Breast tumours maintain a reservoir of subclonal diversity during expansion

https://doi.org/10.1038/s41586-021-03357-x

Received: 13 May 2020

Accepted: 12 February 2021

Published online: 24 March 2021

Check for updates

Darlan C. Minussi[1,2,13], Michael D. Nicholson[3,4,5,13], Hanghui Ye[1,2,13], Alexander Davis[1,2], Kaile Wang[1], Toby Baker[6], Maxime Tarabichi[6], Emi Sei[1], Haowei Du[1,7], Mashiat Rabbani[1,7], Cheng Peng[1,7], Min Hu[1], Shanshan Bai[1], Yu-wei Lin[1,2], Aislyn Schalck[1,2], Asha Multani[1], Jin Ma[1], Thomas O. McDonald[3,4,5,8], Anna Casasent[1,2], Angelica Barrera[9], Hui Chen[10], Bora Lim[9], Banu Arun[9], Funda Meric-Bernstam[9], Peter Van Loo[6], Franziska Michor[3,4,5,8,11]✉ & Nicholas E. Navin[1,2,12]✉

Our knowledge of copy number evolution during the expansion of primary breast tumours is limited[1,2]. Here, to investigate this process, we developed a single-cell, single-molecule DNA-sequencing method and performed copy number analysis of 16,178 single cells from 8 human triple-negative breast cancers and 4 cell lines. The results show that breast tumours and cell lines comprise a large milieu of subclones (7–22) that are organized into a few (3–5) major superclones. Evolutionary analysis suggests that after clonal *TP53* mutations, multiple loss-of-heterozygosity events and genome doubling, there was a period of transient genomic instability followed by ongoing copy number evolution during the primary tumour expansion. By subcloning single daughter cells in culture, we show that tumour cells rediversify their genomes and do not retain isogenic properties. These data show that triple-negative breast cancers continue to evolve chromosome aberrations and maintain a reservoir of subclonal diversity during primary tumour growth.

Aneuploidy is a salient feature of human breast cancers and is particularly prevalent in patients with triple-negative breast cancer (TNBC) harbouring *TP53* mutations[3,4]. Although the underlying molecular mechanisms of aneuploidy have been elucidated in model systems[5], our knowledge of when and how chromosomal rearrangements emerge and are maintained during the growth of primary tumours in humans remains limited. A long-standing paradigm for tumour progression is that mutations and chromosomal aberrations accumulate gradually and sequentially over time, leading to more malignant stages of cancer[6]. However, an alternative model is punctuated copy number evolution (PCNE), in which many chromosomal rearrangements are acquired together in short bursts of genomic instability early in tumour evolution[7–12]. Evidence for this model has been reported in breast tumours[7,8], colon cancers[9] and prostate cancers[10] and may be common in many types of human cancer[13].

Whether there is also ongoing copy number evolution after the initial burst of genome instability remains unresolved[1,7,14]. In our previous work, we found that PCNE is common in patients with TNBC[7], but were unable to ascertain whether copy number profiles continued to evolve after the initial catastrophic event, when tumour cells undergo clonal expansions. Resolving these models has been difficult owing to the limited number of cells that could be sequenced, as well as extensive technical noise in first-generation single-cell DNA sequencing (scDNA-seq) technologies[8]. Here we report on a key technical advance that enabled us to sequence thousands of single cells and address fundamental questions regarding the natural history of chromosome evolution in patients with TNBC.

Single-molecule single-cell sequencing

We developed a method called acoustic cell tagmentation (ACT), which combines fluorescence-activated cell sorting (FACS) of single nuclei, tagmentation and acoustic liquid transfer (ALT) technology to perform high-throughput scDNA-seq at single-molecule resolution (Fig. 1a). To perform ACT, nuclear suspensions are prepared from fresh or frozen tissues and stained with DAPI for flow-sorting into high-density ($N = 384$) plates. The isolated nuclei undergo a three-step amplification chemistry, which involves: (1) nuclear lysis, (2) direct tagmentation of genomic DNA using a Tn5 transposase to add universal adapters, and (3) PCR to incorporate dual barcodes for cell library multiplexing. The chemistry steps are robotically automated and the Tn5 enzyme is scaled down (1:20) to nanolitre volumes using ALT[15]. This approach generates barcoded single-cell DNA libraries with a mean size of 312 base pairs (bp) that are pooled together for next-generation sequencing (Extended

[1]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center UTHealth, Houston, TX, USA. [3]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [5]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. [6]Cancer Genomics Laboratory, The Francis Crick Institute, London, UK. [7]Graduate Program in Diagnostic Genetics, School of Health Professions, MD Anderson Cancer Center, Houston, TX, USA. [8]Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA, USA. [9]Department of Breast Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [10]Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [11]The Ludwig Center at Harvard, Boston, MA, and the Broad Institute of MIT and Harvard, Cambridge, MA, USA. [12]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [13]These authors contributed equally: Darlan C. Minussi, Michael D. Nicholson, Hanghui Ye. ✉e-mail: michor@jimmy.harvard.edu; nnavin@mdanderson.org

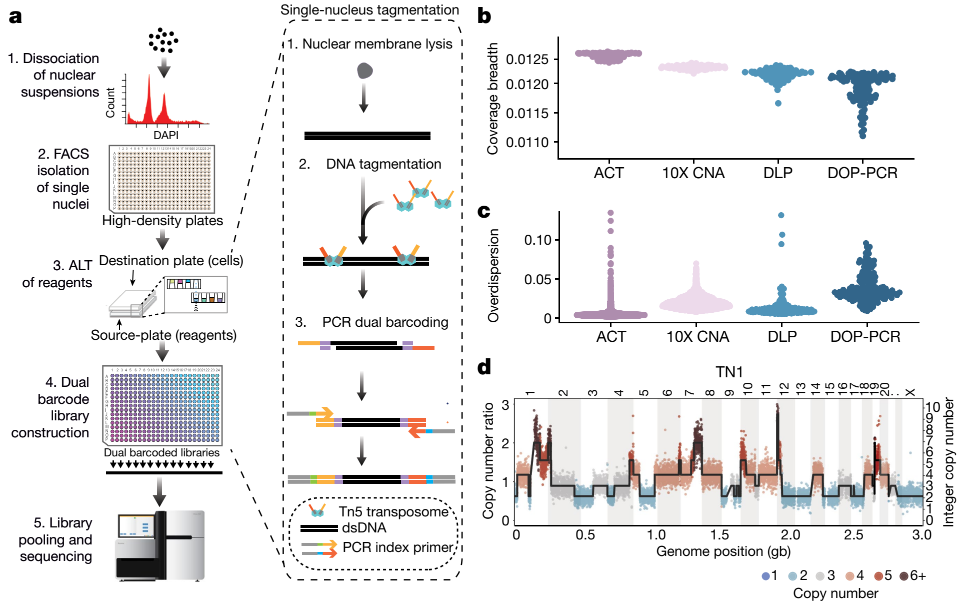


Fig. 1 | The ACT method and technical performance. a, Schematic of the ACT protocol, including the dissociation of nuclei from tissues, isolation of single nuclei into high-density 384-well plates by FACS, ALT of tagmentation reagents, PCR addition of dual barcodes and pooling of single-cell libraries for multiplexed sequencing. dsDNA, double-stranded DNA. **b**, Breadth of coverage for sparse scDNA-seq data from four different methods, including ACT, 10X Genomics CNV, DLP and DOP-PCR using 100 sampled cells. **c**, Overdispersion of bin counts in sparse scDNA-seq data from ACT, 10X Genomics CNV, DLP and DOP-PCR using 100 sampled cells. **d**, Copy number ratio (dots) and segmentation plots (line) for a single cell from sample TN1.

Data Fig. 1a). ACT has several advantages over first-generation scDNA-seq methods[8] that rely on whole-genome amplification steps, including fewer experimental steps and a shorter time frame (reduced from 3 days to around 3 h), increased cell throughput and the ability to measure single-molecule DNA information by positional barcoding (Extended Data Fig. 1b).

Technical properties of single-cell data

The technical performance of ACT was evaluated by comparing sparse data (about 1 million reads per cell) with those from three other scDNA-seq methods, including a microdroplet platform (10X Genomics CNV), two datasets previously generated using the direct library preparation (DLP) method[16] and data from a first-generation scDNA-seq method (DOP-PCR)[7,8]. We evaluated the coverage breadth and technical noise by overdispersion, which showed that ACT achieved a significant improvement ($P < 0.05$, Kruskal–Wallis test) over the other three methods (Fig. 1b, c, Extended Data Fig. 1c, d, Supplementary Methods). To further evaluate the coverage performance of ACT data, we sequenced two SK-BR-3 breast cancer cells at high depth (8.28× and 7.72×). To avoid the influence of copy number changes on coverage, we restricted our analysis to two diploid regions on chr4p and chr10q (Extended Data Fig. 1e, f). We compared the read counts of genomic bins, in which the duplicate molecules were retained or removed by positional barcoding, revealing an increase in uniformity in the single-molecule data and at most one or two reads for most genomic regions, whereas the duplicate-retained data had higher (8×) mean coverage depths (Extended Data Fig. 1f). From these data, we estimated that 97% of the reads were resolved to a single-molecule depth of 1 or 2. Lorenz curves showed that the coverage uniformity of the ACT single cells (Gini coefficient, $G = 0.728$ and 0.678) is similar to that of bulk DNA sequencing (DNA-seq) data ($G = 0.678$) and is more uniform than DOP-PCR data ($G = 0.957$) (Extended Data Fig. 1g). The physical coverage of the two SK-BR-3 cell libraries showed saturation on nearly 50% of the genome (Extended Data Fig. 1h). Finally, we observed that the genomic bin count data (220-kb resolution) are distributed close to those of the integer copy number segments, as exemplified in a representative aneuploid cell from a ductal carcinoma in situ (DCIS) tumour (TN1) (Fig. 1d). These findings led us to conclude that ACT represents a technical improvement over existing scDNA-seq methods.

Copy number substructure of tumours

We applied ACT to sequence 9,765 cells from 8 TNBC tumours, including the TN1 DCIS sample, three untreated invasive ductal carcinoma (IDC) tumours (TN2, TN6 and TN7), and four untreated synchronous DCIS–IDC samples (TN3–TN5 and TN8) (Supplementary Table 1). Nuclear suspensions were generated from frozen tissues and flow-sorted by ploidy distributions ranging from 2.65–3.95N, suggesting that whole-genome duplication (WGD) events had probably occurred in all of the tumours (Extended Data Fig. 2a, Supplementary Table 1). Clustering of the ACT data identified 7–22 subclones that were organized into 3–5 superclones across the 8 tumours (Fig. 2a, b). We define 'subclones' as clusters of cells that share highly similar copy number profiles, representing a clonal expansion from a single genotype, and 'superclones' as a higher-order organization of subclone groups that share a subset of copy number aberration (CNA) events. TN3 and TN5 showed the lowest number of subclones, whereas the remaining tumours had higher subclone numbers and genomic diversity indices (Fig. 2b, Extended Data Fig. 2b).

We define CNAs as segments of the genome in which two sets of chromosome breakpoints have increased or decreased integer copy number values relative to the ground state or 'neutral' copy number that corresponds to the mean DNA ploidy of the tumour (Methods). CNA analysis identified three major classes on the basis of the frequency of the subclones in the population of tumour cells: (1) clonal CNAs (cCNAs) that were shared by all subclones; (2) subclonal CNAs (sCNAs) that occurred in a subset of the tumour cells and were present in two or more subclones; and (3) unique CNAs (uCNAs) that had exclusive copy number states or breakpoints in one subclone (Methods). Of note, the uCNAs represent a subclass of sCNAs with a unique copy number state at a given segment identified in only one subclone. The CNA classes varied across the tumours, with TN5 having the highest number of cCNA events and TN4 having the highest uCNA count (Fig. 2c). Most of the genomic

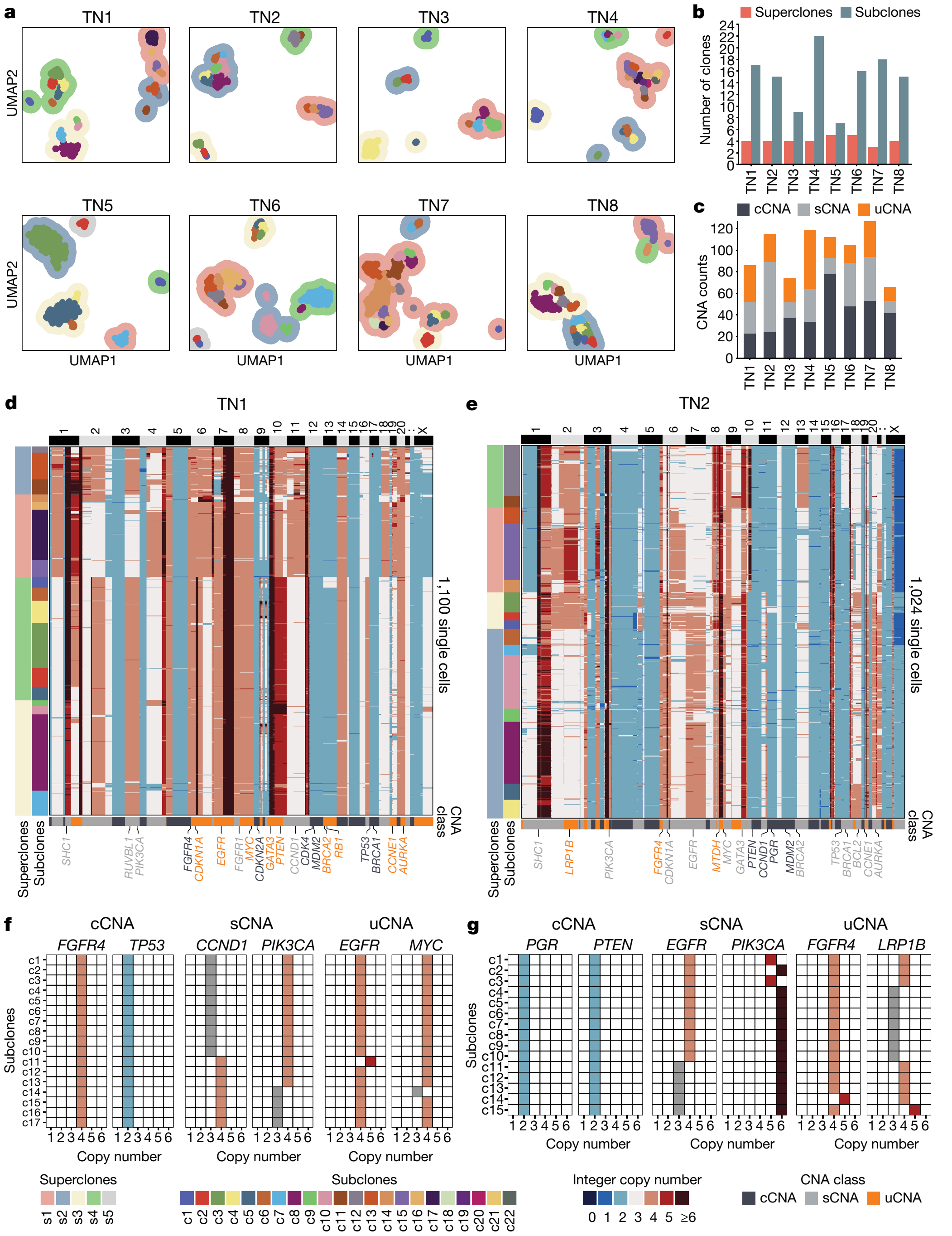


Fig. 2 | Clonal substructure of eight triple-negative breast tumours. **a**, High-dimensional uniform manifold approximation and projection (UMAP) clustering of single-cell copy number data from eight triple-negative breast tumours, in which contour colours represent superclones and coloured points represent subclones. **b**, Number of superclones and subclones detected in each tumour. **c**, Number of clonal, subclonal and unique CNAs detected in each tumour. **d**, **e**, Clustered heat maps of single-cell copy number profiles for TN1 (**d**; *n* = 1,100 cells) and TN2 (**e**; *n* = 1,024 cells). **f**, **g**, Integer copy number states of selected breast cancer genes for TN1 (**f**) and TN2 (**g**) according to clonal, subclonal and unique CNA classes.

regions of subclonal CNAs were not shared across individuals and the three CNA classes had similar genomic size distributions, with the exception of TN2 (Extended Data Fig. 2c, d). Furthermore, the fraction of cells with CNA gains, losses and copy-neutral (ground state) events showed variation across the subclones in each tumour (Extended Data Fig. 2e).

In sample TN1, the single-cell data revealed 17 subclones that were organized into 4 superclones (Fig. 2d). The superclones were distinguished by 29 sCNAs, whereas the subclones were distinguished by 34 uCNAs, of which many events intersected breast cancer genes (Fig. 2f). In sample TN2, the ACT data identified 15 subclones that were organized into 4 superclones (Fig. 2e). The superclones were distinguished by 65 sCNAs, whereas the subclones were distinguished by 26 uCNAs and intersected several breast cancer genes (Fig. 2g). Similarly, the 6 other TNBC tumours harboured a large number (7–22) of subclones that were organized into a few (3–5) major superclones (Extended Data Fig. 3).

To assess the robustness of subclone clustering, we performed bootstrapping, which showed that most clusters were stable (0.702 ± 0.15 (mean ± s.d.), Jaccard similarity) (Extended Data Fig. 2f). These data further revealed a relationship between the stability of a cluster and

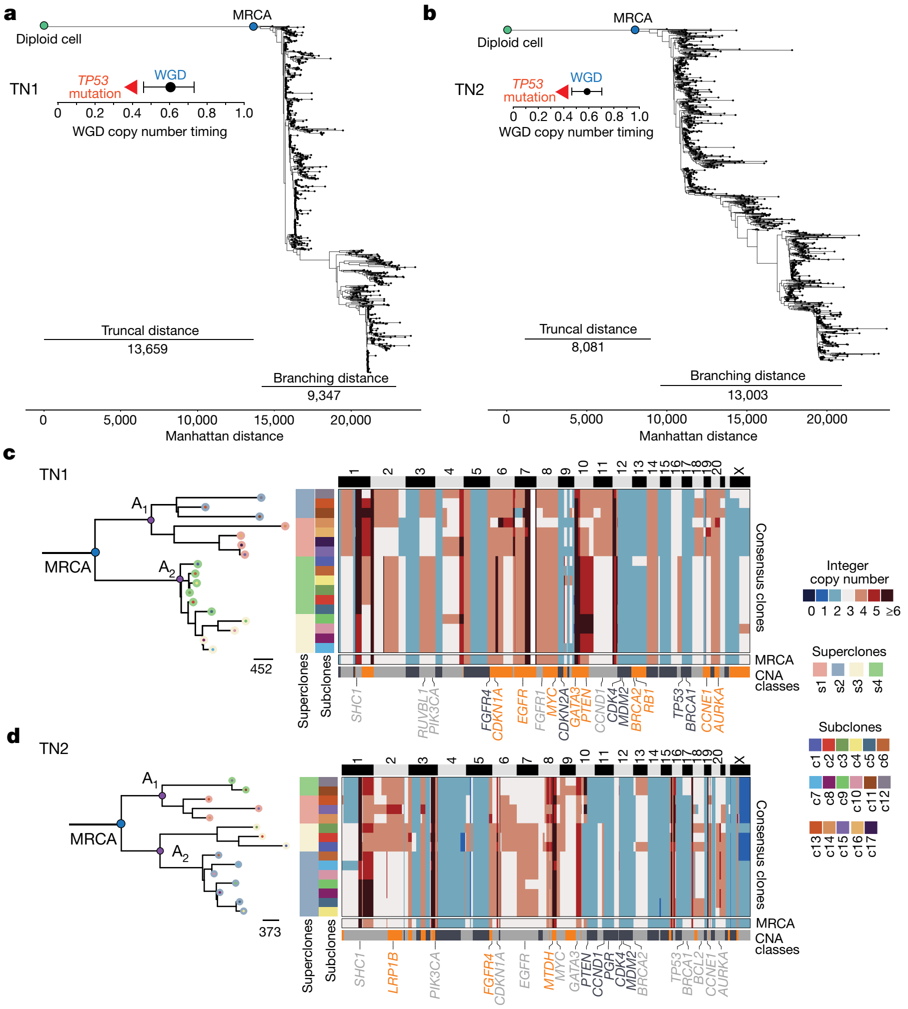


Fig. 3 | Evolutionary analysis of clonal lineages in patients with TNBC. a, b, Minimum evolution trees of single-cell copy number data for TN1 (**a**) and TN2 (**b**), with annotations indicating the time of the WGD events and confidence intervals, as well as the timing of *TP53* mutations, **c**, **d**, Left, minimum evolution trees after the MRCA, generated using consensus CNA profiles of subclones for TN1 (**c**) and TN2 (**d**) and rooted by a neutral node to the MRCA, with common ancestors (A_1, A_2). Right, consensus copy number profile heat maps of subclones of TN1 (**c**) and TN2 (**d**), in which bottom rows represent the inferred MRCA profile and different CNA classes.

the number of cells (Extended Data Fig. 2g). To orthogonally validate the clonal substructure, we performed scDNA-seq of 1,946 cells from two tumours using a different platform (10X Genomics CNV; Methods). The 10X data validated our ACT copy number state distributions and showed that all subclones were composed of a mixture of cells from both platforms, suggesting a high concordance across the orthogonal technologies, despite some variation in the clonal frequencies (Extended Data Fig. 4; Methods).

Clonal lineages during evolution

We next reconstructed the evolution of CNAs before the expansion of the primary tumour mass. Exome sequencing was performed on 8 tumours (107× mean depth) and matched normal tissues (76.3× mean depth), which showed a median of 102 somatic mutations, including *TP53* driver mutations in all tumours (Extended Data Fig. 5a, b, Supplementary Table 2; Methods). To infer the evolutionary history of the tumours up to the most recent common ancestor (MRCA), we classified mutations as either clonal or non-clonal (Extended Data Fig. 5b; Methods). We then selected clonal mutations and copy number changes to reconstruct which events occurred before versus after WGD in seven tumours (Methods). The resulting data showed that *TP53* mutations occurred consistently before WGD in seven tumours and that WGD occurred late in mutational time in most (five out of seven) individuals (Extended Data Fig. 5c).

To investigate tumour evolution after the MRCA, we used the ACT data to infer phylogenetic trees (Fig. 3a, b, Extended Data Fig. 5d). While, as expected, a large number of CNAs were clonal[7], the resulting trees further revealed branching lineages with large distances after the MRCA. Notably, the branching distances from the MRCA to the extant node (11,193 ± 4,106 (mean ± s.d.)) were similar to the truncal distances from the root diploid node to the MRCA (10,063 ± 2,504 (mean ± s.d.); $P = 0.52$, two-sided *t*-test), suggesting ongoing copy number evolution after the MRCA in all eight tumours (Extended Data Fig. 5e).

We then performed a more detailed analysis of the branching phylogenies after the MRCA by computing consensus CNA profiles of the subclones to construct balanced minimum evolution trees (Fig. 3c, d, Extended Data Fig. 6a). In TN1, the MRCA underwent an initial lineage split leading to 2 ancestral clones (A_1 and A_2) that further diverged into 4 clades corresponding to the 4 superclones that split into 17 distinct subclones (Fig. 3c). Similar branching phylogenies were observed after the MRCA in the seven other individuals (Fig. 3d, Extended Data Fig. 6a). In addition, we merged single-cell data by superclone groups and computed allele-specific copy numbers, which showed that most loss-of-heterozygosity (LOH) regions were consistent with the bulk exome data (median 96.1% region overlap), suggesting that they occurred before the MRCA (Extended Data Fig. 6b; Methods). On average, 41.21% of the genome (range 18.1–59.8%) showed LOH events in the 8 individuals. Collectively, these data show a large number of sCNA and uCNAs that were acquired after the MRCA, continuing to diversify the clonal genotypes during the expansion of the primary tumour mass.

Mathematical modelling of evolution

We next aimed to quantitatively investigate two alternative models of genomic evolution: a model in which the PCNE event is followed by the

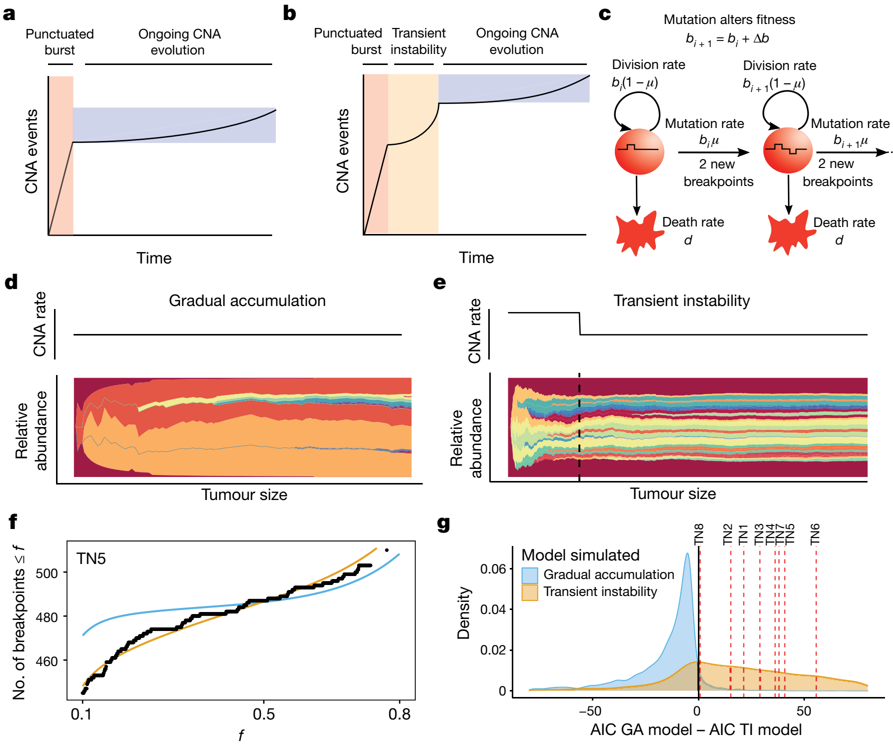


Fig. 4 | Mathematical modelling of transient instability after punctuated copy number evolution. a, **b**, Representations of CNA accumulation under gradual evolution after the punctuated burst (**a**) or transient instability after the punctuated burst (**b**). **c**, Schematic of the branching model for the chromosome-breakpoint accumulation that incorporates cell fitness and cell birth and death rates; replicating cells acquire heritable breakpoints in their copy number profiles with a probability that depends on the tumour size in the transient instability scenario, or is constant under gradual evolution. **d**, **e**, Muller plots of clonal frequencies obtained from stochastic simulations of the gradual accumulation model (**d**) and the transient instability model (**e**). **f**, Maximum-likelihood fits for the chromosome-breakpoint frequency spectra obtained for TN5 under both scenarios (colours as indicated in **g**). **g**, Difference of AICs for the transient instability (TI) and gradual accumulation (GA) models from simulated data from a large parameter range; difference of AICs obtained from the single-cell data indicated by the red lines.

gradual accumulation of CNAs at a constant baseline rate; and a model in which the PCNE event leads to a transient period of elevated genomic instability, followed by a return to gradual evolution at a constant baseline rate (Fig. 4a, b). To describe the accumulation of chromosomal breakpoints, we used a stochastic branching-process model (Fig. 4c; Supplementary Methods). To model transient instability, we considered the CNA rate to be elevated until the tumour exceeds a threshold size, after which the rate decreases to a baseline value (Fig. 4d, e). The alternative, gradual model assumes that the CNA rate remains at the baseline value. All else being equal, transient instability would lead to an enrichment of high-frequency breakpoints (that is, in many cells). To investigate these scenarios, we derived formulas for the number of breakpoints expected to be present at a given frequency for both cases (Extended Data Fig. 7a; Methods, Supplementary Methods). We then embedded these formulae into a likelihood framework incorporating breakpoint-detection errors, which enabled a quantitative assessment of which scenario provides a superior fit to the ACT data. We used the Akaike information criterion (AIC) obtained under each scenario as a summary statistic, which we validated on simulated data and was observed to be conservative for calling transient instability. Applying our method to the eight TNBC tumours, we obtained a lower AIC for the transient instability model for all eight cases, suggesting that an early elevated CNA rate is more likely (Fig. 4f, g, Extended Data Fig. 7b). These results indicate that a transient period of elevated genomic instability early in tumorigenesis explains the patient data better than a gradual evolution model.

Copy number substructure of cell lines

We next investigated whether the extensive copy number diversity observed in human TNBC tumours also exists in TNBC cell lines. We selected four TNBC cell lines with *TP53* mutations and aneuploid karyotypes[17] (MDA-MB-231, BT-20, MDA-MB-157 and MDA-MB-453) and applied ACT to sequence a total of 6,413 cells, after which clustering was used to delineate their clonal substructure (Fig. 5a, Extended Data Fig. 8a, b). Similar to the primary tumours, the four cell lines showed 11–20 subclones, organized into 3–5 superclones (Fig. 5b, c, Extended Data Fig. 8a–c). Furthermore, the Shannon diversity indices and frequencies of cCNA (47.3%), sCNA (27.4%) and uCNA (25.3%) events were in a similar range to the TNBC tumours, as were the segment size distributions (Extended Data Fig. 8d–f). To validate the subclonal copy number states, we designed probes to target 9 breast cancer genes in MDA-MB-231 and performed DNA-FISH to quantify the copy number values for a similar number of cells (N = 1,000) that were sequenced by ACT, confirming the clonality of all CNA events detected (Extended Data Fig. 8g, h; Methods). Collectively, our data suggest that these cell lines are representative of the copy number substructure of human TNBC tumours.

Estimating copy number evolution rates

To estimate the rate of CNA evolution, we physically subcloned and expanded 2 single daughter cells (MDA231-EX1 and MDA231-EX2) from the MDA-MB-231 parental cell line for 19 cell doublings and measured the number of de novo CNA events that were acquired (Fig. 5d; Methods). These data showed that the 2 expanded daughter cells rediversified their genomes into 7–12 subclones in the time it took a single cell to fill a 10-cm culture plate (Fig. 5e, f). During the two expansions, 7 sCNAs and 9 uCNAs were acquired in MDA231-EX1, while 5 sCNAs and 10 uCNAs were acquired in MDA231-EX2 (Fig. 5g, Extended Data Fig. 8d, i). In contrast to the parental TNBC cell lines, the new expansions showed fewer sCNA events compared with cCNAs and uCNAs (Extended Data Fig. 8d). We used the chromosome-breakpoint data from the expanded cells to estimate the de novo CNA rate per cell division[18], and obtained an average rate of 0.242 CNAs per cell division (0.235, 95% confidence interval (0.189, 0.288) for EX1 and 0.249, 95% confidence interval (0.204, 0.3) for EX2) (Methods). Our mathematical modelling framework showed that, in contrast to the primary tumours, a gradual model was more likely to explain the data from both cell-line expansions (Extended Data Fig. 7c). These data show that single cancer cells do not maintain a stable clonal genotype after expansion, even during a relatively short time frame.

Effect of subclonal CNAs on gene dosage

We further investigated whether the subclonal CNAs resulted in gene dosage effects that influenced gene expression levels by expanding

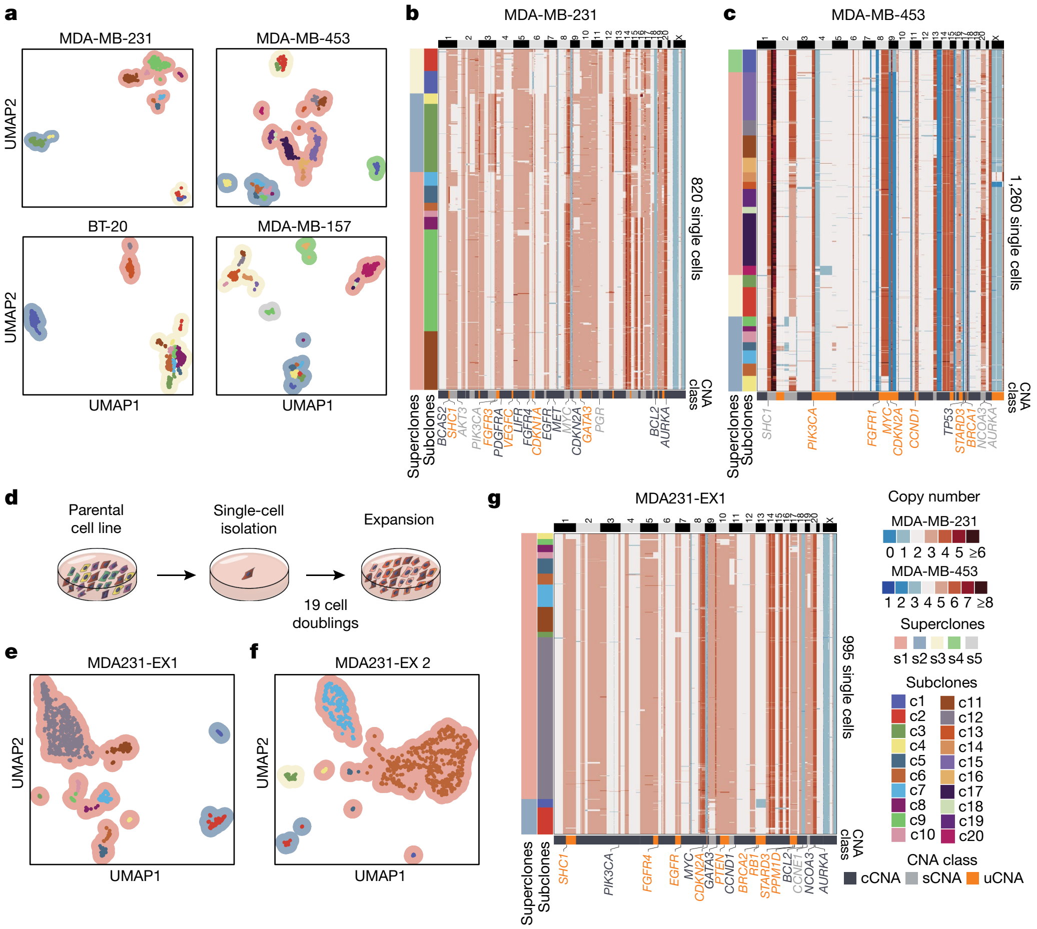


Fig. 5 | Clonal substructure of TNBC cell lines and single-cell expansions. **a**, UMAP and clustering of single-cell copy number data from four TNBC cell lines, including MDA-MB-231 (n = 820 cells), MDA-MB-453 (n = 1,260 cells), BT-20 (n = 1,231 cells) and MDA-MB-157 (n = 1,210 cells), in which contour colours represent superclones and coloured points represent subclones. **b**, **c**, Clustered heat maps of ACT data from the MDA-MB-231 (**b**) and MDA-MB-453 (**c**) cell lines. **d**, Schematic of subcloning experiments for expanding single daughter cells from the parental MDA-MB-231 cell line. **e**, **f**, High-dimensional UMAP clustering of single-cell copy number data from two expanded daughter cell populations after 20 cell doublings. **g**, Clustered heat maps of ACT data from the EX1 expanded cells from MDA-MB-231, with CNA classes indicated below.

78 single daughter cells (e1–e78) from MDA-MB-231 for 19 generations and performing matched bulk DNA-seq and RNA sequencing (RNA-seq) (Extended Data Fig. 9a; Methods). By co-clustering the bulk DNA-seq data with the ACT data (820 cells), we found that 10 out of 13 of the subclones in the parental MDA-MB-231 cell line were reflected in the expansions, which we refer to as expanded clusters (Extended Data Fig. 9b–d). Principal component analysis of the expanded clone bulk RNA-seq data alone revealed groups of expansions that corresponded to the superclone genotypes (Extended Data Fig. 9c). A global analysis of CNA events across the entire genome showed that copy number states in MDA-MB-231 were significantly correlated ($R^2 = 0.45$, $P < 2.2 \times 10^{-16}$) with gene expression levels (Extended Data Fig. 9e; Methods). Similarly, when this analysis was restricted to subclonal regions, we found that 68% of chromosome segments were significantly associated with expression changes ($P < 0.05$, Kruskal–Wallis test), as exemplified in selected CNA regions (Extended Data Fig. 9f, g). We further investigated the effects of subclonal CNAs across larger chromosomal regions, which showed that 100-gene expression windows tracked well with subclonal copy number changes and affected the expression of many cancer genes (Extended Data Fig. 9h, i; Methods). Beyond the localized effects of gene dosage, the subclonal CNA events also had a broader effect on the expression of many genes in pathways and cancer hallmark signatures[19] across the entire genome (Extended Data Fig. 9j).

Discussion

Our data show that the copy number substructure of human TNBC tumours consists of a large milieu of subclones (7–22) that are organized into a few major superclones (3–5) and share a common evolutionary lineage. Although the number of superclones is consistent with previous studies of breast cancer[7,8,20], the number of subclones vastly exceeds previous estimates. Our study extends previous findings of TNBC evolution[7] by showing that *TP53* mutations, genome doubling and extensive LOH are important early evolutionary events that occurred before the MRCA. Our data further show that after the MRCA, a period of transient instability generates a large number of subclones before transitioning to a basal rate of ongoing copy number evolution that persists during the expansion of the primary tumour mass. These data suggest that while there may be some stabilizing selection[21], the tumour cells continue to explore the fitness landscape during the growth and expansion of the primary tumour. On the basis of these results, we propose a revised model for TNBC evolution after PCNE (Extended Data Fig. 10).

By sequencing DNA and RNA from the same expanded subclones, we showed that the subclonal CNAs can influence gene expression, consistent with bulk CNA and RNA data across many human cancers[22]. By expanding single daughter cells in vitro, we showed that cancer cells can quickly rediversify their genomes at a rate of approximately one new CNA per four cell divisions. Our results are consistent with a previous study that reported extensive copy number and mutational evolution during the passaging and subcloning of cancer cell lines[23]. These data serve as an important warning for the research community, namely that isogenic subcloning, a widely used procedure in molecular biology[24], can still result in heterogeneous cell populations when used in downstream functional assays.

ACT represents a major technical improvement over first-generation scDNA-seq methods[8,25]. A few other studies have also implemented tagmentation-based approaches to perform scDNA-seq, including two lower-throughput methods using microfluidic chips (around 100 cells)[16,26], and one high-throughput method that was scaled up using a nanowell system[27]. Another study developed a combinatorial-indexing approach that uses tagmentation and is highly scalable but has limited genomic resolution[28]. Other work has developed a whole-genome amplification-based approach on a microdroplet platform (10X Genomics CNV) that is scalable but does not achieve single-molecule resolution. Compared with these methods, ACT represents an improvement in technical performance and is cost-efficient.

A notable limitation to our study is that the number of subclones that we detected is an 'operational definition' and is dependent on the total number of cells that are sequenced, and therefore probably represents an underestimate of clonal diversity. Finally, we postulate that PCNE and subclonal reservoirs may not be unique to patients with TNBC and may exist in other solid tumours, particularly in aneuploid cancers that harbour *TP53* mutations. Beyond cancer, we expect that ACT will have broad applications for investigating aneuploidy in diverse fields of biology and biomedicine.

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/s41586-021-03357-x.

1. Davis, A., Gao, R. & Navin, N. Tumor evolution: linear, branching, neutral or punctuated? *Biochim. Biophys. Acta Rev. Cancer* **1867**, 151–161 (2017).
2. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
3. Pfister, K. et al. Identification of drivers of aneuploidy in breast tumors. *Cell Rep*. **23**, 2758–2769 (2018).
4. Xu, J., Huang, L. & Li, J. DNA aneuploidy and breast cancer: a meta-analysis of 141,163 cases. *Oncotarget* **7**, 60218–60229 (2016).
5. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nat. Rev. Genet*. **13**, 189–203 (2012).
6. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
7. Gao, R. et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet*. **48**, 1119–1130 (2016).
8. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
9. Cross, W. et al. The evolutionary landscape of colorectal tumorigenesis. *Nat. Ecol. Evol*. **2**, 1661–1672 (2018).
10. Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
11. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol*. **30**, 413–421 (2012).
12. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet*. **45**, 1134–1140 (2013).
13. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
14. Cross, W. Ch., Graham, T. A. & Wright, N. A. New paradigms in clonal evolution: punctuated equilibrium in cancer. *J. Pathol*. **240**, 126–136 (2016).
15. Hadimioglu, B., Stearns, R. & Ellson, R. Moving liquids with sound: the physics of acoustic droplet ejection for robust laboratory automation in life sciences. *J. Lab. Autom*. **21**, 4–18 (2016).
16. Zahn, H. et al. Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods* **14**, 167–173 (2017).
17. Chavez, K. J., Garimella, S. V. & Lipkowitz, S. Triple negative breast cancer cell lines: one tool in the search for better treatment of triple negative breast cancer. *Breast Dis*. **32**, 35–48 (2010).
18. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet*. **48**, 238–244 (2016).
19. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. **1**, 417–425 (2015).
20. Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
21. Cross, W. et al. Stabilising selection causes grossly altered but stable karyotypes in metastatic colorectal cancer. Preprint at https://doi.org/10.1101/2020.03.26.007138 (2020).
22. Fehrmann, R. S. et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet*. **47**, 115–125 (2015).
23. Ben-David, U. et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).
24. Greenfield, E. A. Single-cell cloning of hybridoma cells by limiting dilution. *Cold Spring Harb. Protoc*. https://doi.org/10.1101/pdb.prot103192 (2019).
25. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
26. Xi, L. et al. New library construction method for single-cell genomes. *PLoS ONE* **12**, e0181163 (2017).
27. Laks, E. et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179**, 1207–1221 (2019).
28. Vitak, S. A. et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017).

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

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Human samples

The 8 breast tumour samples were obtained as frozen de-identified samples from the MD Anderson Breast Tissue Bank under an Institutional Review Board (IRB)-approved protocol. All individuals consented to have their tissue used for research studies. The triple-negative status of the tumour samples was determined by immunohistochemistry for oestrogen receptor (<1%) and progesterone receptor (<1%), and FISH analysis of *HER2* (also known as *ERBB2*) amplification using the centromere control probe CEP-17 (ratio of *HER2*/CEP17 < 2.2). TN1 was classified as DCIS by histopathology, while all other samples were invasive ductal carcinomas or synchronous DCIS-IDC (Supplementary Table 1). Most of the tumour samples were untreated, with the exception of TN1 which was treated with adriamycin cyclophosphamide before the collection of the tissue sample. Approximately 0.5 × 0.5 × 0.5 cm of total tissue was used in each experiment, combining macrodissected pieces from multiple sectors in each tumour. More information on the tumour sizes, grades and histopathology are provided in Supplementary Table 1.

Cancer cell line samples

The TNBC breast cancer cell lines were obtained from the Characterized Cell Line Core (CCLC) Facility at the University of Texas MD Anderson Cancer Center. The cell line identities were confirmed by restriction fragment length polymorphism analysis and sparse whole-genome sequencing to determine copy number profiles. All cell lines tested negative for mycoplasma contamination before running the experiments.

Generation of expanded subclonal cell lines

Expanded clones from a parental MDA-MB-231 (80% confluency) were isolated by FACS (BD Melody) into 96-well flat-bottom culture plates containing 100 µl of cell culture medium, followed by visual confirmation by light microscopy after 0 and 24 h. Wells with multiple cells or doublets were eliminated, while wells with confirmed single cells were used for subsequent expansions. The single cells were propagated until ~80% confluency in a 10-cm dish, after which the cells were used for scDNA-seq or bulk DNA and RNA sequencing.

Isolation of single nuclei by FACS

Nuclear suspensions from frozen tumour tissue were prepared using an NST-DAPI lysis buffer as previously described[8,29]. Suspensions were filtered through a 40-µm mesh and single nuclei were flow-sorted (BD FACSMelody, BD FACS AriaII or Beckman MoFlo Astrios). The DAPI intensity was used to set gates on aneuploid cells populations for all tumours. Single nuclei from TN5 were sorted from the aneuploid G2M peak. Single nuclei were then deposited into individual wells of 384-well plates (Eppendorf 951020702). The sorting instrument alignment was assessed under a microscope before each experiment to ensure single nuclei were accurately deposited into the centre of each well using a film-bottom 384-well plate (Greiner 781091). After flow sorting, plates were spun at 1,500*g* for 4 min, sealed and stored at −20 °C until ready for ACT processing. Bulk nuclei were FACS sorted into LoBind tubes (Eppendorf 022431021) for 10X Genomics CNV or exome-capture reactions.

ACT procedure

FACS-sorted 384-well plates were spun at 1,500*g* for >4 min. The Echo525 system (Labcyte) was used to dispense tagmentation reagents (Illumina FC-131-1096) at nanolitre scale, with plate and liquid types detailed in the following steps. Thorough mixing and spinning of each plate after every dispense and incubation period is crucial to maximizing assay performance. Nuclei were lysed in 200 nl (384PP_SPHigh) of freshly prepared Tx Lysis buffer (Protease (1.36 AU ml^{-1}) diluted 1:9 in 5% Tween 20, 0.5% Triton X-100 and 30 mM Tris pH 8.0). Lysis thermocycler settings were programmed as: 55 °C for 10 min, 75 °C for 15 min, and hold at 4 °C, lid temperature 80 °C and volume 1 µl. After lysis, 600 nl of tagmentation reaction mixture (TD:ATM 2:1, 384PP-Plus_GPSA) was dispensed. The ACT reaction settings on the Thermocycler were: 55 °C for 5 min, hold at 4 °C, lid temperature 60 °C and volume 1 µl. The ACT reaction was neutralized with 200 nl (384PP_SPHigh) of NT buffer for 5 min at room temperature. The final PCR reaction included 1.11uM N7XX (5′-CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTCTCGTGGGCTCGG-3′) and S5XX (5′-AATGATACGGCGACCACCGAGATCTACACXXXXXXXXTCGTCGGCAGCGTC-3′) primers (384PP_AQBP) in 2X HiFi HotStart Ready Mix (Roche# KK2602, 6RES_GPSA). XXXXXXXX denotes dual barcode sequences in primers. Unique dual barcode combinations for each well in the 384-well plate were achieved by dispensing 16 unique N7XX barcodes across each row and 24 unique S5XX barcodes across each column (Supplementary Table 3). The PCR reaction was performed using the following conditions: 72 °C for 3 min, 98 °C for 30 s, (98 °C for 10 s, 63 °C for 30 s, 72 °C for 30 s) for 15–18 cycles, 72 °C for 5 min, hold at 4 °C, lid temperature 105 °C and volume 6 µl. ACT performance was evaluated by Qubit fluorometer and TapeStation (Agilent) from selected cell libraries. Final libraries were pooled together and purified with 1.8X AMPURE XP beads. The final libraries were sequenced at 50 or 76 single-read cycles with dual barcodes on the Illumina HiSeq4000 system.

10X Genomics CNV single-cell sequencing

Nuclear suspensions were stained with NST-DAPI and sorted by FACS. The DAPI intensity was used to set gates on aneuploid cell populations (see 'Isolation of single nuclei by FACS'). The resulting aneuploid nuclei suspensions were used as input material for the Chromium (10X Genomics CNV) single-cell DNA cell bead kit (cat. no. 1000056) as described in the user guide with a target capture of 1,000 cells using chromium single-cell chips C and D (cat. nos 1000022 and 1000042, respectively). DNA libraries were prepared using chromium single-cell DNA library and gel bead kit (cat. no. 1000040) and were sequenced at 200 cycles on the NovaSeq6000 S1 flowcell (Illumina).

Fluorescence in situ hybridization

MDA-MB-231 cells were cultured until 80% confluency in a 10-cm dish and transferred to 15-ml conical tubes and centrifuged at 1,500 rpm for 7 min. Cells were subjected to hypotonic treatment (0.075 M KCl) for 20 min at room temperature and fixed in methanol and acetic acid mixture (3:1 v/v) for 15 min, washed three times with the fixative and air-dried. DNA fluorescence in situ (DNA-FISH) hybridization was performed on the above cytological preparations using SHC1-20-GR, EGFR-20-GR, VEGFC-20-GR, PIK3CA-20-GR, AKT3-20-GR, FGFR3-20-GR, MET-20-OR, PDGFRA-20-OR and BCAS2-20-OR probes (Empire Genomics). Slides were hybridized with the FISH probes according to the manufacturer's instructions (Empire Genomics) with slight modifications. In brief, 2 µl of each of the two probes were mixed with 6 µl of the in situ hybridization buffer. The probe was applied on the slide and covered with a glass coverslip (22 × 22 mm) and sealed with rubber cement. The slides were then denatured at 72–73 °C using Thermobrite system (Abbott Laboratories) and incubated at 37 °C overnight. The slides were then washed using 2× SSC at 45–70 °C for 1–2 min, counterstained with DAPI and analysed using a Nikon 80i microscope on the green and orange fluorescent channels. The copy number states of each probe were counted across 1,000 cells and multiple imaging fields for each experiment.

Bulk DNA-seq and RNA-seq of MDA-MB-231

Expanded subclones from MDA-MB-231 were cultured until ~80% confluency in a 10-cm dish and split into triplicates. From each triplicate, a portion of cells was separated for DNA copy number analysis and a second portion was used for RNA extraction using TRIzol (Fisher, cat. no. 15596-018) from the same plates. Genomic DNA was isolated from each expanded subclone with the QIAamp DNA Blood Mini Kit (Qiagen, cat. no. 51106). Recovered DNA was sonicated to 250 bp using the S220 acoustic sonicator (Covaris) and libraries for each sample were prepared with the Kapa HyperPrep Kit (Roche, cat. no. KK8504) and NEXTflex-96 barcodes (Bioo Scientific). The NEBNext Ultra RNA library prep kit for Illumina with poly(A) mRNA magnetic isolation module (NEB, cat. nos E7530 and 7490) was used for the bulk RNA libraries according to the manufacturer's instructions. The protocol was modified to include the NEXTflex-96 barcodes with 14 PCR cycles. DNA-seq and RNA-seq libraries were sequenced on 76 paired-end cycles on the Illumina HiSeq4000 platform.

Bulk DNA exome capture

Genomic DNA from aneuploid tumour nuclei sorted by FACS (see 'Isolation of single nuclei by FACS') was isolated using Qiagen DNA blood mini kit (cat. no. 51106) and matched normal tissue genomic DNA was isolated using Qiagen DNA micro kit (cat. no. 56304). Recovered DNA was sonicated to 250 bp using a S220 acoustic sonicator (Covaris) and libraries for each sample were prepared with the Kapa HyperPrep Kit (Roche cat. no. KK8504) and NEXTflex-96 barcodes (Bioo Scientific), purified with 0.8X AMPure XP beads and amplified by PCR following the manufacturer instructions. Exome libraries were captured with SeqCap EZ Exome V2 kit following the manufacturer's instructions (Roche cat. no. 05860482001) and sequenced with 100 paired-end kits on HiSeq4000 or NextSeq2000 300 cycles kit (Illumina).

Inference of DNA copy number

Sequencing reads were demultiplexed into single-cell FASTQ files allowing 1 mismatch of the 8-bp barcode. FASTQ files were aligned to hg19 (NCBS build 36) using bowtie2 (v2.2.6)[30] and converted from SAM to BAM files with SAMtools (v1.2)[31]. Positional barcoding was performed by marking fragments with equal start position as PCR duplicates and removed from subsequent analysis to obtain single-molecule data. Copy number profiles were inferred with the variable binning pipeline as previously described[7]. In brief, aligned reads were counted in variable bins averaging 220 kb. Bin counts were normalized for GC content with lowess regression and bin-wise ratios were calculated by computing the ratio of bin counts to the sample mean bin count. Segmentation was performed with circular binary segmentation (alpha = 0.0001 and undo.prune = 0.05) from R Bioconductor DNACopy package[32]. MergeLevels was applied to join adjacent segments with non-significant differences in segmented ratios. Cells with excessive noise were excluded according to the following criteria: (1) removal of cells with bin counts that were 2× s.d. below the mean, (2) removal of cells with large breakpoint counts that were 2× s.d. above the mean, and (3) removal of outliers using density-based spatial clustering R package dbscan (v1.1-5)[33] (minPts = 5, bucketSize = 10, $k = 5$, eps parameter was determined by the elbow method from the k-nearest neighbours distance matrix).

Calculation of technical metrics

The Gini coefficient for high-depth sequencing of single-cells from SK-BR-3 for ACT, DOP-PCR and bulk sample was calculated as follows. Let x_i be the set of depths observed and let n_i be the number of sites with depth x_i,

$$\text{Gini} = 1 - \frac{\sum_{i=1}^{n}\left(\frac{n_i}{\sum_{i'} n_{i'}}\right) \times (s_{(i-1)} + s_i)}{s_n}, \text{ where } \sum_{i'=1}^{i} n_{i'} x_{i'}.$$

Single-cell coverage breadth was calculated from BAM files with duplicates removed. We sampled 100-sparse-single-cell sequencing data from BAM files from each scDNA-seq method, ACT (TN1-TN4), 10X-CNA, DOP-PCR[34] and DLP[16] and downsampled the data to 800k reads trimmed to 50 bases to match the lowest read length and depth across all samples. Coverage from all sites was calculated using bedtools (v2.26.0) genomeCoverageBed[35]. Overdispersion was calculated by the index of dispersion of bin counts, that is, the variance over mean, normalized by the mean bin counts for each single cell. Let ϕ be the overdispersion parameter, b be the mean bin counts and iod the index of dispersion, $\phi = (\text{iod} - 1)/\text{mean}(b)$.

Multi-sample segmentation and integer copy number estimation

We used the R bioconductor package with the 'copynumber' (v1.26) function 'multipcf' (gamma = 30)[36] to perform joint segmentation and determine common break points for all single cells on the bin count matrices with an added pseudocount of 5, followed by 'MergeLevels' to join adjacent segments with non-significant differences in segmented ratios. Average tumour ploidy was calculated with DAPI fluorescence values from FACS data. The first peak from the DAPI fluorescence histogram was assumed to be normal (2N) diploid stromal cells. The ratio of the mean DAPI fluorescence from the gated aneuploid population over the mean DAPI fluorescence of the 2N population was multiplied by 2, resulting in the average tumour ploidy, that is, ground state. Segment ratios from joint segmentation were multiplied by the FACS-derived average tumour ploidy and rounded to the nearest integer value.

Clustering of superclones and subclones

Integer single-cell copy number data from multi-sample segmentation was embedded in two dimensions using UMAP[27,37] with R package 'uwot' (v.0.1.8, min dist = 0, n neighbours = 40, seed = 55 for TNBC tumours and n neighbours = 25, seed = 206 for cell-lines, distance = "manhattan"). To identify superclones, the resulting embedding was used to create a shared nearest neighbour (SNN) graph with R Bioconductor package 'scran' (v1.14.6)[38]. For each superclone SNN graph, different k values were used (TN1, 45; TN2, 63; TN3, 65; TN4, 75; TN5, 41; TN6, 51; TN7, 35; TN8, 43; MDA-MB-231, 93; MDA-EX1, 55; MDA-EX2, 17; BT-20, 55; MDA-MB-453, 65; MDA-MB-157, 75), the connected components of the SNN graph were identified using the R package 'igraph' (v1.2.5)[39] and classified as superclones. To identify subclones the UMAP embedding was used as input for the clustering algorithm hdbscan (minPts = 17 for TNBC tumours and 15 for cell lines) from R package 'dbscan' (v1.1-5)[27,40]. Hdbscan is an outlier aware clustering algorithm, since extensive filtering of the dataset was applied before clustering (see 'Inference of DNA copy number'), any cell classified as an outlier was inferred to the same cluster group as its closest, non-outlier, nearest neighbour according to Euclidean distance. Subclones were further organized with hierarchical clustering (Manhattan distance, ward.D2 linkage), further substructures identified by hierarchical clustering were not considered additional subclones. Jaccard similarity for clusters was computed by bootstrap with R package 'fpc' (v2.2-7) with mean Jaccard similarities being reported. Heat maps were plotted with R package ComplexHeatmap (v2.2.0)[41]. Clonal structure on heat maps was organized according to the clonal lineage from the subclonal consensus copy number profiles (see 'Calculating consensus copy number profiles of subclones' and 'Phylogenetic reconstruction of single-cell and clonal lineage trees').

Co-clustering of ACT and 10X Genomics CNV single-cell data

ACT and 10X genomics single-cell CNV resulting bin counts were merged and co-segmented with multipcf (gamma = 30) (see 'Multi-sample segmentation and integer copy number estimation'), followed by MergeLevels to join adjacent segments with non-significant differences in segmented ratios. Segment ratios were scaled by tumour FACS-inferred ploidy and rounded to the nearest integer. Co-clustering of ACT and 10X genomics single-cell CNV datasets was performed as previously described with hdbscan and parameters adjusted to match the original number of subclonal populations from ACT clustering (seed = 55, n neighbours = 40, minPts = 35, 80 for TN1 and TN3, respectively) (see 'Clustering of superclones and subclones from single-cell copy number data').

Calculating consensus copy number profiles of superclones and subclones

For each tumour sample, the integer copy number consensus profiles were calculated by taking the median of the ith segment of all single cells assigned to the same superclone or subclone, the ploidy was scaled by the average tumour ploidy derived by FACS and rounded to the nearest integer value.

Inference of most recent common ancestral profile

The consensus profile of each superclone (see 'Calculating consensus copy number profiles of subclones and superclones') was used to derive the most recent common ancestor (MRCA). For every segment, we selected the copy number (CN) value among the consensus CN values from each superclone that is closest (L1 norm) to the average tumour ploidy as the ancestral segment.

Classification of clonal, subclonal and unique cna segments

cCNA and sCNA segments were identified from the subclonal consensus matrices. sCNAs were further classified into uCNAs if one subclone presented at least one distinct copy number event compared to all others, formally:

Let n_i be the frequency of subclones CNA$_i$ is in.
Let N be the total number of consensus subclones for the sample.
cCNAs are defined as $n_i = 1$
sCNAs are defined as $1/N < n_i < 1$
uCNAs are defined as $n_i = 1/N$

Construction of CNA breakpoint spectra

To construct a frequency spectrum of CNAs using breakpoint frequencies across all single cells, we performed segmentation with the R package 'Piet' (GFL) (v0.1.0)[42] (rho1 = 0, rho2 = 0, rho3 = 70, obj_c = 10^-10, max_iter = 1^5). A matrix of log ratios from the variable binning copy number pipeline (see 'Inference of DNA copy number') and bin-wise variance estimation where: let $x(i)$ be the log ratio bin count at bin i, the variance estimate is $\text{median}\,((x(i+1)-x(i))^2)\Big/\left(2\times\left(1-\frac{2}{9}\right)^3\right)$, was used as input for GFL. GFL returns piecewise constant curves with discontinuities across breakpoints. To account for discontinuities, we built interval estimates at intersecting breakpoints and constructed a graph to verify overlap across genomic positions over all single cells. Discontinuities higher than 10 bins were discarded and connected components were obtained from the resulting graph. Breakpoints that did not reach a ratio difference ≥0.6 between the median of two adjacent segments were not counted. Accuracy of resultant breakpoint frequency calls were assessed by simulation (Supplementary Methods). Resulting segments were ploidy scaled by the average FACS-derived ploidy and rounded to the nearest integer values. Finally, we counted the frequency of each chromosome breakpoint across all cells from the sample resulting in a frequency spectrum.

Calculation of subclonal diversity indexes

For each tumour sample we calculated the proportion (p) of cells that belong to a distinct subclone. Diversity was calculated as Shannon index: $D_c = -\sum_i(p_i \times \ln(p_i))$, with 95% confidence intervals calculated by bootstrapping (B = 3,000).

Phylogenetic reconstruction of single-cell and clonal lineage trees

Pairwise distances of single cells were calculated using Manhattan distance to obtain a distance matrix for each tumour. Phylogenetic inference for single-cell trees and consensus trees were performed with the balanced minimum evolution algorithm[43] from R package ape (v5.3)[44]. Root diploid nodes for phylogenetic inference were constructed from simulated variable binning profiles in which bins presented an integer copy number equal to 2. Distances were calculated from the diploid root to the most recent common ancestral (MRCA) and from the MRCA to the terminal aneuploid node. Terminal aneuploid node was defined by the largest branch length from the MRCA on the aneuploid subtree. Consensus phylogenetic trees were rooted from simulated variable binning profiles equal to the integer average tumour ploidy (see Supplementary Table 1, 'ploidy'). Root nodes from consensus phylogenetic trees were removed for visualization purposes. Trees were plotted using R package ggtree (v2.0.3)[45].

Mathematical modelling of CNA evolution

A branching process model for the accumulation of chromosomal breakpoints was used, in which a tumour cell can replicate, die, or replicate such that one of the daughter cells acquires two new breakpoints in its copy number profile and its ability to replicate is altered according to a fitness distribution. Under a reduced fitness distribution considering neutral and lethal aberrations only, we derived formulas for the expected number of breakpoints present at a given frequency, which were used in a likelihood analysis to determine whether an elevated breakpoint rate early in tumour growth provided a superior explanation of the data. Full details are given in Supplementary Methods, 'Mathematical modelling'.

Estimation of cell doubling rates

The expanded subclones were grown from a single cell (I = 1) to a 90% confluent 10-cm cell culture dish. MDA-MB-231 EX1 and EX2 remained in culture for 26 days (t), reaching a final number of ~5.86×10^5 cells (F). Doubling time (Dt) of the expanded subclones was calculated as: $\text{Dt} = (t \log 2)/(\log F - \log I)$ and number of generations (G) of cell divisions in each expanded population of cells was determined by $G = t/Dt$.

Estimation of the de novo copy number rates

Estimation of de novo copy number rates was carried out with intra-arm breakpoints, and do not include arm level events (see 'Construction of CNA breakpoint spectra'). We assume exponential expansions and no cell death. For expansion i let the number of cells sequenced be $n(i)$. Further let the number of CNAs expected in the frequency range $[2/n(i), 0.5)$ be nCNA(i). Then an analytic formula, which contains the CNA rate as a prefactor, can be obtained for the expectation of nCNA(i) (E[nCNA(i)) (Supplementary Methods). Assuming each new CNA leads to two new breakpoints, we adopt the statistical model that the number of breakpoints at frequencies $[2/n(i), 0.5)$ is Poisson distributed with parameter $2E$[nCNA(i)]. Further, we include that the probability of not observing a breakpoint present in y cells, which based on simulated data we approximated as $0.57 \times \exp(-7.5y \times 10^{-4})$ (the estimated rates decrease by a factor of ~2 without this assumption). For each expansion, the observed number of breakpoints in the frequency range $[2/n(i), 0.5)$ is called with Piet as described in 'Construction of the CNA breakpoint spectrum'. The point estimate for the CNA rate in each cell expansion is then calculated via maximum likelihood (Supplementary Methods, section 7) and the confidence intervals are based on the assumed Poisson distribution and obtained numerically.

Somatic mutation variant calling

Sequencing reads from bulk tumour tissue and matched normal tissues were demultiplexed into FASTQ files allowing 1 mismatch out of the 8-bp barcode. FASTQ files were aligned to hg19 (NCBS build 36) using bowtie2 (v2.2.6)[30], sorted and converted from SAM to BAM files with SAMtools (v1.2)[31]. Duplicates were marked with Picard tools (v2.20.4) and BAM files were recalibrated for base quality scores using Genome

Analysis Toolkit (GATK v4.1.3)[46] Base Recalibrator. Somatic variants from tumour tissue were identified with MuTect2[47] and filtered using GATK FilterMutectCalls. Bcftools (v1.11-3) was used to retain PASS variants. Additionally, variants with allele frequency higher than 0.05 in matched normal samples were excluded. Variants on bulk tissue required a minimum depth of 10×, 5× of the alternative allele and allele frequencies >0.1. Variants <1,000 base pairs apart were excluded from the analysis. VCF analysis was performed with the help of the R package 'vcfR' (v.1.12.0)[48]. Variants were annotated with ANNOVAR[49] and excluded if present in dbsnp129. Mutations were considered to have a damaging impact using SIFT[50] and PolyPhen-2[51] prediction algorithms, in which mutations with SIFT scores <0.05, and PolyPhen-2 scores >0.85 were considered to be significant (http://picard.sourceforge.net/).

Allele-specific copy number with ASCAT on exomes

We counted the reads with each genotype at the 1000-genome single nucleotide polymorphism (SNP) positions[52] in the normal and tumour exome sequencing data using alleleCounter (v.4.0.0). SNP positions overlapping the genomic ranges defined by {start − 100} and end {end + 100} target regions of the exome panel bed file (SeqCap EZ Exome v2, Roche, cat. no. 05860482001); SNP positions <20X depth in the normal tissue were excluded.

From the read counts at those positions we derived the

$$\mathrm{BAF} = \frac{\text{No. of reads allele } B}{\text{No. of reads allele } A + \text{No. of reads allele } B}$$

and

$$\log R = \log_2\left(\frac{\text{No. of reads in tumour}}{\text{Average depth of coverage in tumor}}\right) - \log_2\left(\frac{\text{No. of reads in normal tissue}}{\text{Average depth of coverage in normal}}\right)$$

as input to ASCAT. We ran ASCAT (v.2.5.2) on the B-allele frequency (BAF) and LogR tracks[53]. We refitted the profiles by selecting the local optima (that is, the minima in the total distance to integer DNA copy numbers) corresponding to the tumour ploidy that best matched the FACS-derived ploidy.

Estimation of whole-genome doubling timing

The timing of whole-genome duplications in relative mutational time was determined by inferring the proportion of clonal single-nucleotide variants (SNVs) present on two allelic copies p_2. Clonal SNVs were identified by running DPClust[54] on its default settings to produce clustering estimates. SNVs assigned to clusters with a cancer cell fraction of between 0.9 and 1.1 were labelled as clonal.

A mixture model on the observed alternate reads from clonal SNVs described[13] and was used to calculate the probability distribution on p_2. The mixture model was composed of two binomial distributions with frequencies $\frac{\rho}{(\rho T + 2(1-\rho))}$ and $\frac{2\rho}{(\rho T + 2(1-\rho))}$ corresponding to mutations on one and two alleles, respectively, where ρ is the purity of the tumour and T is the total copy number of the segment under consideration. A probability distribution on p_2 was calculated for SNVs in segments with allele-specific copy number 2+0/2+2 and 2+1 separately. A probability distribution on p_2 was calculated for SNVs in segments with allele-specific copy number 2+0/2+2 and 2+1 separately.

The distributions on p_2 were then used to calculate a timing distribution for the whole-genome doubling (WGD) in relative mutational time. In 2+0 and 2+2 copy number regions the whole-genome doubling timing π is given by: $\pi = \frac{2p_2}{1+p_2}$ and in 2+1 regions it is given by $\pi = \frac{3p_2}{1+p_2}$. A combined probability distribution on π was calculated from combining the estimates derived from the 2+0/2+2 and 2+1 segments.

TP53 mutation timing

The cluster profiles produced by DPClust were used in MutationTimeR[13] to estimate the probability that each SNV was clonal or subclonal and whether it occurred before the WGD.

Calculation of CNA ratios from exome data

The fraction of clonal copy number events that occurred before the WGD was calculated using the allele-specific-exome copy number. Adjacent segments with identical allele-specific copy number were first merged and segments smaller than 100 kb were filtered. Clonal copy number events were selected by filtering out segments with a total copy number different to the ancestral total copy number. Maximum parsimony was used to infer the copy number event history that led to each segment. Given that a WGD occurred in a tumour, the smallest combination of gains and losses of parental alleles that result in the final copy number state is assumed to have transpired. The proportion of copy number events occurring before and after the WGD across all segments in a tumour sample was calculated from these route histories. Confidence intervals were calculated by bootstrapping the filtered segments.

Allele-specific copy number in superclones and agreement with exome bulk

To obtain parental allele-specific copy number in the superclones we merged single-cell BAM files according to their superclones (see 'Clustering of superclones and subclones') using Sambamba (v0.7.0) and then proceed in three steps: 1. phasing of heterozygous SNPs to the major allele. First, we define heterozygous SNPs in the exome as those having at least 20 reads and a BAF between [0.2, 0.8] in the matched normal sample. We then phase the genotype with the maximum of the two read counts to the major parental allele. Second, we pool read counts per genotype across all single cancer cells at the 1000-genome SNP positions. We identify heterozygous SNPs with allele counts for genotype A and B, c_A and c_B, with $P(\mathrm{Bin}(c_A + c_B, 0.99) \leq c_A) < 0.01$ and $P(\mathrm{Bin}(c_A + c_B, 0.99) \leq c_B) < 0.01$. We then phase the genotypes with the maximum of the two read counts to the major allele. Finally, we pool the phased SNPs identified from the exome and the single cells. Although exome SNPs can in theory also be identified in the superclones, including SNPs from the matched normal exome ensures that enough SNPs are still covering regions with LOH that would be mistaken as homozygous in the single cells. 2. Maximum-likelihood estimate of the BAF of each copy number segment. For each copy number segments i, we model the read counts of the genotype phased to the major allele at each heterozygous SNP positions k_i as a Binomial: $k_i \sim \mathrm{Bin}(n_i, p_i)$, where n_i is the total read count and p_i is the BAF. We compute the likelihood across all N heterozygous SNP positions $\mathcal{L} = \prod_{i=1}^{N} \binom{n_i}{k_i} p_i^{k_i} \left(1 - p_i\right)^{n_i - k_i}$ for BAF values $p_i \in 0.5 + 0.001 \times \{0, 1, 2, \ldots,\ 500\}$ and normalize the likelihoods to get a probability distribution over the BAF values. The BAF is taken as the maximum-likelihood estimate and we also derive the [5%, 95%] confidence intervals. 3. Deriving parental-allele-specific copy number in superclones. For each copy number segment and its inferred total copy number n_t, we derive the number of copies of the major allele as $N_{\mathrm{maj}} = \mathrm{round}(\mathrm{BAF} \times n_t)$ and the number of copies of the minor allele as $N_{\mathrm{min}} = n_t - N_{\mathrm{maj}}$.

Analysis of bulk DNA-seq copy number data

Bulk DNA-seq copy number data from the expanded subclones was processed with the variable binning copy number pipeline at a genomic resolution averaging 200 kb as described in 'Inference of DNA copy number' and segmented as described in the section 'Multi-sample segmentation and integer copy number estimation'.

Analysis of bulk RNA-seq expression data

Transcript abundances for expanded clones triplicates were quantified using Salmon (v.0.14)[55] with GENCODE transcript v30[56] and options -l A -1 read1 -2 read2 -p 40–validateMappings–seqBias–gcBias. Quantified transcripts were imported into R with 'tximport' (v 1.14)[57]. Expanded clones e7, e39 and e71 had one technical replicate excluded due to poor RNA quality. Genes with a read count of <5 in 3 or more samples were excluded from the analysis. Samples were normalized for differences in sequencing depth by computing size factors and further variance stabilizing transformation with DESeq2 (v 1.26.0)[58].

Integrated analysis of DNA and RNA in subclonal regions

MDA-MB-231 DNA copy number data from single cells of the parental cell line and from bulk expanded single daughter cells were jointly segmented and co-clustered as described in 'Multi-sample segmentation and integer copy number estimation' (gamma = 20) and 'Clustering of superclones and subclones' (minPts = 14, n neighbours = 25, seed = 5, *k* superclones = 43). In brief, segment ratio copy number profiles were embedded into two dimensions using UMAP followed by construction of an SNN graph. Matching DNA–RNA pairs from the bulk expanded single daughter cells dataset were assigned identities according to their subclonal classification from the DNA co-clustering results. The group of expanded single daughter cells belonging to the same subclone were designated as expanded clusters. Variance stabilized gene counts from RNA triplicates (see 'Bulk DNA-seq and RNA-seq of MDA-MB-231' and 'Analysis of bulk RNA-seq expression data') for each expanded single daughter cell were averaged and a gene-wise *z*-score was calculated. Gene-wise *z*-scores were further averaged according to their assigned expanded clusters. Genes were organized by their corresponding genomic positions and moving windows of 100 genes were calculated for each chromosome. DNA copy number profiles from the expanded clusters are shown by taking the mode of the *ith* segment from their profiles according to the co-clustering identities.

Gene set enrichment analysis

Differential expression analysis was performed with DESeq2. Comparisons were made by contrasting each subclonal identity against all others. Fast Gene Set Enrichment Analysis was performed using R package 'fgsea' (nperm = 2000)[59] with the msigdb h.all.v6.2.symbols cancer hallmark gene sets[19]. Gene sets that were not significant (p-value < 0.05) in at least 6 subclonal identities were excluded from the analysis. Gene set pathways and expanded clusters were clustered with hierarchical clustering (Euclidean distance, ward.D linkage).

Statistical analysis

Statistical analysis and plotting were performed in the R software (v.3.6.2)[60] with 'base', 'Rstatix'[61], 'ggplot2' (v.3.2.1)[62], SciPy (v.1.4.1)[63] and pandas (v.1.01)[64].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data from this study were deposited in the NCBI Sequence Read Archive under accession number PRJNA629885.

Code availability

Code used in this study is available at https://github.com/navinlabcode/ACT_paper.

29. Baslan, T. et al. Genome-wide copy number analysis of single cells. *Nat. Protoc.* **7**, 1024–1041 (2012).
30. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
31. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
32. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
33. Hahsler, M., Piekenbrock, M. & Doran, D. Fast density-based Clustering with R. *J. Stat. Softw*. https://doi.org/10.18637/jss.v091.i01 (2019).
34. Leung, M. L. et al. Highly multiplexed targeted DNA sequencing from single nuclei. *Nat. Protoc*. **11**, 214–235 (2016).
35. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
36. Nilsen, G. et al. Copynumber: efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
37. L. McInnes, J. Healy & J. Melville. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
38. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res*. **5**, 2122 (2016).
39. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **1695** (2006).
40. McInnes, L. Healy, J. & Astels, S. hdbscan: hierarchical density based clustering. *JOSS* **2**, 205 (2017).
41. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
42. Zhang, Z., Lange, K. & Sabatti, C. Reconstructing DNA copy number by joint segmentation of multiple sequences. *BMC Bioinformatics* **13**, 205 (2012).
43. Desper, R. & Gascuel, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol*. **9**, 687–705 (2002).
44. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
45. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol*. **8**, 28–36 (2017).
46. McKenna, A. et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297–1303 (2010).
47. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol*. **31**, 213–219 (2013).
48. Knaus, B. J. & Grünwald, N. J. vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour*. **17**, 44–53 (2017).
49. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. **38**, e164 (2010).
50. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. **31**, 3812–3814 (2003).
51. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet*. **7**, 20 (2013).
52. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
53. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
54. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
55. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
56. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. **47**, D766–D773 (2019).
57. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*. **4**, 1521 (2015).
58. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**, 550 (2014).
59. Korotkevich, G., Sukhov, V. & Sergushichev, A. Fast gene set enrichment analysis. Preprint at https://doi.org/10.1101/060012 (2019).
60. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, 2013).
61. Kassambara, A. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* https://CRAN.R-project.org/package=rstatix (2020).
62. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016).
63. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
64. McKinney, W. Data structures for statistical computing in Python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 56–61 (2010).

Acknowledgements This work was supported by grants to N.E.N. from the American Cancer Society (129098-RSG-16-092-01-TBG), the National Cancer Institute (R01CA240526, R01CA236864), the Emerson Collective Cancer Research Fund (144300-80-121457-19) and the CPRIT Single Cell Genomics Center (RP180684). N.E.N. is an AAAS Wachtel Scholar and AAAS Fellow. This study was supported by core facility grants (CA016672, CA016672) from the NIH. This work was supported by the Dana-Farber Cancer Institute Physical Sciences Oncology Center (grant no. U54CA143798 to F.M.) and the DFCI Center for Cancer Evolution (to F.M.). T.B., M.T. and P.V.L. are supported by the Francis Crick Institute, Cancer Research UK (FC001202), the UK Medical Research Council (FC001202) and the Wellcome Trust (FC001202). P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of the Francis Crick Institute. T.B. is supported by a fellowship from the Boehringer Ingelheim Fonds. We thank K. Polyak, H. Tang,

L. Ramagli, E. Thompson, S. Shah, A. McPherson, N. Ramesh and A. Kalia for their assistance on this project.

Author contributions D.C.M. was involved in all aspects of the work, M.D.N., T.O.M. and F.M.-B. performed mathematical modelling. T.B., P.V.L. and M.T. performed WGD and LOH analysis. M.H. and A.D. performed data analysis. H.Y., K.W., M.R., C.P., H.D., E.S., S.B., A.S. and Y.-w.L. performed single-cell experiments. A.M. and J.M. performed cytogenetics experiments. M.H. and A.C. performed data processing. A.B., H.C., B.L., B.A. and F.M.-B. provided tissue samples, managed IRBs and contributed clinical expertise. F.M., P.V.L. and N.E.N. managed the project, analysed data and wrote the manuscript.

Competing interests F.M. is the co-founder of an oncology company. The other authors declare no competing interests.

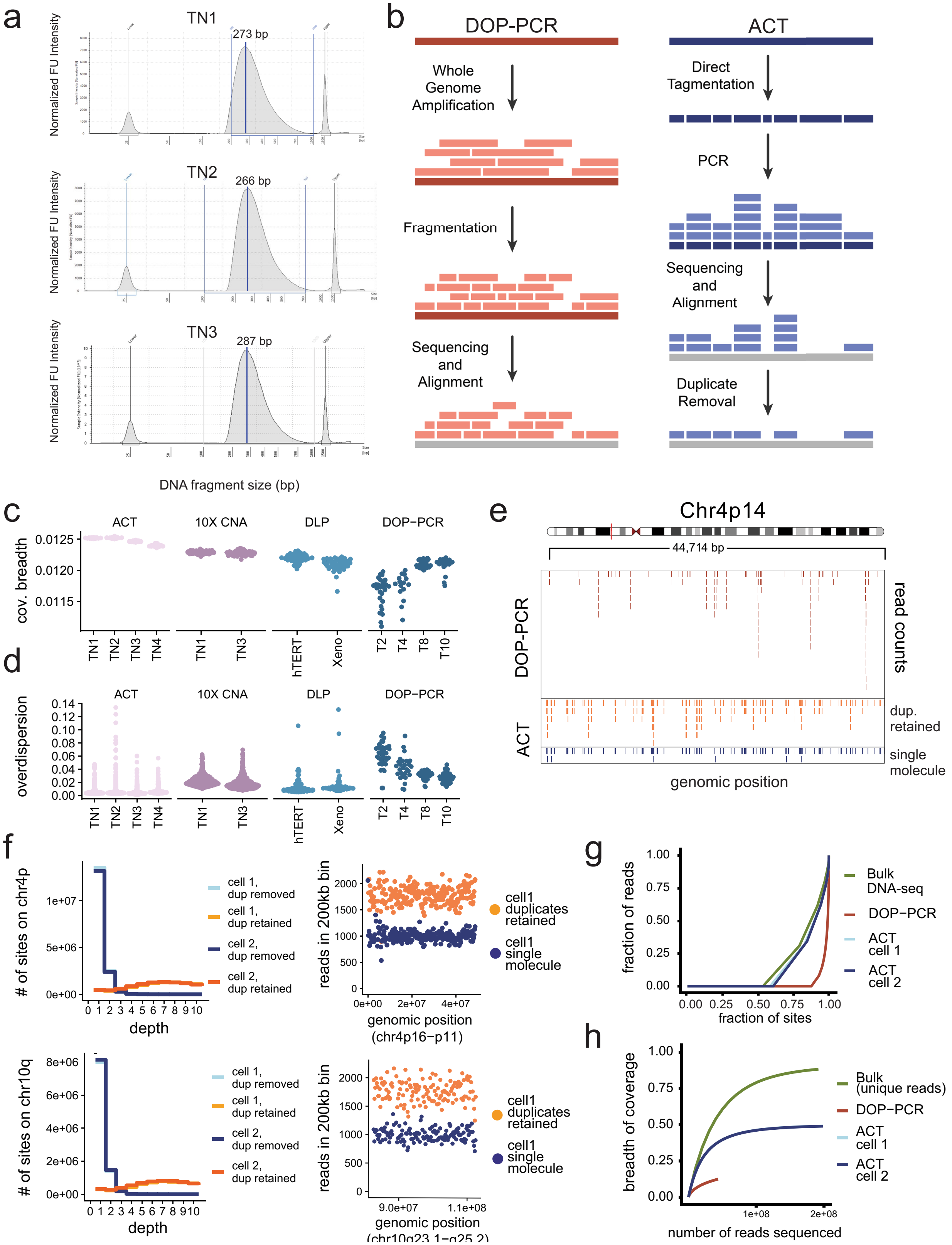
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03357-x.

Correspondence and requests for materials should be addressed to F.M. or N.E.N.

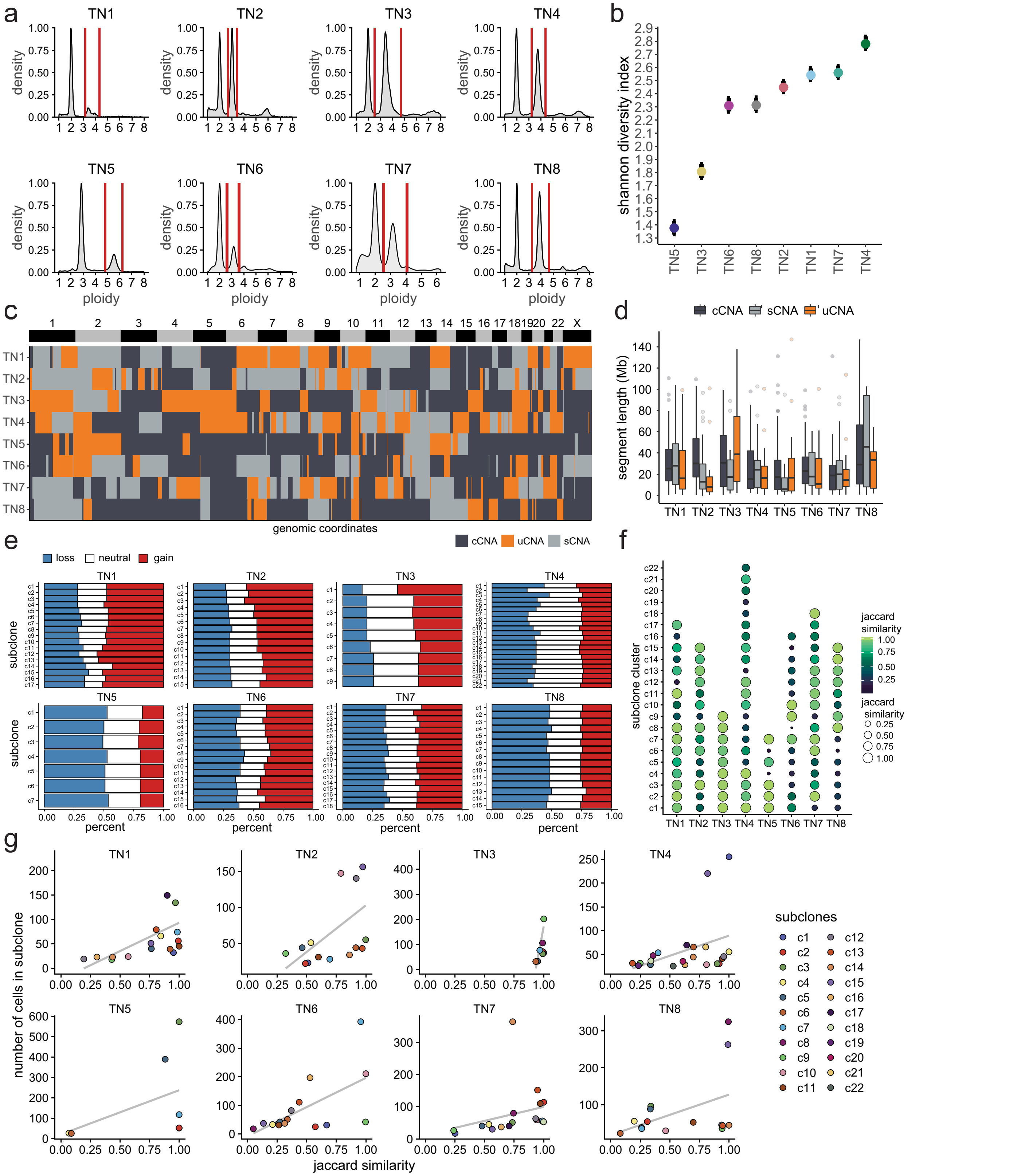
Peer review information *Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

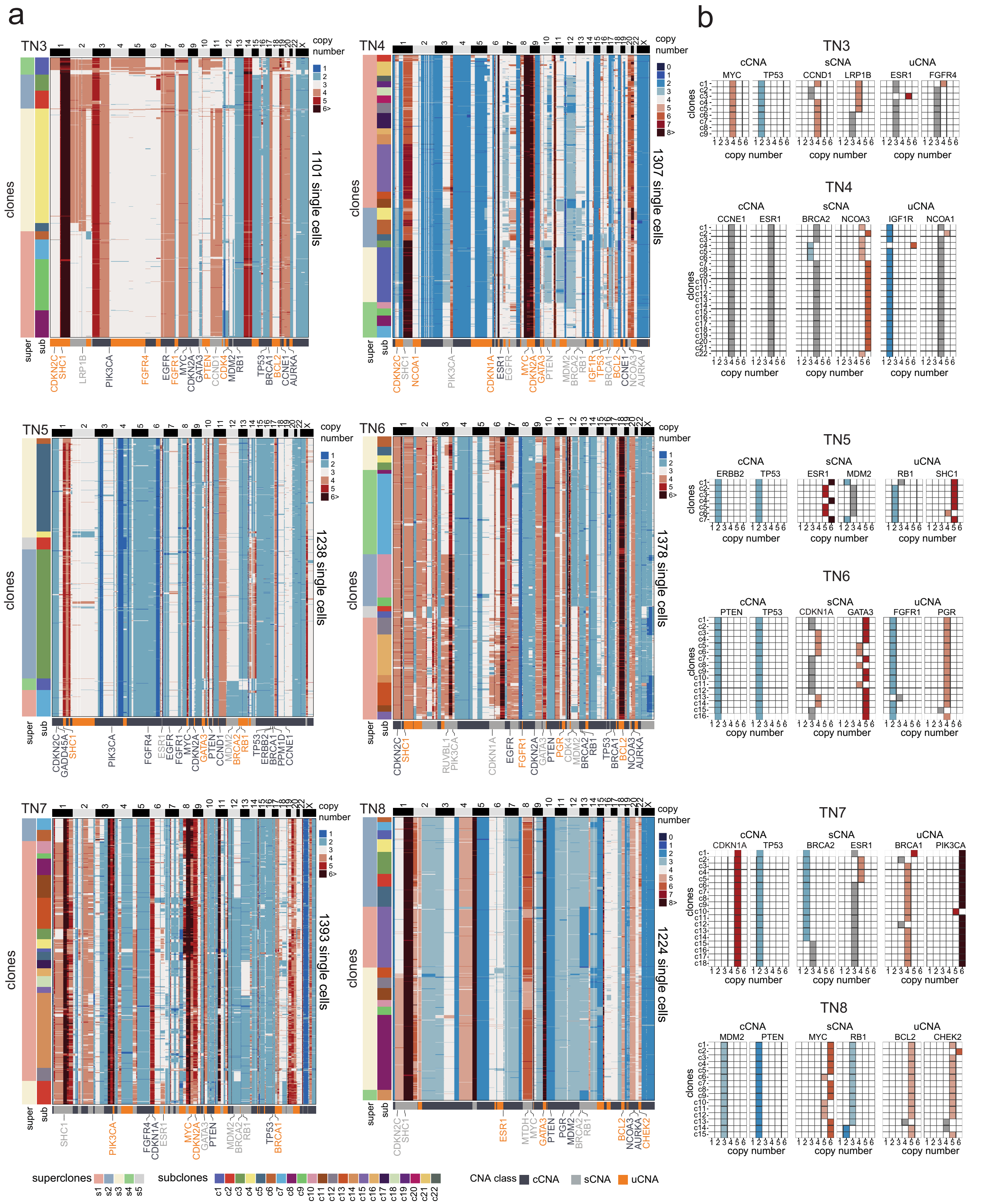


Extended Data Fig. 1 | See next page for caption.

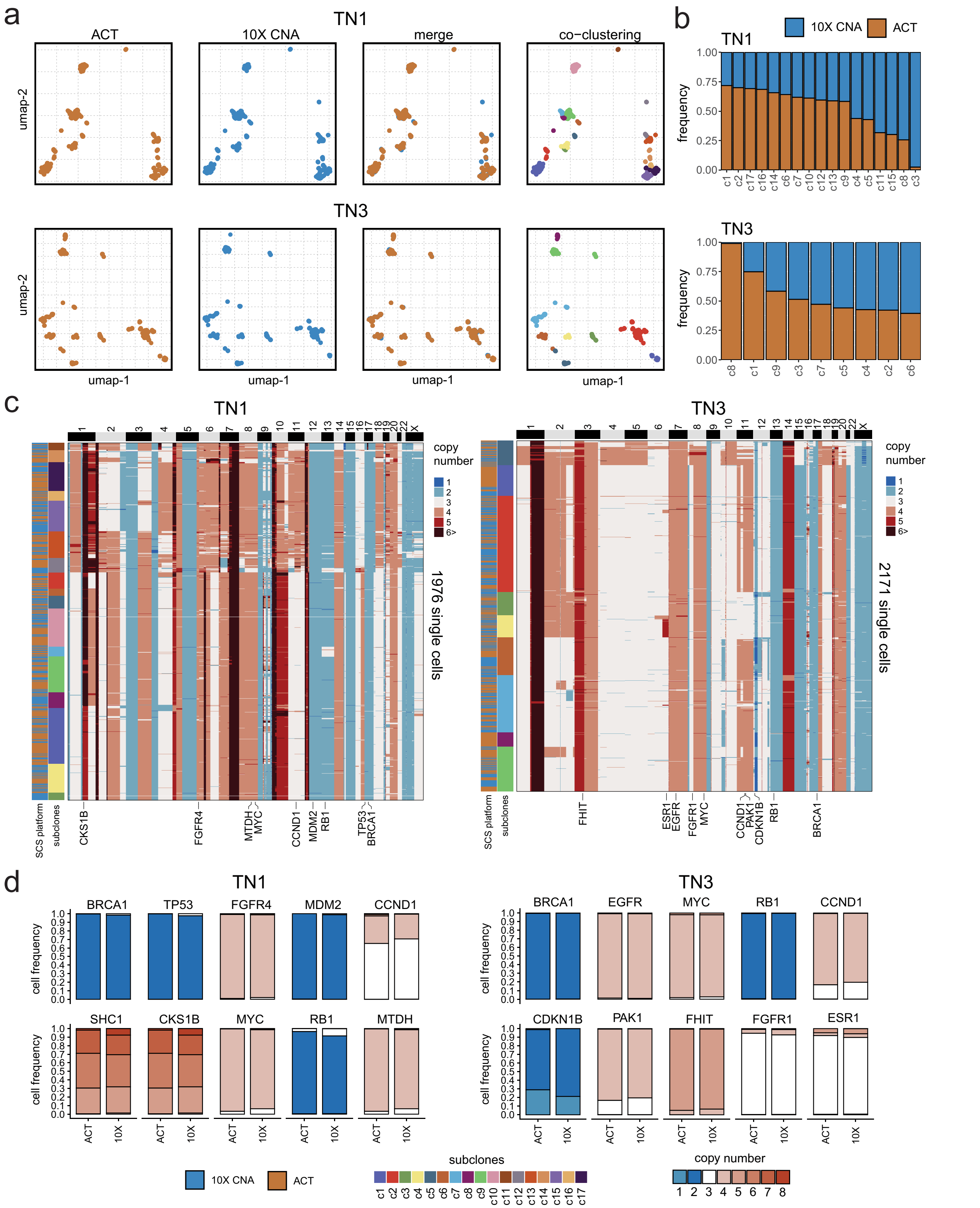
Extended Data Fig. 1 | Technical metrics and performance of ACT. a, ACT single-cell DNA library size distributions for TN1, TN2 and TN3 after pooling 384 cell libraries. **b**, Schematic showing the use of positional barcoding information to determine single-molecule information by tagmentation during ACT, compared to whole-genome amplification using DOP-PCR, where the original DNA fragmentation sites of single molecules cannot be resolved. **c**, Breadth of coverage for sparse depth data from different scDNA-seq methods plotted by individual samples, using $n = 100$ random cells per sample. **d**, Overdispersion of bin counts for sparse depth data from different scDNA-seq methods plotted by individual samples, using $N = 100$ random cells per sample. **e**, Distribution of sequencing reads across a diploid region of chromosome 4p14 for a single SK-BR-3 cell sequenced by DOP-PCR compared to ACT, in which the PCR duplicates were retained or removed to obtain single-molecule data. **f**, Distribution of sequencing reads across a diploid region of chromosome 4p (top) and 10q (bottom) for a single SK-BR-3 cell sequenced by DOP-PCR compared to ACT, with or without duplicate molecules retained. **g**, Lorenz curves of coverage uniformity for ACT, DOP-PCR and one bulk DNA-seq data from SK-BR-3 single cells, downsampled to equal coverage depth. **h**, Breadth of coverage as a function of pseudo-bulk reconstruction by combining multiple cells for ACT, DOP-PCR and bulk sequencing.



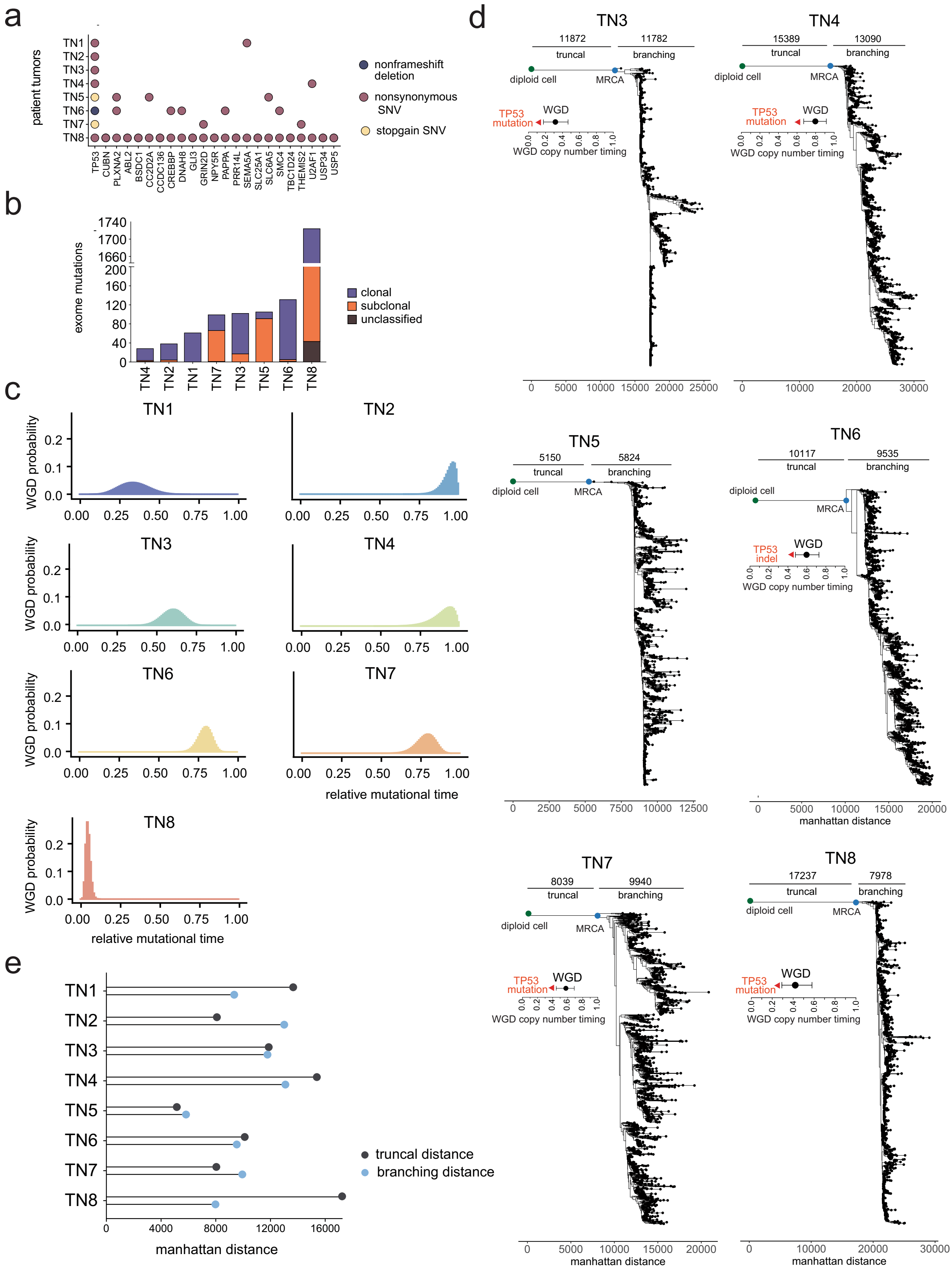
Extended Data Fig. 2 | Molecular properties of subclonal chromosome aberrations. a, FACS profiles of DAPI-stained nuclei flow-sorted for ACT from eight patients with TNBC, showing ploidy distributions, with vertical red lines showing the sorting gates. **b**, Shannon diversity indexes calculated from the single-cell copy number data from each of the eight individuals with 95% confidence intervals indicated. **c**, Heat map of the genomic regions of cCNAs, sCNAs and uCNAs across the eight tumour samples. **d**, Distributions of the genomic segment sizes of clonal, subclonal and unique CNAs across the eight tumours. **e**, Proportion of genome altered relative to the tumour ploidy classified as copy number losses in blue, neutral ground state copy number in white and gains in red. **f**, Bootstrapping of subclone clusters showing the mean Jaccard similarity for each subclone across the eight tumours. **g**, Scatter plots of number of cells in each subclone cluster by mean Jaccard similarity for each of the eight tumours.



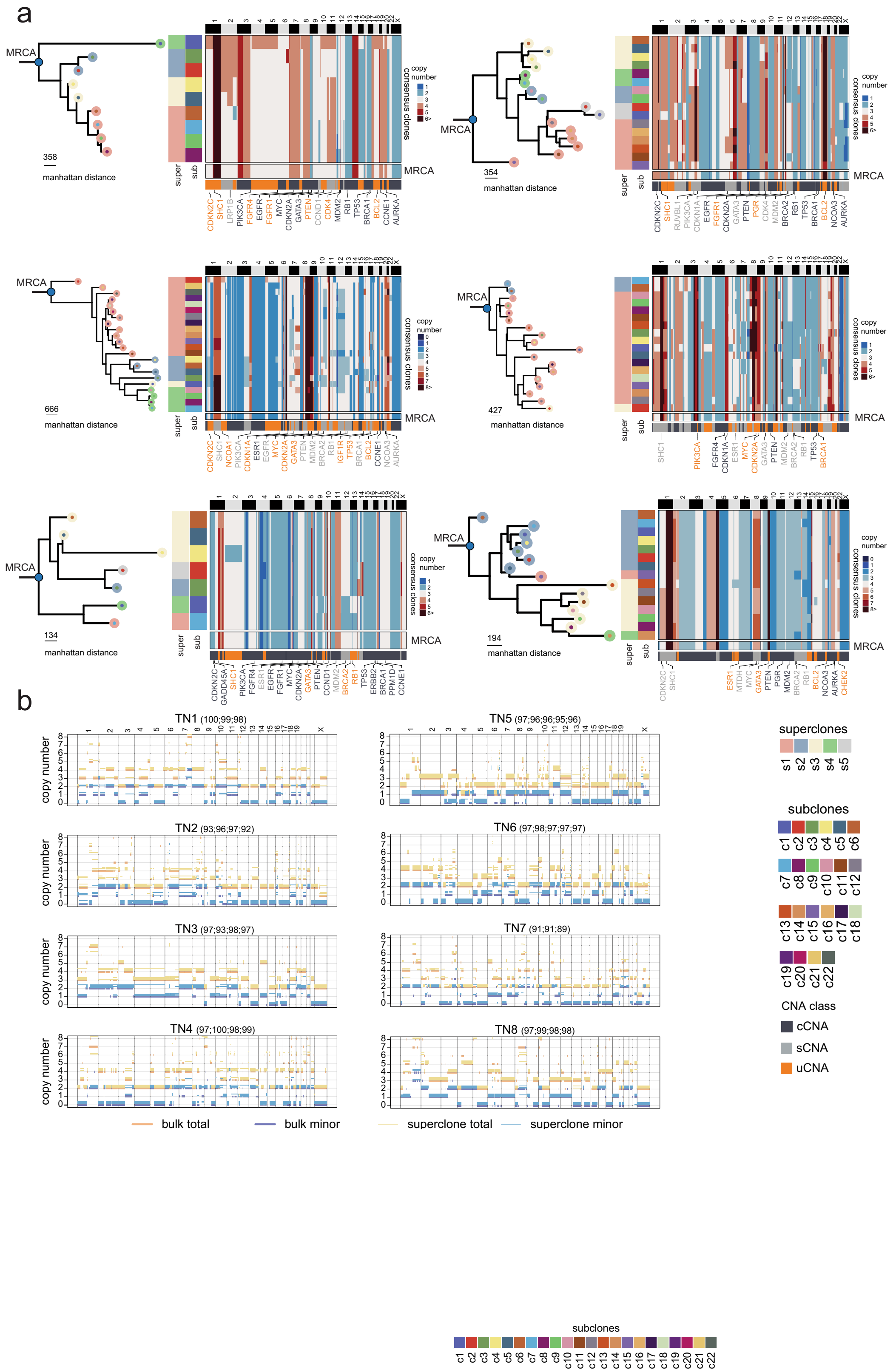
Extended Data Fig. 3 | Copy number substructure of additional patients with TNBC. a, Clustered heat maps of single-cell copy number profiles for TN3–TN8 with left annotation bars representing superclones and subclones, and bottom annotation bars representing different genomic regions of CNA classes as well as annotations for selected breast cancer genes. **b**, Matrix plots for TN3–TN8, showing integer copy number states for selected breast cancer genes in regions of cCNAs, sCNAs and uCNAs across the different subclones in each tumour.



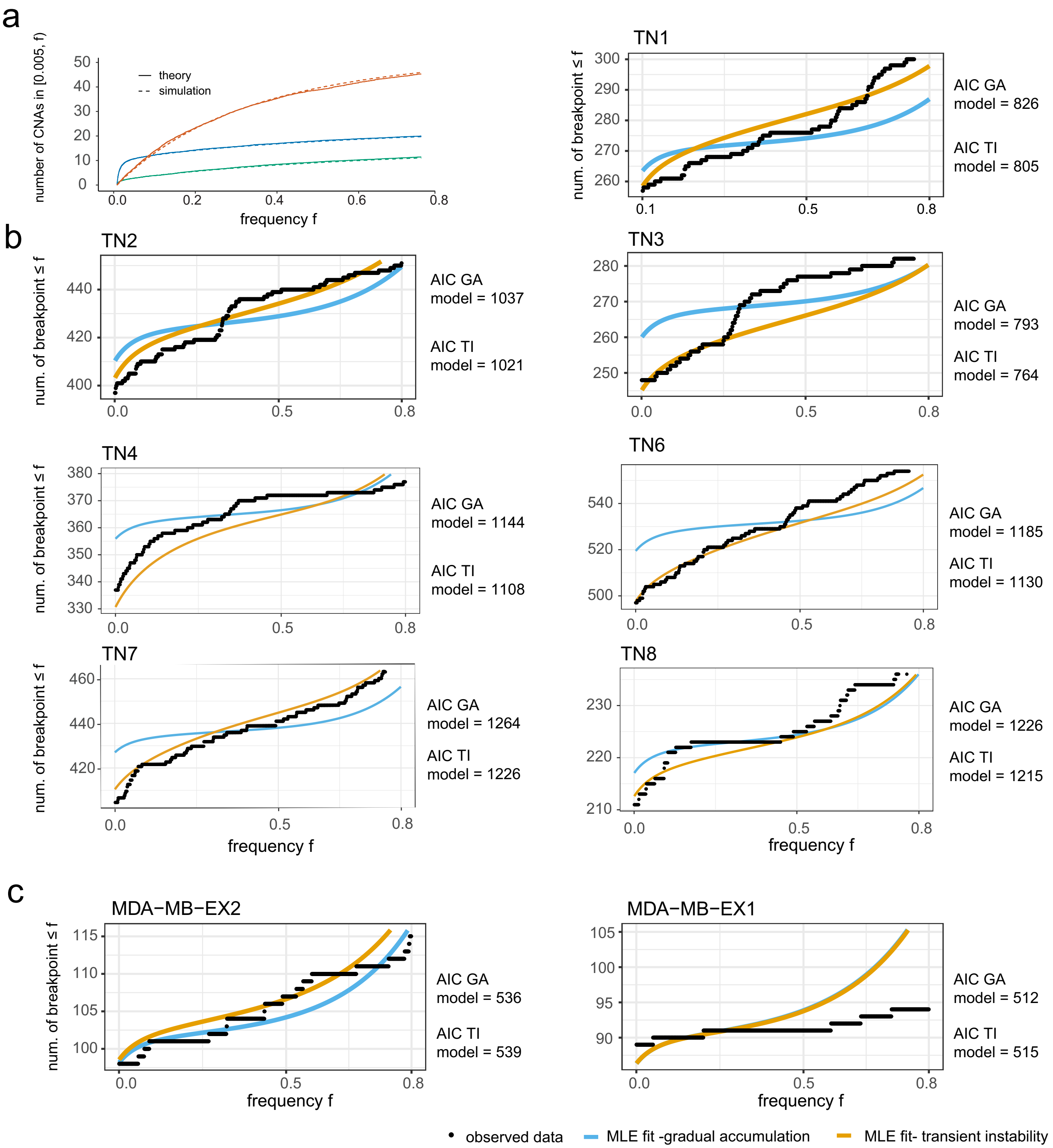
Extended Data Fig. 4 | Validation of clonal substructure using a microdroplet approach. a, Co-clustering of ACT and 10X Genomics copy number data for samples TN1 (n = 1,976 cells) and TN3 (n = 2,171 cells), showing subclones detected in the merged datasets. **b**, Frequency of subclones detected on each platform in the merged datasets from 10X and ACT. **c**, Clustered heat maps of single-cell copy number profiles for TN1 and TN3 with left annotation bars representing the scDNA-seq technology platform and the different subclones, with annotations for selected breast cancer genes indicated below. **d**, Bar plots of copy number state frequencies of selected breast cancer genes for ACT and 10X CNV showing the proportion of copy number states for all cells separated by platform.



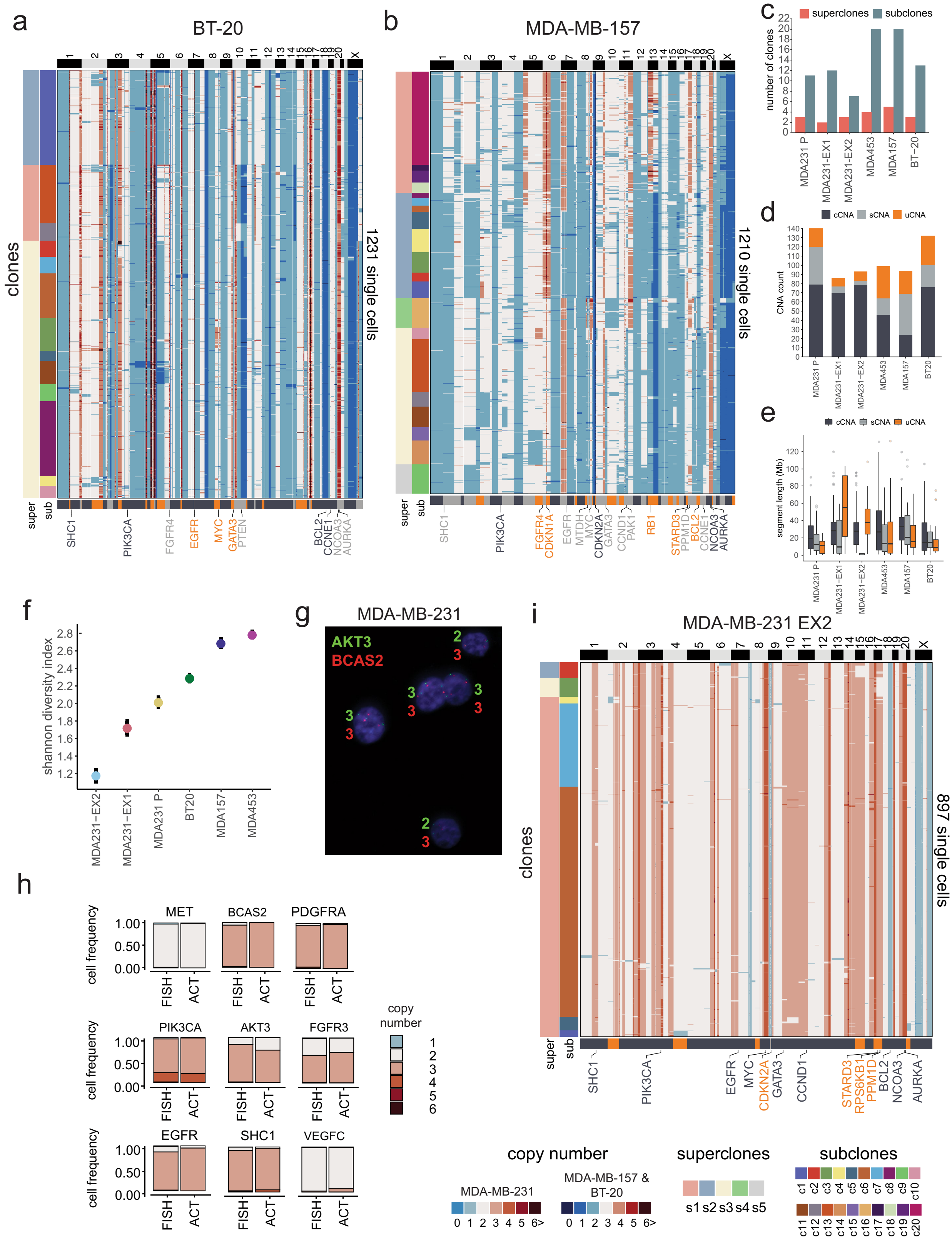
Extended Data Fig. 5 | WGD estimates and additional copy number lineages. a, Most frequent exonic mutations in genes with significant SIFT (<0.05) and PolyPhen-2 (>0.85) scores. **b**, Exome mutation counts of each tumour indicating mutations that were classified as clonal or subclonal based on allele-specific copy number frequencies. **c**, Density plots showing the probability of genome doubling as a function of relative mutational time for 7 out of the 8 patients with TNBC with sufficient number of truncal exome mutations. **d**, Minimum evolution trees of single-cell copy number profiles using Manhattan distances for TN3–TN8, indicating the distance from the diploid root node to the MRCA and the distance from the MRCA to the terminal nodes. Annotations indicate the timing of genome doubling and timing of *TP53* mutations before WGD in all of the tumours. **e**, Summary of the truncal distances from the diploid root node to the MRCA and the branching distances from the MRCA to the last terminal node.



Extended Data Fig. 6 | Evolutionary analysis of clonal lineages in additional patients with TNBC. a, Left, minimum evolution trees after the MRCA generated using the consensus CNA profiles of subclones for TN3–TN8 rooted by a neutral node to the MRCA and coloured by superclones and subclones. Right, heat maps of consensus subclones profiles, with annotations for the superclones and subclones on left annotation bars and bottom annotation bars showing different CNA classes, as well as selected breast cancer genes. The last row in the clustered heat maps shows the inferred MRCA copy number profiles. **b**, Genome-wide copy number profiles of TNBC tumours with segments of the rounded total copy number (orange) and the rounded number of copies of the minor allele (blue). Thick segments are ASCAT profiles from the exome bulk, and thinner segments are from the superclones with slight offset relative to integer values for visualization. For each superclone, parentheses show the percentage of the genomic region in which both the minor and major allele copy numbers are the same as in the exome, restricting analysis to the genomic region where the total is also the same.

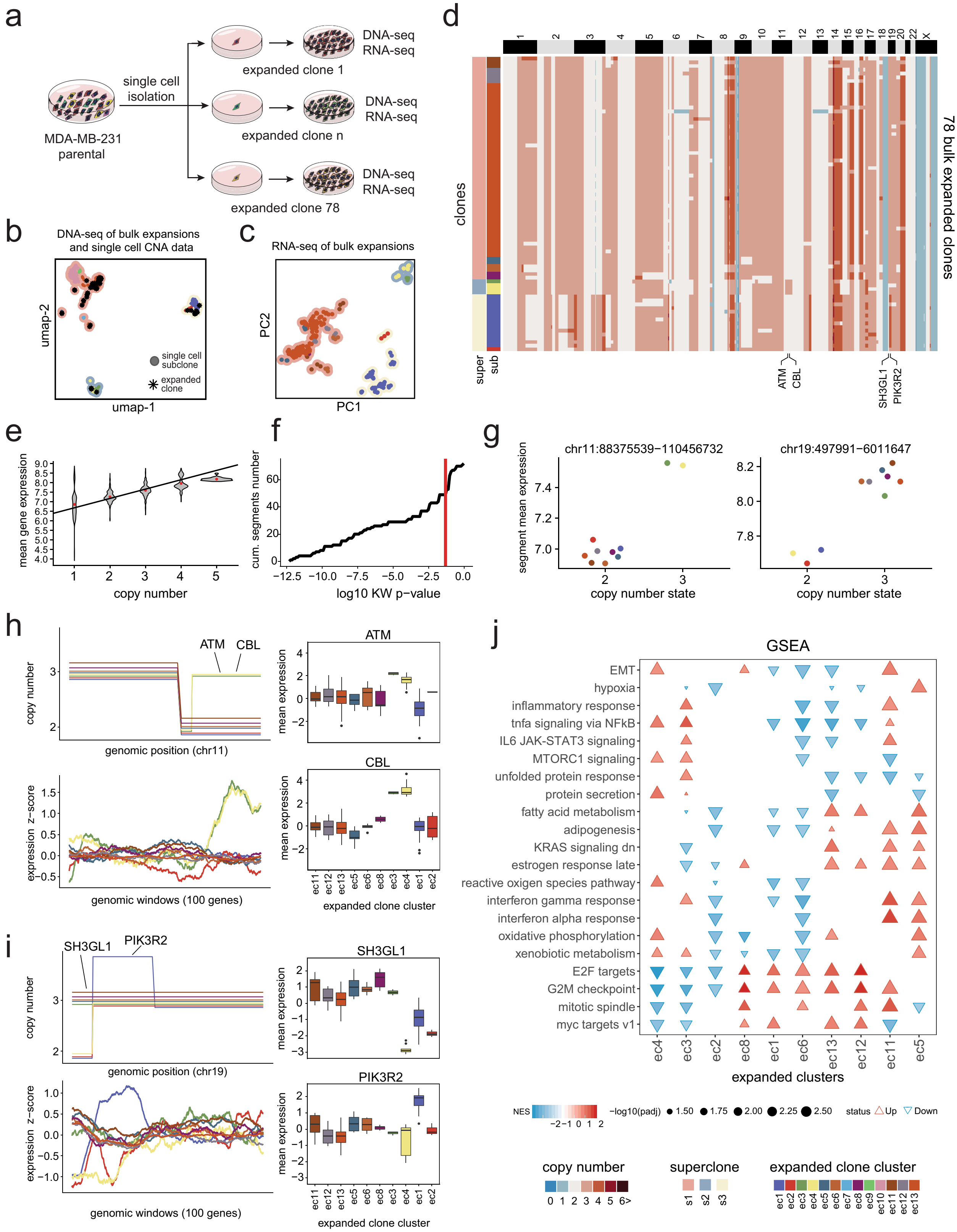


Extended Data Fig. 7 | Chromosome-breakpoint frequency spectra of additional tumours. **a**, Comparison of the expected CNA frequency spectrum obtained from theory and simulation. Simulations include a flexible fitness distribution, whereas the theoretical analysis considers neutral and lethal changes only. Different colours correspond to varying the increase in CNA rate during the transient instability phase, and the tumour size at which the instability subsides. Exact parameters are provided in the Supplementary Methods. **b**, Maximum-likelihood fits for the breakpoint frequency spectra obtained for TNBC tumours under models of gradual and transient instability after PCNE; parameter values for simulations and further details are provided in the Supplementary Methods. **c**, Maximum-likelihood fits for the breakpoint frequency spectra obtained from expanded clones of MDA-MB-231 under models of gradual and transient instability. Further details are provided in the Supplementary Methods.



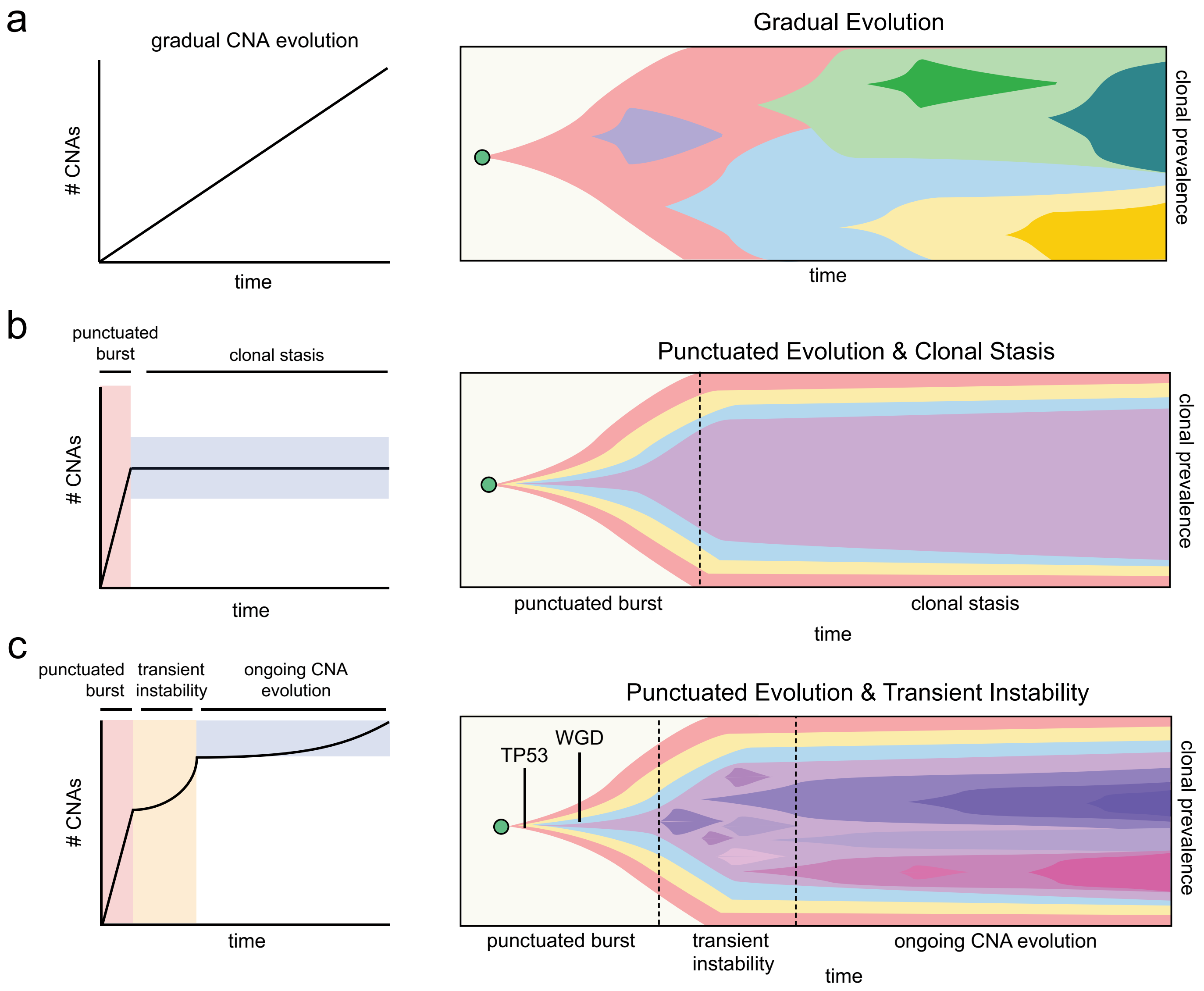
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Clonal substructure of additional TNBC cell lines and single-cell expansions. a, **b**, Clustered heat maps of single-cell copy number data from the BT-20 (*n* = 1,231 cells) and MDA-MB-157 (*n* = 1,210 cells) cell lines, in which left annotation bars represent superclones and subclones, and the bottom annotation bar represents different classes of CNA types. **c**, Number of superclones and subclones identified in the TNBC cell lines. **d**, Number of clonal, subclonal and unique CNAs detected in the four TNBC cell lines, as well as the two MDA-MB-231 expanded daughter cells. **e**, Distributions of the genomic sizes of clonal, subclonal and unique CNAs across the four TNBC cell lines and the two MDA-MB-231 expanded daughter cell lines. **f**, Shannon indexes calculated from the single-cell copy number profiles from the four TNBC cell lines and the two expanded MDA-MB-231 daughter cells with 95% confidence intervals. **g**, Microscopic field of DNA-FISH experiments of MDA-MB-231 using AKT3 and BCAS2 probes at 60× original magnification. **h**, Bar plots showing the results of DNA-FISH copy number states counted across 1,000 cells for each of the probes compared to the ACT data. **i**, Clustered heat map of single-cell copy number data for MDA-MB-231 EX2 cell line expansion (*n* = 897 cells), in which left annotation bars represent superclones and subclones, and the bottom annotation bar represents different classes of CNA types.



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

Extended Data Fig. 9 | DNA and RNA analysis of expanded clones from MDA-MB-231. a, Schematic of physical single-cell subcloning experiments of daughter cells to generate 78 expansions from the MDA-MB-231 parental cell line. **b**, Co-clustering of the single-cell copy number data from the parental MDA-MB-231 cell line (n = 820 cells) with the 78 expanded clone bulk DNA-seq copy number profiles. **c**, Principal component analysis of bulk RNA-seq profiles of the 78 expanded daughter cell lines triplicates, with contour colour representing superclones and point colour representing the subclone clusters from the genotypes of the single-cell and bulk DNA-seq co-clustering. **d**, Clustered heat map of bulk DNA copy number profiles from the 78 expanded clones, with left annotation bars representing superclones and subclones, as determined by co-clustering with the parental single-cell copy number data. **e**, Mean gene expression levels of different copy number states for 78 expansions from the MDA-MB-231 parental cell line. **f**, Cumulative number of subclonal segments as a function of Kruskal–Wallis test *P*-value, in which the red line denotes a *P*-value of 0.05. **g**, Mean gene expression as a function of copy number segments with points representing expanded clusters for two subclonal CNAs on chr11 and chr19. **h**, **i**, Consensus integer copy number profiles of the 10 expanded clone clusters on chromosome 11 (**h**) and chromosome 19 (**i**) (top) with matched RNA-seq expression (bottom) using moving windows of 100 genes. Right, selected breast cancer genes in subclonal CNA regions and their corresponding box plots of RNA expression for each expanded cluster. Box plots show the median, box edges represent the first and third quartiles, and the whiskers extend to 1.5× interquartile range. **j**, Cancer hallmark signatures with significant variability of normalized enrichment scores (NES) across the expanded clone clusters.



Extended Data Fig. 10 | Models of chromosome evolution during primary tumour expansion. a–**c**, Three models of chromosome evolution dynamics during the expansion of primary TNBC tumours, with schematic plots of chromosome accumulation over time (left) and Muller plots of clonal frequencies (right). **a**, Gradual model of copy number evolution, in which CNAs are acquired sequentially throughout tumour progression leading to the expansion of successive subclones over time. **b**, Punctuated copy number evolution model, in which an initial burst of instability generates a large number of CNAs and subclones that undergo stable expansions to form the primary tumour mass, with no (or few) new CNAs acquired after the initial burst. **c**, Model of punctuated evolution and transient instability, in which the early acquisition of *TP53* mutations and genome doubling lead to a burst of genomic instability in which a large number of CNA events are acquired and subclones are generated. These events are followed by a period of transient instability and ongoing copy number evolution during the expansion of the primary tumour mass, which leads to the generation of additional subclones and genomic diversity.

nature research

Corresponding author(s): Franziska Michor and Nicholas Navin

Last updated by author(s): Nov 24, 2020

Reporting Summary

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Data was collected by sequencing on Illumina HiSeq4000 and NextSeq500 systems
Data analysis	The data analyzed in this study was performed using custom python and R code, as well as open-source software and R packages. All of the code used to analyze the data is available upon request from the authors.

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

Data

Policy information about availability of data

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

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

The data from this study was deposited in NCBI Sequence Read Archive under accession number
PRJNA629885.

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	This is an exploratory study that is does not involved samples sizes estimated by power calculation for specific statistical tests.
Data exclusions	Some single cell data with poor sequencing metrics were excluded from the study in filtering steps, as described in the methods section
Replication	Single cells serve as replicates for detecting clonal subpopulations in each TNBC patient.
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
□	☒ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
□	☒ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
□	☒ Flow cytometry
☒	□ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	TNBC cell lines in this study were obtained from the Characterized Cell Line Core (CCLC) Facility at the University of Texas MD Anderson Cancer Center, Houston, TX.
Authentication	The cell line identities were confirmed by RFLP analysis and sparse WGS sequencing to determine copy number profiles
Mycoplasma contamination	All cell lines tested negative for mycoplasm contamination prior to running the experiments.
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Human research participants

Policy information about studies involving human research participants

Population characteristics	All 8 patients are female and range in age from 37-79 years old. Detailed clinical information on the patients is provided in supplementary table 1.
Recruitment	Samples were collected as retrospective frozen samples from our institutional tissue bank
Ethics oversight	IRB

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

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Nuclear suspensions were isolated from frozen tissue samples or cell lines using a DAPI/NST buffer prior to FACS
Instrument	FACS Melody System, Illumina HiSeq4000 system, Illumina NextSeq500 system, Echo550 Labcyte
Software	FlowJoe
Cell population abundance	NA
Gating strategy	Nuclei stained with DAPI were gated from Aneuploid distributions during FACS

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.