Single-cell genomic variation induced by mutational processes in cancer

How cell-to-cell copy number alterations that underpin genomic instability in human cancers drive genomic and phenotypic variation, and consequently the evolution of cancer, remains understudied. Here, by applying scaled single-cell whole-genome sequencing to wild-type, TP53-deficient and TP53-deficient; BRCA1-deficient or TP53-deficient; BRCA2-deficient mammary epithelial cells (13,818 genomes), and to primary triple-negative breast cancer (TNBC) and high-grade serous ovarian cancer (HGSC) cells (22,057 genomes), we identify three distinct ‘foreground’ mutational patterns that are defined by cell-to-cell structural variation. Cell- and clone-specific high-level amplifications, parallel haplotype-specific copy number alterations and copy number segment length variation (serrate structural variations) had measurable phenotypic and evolutionary consequences. In TNBC and HGSC, clone-specific high-level amplifications in known oncogenes were highly prevalent in tumours bearing fold-back inversions, relative to tumours with homologous recombination deficiency, and were associated with increased clone-to-clone phenotypic variation. Parallel haplotype-specific alterations were also commonly observed, leading to phylogenetic evolutionary diversity and clone-specific mono-allelic expression. Serrate variants were increased in tumours with fold-back inversions and were highly correlated with increased genomic diversity of cellular populations. Together, our findings show that cell-to-cell structural variation contributes to the origins of phenotypic and evolutionary diversity in TNBC and HGSC, and provide insight into the genomic and mutational states of individual cancer cells.

The identification and characterization of endogenous mutational processes has transformed our understanding of cancer genomes, and have led to improved prognostic and therapeutic stratification of cancers with genomic instability. However, mutational processes are typically inferred from bulk whole-genome sequencing (WGS), which yields aggregate signals from pools of DNA composed of millions of cells. Thus, contemporaneous post-mitotic cell-to-cell variation due to genomic instability is not detectable in bulk sequencing, and has been understudied. Single-cell WGS can readily decompose clone-specific and cellular genomic events, enabling the calculation of copy number alteration (CNA) and structural variation (SV) accrual rates and mutational patterns over thousands of individual cells. This allows for the separation of evolutionary vestigial events, which are present in initial clonal expansions, from contemporaneous ‘foreground’ events, which reflect ongoing mechanisms of cell-to-cell genomic diversification. For example, breakage–fusion–bridge cycles (BFBCs) and homologous recombination deficiency (HRD) are endogenous mutational processes that accrue SVs with specific patterns including tandem duplications, interstitial deletions and fold-back inversions (FBIs) that generate high-level copy number amplifications. Because HRD and BFBCs...
are predicted to induce cell-specific structural changes on individual maternal or paternal alleles, a haplotype-specific analysis is essential for a comprehensive account of genome-scale structural variation. Here we combine single-cell approaches with haplotype-specific analysis to reveal how different mutational processes diversify the genomes of individual cancer cells and thereby determine phenotypic variation and evolutionary selection in human tumours. We apply scaled single-cell WGS and haplotype-specific analysis to an in vitro cell line system with experimentally induced HRD-associated genomic instability and human breast and ovarian tumours defined by SV-associated mutational processes\textsuperscript{4,10,12,14-17}. Our study reveals three sources of cell-to-cell variation in cancer genomes, with implications for interpreting phenotypic diversity and evolutionary selection in cancers with genomic instability.

**Induced single-cell genomic instability**

We first developed a combined experimental and computational approach for studying genome-scale cell-to-cell variation in human cells, by establishing an in vitro isogenic system of breast epithelium approach for studying genome-scale cell-to-cell variation in human tumours. We apply scaled single-cell WGS and haplotype-specific analysis to an in vitro cell line system with experimentally induced HRD-associated genomic instability and human breast and ovarian tumours defined by SV-associated mutational processes\textsuperscript{4,10,12,14-17}. Our study reveals three sources of cell-to-cell variation in cancer genomes, with implications for interpreting phenotypic diversity and evolutionary selection in cancers with genomic instability.

We first developed a combined experimental and computational approach for studying genome-scale cell-to-cell variation in human cells, by establishing an in vitro isogenic system of breast epithelium approach for studying genome-scale cell-to-cell variation in human tumours. We apply scaled single-cell WGS and haplotype-specific analysis to an in vitro cell line system with experimentally induced HRD-associated genomic instability and human breast and ovarian tumours defined by SV-associated mutational processes\textsuperscript{4,10,12,14-17}. Our study reveals three sources of cell-to-cell variation in cancer genomes, with implications for interpreting phenotypic diversity and evolutionary selection in cancers with genomic instability.

We first developed a combined experimental and computational approach for studying genome-scale cell-to-cell variation in human cells, by establishing an in vitro isogenic system of breast epithelium approach for studying genome-scale cell-to-cell variation in human tumours. We apply scaled single-cell WGS and haplotype-specific analysis to an in vitro cell line system with experimentally induced HRD-associated genomic instability and human breast and ovarian tumours defined by SV-associated mutational processes\textsuperscript{4,10,12,14-17}. Our study reveals three sources of cell-to-cell variation in cancer genomes, with implications for interpreting phenotypic diversity and evolutionary selection in cancers with genomic instability.
HLAMP, parallel allele alteration and serriform patterns of copy number breakpoints in the cellular population.

**Cell-level CNA variation in HGSC and TNBC**

On the basis of observing foreground mutational patterns defined by cell-to-cell variation, we next asked how the foreground event types distributed as a function of HRD and non-HRD mutational processes in TNBC and HGSC cancers. To identify appropriate patient tumour samples for this comparison, we first constructed a ‘meta-cohort’ of 309 patients comprising 170 patients with HGSC and 139 patients with TNBC with bulk tumour–normal paired WGS to infer the distribution of established mutational processes (106 TNBC and 22 HGSC genomes were newly sequenced for this study and combined with published...
HGSC\textsuperscript{12,30–32} and TNBC\textsuperscript{3,13,33–35} datasets (Extended Data Fig. 1). We applied a previously described correlated topic model machine learning approach (MMCTM)\textsuperscript{30} and recapitulated previously described groups of tumours. Distinct structural copy number mutational features in both TNBC and HGSC were observed as follows: HRD-Dup (enriched in small tandem duplications and \textit{BRCA1} mutations), HRD-Del (enriched in deletions, \textit{BRCA2} mutations), FBI (enriched in FBIs and \textit{CCNE1} amplification) and TD (enriched in large tandem duplications, \textit{CDK12} mutations) (Extended Data Figs. 6 and 7 and Supplementary Tables 4 and 5). Prognostic association of these groups in the meta-cohort of patients with HGSC was consistent with previous findings\textsuperscript{10,12} (\textit{P} = 0.0038; Extended Data Fig. 7d), with HRD-Del at a higher median survival than HRD-Dup, followed by FBI and TD with the worst median survival (Extended Data Fig. 7e; \textit{P} = 0.0022). We then selected 23 cases (16 HGSC and 7 TNBC) from the meta-cohort to assess the associations derived from patient-derived xenografts (PDXs), passaged over a multi-year period using subcutaneous engraftment\textsuperscript{33} (Extended Data Fig. 2c and Methods). DLP+ libraries from HRD-Dup (\textit{n} = 8), TD (\textit{n} = 3), and FBI (\textit{n} = 12) PDXs and patient tissues yielded a total of 22,057 genomes (median S56 per series), and a median of 1.96 million reads per genome (median 0.05× coverage, IQR 0.05; Extended Data Fig. 4 and Supplementary
Tables 2 and 3). Single-nucleotide variant (SNV) and SV mutational signature profiles that were inferred from DLP+ derived pseudobulk from the PDXs clustered with their bulk WGS counterparts (Extended Data Fig. 8a), indicating consistent mutational signature types without significant distortion of the signals from the original source tumour. In addition, SIGNALS analysis from DLP+ showed that the proportion of the genome identified as homozygous was highly correlated with VAFs of somatic mutations (Extended Data Fig. 8b), and that VAFs of somatic mutations were distributed as expected (Extended Data Fig. 8d). However, FBI cells exhibited higher overall rates of polyploidy (Fig. 3c; \(P = 0.02\)) and chromosomal missegregation relative to HRD-Dup (Fig. 3d; \(P = 0.0015\)). In addition, FBI tumours accrued gains at a significantly higher rate than did HRD-Dup tumours, with more skewing of the gain/loss ratio (4.9 versus 2.1; \(P = 0.04\); Fig. 3e and Extended Data Fig. 8e,f). This was more pronounced when considering the baseline ploidy of the tumours (\(P = 0.0012\); Extended Data Fig. 8g). Indeed, higher rates of polyploidy and segmental copy number gains may provide a greater opportunity for—and greater tolerance of—the large interstitial deletions that are found in some FBI cancers (Extended Data Fig. 7b). Pairwise HSCN distances between cells, reflecting cell-population diversity, yielded highly variable distributions across samples, ranging from a median value of 2 for the diploid TD case SA1047 to more than 123 for the pentaploid FBI case SA604.

---

**Fig. 3 | Single-cell genome properties of PDX models and patient tissues.**

a. Total copy number (top) and HSCN (bottom) profiles for two single cells from a case of TNBC HRD-Dup (SA501) and a case of HGSC FBI (SA1049). b. Total copy number heat map (left) and HSCN heat map (right) for 1,283 cells from SA1049.

c, d. Proportion of polyploid cells (c) and cells with missegregation events (d).

e. Ratio of chromosomal gains versus losses across different ploidy states and mutational signature groupings (number of cells shown below each violin plot). All box plots indicate the median, first and third quartiles (hinges), and the most extreme data points no farther than 1.5× the IQR from the hinge (whiskers).
HLAMP amplitude varies within FBI tumours

Next, we determined whether the CNA patterns that gave rise to single-cell variation in the cell lines could also explain cell-to-cell variation in the amplitude of HLAMPs in the tumours. We found extensive heterogeneity in the amplitude of HLAMPs across clonal populations within tumours, that would otherwise be obscured in bulk sequencing. By first focusing on a specific example, we assessed the phenotypic effect of a clone-specific HLAMP in the KRAS locus, present with average copy number 16.1 in a clone with 55 cells relative to a sibling clone composed of 230 cells that lacked the amplification (Fig. 4a). KRAS was differentially expressed between cell clusters from matched scRNA-seq data (maximum log-transformed fold change (logFC) = 0.346, q < 0.05; Fig. 4b), and immunohistochemistry for KRAS both in tissue from the primary patient and in PDX tissue corroborated a punctate pattern of expression across spatially separated regions within tumour sections (Fig. 4c). Thus, in a specific example, clone-specific HLAMP of an oncogene in a minor clone—otherwise not detectable with bulk methods—revealed co-associated clone-specific phenotypic variation. Across the dataset as a whole, FBI tumours had a 1.9-fold higher median HLAMP copy variance than did the other tumours ($P = 0.00096$; Fig. 4d,e), consistent with continual plasticity of HLAMP amplitude as a general property of FBI. Most events were less than 10 Mb in width (56%; Fig. 4f) and exhibited a distribution of maximum observed copy number with median 16.1 and IQR 8.7 (Fig. 4g). Furthermore, we noticed that amplitude variation in HLAMPs affected numerous other known oncogenes, including ERBB2 (DG1197), KIT (DG1197), KRAS (SA1049 and SA604), MYC (SA1184 and SA1051), CCNE1 (DG1134, SA1162 and SA604) and FGFR1 (SA1049 and SA355) (Fig. 4h). Notably, oncogenes with a variable copy number between cells and clones also exhibited greater variability in expression than did other genes, as measured by matched scRNA-seq ($P = 0.012$; Fig. 4i).

To determine the structural processes that lead to these events, we found that the rearrangement properties of variable HLAMPs were enriched for FBIs, consistent with BFBCs being a central mechanism of variable HLAMPs in FBI tumours. We also found clusters enriched for simple tandem duplications driving variable HLAMPs in HRD-Dup tumours (Methods and Extended Data Fig. 9a), providing further evidence for the different aetiological origin of these events in FBI and HRD-Dup tumours. This analysis also revealed that in many cases, clone-specific HLAMPs were part of complex genomic structures involving multiple chromosomes. For example, variable amplification around the CTNNB1 locus in SA1096 coincided with a translocation between chr. 3 and chr. 6 (Fig. 4j). Long-read single-molecule nanopore sequencing of the same samples validated the presence of this rearrangement (Extended Data Fig. 9d). Other examples of complex inter-chromosomal HLAMPs with orthogonal long-read sequencing included fixed non-variable amplification of CCNE1 in SA530 (chr. 4 and chr. 19), variable MYC amplification in SA1184 (chr. 3 and chr. 8) and amplification on q5 in SA1184 (Extended Data Fig. 9b,c and Supplementary Note).

Thus, cell-to-cell variability in HLAMP—which is not observable with bulk sequencing—is a pervasive mutational pattern that is most pronounced in FBI tumours, and consists of clone-specific complex rearrangements that influence phenotype through variable oncogene expression.

Haplotype-specific parallel evolution

We next investigated the extent of haplotype-specific parallel copy number evolution in tumours (Fig. 5). Phylogenetic tree analysis using breakpoints inferred from total copy number across the whole genome$^{16,17}$ (see Methods) revealed that in some cases, alleles segregated into distinct clades on the tree; for example, gains of 1q in SA1049 (Fig. 5a) and losses at the terminal end of chr. 10 in SA1053 (Fig. 5b). In other cases, gains and losses of different alleles were sporadic and were distributed more randomly across the tree, such as chr. 8 in SA1093 (Fig. 5c). Parallel copy number events were validated using VAFs of mutations found in these regions, in which—as expected—the VAF distribution inverted between two expected values, depending on allelic composition (Fig. 5d). We contend that in bulk sequencing, represented here by pseudobulk with mixtures (see Methods), the computed VAF no longer reflected the underlying copy number state in a heterogeneous mix of cells (Fig. 5e). We therefore suggest that accurate cancer cell fraction (CCF) inference, which depends on accurate VAF values, may be challenging in tumours with parallel copy number evolution.

We confirmed that parallel CNAs influence transcription with matched scRNA-seq. Inactivation of TP53 is invariably mediated by LOH of chr. 17 in these cancer types, and chr. 17 was indeed mono-allelically expressed across all tumours—in contrast to the hTERT wild-type cell line, which was used here as a control population (Fig. 5i). In addition, genes located at the terminal end of chr. 10 in SA1053 (Fig. 5b), were mono-allelically expressed in 100% of cells, with one cluster of cells expressing the B allele and another group of cells expressing the A allele (Fig. 5f,g). Across all data with matched scRNA-seq, mean BAF values per segment per sample measured in single-cell DNA sequencing (scDNA-seq) were strongly correlated ($R = 0.91, P < 10^{-5}$) with those measured in scRNA-seq (Fig. 5h), consistent with allele bias at the DNA level translating to consequent allele bias in expression.

Notably, nearly all tumours exhibited parallel CNA evolution. We classified genomic segments as parallel CNAs if more than 1% of cells had gain or loss of both the A and B alleles and assigned the clonality using total copy number as follows: clonal (CCF > 80%, as in Fig. 5a,b), subclonal (20% < CCF ≤ 80%) or rare (CCF ≤ 20%, as in Fig. 5c). Every tumour had at least one parallel CNA event, with most containing parallel CNAs at different clonalities (Fig. 5j). Across all samples, an average of 6% of clonal segments, 15% of subclonal segments and 7% of rare segments contained parallel CNAs, with a trend for higher event rates in FBI relative to HRD-Dup (Extended Data Fig. 8j,k; $P = 0.02$ for subclonal, $P > 0.05$ for clonal and rare). Motivated by the sporadic pattern of losses of both alleles observed in Fig. 5c, we then tested whether parallel CNAs due to losses were more common on a tetraploid versus a diploid background, using ancestral state reconstruction to estimate the event rate across the phylogenetic tree (see Methods). We found that on a diploid (1|1) background, parallel gains were more common than losses, but that on a tetraploid (2|2) background, parallel losses became more common than gains. This was true for whole chromosome, chromosome arm and segmental aneuploidies (Fig. 5j). The number of parallel CNAs was significantly correlated with both copy number and phylogenetic distances computed using total copy number (Fig. 5k,l). Thus, parallel copy number evolution was a pervasive feature, affecting the interpretation of somatic mutations, haplotype-specific expression and overall levels of genomic diversity in TNBC and HGSC tumours.

Increased CNA serriformity in FBI tumours

SSVs first identified in cell lines from single-cell WGS represent a structural mutation type that is not identifiable using bulk WGS, as the cell-to-cell variation in copy number breakpoints is obscured. We analysed the tumour DLP+ data for the presence of SSVs (Fig. 6a; heat maps 2–3 from left). SSVs, occurring at a megabase length scale, were distinct from small cell-to-cell variations in copy number breakpoint (Extended Data Fig. 8h). FBI tumours were more diverse than HRD-Dup or TD samples, with average HSCN distances of 71 (FBI), 46 (HRD-Dup) and 26 (TD) ($P = 0.047$ FBI versus HRD-Dup, $P = 0.031$ FBI versus TD; Extended Data Fig. 8i). Thus, considering whole-genome duplication, overall rates of segmental aneuploidy and the gain/loss ratio, FBI and HRD-Dup tumours showed markedly different patterns of CNA accrual at the single-cell level.
localization, which may occur owing to fluctuations in sequencing coverage rather than true changes in copy number (for example, Fig. 6a). Cell numbers indicate the number of cells used for consensus copy number evaluation. SSVs were also visible in single cells comprising the serration pattern (Fig. 6b). Additional confirmation of the SSV scale was obtained from allele phasing, in which the concomitant loss of heterozygosity was observed (Fig. 6b, bottom track). We computed serration scores per breakpoint event to identify the relative degree of variation in breakpoints across cells in each cancer. Scores were calculated as the fraction of event-containing cells with rare (less than 5% of cells) event breakpoint positions (see Methods) and with breakpoint regions that met size and prevalence criteria (≥20 Mb, that is, 40 genomic bins; ≥100 cells with breakpoint event) to permit the detection of positional and cell-to-cell variation. Variable cell-to-cell breakpoints were common, with 6.6% of regions having serration scores of 0.15 or higher (that is, 15% of cells or more have a rare event breakpoint position) across all cases, with FBL cases having the highest (12.1%), HRD-Dup cases the lowest (1.2%), and TD cases intermediate (10.5%) rates. Comparison of distributions of serration using a mixed effects linear model accounting for individual variation indicated that FBL cases had higher degrees of breakpoint variance in breakpoint regions ($P = 0.0096$; two-sided Wilcoxon test); **$P < 0.01$. Notches show individual data points. Width of genomic segment containing the amplification; dashed red line indicates a width of 10 Mb. Maximum cell copy number (CN) per gene. Clone maximum/minimum copy number ratio of cancer genes overlapping HLAMP regions. Genes across all cancer datasets with ratio > 2 are shown (n = 296). Colours as in d. For f, h, distributions of values are shown in a violin plot on the right. i, Distribution of the maximum log FC between gene expression clusters in matched scRNA-seq for variable oncogenes (n = 140) versus non-variable oncogenes (n = 159). $P = 0.019$ (two-sided Wilcoxon test). j, Consensus copy number profiles in two clones in SA1096 overlaid with lines indicating SVs. All box plots indicate the median, first and third quartiles (hinges), and the most extreme data points no farther than 1.5× the IQR from the hinge (whiskers).
Fig. 5 | Haplotype-specific parallel copy number evolution. a–c. Heat maps of chromosomes 1q (a), 10q (b) and 8 (e) ordered by a phylogenetic tree. The tips of the phylogeny are coloured according to the allelic phase of the region of interest. Arrows indicate single cells, the copy number profiles of which are shown below each heat map. d. VAFs of SNVs in parallel copy number events in two haplotype-specific states in which the dominant allele switches between the two states. Each point is the VAF of a single SNV; lines connect the same SNV in the two states. Dashed lines indicate the expected VAF on the basis of the proportion of cells in state 1|2. e. Correlation of the number of parallel copy number events with copy number distance (total copy number) across all datasets (n = 828) computed in RNA versus DNA. f. Scatter plot of mean BAF per segment across all datasets (n = 828) computed in RNA versus DNA. g. BAF distribution on chromosome 17 in all tumours and cell lines with matched scRNA (n = 21,347 cells DNA; n = 70,553 cells RNA) versus wild-type cell line (n = 1,963 cells DNA; n = 5,752 cells RNA). h. Rate of gains and losses within whole chromosomes (n = 35 events), chromosome arms (n = 31 events) and segments (n = 341 events) on diploid (1,1) and tetraploid (2,2) backgrounds. WGD, whole-genome duplication. i. Correlation of the number of parallel copy number events with copy number distance (P = 0.0008) (k) and phylogenetic distance (P = 0.0003) (l). Annotations at the top indicate the correlation coefficient (R) and P-value derived from a linear regression; shaded areas in plots show the 95% confidence interval of the linear regression. All box plots indicate the median, first and third quartiles (hinges), and the most extreme data points no farther than 1.5× the IQR from the hinge (whiskers).

Discussion
Our findings show that cell-to-cell variation at the level of structural and copy number alteration is a pervasive ‘foreground’ feature of TNBC and HGSC cancers that is exhibited against distinct endogenous mutational processes of genomic instability. Because CNAs can influence the expression levels of hundreds of genes, each of the foreground mutational patterns provides extensive and distinct genomic diversity upon which selection may act. Oncogenic HLAMPs are understood to be key drivers of tumour progression and are prognostic in HGSC when co-localized with FBIs22. Here we reveal an additional layer of complexity, finding that the amplitude of HLAMPs can vary substantially between cells. Although this has been recognized as a defining feature of extrachromosomal DNA amplifications30, we propose that it is also a general property of other classes of HLAMPs, such as those mediated by BFBCs and by complex inter-chromosomal rearrangement processes39. This has important implications for therapeutic strategies to target frequently altered oncogenes, as cancer types of high genomic instability may be predisposed to containing treatment-resistant clones. Multi-allelic variation within the same locus is also a highly prevalent feature of breast and ovarian cancers, consistent with some previous observations in other cancers22,30.
Notably, events that appeared clonal at the total copy number level were often composed of distinct clades with different alleles gained or lost; this might reflect evolutionary convergence for favourable karyotypes at the total copy number level, as shown by transcriptional phenotypic convergence. Evolutionary time series modelling is likely to further help to resolve patterns of phenotypic selection from parallel CNAs. We also highlight that sporadic gains and losses happen on both alleles, with rates increased on a whole-genome-doubled background relative to diploid, potentially reflecting increased fitness tolerance owing to genomic redundancy. Finally, megabase-scale copy length variation at a single-cell level (SSV) has been observed in vitro with cell-selected single-cell sequencing. Here we show with single-cell genome sequencing at the cell-population level that SSVs are in fact prevalent in TNBC and HGSC and distribute across clones within tumours. Although the underlying mechanisms that generate SSVs are unknown, they represent a new class of variation that may contribute to the structural copy evolution of tumours enriched in the FBI background and in polyploid genome states. We observed each of the foreground mutational patterns in all mutational processes, as such, FBI may comprise a distinct phenotypic class in which foreground mutational patterns generate diversity that could underlie poor prognostic significance. We conclude that SSVs in cancer genomes are a useful means to reveal hidden cellular states of structural copy number diversity in genomically unstable tumours. The data that we present here show that foreground...
mutational patterns are key determinants of genomically encoded phenotypic diversity and consequent 'evolvability' in cancer.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05249-0.

1. Umbreit, N. T. et al. Mechanisms generating cancer genome complexity from a single cell division error. Science 368, eaba0712 (2020).
2. Minussi, D. C. et al. Breast tumours maintain a reservoir of subclonal diversity during expansion. Nature 592, 302–306 (2021).
3. Laks, E. et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. Cell 179, 1207–1221 (2019).
4. Hakeem, R. DNA-damage repair, the good, the bad, and the ugly. EMBO J. 27, 589–605 (2008).
5. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. Nature 578, 112–121 (2020).
6. Macintyre, G. et al. Copy number signatures and mutational processes in ovarian carcinomas. Nat. Genet. 50, 1262–1270 (2018).
7. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–423 (2013).
8. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016).
9. Shirahashi, Y., Tremmel, G., Miyano, S. & Stephens, M. A simple model-based approach to inferring and visualizing cancer mutation signatures. PLoS Genet. 11, e1005657 (2015).
10. Funnell, T. et al. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. PLoS Comput. Biol. 15, e1006799 (2019).
11. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346–352 (2012).
12. Wang, Y. et al. Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. Nat. Genet. 49, 856–865 (2017).
13. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. 23, 517–525 (2017).
14. Staaf, J. et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nat. Med. 25, 1526–1533 (2019).
15. Zahn, H. et al. Scalable whole-genome single-cell library preparation without preamplification. Nat. Methods 14, 167–173 (2017).
16. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. Nature 578, 94–101 (2020).
17. Nguyen, L., W M Martins, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. Nat. Commun. 11, 5584 (2020).
18. Salehi, S. et al. Clonal fitness inferred from time-series modelling of single-cell cancer genomes. Nature 595, 585–590 (2021).
19. Burleigh, A. et al. A co-culture genome-wide RNA screen with mammary epithelial cells reveals transmembrane signals required for growth and differentiation. Breast Cancer Res. 17, 4 (2015).
20. Schwarz, R. F. et al. Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput. Biol. 10, e1003535 (2014).
21. Wu, C.-Y. et al. Integrative single-cell analysis of allele-specific copy number alterations reveals targetable vulnerabilities for difficult-to-treat tumors. Nature 565, 310 (2019).
22. Patch, A.-M. et al. Whole-genome characterization of chemoresistant ovarian cancer. Nature 521, 489–494 (2015).
23. Eirew, P. et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature 518, 422–426 (2015).
24. Shah, S. P. et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature 486, 395–399 (2012).
25. Savage, P. et al. Chemogenomic profiling of breast cancer patient-derived xenografts reveals targetable vulnerabilities for difficult-to-treat tumors. Commun. Biol. 3, 310 (2020).
26. Branton, D. et al. The potential and challenges of nanopore sequencing. Nat. Biotechnol. 26, 1146–1153 (2008).
27. Salehi, S. et al. Cancer phylogenetic tree inference at scale from 1000x of single cell genomes. Preprint at bioRxiv https://doi.org/10.1101/2020.05.06.058190 (2020).
28. Verhaak, R. G. W., Bafna, V. & Mischel, P. S. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nat. Rev. Cancer 19, 283–288 (2019).
29. Hadi, K. et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. Cell 183, 197–210 (2020).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2022

IMAX Consortium
Gregory J. Hannon1,2, Geoffrey Battistoni3, Dario Bressan1, Ian G. Cannell1, Hannah Cosbolt2, Cristina Jaussel2, Tatjana Kovacevic2, Claire M. Mulvey2, Fiona Nugent2, Marta Paez Ribes2, Isabella Pearson1, Fatime Qasmi2, Kirsty Sawicka2, Sophia A. Wild1, Elena Williams1, Samuel Aparicio3, Emma Laks4, Yangguang Liu5, Clara H. O’Flanagan6, Austin Smith7, Teresa Ruiz de Algarda8, Daniel Lai9, Andrew Roth10, Shankar Balasubramanian6,7, Maximilian Lee11, Bernd Bodenmiller12, Marcel Burger12, Laura Kuett1, Sandra Tietscher1, Jonas Windhager1, Edward S. Boyden1, Shahar Aloni4, Yi Cu5, Amauche Ememari6, Daniel R. Goodwin7, Emmanouil D. Karagiannis1, Anubhav Sinha1, Asmamaw T. Wassie1, Carlos Caldas3, Alejandra Bruna1, Maurizio Callari2, Wendy Greenwood9, Giulia Lerda7, Yariv Eyal-Lubling9, Oscar M. Rueda1, Abigail Sheal10, Owen Harris10, Robby Becker10, Flaminia Grimaldo10, Suvi Sara Vogi9, Johanna A. Joyce9, Spencer S. Watson9, Sohrab P. Shah11, Andrew McPherson1, Ignacio Vázquez-García2, Simon Tavaré12, Oscar Harris13, Khanh N. Dinh14, Eyal Fisher1, Russell Kunes15, Nicholas A. Walton16, Mohammed Al Sa’d17, Nick Chornay17, Ali Darwish17, Eduardo A. González-Solares17, Carlos González-Fernández17, Aybüke Küçükbül1, Nick Miller18, Xiaoxi Zhaung18,24,26, Jian Fan18,24,26, Hsuan Lee18,24,26, Leonardo A. Sepúlveda18,20, Chenglong Xia18,20, Pu Zheng18,20

1Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK. 2UBC Data Science Institute, University of British Columbia, Vancouver, British Columbia, Canada. 3Department of Chemistry, University of Cambridge, Cambridge, UK. 4School of Clinical Medicine, University of Cambridge, Cambridge, UK. 5Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland. 6McGovern Institute, Departments of Biological Engineering and Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. 7Department of Oncology and Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. 8Sgl Interactive, Dublin, Ireland. 9Department of Oncology and Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland. 10Herbert and Florence Irving Institute for Cancer Diagnostics, Columbia University, New York, NY, USA. 11New York Genome Center, New York, NY, USA. 12Institute of Astronomy, University of Cambridge, Cambridge, UK. 13Hughes Medical Institute, Harvard University, Cambridge, MA, USA. 14Department of Physics, Harvard University, Cambridge, MA, USA. 15Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA.
Methods

Generation and culture of human mammary epithelial cell lines

The wild-type human mammary epithelial cell line 184-hTERT L9 (SA039) and isogenic 184-hTERT TP53 knockouts (SA906) cell line, generated from 184-hTERT L9, were cultured as previously described18,19,40 in Mammary epithelial cell growth basal medium (MEMB) (Lonza) supplemented with the SingleQuots mix (Lonza), 5 µg ml−1 transferrin (Sigma-Aldrich) and 10 µM Lipo spondin (Sigma-Aldrich). Additional truncation mutations (Supplementary Table 4) of BRCA1 (SAI054: c. [427_441+3delGAATAATTCTTTGTTGAAACCTGGTTTCTTC]; [437_466+3delCCTTTGTTGAAACCTGGTTTCTTC]; [412_441+2delGTAGTTTCTTC]; [412_437delCGTAGTTTCTTC]); BRCA2 (SAI056: c.[6997delG];[6997_6998delGT]; SA1188; c.[6997_6998delGT]; SAI055; c.[3507_3522delinsGA];[3509_3520delinsT]) (hg19) were introduced by CRISPR–Cas9 nucleases (pX330 hSpCas9) with an rFP reporter gene using Mirus TransIT LT1 transfection (Mirus Bio). Clonal populations were generated by flow sorting and propagating single RFP-positive cells. Mutations were verified by TOPO cloning and Sanger sequencing of both alleles for genotypes, protein expression by western blotting and absence of off-target effects by sequencing of the top hits. SNV positions from Sanger sequencing data were annotated with information from GENCODE v.19 (ref. 41). Variant sequence and position was used to annotate variant calls with records from Clinvar 20200206_data_release_42 and COSMIC v. 91 (ref. 43).

Verification of mutations in the 184-hTERT cell line

Immunoblotting

184-hTERT cells were lysed directly in 1× Laemmli buffer supplemented with 7.5% β-mercaptoethanol and proteins were denatured at 95 °C for 15 min. Protein from 250,000 cells was resolved on a 4–15% acrylamide gel (Biorad) or 3–8% Tris-acetate acrylamide gel (Novex) and transferred to a nitrocellulose membrane with Towbin transfer buffer overnight at 30 V at 4 °C. Blots were blocked with 5% milk in TBST for 1 h and incubated overnight at 4 °C with mouse anti-p53 (Santa Cruz SC-126, 1:500 in 5% bovine serum albumin (BSA)), mouse anti-BRCA1 (Santa Cruz SC-6954, 1:200 in 5% BSA), mouse anti-BRCA2 (Millipore OP95, 1:200 in 5% BSA) and goat anti-GAPDH (SC-48166, 1:500 in 5% bovine serum albumin (BSA)). Blots were washed five times for 5 min in TBST and incubated with anti-goat HRP-conjugated secondary antibody (Abcam ab6721, 1:5000 in 5% BSA) or goat anti-GAPDH (SC-48166, 1:500 in 5% BSA). Blots were washed five times for 5 min and the signal was imaged using Immobilon Western Chemiluminescent HRP Substrate (MilliporeSigma, WBKLL20500) and the ImageQuant LAS 4000 (GE Healthcare) using the ImageQuant TL software.

Xenografting

Fragmentations of tumours from patients were chopped finely with scalpels and mechanically disaggregated for one minute using a Stomacher 80 Biomaster (Seward Limited) in 1 ml cold DMEM-F12 with glucose, l-glutamine and HEPES (Lonza 12–179). Two hundred microlitres of medium containing cells or organoids from the resulting suspension was used equally for transplantation in four mice. The remaining tissue fragments were cryopreserved viably in DMEM-F12 supplemented with 47% FBS and 6% dimethyl sulfoxide (DMSO). Tumours were transplanted in mice as previously described (Eirew) in accordance with SOP BBCCR 009. Female NOD/SCID/IL2rg−/− (NSG) and NOD/ Rag1−/−Il2rg−/− (NRG) mice were bred and housed at the Animal Resource Centre (ARC) at the British Columbia Cancer Research Centre. For subcutaneous transplants, mechanically disaggregated cells and clumps of cells were resuspended in 150–200 µl of a 1:1 v/v mixture of cold DMEM-F12:Matrigel (BD Biosciences). Female mice (8–12 weeks old) were anaesthetized with isoflurane and the mechanically disaggregated cell clump suspension was transplanted under the skin on the left flank using a 1-ml syringe and 21-gauge needle. Mice were housed at a 18–25 °C temperature range and 50–70% humidity range, with a 12-h daylight cycle (on at 06:00; off at 18:00). All animal experimental work was approved by the animal care committee (ACC) and animal welfare and ethical review committee at the University of British Columbia (UBC) under protocol A19–0298.

Verification of mutations in the 184-hTERT cell line

Tissue processing

Xenograft-bearing mice were euthanized when the size of the tumours approached 1,000 mm3 in volume, according to the limits of the experimental protocol. The tumour material was excised aseptically and processed as described for primary tumour. A section of tumour was fixed in 10% buffered formalin for 24 h, dehydrated in 70% ethanol and paraffin-embedded before duplicate 1-mm cores were used to generate tissue microarrays for staining and pathological review. Remaining tumour was finely chopped and gently paddle-blended, and released single cells and fragments were viably frozen in DMEM supplemented with 47% FBS and 6% DMSO.

Histopathology of PDX tumours

Deparaffinized 4-µm sections of tissue microarrays were stained with haematoxylin and eosin or KRAS (Lifespan Bioscience, LS-B4683, 1:50), performed using the Ventana Discovery XT platform and the UltraMap DAB detection kit. HGSC pathology was confirmed by an anatomical pathology resident at University of British Columbia, under the supervision of a certified staff pathologist.

WGS

Genomic DNA was extracted from frozen tissue fragments using the DNeasy Blood and Tissue kit (Qiagen) and constructed libraries for whole genomes of 309 tumour–normal pairs were sequenced on the Illumina HiSeqX according to Illumina protocols, generating 100-bp paired-end reads for an estimated coverage of sequencing between 40× (normal) and 80× (tumour). Sequenced reads were aligned to the human reference GRCh37 (hg19) using BWA-MEM.
Long-read sequencing

High-molecular weight (HMW) DNA was extracted from fresh and/or frozen tissue fragments using the MagAttract HMW DNA Kit (Qiagen) and size-selected using Blue Pippin for single long-molecule sequencing on the PromethION (Oxford Nanopore Technologies).

Generation of single-cell suspensions and nuclei for scDNA-seq

Viable frozen aliquots of patient tissues and PDX tumours were thawed and either homogenized and lysed using Nuclei EZ Buffer (Sigma) or enzymatically dissociated using a collagenase/hyaluronidase 1:10 (10×) enzyme mix (STEMCELL Technologies), as described previously. Cells and nuclei were stained with CellTrace CFSE (Life Technologies) and LIVE/DEAD Fixable Red Dead Cell Stain (Thermo Fisher Scientific) in a 0.04% BSA/PBS (Miltenyi Biotec 130-091-376) incubated at 37°C for 20 min. Cells were pelleted and resuspended in 0.04% BSA/PBS. This single-cell suspension was loaded into a contactless piezoelectric dispenser (Cellenone or sciFLEXARRAYER S3, Scienion) and spotted into open nanowell arrays (SmartChip, TakaraBio) preprinted with unique dual index sequencing primer pairs. Occupancy and cell state were confirmed by fluorescent imaging and wells were selected for single-cell copy number profiling using the DLP+ method. In brief, cell dispensing was followed by enzymatic and heat lysis. After cell lysis, tagmentation mix (14.335 nl TD buffer, 3.5 nl TDE1 and 0.165 nl 10% Tween-20) in PCR water was dispensed into each well followed by incubation and neutralization. For BCA+ cells, the tagmentation mix consisted of 10 nl TB1 buffer and 10 nl BT1 enzyme without Tween-20 in PCR water. Final recovery and purification of single-cell libraries was done after eight cycles of PCR. Pooled single-cell libraries were analysed using the Agilent Bioanalyzer 2100 HS kit.

Generation of 10X scRNA-seq data

The 184-hTERT cells were pelleted and gently resuspended in 200 μl PBS followed by 800 μl 100% methanol and incubation at −20°C for 30 min to fix, dehydrate and shrink cells. PDX tumour fragments were dissociated into single cells using collagenase/hyaluronidase at 37°C for 2 h for TNBC tumours or with cold active Bacillus licheniformis (Creative Enzymes NATE0633) in PBS supplemented with 5 mM CaCl2 and 125 U ml−1 DNase for HGSC tumours, as described previously. Cells were then pelleted and resuspended in 0.04% BSA/PBS and immediately loaded onto a 10X Genomics Chromium single-cell controller targeting 3,000 cells for recovery. Libraries were prepared according to the 10X Genomics Chromium single-cell protocol standard protocol. Libraries were then sequenced on an Illumina Nextseq500/550 with 42-bp paired-end reads or HiSeq2500 v4 with 125-bp paired-end reads. 10X Genomics Cell Ranger 3.0.2 was used to perform demultiplexing, counting and alignment to GRCh38 and mm10.

Processing of bulk whole-genome data

SNV and SV calls for 121 HGSC samples were acquired from a previous study. For new samples, reads were aligned to the hg19 reference genome using BWA-MEM. Processing proceeded as per the aforementioned study to maintain consistency. SNVs were called with MutationSeq (probability threshold = 0.9) and Strelka27. The intersection of calls from these methods were retained; however, SNVs falling in blacklist regions were removed. The blacklist regions include the UCSC Genome Browser Duke and DAC blacklists, and those in the CRG Alignability 36mer track that had more than two mismatched nucleotides. SNVs were then annotated with OncoKB for variant impact. SVs were called using deStruct and LUMPY, and breakpoints called by both methods were retained. We then filtered events with the following criteria: any breakpoints falling in the blacklist described above; ≤30-bp inter-breakpoint distance; ≤1,000-bp deletion; any breakpoints with fewer than 5 supporting reads in the tumour sample or any read support in the matched normal sample.

Gene mutation enrichment analysis was performed using the hyper-geometric test for SNVs, amplifications and deep deletions separately, comparing each signature stratum to all other samples.

Nanopore data analysis

For Nanopore sequence data, base calling and read alignment were performed using Guppy v.3 and Minimap2, respectively. Reads that were likely to be derived from mouse were filtered by first aligning to a concatenated hg19 and mm10 reference, removing reads with alignments to mm10 and re-aligning the remaining reads to hg19. Signal artefact regions as well as alignments with mapping quality of less than 60 were excluded from the final alignments. Alignments were then phased using the PEPPER-Margin-DeepVariant pipeline, after which WhatsHap was used to tag reads in the filtered alignments using phasing information. SV calling was performed using Sniffles v.1.0.12 and cuteSV v.1.0.11 with 5 read support, and subsequently merged using SURVIVOR for a union set of predicted variants. Alignments and variants were visualized using IGV, Ribbon and the karyoplotE R package.

HGSC and TNBC meta-cohort signature analysis

Signature analysis was performed according to a previous study. The MMCTM model was run on the sample SNV and SV count matrices. The number of signatures to estimate in the HGSC and TNBC integrated cohort was chosen by running the above fitting procedure for k = 2–16 for both SNV and SV signatures with the number of restarts set to 500, in which k is the number of signatures. We performed this step on approximately half the mutations in each sample, then computed the average per-mutation log likelihood on the other held-out half of the mutations. The elbow curve method on log-likelihood values was used to select the final number of signatures to fit to the entire dataset.

To estimate MMCTM parameters on the full dataset, a hyper-parameters were set to 0.1. The model was initially fit to the data 1,500 times. Each restart was run for a maximum of 1,000 iterations or until the relative difference in predictive log likelihood on the training data was less than 10−4 between iterations. The restarts with the best predictive log likelihoods for SNVs and SVs were selected as seeds for the final fitting step. The model was again fit to the data 1,500 times. The model parameters for each restart were set to the parameters of the optimal models from the previous step described above, then run for a maximum of 1,000 iterations or until the relative difference in predictive log likelihood on the training data was less than 10−5 between iterations. The restart with the best mean rank of the SNV and SV predictive log likelihoods from this round was selected as the final model.

MMCTM estimated SNV signatures were matched to COSMIC signatures by solving the linear sum assignment problem for cosine distances between the MMCTM and COSMIC signatures v.3 (minus tobacco smoking-associated COSMIC SBS4) using the clue R package. Samples were clustered by first applying UMAP to the normalized signature probabilities for the HRD SNV signature and all SV signatures with n_neighbors = 20 and min_dist = 0 to produce two-dimensional sample embeddings. Next, HDBSCAN was run on the sample embeddings with min_samples = 5, min_cluster_size = 5 and cluster_selection_epsilon = 0.75 to produce the sample clusters (strata).

Survival analysis

For each patient, the number of days between diagnosis and death or last follow-up was collected. Patients were segregated into groups,
and a Kaplan–Meier curve was fitted for each group. Each cancer type was analysed separately and in two distinct grouping schemes. First, patients were split into HRD and ‘Other’ groups, in which the HRD group included patients whose cancers were identified as being in either the HRD-Dup or HRD-Del groups, and the ‘Other’ group included all other patients. Next, patients were grouped according to their assigned signature types: HRD-Dup, HRD-Del, TD or FBI.

**DLP+ WGS quantification and analysis**

Single-cell copy number, SNV and SV calls were generated using a previously described approach, except that BWA-MEM was used to align DLP+ reads to the hg19 reference genome. The genome was segregated into 500-kb bins, and GC-corrected read counts were calculated for DLP+ reads to the hg19 reference genome. The genome was segregated into 500-kb bins, and GC-corrected read counts were calculated for each bin. These read counts were then input into HMMCopy to produce integer copy number states for each bin.

**DLP+ data filtering**

Cells were retained for further analysis if the cell quality was at least 0.75 (ref. 3), and they passed both the S-phase and the contamination filters. The contamination filter uses FastQ Screen to tag reads as matching human, mouse or salmon genomes. If more than 5% of reads in a cell are tagged as matching the mouse or salmon genomes, then the cell is flagged as contaminated. The S-phase filter uses the cell-cycle state Random Forest classifier from ref. 3 and removes cells for which S-phase is the most probable state. The HGSC and TNBC cells were also filtered to remove small numbers of contaminating diploid cells.

Finaly, cell filtering was performed to remove putative early and late S-phase cells that passed the initial S-phase filter. This involved two steps: first, building a cell phylogeny with sitka and manually identifying the minimal phylogeny branches in which the cycling cells have been clustered. The cells in these branches were then removed. Next, clustering cells according to their copy number profiles and removing manually identified clusters of cycling cells.

We removed potentially problematic genome bins from our copy number results that had a mappability score of 0.99 or below, or that were contained in the ENCODE hg19 blacklist.

To detect SNVs and SVs in each dataset, reads from all cells in a DLP+ library were merged to form ‘pseudobulk’ libraries. SNV calling was performed on these libraries individually using MutationSeq (probability threshold = 0.9) and Strelka (score > 20) (ref. 3). Only SNVs that were detected by both methods were retained. For each dataset, the union of SNVs was aggregated, then for each cell and each SNV, the sequencing reads of that cell were searched for evidence of that SNV. SV calling was performed in a similar manner, by forming pseudobulk libraries, then running LUMPY and deStruct on each pseudobulk library, and retaining events that were detected by both methods. LUMPY and deStruct predictions were considered matched if the breakpoints matched in orientation and the positions involved were each no more than 200 nucleotides apart on the genome. Only deStruct predictions with a matching LUMPY prediction were retained. Sparse per-cell breakpoint read counts were extracted from deStruct using the cell identity of read evidence for each predicted breakpoint. SNV and SV calls were further post-processed according to a previous study. When performing pseudobulk analysis on groups of cells, a breakpoint is considered present in a clone if at least one cell that constitutes the clone contains evidence of the breakpoint. A subsampling experiment determined that this approach has 80% power to recover breakpoints at a cumulative coverage of 5× (100–150 cells) (see Supplementary Note).

**Analysis of mutation signatures in DLP+ data**

Mutation signature probabilities were fit to DLP+ pseudobulk-derived SNV and SV counts for each patient using the MMCTM method and pre-computed mutation signatures from the HGSC and TNBC meta-cohort. Inference was performed as per the bulk sequencing data, until the relative difference in predictive log likelihood was < 10⁻⁶ between iterations.

**Identifying clones in DLP+ WGS by clustering copy number profiles**

For most datasets, clones were detected by first using UMAP on per-cell GC-corrected read count profiles, producing a two-dimensional embedding of the cell profiles. We then ran HDBSCAN on the two-dimensional embedding from UMAP to detect clusters of cells with similar copy number profiles.

UMAP was run with min_dist = 0.0, and metric = “correlation”, whereas HDBSCAN was run with approx_min_span_tree = False, cluster_selection_epsilon = 0.2, and gen_min_span_tree = True. Dataset specific UMAP and HDBSCAN parameter settings are listed in Supplementary Table 3.

**Calculating cell ploidy**

Cell ploidy was calculated by taking the most common copy number state. Copy number states were those determined by HMMCopy.

**Identifying missegregated chromosomes**

The approach taken to identify putative chromosome missegregation events is similar to a previous one. Cells were split into groups corresponding to their clones. Clone copy number profiles were generated for each clone. Cells with ploidy not equal to the clone consensus profile were normalized to match the clone ploidy. Cell copy number profiles were compared to the clone copy number profile for the matching clone to which the cell belongs. The result was assignment of an offset value for each genomic bin in each cell, that represented the copy number difference between the cell and the clone-level consensus profile. For each chromosome in each cell, if a particular copy number difference (that is, -1, 1, and so on) represented at least 75% of the chromosome, then we labelled that chromosome as having a missegregation event.

**Identifying CNA segments**

Gain and loss segments in each cell were found by comparing the copy number state in each 500-kb bin to that cell’s ploidy. A copy number higher than ploidy was labelled as a gain, and a copy number lower than ploidy was labelled as a loss. Gain and loss segments are a set of consecutive bins with the same gain–loss label. Segments ≤ 1.5 × 10⁶ bp were excluded to reduce segments potentially resulting from noise in the HMMCopy copy number states.

**Computing serriform variability scores in CNA breakpoints**

For each dataset, consensus copy number profiles were generated for each clone. Copy number segments were identified as above for each consensus profile. Copy number segments were then identified for single-cell copy number profiles. The copy number profiles of each cell were normalized so that the adjusted cell ploidy matched the ploidy of the clone to which the cell belonged using the following formula: cell_state = cell_state /cell_ploidy × clone_ploidy

Cell copy number segments were matched to segments in the clone copy number profile as follows: for each segment in the clone copy number profile, inspect the copy number states of the adjacent segments. If the segment state was less than both adjacent states, then only cell segments whose state was less than both of the two adjacent clone segment states could be matched to that segment. If the clone segment state was higher than both adjacent states, only cell segments whose state was higher than both of the adjacent clone segment states could be matched to that segment. If the clone segment state was in between the two adjacent states, only cell segments whose state was in between the two adjacent clone segment states could be matched to that segment. Finally, each cell segment was matched to the compatible clone segment that it overlapped the most, in which compatibility
means that the cell segment state met the criteria described above and
the cell belonged to the relevant clone.

Next, clone segment breakpoints were aggregated across all clones.
For each breakpoint, matched cell breakpoints were identified. Sta-
table cell breakpoints (that is, those cell breakpoints that matched
the clone-level breakpoint position) and unstable cell breakpoints (all other cell
breakpoints) were queried for their raw GC-corrected read count
values up to five bins to the left and right of the breakpoint position.
Breakpoint noise values were computed as the mean absolute value
of the difference between these values and the integer copy number
state inferred by HMMCopy. For each clone-level breakpoint event,
cells were removed if their breakpoint noise values were higher than
a threshold value, which was computed as the mean noise value of
the stable cell breakpoints.

Serration scores for each event were calculated by first computing
the frequency of cell-specific breakpoint positions. Each cell breakpoint
position was considered 'rare' if it occurred in less than 5% of cells with
the considered event. The final serration score was computed as the
fraction of event cells whose breakpoint position was considered 'rare'.

Comparing segmentation profiles across cells
To facilitate comparisons of genomic profiles across cells, we inferred
a set of disjoint segments from the consensus copy number profiles
of clusters. For each clone or cluster we generated a consensus segmenta-
tion profile, and then used the 'disjoin_ranges' function from ply-
ranges67 to generate a non-overlapping disjoint segmentation profile.
Each segment was then genotyped in each cell by taking a consensus
across the bins within each segment, producing a consistent set of
genomic segments and states that could be compared across cells.

Comparison of HLAMP copy number variance
HLAMPs were identified by first selecting 500-kb genomic bins in which
at least 10 cells have a raw copy number (adjusted per-bin read counts)
of at least 10. Copy number variance for each bin was calculated using
the raw copy number that was adjusted for cell ploidy and cell clone
by first dividing the copy number by the cell ploidy, then subtracting
the mean clone raw copy number. The cell ploidy is the most common
HMMCopy copy number state as described above, and the mean clone
copy number is computed for each bin in each clone across all cells in
that clone. Mean HLAMP copy number variance was calculated for
each dataset across all HLAMP bins, and these values were compared
between signature type dataset groups.

Clustering HLAMP genomic features
To explore plausible mechanistic origins of oncogenic HLAMPs we
extracted genomic features proximal to the locus of interest. We took
a region 15 Mbp either side of the locus of interest and pulled out copy
number and SV features. We extracted the following features: entropy
of haplotype-specific states; total number of SVs identified; proportion
of SVs of each type (fold-back inversions, duplications, deletions and
translocations); number of chromosomes involved in translocations;
ratio of copy numbers between the bin containing the oncogene and the
average copy number across the chromosome; average copy number
state; average size of segments; average number of segments; and aver-
age minor allele copy number. All averages are across cells. We then
performed hierarchical clustering on a scaled matrix of all features, using
the silhouette width to determine the appropriate number of clusters.

HSCN analysis
See the Supplementary Note for a detailed discussion of our method,
 SIGNALS, for HSCN analysis. This includes validation of the method and
benchmarking against other methods. In brief, SIGNALS uses haplotype
blocks genotyped in single cells and implements an hidden Markov
model (HMM) based on a Beta-Binomial likelihood to infer the most
probable haplotype-specific state. We used default parameters for all
datasets apart from SA1292, in which we increased the self transition
probability from 0.95 to 0.999 to mitigate against the noisier copy
number data in this sample.

Pseudobulk HSCN profiles
In numerous places in this study we construct 'pseudobulk'
haplotype-specific or total copy number profiles either across all cells
in a sample or subsets of cells that share some features of interest. To do
this, we group the cells of interest and then compute an average profile
by taking the median values of copy number and BAF and the mode of
the haplotype-specific state. The function 'consensuscopynumber'
provided in SIGNALS was used for this.

Comparing segmentation profiles across cells
To facilitate comparisons of genomic profiles across cells, we inferred
a set of disjoint segments from the consensus copy number profiles
of clusters. For each clone or cluster we generated a consensus segmenta-
tion profile, and then used the 'disjoin_ranges' function from ply-
ranges67 to generate a non-overlapping disjoint segmentation profile.
Each segment was then genotyped in each cell by taking a consensus
across the bins within each segment, producing a consistent set of
genomic segments and states that could be compared across cells.

Identification of parallel copy number events
The set of genotyped disjoint segmentation profiles was used to cal-
culate the number of parallel copy number CNAs. Parallel CNAs were
defined as genomic regions greater than 4 Mbp in which gain or loss
of both the maternal and paternal haplotype was observed in more
than 1% of cells. Copy number breakpoints of segments do not need
to match to be included.

Phylogenetic analysis
We computed phylogenetic trees using sitka as previously described67,
using the consensus tree from the posterior distribution for down-
stream analysis. For visualization, clades with a high fraction of sin-
gletons (nodes with a single descendant) were removed. To remove
nodes, nodes were ordered by the fraction of descendants that were
singletons, and nodes were removed iteratively until a maximum of 3%
of cells in the tree were removed. Trees were visualized using ggtree68
and functionality in SIGNALS. Phylogenetic distances were computed
as the mean pairwise distances between phylogenetic tips (cells) using
the cophenetic function in APE69. Distances represent the number of
copy number change points between two cells on the phylogeny.

Event rates inferred from single-cell phylogenies
To compute the rates of gains and losses of whole chromosomes, chro-
mosome arms and segmental aneuploidies we enumerated the number
of events from our single-cell phylogenies using parsimony-based
ancestral state reconstruction. We used the genotyped disjoint seg-
mentation profiles for this.

We first defined states for each segment in each cell relative to the
most common state across all cells. For each segment, cells can have
one of two possible states for each class of interest: (gain, not gained),
(loss, not lost). By casting the problem as reconstructing the ancestral
states within the phylogeny, we can then compute the number of tran-
sitions between these states that most parsimoniously explains
the phylogenetic tree. We used a simple transition matrix in which transi-
tions between states incur a cost of 1. Ancestral state reconstruction
then amounts to finding the reconstruction that minimizes this cost.
The event frequency per sample is then calculated by dividing the par-
simony score (number of events) by the number of cells. We used castor
v.1.6.6 in R to perform the ancestral state reconstruction69. The unit
of this quantity is the number of events per cell division, assuming
no cell death. It is possible (perhaps likely) that many cells get segmental
gains or losses but then die, we never sample such cells and our phy-
logenetic tree reconstructs ancestral relationships between cells that
survive and that we sample. It is challenging to decouple the death rate of cells from the true event rate per cell division\textsuperscript{71}; thus, our event rate is an effective event rate; that is, the event rate scaled by the (unknown) death rate of cells. To contrast the rates across different types of events, we classified segments as whole chromosomes, chromosome arms or segmental aneuploidies.

**Calculation of copy number distance**

The copy number distance calculates the number segments that need to be modified to transform one copy number profile into another\textsuperscript{20}. We use this measure to compute cell-to-cell variation in our dataset. To compute this measure, we modified the code provided in a previous study\textsuperscript{29} to take into account whole-genome doubling of cells (https://github.com/raphael-group/WCND). We did this as follows: given two copy number profiles (integer copy number states of individual haplotypes in bins across the genome) \( CNP_A \) and \( CNP_B \), we computed the following distances:

\[
\begin{align*}
    d_1 &= f(CNP_A, CNP_B) \\
    d_2 &= f(2 \times CNP_A, CNP_B) \\
    d_3 &= f(CNP_A, 2 \times CNP_B)
\end{align*}
\]

in which \( 2 \times \) refers to doubling the copy number state across the whole genome. We then took the copy number distance to be

\[
d = \min(d_1, d_2, d_3).
\]

If the minimum was \( d_1 \) or \( d_3 \), we increased \( d \) by 1 (that is, counting WGD as an additional event). Calculating all pairwise comparisons is computationally expensive, so for each dataset we subsampled 250 cells and calculated all pairwise distances for these 250 cells.

**10X scRNA-seq processing**

CellRanger software (v.3.1.0) was used to perform read alignment, barcode filtering and quantification of unique molecular identifiers (UMIs) using the 10X GRCm38 transcriptome (v.3.0.0) for FASTQ inputs. CellRanger filtered matrices were loaded into individual Seurat objects using the Seurat R package (v.4.1.0) (refs. \textsuperscript{72,73}). The resulting gene-by-cell matrix was normalized and scaled for each sample. Cells retained for analysis had a minimum of 500 expressed genes and 1,000 UMI counts and less than 25% mitochondrial gene expression. Genes expressed in fewer than three cells were removed. Cell-cycle phase was assigned using the Seurat\textsuperscript{73} CellCycleScoring function. Scrublet\textsuperscript{74} (v.0.2.3) was used to calculate and filter cells predicted to be doublets. We then applied the standard Seurat processing pipeline using default parameters apart from using the first 20 principal component analysis (PCA) dimensions for nearest neighbour and UMAP calculations.

**Allelic imbalance in scRNA-seq**

We called heterozygous SNPs in the scRNA-seq data using cellSNP v.1.2.2 (ref. \textsuperscript{75}). As input, we used the same set of heterozygous SNPs identified in the scDNA-seq and the corresponding normal sample for each sample. The liftover script provided in cellSNP was used to lift over SNPs from hg19 to hg38. After genotyping, we phased the SNPs using the phasing information computed from the haplotype-specific reference in the scDNA-seq. As SNP counts are much more sparse in scRNA-seq than in scDNA-seq (around two orders of magnitude lower), we aggregated counts across segments (minimum size = 10 Mbp), computing the BAF for each segment. Then we generated a cell by segment BAF matrix and incorporated this into our gene expression Seurat objects. We applied an additional filtering criterion here, removing cells with fewer than 200 SNP counts. Functionality to map scDNA-seq to scRNA-seq and call allelic imbalance is provided in SIGNALS.

**Differential expression analysis**

Differential expression analysis between gene expression clusters was computed using the Wilcoxon rank sum test with the presto R package. Gene expression clusters were computed using the FindClusters function in Seurat. Only cells in G1 phase were included. To compare gene expression variability for oncogenes, we took the absolute maximum log-transformed fold change for each sample for each oncogene and contrasted this value in cases in which oncogene copy number was determined to be fixed or variable from DLP+ single-cell sequencing of the same samples. ‘Variable’ oncogenes were defined as those that had a minimum ratio of 2 between the maximum to minimum clone-level copy number, and ‘non-variable’ oncogenes as those that had a ratio of less than 2.

**Nearest neighbour gene expression analysis**

To assess transcriptional convergence of losses of alleles we made use of the shared nearest neighbour graph computed using Seurat. This was done for chr. 2q in sample SA906b. For a given cell, an enrichment score was defined as the observed fraction of nearest neighbours divided by the expected fraction of nearest neighbours. Here, the expected fraction of neighbours with the same allelic state was defined as the global fraction of cells in each state. Hence, a positive enrichment score indicates an overrepresentation of cells in the allelic state of interest among its nearest neighbours, a negative score indicates an underrepresentation and a score of 0 would reflect a perfectly mixed neighbourhood of cells with different allelic states. To mitigate the influence of other technical or biological variability, for this analysis we only included cells in G1 phase, and removed cells with greater than 7.5% mitochondrial gene expression as we found that this was variable between gene expression clusters.

**Statistical tests**

The statistical tests used were two-tailed unequal-variance \( t \)-tests unless otherwise specified: log-rank tests were used for comparing survival curves; Wilcoxon rank sum two-tailed tests were used for comparing segment lengths, segment counts, miss segregations and ploidy percent-ages, copy variances, bin counts, gene copy number distributions, gene expression log-transformed fold changes, parallel copy number counts and breakpoint counts; and hypergeometric tests were used to identify enrichment of gene mutations. \( p \) values from multiple comparisons were corrected using the Benjamini–Hochberg method\textsuperscript{76}.

**Box plot statistics**

All box plots indicate the median, first and third quartiles (hinges), and the most extreme data points no farther than 1.5\( \times \) the IQR from the hinge (whiskers).

**Data availability**

All data are available for general research use. Processed data including somatic mutation data for bulk WGS, total (and allele-specific) copy number profiles for DLP+ data and filtered count matrices for scRNA-seq data are available for download at https://zenodo.org/record/6998936. Raw scRNA-seq data are available for download at https://ega-archive.org/studies/EGAS00001006343. Raw single-cell sequencing data generated for this study are available from https://ega-archive.org/studies/EGAS00001006343, and previously published single-cell sequencing data used in this study are available at https://ega-archive.org/studies/EGAS00001004448 and https://ega-archive.org/studies/EGAS00001003190. Somatic mutation calls from bulk
Extended Data Fig. 1 | Study overview. Experimental and cohort design. Single-cell genomes, transcriptomes and long-read sequencing libraries were generated from isogenic 184-hTERT cell lines (WT, or with *TP53*, *BRCA1* or *BRCA2* mutations) a) or PDX tissue from patients with TNBC and patients with HGSC from a meta-cohort with assigned SV or SNV mutational signatures b). Single-cell and long-read sequencing (c) was used to examine mutational processes and haplotype-specific genomic diversity, including HLAMPs or rearrangements d), parallel events e) and SSVs f) at single-cell resolution and within clonal and subclonal populations. Generated using Biorender.com.
Extended Data Fig. 2 | Sanger sequencing of cell lines and tumour histology. a,b) Verification of CRISPR-Cas9 induced genotypes of 184-hTERT cell lines. a) Sanger sequencing of TOPO cloned BRCA1 and BRCA2 regions. b) Western blotting for p53, BRCA1 and BRCA2 proteins for 184-hTERT cell lines and including an additional BRCA2−/− clone, 112.72 and TP53−/− clone, SA1101. GAPDH and vinculin loading controls were performed on the same blot as p53, BRCA1 or BRCA2 probes. Blots shown are representative of n = 3 (WT), n = 3 (BRCA1) and n = 6 (BRCA2) independent experiments. For source blots, refer to Supplementary Fig. 1. c) Histology of HGSC PDXs in the dataset. Scale bars 300 μm and 50 μm as indicated. Images are representative of two cores stained from each PDX tissue.
Extended Data Fig. 3 | Validation of the SIGNALs method in an ovarian cancer cell line. a, b) HSCN from 2 individual cells from the OV2295 cell line. c) Total copy number heat maps and HSCN heat map for 1084 cells. Each row is an individual cell. Rows are ordered according to a UMAP + HDBSCAN clustering, with clusters annotated on the left hand side. d) Distribution of VAFs as a function of haplotype-specific state. e, f, g) VAFs in clones where the dominant allele switches between A and B. Each point is the VAF of a mutation, with lines connecting the same mutation in different clones.
Extended Data Fig. 4 | DLP summary statistics. Summary of DLP+ sequencing statistics of data for 184-hTERT cell lines (a–d) and HGSC and TNBC tumours (e–h). Number of cells for each box plot as indicated in panels a) and e). Shared y axis labels shown at left. The legend for e–h) indicates the number of samples for each cancer and signature type. All box plots indicate the median, 1st and 3rd quartiles (hinges), and the most extreme data points no farther than 1.5x the IQR from the hinge (whiskers).
Extended Data Fig. 5 | Haplotype-specific analysis reveals breakage–fusion–bridge processes and parallel losses. a) Diagram of BFBCs b) Heat maps of the copy number of each homolog in SA1188. c–f) HSCN and structural variation in clusters B, I, F and the small subpopulation with PIK3CA amplification. Here we plot the copy number for each homologous chromosome in purple for homologue B and green for homologue A. g) Parallel copy number losses in SA906b: total copy number (top) and HSCN (bottom) heat maps for chr2 in SA906b h) two individual cells from g). i) UMAP dimensionality reduction plots of scRNA-seq data generated from SA906b, colours indicate the density of loss of chr 2q A vs. B haplotype. j) Enrichment of the haplotype-specific state on chr 2q of nearest neighbour cells (# cells with loss of A = 175, # of cells with loss of B = 34, # Balanced = 2066). All box plots indicate the median, 1st and 3rd quartiles (hinges), and the most extreme data points no farther than 1.5x the IQR from the hinge (whiskers).
Extended Data Fig. 6 | SNV and SV signatures. a) SNV and b) SV mutation signatures estimated from HGSC and TNBC bulk tumour mutation catalogues using the MMCTM method. The x axis in a) is the 96-channel (i.e. A>C>A, …, T>T>G) SNV types. SV types are DEL: interstitial deletions, DUP: tandem duplications, INV: inversions, FBI: fold-back inversions, TR: translocations. SV signature labels are S-Dup: small duplications, M-Dup: medium duplications, L-Dup: large duplications, S-Del: small deletions, L-Del: large deletions, Clust-FBI: clustered fold-back inversions, Clust-SV: clustered other structural variants, Tr: translocations, FBI/Inv: fold-back inversions and inversions.
Extended Data Fig. 7 | Meta-cohort signature analysis of 139 TNBC and 170 HGSC bulk whole genomes. a) Heat map representing individual patients as columns, annotation tracks (top) including cancer type and mutation status of key genes (strata with adjusted p-values ≤ 0.1 shown as coloured bars on left), standardized signature probabilities of SNVs and SVs (middle) and event counts (bottom). b) Signature type (see stratum annotation track) proportions by cancer type. c) SNV and SV count distributions per signature type (number of samples shown below each violin, data points shown left of violins). Kaplan–Meier survival probability of HGSCs faceted by d) HRD and e) more granular signatures (p-values computed using the log-rank test, p = 0.0038 for d) and p = 0.0022 for e). All box plots indicate the median, 1st and 3rd quartiles (hinges), and the most extreme data points no farther than 1.5x the IQR from the hinge (whiskers).
Extended Data Fig. 8 | See next page for caption.
Extended Data Fig. 8 | Summary, quality control and features of single-cell WGS of tumours. a) UMAP of meta-cohort signature probabilities. Lines connect DLP-pseudobulk to their bulk data counterpart. b) Correlation of proportion of the genome that is LOH between DLP-pseudobulk (horizontal) and matched bulk WGS (vertical). Correlation coefficient (R) and p-value (p) derived from a linear regression in inset, shaded area shows the 95% CI of the linear regression. c) VAF distributions (horizontal) for somatic mutations called in single cells as a function of haplotype-specific state (vertical), coded as integer copy level allele A | integer copy level allele B. Data from all DLP samples are included. d) Heat map showing total copy number (left) and HSCN (right) of single cells from a TNBC HRD-Dup case (SA501). e) Chromosomal gains and losses across different ploidy states and mutational signature grouping. Total counts (black), gains (red), and losses (blue) shown.

f) Relationship between gain/loss ratios and number of gained or lost segments for representative datasets from each signature type (left) and all HRD-Dup, TD, or FBI cases (right).

g) Differences in copy number segmental gain and loss counts (n = 12 FBI, n = 8 HRD-Dup, n = 3 TD), comparing ploidy-relative case-level consensus copy number profiles (green) and mean cell-level changes relative to clone copy number profiles (purple).

h) HSCN distance distributions for all PDX samples. Distribution is over n = 1,000 sampled pairwise HSCN distances. Horizontal black line shows the mean value of the distribution.

i) HSCN distance distributions as a function of signature type, each dataset is summarized as the mean of the distributions on the left.

P-values indicate per group comparisons using the two-sided Wilcoxon test (n = 12 FBI, n = 8 HRD-Dup, n = 3 TD).

j) Number of parallel copy number segments (n, size of circle) and the proportion of segments containing parallel events (f, colour of circle) across all datasets as a function of clonality. Clonal: CCF > 80%, Subclonal: 20% < CCF ≤ 80%, Rare: 1% < CCF ≤ 20%. k) Proportion of segments with parallel CNA in HRD-Dup vs FBI, * = p < 0.05, ns = p > 0.05, two-sided Wilcoxon test (n = 12 FBI, n = 8 HRD-Dup). Exact p-values from left to right, p = 0.85, p = 0.031, p = 0.1. All box plots represent the median, 1st and 3rd quartiles (hinges), and the most extreme data points no farther than 1.5x the IQR from the hinge (whiskers).
Extended Data Fig. 9 | See next page for caption.
Extended Data Fig. 9 | Genomic features of HLAMPS and long read sequencing validation. a) Each column is a HLAMP that amplifies an oncogene. Each row is a feature extracted from a region 15Mb either side of the amplification. Complexity = entropy of haplotype-specific states, #SV = total number of structural variants identified, proportion of SVs of each type: foldback inversions, duplications, deletions and translocations. #chr = number of chromosomes involved in translocation. bin/chr ratio copy number of the bin containing the oncogene to the average copy number across the chromosome. Ratio is the copy number ratio between the clone with the maximum copy number state and the minimum copy number state. b–c) HLAMPS involving multiple chromosomes. Left plot shows copy number profiles from pseudobulk clones derived from DLP, lines indicate rearrangement breakpoints. Right plot shows example long reads from Oxford nanopore technologies that support inter-chromosomal translocations. Example reads and their mapping to chromosomes of interest (top right), long-read coverage of genomic region and alignment of all supporting reads (bottom right). d) SA1184 MYC amplification c) SA1181 chr5q amplification. d) Long-read support for inter-chromosomal alterations involving chromosomes 3 and 6 in SA1096, DLP clone-level plots shown in Fig. 4j.
Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection
- DLP: single cell WGS pipeline: https://github.com/shahcompbio/single_cell_pipeline
- Bulk WGS pipeline: https://github.com/shahcompbio/wgs

Data analysis
- MMCTV method: https://github.com/shahcompbio/MultiModalMuSig.jl v0.3.0
- SIGNALS: https://github.com/shahcompbio/signals v0.7.2
- CELLRANGER: https://support.10xgenomics.com/single-cell-gene-expression/software v3.1.0
- cellSNP: https://github.com/single-cell-genetics/cellsnlp-v1.2.2
- seurat https://satijalab.org/seurat/v4.1.0
- scrublet: https://github.com/swolock/scrublet v0.2.3
- CHISEL, https://github.com/raphael-group/chisel v1.0.0
- Allelescope https://github.com/seasoncloud/Alelescope
- sirska https://github.com/UBC-Stat-MI/sirstka
- castor https://cran.r-project.org/web/packages/castor/index.html v1.6.6
- Guppy https://nanoporetech.com/nanopore-sequencing-data-analysis v3
- minmap2 https://github.com/lh3/minimap2 v2.24
- sniffles https://github.com/fritzsedlazeck/Sniffles v1.0.12
- cuteSV https://github.com/minghui11/cuteSV v1.0.11

For manuscripts utilizing custom algorithms or software that are not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.
Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data are available for general research use. Processed data including somatic mutation data for bulk WGS, total (and allele-specific) copy number profiles for DLP+ data and filtered count matrices for scRNA-seq data are available for download at https://remodo.org/record/6998936. Raw scRNA-seq data are available for download at https://ega-archive.org/studies/EGAS00001006343. Raw single-cell sequencing data generated for this study are available from https://ega-archive.org/studies/EGAS00001006343, and previously published single-cell sequencing data used in this study are available at https://ega-archive.org/studies/EGAS00001004448 and https://ega-archive.org/studies/EGAS00001003190. Somatic mutation calls from bulk WGS for 16 patients with TNBC for whom the IRB consent does not include publication of raw sequencing data are available at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs0030338.v1.p1, and raw sequencing data can be provided upon request under material transfer agreement to shahs3@mskcc.org. Bulk WGS 8AM files from patients under IRB consent protocols for public release of raw data are available for download at https://ega-archive.org/studies/EGAS00001006343, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003036.v1.p1 and https://ega-archive.org/datasets/EGAD00001003268 (for previously published data) or by request under material transfer agreement to shahs3@mskcc.org and saparicio@bccrc.ca.

Field-specific reporting

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

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

| Sample size | 309 patients, 309 with bulk WGS and 23 with scWGS or RNAseq. 8 hTERT lines, 2 ovarian cancer cell lines. Sample size was determined by the number of patient tissue accrued and number of 184hTERT clones with confirmed knockout identified. |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | Poor quality and s-phase cells were removed from analysis as per Methods |
| Replication | Where possible, biological duplicates were sequenced by scDNAseq. 11 patients with scWGS have more than 1 replicate tissue sequenced by DLP+ |
| Randomization | Patients were stratified into groups according to their mutational signature as described in the manuscript. No grouping of animals was performed, so no randomization was necessary |
| Blinding | Patient tissues were stratified by mutational signature during analysis, blinding was not necessary during data collection or analysis as no groups were assigned. |

Reporting for specific materials, systems, and methods

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

| Materials & experimental systems | Methods |
|---------------------------------|---------|
| n/a | Involved in the study |
| ☐ Antibodies | ☒ ChIP-seq |
| ☒ Eukaryotic cell lines | ☐ Flow cytometry |
| ☐ Palaeontology and archaeology | ☒ MRI-based neuroimaging |
| ☒ Animals and other organisms | |
| ☐ Human research participants | |
| ☐ Clinical data | |
| ☒ Dual use research of concern | |
Antibodies

Antibodies used: mouse anti-p53 (Santa Cruz SC-126), mouse anti-BRCA1 (Santa Cruz SC-6954), mouse anti-BRCA2 (Millipore OP95), goat anti-GAPDH (SC-48166), rabbit K Ras (Lifespan Bioscience, LS-84683)

Validation: K Ras - Validated by our lab using human uterine tissue and by Lifespan Bioscience using human uterine and placental tissues. GAPDH - Validated by Santa Cruz. p53 - Validated by Santa Cruz using A549, Daudi, MDA-MB-231, SW480 cell lines and human bladder carcinoma tissue. BRCA1 - Validated by Santa Cruz using A-431, Hela, u705 and MCF7 cell lines. BRCA2 - Validated by Millipore using MCF7 cell line.

Eukaryotic cell lines

Policy information about cell lines:

Cell line source(s): 184-hTERT L9 (SA039), 184-hTERTTPS3-/- (SA906a and SA906b), 184-hTERTTPS3-/-;BRCA1+/- (SA1292), 184-hTERTTPS3-/-;BRCA1-/- (SA1054), 184-hTERTTPS3-/-;BRCA2-/- (SA1055 and SA1056) and 184-hTERTTPS3-/-;BRCA2+/- (SA1188), OV2295, TOV2295

Authentication: All cell lines were authenticated by short tandem repeat testing by Genetica Lab Corp DNA Identity

Mycoplasma contamination: All cells were tested for mycoplasma contamination by Genetica Lab Corp DNA Identity and tested negative

Commonly misidentified lines: No commonly misidentified cell lines were used in this study.

Animals and other organisms

Policy information about studies involving animals: ARRIVE guidelines recommended for reporting animal research

Laboratory animals: NOD/SCID/IL2-/- (NSG) and NOD/IL2-/- (NRG) mice were used in this study. Housing was maintained in a 18–25 °C temperature range and 20–70% humidity range, with a 12 hour daylight cycle (on at 6:00am, off at 6:00pm). Surgery was carried out on female mice between the ages of 5–12 weeks. All experimental procedures were approved by the University of British Columbia Animal Care Committee

Wild animals: This study did not involve wild animals

Field-collected samples: This study did not involve samples collected from the field

Ethics oversight: The animal care committee of University of British Columbia approved all experimental procedures in this study

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

Human research participants

Policy information about studies involving human research participants

Population characteristics: Age, gender (female), past and current diagnosis, hormone receptor status, past and current treatment, survival. Only survival was used in this manuscript.

Recruitment: Patients were recruited with informed consent during surgery or biopsy procedures. Patients were not selected for or excluded based on genomic information, so would not impact results.

Ethics oversight: approved by the Ethics Committees at the University of British Columbia

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