 3: Distinct SNVs are associated with a number of SV metrics. A) The total number of translocations can differ by 2-3 folds across tumour. B) To identify potential patterns in overall SV trend, SVs were first binned and subsequently filtered based on recurrence, this unveiling some interesting patterns in translocations for chromosomes 1, 12, 14 and 18. C) Linear mixed effects modelling identified 9 genes that were statistically significantly associated with translocation count following adjustment for multiple testing.

Figure 4: Summary of kataegic events. A) Frequency of kataegis events or B) maximum SeqKat score per patient, stratified by tumour type. C) Rainfall plot depicting kataegis event on chromosome 6 for a single BRCA (UK) patient. D) Circos plot displaying the chromosomal distribution of kataegic events in the pan-cancer cohort. E) Top frequent kataegic genes.

Figure 5: Specific SNVs and SCNAs are associated with an increased in kataegic rate. A) RNA abundance levels of *MYC* were significantly higher in patients demonstrating kataegis within this gene than those without. B) Circos plot displaying the chromosomal distribution of translocations in the lymphoma samples alone. Volcano plots demonstrating the results of the mixed effects models for C) SNVs in known driver genes or D) CN losses to predict the presence of kataegis; a single gene (*TP53*) was significantly associated (FDR < 0.01) with the presence of at least one kataegic event while multiple genes with CN deletions were significantly associated (FDR < 0.01) with the presence of at least one kataegic event. Kaplan-

Meier plots demonstrating the overall survival differences between kataegic and kataegis-free patients in E) Lymph-CLL and F) Panc-AdenoCA, stratified by binary kataegis status.

Figure 6: Top associated variants across the cohort. A) For each mutation metric, the percent of total genes tested that were found to be statistically significantly associated with that metric alone (diagonal) or overlapping with other metrics (45% of genes tested had a CN gain associated with SNV/Mbp; alternatively, 6% of genes tested were also significant by CN loss with this metric). Top 10 recurrent genes containing B) SCNAs or C) SNVs across the dataset; (left panels) dot colour represents direction of association, while background shading indicates FDR-adjusted p-value; (middle panels) clonal status of each SCNA (amplification or deletion) or SNV is (where 0 indicates all samples had a clonal mutation and 1 indicates all samples had a subclonal mutation); (right panels) associations with trinucleotide mutation signatures; those signatures with a significant association (FDR < 0.01) are shown, where green indicates a positive association and purple a negative association; dots indicate type of SCNA with a significant association.

Supplementary Figure Legends

Supplementary Figure 1: Experimental Design.

Supplementary Figure 2: Summary of single-stranded break variants. The distribution of SNVs/Mbp before A) and after B) transformation. The distribution of SNVs/Mbp according to C) patient sex and D) patient age, coloured to distinguish different tumour types.

Supplementary Figure 3: SCNAs in known driver genes associated with SNVs/Mbp. Model results for 46 common driver genes are shown within the pan-cancer cohort (left two columns) or within a tumour-type specific manner. Genes were selected if they contained CN losses (top 14 genes) or CN gains (lower 32 genes) found to be significantly associated with SNVs/Mbp (FDR < 0.001) in the pan-cancer analysis. Dot size represents SNVs/Mbp, while background shading indicates FDR-adjusted p-value. Covariates along the top indicate data used (pan-cancer or individual tumour type) or CN direction.

Supplementary Figure 4: The presence of SCNAs, regardless of direction, are associated with SNV density. Volcano plots showing the coefficient and FDR-adjusted p-value for all 3,200 SCNA segments in the A) pan-cancer analysis and B) within the 5 most well powered tumour types.

Supplementary Figure 5: Several tumours present with notable differences in PGA gain:loss ratio. The SCNA profiles of five tumour types are shown; CNS-PiloAstro consistently has high PGA gain:loss while Prost-AdenoCA generally has low PGA gain:loss ratio. Within each tumour type, samples are ordered by the PGA gain:loss ratio, calculated using autosomes only, as depicted in top plots. Genomic regions are collapsed by gene and organized along the y-axis by their chromosomal locations.

Supplementary Figure 6: Summary of DSB mutation metrics (SCNA). The distribution of 14 SCNA-derived measures of DSB density across tumours. For each metric, tumour types were ranked according to the median. Dot size indicates rank (where the tumour type with the highest median metric is ranked 1). Background shading indicates variance for each metric in $\log_{10}$ space.

Supplementary Figure 7: Associations between variant genes and SCNA metrics. The results of linear mixed effects modelling for assessing gene specific associations with various SCNA-derived metrics of DSB density. Average SCNA length: A) total, B) gains, C) losses or D) gain to loss ratio. E) PGA gain to loss ratio. PGA by F) gains only or G) losses only. H) Estimated tumour ploidy and I) purity. Total SCNA count divided by J) gains only or K) losses only. P-values were adjusted for multiple testing using FDR and statistical significance was defined at FDR < 0.1.

Supplementary Figure 8: *ERBB4* associated with multiple metrics and tumour types. *ERBB4* was significantly associated with 5 individual tumour types for 7 distinct metrics of SCNA density. Dot size indicates model coefficient while background shading indicates FDR-adjusted p-value for each.

Supplementary Figure 9: Summary of DSB mutation metrics (SV). The distribution of 5 SV-derived measures of DSB density across tumours. For each metric, tumour types were ranked according to the median. Dot size indicates rank (where the tumour type with the highest median metric is ranked 1). Background shading indicates variance for each metric in $\log_{10}$ space.

Supplementary Figure 10: Associations between variant genes and SV metrics. Model results for assessing SNV associations with SV-derived metrics of DSB density: A) total SV count, B) total number of inversions and inversions subdivided by C) head to head and D) tail to tail inversions. FDR adjustment of the p-values was applied to correct for multiple testing. Significant hits are labelled where appropriate.

Supplementary Figure 11: Point mutations in *TP53* are associated with TRA burden. The effect and statistical significance of 11 genes were identified to be significantly associated with SV TRA burden at the pan-cancer level are shown for the five highest powered tumour types. Dot sizes represent the gene coefficient from modelling fitting and background shading denotes the FDR-adjusted p-values.

Supplementary Figure 12: SeqKat algorithm development. A) Algorithm workflow highlighting the tuned parameters. B) Overlapping significant windows are stitched together to establish kataegic boundaries. C) Distribution of tumour types used for parameter tuning. D) Performance results following parameter tuning using cross validation. E) Kataegic events can be visualized using rainfall plots; each event has an associated hypermutation and kataegic score.

Supplementary Figure 13: Application of SeqKat in pancreatic cancer. Kaplan-Meier plot demonstrating the overall survival differences between kataegic and kataegis-free patients in a pancreatic patient cohort.

Supplementary Figure 14: Visualization of strong kataegic events through rainfall plots. Examples of rainfall plots for some of the strongest kataegic events in the pan-cancer study: A) whole genome and B) chromosome 8 plots for a single BRCA (USA) tumour and C) whole genome rainfall plot for a single tumour from the BRCA (UK) cohort.

Supplementary Figure 15: mRNA expression profiles of APOBEC family genes. Gene expression across the APOBEC family of proteins using RNA-seq technology stratified by kataegis status; groups were compared using a Student's t-test.

Supplementary Figure 16: Associations between variant genes and kataegis metrics. Mixed effects linear modelling was applied to assessing gene associations with kataegis-based metrics. A) Maximum kataegis score as a continuous variable or B) counts of kataegis events were evaluated using SNV-based features. Similarly, C) presence of kataegis was evaluated using CN gains.

Supplementary Figure 17: *TP53* status is associated with kataegis events. The proportion of patients with and without kataegis events is significantly different between patients with and without SNVs in *TP53* (Proportion Test P-value = $5.35 \times 10^{-5}9$) in a pan-cancer setting.

Supplementary Figure 18: Kataegis is associated with overall survival. Kaplan-Meier plots demonstrating the overall survival differences between kataegic and kataegis-free patients across different tumour types: A) CLL stratified by frequency of kataegis events; B) GBM and C) OV stratified by binary kataegis status; D) pancreatic adenocarcinoma stratified by frequency of kataegis events; prostate adenocarcinoma stratified by E) binary kataegic classification or F) frequency of kataegis events.

Supplementary Figure 19: Events associated with PCAWG mutation signatures. A) Venn diagram demonstrating genes significantly associated with Signature 5 for various genomic events.

Supplementary Figure 20: Summary of mutation metrics. A) Each chromosome was divided into 1 Mbp bins, and the total number of events across patients were calculated for each mutation type, including functional SNV counts, total SCNA counts (divided into gains/losses), total number of transversions, and total number of APOBEC-mediated kataegic events (from bottom to top); counts were then ranked, with the scaled rank product (top) showing mutation enrichment in specific bins. B) Pairwise Spearman's correlation analyses were performed to assess metric similarity. Dot size indicates the strength of the correlation, with colour indicating direction (red for positive correlation, blue for negative). Background shading indicates the p-value of the correlation. For simplicity, p-values are truncated to 10^{-4}.

Table Legends

Supplementary Table 1: Mutation density metrics. All available mutation metrics for each patient.

Supplementary Table 2: Summary of mutation density metrics per tumour type. Median, standard deviation and interquartile range for each metric, stratified by tumour type; values were determined after removing patients with unknown sex/age.

Supplementary Table 3: Gene-wise model results. Results of linear mixed effects models for all SSB (using gene-wise ternary SCNAs), DSB (using functional SNVs in driver genes only) and kataegis (ternary SCNAs and functional SNVs in driver genes) metrics in pan-cancer analyses.

Supplementary Table 4: Collapsed, binary SCNA based model results. Results of linear mixed effects models for binarized SCNA segments predicting SNVs/Mbp.

Supplementary Table 5: Pathway analysis for genes associated with SNVs/Mbp. Genes containing statistically significantly associated SCNAs were utilized for pathway analyses. Statistically significantly enriched pathways are shown.

Supplementary Table 6: Chromosome Enrichment. Genes that were found to be significantly associated with SNVs/Mbp (when considering CN gains/losses) were assessed for chromosome enrichment.

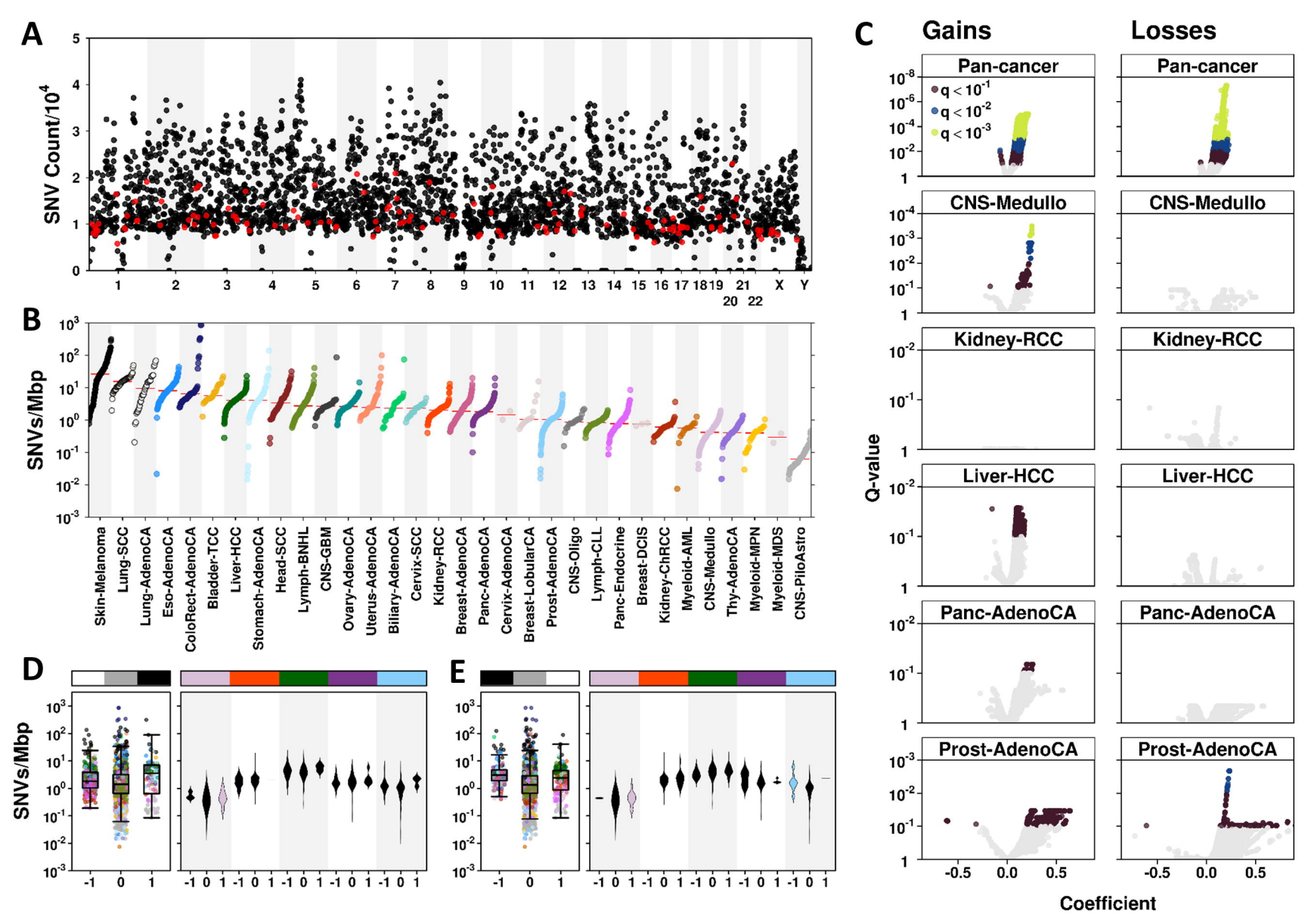
Supplementary Table 7: Consensus of clonal/subclonal classification of gene-wise variants. The proportion of samples demonstrating a subclonal SCNA or SNV was calculated for each gene (proportion ≤ 0.5 = probable clonal variant).

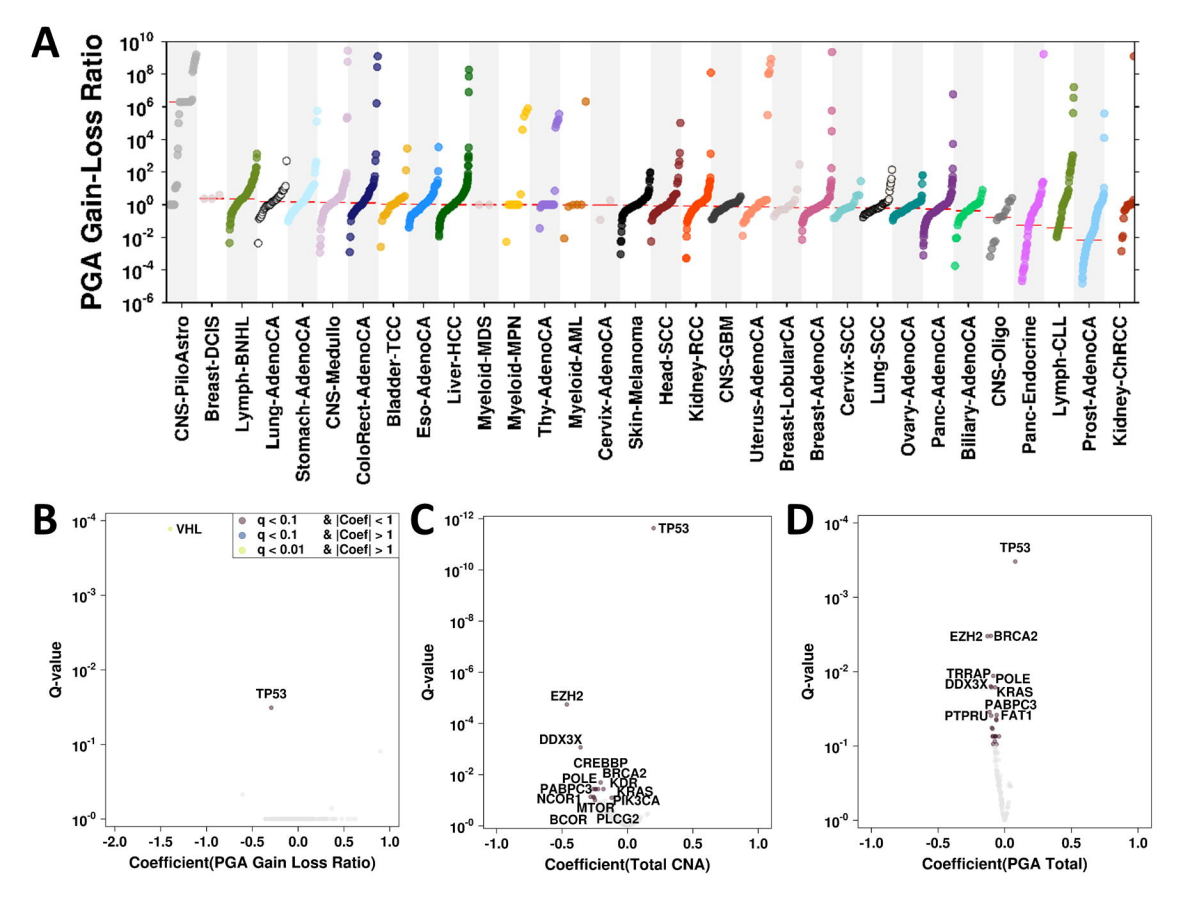
Supplementary Table 8: Model replication in individual tumour types. Results of linear models for each powered (n ≥ 100 patients) tumour type with the required data (medulloblastoma, RCC, HCC, pancreatic and prostate adenocarcinomas) for predicting metrics of SSBs (using gene-wise ternary SCNAs), DSBs (using functional SNVs in driver genes only) or kataegis (ternary SCNAs or functional SNVs in driver genes).

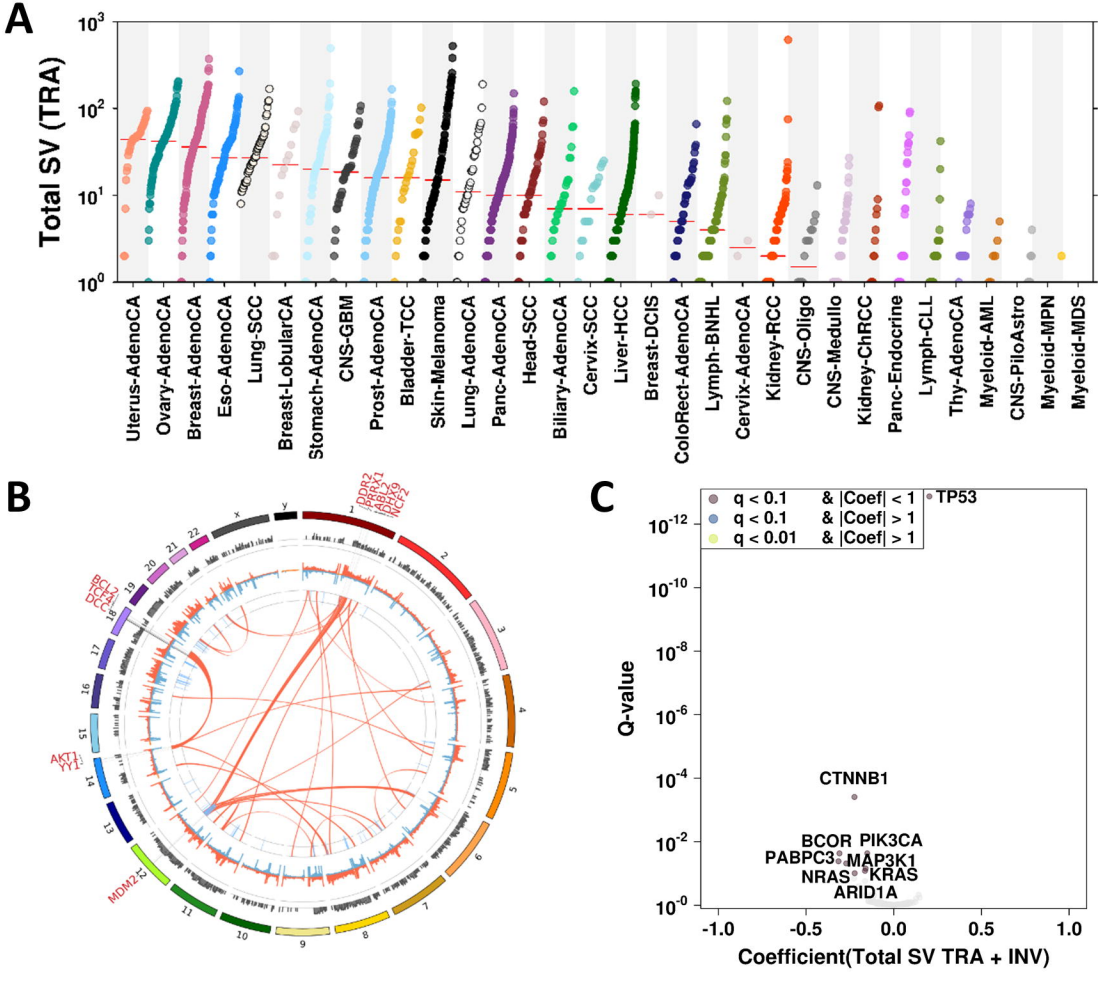
Supplementary Table 9: Mutation signatures are associated with SNVs, SCNAs and kataegis events. Results of linear mixed effects models for each variant type (genes containing functional SNVs or kataegis events as well as collapsed SCNA segments) modeled independently.

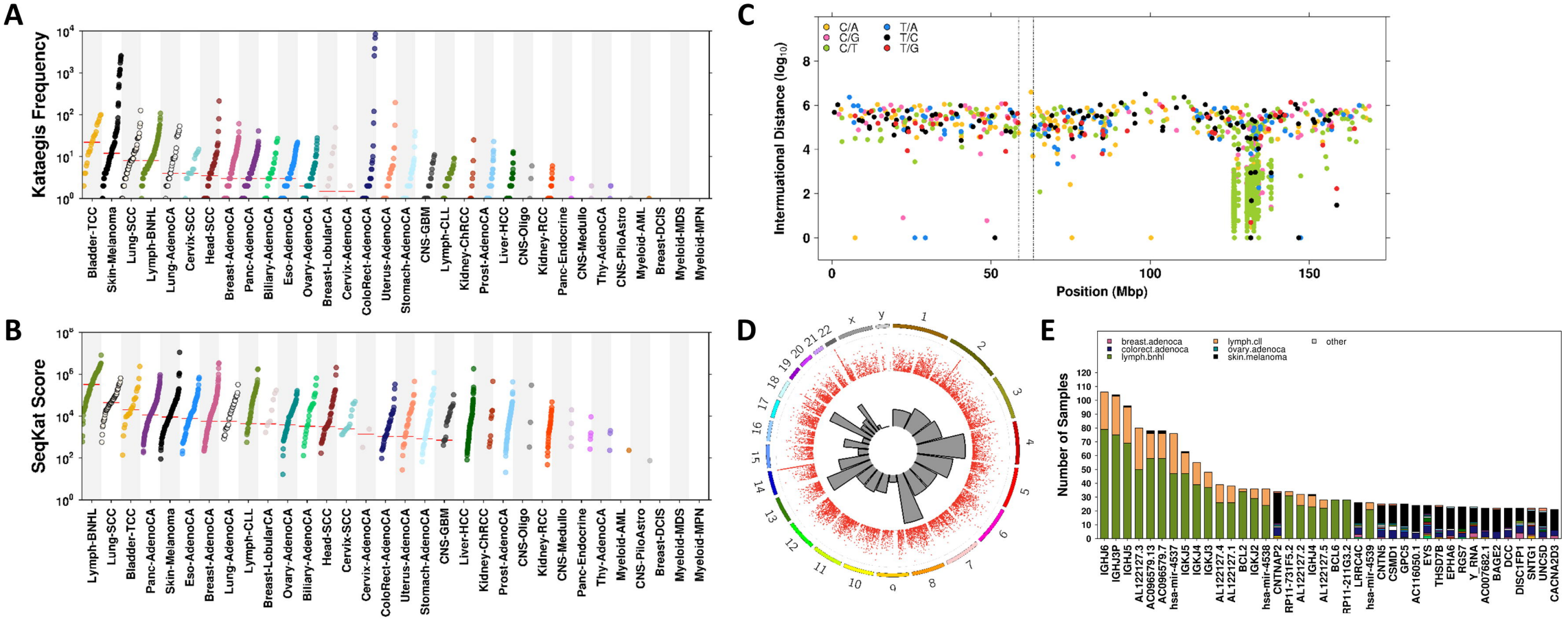
Supplementary Table 10: Kataegic Enriched Genes. Summary of genes that are enriched in kataegic events along with the number of samples and tumour types affected.

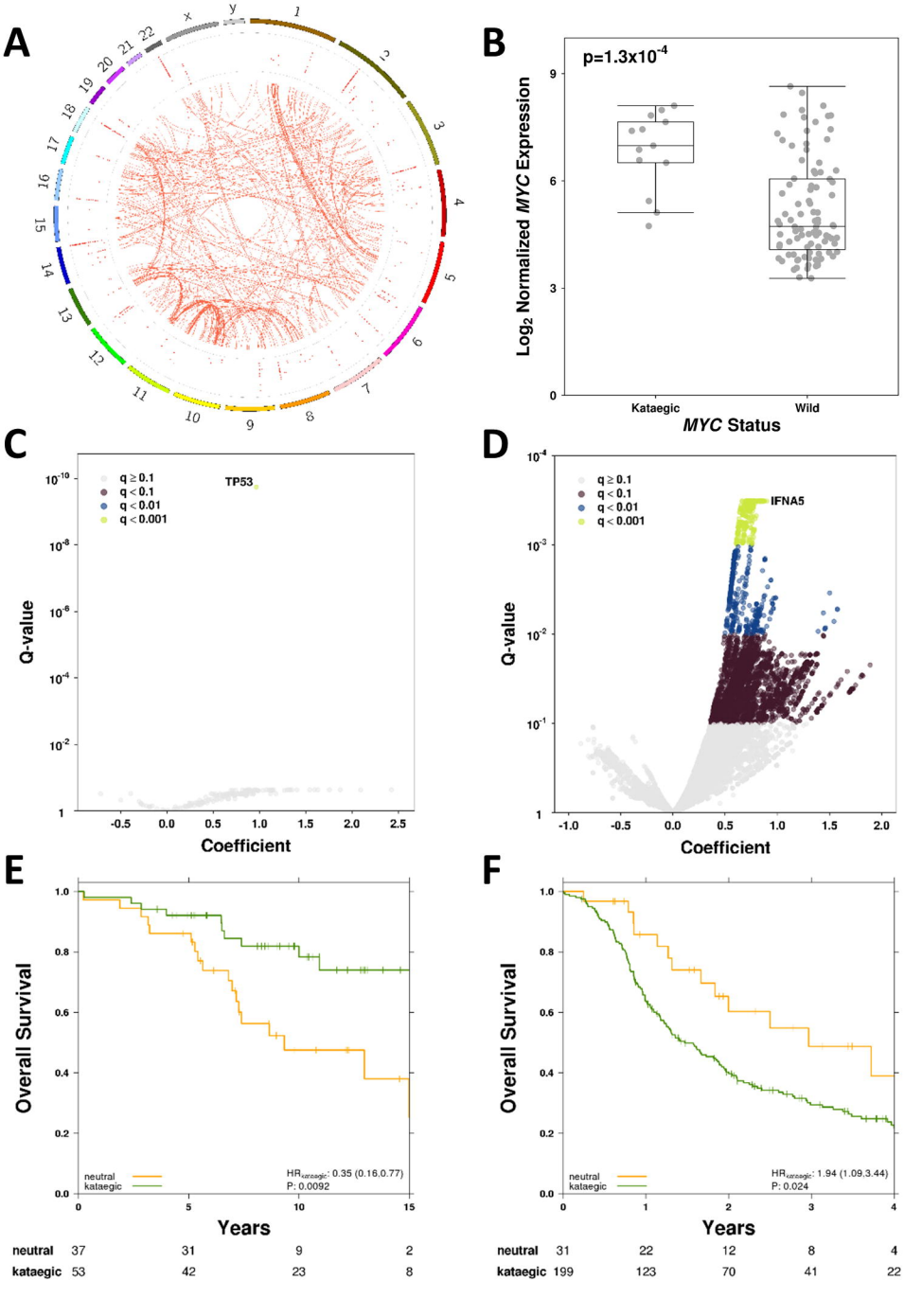
Supplementary Table 11: Kataegis association with overall survival. Summary of Coxph models testing for overall survival differences between kataegic and kataegis-free patients across tumour types with sample size larger than 25 patients.

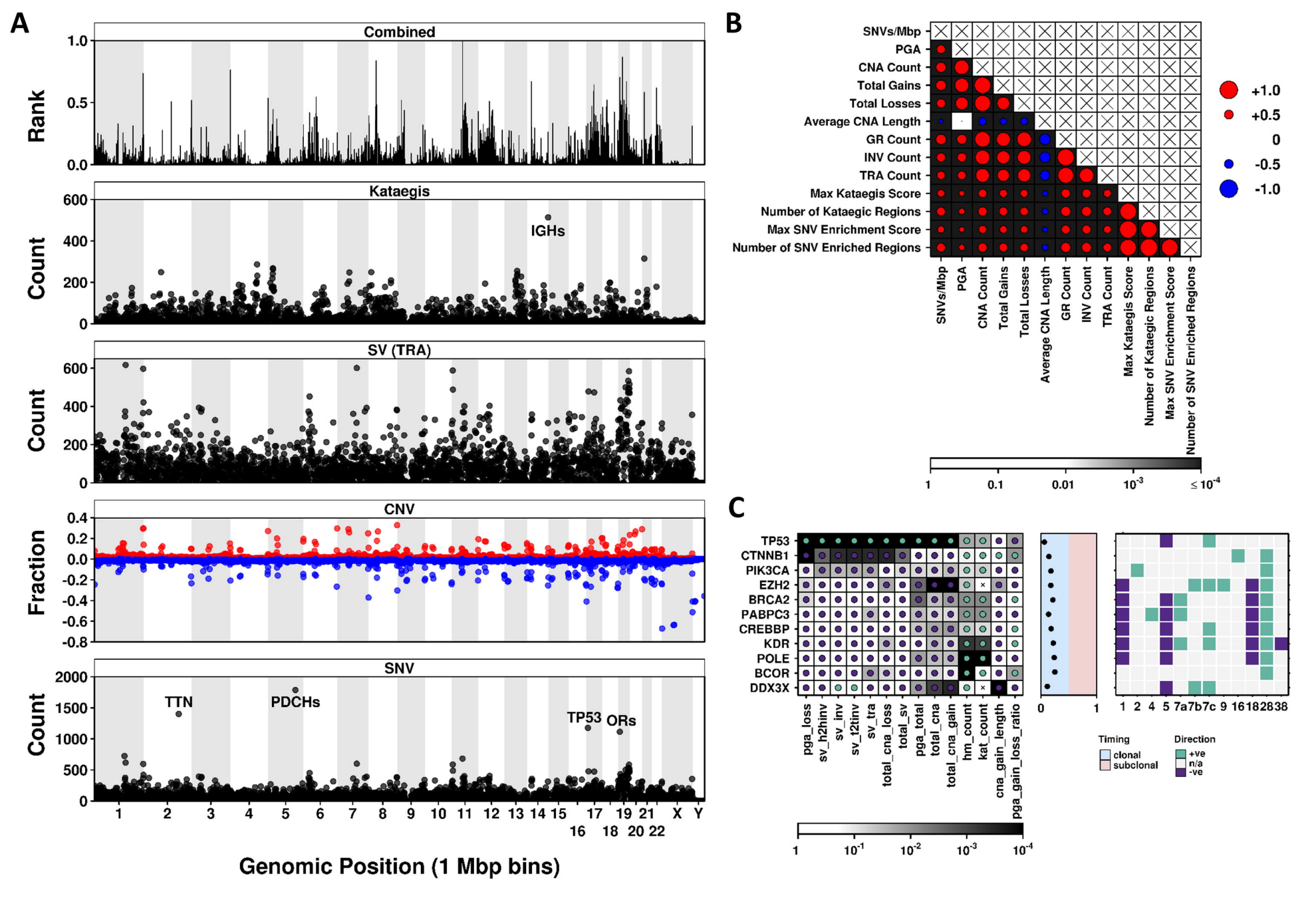

A
SNV Count/10^4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
B
SNVs/Mbp
Skin-Melanoma
Lung-SCC
Lung-AdenoCA
Eso-AdenoCA
ColoRect-AdenoCA
Bladder-TCC
Liver-HCC
Stomach-AdenoCA
Head-SCC
Lymph-BNHL
CNS-GBM
Ovary-AdenoCA
Uterus-AdenoCA
Biliary-AdenoCA
Cervix-SCC
Kidney-RCC
Breast-AdenoCA
Panc-AdenoCA
Cervix-AdenoCA
Breast-LobularCA
Prost-AdenoCA
CNS-Oligo
Lymph-CLL
Panc-Endocrine
Breast-DCIS
Kidney-ChRCC
Myeloid-AML
CNS-Medullo
Thy-AdenoCA
Myeloid-MPN
Myeloid-MDS
CNS-PiloAstro
C
Gains
Losses
Pan-cancer
CNS-Medullo
Kidney-RCC
Liver-HCC
Panc-AdenoCA
Prost-AdenoCA
q < 10^-1
q < 10^-2
q < 10^-3
Q-value
Coefficient
D
E
SNVs/Mbp


A
PGA Gain-Loss Ratio
CNS-PiloAstro
Breast-DCIS
Lymph-BNHL
Lung-AdenoCA
Stomach-AdenoCA
CNS-Medullo
ColoRect-AdenoCA
Bladder-TCC
Eso-AdenoCA
Liver-HCC
Myeloid-MDS
Myeloid-MPN
Thy-AdenoCA
Myeloid-AML
Cervix-AdenoCA
Skin-Melanoma
Head-SCC
Kidney-RCC
CNS-GBM
Uterus-AdenoCA
Breast-LobularCA
Breast-AdenoCA
Cervix-SCC
Lung-SCC
Ovary-AdenoCA
Panc-AdenoCA
Biliary-AdenoCA
CNS-Oligo
Panc-Endocrine
Lymph-CLL
Prost-AdenoCA
Kidney-ChRCC
B
VHL
q < 0.1 & |Coef| < 1
q < 0.1 & |Coef| > 1
q < 0.01 & |Coef| > 1
TP53
Q-value
Coefficient(PGA Gain Loss Ratio)
C
TP53
EZH2
DDX3X
CREBBP
BRCA2
POLE
KDR
PABPC3
KRAS
NCOR1
PIK3CA
MTOR
BCOR
PLCG2
Q-value
Coefficient(Total CNA)
D
TP53
EZH2
BRCA2
TRRAP
DDX3X
POLE
KRAS
PABPC3
PTPRD
FAT1
Q-value
Coefficient(PGA Total)


A
Total SV (TRA)
Uterus-AdenoCA
Ovary-AdenoCA
Breast-AdenoCA
Eso-AdenoCA
Lung-SCC
Breast-LobularCA
Stomach-AdenoCA
CNS-GBM
Prost-AdenoCA
Bladder-TCC
Skin-Melanoma
Lung-AdenoCA
Panc-AdenoCA
Head-SCC
Biliary-AdenoCA
Cervix-SCC
Liver-HCC
Breast-DCIS
ColoRect-AdenoCA
Lymph-BNHL
Cervix-AdenoCA
Kidney-RCC
CNS-Oligo
CNS-Medullo
Kidney-ChRCC
Panc-Endocrine
Lymph-CLL
Thy-AdenoCA
Myeloid-AML
CNS-PiloAstro
Myeloid-MPN
Myeloid-MDS
B
C
q < 0.1 & |Coef| < 1
q < 0.1 & |Coef| > 1
q < 0.01 & |Coef| > 1
TP53
Q-value
CTNNB1
BCOR
PIK3CA
PABPC3
MAP3K1
NRAS
KRAS
ARID1A
Coefficient(Total SV TRA + INV)


A
Kataegis Frequency
Bladder-TCC
Skin-Melanoma
Lung-SCC
Lymph-BNHL
Lung-AdenoCA
Cervix-SCC
Head-SCC
Breast-AdenoCA
Panc-AdenoCA
Biliary-AdenoCA
Eso-AdenoCA
Ovary-AdenoCA
Breast-LobularCA
Cervix-AdenoCA
ColoRect-AdenoCA
Uterus-AdenoCA
Stomach-AdenoCA
CNS-GBM
Lymph-CLL
Kidney-ChRCC
Prost-AdenoCA
Liver-HCC
CNS-Oligo
Kidney-RCC
Panc-Endocrine
CNS-Medullo
Thy-AdenoCA
CNS-PiloAstro
Myeloid-AML
Breast-DCIS
Myeloid-MDS
Myeloid-MPN
B
SeqKat Score
Lymph-BNHL
Lung-SCC
Bladder-TCC
Panc-AdenoCA
Skin-Melanoma
Eso-AdenoCA
Breast-AdenoCA
Lung-AdenoCA
Lymph-CLL
Breast-LobularCA
Ovary-AdenoCA
Biliary-AdenoCA
Head-SCC
Cervix-SCC
Cervix-AdenoCA
ColoRect-AdenoCA
Uterus-AdenoCA
Stomach-AdenoCA
CNS-GBM
Liver-HCC
Kidney-ChRCC
Prost-AdenoCA
CNS-Oligo
Kidney-RCC
CNS-Medullo
Panc-Endocrine
Thy-AdenoCA
Myeloid-AML
CNS-PiloAstro
Breast-DCIS
Myeloid-MDS
Myeloid-MPN
C
C/A
C/G
C/T
T/A
T/C
T/G
Intermuational Distance (log10)
Position (Mbp)
D
E
breast.adenoca
colorect.adenoca
lymph.bnhl
lymph.cll
ovary.adenoca
skin.melanoma
other
Number of Samples
IGHJ6
IGHJ3P
IGHJ5
AL122127.3
AC096579.13
AC096579.7
hsa-mir-4537
IGKJ5
IGKJ4
IGKJ3
AL122127.4
AL122127.1
BCL2
IGKJ2
hsa-mir-4538
CNTNAP2
RP11-731F5.2
AL122127.2
IGHJ4
AL122127.5
BCL6
RP11-211G3.2
LRRC4C
hsa-mir-4539
CNTN5
CSMD1
GPC5
AC116050.1
EYS
THSD7B
EPHA6
RGS7
Y_RNA
AC007682.1
BAGE2
DCC
DISC1FP1
SNTG1
UNC5D
CACNA2D3


A
B
p=1.3x10-4
Log2 Normalized MYC Expression
Kataegic
Wild
MYC Status
C
TP53
q ≥ 0.1
q < 0.1
q < 0.01
q < 0.001
Q-value
Coefficient
D
IFNA5
E
Overall Survival
neutral
kataegic
HRkataegic: 0.35 (0.16.0.77)
P: 0.0092
Years
neutral 37 31 9 2
kataegic 53 42 23 8
F
HRkataegic: 1.94 (1.09.3.44)
P: 0.024
neutral 31 22 12 8 4
kataegic 199 123 70 41 22


A
Combined
Rank
1.0
0.5
0.0
Kataegis
Count
600
400
200
0
IGHs
SV (TRA)
Count
CNV
Fraction
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
SNV
Count
2000
1500
1000
500
TTN
PDCHs
TP53
ORs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Genomic Position (1 Mbp bins)
B
SNVs/Mbp
PGA
CNA Count
Total Gains
Total Losses
Average CNA Length
GR Count
INV Count
TRA Count
Max Kataegis Score
Number of Kataegic Regions
Max SNV Enrichment Score
Number of SNV Enriched Regions
+1.0
+0.5
0
-0.5
-1.0
1
0.1
0.01
10^{-3}
$\leq 10^{-4}$
C
TP53
CTNNB1
PIK3CA
EZH2
BRCA2
PABPC3
CREBBP
KDR
POLE
BCOR
DDX3X
pga_loss
sv_h2hinv
sv_inv
sv_t2tinv
sv_tra
total_cna_loss
total_sv
pga_total
total_cna
total_cna_gain
hm_count
kat_count
cna_gain_length
pga_gain_loss_ratio
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
1
1 2 4 5 7a7b7c 9 16182838
Timing
clonal
subclonal
Direction
+ve
n/a
-ve