Published in final edited form as:

Nature. 2023 July ; 619(7971): 793–800. doi:10.1038/s41586-023-06266-3.

Cancer aneuploidies are shaped primarily by effects on tumor fitness

Juliann Shih[1,2,3,4], **Shahab Sarmashghi**[1,2,5], **Nadja Zhakula-Kostadinova**[6,7], **Shu Zhang**[1,2], **Yohanna Georgis**[6], **Stephanie H. Hoyt**[1,8], **Michael S. Cuoco**[8], **Galen F. Gao**[1,2], **Liam F. Spurr**[1,2,8], **Ashton C. Berger**[1,8], **Gavin Ha**[1,8,9], **Veronica Rendo**[1,2,5], **Hui Shen**[10], **Matthew Meyerson**[1,8,11], **Andrew D. Cherniack**[1,5,8], **Alison M. Taylor**[1,6,8,*], **Rameen Beroukhim**[1,2,5,8,*]

[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]Tufts University School of Medicine, Boston, MA, USA

[4]Department of Internal Medicine, Kirk Kerkorian School of Medicine at UNLV, Las Vegas, NV, USA

[5]Department of Medicine, Harvard Medical School, Boston, MA, USA

[6]Department of Pathology and Cell Biology, Herbert Irving Comprehensive Cancer Center, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

[7]Department of Genetics and Development, Columbia University Vagelos College of Physicians and Surgeons, New York, NY, USA

[8]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

[9]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[10]Department of Epigenetics, Van Andel Institute, Grand Rapids, MI, USA

[11]Department of Genetics, Harvard Medical School, Boston, MA, USA

Correspondence: Alison M. Taylor, at3488@cumc.columbia.edu; Rameen Beroukhim, rameen_beroukhim@dfci.harvard.edu.

[*]These authors contributed equally to this work.

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 analytic methods. J.S. and S.S. implemented analytic methods. J.S., S.S., S.Z., S.H.H., N.Z.K., Y.G., H.S., and A.M.T. performed computational analyses. N.Z.K., Y.G., M.S.C., and A.M.T. developed and performed *in vitro* experiments. 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., N.Z.K, S.S., S.Z., A.M.T., and R.B. wrote the manuscript with input from all coauthors.

COMPETING INTERESTS

G.F.G., A.C.B, A.D.C., and M.M. receive or received research support from Bayer AG. M.M. and A.M.T. received research support from Ono Pharmaceutical. M.M. is an equity holder, consultant for, and Scientific Advisory Board chair for OrigiMed. M.M. 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. R.B. consults for and owns equity in Scorpion Therapeutics and receives research support from Novartis. J.S., S.S., S.Z., N.Z., Y.G., S.H.H., M.S.C., L.F.S., G.H., V.R., and H.S. declare no competing interests.

ADDITIONAL INFORMATION

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Rameen Beroukhim (rameen_beroukhim@dfci.harvard.edu) and Alison M. Taylor (at3488@cumc.columbia.edu). Reprints and permissions information is available at www.nature.com/reprints.

Abstract

Aneuploidies—whole-chromosome or whole-arm imbalances—are the most prevalent alteration in cancer genomes[1,2]. However, it is still debated whether their prevalence is due to selection or because they are readily generated passenger events[1,2]. We developed a method, BISCUT, that identifies loci subject to fitness advantages or disadvantages by interrogating length distributions of telomere- or centromere-bounded copy-number events. 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. BISCUT identified the helicase *WRN* as a haploinsufficient tumor suppressor gene on chromosome 8p, which is supported by several lines of evidence. We also formally quantified the role of selection and mechanical biases in driving aneuploidy, finding that rates of arm-level copy-number alterations are most highly correlated with their effects on cellular fitness[1,2]. These results provide insight into the driving forces behind aneuploidy and its contribution to tumorigenesis.

INTRODUCTION

Although aneuploidy, which we define as whole-chromosome or whole-arm DNA imbalance, is observed in ~90% of tumors[1,2] and was the first proposed somatic alteration in cancer[3], the reasons for its prevalence and role in driving cancer remain unclear. Its prevalence may reflect frequent chromosome missegregation, rearrangements, or centrosome aberrations (mechanical biases)[4], or fitness advantages associated with aneuploidy (selection biases). However, in conditions of low cellular stress, 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[5]. In yeast and human cells, aneuploidy has been found to be beneficial in the context of some types of cellular stress or gene deficiencies[6–12], and the general state of aneuploidy has been found to contribute to tumor evolution[13,14] and drug resistance[15]. However, we do not have a comprehensive understanding of the positive or negative effects of individual arm-level somatic copy-number alterations (arm-SCNAs) on fitness in the natural context of human tumors. Experimental methods to assess functional consequences of arm-SCNAs are technically challenging and have rarely been performed in human cells[2,16,17]. In the context of focal SCNAs, mapping minimal common regions of amplification or deletion can point to relevant oncogenes and tumor suppressor genes[18]. Arm-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 arm-SCNAs, except for the small region that they lack immediately adjacent to the centromere (or telomere). Similarly, slightly shorter SCNAs would also be expected to have the same fitness effects, except for the small region they lack – and so on. In this fashion, one may explore fitness effects of each successive portion of the chromosome arm on the telomeric (or centromeric) alterations that encompass it[9]. The effects of an entire arm-SCNA can therefore be inferred as the sum of effects across all regions it encompasses. In this study, we develop an algorithm called BISCUT (Breakpoint Identification of Significant Cancer

Undiscovered Targets) that exploits this source of information – the length distributions of telomere- and centromere-bounded SCNAs – to better understand the effects of arm-SCNAs 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 selective and mechanical biases on patterns of chromosome arm aneuploidies within and across cancers.

RESULTS

Impact of arm-level SCNAs

SCNAs that extend from telomere to centromere (arm-SCNAs) are among the most frequent somatic genetic alterations in cancer. Across 10,872 TCGA tumors spanning 33 cancer types[2], arm-SCNAs constitute 23 of the 25 most frequent events (Figure 1a). Arm-SCNAs also encompass more of the cancer genome by far – 22.5% – than any other type of somatic genetic alteration (Figure 1b and Supplementary Table 1). In contrast, focal SCNAs, which affect the next largest fraction of the genome, encompass only 11.3% of the cancer genome, or half that of arm-SCNAs.

Despite their impact on the cancer genome, it is uncertain whether the frequencies of different arm-SCNAs are primarily determined by their mechanical ease of generation or by their effects on evolutionary fitness. We also do not know which loci contribute to the effects of arm-SCNAs on evolutionary fitness. Among the 23 most frequent arm-SCNAs, 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 arm-SCNA (Supplementary Table 1a). Even among cases with such drivers, these drivers do not always explain the observed frequencies of arm-SCNAs. For example, although 10q is lost in ~80% of glioblastomas, the presumed target of these losses – the tumor suppressor *PTEN* – is only biallelically inactivated by homozygous deletion or mutation in ~40% of cases[19]. Glioblastomas without biallelic inactivation express *PTEN* at similar levels to those without 10q loss. Epigenetic alterations could provide more information as to drivers that contribute to the frequencies of these arm-SCNAs, but current tools are insufficient to fully address this question.

Centromeric breaks are favored overall

We thus set out to quantify the extent to which arm-SCNA frequencies can be attributed to their mechanical ease of generation versus effects on cell fitness. We approached this question by examining the locations of the breakpoints of SCNAs that begin at either the telomere or centromere (tel-SCNAs and cent-SCNAs, respectively, collectively termed partial-SCNAs; Figure 1c). We hypothesized that the enrichment of breakpoints in specific loci might provide insight on the roles that mechanical biases and selection play, and to the specific genetic elements undergoing selection.

We first considered the relative densities of breakpoints within centromeres versus within chromosome arms to indicate mechanical biases favoring or disfavoring arm-SCNAs (Figure 1d and Extended Data Figure 1a). Across all chromosome arms and tumor types, 39% of tel-SCNAs end in centromeres – a four-fold enrichment of breakpoint density in centromeres relative to within chromosome arms (86.5 and 19.7 per Mb, respectively; Figure 1e and Extended Data Figure 1b). Reasons for this enrichment may include the role of the kinetochore in mitosis and defects in mitotic checkpoint signaling, cohesion, merotelic attachment, or toxicity of acentric DNA to cells[20,21]. The frequency of centromeric breakpoints appears to be unrelated to the length of the centromere (Figure 1f). We explore differences in rates of centromeric breakage across chromosomes below.

SCNA lengths inform fitness effects

Within chromosome arms, we evaluated partial-SCNAs (tel-SCNAs and cent-SCNAs) as sources of information about fitness effects of arm-SCNAs (Supplementary Table 2). Like arm-SCNAs, partial-SCNAs tended to have lower amplitudes (number of copies gained or deleted) than interstitial SCNAs ($p < 2.2e\text{-}16$; Figure 1g), indicating the effects of partial-SCNAs on loci they encompass are similar to those of arm-SCNAs. When summed across all chromosome arms and cancer types, partial-SCNAs followed near-uniform length distributions (Figure 1h and Extended Data Figure 1c). We consider these to be “background” partial-SCNA distributions in the absence of fitness effects (see also Supplementary Note 1).

Next, we compared chromosome arm-specific partial-SCNA length distributions to these background distributions to detect genomic loci subject to selection. Specifically, we hypothesized that more partial-SCNAs would encompass a locus if its alteration increased cellular fitness (i.e. positively selected “driver” events), and fewer partial-SCNAs would encompass a locus if its alteration decreased fitness. We would therefore observe a sudden jump or fall in partial-SCNA breakpoint frequencies adjacent to loci under selection. Indeed, when comparing the near-uniform background model of tel-SCNA lengths with tel-SCNAs from individual chromosome arms across cancer types, 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 selection; and 4) loci that deviate in opposite directions from the background, indicating balanced selection (Figure 2a–b).

We therefore developed a method (“BISCUT”; see Methods for a detailed description; Figure 2c) based on these principles. BISCUT first determines whether the distribution of partial-SCNA lengths on a given chromosome arm differs significantly from the empirical background distribution. If yes, BISCUT identifies the genomic locus at which the arm-specific and background distributions diverge most and sets boundaries for a “peak region” that would be expected to encompass genes driving this divergence. Once this locus is identified, the chromosome arm is divided at the locus and BISCUT is repeated on both its telomeric and centromeric side (Extended Data Figure 1d). This process is repeated until no significant divergence between the observed and expected data is detected. Detailed results of simulations to test BISCUT’s power to detect loci under selection and its robustness to artifact are in Supplementary Figure 1 and Supplementary Note 2.

Loci under selection

When applied to the 10,872 cancer copy-number profiles generated by TCGA across 33 cancer types, BISCUT detected 193 genomic loci under apparent selection: 90 regions of positive selection (39 from amplifications [amp-pos] and 51 from deletions [del-pos]) and 103 regions of negative selection (41 from amplifications [amp-neg] and 62 from deletions [del-neg]) (Figure 2d and Extended Data Figure 2a–b). These peak loci were significantly enriched for known oncogenes and tumor suppressors from the COSMIC Cancer Gene Census ($p < 3.6e-5$; Extended Data Figure 2c and Supplementary Table 3a–b). This finding held in both genes subject to positive ($p < 9.5e-5$) and negative selection ($p = 0.015$). Intriguingly, many of the COSMIC genes detected by BISCUT are not targets of focal SCNAs, possibly due to combinatorial effects of large SCNAs evaluated by BISCUT and possibly because BISCUT is more sensitive to selection acting on low-amplitude amplifications and deletions (Extended Data Figure 2d–f, Supplementary Note 3, and Supplementary Table 3c–f).

BISCUT results aligned with fitness estimates from normalized ratios of non-synonymous to synonymous mutations (dN/dS ratios)[22]. Genes in del-neg peaks tended to have low dN/dS scores (indicating negative selection) for all truncating mutation types (nonsense and splice site), whereas genes in del-pos peaks had high dN/dS scores (indicating positive selection) for all mutation types (Figure 3a). We conclude that BISCUT peaks are enriched for cancer drivers and provide valuable information beyond a focal SCNA analysis.

We next asked whether analyses specific to genetic or tissue contexts would uncover additional loci under selection. Among genetic contexts, we found little evidence that arm-SCNAs of one chromosome arm altered fitness effects of partial-SCNAs of other arms (Supplementary Note 4). Lineage-specific analyses of the 33 TCGA tumor types each indicated far fewer significant peaks than the pan-cancer analysis (median = 9), ranging from none (in DLBC, KICH, LAML, THCA, and THYM) to 59 (in OV). Altogether, we detected 397 peaks across the 33 lineages. We also analyzed combined groups of shared lineage (COADREAD, ESCASTAD, GBMLGG, KIPAN, and PANSCC) and notable sub-lineages (BRCA-basal, BRCA-luminal, ESCASTAD-CIN, and ESCASTAD-GS) (Supplementary Table 4a–d). Across all groups, we detected a total of 609 peaks. Among peaks in independent lineages, 331 peaks, or 83%, overlapped with at least one peak in another lineage, leaving 66 distinct non-overlapping peaks across all lineages (excluding the pan-cancer analysis; Supplementary Table 4e–f). Among independent cohorts, overlapping peaks occurred more often among related developmental lineages[23] than expected by chance ($p = 0.001$; Extended Data Figures 3–5, Supplementary Note 4, and Supplementary Table 4g–h), mirroring the association between arm-SCNA rates and developmental lineage[2,24].

BISCUT results appeared relatively robust to slight modifications to the method (Supplementary Note 5 and Supplementary Table S4i–k) and are reproducible in other datasets. However, we expect that tumor impurity can limit power to detect loci under selection. We also detected a lower magnitude of selection against deletions in WGD samples (see below). To assess the reproducibility of our results, we also applied BISCUT to an entirely separate cohort and data generation platform: 1,765 tumors that had undergone whole-genome sequencing within the International Cancer Genome Consortium (ICGC)[25] (Supplementary Note 5 and Supplementary Table S4l–m). Among 55 peaks detected in this analysis, 36 (65%) overlapped with at least one of the 185 most significant original peaks ($p = 4.2e\text{-}8$). The additional 19 peak loci may reflect differences in tumor types represented in the two cohorts.

Validation of positive selection peaks

To validate genes in BISCUT peaks without known oncogenes or tumor suppressors, we used immortalized lung epithelial cells in which we isogenically engineered an arm-level loss of 8p[2] (Extended Data Figure 6a–b). Chromosome arm 8p is frequently deleted across cancer types, but canonical tumor suppressors have not been detected on 8p[16,26]. We observed less cell death by caspase activity ($p = 0.02$) and flow cytometry ($p = 0.004$) in cells with engineered 8p deletion (Extended Data Figure 6c–d and Supplementary Figure 2). Of the three 8p del-pos peaks (Figure 2d), the smallest peak, in 8p12, contained two protein-coding genes, *WRN* and *NRG1*. We performed RNA sequencing of clones with or without 8p deletion (Supplementary Table 5a) and found that expression of *WRN* is significantly lower in both cells with engineered 8p deletion and TCGA tumors with 8p deletion ($p = 0.014$ and 2.2e-16 respectively). However, *NRG1* expression was low both *in vitro* and in TCGA tumors, with little correlation between its expression and copy number (Extended Data Figure 6e). We therefore focused on *WRN*.

WRN encodes a RecQ DNA helicase involved in DNA damage repair, and its inactivation is synthetic lethal with microsatellite instability (MSI)[27]. To assess the functional relevance of non-homozygous loss of *WRN* due to 8p deletions, we asked whether these losses were negatively associated with MSI and associated with any mutational signatures of their own. Across TCGA, about 1% of tumors have MSI[28]. None of these tumors harbor *WRN* copy-loss or arm-level deletion of 8p – a significant anticorrelation even after controlling for lineage and overall aneuploidy[2,29] ($p < 1e\text{-}4$ for both *WRN* and 8p loss; Figure 3b). Among samples without MSI or *POLE* mutations (which are both associated with high tumor mutational burdens, or TMBs), non-homozygous loss of *WRN* was associated with higher-than-average TMBs after controlling for lineage and overall aneuploidy ($p = 0.01$). Specifically, it was associated with a lower level of COSMIC mutational signature SBS6 (reflecting MSI), SBS42 (haloalkane exposure), and SBS46 (sequencing artifact) (all $q = 0.09$) and higher levels of SBS39 (etiology unknown) ($q = 0.15$) (Figure 3c and Supplementary Table 5b). These signature associations were recapitulated in tumors with arm-level loss of 8p (Supplementary Table 5b). Although MSI-annotated tumors were excluded in this mutational signature analysis, we hypothesize that the anticorrelation between *WRN* loss and SBS6 reflects mutual exclusivity between *WRN* loss and samples with unrecognized MSI. The etiology of the other signature associations is unclear. However, we also detected an increase in SBS39 in RNA sequencing data from cell lines with engineered 8p loss ($n = 8$) relative to their isogenic WT counterparts ($n = 8$) ($p = 0.02$, two-tailed Mann-Whitney U test) (Extended Data Figure 6f and Supplementary Table 5c). In contrast, these cell lines exhibited no significant differences for any of the 37 other COSMIC signatures detected in a *de novo* analysis of TCGA cancers. These data suggest that SBS39 results from 8p loss, possibly due to *WRN* inactivity.

Consistent with these findings, we found that siRNA knockdown of *WRN* (Extended Data Figure 6g) in cells with 8p disomy resulted in significant decreases in cell death by both flow cytometry and trypan blue staining (p = 0.003 and 0.04, respectively; Figure 3d and Extended Data Figure 6h–i). The degree of *WRN* knockdown and cell death were comparable to that of cells with engineered 8p loss (Extended Data Figure 6c–d). We also found that transfection-induced exogenous expression of *WRN* (Extended Data Figure 6j) decreased cell growth (p = 0.002; Figure 3e). We conclude that copy-loss of *WRN* likely accounts for some of the increased fitness of cells with 8p loss.

Validation of negative selection peaks

We next interrogated genes in del-neg peaks, which we hypothesized would be cancer cell-essential, for evidence that their loss would decrease cellular fitness. Among the 789 genes in the 45 “restricted” peaks (fewer than 50 genes each), 60 genes across 24 peaks were among a list of 1,482 genes that were previously found to be cell-essential based upon an integrated analysis of genetic and CRISPR screening data[30], which constitutes significant enrichment (p < 0.02; Supplementary Table 5d). We also analyzed published genome-scale functional screening data to test the hypothesis that the 789 genes in restricted del-neg peaks tend to be essential. Each peak often contains many genes (median of 25); only one of these might be under significant negative selection. For this reason, we anticipated that the average viability scores across all genes in these loci might be only slightly lower than the genome-wide average. Indeed, both RNAi suppression (p = 0.004) and CRISPR knockout (p = 0.001) of these genes led to slightly, but significantly, decreased cell viability compared to other assessed genes (Extended Data Figure 6k)[31,32].

We also picked three del-neg peaks for experimental validation in our cells with 8p disomy, focusing on peaks containing only 1–2 protein-coding genes. These peaks contained *EPN2* (17p11.2), *PPFIA1* (11q13.3), and *KAT6A* and *ANK1* (8p11.21). We performed siRNA knockdown of each of these genes and achieved successful knockdown of *EPN2* and *KAT6A* (Extended Data Figure 6l–m). Knockdown of both *EPN2* and *KAT6A* decreased cell growth relative to cells treated with control-pool siRNA (p = 0.0003 and 0.02 respectively; Figure 3f). From both our experimental and genome-scale analyses, we conclude that del-neg peaks detected by BISCUT often contain genes whose knockdown is harmful to cell growth.

Quantifying effects on fitness

We next extended BISCUT to estimate the magnitude of change in cellular fitness that derives from each peak locus by assessing the change in breakpoint frequency across the peak (Figure 4a and Supplementary Table 6a). In population genetics, fitness effects of genetic variants are often represented by “relative fitness” (RF), which represents the probability of survival of an individual with the variant compared to an identical individual without it. To calculate peak-level RFs, we divide the number of tumors observed to have alterations that include the peak by the number expected to in the absence of selection (only considering deletions for deletion peaks and amplifications for amplification peaks). We then combined peak-level RFs by multiplying them to estimate the total positive, negative, and overall net fitness effects contributing to each arm-SCNA, generating “arm-level” RFs (reported as log2 values; Extended Data Figure 7a and Supplementary Table 6b–c). In much of population genetics, RFs are calculated relative to individuals with the highest fitness. In contrast, we calculate RFs relative to cells without any copy-number alteration so as to explicitly represent positive selection.

For individual peaks, deletions exhibited greater positive and negative effects on fitness (average log RFs of 1.1 and −1.4 respectively) than amplifications (0.65 and −0.95 respectively). However, when summed across the arm, we found no significant difference between net RFs of deletions and amplifications (average −0.30 for both; p = 0.86). The arm-SCNAs under the most net positive selection were amplifications of 10p, 8q, and 7p, and deletions of 8p. Arm-SCNAs under the most net negative selection were deletions of 8q, 3q, and 17q, and amplifications of 8p.

Arm-SCNA RFs modestly correlated with Charm scores[33], a previously published measure of selection that is based on mutational profiles of known oncogenes and tumor suppressors rather than copy-number information. Among arm-level deletions, both the net and positive RFs significantly correlated with several Charm scores, most notably the Charm$^{\text{TSG-OG-Ess}}$ scores (net: Spearman’s rho = 0.45, p = 3.6e-3; positive: rho = 0.40, p = 0.012; Supplementary Table 6d). We did not find any significant correlations between Charm scores and RFs for arm-level amplifications.

Although we had found little difference in which loci BISCUT determined to be under selection between WGD and DIPLOID samples, we also considered the hypothesis that selection acting on these loci differed in magnitude between these two subsets of cancer. Prior studies have shown that tumor cells that have undergone WGD exhibit an increased tolerance to aneuploidy[2,34–36]. Cells adapt to aneuploidy by buffering gene and protein expression[37–39], though buffering is not as strong in cells with WGD[40]. We detected lower RFs, reflecting greater negative selection, in DIPLOID versus WGD samples (mean log net RFs of −0.49 and 0.04 respectively; paired t-test p = 0.0001; Figure 4b). However, we did not detect a difference for amplifications (net RFs of −0.20 and −0.19 respectively; p = 0.94). These results suggest that negative selection against deletions, but not amplifications, is more consequential in diploid samples.

Scoring mechanical biases

In addition to RFs, we generated two other metrics: centromeric and telomeric mechanical coefficients. Centromeric mechanical coefficients reflect observed rates of tel-SCNAs ending in a centromere (i.e. arm-SCNAs) compared to rates of tel-SCNAs ending adjacent to it, indicating biases favoring breakage within that centromere. Telomeric mechanical coefficients reflect frequencies of tel-SCNAs that do not encompass loci under selection, indicating mechanical biases favoring tel-SCNAs across different chromosome arms (Supplementary Table 6b–c and Extended Data Figure 7a).

Notably, all log centromeric mechanical coefficients were greater than 0, suggesting that mechanical biases favor breakage in all centromeres relative to elsewhere along the chromosome. The centromeres of the acrocentric chromosomes 13, 14, 15, 21, and 22 had significantly higher mechanical coefficients than other arms (average mechanical coefficients of 3.23 and 1.79; p = 0.0003). We expected this: both whole-chromosome SCNAs and arm-level SCNAs of acrocentric chromosomes appear as arm-SCNAs. Excluding these chromosomes, the centromeres with the highest mechanical coefficients were those of chromosomes 3, 5, and 17, while the centromeres of chromosomes 9, 1, and 16 had the lowest mechanical coefficients. Centromere mechanical coefficients did not correlate with centromere size or arm-SCNA frequency (Extended Data Figure 7b–c), consistent with previous studies suggesting that other DNA-dependent features instead influence chromosome segregation fidelity[41].

We also hypothesized that long telomeres protect chromosomes from telomeric copy-number events, and thus chromosome arms with longer telomeres would have lower telomeric mechanical coefficients. This was indeed the case (Spearman's rho = −0.65; p = 6.8e-6) (Figure 4c, Extended Data Figure 7d, and Supplementary Figure 3).

Only RF values predict arm-SCNA rates

To determine the contribution of selection versus mechanical biases on arm-SCNA formation, we assessed the correlation between their respective values and arm-SCNA rates within and across cancers (Extended Data Figure 7e and Extended Data Figure 8a). The only significant associations were between RFs and arm-SCNA rates. Moreover, as expected, the correlation was strongest between arm-SCNA rates and net RFs that reflect the aggregated effects of both positive and negative selection (Spearman's rho = 0.72 and 0.53 for amplifications and deletions respectively; Figure 4d). We also found no relationship between arm-SCNA rates and missegregation probabilities determined by single-cell sequencing of non-transformed cells[42] (Extended Data Figure 8b and Supplementary Table 6e), nor between chromosome arm length and RF values or arm-SCNA rates (Extended Data Figure 8c–d). All of these hypotheses were further tested in a multivariate model, which found that positive and negative RFs were the only statistically significant predictors of arm-SCNA rates (Extended Data Figure 8e). Charm scores were previously found to correlate with arm-SCNA rates. However, BISCUT RFs appear to convey additional information; RFs and Charm scores each correlate more closely with aneuploidy rates than they do with one another (Supplementary Table 6d)[33].

Selection in cancer can be lineage-specific[22], and our lineage-specific analyses support the notion that arm-SCNA rates are determined largely by their fitness effects. Similar to the pan-cancer analyses, we calculated net RFs for each chromosome arm in each unique TCGA lineage and found that they significantly correlated with arm-SCNA rates in that lineage in 55% (18/33) of cases (Fisher's method p-value 3.5e-39; Figure 4e and Supplementary Table 6f). The 15 cohorts in which a significant correlation was not observed were either small or primarily unaffected by SCNAs.

We conclude that selection is the main determinant of relative arm-SCNA rates both between chromosome arms and across cancer types.

DISCUSSION

This study directly addresses the longstanding question of whether aneuploidy is positively selected for in cancer. We find strong evidence that SCNA-mediated effects on cellular fitness are highly associated with rates of aneuploidy in cancer, often in a tissue-specific manner. We also find evidence of negative selection on arm-level SCNAs. 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- and centromere-bounded SCNAs – previously underappreciated subsets of somatic alterations – contribute new information to detect loci under selection in cancer.

As in all analyses that detect cancer drivers based on their frequency of alteration in cancer, BISCUT compares observed data to a background model that represents expected data in the absence of fitness effects. All of these methods can be biased by inaccuracies in their background models. For this reason, all candidate driver loci indicated by this or any other recurrence analysis require experimental validation. Further studies that improve the background model would greatly aid the detection of loci under selection. Despite these caveats, a key advantage of BISCUT is that it relies on step-function changes in breakpoint frequencies of partial-SCNAs rather than focally recurrent breakpoints, thus limiting the effects of localized fragility. Moreover, fragile loci usually undergo high rates of interstitial SCNAs, which are excluded from BISCUT analyses. None of the top 25 BISCUT deletion loci encompass known fragile sites, in contrast to 8 of the top 25 recurrent focal deletion peaks identified by GISTIC[36].

BISCUT is unlike other recurrence analyses for somatic genetic events because it relies on the distribution of breakpoints across chromosome arms as opposed to 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. 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. Our focus on the breakpoints of low-amplitude SCNAs does leave our analysis somewhat susceptible to false negatives – for example, both *VHL* and *BAP1*/*PBRM1* (all tumor suppressor genes in KIRC) are within a highly recurrent telomere-bounded deletion on 3p (Extended Data Figure 6b), but only the latter locus is near the tel-SCNA breakpoint and therefore detected by BISCUT. However, BISCUT's ability to detect effects of selection on low-amplitude SCNAs also allows it to identify driver genes that are missed by focal SCNA analyses. Future iterations of BISCUT can be adapted to detect loss-of-heterozygosity events, distinguish between different absolute copy-number states, and possibly incorporate gene expression.

Our findings indicate novel biology with respect to *WRN*. Germline mutations of *WRN* cause Werner syndrome, which is marked by premature aging including an increased risk of cancer. However, Werner syndrome only arises in the setting of biallelic *WRN* loss, and somatic alterations of *WRN* have not been shown to drive cancer. We find that partial suppression of *WRN* is sufficient to decrease cell death. We also find that hemizygous loss of *WRN* is associated with specific mutational signatures including a lack of MSI, consistent with prior studies showing that ablation of *WRN* is synthetic lethal with MSI[27]. These findings support *WRN* as a haploinsufficient tumor suppressor gene.

Our analyses indicate that the ubiquity of arm-SCNAs in cancer is due to both selective and mechanical pressures. The fragility of centromeres supports the frequency with which arm-SCNAs are observed overall; the relative frequencies of different arm-SCNAs appear to be determined primarily by fitness effects according to our analyses. Centromeric breakpoint frequencies vary widely – from two-fold to ten-fold – across chromosomes. The sources of these discrepancies are unclear and may include: 1) differences in centromeric and pericentromeric DNA sequences, some of which have been shown to be overexpressed in epithelial cancers[43], 2) three-dimensional folding of DNA[44], or 3) how different centromeres interact with the centrosome via kinetochores[45]. Historically, it has not been possible to precisely map centromeric breaks[24,46,47]. Hopefully, when recently developed long-read shotgun sequencing and analysis methods to map peri/centromeric regions[48,49] are applied en masse to tumor data, it will also be possible to determine whether breakpoints occur within the centromere or in the pericentromeric region. Future work to computationally model the likelihood of specific mechanical processes (e.g. merotelic attachment versus telomeric erosion)[50,51] 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 (mostly primary tumors except for 392 metastases, mostly melanomas, and 53 recurrent 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[53]. Briefly, from raw .CEL files, Birdseed was used to infer a preliminary copy number at each probe locus[54]. 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[55,56]. 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 Circular Binary Segmentation (CBS) as implemented by the DNAcopy R package (http://www.bioconductor.org/packages/release/bioc/html/DNAcopy.html), which segments DNA copy-number data and estimates copy-number ratios for each segment[57]. 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 standardized copy-number profiles generated as above are stored in the NCI Genomic Data Commons portal: https://docs.gdc.cancer.gov/.

The ABSOLUTE algorithm[58] 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[36]. These are log2-corrected relative to the average ploidy of the sample, regardless of whole-genome doubling status.

Generation and post-processing of segmented data from whole-genome sequencing platforms

Illumina HiSeq whole-genome sequencing DNA from 2,658 tumors and their matched germline genomes from ICGC was aligned to human reference hs37d5. Then, consensus copy-number alterations were constructed from outputs of six different SCNA callers as previously described for the PCAWG (Pan-cancer Analysis of Whole Genomes) project[25,59]. In brief, each cancer's genome was segmented into regions of constant copy number, separated by breakpoints denoting copy-number shifts. These breakpoints were based on PCAWG's consensus structural variants (SVs), also previously described, which were complemented with high-confidence breakpoints (the intersection of genomic regions where these SCNA callers agreed a breakpoint must exist). Allele-specific copy number for each consensus segment was determined and combined. Finally, segments were assigned a level of confidence based on the degree of consensus on the major and minor allele copy-number states. On average, a strict majority of callers agreed on 93% of each cancer's genome. The allelic copy-number profiles generated as above are stored in the ICGC Data Portal (PCAWG publication freeze): https://dcc.icgc.org/releases/PCAWG/consensus_cnv.

Samples were then filtered to only include those from the International Cancer Genome Consortium as opposed to The Cancer Genome Atlas. Duplicate samples were removed, and if there were multiple tumor samples from the same individual (e.g. primary and metastatic), then only the primary was kept. This left a total of 1,765 samples. Then, absolute copy numbers (i.e. sum of allelic copy numbers) were logarithmized relative to the sample's calculated ploidy: $CN_{relative} = log_2 \frac{CN_{absolute}}{ploidy}$ for consistency with copy-number data derived from SNP arrays (see above). For segments with absolute copy number of 0 (denoting a homozygous deletion), 0.1 was added to the absolute copy number to avoid the undefined value of $log_2(0)$.

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; Figure 1c), as well as determine background rates for these subtypes (Figure 1d). In some cases, partial-SCNAs 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 partial-SCNAs, 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 partial-SCNA lengths was uniform while the distribution of interstitial SCNA lengths decreases as 1/segment length (Figure 1g). In cases where a telomere- or centromere-bounded segment neighbored another segment in the same direction (i.e. gain or loss), we accounted for two possibilities: *first*, that these represented separate SCNAs (a partial-SCNA and a neighboring interstitial SCNA in the same direction), and *second*, that they represented a single partial-SCNA with an intervening interstitial SCNA in the opposite direction. In either case, the partial-SCNA would have the same probability, due to the near-uniform length distribution of partial-SCNAs. 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.

Based on the logic above, we implemented a segment-joining algorithm as follows: we 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. In the case of telomere-bounded deletions, we then calculated the distance between the centromeric end of the telomere-bounded deletion and the end of the last altered interstitial deletion (i.e. closest to the centromere). If the total length of copy-number-altered DNA was greater than that of the intermediate non-deleted DNA, we recorded the end of the last altered segment as the breakpoint location of the telomere-bounded deletion. However, if it was not, we iteratively removed the last altered segment until this is true. We then join the remaining deleted segments, and the end of the last deleted segment is recorded as the breakpoint location. This is equivalent to the length of the partial-SCNA, as a fraction over the length of the arm. The same approach applies for amplifications and centromeric-bounded events. An example is depicted in Supplementary Figure 4. As a test of the validity of this approach, we also implemented an alternative “half-joining” algorithm, in which we joined adjacent segments with the same copy-number direction – for example, if two segments were immediately adjacent and both represented amplifications, we would not end the partial-SCNA at the end of the first segment, but rather extend it to the end of the second segment. However, we stopped extending the partial-SCNA as soon as we encountered a segment with no copy-number change or a change in the opposite direction (e.g. a deletion following an amplification).

If this approach yielded a tel-SCNA that ended at the centromere, a cent-SCNA that ended at the telomere, or an overlapping tel-SCNA/cent-SCNA (e.g. tel-SCNA ending close to the centromere and cent-SCNA ending close to the telomere in the same direction), we designated an arm-SCNA on this arm. Arm-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” arm-SCNAs were included in the calculation of the centromeric mechanical coefficient; 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 arm-SCNAs were considered. Somatic mutations were called from 9,423 tumor exomes across 33 TCGA tumor types[28]. 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 (v2.0.23 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[60]. 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* versus *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[61]. 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.

BISCUT peak-finding algorithm

BISCUT's peak-finding algorithm performs two main functions: 1) it detects loci that appear to underlie fitness effects of arm-SCNAs; 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 affected by fitness effects, we first sort our set of partial-SCNA breakpoints $T_i = \{t_1, t_2, \cdots, t_n\}$ by increasing order. If there were multiple breakpoints in the cohort ending at the same SNP array probe (suggestive of a region lacking SNP coverage), these were linearized uniformly to the next covered probe (e.g. if bases 10 and 19 are present in the array, but there are three breakpoints ending at base 10, then these would be analyzed as 10, 13, and 16). We then use a two-sample Kolmogorov-Smirnov test to compare T_i 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 partial-SCNAs.

The boundaries of the “peak region” are detected such that they include the gene, set of genes, or other genomic elements that confer fitness effects when altered at a confidence level of $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ 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 partial-SCNA 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 (v1.0–14)[62] (Supplementary Table 2e).

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

$$direction\ of\ selection = \{positive\ if\ |(T_i - B_i)| > 0,\ else\ negative\}$$

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* has a 95% chance of belonging to a distribution unaffected by fitness effects. For this purpose of determining confidence intervals, we use an approximation of the background distribution for computational efficiency; specifically a beta distribution that has been fit to the empirical background: B_i. The helper function first approximates the generalized extreme value distribution $GEV_{left}(peak_0)$, which comprises 1000 independently generated maximum values of n_{left} random variates following the background distribution B_i, where n_{left} is the number of tumors to the left of the peak. This is repeated on the right. In order to define the right boundary of our peak region, we repeat $GEV_{left}(peak_x)_{x=1}^{\infty}$ where x represents the xth tumor to the *right* of the original *peak*. If $peak_{x-1}$ is greater than the 97.5% percentile of $GEV_{left}(peak_x)$, then we call $peak_x$ as the right boundary of the BISCUT peak region. If not, we infer that $peak_{x-1}$ is unlikely to belong to the left-sided distribution, and we continue $GEV_{left}(peak_x)_{x=1}^{\infty}$ until the former is true. To define the left boundary of our peak region, we perform $GEV_{right}(peak_x)_{x=-1}^{-\infty}$ where x represents the xth tumor to the *left* of the original *peak*. If $peak_{x-1}$ is greater than the 97.5% percentile of $GEV_{right}(peak_x)$, then we call $peak_x$ as the left boundary of the BISCUT peak region. If not, then we infer that $peak_{x+1}$ is likely to belong to the right-sided distribution, and we continue $GEV_{right}(peak_x)_{x=-1}^{-\infty}$ until the former is true. This corresponds to a 95% confidence interval; this confidence level can be adjusted by the user.

We repeat the entire BISCUT 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[63], 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.

Assessment of breakpoint frequency by genome properties

Sources for each of the following genomic properties are listed:

Alu Repeat	Li et al., 2020[59]
LINE Repeats	Li et al., 2020
SINE Repeats	Li et al., 2020
LTRs	Li et al., 2020
Simple Repeats	Li et al., 2020
Satellite Repeats	Li et al., 2020
Common TAD Boundaries	Li et al., 2020
LAD Domains	Li et al., 2020
CTCF Bound Domains	Li et al., 2020
Common Fragile Sites	Li et al., 2020
Evolutionary Conserved CpG Islands	Li et al., 2020
Genes	Li et al., 2020
ORC Binding Sites	Miotto et al., 2016[64]

Enrichment of known cancer genes in BISCUT peak regions

We calculated the statistical significance for the overlap between unique genes in BISCUT peak regions and those reported to be cancer-driving genes from the COSMIC Cancer Gene Census[65] using a hypergeometric test implemented at the following website: http://nemates.org/MA/progs/overlap_stats.html. Specifically, we used "restricted" peak lists; i.e. BISCUT peaks that contained 50 genes or fewer. 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.

Estimating dN/dS (non-synonymous to synonymous mutation) ratios in BISCUT peak genes

The dNdScv method[22] was used to obtain global estimates of the ratio of non-synonymous to synonymous mutations in a) the 548 genes that are protein-coding (RefCDS) and are identified by BISCUT to be in restricted del-neg peaks (50 genes or fewer), and b) the 446 protein-coding genes that are identified by BISCUT to be in restricted del-pos loci. In both cases, dN/dS ratios were compared to 1000 randomly selected sets of the same number of genes (548 and 446, respectively) to calculate the statistical significance of their deviations from what would be expected by random chance.

Lineage-corrected comparisons of breakpoint distributions

A few analyses throughout this manuscript involve comparison of partial-SCNA distributions between different groups or characteristics of tumors (e.g. diploid versus whole-genome doubled samples). Because these characteristics were often confounded by tumor type, we performed KS-statistics between the partial-SCNA distributions of these groups separately for each independent tumor type, and combined p-values using Fisher’s method. Comparisons that did not have at least four samples in each group were excluded. FDR correction was performed on p-values derived from Fisher’s method.

Lineage specificity of breakpoint distributions

To determine the relative lineage specificity of partial-SCNAs involving specific chromosome arms, we first compared the breakpoint vector of a partial-SCNA (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)[66,67] 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 partial-SCNA (amplification or deletion in each chromosome arm) as the “lineage-specificity score” (Supplementary Table 4g).

Overlap of BISCUT peak regions and clustering analysis

Two peak regions from different cohorts were considered to overlap if their 95% confidence intervals intersected. When assessing for significant peak overlap between two cohorts (i.e. sets of peak regions), peaks were only compared to other peaks sharing the same directionality (i.e. amplification versus deletion) and selection type (i.e. positive versus negative) using the R packages GenomicRanges[68] and regioneR[69]. Specifically, we obtain p-values from the overlapPermTest function from the latter package, which performs a permutation test to see if the overlap between two sets of regions is more or less frequent than expected by chance. We then combine p-values between each of the four categories (amp-pos, amp-neg, del-pos, and del-neg) using Fisher's method.

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 BISCUT 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 BISCUT's ability to detect driver events, we generated *in silico* sets of partial-SCNA breakpoints, with various degrees of simulated selective advantage or disadvantage, and at multiple locations across a chromosome arm (Supplementary Figure 1a). Background distributions for tel-SCNAs and cent-SCNAs were assessed independently; within those, amplifications and deletions were combined (Supplementary Table 2e).

To create the breakpoint data, we started by generating a set of n = [200, 500, 1,000, 10,000] 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 tel-SCNAs and cent-SCNAs.

We then ran BISCUT at a confidence level of 0.95 on the simulated sets of tel-SCNAs and cent-SCNAs separately. For each combination of n, l, s, and type of partial-SCNA (tel-SCNA versus cent-SCNA), we report the frequency at which BISCUT 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 BISCUT 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: $F_1 = 2 * \frac{precision \quad * \quad recall}{precision \quad + \quad recall}$. For each statistic, we generated a "combined" statistic by taking the weighted average of the statistic from tel-SCNAs and cent-SCNAs, where weights are defined as the total number of tel-SCNAs and cent-SCNAs in our dataset (n = 51,588 and 34,007 respectively).

Simulations of delta functions

Similar to above, we also wished to test BISCUT's performance on simulated "easily breakable" loci that might interfere with our background distributions. Each breakable locus would be observed as a delta function in breakpoint density, so we simulated such delta functions to assess how such loci would affect BISCUT results.

To create the breakpoint data, we started by generating a set of $n = [50, 100, 200, 500, 1{,}000]$ random samples from the averaged beta background distribution. We defined several locations for the delta function $l = [0.1, 0.2, \ldots, 0.9]$, which represents the distance across the chromosome arm. We also tested several frequencies at each set location $f = [0.01, 0.02, \ldots, 0.5]$, which represents the fraction of total samples that is designated to break at l (e.g. if n is 200, l is 0.7, and f is 0.24, then there will be exactly 48 simulated tumors with a breakpoint at 0.7 of the chromosome arm). This was repeated 100 times for each combination of n, l, and f.

We then ran BISCUT at a confidence level of 0.95 on the simulated datasets. For each combination of n, l, and f, we report the frequency at which BISCUT includes the locus l in its peak region, which reflects false positive hits. Specifically, we also report the minimum fraction of samples at each l required to cause false positive results 5% and 50% of the time (Supplementary Figure 1e–g).

Simulations of SNP "coverage deserts"

To assess the issue of non-uniform coverage of the genome in the experimental platform used to assess copy numbers, we simulated "coverage deserts" ranging from 1% to 10% of a chromosome arm.

To create the breakpoint data, we started by generating a set of $n = [100, 500, 10{,}000]$ random samples from the averaged beta background distribution. We set l (location of theoretical driver gene) at 0.5, and used several levels of selective pressure $s = [0.1, 0.2, 0.25, 0.5, 1, 2, 4, 5, 10]$. We then generated SNP "coverage deserts" of size 0.01, 0.05, and 0.1 of the chromosome arm starting from desert location $d = [0.1, 0.2, \ldots, 0.9]$. This was repeated 100 times for each combination of n, s, d, and desert size.

We then ran BISCUT at a confidence level of 0.95 on the simulated datasets. For each combination of n, s, d, and desert size, we report the frequency at which BISCUT successfully includes the locus l in its peak region (power). We also report the specificity, calculated as 1-false positives; we count a false positive when BISCUT reports d as the location of the peak (Supplementary Figure 1h).

Mutational signatures in TCGA samples with WRN/8p loss

We wished to compare the mutational signature profiles between TCGA tumors with and without *WRN* copy loss and with and without 8p arm-level deletions. Only tumors with whole-genome sequencing data included in the PCAWG (Pan-cancer Analysis of Whole Genomes) project[25] were analyzed (total n = 828). Consensus somatic SNV and indel calls, which were derived from merging calls from various caller pipelines using the SNV-MERGE script[70], were downloaded from the ICGC Data Portal (https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel).

We first extracted *de novo* SBS signatures from the samples with SignatureAnalyzer (https://github.com/getzlab/SignatureAnalyzer)[71,72]. Each *de novo* signature was then assigned to the COSMIC v3 SBS signature that had the highest cosine similarity with the *de novo* signature (https://cancer.sanger.ac.uk/signatures). Since multiple *de novo* signatures mapped to a COSMIC signature, we then ran supervised signature extraction with all unique COSMIC SBS signatures (38 total) identified from the *de novo* analysis (Extended Data Figure 6f).

We compared the relative activity of each COSMIC signature between TCGA tumors with and without *WRN* copy loss, and with and without 8p arm-level deletions. Samples associated with MSI and *POLE* mutation were removed[73], and the remaining samples were grouped by *WRN* copy-number status (WT versus *WRN* copy loss) or 8p status (WT versus 8p arm-level deletion). We performed two-tailed Mann-Whitney U tests to compare the relative signature activity between samples with and without *WRN*/8p loss. P-values were calculated per independent tumor type, combined using Fisher's method, and FDR corrected.

Cell maintenance and genome engineering

8p was deleted in XX (female) immortalized lung epithelial cells (AALE cells) by SV40 large-T antigen[74] which tested negative for mycoplasma and were authenticated using DNA fingerprinting. All AALE cells were maintained at 37 degrees Celsius and 5% CO_2 in Lonza small airway growth medium (CC-3118). Deletion methods were as previously described[2]. Briefly, two different Cas9 guide sequences were used: TATGCTATACGGAATTCCAT and TAATAAAGAACTATGCTATA . Guides were cloned into px330 (Addgene 42230). One kilobase of sequence homologous to the 8p pericentromeric region adjacent to the CRISPR cut sites was amplified by primers AATGGCACAGTGCTTTACAG and GCAGCTTAGCCAATGGAAGC and cloned via Gibson assembly (New England Biolabs E2611) into a telomere-containing plasmid (TCP) with puromycin resistance. Cells were transfected with 1.2 µg guide plasmid and 1.2 µg linearized or digested TCP. Puromycin selection at 2 µg/mL started one day post transfection. Genomic DNA was isolated from puro-resistant clones using the QiaAmp Mini DNA kit (Qiagen) and analyzed for recombination via PCR. Single-cell cloning was used to isolate clones with 8p deletion. Validation of deletion was performed by PCR, qPCR, and low-pass whole-genome sequencing.

DNA/RNA sequencing of genome engineered cell lines

DNA was isolated from cells using the QiaAmp Mini DNA kit (Qiagen) following manufacturer instructions. Sequencing library preparation was performed using the Lotus DNA Library Kit (IDT) following kit protocols. The index sequences used were the xGen™

Stubby Adapter and UDI Primer Pairs (IDT) Version 2. Samples were pooled and sequenced by miSeq, 300 bp paired end. Copy-number profiles were generated via HMMCopy (https://bioconductor.org/packages/HMMcopy/). Subclonal analysis was performed using ichorCNA v0.2.0 (https://github.com/broadinstitute/ichorCNA)[75].

RNA isolation was performed using the RNeasy® Mini Kit (Qiagen), including an optional DNase digestion step and an optional extra spin. The amount of RNA submitted for each sample ranged from 200–800 ng. All library preparation and subsequent RNA-seq was completed by the JP Sulzberger Columbia Genome Center. Polyadenylated RNA was used for library preparation, and each sample had at least 20 million reads on an Illumina NovaSeq.

Mutational signatures in 8p engineered cell lines

RNA-sequencing FASTQ files were obtained from cell lines with engineered 8p arm-level deletions and their disomic counterparts (n = 8 per category). Following GATK best practices, the FASTQs were aligned to hg38 with STAR v2.7.10b, and processed with MarkDuplicates v3.0.0, and SplitNCigarReads v4.0.11.0[76,77]. Variants were then extracted with Mutect2 and annotated with Funcotator. We removed variants that were in the gNOMAD and 1000 Genomes databases, and also those that occurred in more than one cell line. We then ran supervised signature extraction with the 38 unique COSMIC SBS signatures previously identified in the TCGA samples (see Mutational signatures in TCGA samples with WRN/8p loss section above). Two-tailed Mann-Whitney U statistics were calculated to compare the relative signature activity between the cell lines with and without 8p loss.

Cell transfection, and proliferation assays

Transfections were performed using Fugene-6 (Promega E2691) following manufacturer’s instructions at a 3:1 ratio in 12-well plates and 96-well plates. *WRN* siRNA-mediated knockdown was performed with siGENOME Human *WRN* siRNA (Horizon Discovery D-010278–02-0020) for a final concentration of 90nM. In parallel, siGENOME Non-Targeting siRNA Pool #1 (Horizon Discovery D-001206–13-20) was used as a negative control. *WRN* overexpression was performed by transfecting plx209neo-WRN[27] and plx209neo-EGFP as a control at a final concentration of 2μg. Cells were selected with neomycin at a final concentration of 1mg/mL starting 48 hours after transfection for ten days.

For cell proliferation assays, 1500 cells were plated per well in a 96-well plate in 100μL of media. For cells with *WRN* overexpression, at days 0, 3, and 5, CellTiter-Glo Reagent (Promega G7570) (20μL) was added. Plates were incubated on a benchtop shaker at room temperature for 10 minutes before luminescence readings. For cells with siRNA transfection, cells were transfected one day after plating and incubated with CellTiter-Glo 2–6 days post transfection, using the same incubation protocol.

Flow cytometry and caspase assays

For apoptosis analysis, adherent cells were first washed, trypsinized, and collected. When possible, floating cells pre-wash were also collected and analyzed. Pre-staining, 10μL was removed for trypan blue staining at a 1:2 ratio and percent of living cells counted using the Invitrogen Countess II. Cells were stained with annexin V (Tonbo Biosciences 35–6409-T100) following manufacturer's instructions for 15–20 minutes. Cells were then stained with propidium iodide (PI) solution. PI and annexin V levels were measured on the BD Fortessa. Gating cut-offs for PI and annexin V stained samples were determined by comparison to unstained or singly stained samples using FlowJo (strategy detailed in Supplementary Figure 2). For caspase assays, 1500 cells were plated per well in a 96-well plate in 100 μL media. To read caspase signal, Caspase-Glo Reagent (Promega G8091) (20μL) was added to each well two days post siRNA transfection.

Quantitative PCR

Quantitative PCR (qPCR) was performed using the Power SYBR Green PCR Master Mix (ThermoFisher 4367659), with the listed primers. Normalizations were performed against an endogenous *ACTB* or *UBC* control. qPCR Primer sequences:

WRN F: GCGACATGAACAAACAGTTGA

WRN R: GCTGGGCCTCAGTTCAGTCT

KAT6A F: TGGCTCCAGTCAGTTCTACAC

KAT6A R: TGAGAATTGGTGGCGAGCTT

ANK1A F: CTTCTTAGGGGGTGTCGCC

ANK1A R: GTGAAATTGACGCTGGCTCC

EPN2 F: GGGTGTCAAACTGAGCCAGA

EPN2 R: CAGTGAGCACCCAGCACTTA

PPFIA1 F: GAGACTAAGAGCCGACCCCA

PPFIA1 R: AGACTTCCACTGCCAACTCG

ACTB F: TGGAGAAAATCTGGCACCAC

ACTB R: AGGGATAGCACAGCCTGGAT

UBC F: CGGGATTTGGGTCGCAGTTCTTG

UBC R: CGATGGTGTCACTGGGCTCAAC

Peak-level and arm-level RFs

To calculate the "peak-level" RF of each BISCUT 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 versus deletion, and telomere-bounded versus 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 BISCUT peak, we considered the tumors with partial-SCNAs smaller than the length ascribed to the leftmost (i.e. smallest) BISCUT peak to be under no selective advantage or disadvantage (i.e. "reference"), where the empirical number of partial-SCNAs 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 partial-SCNAs expected to be in this segment in the absence of selection as:

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

where $I_x(x; \alpha, \beta)$ is the incomplete beta function and α and β are the corresponding probability parameters. We then reported the "peak-level" RF for a peak *p* as the number of empirical partial-SCNAs 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 7a).

The "arm-level" RF is a representation of the density and "peak-level" RFs of BISCUT peaks on each arm, calculated separately for amplifications and deletions. For each arm-SCNA, positive and negative selection peaks are assessed both separately (positive and negative RFs respectively) and together (net RF). We then define the log RFs as the log2 value of the product of the relevant RFs (only those greater than 1 for positive selection, only those smaller than 1 for negative selection, and all RFs for net selection), such that net log RFs 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 mechanical coefficients

We developed a centromeric mechanical coefficient 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): mechanical coefficient for amplifications and deletions of the short and long arms (only long arms for acrocentric chromosomes).

To calculate centromeric mechanical coefficients for each arm and direction of SCNA (i.e. amplification versus deletion), we divided the density of arm-SCNA breakpoints within the centromere (the number of "centromeric" arm-SCNAs, see above) by the density of breakpoints within the region immediately flanking the centromere up to the nearest BISCUT peak (Extended Data Figure 7a). 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 partial-SCNA breakpoints only out to the nearest BISCUT peak because these partial-SCNAs likely underwent similar selection as the centromeric arm-SCNAs and we wanted to exclude effects of selection when calculating these mechanical coefficients. For consistency with RFs, we reported the log2 value of this quotient, such that a value greater than 0 represents positive mechanical bias and a value less than 0 represents negative mechanical bias.

We calculated telomeric mechanical coefficients for each chromosome arm and direction of copy-number change as the number of tel-SCNA breakpoints occurring prior to the start of the BISCUT peak closest to the telomere, divided by the incomplete beta function (i.e. the cumulative probability distribution) at the start of the first BISCUT peak, measured as the fraction of the length of its chromosome arm (Extended Data Figure 7a). These coefficients were designed to reflect the propensity for telomeric events to occur in the absence of selection. Telomeric mechanical coefficients were recalculated using the “2/4 analysis,” in which relative copy-number segments were renormalized to a baseline ploidy of 2 if the sample did not undergo whole-genome doubling, or 4 if it underwent one whole-genome doubling. Samples designated as having undergone two or more whole-genome doubling events were removed. The original log2 copy ratio r (after ISAR correction as described above) was renormalized to a baseline ploidy of 2 or 4 (r') using the equation:

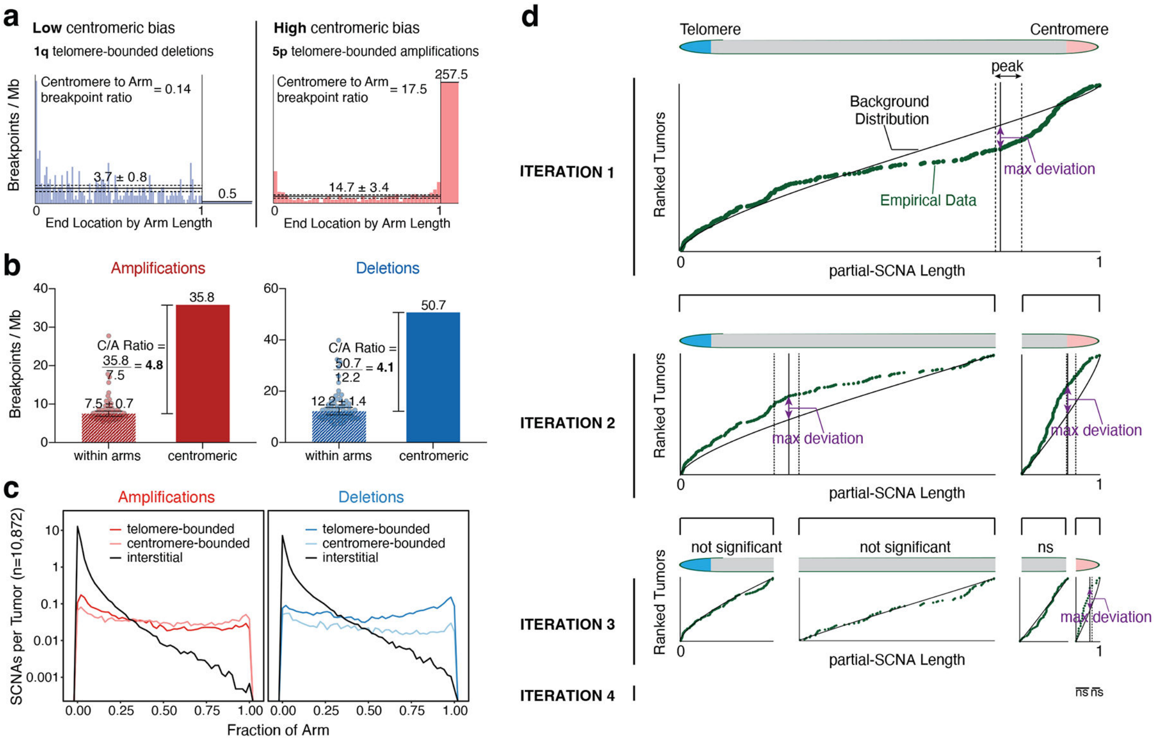
$$r' = log2\left(\frac{2^{r} * ploidy}{2 \quad or \quad 4}\right)$$

Lastly, in order to test which coefficients correlated best with rate of arm-SCNAs, we used a Generalized Linear Model (GLM) framework for a multivariate analysis (Extended Data Figure 8e). Specifically, we fit separate GLMs to arm-SCNA rates of amplifications and deletions across all cancer types. The arm-SCNA rate along each chromosome arm was modeled as a random sample from a normal distribution where the mean is a linear function of predictors (i.e., positive RF, negative RF, telomeric mechanical coefficient, and centromeric mechanical coefficient (all in logarithmic scale):

$$E[Y] = \sum_{i} \beta_i X_i$$

where Y is the random variable for modeling arm-SCNA rates, X_i is a predictor, and β_i encodes the strength of correlation between X_i and Y. A p-value was also computed for each predictor’s correlation with arm-SCNA rates. We then used Fisher’s method to combine p-values from independent tests (individual cancer types) to obtain a single p-value for the pan-cancer meta-analysis.

Extended Data



Extended Data Figure 1: Additional information on different types of SCNA and the BISCUT method.

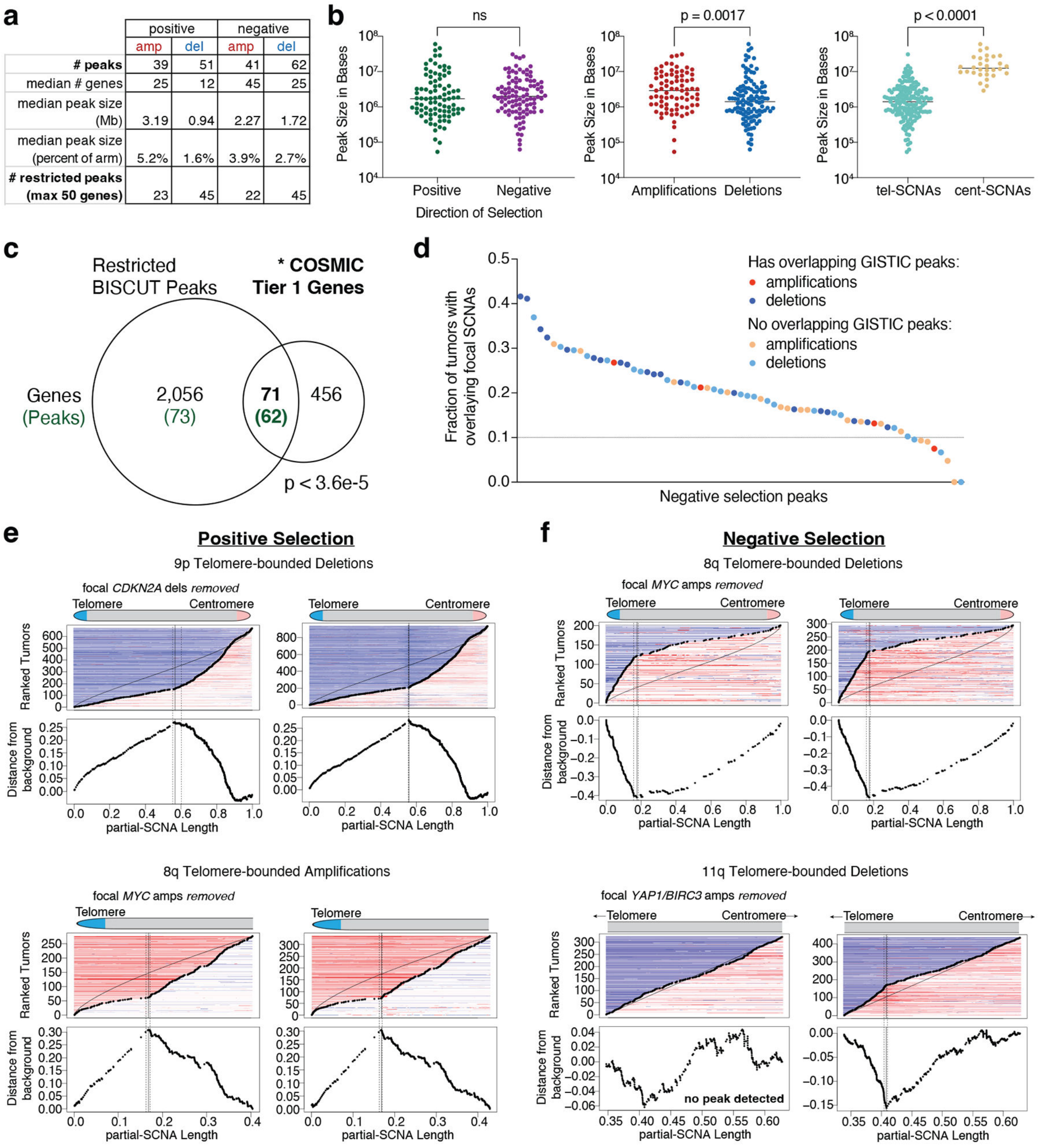
(a) Empirical examples of low centromeric mechanical bias (1q telomere-bounded deletions, for which the ratio of breakpoints occurring in the centromere over those occurring in the arm is less than 1), and high centromeric mechanical bias (5p telomere-bounded amplifications, for which the centromere/arm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

(b) Mean amplification and deletion breakpoint density within chromosome arms, aggregated across all tumors and all chromosome arms (n = 67; binned by Mb), 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.

(c) Comparison of length distributions of telomere-bounded, centromere-bounded, and interstitial amplifications and deletions, aggregated across all chromosome arms.

(d) Example depicting BISCUT's recursion steps. From top to bottom: BISCUT 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 Figure 2c and Methods for details.

Extended Data



	positive		negative	
	amp	del	amp	del
# peaks	39	51	41	62
median # genes	25	12	45	25
median peak size (Mb)	3.19	0.94	2.27	1.72
median peak size (percent of arm)	5.2%	1.6%	3.9%	2.7%
# restricted peaks (max 50 genes)	23	45	22	45

Extended Data Figure 2: Pan-cancer BISCUT analysis.

(a) Summary statistics of the four types of BISCUT peaks in pan-cancer.

(b) Sizes of peaks (in bases) from the pan-cancer BISCUT analysis. From left to right, peaks are categorized by direction of selection (n = 90 and 103 for positive and negative selection respectively), direction of copy number imbalance (n = 80 and 113 for amplifications and deletions respectively), and origin of partial-SCNA (n = 163 and 30 for telomere-bounded and centromere-bounded respectively). Two-tailed p-value was calculated using a Mann-Whitney U test.

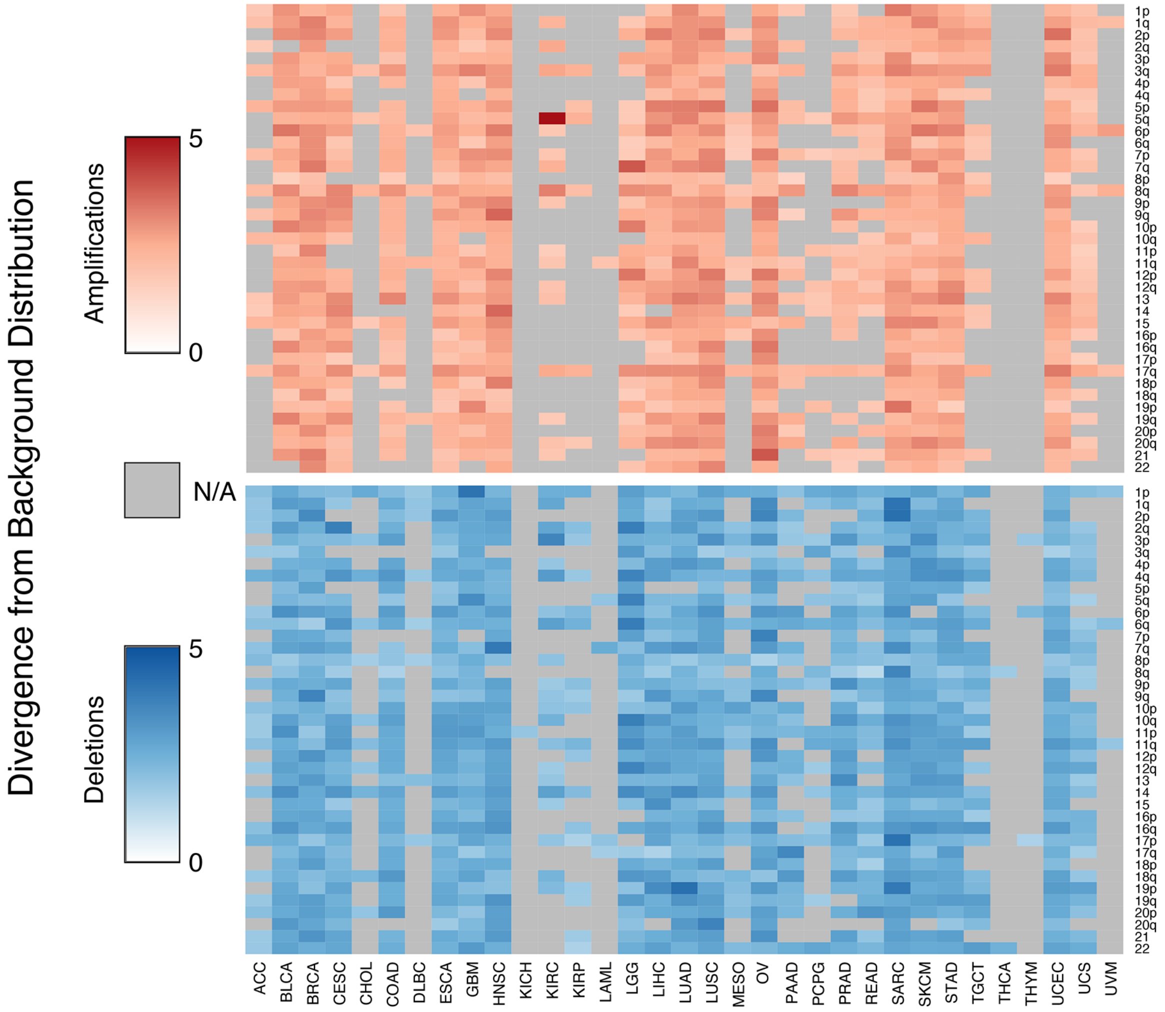
(c) Overlap between genes in BISCUT peaks and Tier 1 COSMIC cancer genes. The numbers of peaks containing these genes are depicted in green. A one-tailed p-value was calculated using a permutation test as outlined in the Methods.

(d) Negative selection peaks from the pan-cancer BISCUT analysis, sorted from highest to lowest by fraction of samples subject to these fitness effects 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.

(e) BISCUT 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).

(f) BISCUT 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).

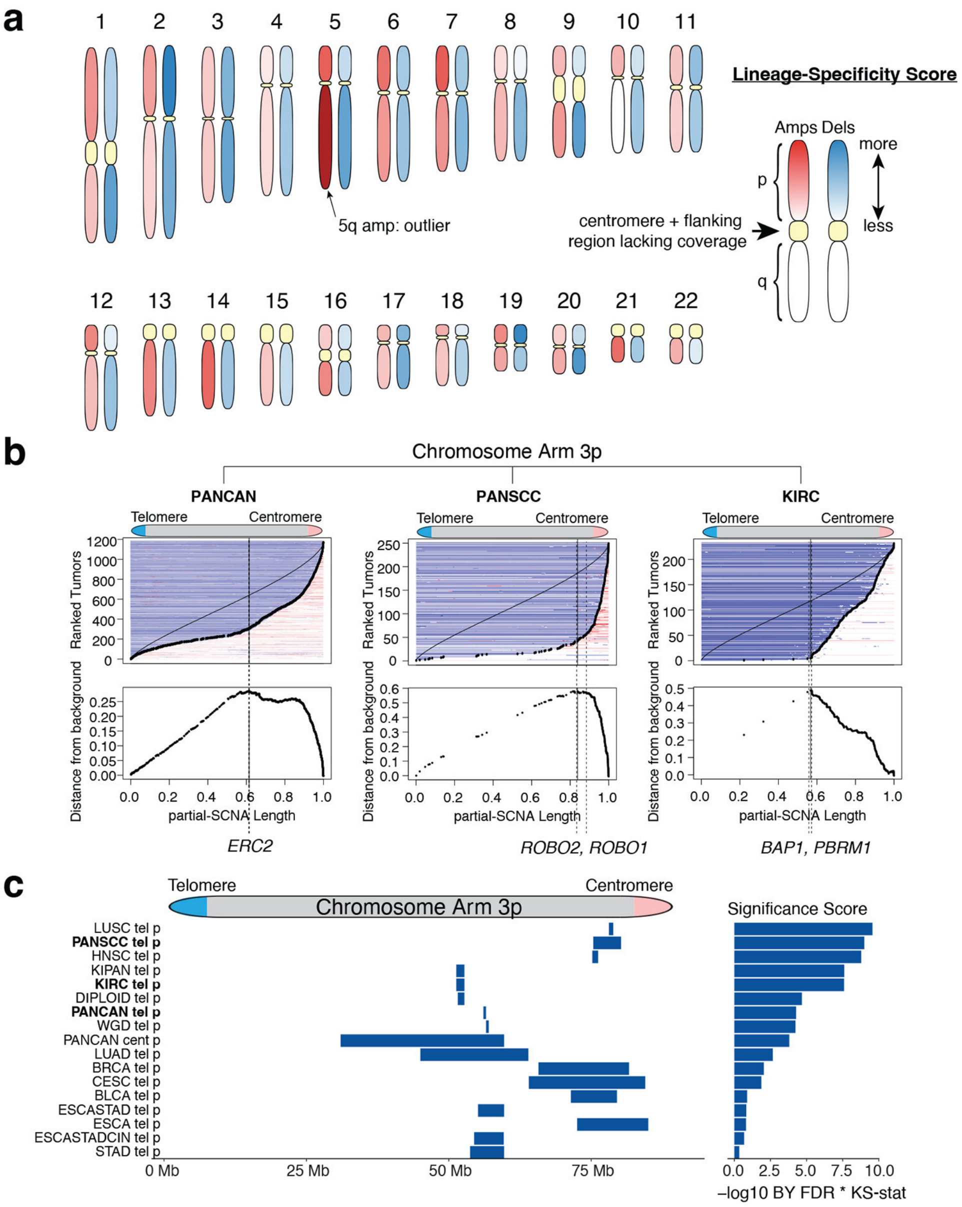
Extended Data



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 a higher divergence score.

Extended Data



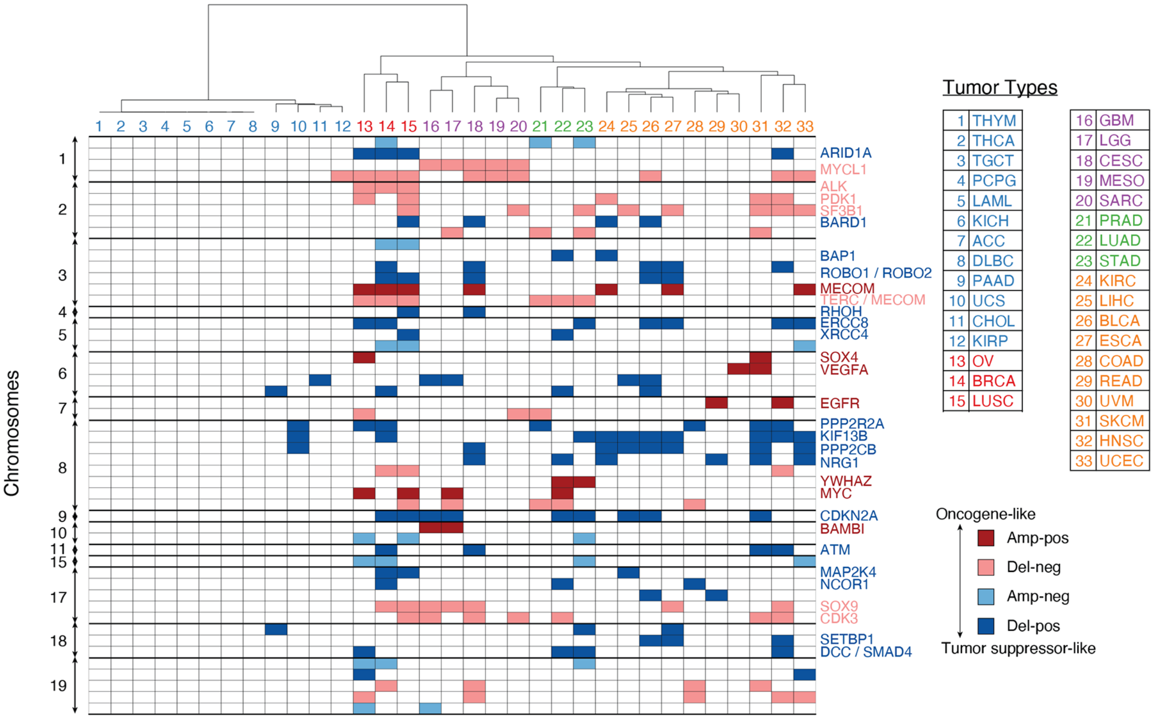
Extended Data Figure 4: Patterns of chromosome 3p deletions are highly lineage-specific.

(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) BISCUT analysis of telomere-bounded deletions on chromosome 3p in three different cohorts. The top panels display telomere-bounded deletions, sorted by length. The bottom panels show the vertical distance of each tel-SCNA 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 BISCUT peaks on chromosome 3p across lineages. See Supplementary Table 2a for tumor type abbreviations.

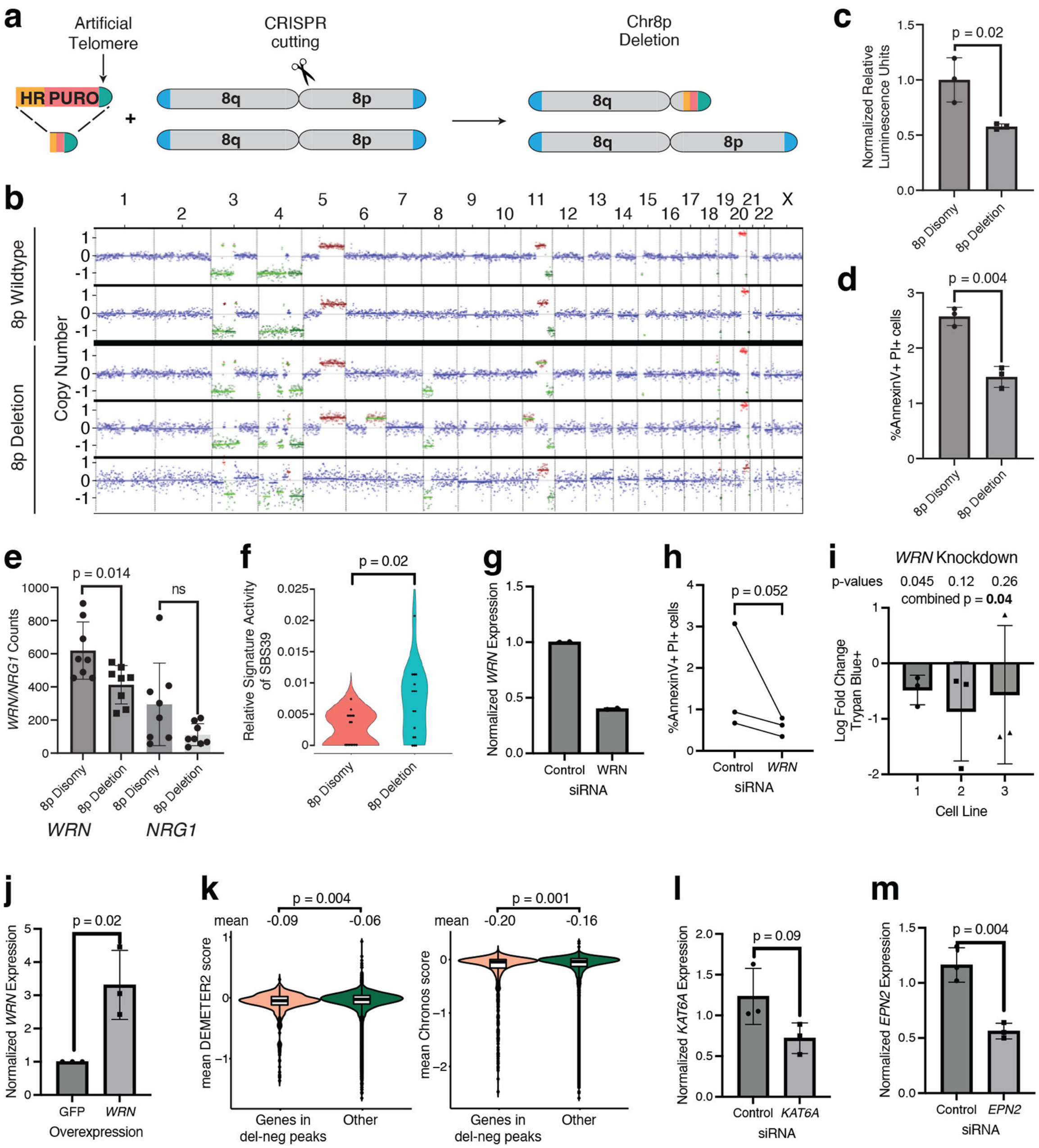
Extended Data



Extended Data Figure 5: Hierarchical clustering of BISCUT peaks across lineages.

Matrix of significantly recurring BISCUT peaks across 33 independent tumor types. Peaks are sorted by genomic location (vertical axis), with four distinct classes of peaks in dark red (positive selection in amplifications), light red (negative selection in deletions), light blue (negative selection in amplifications), and dark blue (positive selection in deletions). Tumor types are sorted and color-coded (k = 5) according to hierarchical clustering by Ward's method (horizontal axis).

Extended Data



Extended Data Figure 6: Cells engineered with chr8p deletion for validation of genes in BISCUT selection peaks.

(a) Schematic for 8p deletion approach. Cells were transfected with a CRISPR targeting 8p just outside the centromere and with a linearized plasmid containing an artificial telomere, puromycin selection cassette, and 1 kilobase of sequence homologous to the 8p pericentromeric sequence. Puromycin selection was used to isolate cells with 8p replaced by the artificial telomere.

(b) ichorCNA output of ultra-low-pass whole genome sequencing data from five AALE cell clones with 8p disomy or 8p monosomy. Horizontal axis is chromosome number, vertical axis is log copy number ratio. Green denotes copy number loss, red denotes copy number gain.

(c) Caspase-glo for cells with 8p deletion compared to cells with 8p disomy (n = 3 for both). Each point represents one biological replicate from a representative experiment. One-tailed p-values from two different experiments were combined using Fisher's method.

(d) Flow cytometry analysis of cells with 8p deletion compared to cells with 8p disomy. Bar graphs represent the percentage of apoptotic cells dually stained for Annexin V and PI. One representative experiment is shown. One-tailed p-values from three independent experiments were combined using Fisher's method.

(e) Vertical axis represents normalized read counts from RNA sequencing of cells with 8p disomy or 8p deletion. Each point is an individual clone (n = 8 for all columns). Two-tailed p-values are reported.

(f) Relative COSMIC SBS39 mutational signature activity (vertical axis) of engineered cells with 8p disomy versus 8p monosomy. Two-tailed p-values are calculated from a Mann-Whitney U test.

(g) *WRN* qPCR for cell clones with 8p disomy after siRNA treatment. Cells were treated with either control siRNA or siRNA against *WRN* for 3 days prior to qPCR (n = 2 for each condition). Each point represents the average value across technical replicates in an individual biological replicate.

(h) Percentages of apoptotic cells detected by flow cytometry for Annexin V and propidium iodide (PI) across three 8p wild-type cell lines, on day three after transfection with *WRN* versus control siRNAs. This is representative data from one of four experiments. A ratio paired t-test was used to calculate one-sided p-values for all four experiments, which were combined using Fisher's method.

(i) Log-fold changes in apoptotic cells detected by trypan blue across these three 8p wild-type cell lines (n = 3 for all cell lines), on day three after transfection with *WRN* vs control siRNAs. Each point represents a different experiment. One-tailed p-values from all experiments were combined using Fisher's method.

(j) *WRN* qPCR in 8p disomic cell clones with overexpression of *WRN* or *GFP* (n = 3 for both). A two-tailed p-value is reported.

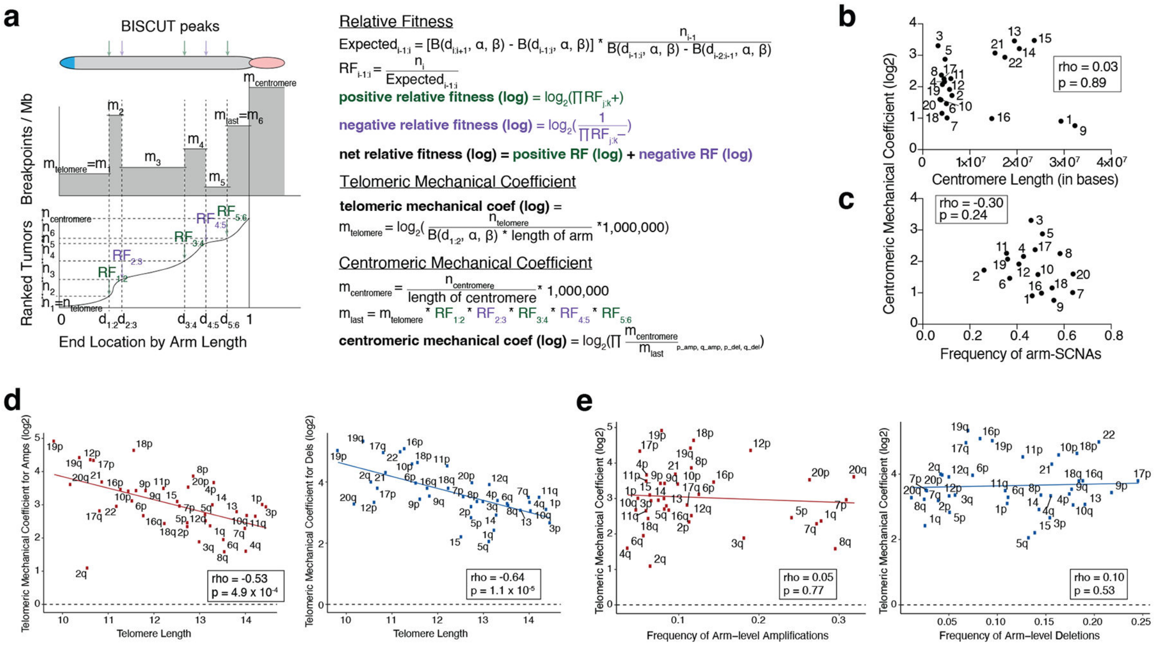
(k) Cell viability is significantly lower when genes in del-neg peaks are knocked down by RNAi (left, DEMETER2 score) or knocked out by CRISPR (right, Chronos score) in Dependency Map screens[31,32], compared to all other genes. The reported p-value is two-tailed. Box plots center on median values and extend to the first and third quartiles; the whiskers extend to 1.5 times the interquartile range.

(l) *KAT6A* qPCR for cell clones three days after siRNA-mediated knockdown (n = 3 for both conditions). A two-tailed p-value is reported.

(m) *EPN2* qPCR for cell clones three days after siRNA-mediated knockdown (n = 3 for both conditions). The reported p-value is two-tailed.

All p-values in this figure were calculated using Student's t-test except as otherwise noted; no adjustments were made for multiple comparisons.

Extended Data



Extended Data Figure 7: Quantitative assessment of selective and mechanical pressures driving aneuploidy.

(a) Calculation of peak-specific relative fitness (RF), arm-level RF, telomeric mechanical coefficients, and chromosome-level centromeric mechanical coefficients. See Methods for further details.

(b) Centromeric mechanical coefficients (log) plotted against centromere length (in bases).

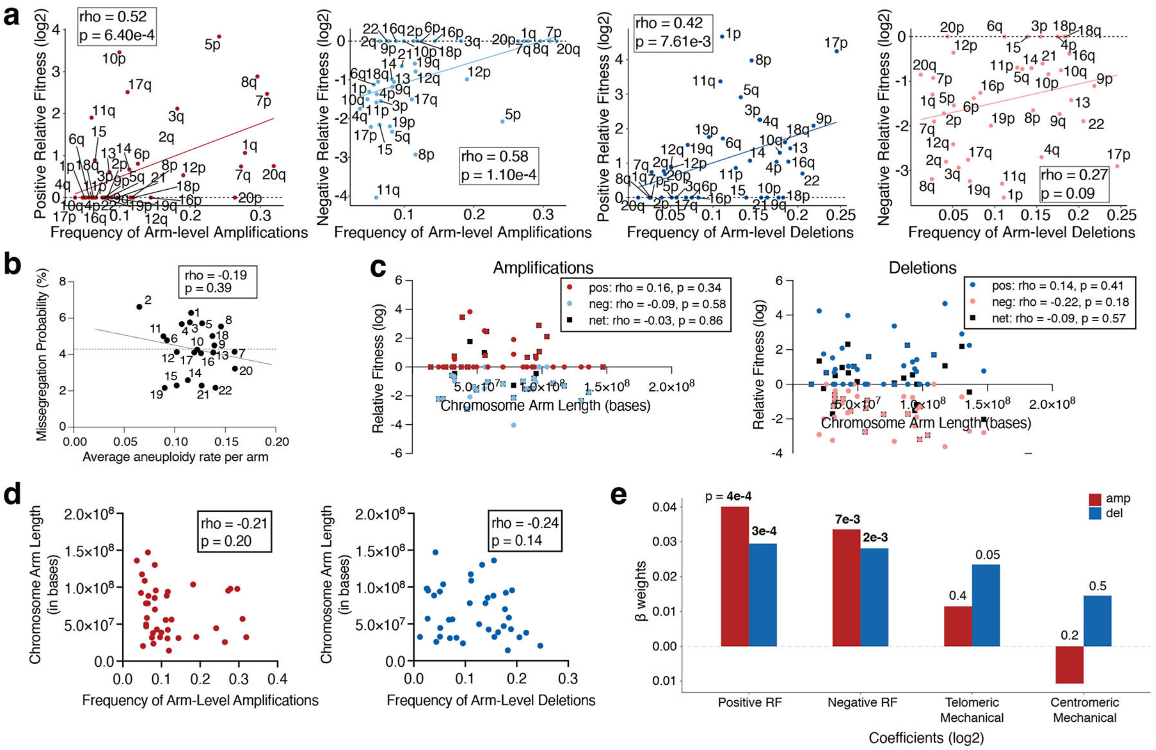
(c) Centromeric mechanical coefficients (log) plotted against total frequency of arm-SCNAs affecting a specific chromosome (i.e. amplifications and deletions of the p and q arms in aggregate). Acrocentric chromosomes are excluded from analysis.

(d) From the original BISCUT analysis: telomeric mechanical coefficients (log) plotted against telomere length, in RTLU, for amplifications (left; in red) and deletions (right; in blue).

(e) From the original BISCUT analysis: telomeric mechanical coefficients (log) plotted against frequency of arm-level amplifications (left; in red) and deletions (right; in blue).

For all panels, two-tailed p-values and rho correlation coefficients were calculated using Spearman's rank correlation. No adjustments were made for multiple comparisons.

Extended Data



Extended Data Figure 8: Telomeric mechanical pressures are better reflected when using baseline ploidies of 2 or 4.

(a) Relative fitness (log) plotted against frequency of arm-SCNAs. From left to right: positive selection in amplifications (dark red), negative selection in amplifications (light blue), positive selection in deletions (dark blue), and negative selection in deletions (light red).

(b) Missegregation probability in percentage (determined by single-cell sequencing of RPE1-hTERT non-transformed cells[42]) plotted against frequency of all arm-SCNAs affecting each chromosome, averaged across arms. The horizontal black line at 4.3% reflects the expected random chance of missegregation of each chromosome.

(c) Relative fitness (log) plotted against arm length.

(d) Chromosome arm length plotted against frequency of arm-SCNAs.

(e) Strength of correlation (β; vertical axis) between various coefficients (horizontal axis) and arm-SCNA rates from a multivariate Generalized Linear Model (GLM), with p-values above each predictor (significant values in bold). Amplifications are in red, and deletions are in blue.

All p-values in this figure were calculated using Spearman's correlation except as otherwise noted; no adjustments were made for multiple comparisons.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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 also thank Geoff Macintyre for suggesting the stick-breaking model to explain partial-SCNA length distributions, Iñigo Martincorena for personal communications regarding negative selection in cancer, Ed Chan for the *WRN* overexpression construct, and Michael Huang for the name of the main method. For flow cytometry analyses, we acknowledge NIH P30CA013696 (HICCC flow cytometry core) and for RNA-sequencing of cell lines we acknowledge the Sulzberger Genome Center. We would also like to acknowledge the following funding sources: the National Institutes of Health and National Cancer Institute (R.B., A.M.T.), National Institute for General Medical Sciences (A.M.T.), Fund for Innovative Cancer Informatics (R.B.), Gray Matters Brain Cancer Foundation (R.B.), Pediatric Brain Tumor Foundation (R.B.), Brain Tumour Charity (R.B.), and St. Baldrick's Foundation (R.B.).

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 DNA and RNA sequencing data generated in this study are available at the NIH Sequence Read Archive (SRA), Accession PRJNA976303.

CODE AVAILABILITY

The code used to merge copy-number segments, call partial-SCNAs, detect loci under selection, and determine relative fitness values and mechanical coefficients are freely available for download at https://github.com/beroukhim-lab/BISCUT-py3, DOI: 10.5281/zenodo.7896522

REFERENCES

1. Weaver BA & Cleveland DW Does aneuploidy cause cancer? Curr. Opin. Cell Biol 18, 658–667 (2006). [PubMed: 17046232]
2. Taylor AM et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell 33, 676–689 e3 (2018). [PubMed: 29622463]
3. 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). [PubMed: 18089652]
4. Holland AJ & Cleveland DW Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nat. Rev. Mol. Cell Biol 10, 478–487 (2009). [PubMed: 19546858]
5. Sheltzer JM et al. Single-chromosome Gains Commonly Function as Tumor Suppressors. Cancer Cell 31, 240–255 (2017). [PubMed: 28089890]
6. Pavelka N et al. Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. Nature 468, 321–325 (2010). [PubMed: 20962780]
7. Ly P et al. Characterization of aneuploid populations with trisomy 7 and 20 derived from diploid human colonic epithelial cells. Neoplasia 13, 348–357 (2011). [PubMed: 21472139]
8. Rutledge SD et al. Selective advantage of trisomic human cells cultured in non-standard conditions. Sci. Rep 6, 22828 (2016). [PubMed: 26956415]
9. Sunshine AB et al. The fitness consequences of aneuploidy are driven by condition-dependent gene effects. PLoS Biol 13, e1002155 (2015). [PubMed: 26011532]
10. Ravichandran MC, Fink S, Clarke MN, Hofer FC & Campbell CS Genetic interactions between specific chromosome copy number alterations dictate complex aneuploidy patterns. Genes Dev 32, 1485–1498 (2018). [PubMed: 30463904]
11. Hughes TR et al. Widespread aneuploidy revealed by DNA microarray expression profiling. Nat. Genet 25, 333–337 (2000). [PubMed: 10888885]

12. Liu G et al. Gene Essentiality Is a Quantitative Property Linked to Cellular Evolvability. Cell 163, 1388–1399 (2015). [PubMed: 26627736]
13. Salehi S et al. Clonal fitness inferred from time-series modelling of single-cell cancer genomes. Nature 595, 585–590 (2021). [PubMed: 34163070]
14. Kimmel GJ et al. Intra-tumor heterogeneity, turnover rate and karyotype space shape susceptibility to missegregation-induced extinction. PLoS Comput. Biol 19, e1010815 (2023). [PubMed: 36689467]
15. Lee AJX et al. Chromosomal instability confers intrinsic multidrug resistance. Cancer Res 71, 1858–1870 (2011). [PubMed: 21363922]
16. Cai Y et al. Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. Cancer Cell 29, 751–766 (2016). [PubMed: 27165746]
17. 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). [PubMed: 28986519]
18. Mermel CH 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). [PubMed: 21527027]
19. Brennan CW et al. The somatic genomic landscape of glioblastoma. Cell 155, 462–477 (2013). [PubMed: 24120142]
20. Cimini D Merotelic kinetochore orientation, aneuploidy, and cancer. Biochim. Biophys. Acta 1786, 32–40 (2008). [PubMed: 18549824]
21. Blackford AN & Stucki M How Cells Respond to DNA Breaks in Mitosis. Trends Biochem. Sci 45, 321–331 (2020). [PubMed: 32001093]
22. Martincorena I et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 171, 1029–1041 e21 (2017). [PubMed: 29056346]
23. Hoadley KA 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). [PubMed: 29625048]
24. Beroukhim R et al. The landscape of somatic copy-number alteration across human cancers. Nature 463, 899–905 (2010). [PubMed: 20164920]
25. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020). [PubMed: 32025007]
26. 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). [PubMed: 22566646]
27. Chan EM et al. WRN helicase is a synthetic lethal target in microsatellite unstable cancers. Nature 568, 551–556 (2019). [PubMed: 30971823]
28. Bailey MH et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 173, 371–385 e18 (2018). [PubMed: 29625053]
29. Ciriello G et al. Emerging landscape of oncogenic signatures across human cancers. Nat. Genet 45, 1127–1133 (2013). [PubMed: 24071851]
30. Nichols CA et al. Loss of heterozygosity of essential genes represents a widespread class of potential cancer vulnerabilities. Nat. Commun 11, 2517 (2020). [PubMed: 32433464]
31. McFarland JM et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat. Commun 9, 4610 (2018). [PubMed: 30389920]
32. Dempster JM et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. Genome Biol 22, 343 (2021). [PubMed: 34930405]
33. Davoli T et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell 155, 948–962 (2013). [PubMed: 24183448]
34. Dewhurst SM et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer Discov 4, 175–185 (2014). [PubMed: 24436049]
35. López S et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. Nat. Genet 52, 283–293 (2020). [PubMed: 32139907]

36. Zack TI et al. Pan-cancer patterns of somatic copy number alteration. Nat. Genet 45, 1134–1140 (2013). [PubMed: 24071852]
37. Liu Y et al. Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. Nat. Commun 8, 1212 (2017). [PubMed: 29089484]
38. Hose J et al. Dosage compensation can buffer copy-number variation in wild yeast. Elife 4, (2015).
39. Stenberg P et al. Buffering of segmental and chromosomal aneuploidies in Drosophila melanogaster. PLoS Genet 5, e1000465 (2009). [PubMed: 19412336]
40. Muenzner J et al. The natural diversity of the yeast proteome reveals chromosome-wide dosage compensation in aneuploids. bioRxiv 2022.04.06.487392 (2022) doi:10.1101/2022.04.06.487392.
41. Dumont M et al. Human chromosome-specific aneuploidy is influenced by DNA-dependent centromeric features. EMBO J 39, e102924 (2020). [PubMed: 31750958]
42. Klaasen SJ et al. Nuclear chromosome locations dictate segregation error frequencies. Nature 607, 604–609 (2022). [PubMed: 35831506]
43. Bersani F et al. Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. Proc. Natl. Acad. Sci. U. S. A 112, 15148–15153 (2015). [PubMed: 26575630]
44. Quinodoz SA et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell 174, 744–757.e24 (2018). [PubMed: 29887377]
45. Kitagawa K & Hieter P Evolutionary conservation between budding yeast and human kinetochores. Nat. Rev. Mol. Cell Biol 2, 678–687 (2001). [PubMed: 11533725]
46. Barra V & Fachinetti D The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. Nat. Commun 9, 4340 (2018). [PubMed: 30337534]
47. Knutsen T et al. Definitive molecular cytogenetic characterization of 15 colorectal cancer cell lines. Genes Chromosomes Cancer 49, 204–223 (2010). [PubMed: 19927377]
48. Nurk S et al. The complete sequence of a human genome. Science 376, 44–53 (2022). [PubMed: 35357919]
49. Altemose N et al. Complete genomic and epigenetic maps of human centromeres. Science 376, eabl4178 (2022). [PubMed: 35357911]
50. Gregan J, Polakova S, Zhang L, Tolić-Nørrelykke IM & Cimini D Merotelic kinetochore attachment: causes and effects. Trends Cell Biol 21, 374–381 (2011). [PubMed: 21306900]
51. 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). [PubMed: 11675499]
52. Wise JL et al. Human telomere length correlates to the size of the associated chromosome arm. PLoS One 4, e6013 (2009). [PubMed: 19547752]
53. McCarroll SA et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet 40, 1166–1174 (2008). [PubMed: 18776908]
54. Korn JM et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet 40, 1253–1260 (2008). [PubMed: 18776909]
55. Cancer Genome Atlas Research, Network. Integrated genomic analyses of ovarian carcinoma. Nature 474, 609–615 (2011). [PubMed: 21720365]
56. Tabak B et al. The Tangent copy-number inference pipeline for cancer genome analyses. bioRxiv 566505 (2019).
57. Olshen AB, Venkatraman ES, Lucito R & Wigler M Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 557–572 (2004). [PubMed: 15475419]
58. Carter SL et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol 30, 413–421 (2012). [PubMed: 22544022]
59. Li Y et al. Patterns of somatic structural variation in human cancer genomes. Nature 578, 112–121 (2020). [PubMed: 32025012]
60. Gao Q et al. Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. Cell Rep. 23, 227–238 e3 (2018). [PubMed: 29617662]
61. 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). [PubMed: 30355485]

62. Delignette-Muller ML & Dutang C fitdistrplus: An R Package for Fitting Distributions. 2015 64, 34 (2015).

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

64. Miotto B, Ji Z & Struhl K Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. Proc. Natl. Acad. Sci. U. S. A 113, E4810–9 (2016). [PubMed: 27436900]

65. Sondka Z et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer 18, 696–705 (2018). [PubMed: 30293088]

66. Endres DM & Schindelin JE A new metric for probability distributions. IEEE Trans. Inf. Theory 49, 1858–1860 (2003).

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

68. Lawrence M et al. Software for computing and annotating genomic ranges. PLoS Comput. Biol 9, e1003118 (2013). [PubMed: 23950696]

69. Gel B et al. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics 32, 289–291 (2016). [PubMed: 26424858]

70. Kim SY, Jacob L & Speed TP Combining calls from multiple somatic mutation-callers. BMC Bioinformatics 15, 154 (2014). [PubMed: 24885750]

71. Kasar S et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat. Commun 6, 8866 (2015). [PubMed: 26638776]

72. Kim J et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet 48, 600–606 (2016). [PubMed: 27111033]

73. Knijnenburg TA et al. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Rep 23, 239–254.e6 (2018). [PubMed: 29617664]

74. Lundberg AS et al. Immortalization and transformation of primary human airway epithelial cells by gene transfer. Oncogene 21, 4577–4586 (2002). [PubMed: 12085236]

75. Adalsteinsson VA et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat. Commun 8, 1324 (2017). [PubMed: 29109393]

76. DePristo MA et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet 43, 491–498 (2011). [PubMed: 21478889]

77. Van der Auwera G & O'Connor B Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. (O'Reilly Media, 2020).

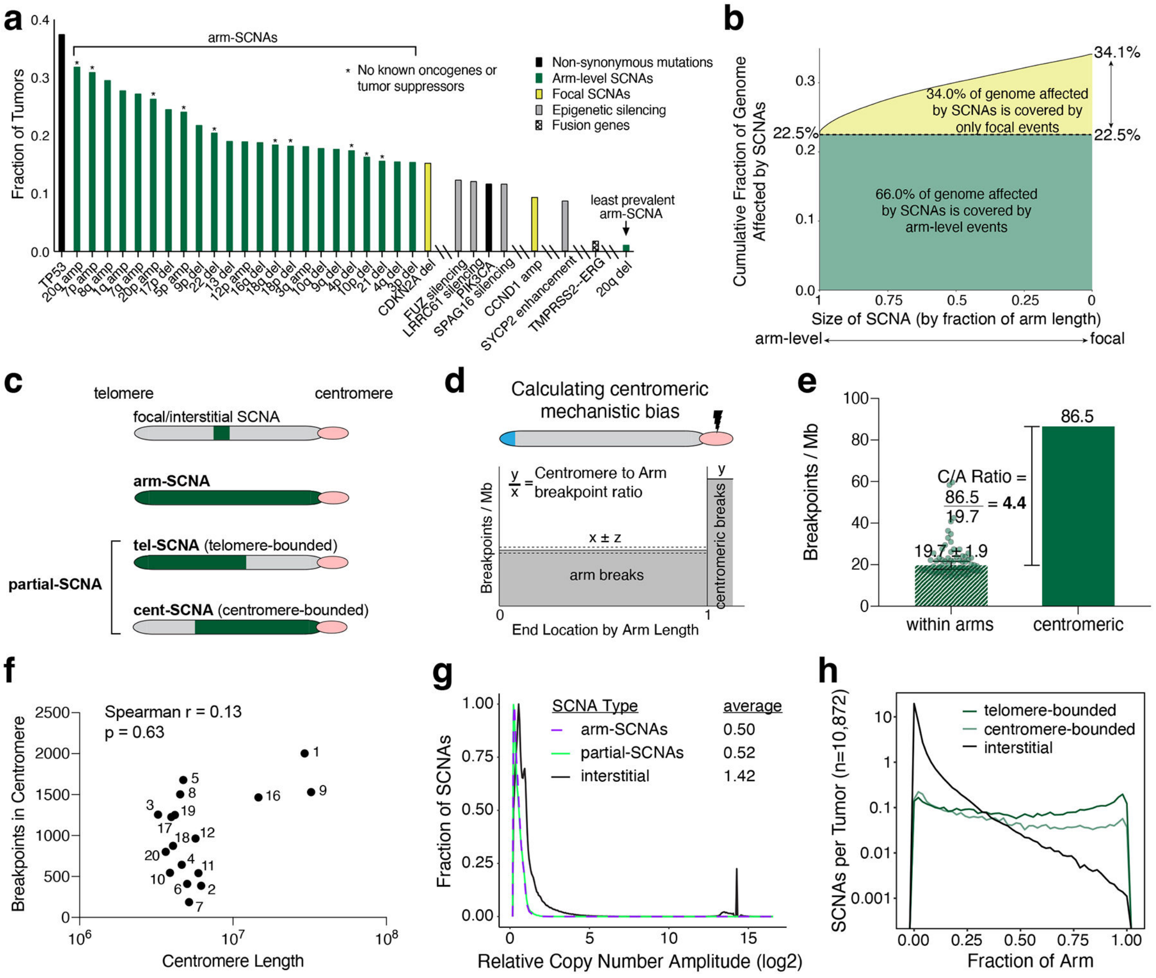


Figure 1: Prevalence and characteristics of different types of SCNAs.

(a) Fraction of TCGA tumors exhibiting frequent somatic genetic alterations. Asterisks indicate arms-SCNAs without known drivers.

(b) Cumulative fraction of cancer genomes affected by SCNAs (y-axis), plotted inversely by size of SCNAs (x-axis). The green region represents the fraction of genome covered by arm-SCNAs, and the yellow region represents the fraction of genome additionally covered by focal SCNAs.

(c) Classes of SCNAs referenced throughout this manuscript. DNA that has undergone copy-number change is colored green.

(d) Schematic representation of centromeric mechanical 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).

(e) Mean breakpoint density within chromosome arms, aggregated across all tumors and all chromosome arms (n = 67; binned by Mb), 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. Two-tailed p-value was calculated using Spearman’s correlation.

(g) 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.

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

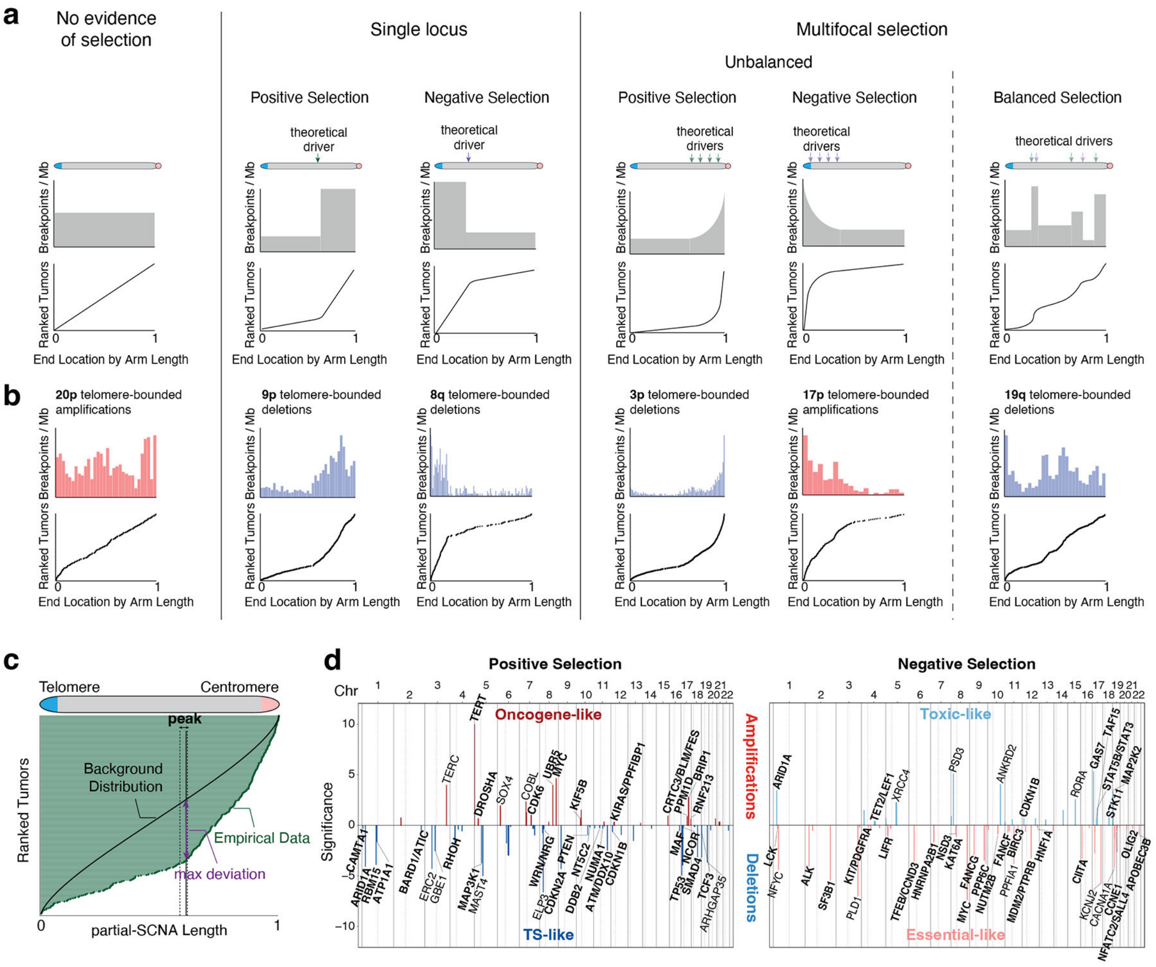


Figure 2: BISCUT identifies known and novel cancer driver genes through analysis of SCNA length distributions.

(a) Daaifferent patterns of SCNA-mediated selection.

(b) Empirical examples of SCNA-mediated selection from the pan-cancer dataset.

(c) BISCUT’s peak-finding function. Tumors (dark green) are ranked along the y-axis by partial-SCNA length. The location at which the empirical data deviates maximally from the background distribution is determined (purple). A peak region encompassing this location (denoted by dashed lines) is calculated; see Methods.

(d) Statistically significant peaks conferring selection as determined by BISCUT are plotted along the genome. The vertical axis indicates the Significance Score, representing 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 in bold.

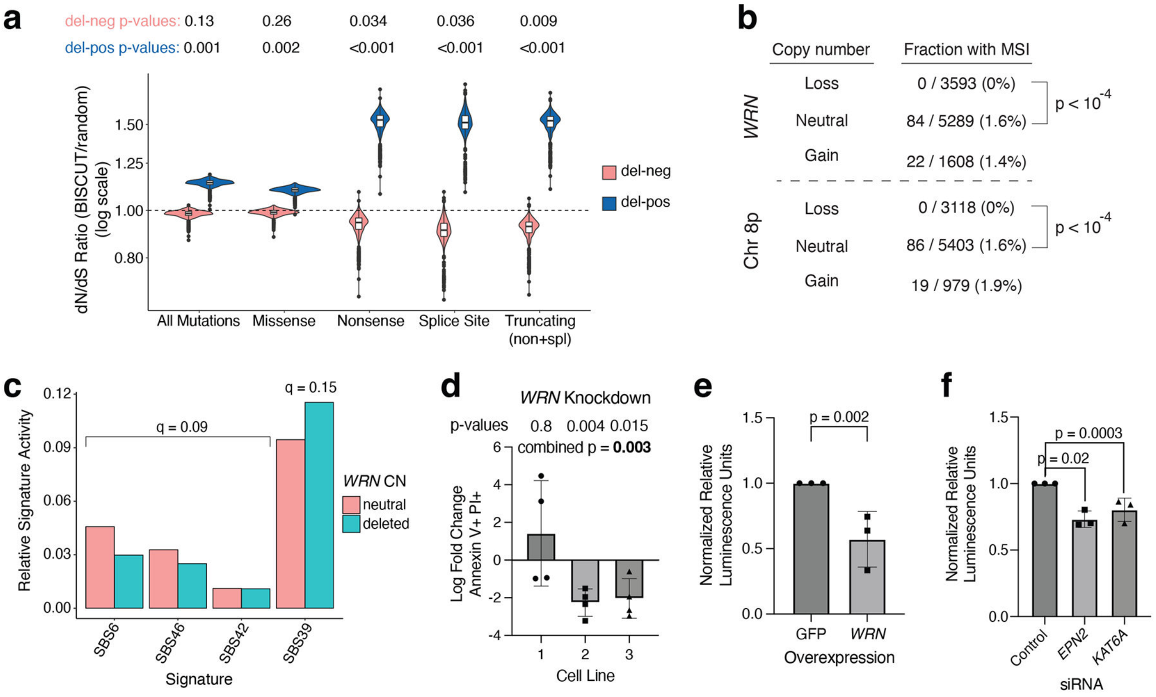


Figure 3: Validation of genes identified by BISCUT for negative and positive selection.

(a) Ratios of dN/dS scores (ratio of non-synonymous to synonymous mutations) between genes in restricted BISCUT del-neg (light red) or del-pos peaks (dark blue), compared to 1000 randomly selected sets of other genes. Comparisons between different types of single nucleotide variants (SNV) are indicated on the horizontal axis. Two-tailed p-values are derived from comparisons between observed and permuted data. Box plots center on median values and extend to the first and third quartiles; the whiskers extend to 1.5 times the interquartile range.

(b) Fraction of TCGA tumors with microsatellite instability (MSI) for different *WRN* or 8p copy-number status. Two-tailed p-values were calculated using Fisher's exact test.

(c) Relative COSMIC mutational signature activity of *WRN* copy-neutral versus *WRN* deleted TCGA tumors. Four statistically significant comparisons are shown, as determined by Mann-Whitney U test and a false discovery rate (q-value) cut-off of 0.2.

(d) Log-fold changes in apoptotic cells detected by flow cytometry for Annexin V and propidium iodide (PI) across three 8p wild-type cell lines (n = 4), on day three after transfection with *WRN* versus control siRNAs. Each point represents a biological replicate.

(e) Cell viability measured using Cell-Titer Glo on day 5 of overexpression of GFP or *WRN*.

(f) Cell viability (Cell-Titer Glo) measured three days after siRNA knockdown of the indicated genes.

In d-f, Student's t-tests were used to calculate one-tailed p-values for each independent experiment, and Fisher's method was used to combine values across the three experiments. For e and f, one representative experiment is shown from three total.

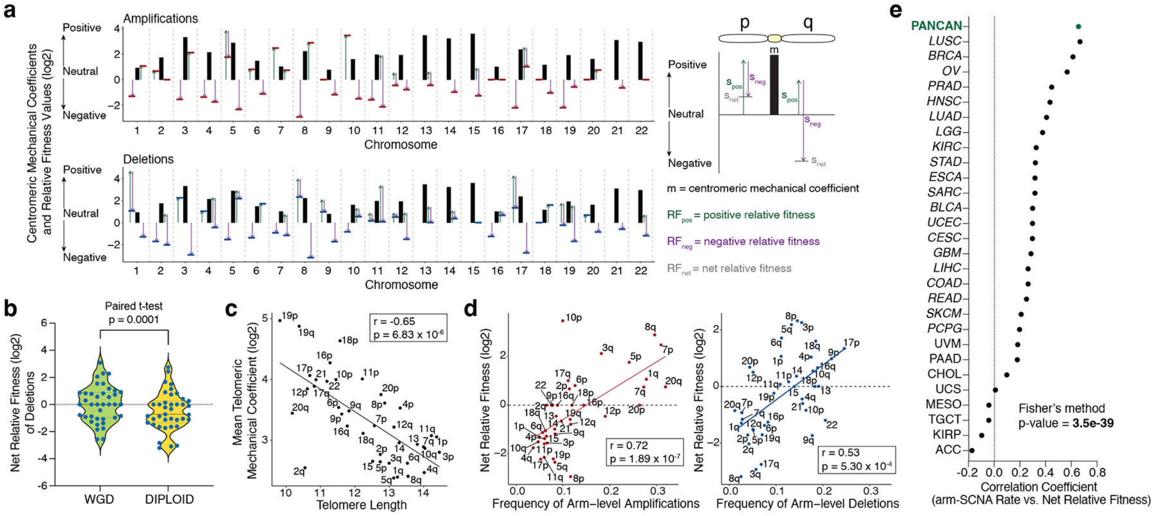


Figure 4: Pan-cancer mechanical coefficients and relative fitness (RF).

(a) Mechanical coefficients for each centromere and relative fitness values for amplifications and deletions of each chromosome arm, both reported as log2 values. Black bars represent centromeric mechanical coefficients. Red and blue horizontal lines represent net relative fitness 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). Relative fitness for both p and q arms are depicted to the left and right of the centromeric mechanical coefficient respectively.

(b) Log net relative fitness of deletions are significantly lower in diploid samples (mean = −0.49) than in WGD samples (mean = 0.04). Each dot represents a chromosome arm (n = 39 in each column), and two-tailed p-value is calculated using a paired t-test.

(c) Log telomeric mechanical coefficients (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)[52].

(d) Log net relative fitness 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.

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

For c-e, p-values were calculated using two-tailed Spearman's correlation except as otherwise noted; no adjustments were made for multiple comparisons.