nature genetics

Punctuated copy number evolution and clonal stasis in triple-negative breast cancer

Ruli Gao[1], Alexander Davis[1,2], Thomas O McDonald[3,4], Emi Sei[1], Xiuqing Shi[5], Yong Wang[1], Pei-Ching Tsai[1], Anna Casasent[1,2], Jill Waters[1], Hong Zhang[6], Funda Meric-Bernstam[7], Franziska Michor[3,4] & Nicholas E Navin[1,2,8]

Aneuploidy is a hallmark of breast cancer; however, knowledge of how these complex genomic rearrangements evolve during tumorigenesis is limited. In this study, we developed a highly multiplexed single-nucleus sequencing method to investigate copy number evolution in patients with triple-negative breast cancer. We sequenced 1,000 single cells from tumors in 12 patients and identified 1–3 major clonal subpopulations in each tumor that shared a common evolutionary lineage. For each tumor, we also identified a minor subpopulation of non-clonal cells that were classified as metastable, pseudodiploid or chromazemic. Phylogenetic analysis and mathematical modeling suggest that these data are unlikely to be explained by the gradual accumulation of copy number events over time. In contrast, our data challenge the paradigm of gradual evolution, showing that the majority of copy number aberrations are acquired at the earliest stages of tumor evolution, in short punctuated bursts, followed by stable clonal expansions that form the tumor mass.

Aneuploidy is pervasive in human cancers[1] and is frequently (>90%) detected in patients with breast cancer[2,3]. DNA copy number aberrations (CNAs) often lead to gene dosage effects that promote tumor growth through the overexpression of oncogenes or downregulation of tumor-suppressor genes. However, most genomic studies have analyzed tumor samples from a single time point (biopsy or surgery), making it difficult to study the natural progression of chromosome evolution during tumorigenesis. Currently, the prevailing model for copy number evolution posits that CNAs are acquired gradually and sequentially over extended periods of time, leading to successively more malignant stages of cancer[4,5]. An alternative model is punctuated copy number evolution (PCNE), in which CNAs are acquired in short bursts of crisis, followed by stable clonal expansions that form the tumor mass (**Supplementary Fig. 1**). Previous work has implicated a punctuated model to explain localized chromosome rearrangements, including chromothripsis[6], chromoplexy[7] and firestorms[2]. However, there has been limited data showing that genome-wide aneuploidy arises in a short punctuated burst at the earliest stages of tumor evolution.

Intratumoral heterogeneity provides a window into time by representing a permanent record of the mutations that occurred during tumor progression. By assuming that mutational complexity increases over time, it is possible to reconstruct the evolutionary history of a tumor[8,9] and investigate PCNE. However, most tumors consist of complex mixtures of single cells with different genotypes, complicating such studies. To address this problem, we previously developed a single-cell DNA sequencing method called single-nucleus sequencing (SNS)[10,11]. We applied this method to sequence single tumor cells from two patients with breast cancer, thus providing initial evidence for PCNE[11]. However, these data were limited to two patients, mainly because of the high costs and low throughput associated with SNS. To address these challenges, we developed a highly multiplexed single-nucleus sequencing (HM-SNS) method that can profile 48–96 single cells in parallel.

In this study, we applied HM-SNS to investigate the clonal substructure and evolution of CNAs in patients with triple-negative breast cancer (TNBC). TNBC is a subtype of breast cancer that is characterized by a lack of estrogen receptor (ER), progesterone receptor (PR) and *HER2* (*ERBB2*) amplification[12]. Patients with TNBC show poor survival and frequently develop resistance to chemotherapy[13]. The majority of patients with TNBC harbor *TP53* mutations[3] and show complex aneuploid rearrangements[2,14]. Genomic studies have shown that patients with TNBC display a large amount of between-patient heterogeneity in somatic mutations[3], in addition to extensive intratumoral heterogeneity within each tumor mass[15–18]. However, most studies of patients with TNBC have been limited to bulk tumor analysis, and thus we investigated the clonal substructure of 12 patients with treatment-naive TNBC at single-cell genomic resolution (**Supplementary Table 1**).

[1]Department of Genetics, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [2]Graduate School of Biomedical Sciences, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [5]Peking Union Medical College, Department of Medical Oncology, Cancer Hospital and Institute, Chinese Academy of Medical Sciences, Beijing, China. [6]Department of Pathology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [7]Department of Surgical Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [8]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. Correspondence should be addressed to N.E.N. (nnavin@mdanderson.org).

Received 17 November 2015; accepted 13 July 2016; published online 15 August 2016; doi:10.1038/ng.3641

RESULTS

Highly multiplexed single-cell copy number profiling

To profile genome-wide copy number in single cells, we developed HM-SNS and applied it to sequence 1,000 single cells from tumors in 12 patients with TNBC (**Fig. 1a**). Nuclear suspensions were prepared from large (tumor volume 0.6–1.0 cm^3) frozen tumor specimens, and DNA was stained with DAPI for flow sorting. Single nuclei were gated by ploidy and deposited into individual wells of a 96-well plate for whole-genome amplification using degenerative-oligonucleotide PCR (DOP-PCR)[10,11]. After amplification, barcoded libraries were

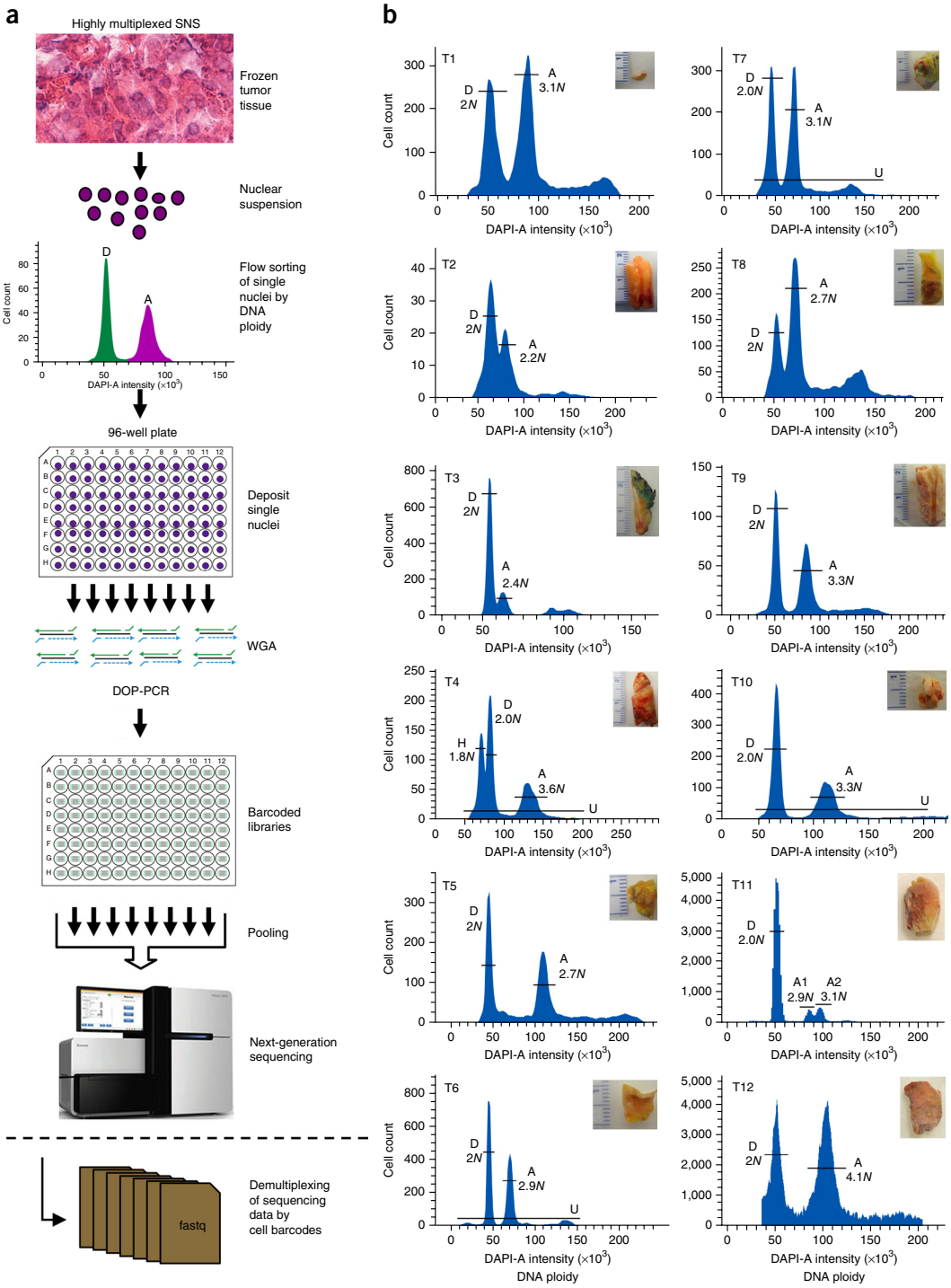

Figure 1 Highly multiplexed single-nucleus sequencing of patients with TNBC. (**a**) The highly multiplexed single-nucleus sequencing (HM-SNS) method. Tumor tissues are dissociated into nuclear suspensions and stained with DAPI for flow sorting by DNA ploidy. Single nuclei are deposited into 96-well plates and subjected to whole-genome amplification (WGA) by DOP-PCR. Single-cell libraries are barcoded with unique 8-bp identifiers, and 48–96 libraries are pooled for sparse next-generation sequencing. Sequence reads are demultiplexed using cell barcodes after sequencing is completed for copy number profile calculations. (**b**) FACS plots of DAPI intensity showing the ploidy distribution for each patient with TNBC. Single cells were isolated from different ploidy groups that were gated as diploid (D), hypodiploid (H), aneuploid (A) or universal (U).

prepared for each single cell, and 48–96 libraries were pooled (Online Methods). The pooled libraries were sequenced on an Illumina platform with 76 single-end cycles. Single nuclei were sequenced with sparse coverage, and copy number profiles were calculated from sequence read depth at 220-kb resolution (Online Methods). On average, 83 single cells (range of 48–120) were sequenced from each patient with TNBC (**Supplementary Table 2**). In each patient, we observed a diploid (2*N*) peak and one or more aneuploid peaks that ranged from 1.8–4.1*N* in the flow sorting histograms (**Fig. 1b**). Single nuclei were isolated from the aneuploid and diploid

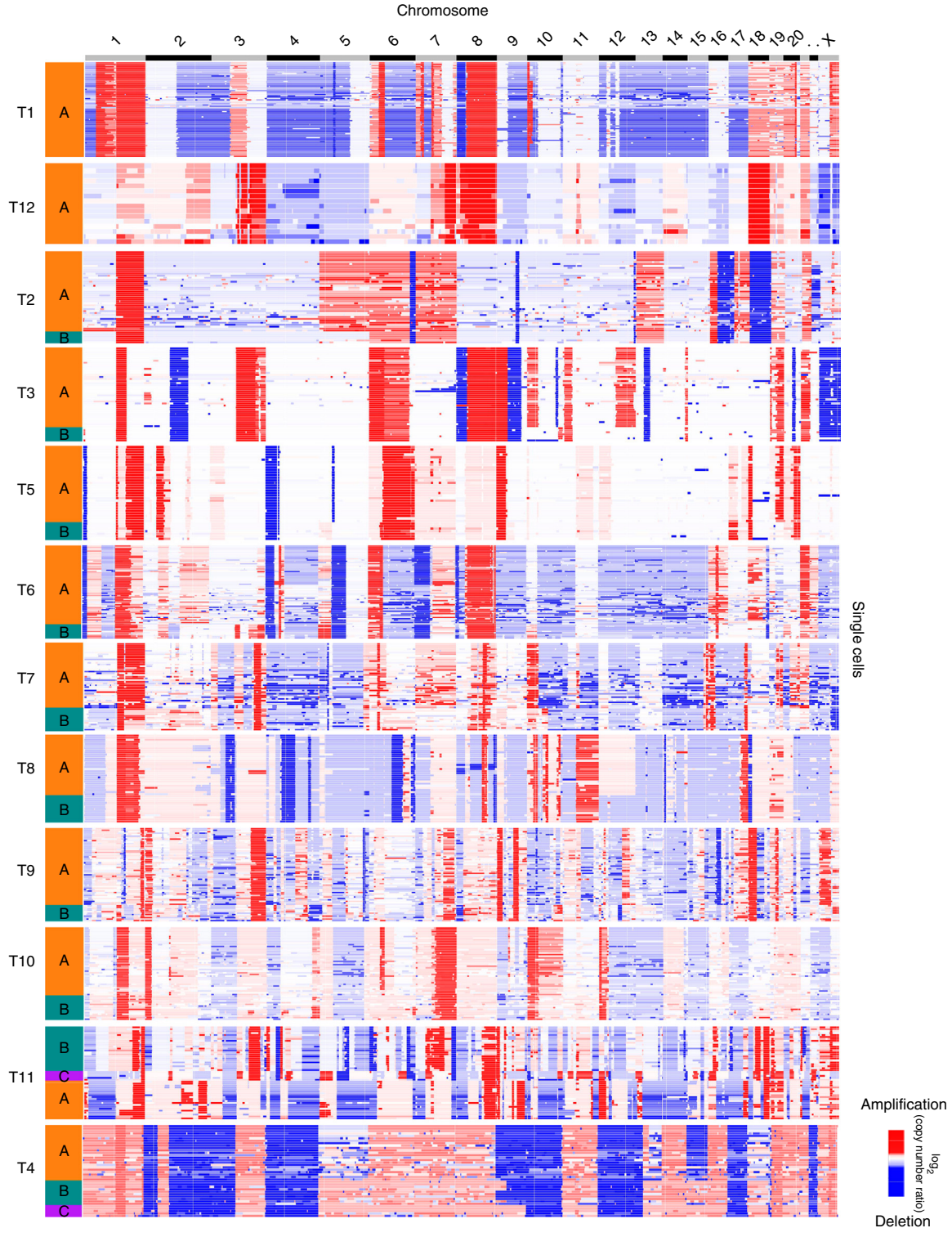

Figure 2 Clonal subpopulations identified by clustering aneuploid cells. Hierarchical 1D clustering is shown for single-cell aneuploidy copy number profiles from each patient with TNBC. Clonal subpopulation identity (clone A, B or C) is indicated on the left. Single cells are plotted along the *y* axis, and CNAs are plotted in genomic order along the *x* axis.

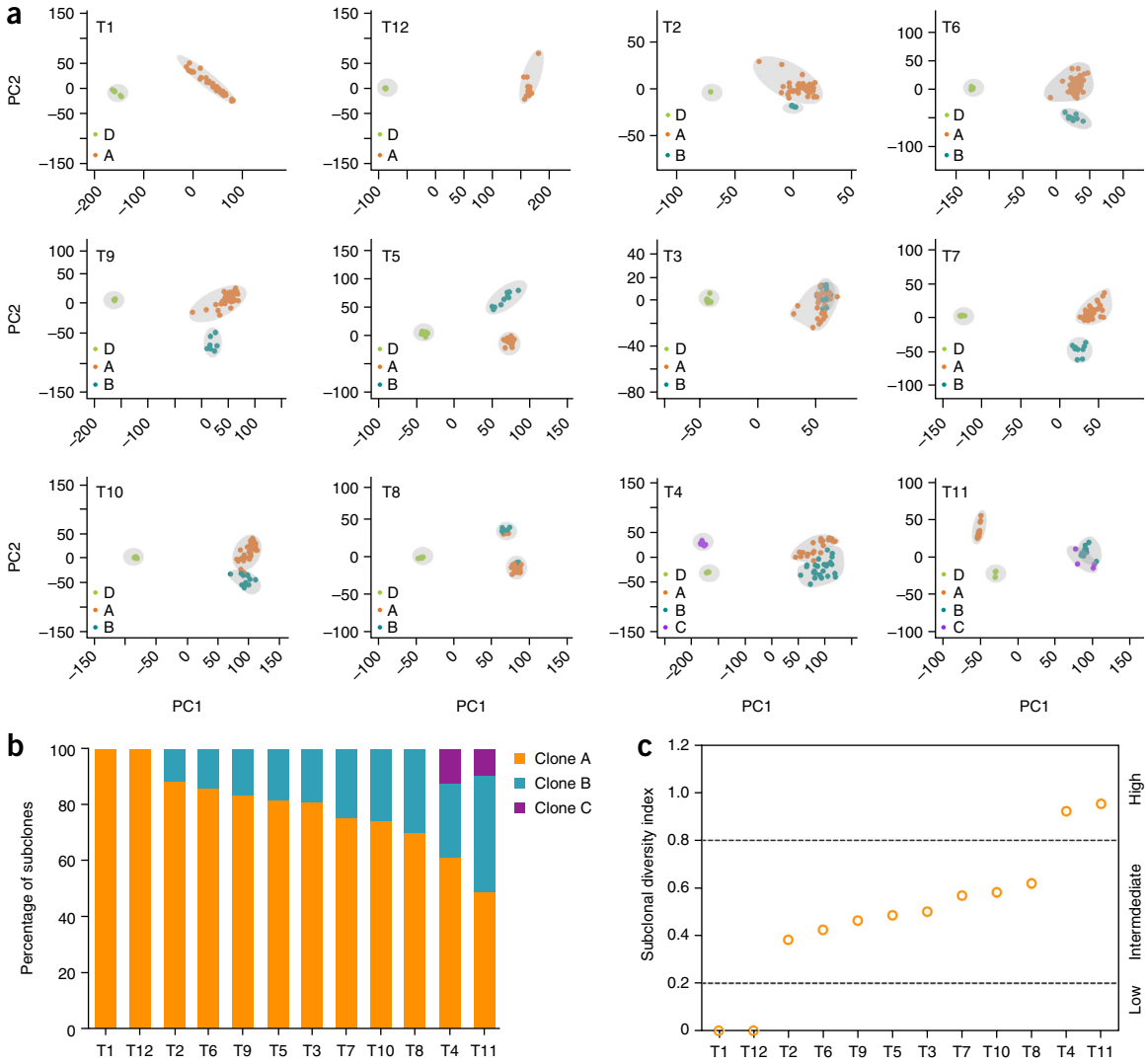

Figure 3 Clonal composition and diversity of TNBC tumors. (**a**) PCA of the single-cell copy number profiles from each TNBC tumor. Single-cell copy number profiles are colored according to their classification in the hierarchical clustering analysis and are labeled as diploid (D) or aneuploid (clone A, B or C). (**b**) Percentage of cells with each subclone genotype in the tumors. (**c**) Shannon diversity index of the copy number profiles from each tumor; dashed lines separate low-, intermediate- and high-diversity groups.

peaks, in addition to broadly gating nuclei across all ploidy distributions using universal gates for a subset of tumors.

Clonal substructure and diversity during tumor growth

To delineate the clonal substructure of each tumor, we performed 1D hierarchical clustering of the aneuploid single-cell copy number profiles. Clustered heat maps identified 1–3 major subpopulations of clones (clones A, B and C) in each tumor (**Fig. 2**). Within each subpopulation, the single cells had highly similar copy number profiles (mean pairwise $r = 0.87$), representing stable clonal expansions that occurred during tumor growth. A similar population substructure was also observed in clustering of all of the aneuploid and diploid cells from each patient with TNBC, where the diploid cells formed another independent cluster (**Supplementary Fig. 2**). To quantitatively determine the optimal number of clusters in each tumor, we applied PAM clustering[19] (**Supplementary Fig. 3**). The PAM results were consistent with the hierarchical clustering results for most patients with TNBC. Principal-component analysis (PCA) was also consistent with the clustering results, showing that 1–3 major clusters were present in each tumor (**Fig. 3a**). We quantified the genotype frequencies of the subpopulations, finding that some clones achieved higher frequencies in the tumor mass (**Fig. 3b**). To calculate a global metric of clonal diversity, we computed Shannon diversity indices for each patient with TNBC (Online Methods). The diversity indices showed a broad range across the cohort and corresponded to the number of clonal subpopulations that were present in each tumor (**Fig. 3c**). These data suggest that most TNBC tumors consist of 1–3 major clonal subpopulations and that complex aneuploid tumor profiles are highly stable (clonal stasis) during tumor growth.

Divergent subpopulations in polyclonal tumors

Polyclonal tumors shared most CNAs across subpopulations, but clones also differed by a few discrete subclonal events that emerged in the later stages of tumor evolution. Subclonal CNAs distinguished clones and often resulted in amplification of oncogenes and deletion of tumor-suppressor genes. In several cases, subclonal CNAs were associated with increased genotype frequencies of the corresponding clones in the tumor mass, suggesting that they may have provided a

fitness advantage. To further investigate this possibility, we calculated clonal frequencies (c_f) in the polyclonal tumors (Online Methods and **Supplementary Table 3**). For instance, in tumor T3, two major clonal subpopulations (clones A and B) were identified, of which clone A acquired additional amplifications of chromosomes 10p and 12q (**Fig. 4a**). The 10p amplification increased the copy number of *GATA3*, while the 12q amplification increased the copy number of *MDM2* as well as several other genes. These amplifications were associated with increased frequency of clone A (c_f = 0.85) in comparison to clone B (c_f = 0.15). In another polygenomic tumor (T2), we identified

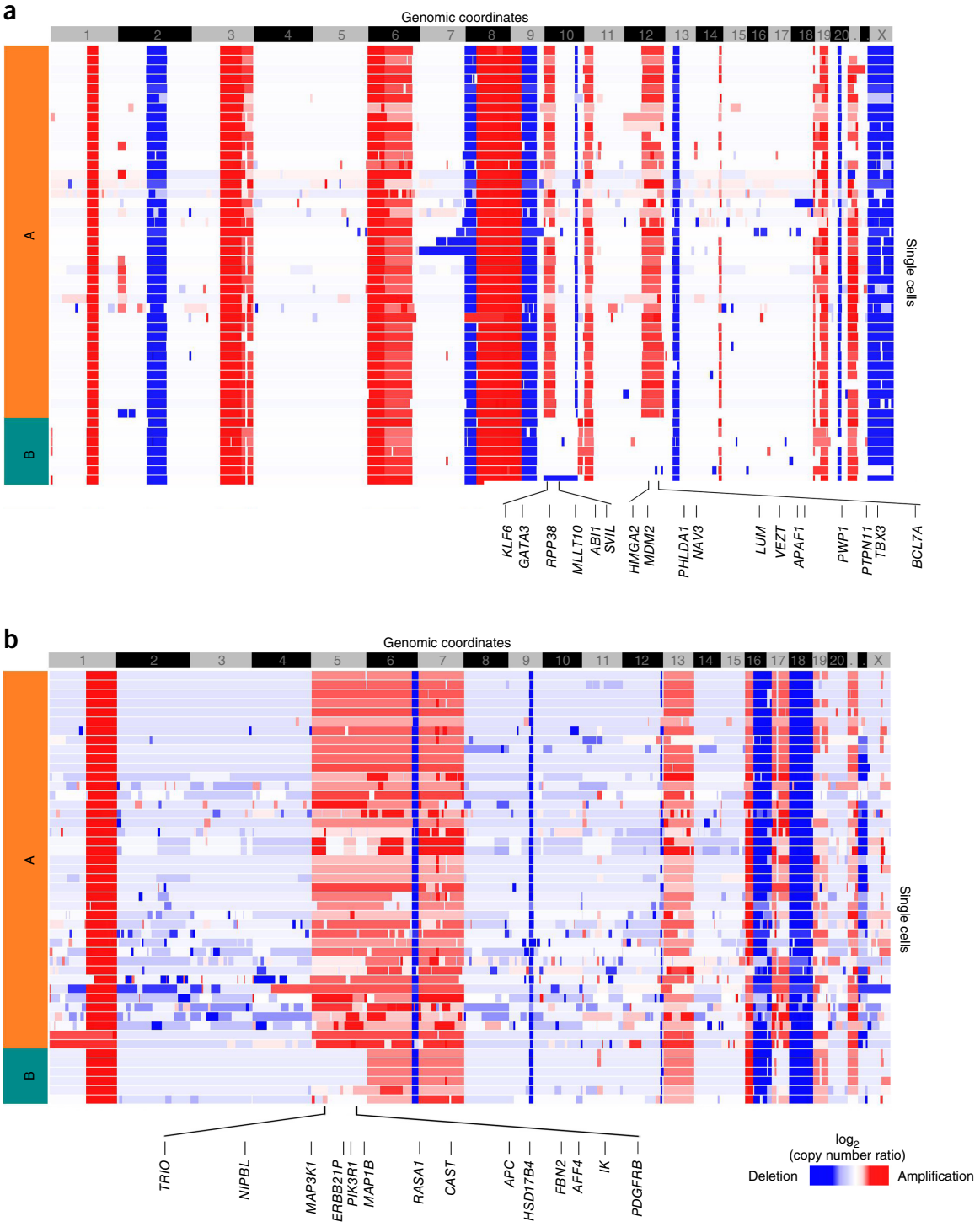

Figure 4 Divergent subpopulations in polyclonal tumors. Clustered heat maps are shown for single-cell aneuploid copy number profiles in polyclonal tumors. (**a**) Tumor T3 heat map with two subpopulations (clones A and B) identified. Subpopulation A diverged from subpopulation B by acquiring additional amplifications on chromosomes 10p and 12q, resulting in the amplification of *GATA3* and *MDM2* in addition to many other genes. (**b**) Tumor T2 heat map with two subpopulations (clones A and B) identified. Subpopulation A diverged from subpopulation B by amplification of chromosome 5, containing many cancer-related genes, including *MAP3K1*, *ERBB2IP* and *PIK3R1*.

two major clonal subpopulations (clones A and B) that differed by a broad amplification on chromosome 5 that encompassed 14 cancer-related genes, including *MAP3K1*, *ERBB2IP* and *PIK3R1* (**Fig. 4b**). This amplification was associated with increased frequency of clone A (c_f = 0.87) relative to clone B (c_f = 0.13). Similar subclonal CNAs were found in tumors from other patients with TNBC (T5 and T8) and were often associated with increased genotype frequencies for the clones that harbored them (**Supplementary Fig. 4**). These data show that, in addition to undergoing stable clonal expansions, tumors in patients with TNBC can continue to acquire single CNAs in the later stages of tumor progression and that these events are associated with the increased prevalence of new subpopulations.

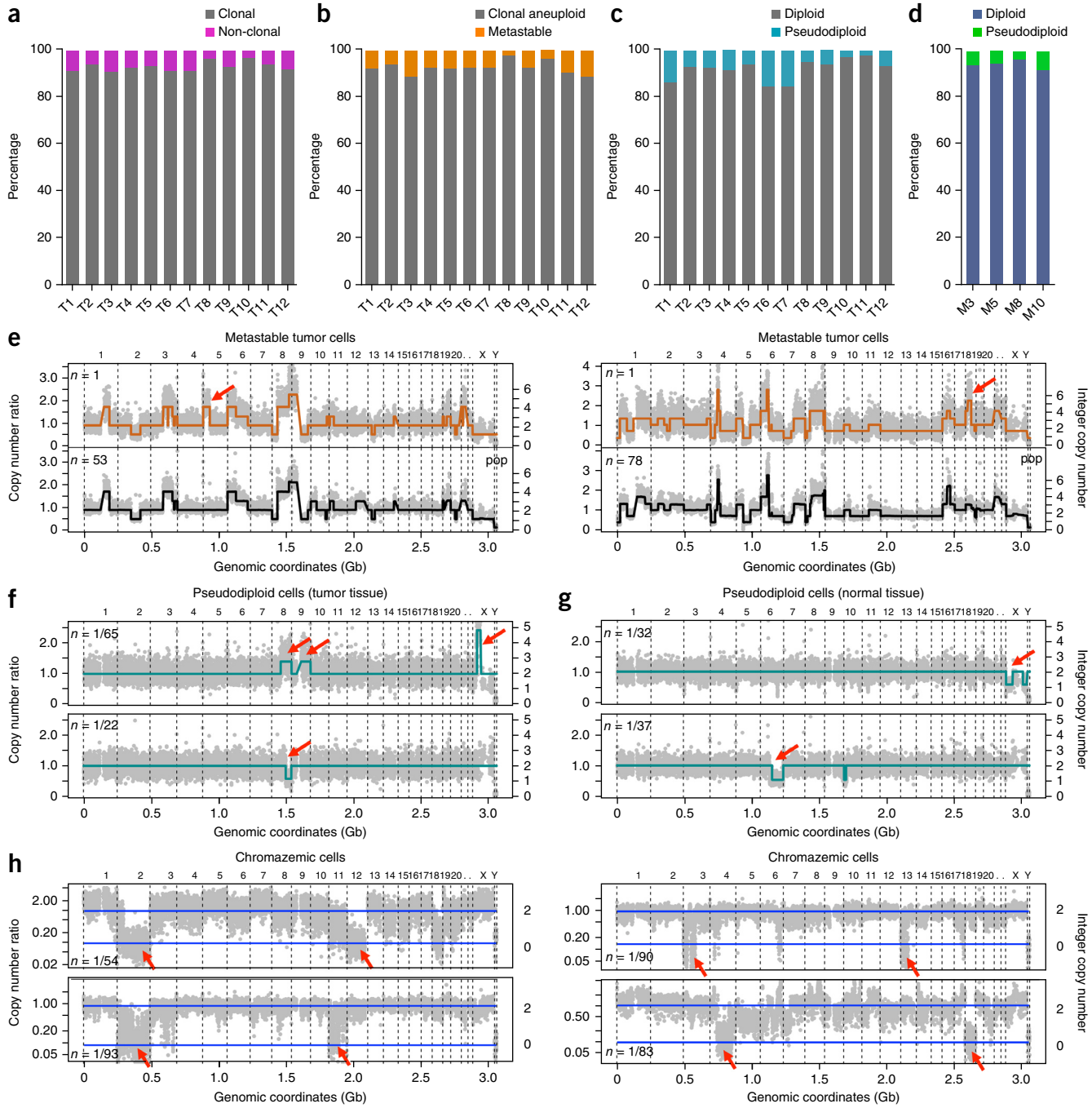

Figure 5 Non-clonal copy number profiles in tumors and normal breast tissues. (**a**) Percentage of non-clonal cells in each tumor. (**b**) Percentage of non-clonal metastable aneuploid cells in the aneuploid fraction of each tumor. (**c**) Percentage of non-clonal pseudodiploid cells in the diploid fraction of each tumor. (**d**) Percentage of pseudodiploid cells in matched normal breast tissues from four patients with TNBC (T3, T5, T8, T10). (**e**) Copy number profiles of two example metastable aneuploid cells (top) in comparison to those of cells from the major aneuploid subpopulation (bottom). (**f**) Copy number profile of an example pseudodiploid cell isolated from the diploid fraction of a tumor. (**g**) Copy number profile of an example pseudodiploid cell isolated from matched normal breast tissue. (**h**) Copy number profiles of four example chromazemic cells with large homozygous deletions of whole chromosomes or chromosome arms. Red arrows indicate non-clonal CNAs. Profiles labeled as "pop" show the consensus copy number profiles of the population of cells. Horizontal lines correspond to diploid copy number (2) and homozygous deletions (0).

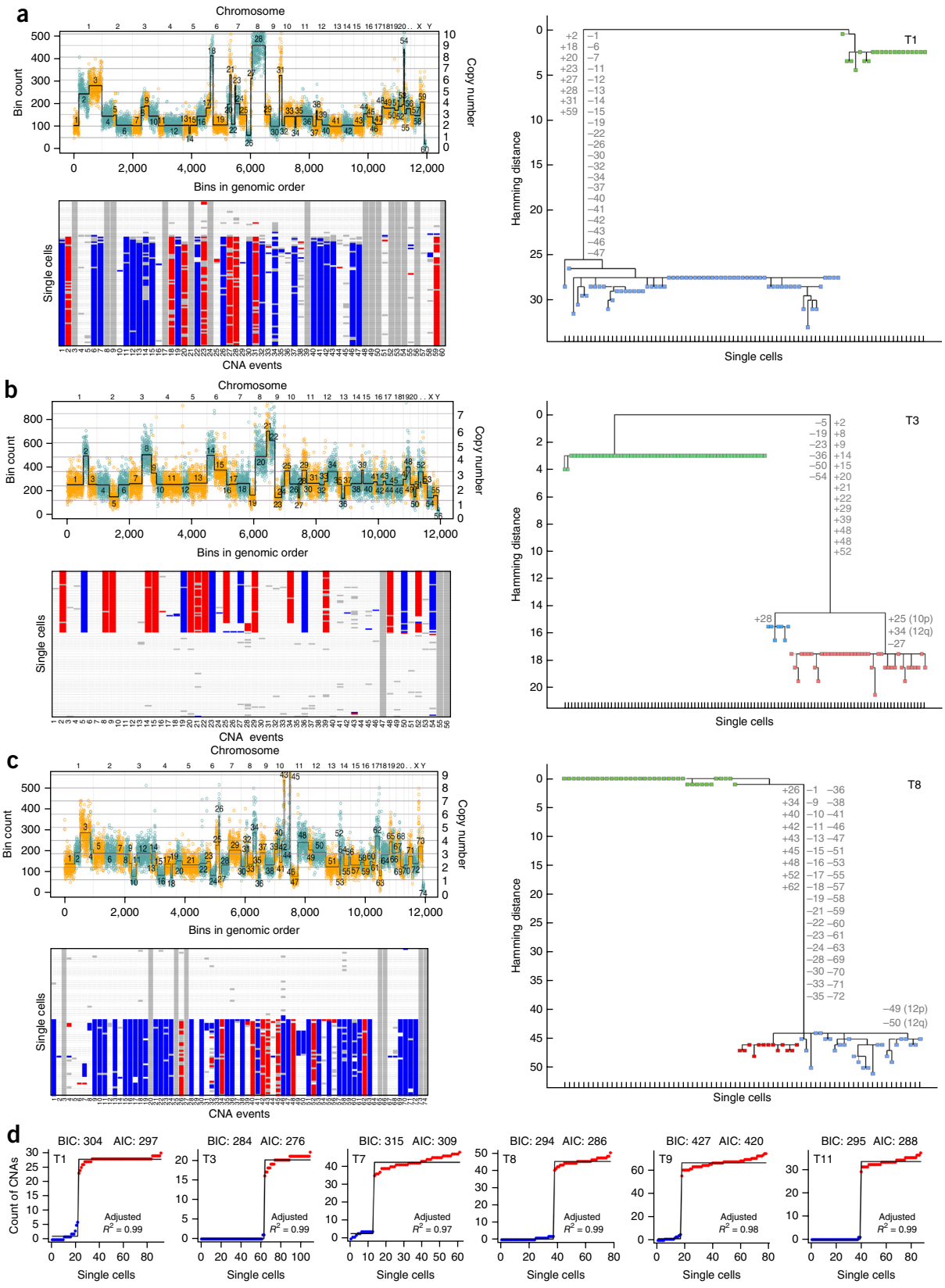

Figure 6 Punctuated copy number evolution and phylogenetic trees. (**a**–**c**) Multiple-cell segmentation (top left), trinary event matrices (white, 0; red, 1; blue, −1) (bottom left) and maximum-parsimony trees (right) for tumors from three patients with TNBC: T1 (**a**), T3 (**b**) and T8 (**c**). Maximum-parsimony trees are rooted by diploid cells; non-clonal profiles were excluded from the analysis. Copy number events with non-integer values were filtered from all cells before tree construction and are shown in gray in the trinary event matrices. (**d**) Step fitting of sorted single-cell CNA count data from six patients with TNBC. Adjusted R^2, BIC and AIC metrics are displayed for each fit. Blue, diploid normal cells; red, aneuploid tumor cells.

Non-clonal copy number profiles in tumors

Whereas most cancer cells clustered into 1–3 major clonal subpopulations, we also identified a minor fraction (<10%) of non-clonal single-cell copy number profiles in each tumor. On average, the non-clonal copy-number profiles occurred at a frequency of 7.4 ± 0.8% (s.e.m.) in the aneuploid fractions, 7.9 ± 1.4% in the diploid fractions and 5.9 ± 1.0% in the adjacent normal tissue cells (**Fig. 5a–d** and **Supplementary Table 4**). On the basis of the patterns of the CNA profiles, we identified three major classes of non-clonal cells: (i) metastable tumor cells, (ii) pseudodiploid cells and (iii) chromazemic cells (**Fig. 5e–h**).

Metastable tumor cells are aneuploid cancer cells that have copy number profiles highly similar to those of the major subpopulations but have evolved additional gains or losses of single chromosomes or chromosome arms (**Fig. 5e**). In tumor T3, we identified 53 single aneuploid tumor cells that shared a copy number profile and 6 unique metastable tumor cells with non-clonal amplifications and deletions. One metastable tumor cell from tumor T3 showed an additional amplification of chromosome 5p in comparison to the tumor cells in the major aneuploid subpopulation (**Fig. 5e–h**, left). In tumor T6, we identified 79 single tumor cells that shared a copy number profile and 6 unique metastable tumor cells with non-clonal CNAs. One metastable tumor cell with an additional amplification of chromosome 18p is shown in comparison to the major aneuploid tumor subpopulation (**Fig. 5e**, right). Metastable tumor cells acquired single CNAs in the later stages of tumor evolution but represent evolutionary 'dead ends' that did not undergo further expansion to achieve prevalence in the tumor mass.

Pseudodiploid cells are single cells with flat 2*N* copy number profiles that have acquired additional gains or losses of single chromosomes or chromosome arms at random genomic locations (**Fig. 5f,g**). Whereas most CNAs were randomly distributed, one exception was a frequent (23%) loss of the X chromosome in multiple cells from different patients ($P < 0.0001$, one-tailed t test) (**Supplementary Table 5**). To determine whether the presence of non-clonal diploid cells was due to a tumor field effect, we profiled normal breast tissues and found that 5.9% of cells also had non-clonal profiles (**Fig. 5d,g**). These data suggest that random copy number gains and losses occur during normal mitosis and are unlikely to be associated with a tumorigenic field effect (**Supplementary Table 6**).

Chromazemic cells (where *zemia* means damage or loss in Greek) are non-clonal cells with large homozygous deletions of whole chromosomes or chromosome arms that occur at random locations in the genome (**Fig. 5h**). These cells are unlikely to be viable as a result of the large homozygous deletions of chromosomes. Chromazemic cells may be the byproduct of asymmetric cell divisions or, possibly, dying cells and were found in diploid fractions, normal tissues and aneuploid fractions.

Punctuated copy number evolution

To trace tumor evolution, we constructed phylogenetic trees from the single-cell copy number data. Intratumoral heterogeneity provides a permanent record of the mutations that occurred during tumor growth, enabling lineages to be reconstructed by assuming that mutational complexity increases with time[8,9]. Copy number segmentation was performed using a multiple-sample breakpoint algorithm[20] to identify common chromosome breakpoints that occurred across single cells within each tumor. We then calculated a trinary event matrix to treat all large and small CNA events equally for phylogenetic analysis using maximum parsimony (Online Methods). The resulting maximum-parsimony trees show that each tumor evolved a long root branch of founder ('truncal') CNAs that were acquired in the early stages of tumor evolution and stably maintained in the clones during tumor growth (**Fig. 6a–c**). Evidence of gradual intermediate branching was not observed as cells progressed from diploid to aneuploid genomes. Although some TNBC tumors showed clear evidence

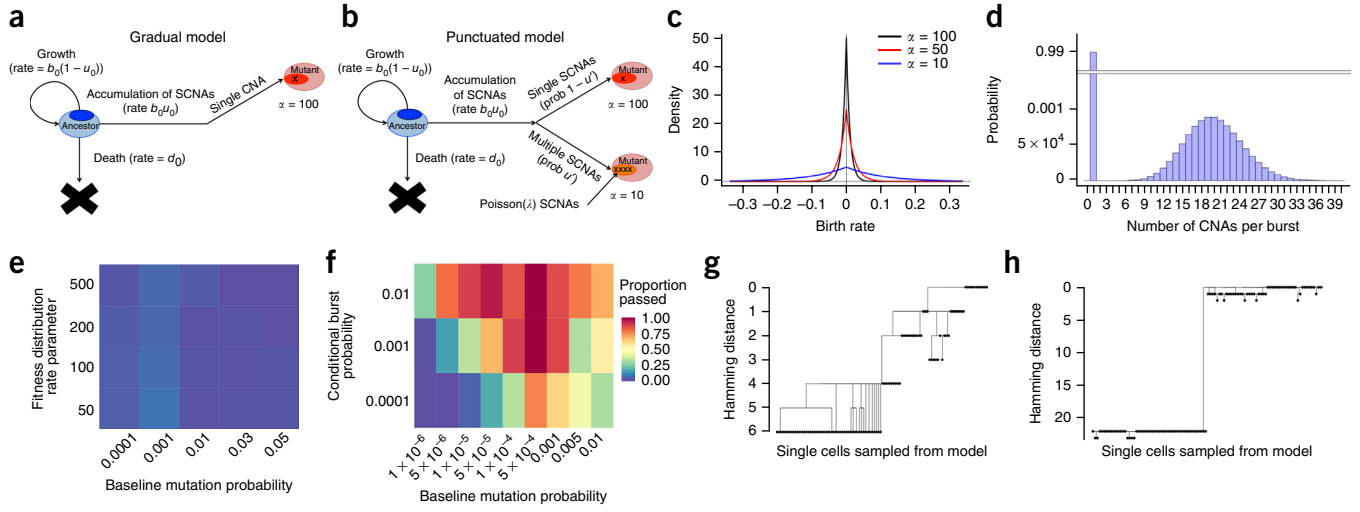

Figure 7 Mathematical modeling of punctuated and gradual tumor evolution. (**a**) Gradual model of the multitype stochastic birth–death–mutation process. The parameter b is the birth rate, u is the mutation rate, d is the death rate and α is the shape parameter for the double exponential distribution (**Supplementary Note**). (**b**) Punctuated model of the multitype stochastic birth–death–mutation process with a Poisson mutation burst probability distribution. SCNA, somatic copy number aberration. (**c**) Fitness distributions with varying shape parameter (α) values used for sampling as new clones emerged during the binary branching process in the gradual or punctuated model. (**d**) Poisson probability distribution for multiple CNA events occurring in the punctuated model, with a single atom (point mass) set at 1 for single CNA events. (**e**) Heat map of AMOVA analysis for different fitness distributions and mutation rates in the gradual model. (**f**) Heat map of AMOVA analysis for different burst and mutation probabilities in the punctuated model. Colors in **e** and **f** correspond to the proportion of simulations passing 'minimal punctuated criteria', with P value <0.05 in AMOVA permutation and >90% of samples having root nodes with at least five CNAs to construct a tree. (**g**) Tree constructed from random sampling of 100 single cells from simulated data in the gradual model. (**h**) Tree constructed from random sampling of 100 single cells from simulated data in the punctuated model.

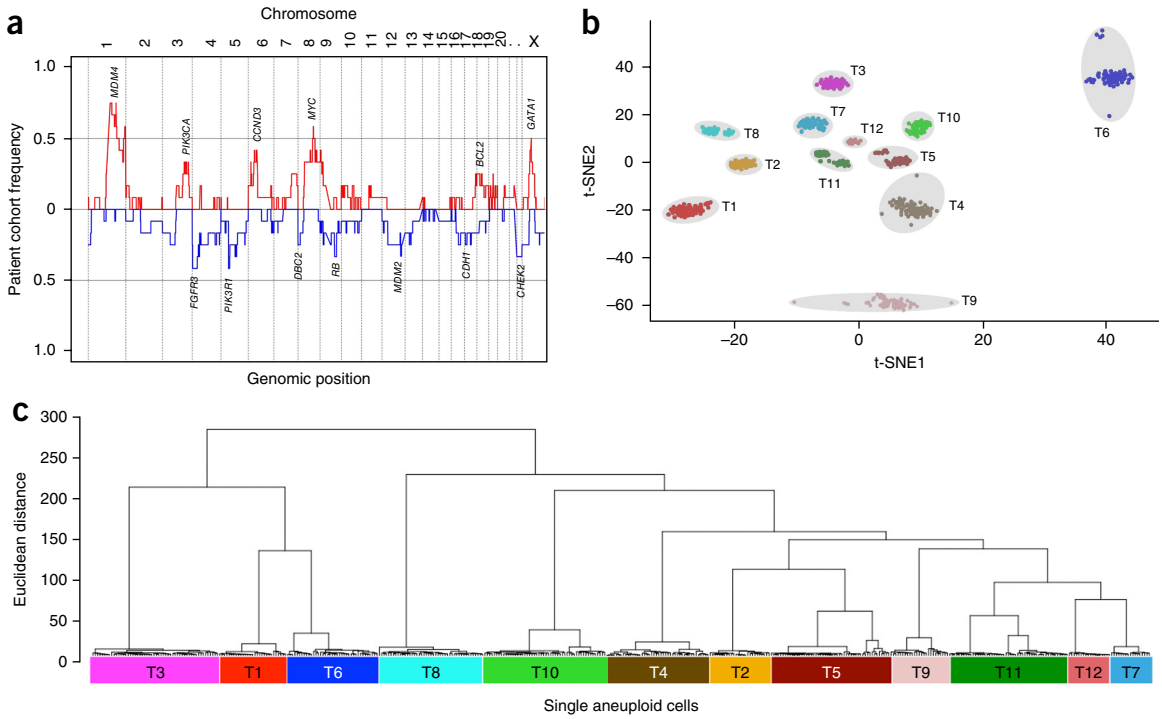

Figure 8 Intertumoral heterogeneity and focal amplifications in TNBCs. (**a**) Frequency plot of CNAs across 12 patients with TNBC: red, amplification; blue, deletion. (**b**) A t-SNE plot was generated using all aneuploid tumor cells from the 12 patients with TNBC. Single cells are colored by individual patient. (**c**) A hierarchical clustering tree using Ward linkage was constructed from the pairwise Euclidean distances between all aneuploid tumor cells from the 12 patients with TNBC.

of divergent subclones in later stages of tumor evolution, these clones typically only diverged by a few (n = 1–3) CNAs, in comparison to the many (n = 24–132) CNAs that were acquired in early punctuated bursts. Another notable characteristic of the phylogenetic trees is that they show that all cancer cells share a common evolutionary origin in each tumor, suggesting that these cells evolved from a single normal cell in breast tissue, not multiple initiating cells.

To further investigate whether the single-cell data were consistent with PCNE, we performed linear (gradual) and multiple-step (punctuated) fitting of the sorted CNA count data from the single cells in each tumor (Online Methods). One-step fit resulted in higher correlation values (adjusted R^2 = 0.977) than linear fitting (adjusted R^2 = 0.704) and was statistically significant ($P = 2.125 \times 10^{-9}$, one-tailed t test) (**Fig. 6d** and **Supplementary Fig. 5**). Similarly, better Bayesian information criterion (BIC) and Akaike information criterion (AIC) values, statistical measurements for model selection, were obtained for all tumors when step fitting was applied. These data support PCNE by showing that a large number of CNA events occurred within a short period of time during tumor evolution.

Absence of gradual intermediate cells in ungated fractions

One possible explanation for the absence of gradual intermediate copy number profiles in the tumor mass is that the gating of ploidy by FACS was too narrow and this analysis may therefore have missed gradual cells with intermediate copy number events that occurred in between ploidy peaks (**Fig. 1b**). To investigate this possibility, we performed universal gating to sample broadly across all ploidy values in 4 of 12 patients with TNBC and flow sorted additional single nuclei for HMSNS. Hierarchical clustering was performed using data from narrowly gated and universally gated nuclei, and heat maps were constructed to compare the clonal substructure (**Supplementary Fig. 6**). Clustering analyses showed similar population substructure in the universally and ploidy-gated populations of tumor cells from each patient, with no evidence of additional intermediate copy number profiles in the universally gated data, suggesting that if intermediate profiles exist and persist in the tumor mass they are very rare. These data are consistent with cell counts in FACS histograms, which showed no evidence of intermediate density between the aneuploid and diploid populations, with the exception of minor S-phase populations (**Fig. 1b**).

Mathematical modeling of gradual and punctuated evolution

To further investigate *in silico* alternative scenarios such as punctuated and gradual evolution, we developed a multitype stochastic branching process model of tumor growth (**Fig. 7**). In this model, during each time step, a cell can divide to produce (i) two daughter cells that are identical to the mother cell, (ii) no cells (death) or (iii) one daughter cell identical to the mother cell and one daughter cell with a new CNA whose fitness advantage is selected from a mutational fitness distribution[21]. In the gradual model, each cell division event leads to the accumulation of a new CNA at a constant rate (**Fig. 7a**) corresponding to the baseline mutation rate for single copy number changes (**Fig. 7c**). In the punctuated model (**Fig. 7b**), each cell division event results in either the accumulation of a single CNA or, at a different rate, a burst of multiple somatic CNAs whose number is chosen from a Poisson distribution (**Fig. 7d**). We implemented both models as exact stochastic computer simulations initiating with a single diploid ancestral tumor cell and continued each instantiation of the model until the total number of cells was equivalent to the total number of cells in each patient with TNBC. From each simulation, we sampled 100 single cells at random and constructed phylogenetic trees (**Supplementary Fig. 7**). We then performed AMOVA[22] to investigate the topologies of the resulting phylogenies. Permutation

testing was applied to obtain *P* values for each sample based on the gradual model (**Fig. 7e**) and the punctuated model (**Fig. 7f**) and to test whether these models were able to recapitulate the tree topologies obtained from the data for patients with TNBC (Online Methods). We investigated a wide range of parameter values by searching through a total of 162 combinations of parameters. In trees resulting from the gradual model, we found evidence of many intermediate subpopulations, suggesting that selective sweeps are unlikely to occur in later stages of tumor evolution, even when clones with high fitness values emerge (**Fig. 7g** and **Supplementary Fig. 7b**). In contrast, the punctuated simulations resulted in tree structures with long root nodes between the ancestral diploid and aneuploid subpopulations, which are consistent with our single-cell data (**Fig. 7h** and **Supplementary Fig. 7a**). We then investigated alternative scenarios for gradual simulations (epistasis, cancer stem cells, increased mutation rates, fixed fitness distributions and mutation rate from a distribution) to test their ability to recapitulate the data (Online Methods and **Supplementary Note**). In total, we investigated these alternative scenarios for a total of 2,097 parameter combinations. However, under all of these scenarios, the trees generated from 100 sampled single cells displayed evidence of many intermediate branching clones (**Supplementary Fig. 7b**). Collectively, these modeling data support PCNE and suggest that selective sweeps occurring in later stages of tumor growth are unlikely to explain the presence of highly clonal subpopulations.

Intertumoral heterogeneity between patients with TNBC

In addition to investigating intratumoral heterogeneity, we also compared copy number differences between patients with TNBC. Consensus profiles were calculated to represent the bulk tumor populations from each patient with TNBC by aggregating the single-cell aneuploid copy number profiles (Online Methods). Frequency plots were generated using data from all patients with TNBC to identify common amplifications and deletions that were recurrent in the patient cohort (**Fig. 8a**). This analysis identified frequent amplifications on chromosomes 1q (*MDM4*), 3q (*PIK3CA*), 6p (*CCND3*), 8q (*MYC*) and 18 (*BCL2* and *SMAD4*), and frequent deletions included chromosomes 4p (*FGFR3*), 5q (*PIK3R1*), 8p (*DBC2*), 9p (*NR4A3*), 12 (*MDM2*) and 22 (*CHEK2*). These genomic regions and oncogenes are consistent with regions previously identified by microarray comparative genomic hybridization (CGH) as having copy number changes in patients with TNBC[14]. In addition to frequent CNAs, we also identified many unique high-level focal amplifications (<10 Mb) that occurred exclusively in individual patients (**Supplementary Fig. 8**). These focal amplifications are consistent with previous reports in patients with TNBC[14,23]. We further investigated between-patient tumor heterogeneity by integrating single-cell data from all patients with TNBC. Dimensionality reduction was performed using t-SNE[24], showing that single cells clustered according to the patient from which they were isolated (**Fig. 8b**). Similarly, hierarchical clustering grouped single cells by patient (**Fig. 8c**). These data show that single cells from individual patients with TNBC are genetically more related to each other than they are to cells from other tumors, suggesting that they share a common ancestral lineage and evolved from a single normal cell in the breast tissue.

DISCUSSION

Collectively, our data support a punctuated model of copy number evolution, in which a large number of CNAs are acquired early in tumor evolution, during a short period of crisis, and remain highly stable as the tumor mass clonally expands (clonal stasis). Despite profiling hundreds of single cells from many spatial regions, we did not detect any intermediate copy number profiles, indicative of gradual evolution, as the tumor cells evolved from diploid to aneuploid genomes. These data challenge the dogma of gradual tumor evolution[4,5] by showing that cancer cells with intermediate copy number profiles are not common during tumor growth. These findings also challenge reports of extensive intratumoral genomic heterogeneity in breast cancer[15,16,18,25] by showing that CNAs are remarkably stable throughout the tumor mass. However, previous studies focused mainly on point mutations, which may correspond to different molecular clocks during tumorigenesis[17]. PCNE is consistent with a 'Big Bang' model for tumor growth[26–28] in which clonal diversification occurs at the earliest stages of tumor progression, leading to the stable expansion of one or more clones.

An analogous model called punctuated equilibrium was originally proposed by Gould and Eldredge in 1972 to explain species evolution[29,30]. This model was mainly supported by evidence in the fossil record and challenged Darwinian gradualism. Several interesting parallels can be drawn between this model and our punctuated model: (i) the occurrence of stasis, (ii) the lack of intermediates corresponding to gradual evolution and (iii) short bursts of rapid evolution. However, it is important to note that the mechanisms underlying punctuated equilibrium in species evolution (for example, allopatric speciation) are likely to be very different from those underlying PCNE in human tumors.

PCNE and clonal stasis have important implications for tumor evolution, diagnostics and therapy. The data suggest that individual tumor cells may be hardwired at the earliest stages of tumor growth and intrinsically preprogrammed to become invasive, metastatic or resistant to chemotherapy[26,31]. This deterministic characteristic may allow oncologists to profile CNAs in early-stage breast cancers (for example, ductal carcinoma *in situ*) to predict whether the tumors should be treated aggressively or, alternatively, not at all ('watchful waiting'). Our single-cell data also have important implications for clinical diagnostics by showing that multiregion sampling may not be necessary to assess CNAs as biomarkers in patients with TNBC, as these aberrations are highly stable throughout the tumor mass.

Although most copy number profiles in patients with TNBC were found to be highly clonal, we also identified a minor (<10%) fraction of cells with non-clonal copy number profiles. These cells were not intermediates in the tumor lineage but instead showed random chromosome gains or losses. To determine whether these cells were present because of a tumor-specific field effect, we also profiled cells from normal breast tissue, which showed percentages of non-clonal cells (5.9%) similar to those in tumors. These data are consistent with recent single-cell genomic data on tissue mosaicism, which have reported 1–5% non-clonal aneuploid cells in different normal tissues, including liver, brain and skin[32]. Because the majority of the non-clonal events involve a single-chromosome gain or loss, we speculate that they are due to the occurrence of lagging chromosomes during asymmetric mitoses[33]. Although such events are unlikely to lead to further cell proliferation in normal tissues, they may provide tumors with additional 'fuel' for evolution, occasionally leading to the emergence of new tumor subpopulations in the later stages of tumor evolution, as we observed in several polyclonal tumors.

Our study has addressed several key questions regarding copy number evolution in patients with TNBC, but it has also raised several new lines of inquiry. How can genome instability be turned on and off at the earliest stages of tumor evolution in a reversible manner? One possibility for a reversible switch is telomerase inactivation and reactivation: telomerase inactivation could lead to complex aneuploid rearrangements within just a few cell divisions, in a manner that

could be reversed by telomerase reactivation. This mechanism has previously been described as 'episodic telomere crisis' and was demonstrated using experimental systems[34–36]. However, we speculate that telomerase inactivation alone is insufficient to cause punctuated evolution, as *TP53* inactivation and genome duplication are also requirements for PCNE. Indeed, previous work using *in vivo* systems has shown that *TP53* and telomere loss can cooperate to drive tumorigenesis[34]. Another important question is how tumor cells with complex aneuploid rearrangements can undergo symmetric cell divisions and stable clonal expansions (clonal stasis). For tumor cells to undergo symmetric cell divisions with supernumerary chromosomes, we speculate that aneuploid cells must cluster multiple centrioles together to align chromosomes equally along the metaphase plate[37,38]. Addressing these interesting questions will require future work, which should be performed using *in vitro* and *in vivo* systems.

We also investigated whether we sequenced a sufficiently large number of cells in each patient with TNBC to detect the major tumor subpopulations. To answer this question, we generated posterior saturation curves with multinomial distributions (**Supplementary Fig. 9**). The resulting data suggest that 20–40 single cells were necessary to detect the major subpopulations with 95% power, suggesting that our sample size was sufficient (mean = 83 cells). Another question we considered is whether an alternative model to PCNE could explain the observed single-cell data, in which evolution is gradual until a clone with high fitness emerges in later stages of progression, leading to a selective sweep. Despite extensive testing with mathematical modeling, we found that selective sweeps were highly uncommon during gradual evolution, even when clones with high fitness emerged. Furthermore, a clonal sweep is inconsistent with studies that support early clonal diversification and selection[26–28].

In summary, our single-cell copy number data and mathematical modeling suggest that clonal stasis and PCNE are common in patients with TNBC. This process leads to complex aneuploid copy number profiles that are remarkably stable during tumor growth and ubiquitous throughout the tumor mass. Our preliminary data in other tumors (colon, prostate, liver and lung) suggest that PCNE may not be restricted to breast cancer and is also likely to operate in other human cancers. This model has important implications for evolutionary understanding of cancer dynamics and for the clinical treatment of patients with TNBC.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. All data from this study have been deposited in the Sequence Read Archive (SRA) under accessions SRP064210 for tumors T1–T10 and SRA018951 for tumors T11 and T12.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Edgerton, J. Kendall, M. Wigler and J. Hicks for their support and discussions. We are also very grateful to the patients with breast cancer at M.D. Anderson for generously donating their tumor tissues to our research studies. This work was supported by a grant from the Lefkofsky Family Foundation. N.E.N. is a Nadia's Gift Foundation Damon Runyon-Rachleff Innovator (DRR-25-13). This work is also supported by grants to N.E.N. from the NCI (1RO1CA169244-01) and the American Cancer Society (129098-RSG-16-092-01-TBG). N.E.N. is a T.C. Hsu Endowed Scholar, an AAAS Wachtel Scholar and an Andrew Sabin Family Fellow. The study is also supported by the Moonshot Knowledge Gap Award and the Center for Genetics and Genomics. This study was supported by the M.D. Anderson Sequencing Core Facility grant (CA016672) and the Flow Cytometry Facility grant (CA016672) from the NIH. Additional funding support includes the Rosalie B. Hite Fellowship (A.C.); a Center for Genetics and Genomics Postdoctoral Fellowship (R.G.); NIH UL1TR000371 (F.M.-B.); the Nellie B. Connally Breast Cancer Research Endowment (F.M.-B.), Susan Komen SAC10006 (F.M.-B.), CPRIT RP110584 (F.M.-B.) and the M.D. Anderson Cancer Center Support grant (NIH/NCI P30CA016672). F.M. gratefully acknowledges support from the Dana-Farber Cancer Institute Physical Science Oncology Center (U54CA193461-01).

AUTHOR CONTRIBUTIONS

R.G. analyzed the data and wrote the manuscript. A.D. analyzed the data. T.O.M. and F.M. performed mathematical modeling and wrote the manuscript. E.S., X.S., P.-C.T. and J.W. performed experiments. A.C. and Y.W. analyzed the data. H.Z. and F.M.-B. provided tumor samples and interpreted the data. N.E.N. analyzed the data, led the project and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
2. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16**, 1465–1479 (2006).
3. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
4. Fearon, E.R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
5. Höglund, M., Gisselsson, D., Hansen, G.B., Säll, T. & Mitelman, F. Multivariate analysis of chromosomal imbalances in breast cancer delineates cytogenetic pathways and reveals complex relationships among imbalances. *Cancer Res.* **62**, 2675–2680 (2002).
6. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
7. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
8. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20**, 68–80 (2010).
9. Navin, N.E. & Hicks, J. Tracing the tumor lineage. *Mol. Oncol.* **4**, 267–283 (2010).
10. Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nat. Protoc.* **7**, 1024–1041 (2012).
11. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
12. Rakha, E.A., Reis-Filho, J.S. & Ellis, I.O. Basal-like breast cancer: a critical review. *J. Clin. Oncol.* **26**, 2568–2581 (2008).
13. Foulkes, W.D., Smith, I.E. & Reis-Filho, J.S. Triple-negative breast cancer. *N. Engl. J. Med.* **363**, 1938–1948 (2010).
14. Turner, N. *et al.* Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. *Oncogene* **29**, 2013–2023 (2010).
15. Shah, S.P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
16. Almendro, V. *et al.* Inference of tumor evolution during chemotherapy by computational modeling and *in situ* analysis of genetic and phenotypic cellular diversity. *Cell Reports* **6**, 514–527 (2014).
17. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
18. Yates, L.R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
19. Reynolds, A.P., Richards, G., de la Iglesia, B. & Rayward-Smith, V.J. Clustering rules: comparison of partitioning and hierarchical clustering algorithms. *J. Math. Model. Algorithms* **5**, 475–504 (2006).
20. Nilsen, G. *et al.* Copynumber: efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
21. Foo, J. *et al.* An evolutionary approach for identifying driver mutations in colorectal cancer. *PLoS Comput. Biol.* **11**, e1004350 (2015).
22. Excoffier, L., Smouse, P.E. & Quattro, J.M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491 (1992).
23. Lips, E.H. *et al.* Next generation sequencing of triple negative breast cancer to find predictors for chemotherapy response. *Breast Cancer Res.* **17**, 134 (2015).
24. Bushati, N., Smith, J., Briscoe, J. & Watkins, C. An intuitive graphical visualization technique for the interrogation of transcriptome data. *Nucleic Acids Res.* **39**, 7380–7389 (2011).

25. Park, S.Y., Gönen, M., Kim, H.J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of *in situ* human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120**, 636–644 (2010).
26. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
27. Ling, S. *et al.* Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proc. Natl. Acad. Sci. USA* **112**, E6496–E6505 (2015).
28. 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).
29. Gould, S.J. & Eldredge, N. Punctuated equilibrium comes of age. *Nature* **366**, 223–227 (1993).
30. Eldredge, N. & Gould, S.J. in *Models in Paleobiology* (ed. Schopf, T.J.M.) 82–115 (Freeman, Cooper and Co., 1972).
31. DePinho, R.A. & Polyak, K. Cancer chromosomes in crisis. *Nat. Genet.* **36**, 932–934 (2004).
32. Knouse, K.A., Wu, J., Whittaker, C.A. & Amon, A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc. Natl. Acad. Sci. USA* **111**, 13409–13414 (2014).
33. Bakhoum, S.F. & Compton, D.A. Chromosomal instability and cancer: a complex relationship with therapeutic potential. *J. Clin. Invest.* **122**, 1138–1143 (2012).
34. Chin, L. *et al.* p53 deficiency rescues the adverse effects of telomere loss and cooperates with telomere dysfunction to accelerate carcinogenesis. *Cell* **97**, 527–538 (1999).
35. Artandi, S.E. & DePinho, R.A. A critical role for telomeres in suppressing and facilitating carcinogenesis. *Curr. Opin. Genet. Dev.* **10**, 39–46 (2000).
36. Chin, K. *et al. In situ* analyses of genome instability in breast cancer. *Nat. Genet.* **36**, 984–988 (2004).
37. Leber, B. *et al.* Proteins required for centrosome clustering in cancer cells. *Sci. Transl. Med.* **2**, 33ra38 (2010).
38. Kwon, M. *et al.* Mechanisms to suppress multipolar divisions in cancer cells with extra centrosomes. *Genes Dev.* **22**, 2189–2203 (2008).

ONLINE METHODS

Triple-negative breast cancer samples. Frozen tumors from 12 patients with TNBC were selected with poorly differentiated and high-grade (grade III) invasive ductal carcinomas as determined by Bloom–Richardson score. The triple-negative status of the tumor samples was determined by immunohistochemistry for ER (<1%) and PR (<1%) and FISH analysis of *HER2* amplification using the CEP-17 centromere control probe (ratio of *HER2*/CEP17 < 2.2). The frozen tumor samples and matched normal breast tissues were obtained from the University of Texas M.D. Anderson Cancer Center Breast Tissue Bank. Two frozen tumor samples (T11 and T12) were obtained from the Cooperative Human Tissue Network (CHTN). This study was approved by the Institutional Review Board (IRB) at the University of Texas M.D. Anderson Cancer Center. Patients were consented by an informed consent process that was reviewed by the IRB.

Highly multiplexed single-nucleus sequencing. Nuclei from frozen tumors were isolated using NST-DAPI buffer (800 ml of NST (146 mM NaCl, 10 mM Tris base at pH 7.8, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 0.05% BSA and 0.2% Nonidet P-40), 200 ml of 106 mM $MgCl_2$, 10 mg of DAPI and 5 mM EDTA). Frozen tumors were dissociated into nuclear suspensions by mincing with no.11 surgical scalpels in 1 ml of NST-DAPI cytoplasmic lysis buffer at 4 °C using ice blocks in a plastic Petri dish. Nuclear suspensions were filtered through 37-µm plastic mesh before flow sorting into a 5-ml polystyrene tube (Falcon). Single nuclei were flow sorted into 96-well plates by FACS using the FACSAria II flow cytometer (BD Biosciences). Ploidy distributions were gated by differences in total genomic DNA content as determined by DAPI intensity. To establish the DAPI fluorescence intensity corresponding to diploid (2*N*) cells, a lymphoblast control cell line (REFM) was flow sorted first to establish gates. Before flow sorting single nuclei, a few thousand cells were sorted to establish the DNA content distributions for gating by ploidy. Single nuclei were collected from both the diploid and aneuploid gated fractions. Additionally, nuclei were collected from each tumor by gating broadly across all ploidy values. Single nuclei were deposited into individual wells of a 96-well plate each containing 10 µl of lysis solution from the Sigma-Aldrich GenomePlex WGA4 kit, along with negative-control reactions, in which no nuclei were deposited.

Whole-genome amplification and barcoded library construction. Whole-genome amplification was performed on single flow-sorted nuclei using DOP-PCR as described in the protocol for the Sigma-Aldrich GenomePlex WGA4 kit (WGA4-50RXN). For quality control of amplification performance, DNA concentration was measured (Thermo Fisher Scientific, Qubit 2.0 fluorometer), and reactions were run out using gel electrophoresis to determine size distributions. To prepare sequencing libraries by TA ligation cloning, 500 ng of DNA was acoustically sonicated to 200 bp in size using the S220 Sonicator (Covaris). Fragmented whole-genome amplification products underwent end repair (New England BioLabs, E6050L) and were purified with the DNA Clean and Concentrator-5 kit (Genesee, 11-303 or 11-306). Libraries were constructed using NEBNext DNA Library Prep enzymes (New England BioLabs, E6050L, E6053L, E6056L/M0202L and M0541L) for end repair, 3′ adenylation, ligation and PCR amplification according to the manufacturer's instructions but using different P7 adaptors to barcode each single-cell library with a unique 8-bp identifier and common P5 adaptors for sample multiplexing. The 96 unique P7 indexes were NEXTflex-96 barcodes that were purchased from Bio Scientific. After ligation, DNA underwent negative and positive selection with AMPure XP beads (Beckman Coulter, A63881), 0.7× and 0.15×, respectively, before PCR amplification. Final library concentrations were measured using a Qubit 2.0 fluorometer, and 48–96 single-cell libraries were pooled in equimolar concentrations. The final concentration of the pooled libraries was measured by quantitative PCR using the KAPA Library Quantification kit (KAPA Biosystems, KK4835) and an ABI PRISM real-time PCR machine (Applied Biosystems 7900HT), as well as a 2100 Bioanalyzer (Agilent Technologies).

Multiplexed Illumina next-generation sequencing. Pooled libraries containing 48–96 barcoded single-cell libraries were sequenced using 76 single-end cycles on the HiSeq 2000 system (Illumina) at the Sequencing Core Facility of the Genetics Department at M.D. Anderson Cancer Center to obtain target coverage of 0.1× for each single-cell library. Data were processed using the CASAVA 1.8.1 pipeline (Illumina), and sequence reads were converted to a master fastq file. Sequencing reads from each single cell were demultiplexed using an in-house Perl script (demultiplex.pl) into 48–96 independent fastq files, where each file represented the sequencing reads from one cell.

Sequence alignment and data processing. After barcodes and sequencing adaptors were trimmed, sequence reads in fastq format were mapped to human genome assembly NCBI Build 37 (hg19/NCBI37), using Bowtie 2 alignment software[39] with default parameters to generate SAM files. SAMtools (0.1.19) was used to convert SAM files to compressed BAM files and sort the BAM files by chromosome coordinates[40]. To eliminate PCR duplicates, SAMtools was used to remove sequence reads with identical start coordinates. Sequence reads with low mapping quality (MQ <40) were also filtered out using SAMtools.

Integer copy number calculation from single-cell data. The sequencing data were counted in 11,927 genomic bins with variable start and stop coordinates, using the variable binning method as previously described[10,11]. The median genomic length spanned by each bin was 220kb. This variable binning approach reduces mappability errors and the number of false deletion events when compared to an approach using scaffolds using uniform, length-fixed bins. A blacklist of aberrant bins was filtered out to remove false positive amplifications in centromeric and telomeric regions. Aberrant bins were defined as bins where 5–95% of ratios were distant from the ground states (|difference of ratio| > 0.5) in at least two-thirds of normal single-cell populations or bins with systematic artifacts where ratios were extremely high (>10) or low (<0.001) across all single cells. Only single cells with ≥50 median reads/bin were included for downstream copy number analysis. We then applied Loess normalization to correct for bias from GC content[10]. Copy number profiles were segmented using circular binary segmentation (CBS)[41] followed by MergeLevels[42] to join adjacent segments with non-significant differences in segmented ratios. The parameters used for CBS segmentation were alpha = 0.0001 and undo.prune = 0.05. Default parameters were used to perform MergeLevels, which successfully joined false positive detections of erroneous breakpoints. exp.mad was calculated as the median distance between the $\log_2$-transformed ratio and segmented values. Only segmentations having at least 1.48 times exp.mad deviation were retained as CNAs. Finally, integer copy number was calculated by scaling segmented ratios with average DNA ploidy determined by flow sorting indexes and rounding to the closest integer values (**Supplementary Code**). When DNA ploidy information was unavailable for universally gated single cells, the least-square rounding method was applied to obtain the optimum scaling factor that had the least sum of deviations from the closet integer after rounding[43]. Lastly, we filtered out diploid single cells with variant coefficients of bin counts larger than 0.4 or having mean_resid values greater than 0.03. We calculated mean_resid as the average deviation of the scaled ratio from the true ground state integer value, which was 2, as cells were diploid (2*N*). We filtered out aneuploid single cells with median absolute deviation (MAD) for genome-wide ratios greater than 0.62 or autocorrelation values for neighboring data points less than 0.53. Autocorrelation values were calculated over sliding windows using a 10-bin interval size across the 11,927 bins in the human genome scaffold. This window size results in low correlation values in regions where adjacent data points have random values. These steps removed single cells with poor whole-genome amplification from the subsequent multivariate data analysis.

Cancer gene annotations. Amplifications and deletions identified in the single-cell copy number profiles were annotated for known cancer-related genes, which consisted of 413 genes that were compiled from multiple databases, including the Cancer Gene Census[44], The Cancer Gene Atlas (TCGA) Project and the National Cancer Institute (NCI) cancer gene index (Sophic Systems Alliance, Biomax Informatics). BEDTools[45] was used with the IntersectBED function to find the intersection of BED files for known cancer-related genes and regions of chromosome amplification and deletion that were detected in the single-cell sequencing data sets.

Clustering for single-cell copy number profiles. To construct the clustering heat maps, Euclidean distances were calculated from the copy number data

matrix, where each column represented one single cell and each row contained $\log_2$ (ratio + 0.1) data for each segment. 1D hierarchical clustering was performed in R using the heatmap.2 function from the gplots package available on CRAN[46]. Columns representing a single cell each were hierarchically clustered using Ward linkage on the basis of pairwise Euclidean distances, and the *x* axis was ordered by genome position. To estimate the optimum number of clusters for each patient, we performed partition around medoids clustering with the optimum Calinski–Harabasz index[47] or average silhouette width using the pamk[48] function from the fpc package. Clusters with singleton cells were collapsed and penalized using pamk criteria to minimize technical artifacts. PAM clustering was performed on a range of *k* values from 1–20.

High-dimensional data analysis methods. For each individual tumor, the numeric matrix containing integer copy numbers was used to perform PCA with the prcomp function in R (ref. 46). The columns in the numeric matrix were segmented bins, and each row was an individual single cell. PC1 and PC2 were plotted on the *x* and *y* axes, respectively, for each plot. Each dot on the PCA plots represents a single-cell copy number profile and is colored according to cells that clustered together into subpopulations that were identified by the hierarchical clustering analysis. To determine the genomic relationship of all aneuploid tumor cells from the 12 patients with TNBC, the t-distribution stochastic neighbor embedding (t-SNE)[49] method was applied on the basis of pairwise Euclidean distances from the ratio data. The t-SNE method is an improved nonlinear dimensionality reduction and visualization method, with which both local and global structure in high dimension can be visualized in low-dimensional plots, while avoiding dramatic masking of very similar data points seen in PCA plots.

Calculation of the subclonal diversity index. To calculate the subpopulation diversity index for each tumor, we performed hierarchical clustering of copy number data to cluster the aneuploid tumor cells into 1–3 major groups ('species') on the basis of Euclidean distances. Cells within each subpopulation were defined as highly correlated with mean R^2 >0.8. We then calculated the proportion (p) of cells that belonged to each distinct group. The subpopulation diversity index was then calculated as the Shannon index: $Dc = -\Sigma_i(p_i \times lnp_i)$, where larger values represent higher subclonal diversity within the tumor.

Clonal frequency of subpopulations. To calculate the clonal frequency of each clonal subpopulation, we first identified clusters of genotypes by hierarchical clustering, and optimal clustering results were selected on the basis of the Calinski–Harabasz index[47] or average silhouette width. We then counted the number of cells that were classified into each subcluster. Relative clonal frequencies were calculated as the number of cells that fell into each specific subcluster divided by the total number of clonal aneuploid cells. Singleton cells that formed the only member of a subcluster were defined as non-clonal and were excluded from this calculation.

Copy number aberration frequency calculation and plots. Consensus copy number integer profiles for each tumor were calculated using the median integer copy number segment values of all aneuploid single cells from each tumor. To calculate the frequency plot of the 12 TNBC samples, the mean copy number values across the genomic bins of each cell were treated as the ground-state copy number and 1.5 s.d. across the genome as deviation cutoff values. If a copy number was higher than mean + 1.5 × s.d., then a significant amplification was designated, while for copy number lower than mean −1.5 × s.d. a significant deletion was designated. The amplification and deletion frequencies across all tumors were calculated by first counting the total number of consensus tumor profiles that had significant amplifications or deletions in each of the 11,927 bins across the genome and then dividing the counts by the total number of consensus profiles.

Multiple-cell segmentation and event matrix construction. To detect common chromosome breakpoints and segments that were shared by single-cell samples, we applied a multiple-sample population segmentation algorithm using a Bioconductor R package (copynumber)[20], with regularization parameter γ = 40 (default). Segments smaller than 20 bins were removed, and their flanking segments were joined, or separated at the center of the removed segment if they differed significantly (Wilcoxon test, Hommel-adjusted $P < 0.05$ in at least two cells)[50]. The ground state of each cell was calculated by rounding its expected ploidy to the nearest integer[43]. For each tumor, a median matrix M was constructed, in which M is the median of the ith segment in the jth cell. From this median matrix, an event matrix E was calculated as follows. Let g_i be the ground state of the jth cell. $E_{ij} = 1$ (amplification) if $M_{ij} - g_i > 0.6$, $E_{ij} = -1$ (deletion) if $M_{ij} - g_i < -0.6$, and $E_{ij} = 0$ (neutral) if $|M_{ij} - g_i| < 0.4$. If $0.4 \leq |M_{ij} - g_i| \leq 0.6$, E_{ij} was treated as missing with systematic artifact. Segments that had missing values systematically across all cells were removed if they satisfied the following criterion

$$\prod_{j=1}^{s} \frac{u(\{M_{ij}\})}{c_{0.2}(\{M_{ij}\})} > 99$$

where u is the probability density function (*p.d.f.*) of a uniform distribution on (0, 1), $c_{0.2}$ is the cardioid *p.d.f.* (ref. 51) with concentration parameter $\rho = 0.2$ and $\{M_{ij}\}$ is the fractional part of M_{ij}. This formula is a Bayes factor for comparing a model in which ploidy-scaled segment medians do not cluster around integer values to one in which they do, and the chosen cutoff represents 99% certainty of the first model assuming equal prior probabilities.

Phylogenetic tree construction using maximum parsimony. Maximum-parsimony trees were calculated from event matrices using the parsimony ratchet algorithm with R package phangorn[52]. Amplification, neutral and deletion were treated as characters, and missing values were treated as ambiguous sites. Events occurring on sex chromosomes were ignored. Metastable cells were removed from each tumor for phylogenetic analyses, as they do not share CNAs in the main tumor lineages. Cells were also removed if at least one-third of events were missing values. Branch lengths and ancestral character probability distributions were inferred using the Acctran algorithm[52]. Altered sites on each edge were estimated as sites such that

$$\forall_c : P(A_i = c) = 0 \vee P(B_i = c) = 0$$

where A and B are the ancestral sequences estimated at each node and c is a character (**Supplementary Code**).

Phylogenic tree visualization. Phylogenic trees were exported in Newick format from R-studio and plotted as square trees using MATLAB (MathWorks). The trees were re-rooted by the top node of diploid cells. Each individual single cell was represented as a tip of the tree; nodes are colored on the basis of subclonal population. Single cells from the same subpopulations were flipped to physically nearby with each other to favor visualization.

Mathematical modeling of gradual and punctuated evolution. Details on mathematical modeling of gradual and punctuated tumor growth are described in the **Supplementary Note**.

Statistical fitting of copy number aberrations. We first counted the total number of CNA events within each single cell from the collapsed trinary event matrix. To minimize technical noise, singleton CNAs that existed in only one cell or CNA events with missing values in >50% of cells, or cells with >40% missing values were excluded from analysis. Subsequently, single cells were sorted on the basis of the total number of CNA events in each cell. For the gradual linear model, we assumed that tumor cells evolved through intermediate genomes and therefore the total number of CNAs increased gradually over time. We therefore fit a linear model of total segments with slope CNAs = time + error. For the punctuated model, we assumed that tumor cells lack gradual intermediate species and therefore the total CNAs changed over one or more critical evolutionary steps reflected in the jumping of events. The punctual model is CNAs = step + error, with no slope, where the fitted values are simply the average numbers of CNAs per cell in each step. Models were fit using the lm function in R (ref. 46). BIC, AIC and adjusted R^2 values were calculated to measure the best fit of each model for the tumor data sets.

doi:10.1038/ng.3641

Saturation analysis to estimate required sample sizes. A *post-hoc* saturation analysis was performed to determine whether we sequenced sufficient cells for the purpose of this study. We first obtained the total number of subpopulations and the fractions of each subpopulation in each tumor by hierarchical clustering single cells with copy number data as described above. We then calculated the accumulative probability of observing at least three single cells in each subpopulation given the numbers of sequenced cells, by assuming the number of observed cells followed a binomial distribution for biclonal tumors and a multinomial distribution for triclonal tumors. The two monogenomic tumors were excluded from this analysis. The cumulative probabilities were calculated in R (ref. 46).

39. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
41. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
42. Willenbrock, H. & Fridlyand, J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084–4091 (2005).
43. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060 (2015).
44. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
45. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
46. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2016).
47. Calinski, R.B. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **3**, 27 (1974).
48. Kaufman, L. & Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, 2005).
49. Maaten, L.J.P.d. & Hinton, G.E. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 27 (2008).
50. Hommel, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–386 (1988).
51. Pewsey, A., Neuhäuser, M. & Ruxton, G.D. *Circular Statistics in R* (Oxford University Press, 2013).
52. Schliep, K.P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).