Selective and mechanistic pressures shaping cancer aneuploidies

Juliann Shih (jshih@broadinstitute.org)
Broad Institute of MIT and Harvard https://orcid.org/0000-0002-4086-8651

Galen Gao
University of Texas Southwestern Medical Center

Liam Spurr
Biological Sciences Division, University of Chicago

Ashton Berger
University of Pennsylvania School of Veterinary Medicine

Gavin Ha
Fred Hutchinson Cancer Research Center https://orcid.org/0000-0001-7578-7272

Veronica Rendo
Dana-Farber Cancer Institute

Matthew Meyerson
Dana-Farber Cancer Institute https://orcid.org/0000-0002-9133-8108

Andrew Cherniack
Broad Institute https://orcid.org/0000-0003-0470-0111

Alison Taylor
Columbia University Vagelos College of Physician and Surgeons

Rameen Beroukhim
Department of Cancer Biology, Dana-Farber Cancer Institute and Harvard Medical School
https://orcid.org/0000-0001-6303-3609

Biological Sciences - Article

Keywords: aneuploidies, cancer, tumorigenesis

Posted Date: May 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-550953/v1

Selective and mechanistic pressures shaping cancer aneuploidies

Juliann Shih[1,2,4], Galen F. Gao[1,3], Liam F. Spurr[1,2,3], Ashton C. Berger[1,3], Gavin Ha[1,2,3,7], Veronica Rendo[1,2,6], Matthew Meyerson[1,2,3,5], Andrew D. Cherniack[1,3,6], Alison M. Taylor[1,3,6,8], Rameen Beroukhim[1,2,3,6]

[1]Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA, USA
[2]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA
[3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
[4]Tufts University School of Medicine, Boston, MA, USA
[5]Department of Genetics, Harvard Medical School, Boston, MA, USA
[6]Department of Medicine, Harvard Medical School, Boston, MA, USA
[7]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[8]Department of Pathology and Cell Biology, Herbert Irving Comprehensive Cancer Center, Columbia University Vagelos College of Physician and Surgeons, New York, NY, USA

ABSTRACT

Aneuploidies, defined as whole-arm or whole-chromosome imbalances, are the most prevalent alteration in cancer genomes. However, the extent to which they are enriched due to selection is unclear, against the alternative hypothesis that they are passenger events that are simply highly prone to occur. We developed a novel method, BrISCUT, that identifies loci under selective advantage or disadvantage due to arm-level copy-number alterations by interrogating length distributions of events that are bounded at either the telomere or centromere. These loci were significantly enriched for known cancer driver genes, including genes not detected through analysis of focal copy-number events, and were often lineage-specific. We also formally quantified the role of selection and mechanistic biases in driving aneuploidy, finding that rates of arm-level SCNAs are most highly correlated with selective pressures. These results provide insight into the causes of aneuploidies and their contributions to tumorigenesis.

INTRODUCTION

Although aneuploidy, which we define as whole-chromosome or whole-arm DNA imbalances, was the first documented somatic alteration in cancer[1], its underlying causes and role in driving cancer remain unclear. The consequences of focal somatic copy number alterations (SCNAs) have been the focus of much computational and functional study[2–6]. However, the consequences of arm-level SCNAs (aSCNAs) are much more difficult to dissect.

Simplistically, stable aneuploidy may be the result of mechanistic breakage due to chromosome missegregation, rearrangements, or centrosome aberrations[7], or to selective fitness advantages these aneuploidies provide. The former is necessary but generally not sufficient: aneuploidy in yeast, mouse, and human cancer cell lines generally decreases proliferation rates and increases cellular senescence, with rescue of proliferation rates only after further evolution[8]. However, aneuploidy is observed in ~90% of solid tumors, with highly tissue-specific patterns[9,10].

This leads to a paradox: experimentally-induced aneuploidy is disadvantageous to cells, but aneuploidy is highly prevalent in cancer, suggesting it confers selective advantages. A robust assessment of the actual fitness effects of different aneuploidies in human tumors is lacking, and experimental methods to assess functional consequences of aSCNAs are technically challenging and have rarely been performed[10–12]. In the context of focal SCNAs, mapping minimal common regions of amplification or deletion can point to relevant oncogenes and tumor suppressor genes[5,13,14]. Arm-level SCNAs, however, always encompass the same hundreds to thousands of genes, so mapping minimal common regions of alteration has no benefit.

However, SCNAs that begin at the telomere and extend almost to the centromere (or vice versa) would be expected to have the same fitness effects as their corresponding aSCNAs, except for the small region that they lack immediately adjacent to the centromere (or telomere). In this study, we develop a new algorithm called BrISCUT (Breakpoint Identification of Significant Cancer Undiscovered Targets) that exploits this source of information, the length distributions of SCNAs that are bounded at the telomere or centromere, to better understand the effects of aSCNAs on fitness and the loci that account for those effects. We apply this approach to over 10,000 tumors in The Cancer Genome Atlas (TCGA) and systematically characterize the influences of selection and mechanistic biases on patterns of chromosome arm aneuploidies within and across cancers.

RESULTS

Impact of arm-level somatic copy number alterations

Arm-level somatic copy number alterations (aSCNAs), SCNAs that extend from telomere to centromere, are among the most frequent and impactful somatic genetic alterations in cancer. Across 10,872 TCGA tumors spanning 33 cancer types, aSCNAs constitute 23 of the 25 most frequent events (**Figure 1a**) and encompass 22.5% of the cancer genome, or 66% of the genome affected by SCNAs in cancer (**Figure 1b**) – more than any other type of somatic genetic alteration (**Extended Data Table 1**). The most common aSCNAs – 20q, 7p, and 8q gains – occur in approximately 30% of cancers, whereas the least common aSCNA, 20q losses, occur in 1.2% of tumors – approximately the same frequency as the most common fusion, *TMPRSS2-ERG* (1.9%). In contrast, focal SCNAs, the genetic events affecting the next largest fraction of the genome, encompass only 11.3% of the cancer genome, or half that of aSCNAs.

Patterns of aSCNAs differ widely between tissue and cancer types, suggesting that they are shaped by epigenetic and environmental factors, but we do not understand how. Many aSCNAs overlap with significantly recurrent somatic mutations, suggesting that these are shaped by individual driver genes. However, they cannot all be explained by the presence of targets such as these: among the 23 most frequent aSCNAs, only 13 encompass a known driver gene that is also altered by either mutation, rearrangement, or focal SCNA in at least 20% of samples with the aSCNA (**Extended Data Table 1a**). Even among cases with purported targets, those targets do not always explain the observed frequencies of aSCNAs. For example, 10q is lost in over 80% of glioblastomas, but the tumor suppressor *PTEN*, the purported target of these losses, is only biallelically inactivated by homozygous deletion or mutation in about 40%[15,16]. It is likely that aSCNAs are prevalent due to the ease of their formation, with frequent breakages in centromeres[7]. Because many genes are affected by each aSCNA, it is also possible that they obtain selective advantage through coordinated disruption of many genes.

High breakpoint densities within centromeres indicate mechanistic biases towards aSCNA formation

Thus, we set out to formally quantify to what extent aSCNAs could be attributed to mechanistic ease of generation versus selective pressures. We approached this question by examining the locations of the breakpoints of SCNAs that begin at either the telomere or centromere, under the assumption that enrichment of breakpoints in specific loci might provide clues to the roles that mechanistic biases and selection play, and to the specific genetic elements undergoing positive and negative selection.

We first considered the relative densities of breakpoints within centromeres versus within chromosome arms to indicate mechanistic biases favoring or disfavoring aSCNAs (**Figure 1c**). Across all chromosome arms and tumor types, 39% of telomere-bounded SCNAs end in centromeres – a four-fold enrichment of breakpoint density in centromeres relative to within chromosome arms (86.5 and 19.7 per megabase, respectively; **Figure 1d**). This observation holds for both amplifications (35.8 breakpoints / Mb in centromere vs. 7.5 in the arm) and deletions (50.7 breakpoints / Mb in centromere vs. 12.2 in the arm). Reasons for this enrichment may include the role of the kinetochore in mitosis and defects in mitotic checkpoint signaling, cohesion, or merotelic attachment[7,17]. It appears that the high frequency of centromeric breakpoints is unrelated to the length of the centromere (**Figure 1e**; **Extended Data Table 5d**). We explore differences in rates of centromeric breakage across chromosomes below.

Use of SCNA length distributions to identify selective pressures contributing to aSCNA rates

Within chromosome arms, we evaluated sub-arm-level SCNAs as sources of information about selective pressures affecting aSCNAs. Consistent with prior findings[2,3], when summed across all chromosome arms and cancer types, SCNAs that start in the telomere or centromere (tSCNAs and cSCNAs, respectively; collectively termed partial SCNAs, or pSCNAs) followed near-uniform length distributions (**Figure 2a**). In contrast, the distribution of interstitial SCNAs was inversely proportional to SCNA length. This was true for amplifications and deletions, though the amplification and deletion distributions – and the tSCNA and cSCNA distributions – differed somewhat (**Figure 2a**). We considered these to be "background" pSCNA distributions in the absence of selective pressures. Moreover, pSCNAs tended to have lower amplitudes (number of copies of change) than within-arm ("interstitial") SCNAs: usually low-level gains and losses and rarely high-level amplifications or homozygous deletions (KS test $p < 2.2e\text{-}16$; **Figure 2b**) – as do arm-level SCNAs. For this reason, we considered the fitness effects of pSCNAs and arm-level SCNAs to be similar for the loci that they both encompass.

Next we compared tumor type-specific, chromosome arm-specific pSCNA length distributions to these background distributions to detect genomic loci subject to selection. Specifically we hypothesized that, compared to the numbers of pSCNAs expected by chance, more pSCNAs would encompass a gene/locus if its alteration conferred selective advantage to the tumor (i.e. positively selected "driver" events), and fewer if its alteration conferred selective disadvantage (**Figure 2c**; **Extended Data Table 2c**). We would therefore observe a sudden jump or fall in pSCNA breakpoint frequencies adjacent to these loci under selection. Indeed, when comparing the near-uniform background model of tSCNA lengths with tSCNAs across several chromosome arms in pan-cancer analyses, we observed four patterns: 1) no deviation from the background model, providing no evidence of selection; 2) a single locus of deviation from the background model, likely representing a single locus subject to detectable (positive or negative) selection; 3) multiple such loci, corresponding to multifocal positive or negative selection; and 4) loci that deviate in opposite directions from the background, indicating balanced selection (**Figure 2c-d**).

We therefore developed a methodology ("BrISCUT", for Breakpoint Identification of Significant Cancer Undiscovered Targets; see **Methods** for a detailed description; **Figure 3a** and **Extended Data Figure 1a**) to formally evaluate evidence for selective pressures from pSCNAs and to identify the loci that are most likely responsible for these selective pressures. BrISCUT first determines whether the distribution of lengths of a specific type of pSCNA on a given chromosome arm differs significantly from its background distribution. If so, it identifies the genomic locus at which the observed and background distributions diverge most. BrISCUT then sets boundaries for a "peak region" around that locus, such that the boundaries would be expected to encompass the genes that drive this divergence. The user pre-specifies the level of confidence desired: a higher level of confidence in encompassing the driver(s) would lead BrISCUT to indicate larger peak regions. Once this locus is identified, the chromosome arm is divided at the locus and BrISCUT is repeated on both the telomeric and centromeric sides of the chromosome arm. This process is repeated until no significant divergence between the observed and expected data is detected. In addition to statistical significance, BrISCUT also estimates effect sizes for each selection peak (see **Methods; Extended Data Table 5a**). We then applied BrISCUT to the 10,872 cancer copy-number profiles generated by TCGA, representing 33 cancer types.

Loci under selection

We detected 193 genomic loci (i.e. peaks) under apparent selection: 90 regions of positive selection (39 from amplifications and 51 from deletions, containing a median of 25 and 12 genes,

respectively) and 103 regions of negative selection (41 from amplifications and 62 from deletions, containing a median of 45 and 25 genes, respectively) (**Figure 3b** and **Extended Data Figure 1b**). These peak loci were significantly enriched for known oncogenes and tumor suppressor genes (**Extended Data Table 3a-b**): among peaks that contained 50 or fewer genes ("restricted peaks", n = 135), 46% (62), encompassed COSMIC Cancer Gene Census Tier 1 genes, which have been strongly implicated in oncogenesis[18] ($p < 1e-5$; **Figure 3c**). This finding held in both peaks subject to positive (31/68; $p < 1e-5$) and negative selection (31/67, $p = 0.00081$). When peaks were further stratified by amplification versus deletion, each category remained significantly enriched for COSMIC genes, with the exception of amplifications under negative selection ($p = 0.384$).

We also found that peaks containing genes affected by focal SCNAs were discovered independent of those focal SCNAs. Among the 23 restricted positive selection amplification peaks, 7 ($p=0.00026$) encompassed recurrently focally amplified genes (*TERT, MYC/PVT1, TERC/PRKCI, SOX4, CDK6, ZNF217, KRAS*)[19]. The 45 restricted positive selection deletion peaks were also significantly enriched for recurrently focally deleted genes (12 peaks; $p < 0.00001$) (*GRID2/ATOH1, CDKN2A, CDKN1B, NCOR1, TP53, ARID1A, PTEN, ATM, SMAD4, IMMP2L/LRRN3, MAP3K1,* and *MACROD2*). However when we repeated our analyses after removing samples with those focal alterations, we often detected the same loci. **Extended Data Figures 1c-d** shows four examples, positively selected deletions encompassing *CDKN2A* and amplifications encompassing *MYC*, and negatively selected deletions encompassing *MYC* and *YAP1*/*BIRC3*. Moreover, the top 25 positive selection deletion peaks from BrISCUT did not contain fragile sites, a marked difference from focal SCNA analyses of these same data, for which 8/25 of the most significant deletion peaks contained known fragile sites ($p = 0.004$; **Extended Data Table 3d**)[20]. Indeed, many of the driver genes identified by BrISCUT were not targets of focal SCNAs. Among the 16 and 33 restricted amplification and deletion peaks without known focal SCNA targets, an additional 7 and 11 respectively contained Tier 1 driver genes from the COSMIC Cancer Gene Census (*DROSHA, NCOA2, CRTC3/BLM/FES, KIF5B, RNF213, UBR5, and PPM1D/BRIP1* among amplifications; *RHOH, RBM15, MAF, NT5C2, CAMTA1, TCF3, NUMA1, WRN/NRG1, BARD1/ATIC, DDB2,* and *ATP1A1* among deletions). This leads us to two conclusions. *First*, genes whose focal amplification or deletion favors oncogenesis often also favor oncogenesis through pSCNAs. This is striking because the focal alterations are often high-amplitude SCNAs, whereas pSCNAs are almost uniformly low-amplitude events (**Figure 2b**). *Second*, BrISCUT provides additional sources of information to detect driver genes.

Prior genome-scale RNAi screening data indicated that knockdown of genes in restricted negative selection deletion peaks led to slightly, but significantly, decreased cell viability compared to other assessed genes ($p = 0.0035$; **Figure 3d**)[21]. Given that each peak tends to contain many genes, of which perhaps only one is under significant negative selection, we would expect that the average viability scores across all those genes would not be much lower than those of other genes, as was observed. Increased sample numbers would be expected to increase the resolution of these peaks, thus decreasing the number of genes they encompass (**Extended Data Figure 2**) and therefore increasing the differences in viability scores.

Recognizing that selection in cancer can be lineage-specific[22], we asked whether differences in pSCNA distributions across lineages predicted differences in aSCNA rates. Specifically, we compared lineage-specific pSCNA distributions for each chromosome arm to the pan-cancer pSCNA distributions for that same chromosome arm, using a Jensen-Shannon divergence metric (**Extended Data Figure 3**). These pSCNA lineage-specificity scores correlated tightly with lineage-specific aSCNA rates on the same chromosome arms (Fisher's method p-value 1.6e-3; **Extended Data Table 4b**). We also found that different chromosome arms differed substantially in the degree to which

their pSCNA distributions varied across cancer types (**Extended Data Figure 4a; Extended Data Table 4a**). We conclude that pSCNA distributions reflect differences in selective pressures across lineages.

We therefore applied BrISCUT to the 33 individual TCGA tumor types to detect lineage-specific aSCNA drivers. The median number of peaks found in these cohorts was 9, ranging from none (in DLBC, KICH, LAML, THCA, and THYM) to 59 (in OV). Additionally, we analyzed combined groups of shared lineage (COADREAD, ESCASTAD, GBMLGG, KIPAN, and PANSCC), notable sub-lineages (BRCA-basal, BRCA-luminal, ESCASTAD-CIN, and ESCASTAD-GS), and distinguished between tumors that underwent whole genome doubling (WGD) and those that did not (DIPLOID) (**Extended Data Table 4c-f**). Altogether, we detected 825 peaks across these varied lineages, with 397 occurring in the 33 independent cohorts. Among these, 331 peaks, or 83%, overlapped with at least one peak in another lineage (not including the pan-cancer analysis), leaving 66 unique non-overlapping peaks across all lineages (**Extended Data Figures 4b-c; Extended Data Table 4g** indicate examples on 3p). Among independent cohorts, overlapping peaks occurred more often among related developmental lineages[23] than expected by chance ($p = 0.001$, **Extended Data Figure 5**; see **Extended Data Note**), mirroring the association between aSCNA rates and developmental lineage.

Quantitative assessment of selective and mechanistic pressures driving aneuploidy

Along with indicating loci that may be subject to selection, these analyses provide estimates of the extent to which both positive and negative selection contribute to observed rates of aSCNAs. Specifically, we considered the change in breakpoint frequency across each BrISCUT peak as indicating the magnitude of selective pressure ("selection effect size") that derives from that locus (**Figure 4a**). By aggregating selection effect sizes of all BrISCUT peaks on each arm, we were able to estimate the total positive, negative, and overall net selective pressures (represented as "selection scores") contributing to that aSCNA.

There was evidence of significant positive and negative selective pressures for 50% (39/78) and 71% (55/78) of aSCNAs respectively; 32% (25/78) exhibited both. Of these, those with the most positive selection were deletions of 1p, 17p, and 8p, and amplifications of 5p (positive selection scores of 4.67, 4.24, 3.98, 3.84, respectively; **Extended Data Table 5b**). The aSCNAs with the most negative selection were deletions of 1p and 19q (negative selection scores of -4.04 and -3.24, respectively) and both amplifications and deletions of 11q (negative selection scores of -4.04 and -3.29, respectively), which might explain why chromosome 11 is affected by fewer aSCNAs (3.8%) than any other non-acrocentric chromosome (range 2.8-6.9%) except chromosome 2 (2.8%). Conversely, there was no evidence of significant selective pressure in either direction for 12% of aSCNAs (9/78; amplifications of 2q, 9p, 16p, 16q, 18p, 20p, and 22, and deletions of 18p and 15). Notably, 18p was the only chromosome arm without evident selection.

Deletions appeared to undergo more positive and negative selective pressures (average scores of 1.12 and -1.42 respectively) than amplifications (0.65 and -0.95 respectively). However, we found no significant difference between net selection scores (i.e. sum of selection effect sizes of all BrISCUT peaks contributing to an aSCNA) of deletions and amplifications (average -0.30 for both; $p = 0.86$). The aSCNAs with the highest net selection scores (i.e. most positive selection) were amplifications of 10p, 8q, and 7p, and deletions of 8p (net selection scores of 3.46, 2.89, 2.47, and 2.32, respectively). Those with the lowest net selection scores (i.e. most negative selection) were deletions of 8q, 3q, and 17q, and amplifications of 8p (net selection scores of -3.19, -2.93, -2.93, and -2.76, respectively).

In addition to selection scores, we also generated two other metrics: 1) "centromeric mechanism scores" reflecting observed rates of aSCNAs ending in the centromere compared to rates of tSCNAs ending adjacent to it, reflecting biases favoring breakage within specific centromeres, and 2) "telomeric mechanism scores" reflecting frequencies of tSCNAs that do not encompass any loci under evident selection, reflecting mechanistic biases favoring tSCNAs across different chromosome arms (**Extended Data Figure 6a**).

Notably, all centromeric mechanism scores were greater than 0, suggesting that there are positive mechanistic biases favoring breakage in all centromeres, relative to elsewhere in the chromosome. We would expect that aSCNAs would be easier to generate in the long arms of acrocentric chromosomes than in non-acrocentric arms, because the former represent both whole-chromosome and arm-level SCNAs. Indeed, the centromeres of acrocentric chromosomes 13, 14, 15, 21, and 22 had significantly higher mechanism scores than other arms (average mechanism scores of 3.23 and 1.79; $p = 0.0003$). Excluding these, the centromeres with the highest mechanism scores were those of chromosomes 3, 5, and 17 (mechanism scores of 3.31, 2.88, and 2.37, respectively), while the centromeres of chromosomes 9, 1, and 16 had the lowest mechanism scores (0.76, 0.90, and 0.99 respectively).

We hypothesized that long telomeres protect chromosomes from telomeric copy number events, and thus chromosome arms with longer telomeres would have lower telomeric mechanism scores. Indeed, we found that this was the case (Spearman's rho = -0.65; p = 6.83e-6) (**Figure 4b**).

To help determine the contribution of selection versus telomeric and centromeric mechanisms on aSCNA formation, we tested the level to which the scores representing each of these correlated with aSCNA rates within and across cancers (**Extended Data Figure 6b-d**). The only significant associations were between selection scores and aSCNA rates. This was true for both amplifications and deletions, and for both negative and positive selection scores, with the sole exception of negative selection scores and deletions of the corresponding chromosome arms (**Extended Data Figure 6d**). However, as expected, the correlation was strongest between aSCNA rates and net selection scores, which reflect the summed effects of both positive and negative selection (Spearman rho = 0.72 and 0.53 for amplifications and deletions respectively; **Figure 4c**). Furthermore, we found evidence of correlation ($p < 0.1$) between aSCNA rates and net selection score in 61% (20/33) of the unique TCGA cohorts (Fisher's method p-value 3.5e-39; **Figure 4d** and **Extended Data Table 5e**); the other 13 cohorts were either small or primarily unaffected by SCNAs. Combined with our finding above that pSCNA lineage-specificity scores correlated tightly with lineage-specific aSCNA rates on the same chromosome arms (**Extended Data Table 4b**), we conclude that selective pressures are the main drivers of relative aSCNA rates both between chromosome arms and across cancer types.

DISCUSSION

This study represents the first systematic, SCNA-based attempt to answer the longstanding question of whether or not aneuploidy is positively selected for in cancer. While this has been widely assumed given the widespread and non-random nature of aneuploidy, *in vitro* models of aneuploidy have not supported this hypothesis[8,10]. In our analysis of tumor samples, however, we find strong evidence that SCNA-mediated selective pressures are indeed highly associated with rates of aneuploidy in cancer, often in tissue-specific manners. The relative frequencies between arm-level SCNAs both within and across cancer types appears to be determined primarily by the selective pressures acting on them. Whereas studies have shown that the effects of positive selection from coding point mutations greatly outweigh those of negative selection in cancer[22], we show that both are significant in the SCNA context. We also demonstrate that length distributions of low-amplitude telomere-bounded and centromere-bounded SCNAs – previously underappreciated subsets of somatic alterations – contribute new information to detect loci under selection in cancer.

As in all driver detection analyses based upon the recurrence of genetic alterations in cancer (e.g. SCNAs, single nucleotide variants, and rearrangements), the BrISCUT methodology requires a comparison of observed data to a background model of the expected distribution of data in the absence of selective pressures, and therefore can be biased by inaccuracies in this background model. In the case of telomere- and centromere-bounded SCNAs, we estimate a near-uniform background distribution based upon the average length distributions of pSCNAs across all chromosome arms. However, if this distribution is heavily altered by selective or mechanistic biases or technical artifacts that are specifically associated with a subset of chromosome arms, the background distribution and therefore the BrISCUT results could reflect these biases rather than selective pressure. Further studies into the background distribution of pSCNAs, for example through analyses of pSCNAs being generated at the single-cell level and detected prior to the effects of selection, would greatly aid the detection of loci under selection. In addition, all candidate driver loci indicated by this or any other recurrence analysis require experimental validation. Novel chromosome engineering techniques that generate pSCNAs *in vitro* with precisely targeted breakpoints[10–12] could enable investigations of the expression and phenotypic (including proliferative) effects of tissue-specific pSCNAs and aSCNAs in appropriate cellular contexts.

Despite the above caveats, a key advantage to the BrISCUT analysis is that it relies on step-function changes in breakpoint frequencies of pSCNAs rather than focally recurrent breakpoints, thus limiting the effects of localized fragility. A single fragile locus with high local SCNA rates will not be falsely identified as undergoing selection both because it will not generate step-function changes, and because our data demonstrate that most SCNAs at fragile sites are interstitial. Indeed, none of the top 25 BrISCUT deletion loci encompass known fragile sites. In contrast, 8 of the top 25 recurrent focal deletion peaks identified by GISTIC do.

BrISCUT is unlike other forms of recurrence analysis for somatic genetic events because it relies on the distribution of breakpoints across chromosome arms and not maximal frequencies of alterations. Because it relies on widely dispersed signals, large sample numbers are required to gain high resolution into the precise loci under selection. Conversely, the spatial resolution of the assay used to profile copy numbers in each sample is less critical to BrISCUT than to methods that rely on focal genetic alterations. Fortunately, both the decreasing costs of sequencing and the increasing prevalence of clinical sequencing are likely to provide very large numbers of samples that have undergone copy number profiling. Moreover, whole genome sequencing maps SCNA breakpoints better than microarray data, as were used here. BrISCUT can also be adapted to detect loss-of-heterozygosity events and to distinguish between different absolute copy number states. For these

reasons, we expect that BrISCUT and similar analysis approaches will have increasing impact over time.

Our analyses indicate that the ubiquity of aSCNAs in cancer is due to both selective and mechanistic pressures. The fragility of centromeres supports the frequency with which aSCNAs are observed overall; the relative frequencies of different aSCNAs appear to be determined primarily by selective pressures according to our analyses. Prior studies have also suggested centromeric fragility[2,24,25], but had not quantified the amplitude of its effects or variations across chromosomes. We find that, on average, breakpoints occur in centromeres four times as frequently relative to immediately adjacent loci that should be undergoing similar selective pressures. This varies widely – from two-fold to ten-fold – across chromosomes. The sources of these discrepancies are unclear and may include differences in (peri-)centromeric DNA sequences[26], three-dimensional folding of DNA[27], or how different centromeres interact with the centrosome via kinetochores[28]. Future work to computationally model the likelihood of specific mechanistic processes (e.g. merotelic attachment versus telomeric erosion)[29,30] underlying not only aneuploidies, but also telomere- and centromere-bounded SCNAs, would further our understanding of these highly common, yet mysterious, events in cancer biology.

METHODS

Generation and post-processing of segmented data from Affymetrix SNP 6.0 arrays

DNA from 10,872 tumors and their matched germline samples was hybridized to Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute as previously described[31]. Briefly, from raw .CEL files, Birdseed was used to infer a preliminary copy number at each probe locus[32]. For each tumor, genome-wide copy number estimates were refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor[33,34]. This linear combination of normal samples tended to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy number profile. Individual copy number estimates then underwent segmentation using Circular Binary Segmentation, or CBS[35]. As part of this process of copy number assessment and segmentation, regions corresponding to germline copy number alterations were removed by applying filters generated from TCGA germline samples.

The ABSOLUTE algorithm[36] was applied to data from these cancers, along with whole-exome sequencing data from the same cancers when available (10,162 samples). Purity and ploidy estimates and allelic copy numbers were called successfully in 10,497 samples. For samples with ABSOLUTE corrected copy number, CBS-derived segmented copy number values were re-centered using the *In Silico* Admixture Removal (ISAR) procedure[3].

Deconstruction of copy number segments into whole-arm, telomere-bounded, centromere-bounded, and interstitial SCNAs

In order to study genomic regions likely to confer survival advantage or disadvantage if copy number altered, we first needed to distinguish between different types of SCNAs in cancer, namely interstitial/focal, whole-arm, and partial SCNAs (further split into telomere-bounded and centromere-bounded SCNAs), as well as determine background rates for these subtypes. In some cases, aSCNAs or pSCNAs might be divided by an interstitial SCNA (e.g. an arm-level gain with a small deletion in the arm). To accurately call the full length of aSCNAs and pSCNAs, we joined copy number-altered segments likely to represent single events and adjust the amplitudes of overlaying focal events accordingly.

Here, we assumed that, to a first-order approximation, the distribution of pSCNA lengths was uniform while the distribution of interstitial SCNA lengths decreases as 1/segment length (**Figure 2a**). In cases where a telomere- or centromere-bounded segment neighbored another segment in the same direction (ie gain or loss), we accounted for two possibilities: *first*, that these represented separate SCNAs (a pSCNA and a neighboring interstitial SCNA in the same direction), and *second*, that they represented a single pSCNA with an intervening interstitial SCNA in the opposite direction. In either case, the pSCNA would have the same probability, due to the near-uniform length distribution of pSCNAs. The probability of the interstitial SCNA, however, would be greater for the smaller SCNA. Therefore, we chose between these possibilities the one involving the smaller interstitial SCNA. A similar analysis applied in cases of three or more neighboring segments. We therefore recorded all altered segments on each specified chromosome arm in each specified direction (amplification and deletion defined as log2 copy number ratio > 0.2 and < -0.2 respectively). If a segment spanned the centromere, we split it into two separate segments. We then calculated the distance between the end of the telomere- or centromere-bounded segment and the end of the last altered segment (i.e. furthest from the telomere or centromere). If the total length of copy-number altered DNA was greater than that of non-copy-number altered DNA, we recorded the

end of the last altered segment as the breakpoint location of the telomere-bounded SCNA. However, if it was not, we iteratively removed the last altered segment until this is true.

We recorded the breakpoint location for each pSCNA, which is equivalent also to the length of the pSCNA, as a fraction over the length of the arm. Arm-level SCNAs were further classified as "centromeric" if they affected only the arm in question and did not extend at all into the other arm of the chromosome. Acrocentric chromosomes 13, 14, 15, 21, and 22 were excluded from this classification. "Centromeric" aSCNAs were included in the calculation of the centromeric mechanism score, whereas non-centromeric ones were not.

Compilation of most frequent somatic alterations in cancer

Rates of aneuploidy were derived for 10,872 tumors across 33 TCGA tumor types from the copy number segment-joining method detailed above; only whole-arm SCNAs (aSCNAs) were considered. Somatic mutations were called from 9,423 tumor exomes across 33 TCGA tumor types[37]. Rates of mutation were reported for 299 likely driver genes.

We determined focal SCNA rates for 10,872 tumors across 33 TCGA tumor types by running GISTIC 2.0 (version 2.0.23 on GenePattern https://cloud.genepattern.org/gp/pages/index.jsf) on segmented data containing only amplitude-corrected interstitial events (see segment-joining method above). We used a noise threshold of 0.2, broad length cutoff of 0.5 chromosome arms, confidence interval of 99%, and copy-ratio cap of 1.5. For the top 25 most significant focal amplifications and deletions separately, we calculated their frequencies of focal alteration, defined as >0.2 or <-0.2 copy number in the output file focal_data_by_genes.txt.

Fusion genes were identified from 9,624 tumors across 33 TCGA tumor types using various fusion calling tools[38]. Pan-cancer fusion gene rates were reported for fusion genes found to be recurrent in any tumor type. Fusions between the same two genes, regardless of pair order (e.g. *TMPRSS2-ERG* vs. *ERG-TMPRSS2*) were considered as the same event, and reported in alphabetical order.

Hyper- and hypo-methylation leading to epigenetic silencing or enhancement were determined from 5,898 tumors across 24 TCGA tumor types/subtypes (ACC, BLCA, BRCA-basal, BRCA-nonbasal, CESC, COADREAD, ESCA, GBM, HNSC, KICH, KIRP, LAML, LGG, LIHC, LUAD, LUSC, OV, PRAD, SKCM, STAD, THCA, UCEC, and UCS) using the RESET method[39]. Pan-cancer rates were calculated for genes that were significantly silenced or enhanced in at least one tumor type by back-calculating the total number of events in each tumor type. Although there was likely methylation of these genes in other tumor types, these events were not included because there was no evidence of correlation of methylation with gene expression.

BrISCUT peak finding algorithm

BrISCUT performs two main functions: 1) it detects loci that appear to be under selective pressure; and 2) it determines confidence intervals ("peak regions") bounding each of these loci, within which the specific site undergoing selection is likely to be present at a preset level of confidence.

To detect loci that are likely to be under selective pressure, we use a two-sample Kolmogorov-Smirnov test, comparing our set of pSCNA breakpoints $T_i = \{t_1, t_2, \cdots, t_n\}$, sorted in increasing order, to the empirical "background" distribution, comprising all telomere-bounded SCNA lengths across the dataset of 10,872 cancer specimens across 33 cancer types. We called loci meeting the criteria $n \geq 4$ and $p_{KS} \leq 0.05$ as under selection, where n is the total number of pSCNAs.

The boundaries of the "peak region" are detected such that they include the gene, set of genes, or other genomic elements that confer selective pressure when altered at a confidence level of γ, where γ is a user-specified parameter. To simplify these calculations, we approximate the empirical background distribution by an incomplete beta function:

$$B_i = I_x(x; \alpha, \beta),$$

where x is the pSCNA breakpoint location and α and β are the parameters of the beta function. We selected a beta function as the best-fit univariate distributions to the empirical data among the following distributions: normal, exponential, Poisson, gamma, logistic, binomial, geometric, beta, and uniform. We determined this fit and α and β for each of the four groups of partial SCNAs: telomere-bounded amplifications, telomere-bounded deletions, centromere-bounded amplifications, and centromere-bounded deletions, using the fitdist function from the fitdistrplus R package (version 1.0-14)[40] (**Extended Data Table 2e**).

We determine whether the strongest genomic region of selection confers positive or negative selection using the following equation:

$$direction\ of\ selection = \{positive\ if\ \ |(T_i - B_i)| > |(T_i - B_i)|, else\ negative\}$$

Then, the "starting peak," or genomic locus from which to expand the peak region to define the final set of boundaries, is the breakpoint location (*peak*) at which T and B maximally diverge.

We then define boundaries on either side of this starting peak using a helper function, $H(p)$, that determines whether the breakpoint directly to the left and right of *peak* have a 95% chance of belonging to a distribution unaffected by selective pressures, i.e. following the background $Beta(\alpha, \beta)$. On the left, this distribution, which approximates the generalized extreme value distribution $GEV(x)$, is created by taking the 25th most distant locus among 1000 independently generated random loci to either the left or right of this peak from the background beta distribution. Note that the 95% parameter can be adjusted by the user.

We repeat the entire BrISCUT method recursively to the right and left of the calculated peak boundaries until one of the following is true: 1) there are not at least 4 breakpoints in the analysis, 2) significance is not reached, 3) a tentatively discovered peak overlaps with one that occurred in a prior iteration, or 4) a tentatively discovered peak covered more than half the length of a chromosome arm.

Provided there were at least 4 breakpoints, all KS p-values are corrected for multiple hypothesis testing using the Benjamini-Yekutieli method[41], which controls the false discovery rate (FDR) under complicated dependence structures including both positive and negative dependencies. Peaks were considered significant if their Benjamini-Yekutieli-corrected q-values were ≤ 0.05. The genes listed in each peak region include all protein-coding genes, microRNAs, and additional noncoding RNAs from NCBI's RefSeq release 85 as of February 3, 2018. If a peak (e.g. iteration 2) is dependent on a previous peak (e.g. iteration 1) that has been removed from significance due to multiple hypothesis correction, the dependent one is also removed from the final results.

Enrichment of known cancer genes in BrISCUT peak regions

To determine whether known cancer genes were enriched in peak regions, we compared the number of regions with genes reported to be cancer-driving genes from the COSMIC Cancer Gene Census[18] to permuted datasets in which each gene in each region was replaced by a gene randomly selected from elsewhere in the genome, repeating this 100,000 times. Two gene lists were used (after filtering for genes only on autosomal chromosomes and covered by the Affymetrix SNP 6.0 array): one containing 663 genes from both Tier 1 and Tier 2, and one containing 527 genes from only Tier 1. Then, for each list, we calculated the p-value as the number of permutations with more peaks containing a driver gene than actually observed, divided by 100,000.

Lineage-specificity of breakpoint distributions

To determine the relative lineage-specificity of pSCNAs involving specific chromosome arms, we first compared the breakpoint vector of a pSCNA (e.g. 3p telomere-bounded deletions) within a specific tumor type (e.g. KIRC) to that of all other samples in our dataset by computing the log2 Jensen-Shannon Divergence (JSD)[42,43] between their quantile values (total of 101 values). The JSD is a measure of similarity between probability distributions in which a low value indicates similarity and a high value indicates dissimilarity. We normalized for the number of tumors contributing to a single score by multiplying the log2 JSD by $\frac{n}{\sqrt{n}}$ to arrive at the "divergence score". We then report the variance of these values across different tumor types within the same pSCNA (amplification or deletion in each chromosome arm) as the "lineage-specificity score" (**Extended Data Table 4a**).

Overlap of BrISCUT peak regions and clustering analysis

Two peak regions from different cohorts were considered to overlap if their 95% confidence intervals intersected. Peaks were only compared to other peaks sharing the same directionality (i.e. amplification vs. deletion), pSCNA type (i.e. telomere-bounded vs. centromere-bounded), and selection (i.e. positive vs. negative).

To determine peak regions that significantly overlap across all of the tumor types, we also ran GISTIC 2.0 (version 2.0.23) on segmented copy-number files generated from the 95% confidence intervals of the BrISCUT peaks. Telomere-bounded and centromere-bounded peaks were combined, but negative and positive selection peaks were separated, such that each row within the segmented file was represented by a combination of a tumor type and direction of selection: e.g. LUAD_n or BRCA_p. For positive selection, peaks derived from amplifications were considered to have positive amplitude equal to their "significance score" (KS statistic * -log10 q-value), and peaks from deletions were considered to have negative amplitude. For negative selection, peaks from deletions were considered to have positive amplitude and peaks from amplifications had negative amplitude. GISTIC 2.0 was run with a confidence interval of 99% for positive selection peaks and negative selection peaks separately. All unique tumor types were clustered based on the thresholded copy number at recurring peaks from the "all_lesions.txt" file from GISTIC. Hierarchical clustering was performed in R using Euclidean distances and Ward's method (Ward.D).

Power and accuracy analysis on simulated datasets

In order to assess BrISCUT's ability to detect driver events, we generated *in silico* sets of pSCNA breakpoints, with various degrees of simulated selective advantage or disadvantage, and at multiple locations across a chromosome arm (**Extended Data Figure 2a**). Background

distributions for tSCNAs and cSCNAs were assessed independently; within those, amplifications and deletions were combined (**Extended Data Table 2e**).

To create the breakpoint data, we started by generating a set of *n* random samples from the corresponding beta background distribution. We defined several locations for theoretical driver genes *l* = [0.1, 0.2, …, 0.9], which represents the distance across the chromosome arm (0 at the telomere and 1 at the centromere for telomere-bounded SCNAs, and vice versa). We introduced several levels of selective pressure *s* = [0.02, 0.05, 0.1, 0.2, 0.25, 0.5, 1, 2, 4, 5, 10, 20, 50], where *s* represents the likelihood of a tumor that contains the driver event relative to a tumor that does not contain the driver event. For each *b* within the set of *n* random samples derived from the background distribution, we then include it at a rate of: $r_b = \frac{1}{1+s}$ if $b < l$, otherwise $r_b = 1 - \frac{1}{1+s}$. This was repeated 100 times for each combination of *n*, *l*, and *s*, separately for tSCNAs and cSCNAs.

We then ran BrISCUT at a confidence level of 0.95 on the simulated sets of tSCNAs and cSCNAs separately. For each combination of *n*, *l*, *s*, and type of pSCNA (tSCNA vs. cSCNA), we report the frequency at which BrISCUT correctly includes the locus *l* in its peak region as power (also known as sensitivity or recall). We also calculate the positive predictive value (PPV, also known as precision) by dividing the number of detected peaks containing the locus *l* by the total number of detected peaks. If BrISCUT did not detect a peak in a particular set of breakpoints, this analysis was removed from the denominator. The F1 score was calculated as the harmonic mean of precision and recall: $F_1 = 2 * \frac{precision * recall}{precision + recall}$. For each statistic, we generated a "combined" statistic by taking the weighted average of the statistic from tSCNAs and cSCNAs, where weights are defined as the total number of tSCNAs and cSCNAs in our dataset ($n = 51{,}588$ and 34,007 respectively).

Peak effect sizes and selection score

To calculate the effect size of each BrISCUT peak (i.e. the amount of positive or negative selection conferred by the partial copy number alteration of a given locus), we first separated peaks by chromosome arm, amplification vs. deletion, and telomere-bounded vs. centromere-bounded, and within these ranked each statistically significant peak by its genomic location as fraction of chromosome arm length. If there was at least one BrISCUT peak, we considered the tumors with pSCNAs smaller than the length ascribed to the leftmost (i.e. smallest) BrISCUT peak to be under no selective advantage or disadvantage (i.e. "reference"), where the empirical number of pSCNAs is equal to the expected number E_0^1. Henceforth, we considered the segment $[S_p^{end},\ S_{p+1}^{start}]$ between each peak *p* and *p+1* to be immediately affected by the selective advantage or disadvantage conferred by peak *p*. For each segment between two peaks, we calculated the number of pSCNAs expected to be in this segment in the absence of selection as:

$$E_p^{p+1} = (I_x\left(S_{p+1}^{start}\ ;\alpha,\beta\right) - I_x\left(S_p^{end}\ ;\alpha,\beta\right)) * \frac{E_{p-1}^{p}}{I_x\left(S_p^{start}\ ;\alpha,\beta\right) - I_x\left(S_{p-1}^{end}\ ;\alpha,\beta\right)},$$

where $I_x(x;\alpha,\beta)$ is the incomplete beta function and α and β are the corresponding probability parameters. We then reported the effect size for a peak *p* as the number of empirical pSCNAs in $[S_p^{end},\ S_{p+1}^{start}]$ over the expected number E_p^{p+1}. This value is greater than 1 for positive selection and smaller than 1 for negative selection (**Extended Data Figure 6a**).

The selection score is a representation of the density and effect sizes of BrISCUT peaks on each arm for amplifications and deletions separately. For each aSCNA, positive and negative selection peaks are assessed both separately (positive and negative selection scores respectively) and together (net selection score). Specifically, we define the selection scores as the log2 value of

the product of the relevant effect sizes (only those greater than 1 for positive selection scores, only those smaller than 1 for negative selection scores, and all effect sizes for net selection score), such that scores greater than 0 represent positive selection, and those less than 0 represent negative selection. Only telomere-bounded peaks were included in this analysis.

Centromeric and telomeric mechanism scores

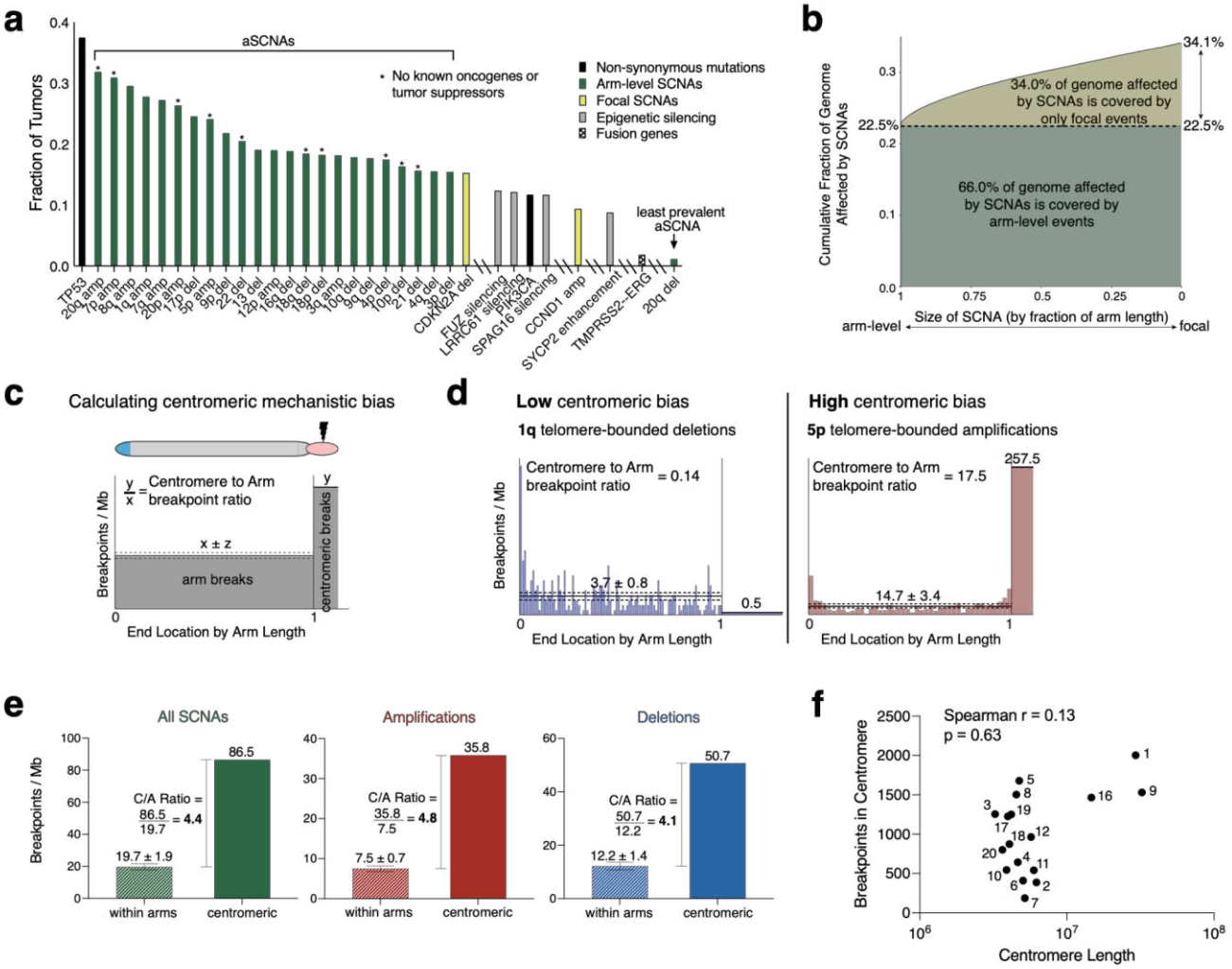
We developed a centromeric mechanism score to represent the likelihood of an SCNA-causing breakpoint occurring in a specific centromere relative to the likelihood of one occurring within its flanking chromosome arm(s), corrected for the selection affecting those arm(s). This was calculated as the average of four individual values (two for acrocentric chromosomes): mechanism scores for amplifications and deletions of the short and long arms (only long arms for acrocentric chromosomes).

To calculate centromeric mechanism scores for each arm and direction of SCNA (i.e. amplification versus deletion), we divided the density of aSCNA breakpoints within the centromere (the number of "centromeric" aSCNAs, see above) by the density of breakpoints within the region immediately flanking the centromere up to the nearest BrISCUT peak. For acrocentric chromosomes, the density of breakpoints within both the centromere and p arm was used as the numerator, since the p arm lacked coverage. We used the density of pSCNA breakpoints only out to the nearest BrISCUT peak because these pSCNAs likely underwent similar selection as the centromeric aSCNAs and we wanted to exclude effects of selection when calculating these mechanism scores.For consistency with selection scores, we reported the log2 value of this quotient, such that a value greater than 0 represents positive mechanism and a value less than 0 represents negative mechanism.

We calculated telomeric mechanism scores for each chromosome arm and direction of copy-number change as the number of tSCNA breakpoints occurring prior to the start of the BrISCUT peak closest to the telomere, divided by the incomplete beta function (i.e. the cumulative probability distribution) at the start of the first BrISCUT peak, measured as the fraction of the length of its chromosome arm. These scores were designed to reflect the propensity for telomeric events to occur in the absence of selection.

MAIN FIGURES

Figure 1: Prevalence and characteristics of different types of SCNAs.



(a) Fraction of 10,872 TCGA tumors exhibiting specific somatic alterations. The 25 most frequent events and a selection of representative alterations are depicted.

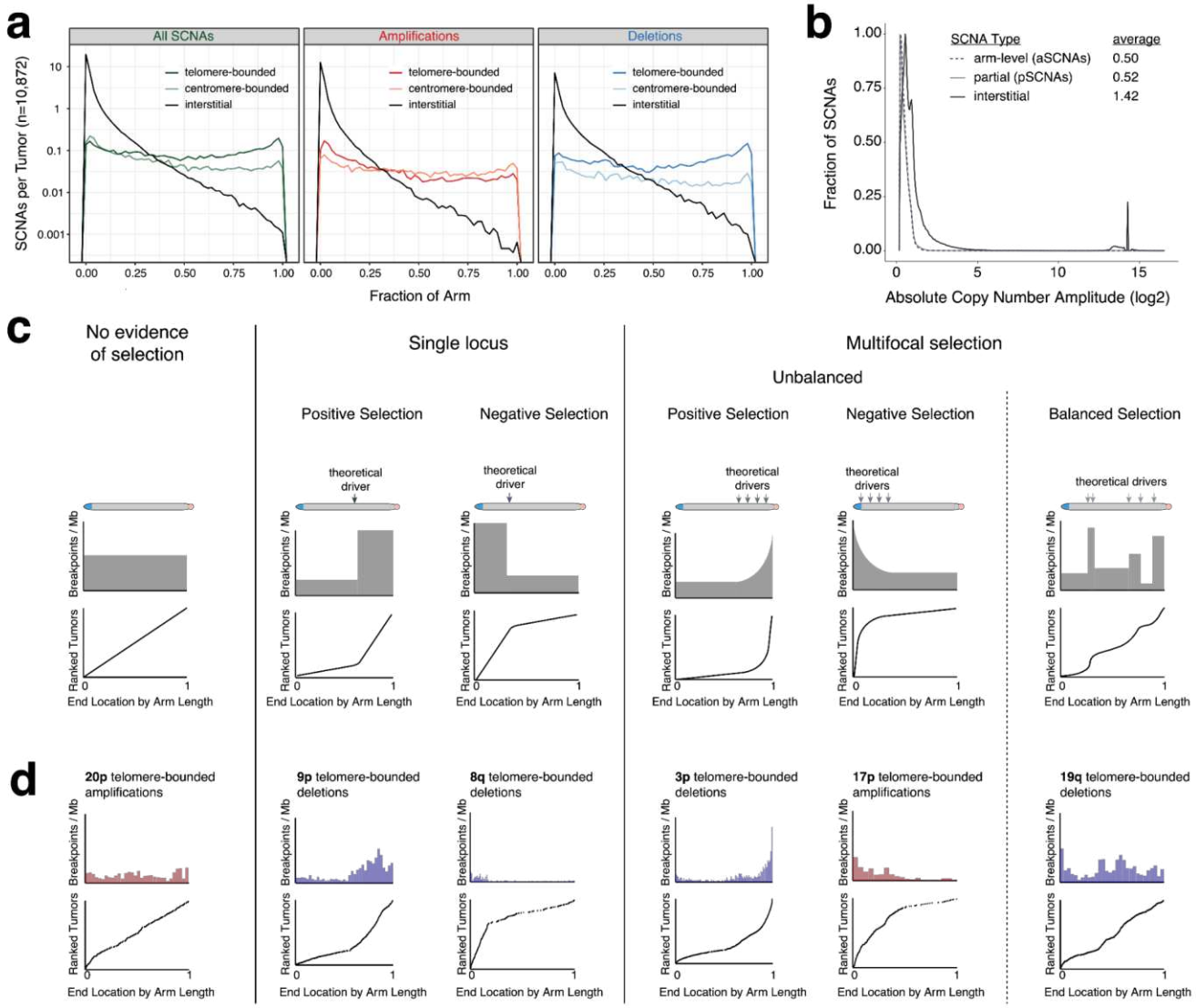
(b) Cumulative fraction of cancer genomes that is affected by somatic copy number alterations (SCNAs, y-axis), plotted inversely by size of SCNAs from arm-level SCNAs (aSCNAs) to small focal events (x-axis). The green region represents the fraction of the genome covered by aSCNAs, and the yellow region represents the fraction of the genome additionally covered by focal SCNAs.

(c) Schematic representation of centromeric mechanistic bias. The line underneath the chromosome arm ($x \pm z$) represents the number of breakpoints per Megabase (Mb) within the chromosome arm (dashed lines are the 95% confidence interval for the mean), and the line under the centromere (y) represents the breakpoints per Mb within the centromere. The quotient of y / x represents the centromere to arm breakpoint ratio (C/A Ratio).

(d) Empirical examples of low centromeric mechanistic bias (e.g. 1q telomere-bounded deletions, for which the ratio of breakpoints occurring in the centromere over those occurring in the arm is less than 1), and high centromeric mechanistic bias (e.g. 5p telomere-bounded amplifications, for which the centromere/arm breakpoint ratio is much greater than 1). Within the chromosome arm, bins are 1 Mb large.

(e) Mean breakpoint density within chromosome arms, aggregated across all tumors and all chromosome arms, versus breakpoint density within all centromeres (values in breakpoints per megabase). Error bars represent the 95% confidence interval for the mean. C/A Ratio represents centromeric breaks over arm breaks.
(f) Total number of breakpoints occurring in the centromere that cause SCNAs plotted against centromere length, which includes pericentromeric regions that lack coverage in the SNP arrays.

Figure 2: SCNA length distributions follow distinct patterns dependent on selective pressures.



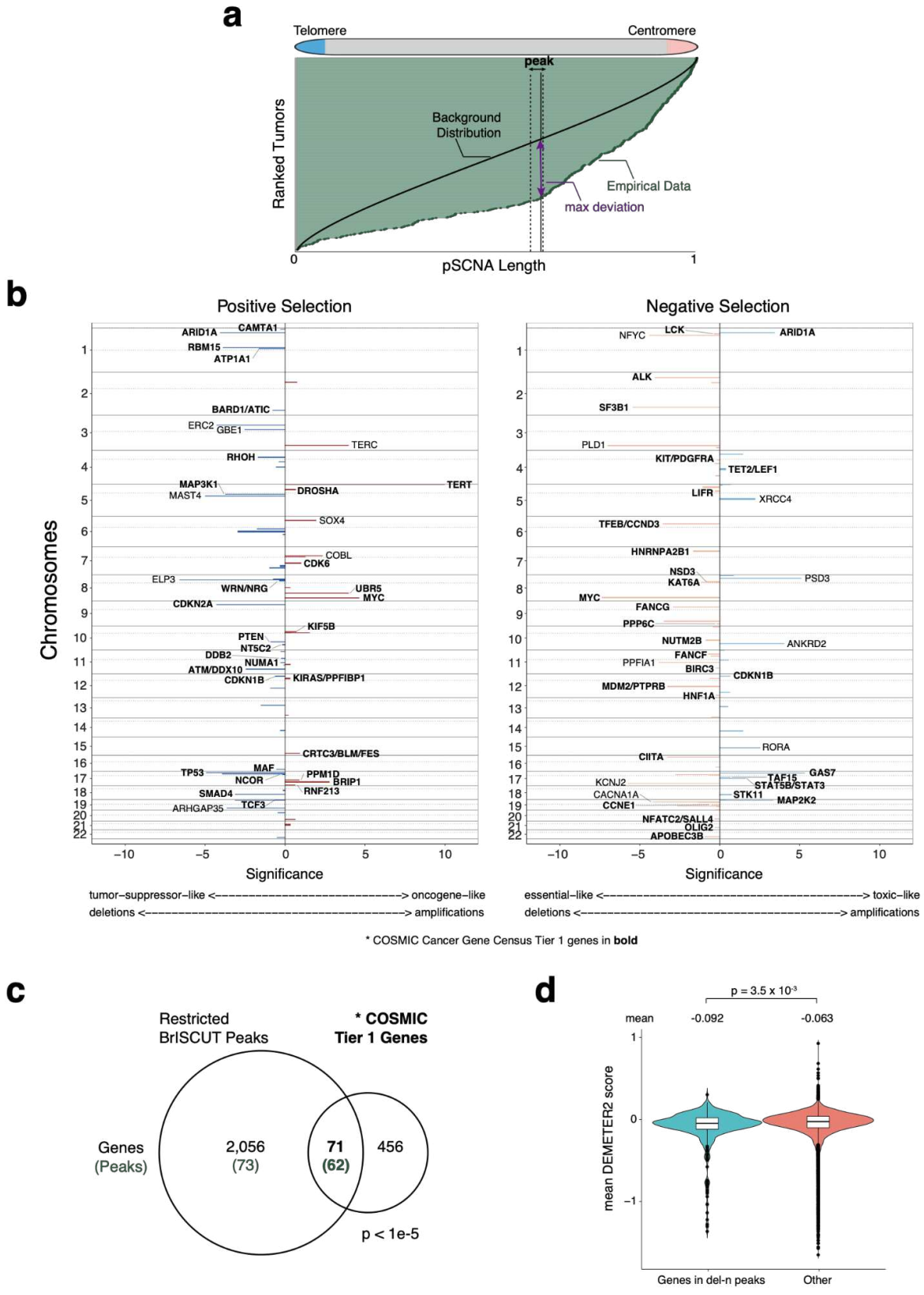
(a) Comparison of length distributions of telomere-bounded, centromere-bounded, and interstitial SCNAs, aggregated across all chromosome arms.

(b) Amplitude distributions and mean log2 copy number of arm-level, partial, and interstitial SCNAs. Amplitudes are calculated as the absolute value of a weighted average of the amplitudes of segments included in the SCNA (see **Methods** for details). Curves are scaled according to the total number of SCNAs within each category, to a maximum of 1.

(c) Schematics of SCNA-mediated selection; specific patterns are explored further in the main text.

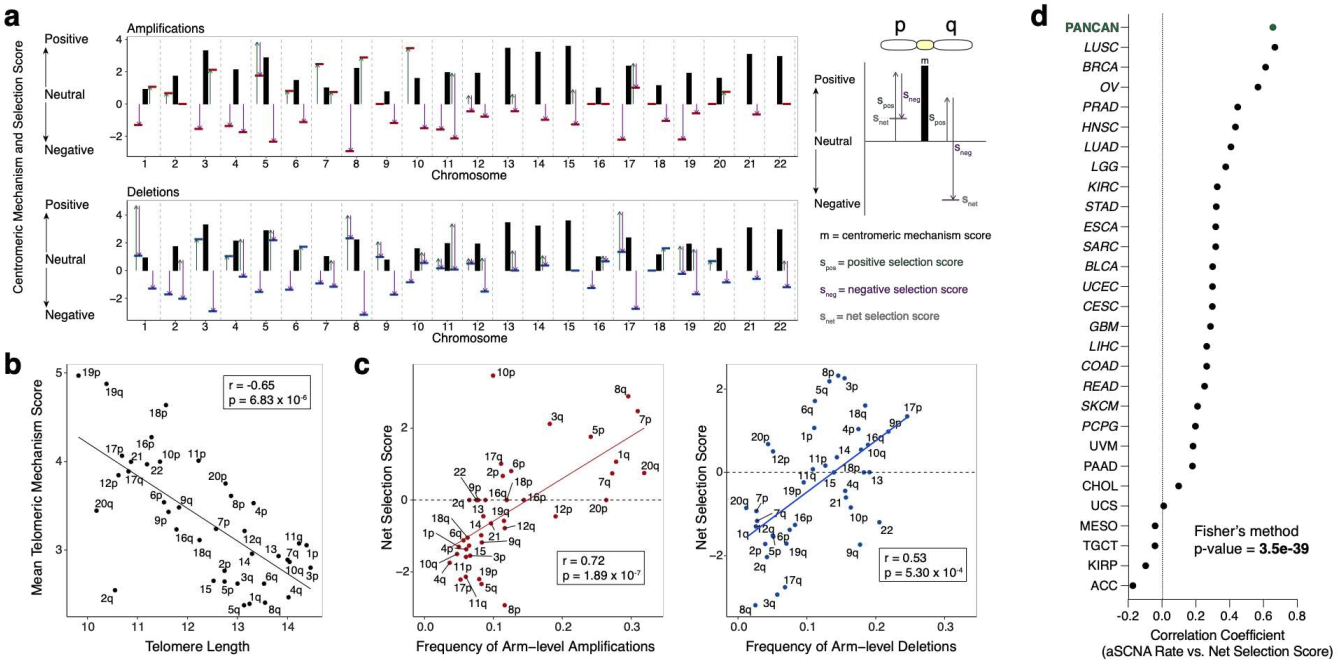
(d) Empirical examples of SCNA-mediated selection from the pan-cancer dataset; specific patterns are explored further in the main text.

Figure 3: BrISCUT identifies known cancer driver genes.



(a) Schematic representation of BrISCUT's peak-finding function. Tumors (in dark green) are ranked along the y-axis by pSCNA length. The location at which the empirical data deviates maximally from the background distribution is determined (in purple). A peak region encompassing this location (denoted by dashed lines) is calculated; see Methods.
(b) Statistically significant peaks conferring selection as determined by BrISCUT are plotted along the genome. Amplitude of peaks is depicted in significance score, or KS statistic * -log10(q-value). Positive selection peaks are in dark red (amplifications) and blue (deletions), and negative selection peaks are in light red (deletions) and blue (amplifications). Genes found in Tier 1 of the COSMIC Cancer Gene Census are bolded.
(c) Overlap between genes in BrISCUT peaks and Tier 1 COSMIC cancer genes. The numbers of peaks containing these genes are depicted in green.
(d) Cell viability (measured by median DEMETER2 score across 712 cell lines) is significantly lower when genes in negative deletion peaks ("essential-like") are knocked down by RNAi, compared to all other genes.

Figure 4: Pan-cancer mechanism and selection scores.



(a) Mechanism scores for each centromere and selection scores for amplifications and deletions of each chromosome arm. Black bars represent centromeric mechanism scores. Red and blue horizontal lines represent net selection scores for amplifications and deletions respectively, and are the sum of the amplitude of positive selection (green arrows, pointing up) and negative selection (purple arrows, pointing down). Selection scores for both p and q arms are depicted to the left and right of the centromeric mechanism score respectively.

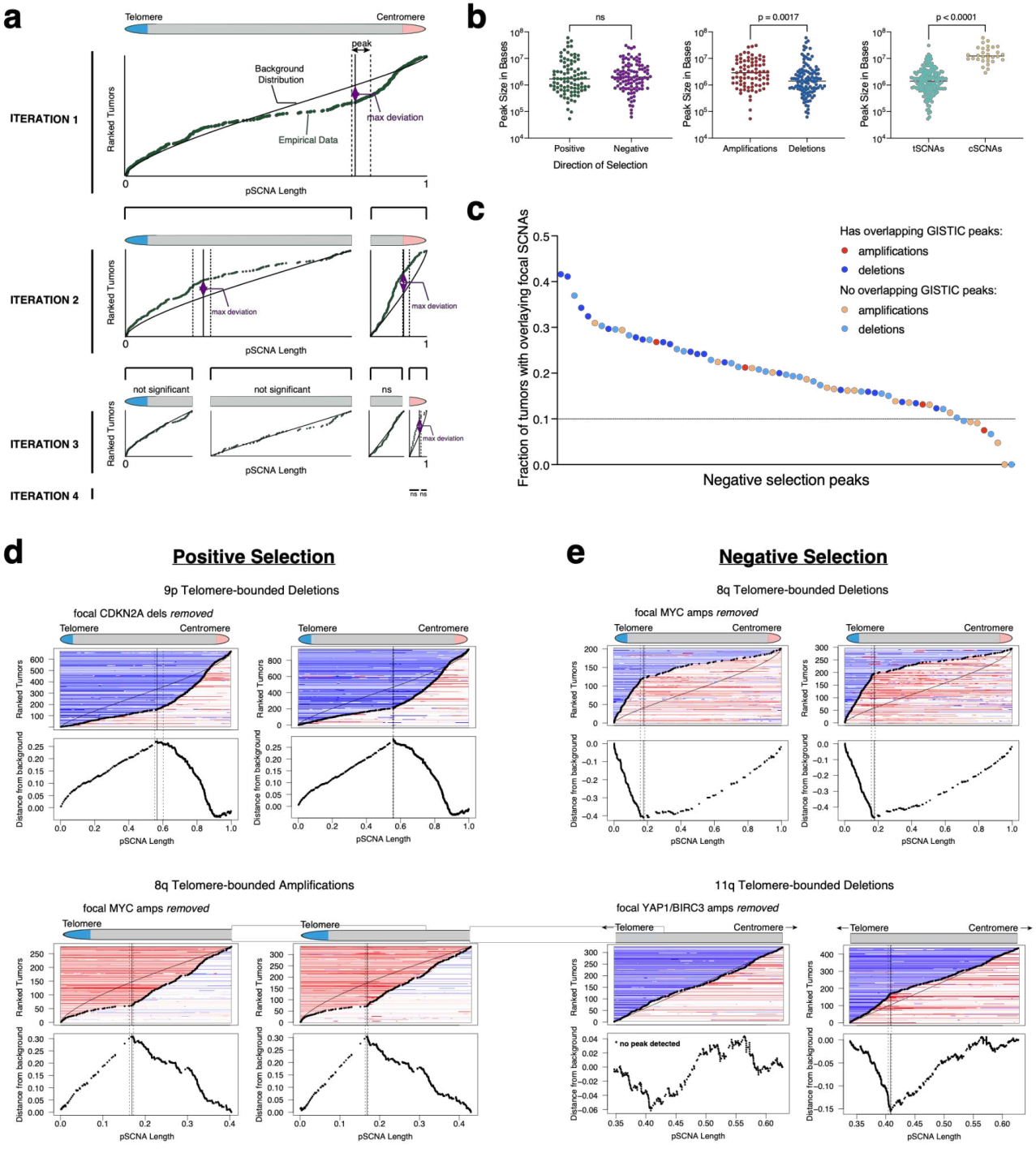
(b) Telomeric mechanism scores (averaged between amplifications and deletions) versus telomere length, in RTLU (Relative Telomere Length Units; a ratio of telomere signal to a reference signal within one genome[55].

(c) Net selection scores versus frequency of arm-level amplifications (left; in red) and deletions (right; in blue). Values above the dashed line represent net positive selection and values below the dashed line represent net negative selection.

(d) Spearman correlation coefficients for net selection score and aSCNA rate across pan-cancer (in green) and unique TCGA tumor types (in black; arranged from largest to smallest). Tumor types in italics have Spearman p-values < 0.1. Fisher's method p-value is calculated from unique TCGA types only.

EXTENDED DATA

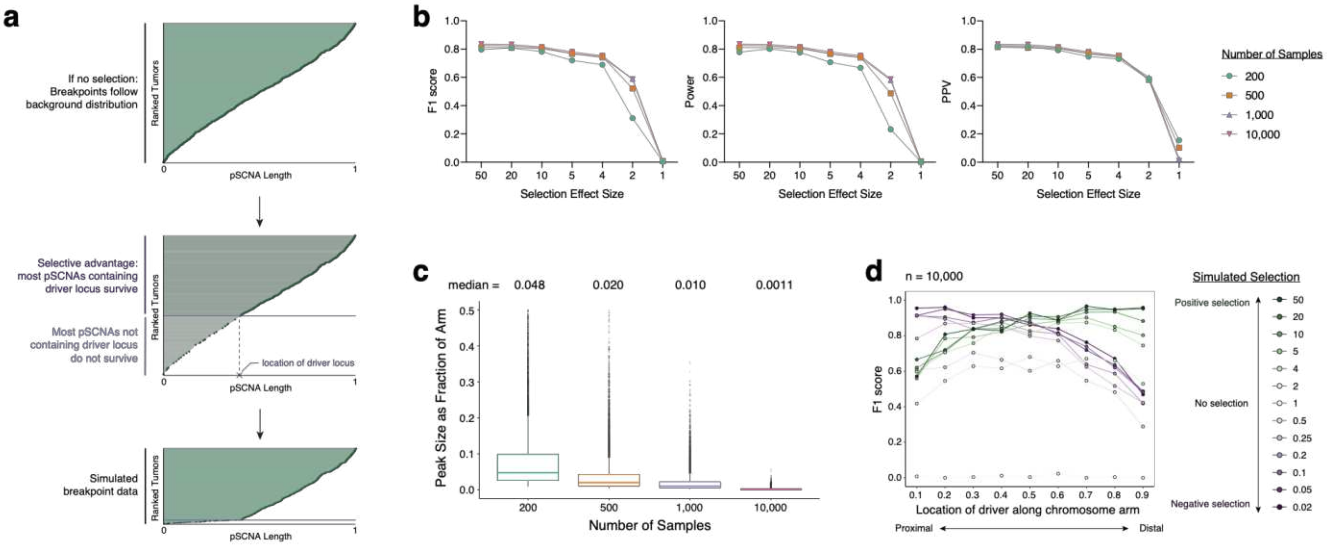
Extended Data Figure 1: Pan-cancer BrISCUT analysis



(a) Example depicting BrISCUT's recursion steps. From top to bottom: BrISCUT detects peaks iteratively, walking both left and right if a significant peak is detected, with the new boundaries including the detected peak. If a peak is not detected, overlaps with a previous peak, or there are fewer than 4 samples, the analysis is stopped. See **Methods** for details.

(b) Sizes of peaks (in bases) from the pan-cancer BrISCUT analysis. From left to right, peaks are categorized by direction of selection, direction of copy-number imbalance, and origin of pSCNA (telomere-bounded vs. centromere-bounded).
(c) Negative selection peaks from the pan-cancer BrISCUT analysis, sorted from highest to lowest by fraction of samples subject to this selective pressure that also possessed overlapping focal SCNAs in the opposite direction. Peaks that overlap with GISTIC 2.0 peaks are denoted in dark red and dark blue.
(d) BrISCUT analysis detecting two positive selection peaks (top: 9p telomere-bounded deletions, overlapping with *CDKN2A* focal deletions; bottom: 8q telomere-bounded amplifications, overlapping with *MYC* focal amplifications) with focal SCNAs removed (left) and with focal SCNAs included (right).
(e) BrISCUT analysis detecting two negative selection peaks (top: 8q telomere-bounded deletions, overlapping with *MYC* focal amplifications; bottom: 11q telomere-bounded deletions, overlapping with *YAP1/BIRC3* focal amplifications) with focal SCNAs removed (left) and with focal SCNAs included (right).

Extended Data Figure 2: BrISCUT power and accuracy analysis



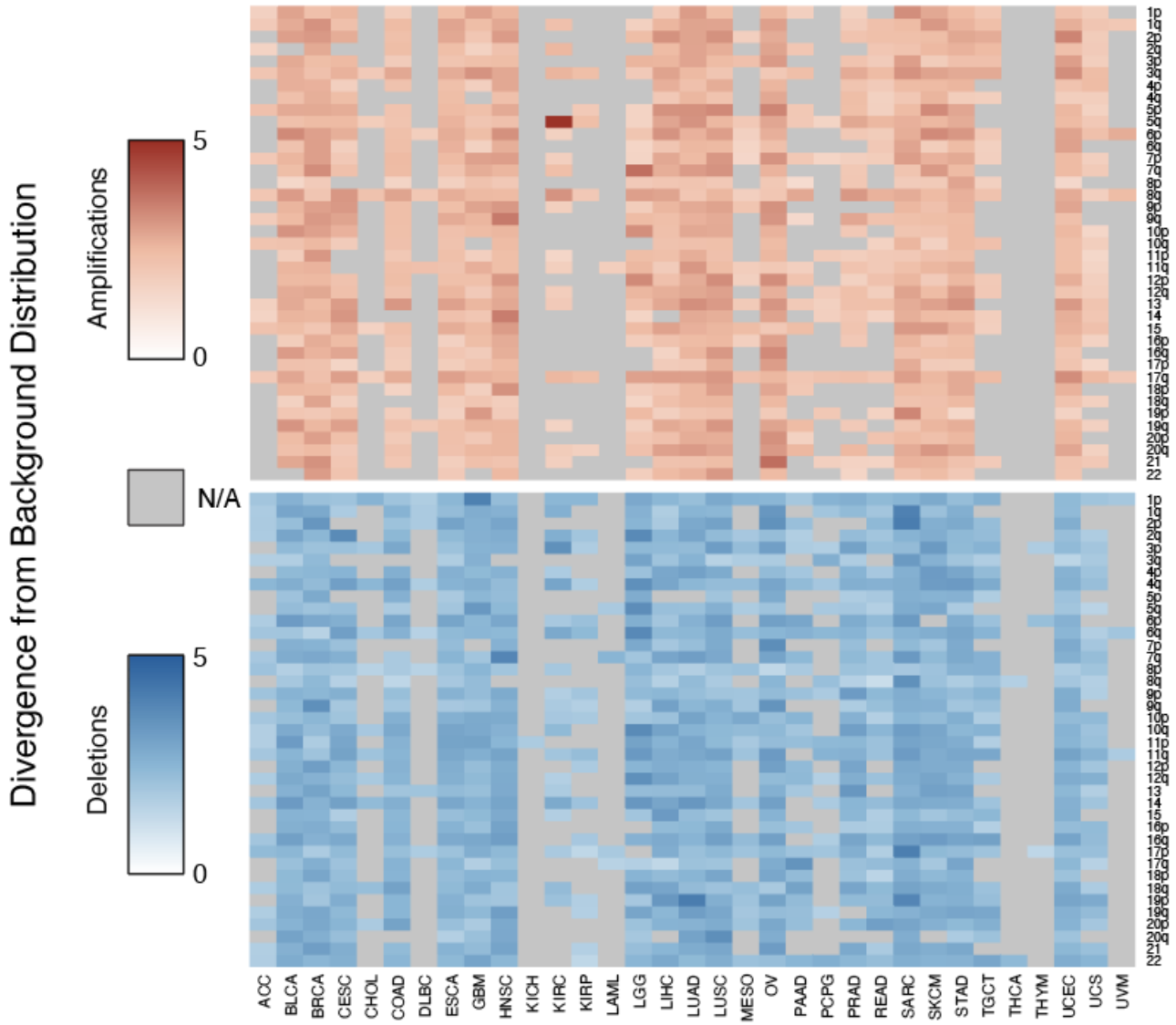
(a) Schematic of simulations for a single locus under selective pressure. Tumors (in dark green) are ranked along the y-axis by pSCNA length. The top panel depicts an example of pSCNAs under no selection, thereby following the background breakpoint distribution. The middle panel shows tumors that survive under a simulated locus conferring a 10:1 selective advantage (dashed purple vertical line). The tumors with pSCNAs encompassing the positive selection locus (i.e. "driver locus", above the solid purple line) are ten times more likely to survive than those that do not (below the line); this is reflected in the number of dark green dots remaining. The bottom panel shows only the surviving tumors.

(b) F1 score, power, and PPV (positive predictive value) of BrISCUT simulations for several selection effect sizes (x-axis). Each line represents the starting number of samples. Values are aggregated from 3,600 simulations each (100 for each combination of location [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9], positive vs. negative selection, and telomere-bounded vs. centromere-bounded).

(c) Distribution and median of peak sizes (as fraction of arm) across starting number of samples. Only statistically significant peaks that encompass the simulated "driver locus" and smaller than 0.5 are included. Box plots show the median, first quartile, and third quartile of peak size.

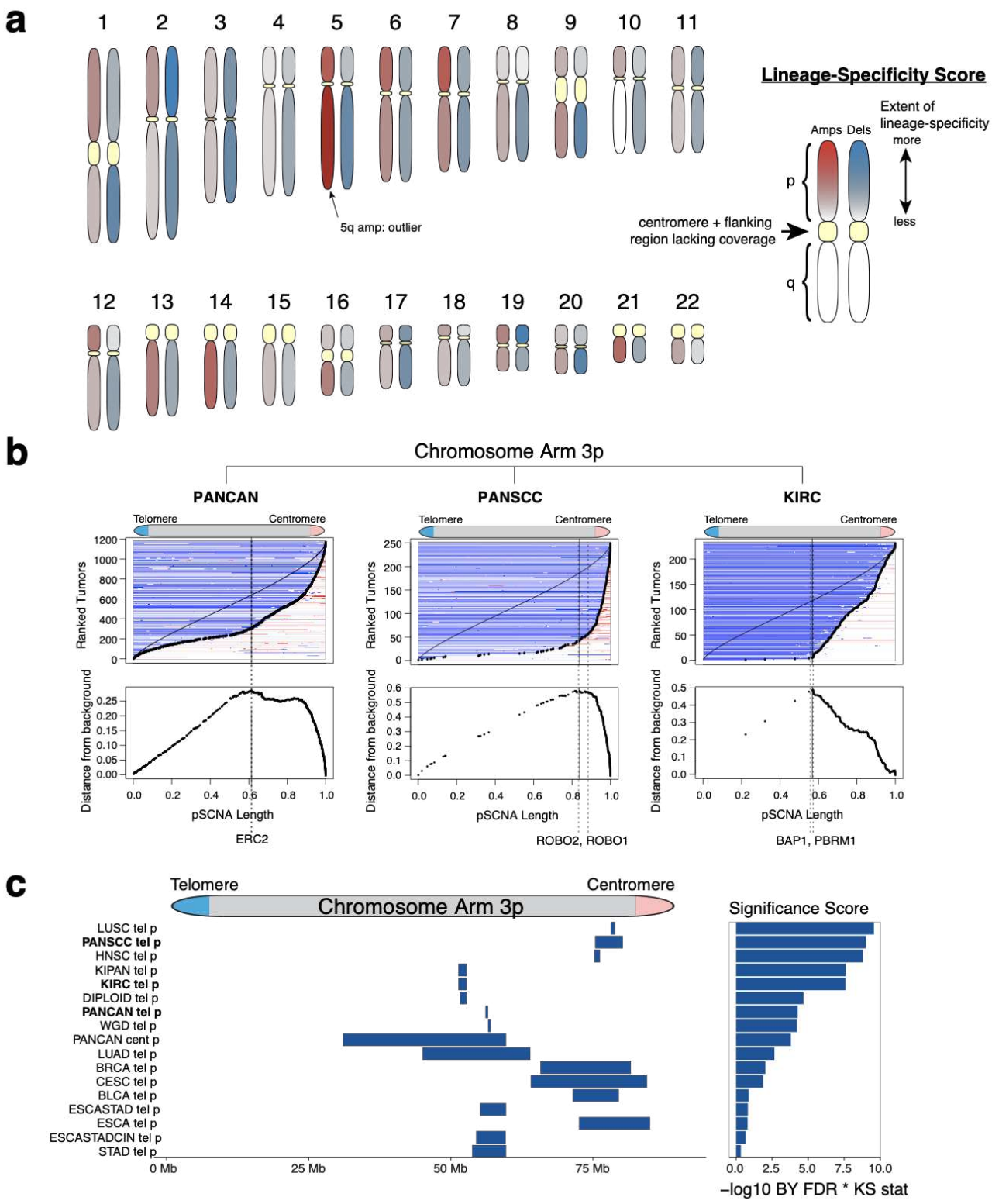
(d) BrISCUT F1 scores across locations of simulated driver loci, separated by effect size. Each line represents a different degree of positive selection (shades of green; darkest representing the highest effect size) or negative selection (shades of purple; darkest representing the highest effect size).

Extended Data Figure 3: Lineage-specific divergence of breakpoint distributions from the background distribution



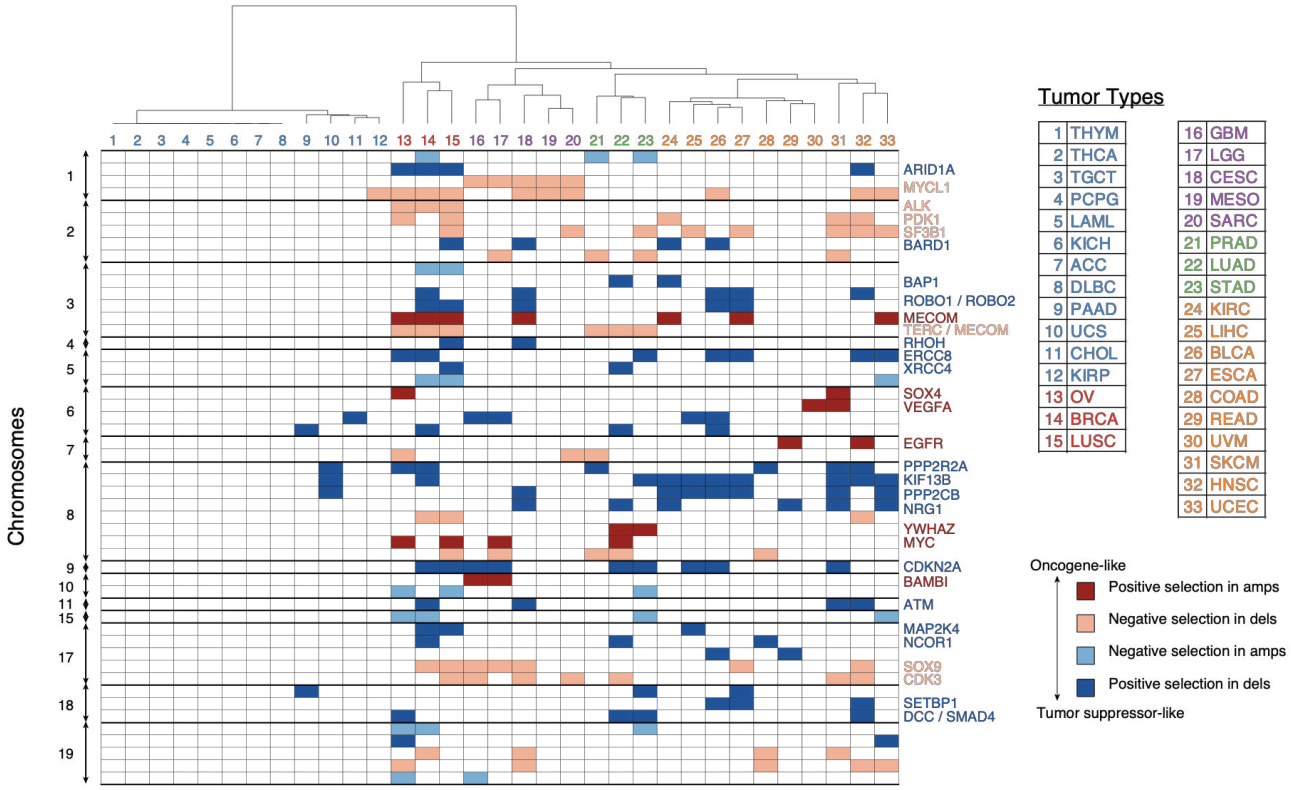
Heatmaps of lineage divergence scores for each tumor type (x-axis) and chromosome arm (y-axis). Amplifications are on top (in red) and deletions are on the bottom (in blue). Darker color represents higher divergence score.

Extended Data Figure 4: Patterns of chr3p deletions are highly lineage-specific



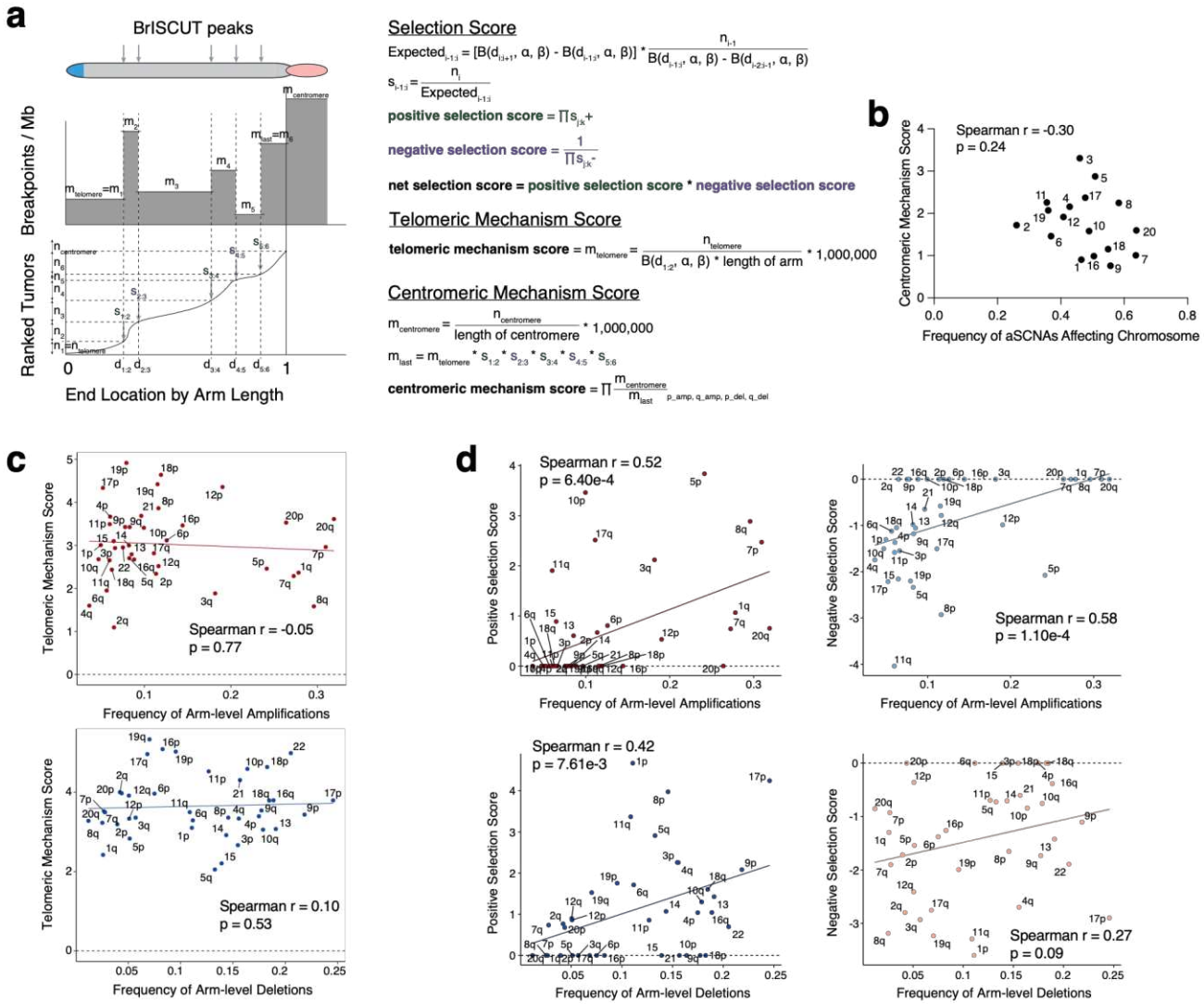
(a) Lineage-specificity scores across chromosomes. The left chromatid is shaded in red and represents amplifications, whereas the right chromatid is shaded in blue and represents deletions. Darker colors indicate greater lineage-specificity.
(b) BrISCUT analysis of telomere-bounded deletions on chr3p in three different cohorts. The top panels depict telomere-bounded deletions, sorted by length, with underlying copy number data visualized using IGV (the Integrative Genomics Viewer)[44]. The bottom panels show the vertical distance of each tSCNA from the background distribution; the maximum deviation is denoted by the solid vertical line. The dashed lines represent the peak regions determined to be under significant positive selection (i.e. conferring survival advantage in this cohort).
(c) Genomic locations and corresponding significance score of positive selection deletion BrISCUT peaks on chr 3p across lineages. See **Extended Data Table 2a** for tumor type abbreviations.

Extended Data Figure 5: Hierarchical clustering of BrISCUT peaks across lineages



Matrix of significantly recurring BrISCUT peaks across 33 independent tumor types. Peaks are sorted by genomic location (y-axis), and tumor types are sorted according to hierarchical clustering by Ward's method (x-axis). Different clusters (k = 5) are in different colors.

Extended Data Figure 6: Quantitative assessment of selective and mechanistic pressures driving aneuploidy



(a) Calculation of effect sizes ($s_{i:i+1}$) of specific BrISCUT peaks, selection scores and telomeric mechanism scores of each arm-level alteration, and centromeric mechanism scores of each chromosome. See **Methods** for further details.

(b) Centromeric mechanism scores plotted against total frequency of aSCNAs affecting a specific chromosome (i.e. amplifications and deletions of the p and q arms in aggregate). Acrocentric chromosomes are excluded from analysis.

(c) Telomeric mechanism scores plotted against frequency of arm-level amplifications (top; in red) and deletions (bottom; in blue).

(d) Selection scores plotted against frequency of aSCNAs. Positive selection scores are plotted on the left, with amplifications on top (in dark red) and deletions on the bottom (in dark blue). Negative selection scores are plotted on the right, with amplifications on top (in light blue) and deletions on the bottom (in light red).

Supplementary Note

Our understanding that chromosome arms are subject to cohort-specific patterns of selection allows for further interrogation into drivers of specific cancers. For example, both chromosome arms 3p and 8p are frequently deleted in cancer at comparable rates (15.5% and 14.6% whole-arm deletions respectively, with an additional 14.0% and 20.8% partial deletions). However, 3p deletions are significantly more lineage-specific than 8p (lineage-specificity scores = 0.26 and 0.10 respectively; F-test p = 0.01). Chromosome arm 3p (chr3p) is frequently deleted in tumors of squamous origin and in renal clear cell carcinomas[45–48]. However, breast cancer and several squamous lineages (pan-squamous, CESC, ESCA, HNSC, LUSC, and BLCA) share a peak on 3p12.3 containing *ZNF717*, *ROBO1*, and *ROBO2*, while kidney cancers and lung adenocarcinoma share a peak on 3p21.2 containing the renal cancer tumor suppressors *BAP1* and *PBRM1* (**Extended Data Figure 3b**)[49]. Germline mutations of *BAP1* cause a tumor predisposition syndrome associated predominantly with uveal and cutaneous melanoma, mesothelioma, and renal cell carcinoma, but have not been related to squamous tumors [50]

Although chr8p has also been the target of extensive genomic and functional study, no strong evidence has been obtained for any single candidate tumor suppressor gene driving loss of the arm[51]. Indeed, chr8p loss has not been shown to contribute to transformation *in vitro*[11]. In our pan-cancer analysis, we detect multiple BrISCUT peaks on this arm, consistent with multiple genes on 8p contributing to its loss[52]. There are several notable lineage-specific peaks, including 8p21.2, which contains putative tumor suppressor *BNIP3L* and is adjacent to *PPP2R2A* (in OV, UCEC, UCS, HNSC, PRAD, and SKCM) and three separate regions located on 8p12: *NRG1*, which encodes ligands that bind to *ERBB3* and *ERBB4* (most strongly represented in COADREAD and LUAD), *KIF13B* and *HMBOX1* (most strongly represented in BRCA), and *PPP2CB* (most strongly represented in pan-squamous and LUSC). However, most peaks occur within 5 Mb of each other and often overlap, consistent with the possibility that these reflect strong multifocal positive selection for partial deletions of chr8p and little variation in these loci across tumor types. In contrast, we do observe stronger evidence for lineage-specificity for deletion negative selection peaks in 8p11.21, which contains *ANK1* and *KAT6A*, in the COADREAD and OV analyses, and 8p11.23, which contains *NSD3* (otherwise known as *WHSC1L1*), in the pan-squamous and BRCA cohorts. Both loci are also focally amplified in these respective cohorts, but the deletion negative selection peaks persist even after removing samples with focal amplifications of these loci.

To compare patterns of pSCNA-driven selection across different lineages, we also performed hierarchical clustering of significantly recurring BrISCUT peaks across the 33 independent tumor types (**Extended Data Figure 5**). We observed four major groups. One was characterized by very few peaks, either due to lack of pSCNA-driven selection or lack of power to detect events. Tumors of squamous morphology (cervical, lung squamous, and head and neck squamous), ovarian carcinomas, and sarcomas clustered together in a second group, largely due to negative selection deletion peaks on 1q34.2, containing *MYCL1*, and on 17q25.1, containing the cyclin-dependent kinase *CDK3*. A third cluster, comprising only glioblastomas and low grade gliomas, was driven by negative selection for 3q27.1 deletions, containing *SOX2*, a transcription factor in the TGF-β pathway whose expression is essential for retention of stemness in glioma-initiating cells[53], and positive selection for 10p12.1 amplification, containing *BAMBI*, which encodes a pseudoreceptor for type I TGF-β receptors that is highly expressed in several tumor types and whose overexpression been shown to decrease tumor responsiveness to TGF-β signaling[54]. The final group contained adenocarcinomas, including lung adenocarcinoma, prostate, stomach, and breast cancers, as well as melanomas, but was not particularly enriched for any single event.

Although some tumor types of shared cell-of-origin did not necessarily cluster together, BrISCUT identified several peaks unique to lineages arising from certain tissues. Examples included negative selection of *SOX4* amplification in ovarian and breast cancers, positive selection of *MATK* (19p13.3) deletion in ovarian and endometrial cancers, positive selection of *VEGFA* (6p21.1) amplification in uveal melanoma and skin melanoma, and positive selection of 5q13.3 deletion in lung adenocarcinoma and lung squamous cell carcinoma. In contrast, some peaks, such as a negative selection amplification peak on 2q33.1 containing *SF3B1* or a positive selection deletion peak on 9p21.3 containing *CDKN2A*, were frequently the only loci under selection on a given chromosome arm, but were prevalent across lineages within multiple clusters.

EXTENDED DATA TABLE LEGENDS

Extended Data Table 1
Most common somatic alterations in cancer

Extended Data Table 2
Cohort information and statistics of different types of SCNAs

Extended Data Table 3
Pan-cancer regions of selection

Extended Data Table 4
Cohort-specific regions of selection and lineage-specificity analysis

Extended Data Table 5
Selection and mechanism scores

DATA AVAILABILITY

All SNP array data used for analysis are publicly available from The Cancer Genome Atlas' Genomic Data Commons Data Portal at https://portal.gdc.cancer.gov/. All data generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

CODE AVAILABILITY

The code used to merge copy-number segments, call pSCNAs, detect loci under selection, and determine selection and mechanism scores are freely available for download at https://github.com/beroukhim-lab/BrISCUT.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the Beroukhim, Taylor, and Meyerson laboratories for their support and helpful discussions, as well as Chip Stewart, Cheng-Zhong Zhang, and Steve Schumacher. We would also like to acknowledge the following funding sources: the National Institutes of Health and National Cancer Institute (RB, AMT), Fund for Innovative Cancer Informatics (RB), the Gray Matters Brain Cancer Foundation (RB), Pediatric Brain Tumor Foundation (RB), Brain Tumour Charity (RB), and the St. Baldrick's Foundation (RB).

AUTHOR CONTRIBUTIONS

J.S., A.D.C., A.M.T., and R.B. conceived of this study. J.S., G.F.G., L.F.S., A.C.B., A.D.C., and R.B. developed methods. J.S. implemented methods and performed computational analyses. G.H., V.R., and M.M. provided feedback and advice on analyses. A.D.C., A.M.T., and R.B. supervised the work. J.S., A.M.T., and R.B. wrote the manuscript.

CORRESPONDING AUTHOR

Correspondence to Alison M. Taylor and Rameen Beroukhim.

COMPETING INTERESTS

Galen F. Gao, Ashton C. Berger, Andrew D. Cherniack, and Matthew Meyerson receive or received research support from Bayer AG. Matthew Meyerson and Alison M. Taylor receive research support from Ono Pharmaceutical. Matthew Meyerson is an equity holder, consultant for, and Scientific Advisory Board chair for OrigiMed. Matthew Meyerson additionally receives research support from Novo Nordisk and Janssen Pharmaceuticals, consults for Interline Therapeutics, and is an inventor of a patent for EGFR mutation diagnosis in lung cancer, licensed to Labcorp. Rameen Beroukhim consults for and owns equity in Scorpion Therapeutics and receives research support from Merck & Co. and Novartis.

REFERENCES

1. Boveri, T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J. Cell Sci.* **121 Suppl 1**, 1–84 (2008).

2. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

3. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

4. Tang, Y.-C. & Amon, A. Gene copy-number alterations: a cost-benefit analysis. *Cell* **152**, 394–405 (2013).

5. 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).

6. Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010).

7. Holland, A. J. & Cleveland, D. W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat. Rev. Mol. Cell Biol.* **10**, 478–487 (2009).

8. Sheltzer, J. M. *et al.* Single-chromosome Gains Commonly Function as Tumor Suppressors. *Cancer Cell* **31**, 240–255 (2017).

9. Weaver, B. A. & Cleveland, D. W. Does aneuploidy cause cancer? *Curr. Opin. Cell Biol.* **18**, 658–667 (2006).

10. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676–689 e3 (2018).

11. Cai, Y. *et al.* Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. *Cancer Cell* **29**, 751–766 (2016).

12. Uno, N. *et al.* CRISPR/Cas9-induced transgene insertion and telomere-associated truncation of a single human chromosome for chromosome engineering in CHO and A9 cells. *Sci. Rep.* **7**, 12739

(2017).

13. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 20007–20012 (2007).

14. Rouveirol, C. *et al.* Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics* **22**, 849–856 (2006).

15. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).

16. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

17. Cimini, D. Merotelic kinetochore orientation, aneuploidy, and cancer. *Biochim. Biophys. Acta* **1786**, 32–40 (2008).

18. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

19. Gao, G. F. *et al.* Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst* **9**, 24–34 e10 (2019).

20. Gao, G. & Smith, D. I. Role of the Common Fragile Sites in Cancers with a Human Papillomavirus Etiology. *Cytogenet. Genome Res.* **150**, 217–226 (2016).

21. McFarland, J. M. *et al.* Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.* **9**, 4610 (2018).

22. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041 e21 (2017).

23. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304 e6 (2018).

24. Barra, V. & Fachinetti, D. The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nat. Commun.* **9**, 4340 (2018).

25. Knutsen, T. *et al.* Definitive molecular cytogenetic characterization of 15 colorectal cancer cell lines. *Genes Chromosomes Cancer* **49**, 204–223 (2010).

26. Lee, C., Wevrick, R., Fisher, R. B., Ferguson-Smith, M. A. & Lin, C. C. Human centromeric DNAs. *Hum. Genet.* **100**, 291–304 (1997).

27. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744–757.e24 (2018).

28. Kitagawa, K. & Hieter, P. Evolutionary conservation between budding yeast and human kinetochores. *Nat. Rev. Mol. Cell Biol.* **2**, 678–687 (2001).

29. Gregan, J., Polakova, S., Zhang, L., Tolić-Nørrelykke, I. M. & Cimini, D. Merotelic kinetochore attachment: causes and effects. *Trends Cell Biol.* **21**, 374–381 (2011).

30. Gisselsson, D. *et al.* Telomere dysfunction triggers extensive DNA fragmentation and evolution of complex chromosome abnormalities in human malignant tumors. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 12683–12688 (2001).

31. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).

32. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).

33. Cancer Genome Atlas Research, Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).

34. Tabak, B. *et al.* The Tangent copy-number inference pipeline for cancer genome analyses. *bioRxiv* 566505 (2019).

35. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).

36. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).

37. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385 e18 (2018).

38. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* **23**, 227–238 e3 (2018).

39. Saghafinia, S., Mina, M., Riggi, N., Hanahan, D. & Ciriello, G. Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. *Cell Rep.* **25**, 1066–1080 e8 (2018).

40. Delignette-Muller, M. L. & Dutang, C. fitdistrplus: An R Package for Fitting Distributions. *2015* **64**, 34 (2015).

41. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).

42. Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **49**, 1858–1860 (2003).

43. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).

44. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

45. Zabarovsky, E. R., Lerman, M. I. & Minna, J. D. Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene* **21**, 6915–6935 (2002).

46. Bass, A. J. *et al.* SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nat. Genet.* **41**, 1238–1242 (2009).

47. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).

48. Cancer Genome Atlas Research, Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).

49. Ricketts, C. J. *et al.* The Cancer Genome Atlas Comprehensive Molecular Characterization of

Renal Cell Carcinoma. *Cell Rep.* **23**, 313–326 e5 (2018).

50. Masoomian, B., Shields, J. A. & Shields, C. L. Overview of BAP1 cancer predisposition syndrome and the relationship to uveal melanoma. *J Curr Ophthalmol* **30**, 102–109 (2018).

51. Tabares-Seisdedos, R. & Rubenstein, J. L. Chromosome 8p as a potential hub for developmental neuropsychiatric disorders: implications for schizophrenia, autism and cancer. *Mol. Psychiatry* **14**, 563–589 (2009).

52. Xue, W. *et al.* A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 8212–8217 (2012).

53. Ikushima, H. *et al.* Autocrine TGF-beta signaling maintains tumorigenicity of glioma-initiating cells through Sry-related HMG-box factors. *Cell Stem Cell* **5**, 504–514 (2009).

54. Sekiya, T. *et al.* Identification of BMP and activin membrane-bound inhibitor (BAMBI), an inhibitor of transforming growth factor-beta signaling, as a target of the beta-catenin pathway in colorectal tumor cells. *J. Biol. Chem.* **279**, 6840–6846 (2004).

55. Wise, J. L. *et al.* Human telomere length correlates to the size of the associated chromosome arm. *PLoS One* **4**, e6013 (2009).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- t1210507.xlsx
- t2210507.xlsx
- t3210510.xlsx
- t4210510.xlsx
- t5210510.xlsx
- s1210519.pdf
- s2210519.pdf
- s3210519.pdf
- s4210519.pdf
- s5210520.pdf
- s6210510.pdf