ARTICLE

doi:10.1038/nature13600

Clonal evolution in breast cancer revealed by single nucleus genome sequencing

Yong Wang[1], Jill Waters[1], Marco L. Leung[1,2], Anna Unruh[1], Whijae Roh[1], Xiuqing Shi[1], Ken Chen[3], Paul Scheet[2,4], Selina Vattathil[2,4], Han Liang[3], Asha Multani[1], Hong Zhang[5], Rui Zhao[6], Franziska Michor[6], Funda Meric-Bernstam[7] & Nicholas E. Navin[1,2,3]

Sequencing studies of breast tumour cohorts have identified many prevalent mutations, but provide limited insight into the genomic diversity within tumours. Here we developed a whole-genome and exome single cell sequencing approach called nuc-seq that uses G2/M nuclei to achieve 91% mean coverage breadth. We applied this method to sequence single normal and tumour nuclei from an oestrogen-receptor-positive (ER^+) breast cancer and a triple-negative ductal carcinoma. In parallel, we performed single nuclei copy number profiling. Our data show that aneuploid rearrangements occurred early in tumour evolution and remained highly stable as the tumour masses clonally expanded. In contrast, point mutations evolved gradually, generating extensive clonal diversity. Using targeted single-molecule sequencing, many of the diverse mutations were shown to occur at low frequencies (<10%) in the tumour mass. Using mathematical modelling we found that the triple-negative tumour cells had an increased mutation rate (13.3×), whereas the ER^+ tumour cells did not. These findings have important implications for the diagnosis, therapeutic treatment and evolution of chemoresistance in breast cancer.

Human breast cancers often display intratumour genomic heterogeneity[1–3]. This clonal diversity confounds the clinical diagnosis and basic research of human cancers. Expression profiling has shown that breast cancers can be classified into five molecular subtypes that correlate with the presence of oestrogen, progesterone and Her2 receptors[4]. Among these, triple-negative breast cancers ($ER^-/PR^-/Her2^-$) have been shown to harbour the largest number of mutations, whereas luminal A ($ER^+/PR^+/Her2^-$) breast cancers show the lowest frequencies[5–7]. These data suggest that triple-negative breast cancers (TNBCs) may have increased clonal diversity and mutational evolution, but such inferences are difficult to make in bulk tissues[8,9]. To gain better insight into the genomic diversity of breast tumours, we developed a single cell genome sequencing method and applied it to study mutational evolution in an ER^+ breast cancer (ERBC) and a TNBC patient. We combined this approach with targeted duplex[10] single-molecule sequencing to profile thousands of cells and understand the role of rare mutations in tumour evolution.

Whole-genome sequencing using G2/M nuclei

In our previous work we developed a method using degenerate-oligonucleotide PCR and sparse sequencing to measure copy number profiles of single cells[11]. Although adequate for copy number detection, this method could not resolve genome-wide mutations at base-pair resolution. We attempted to increase coverage by deep-sequencing these libraries, but found that coverage breadth approached a limit near 10% (Fig. 1a). To address this problem, we developed a high-coverage, whole-genome and exome single cell sequencing method called nuc-seq (Extended Data Fig. 1). In this method we exploit the natural cell cycle, in which single cells duplicate their genome during S phase, expanding their DNA from 6 to 12 picograms before cytokinesis. This approach provides an advantage over using chemical inhibitors to induce polyploidy in single cells[12,13] because it does not require live cells.

We input four (or more) copies of each single cell genome for whole-genome-amplification (WGA) to decrease the allelic dropout and false positive error rates, which are major sources of error during multiple-displacement amplification (MDA)[14,15]. Additionally, we limit the MDA time to 80 min to mitigate false positive (FP) errors associated with the infidelity of the ϕ29 polymerase (Methods). The improved amplification efficiency can be shown using 22 chromosome-specific primer pairs for PCR (Extended Data Fig. 2). In G1/G0 single cells we find that only 25.58% (11/43) of the cells show full amplification of the chromosomes, whereas G2/M cells have 45.34% (39/86). After MDA, we incubate the amplified DNA with a Tn5 transposase, which simultaneously fragments DNA and ligates adapters for sequencing[16]. The libraries are then multiplexed for exome capture or used directly for next-generation sequencing.

Method validation in a monoclonal cancer cell line

To validate our method we used a breast cancer cell line (SK-BR-3) that was previously shown to be genetically monoclonal[11,17]. We evaluated the genetic homogeneity of this cell line using spectral karyotyping and found that large chromosome rearrangements were highly stable in 85.80% of the single cells (Supplementary Table 1). We also performed single nucleus sequencing (SNS)[11,18] on 50 single SK-BR-3 cells and calculated copy number profiles at 220 kilobase (kb) resolution, which showed that the major amplifications of *MET, MYC, ERBB2, BCAS1* and a deletion in *DCC* were stable (mean $R^2 = 0.91$) in all of the 50 cells (Fig. 1b). Next, we deep-sequenced the SK-BR-3 cell population (SKP) at high coverage depth (51×) and breadth (90.40%) and detected single-nucleotide variants (SNVs), copy number aberrations (CNAs) and structural variants (SVs) using our processing pipeline (Methods). We filtered the variants using dbSNP135 and identified 409 non-synonymous variants and 1,452 structural variants (Fig. 1d), several of which occurred in cancer genes (Supplementary Table 2).

We applied nuc-seq to sequence the whole genomes of two single SK-BR-3 cells (SK1 and SK2) and calculated coverage depth, breadth (sites with at least one read) and uniformity (evenness). We found that both SK-BR-3 cells achieved high coverage depth (61× ± 5 s.e.m., $n = 2$)

[1]The University of Texas MD Anderson Cancer Center, Department of Genetics, Houston, Texas 77030, USA. [2]The University of Texas Graduate School of Biomedical Sciences, Houston, Texas 77030, USA. [3]The University of Texas MD Anderson Cancer Center, Department of Bioinformatics and Computational Biology, Houston, Texas 77030, USA. [4]The University of Texas MD Anderson Cancer Center, Department of Epidemiology, Houston, Texas 77030, USA. [5]The University of Texas MD Anderson Cancer Center, Department of Pathology, Houston, Texas 77030, USA. [6]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, and Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02215, USA. [7]The University of Texas MD Anderson Cancer Center Department of Investigational Cancer Therapeutics, Houston, Texas 77030, USA.

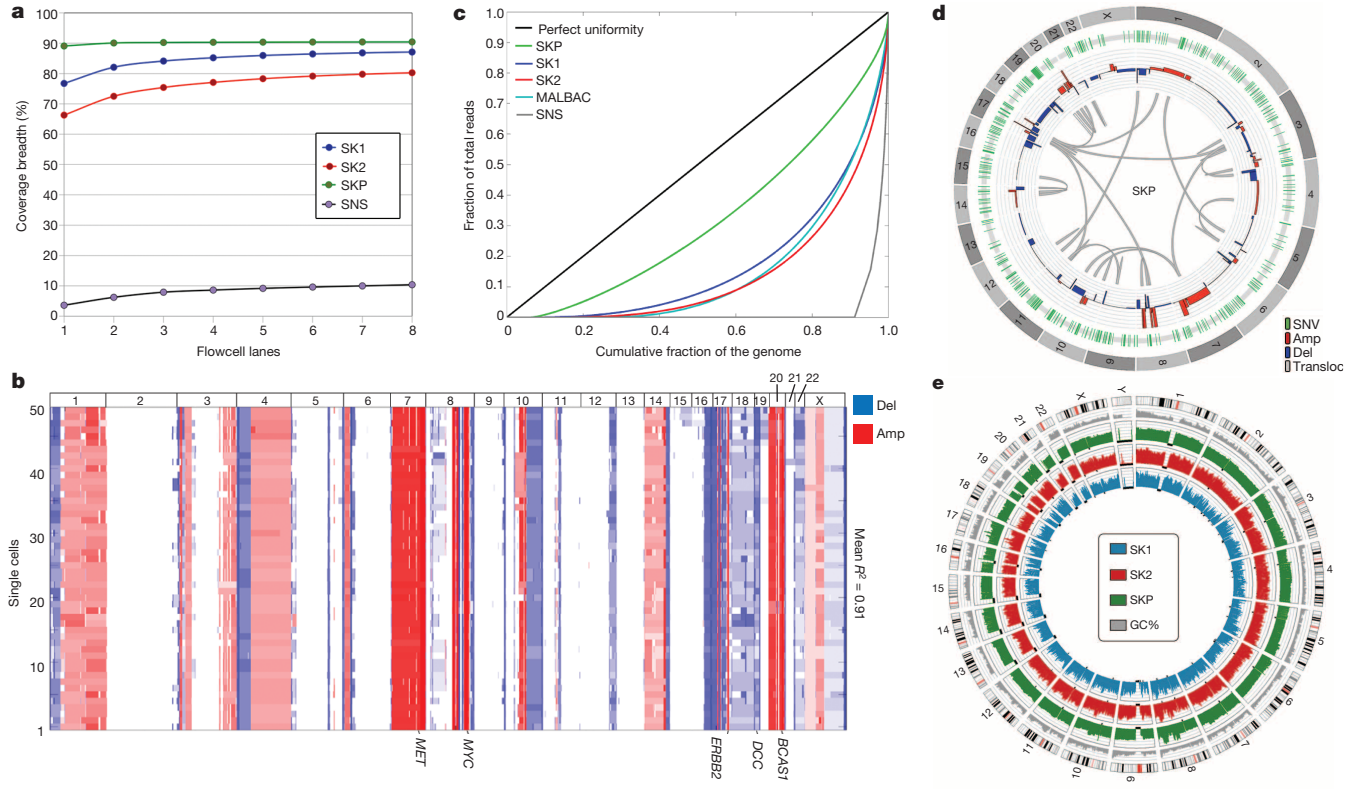


Figure 1 | Method performance in a monoclonal cell line. **a**, Coverage breadth for single cells (SK1, SK2) sequenced by nuc-seq, a single cell SNS library and a SK-BR-3 population (SKP) sample. **b**, Heatmap of 50 single cell SK-BR-3 copy number profiles. **c**, Lorenz curve of coverage uniformity for the single SK-BR-3 cells sequenced by nuc-seq, a cell sequenced by SNS, a population of SK-BR-3 cells, and a cell sequenced by MALBAC. **d**, Circos plot of variants detected by sequencing populations of SK-BR-3 cells. **e**, Coverage depth for the SK-BR-3 population sample and the SK1 and SK2 single cells.

and breadth (83.70 ± 3.40% s.e.m., $n = 2$) (Fig. 1e). In comparison, we re-analysed coverage breadth in single cells sequenced by MALBAC[19] using unique reads and calculated 69.54% coverage breadth. We evaluated coverage uniformity using Lorenz curves[20] which showed highly uniform coverage, representing a major improvement over our previous SNS method[11,18] and is equivalent to the MALBAC data[19] (Fig. 1c). Next, we calculated error rates, including the allelic dropout rate (ADR) and false positive rate (FPR) by comparing single cell variants to the population data (Methods). Our analysis suggests that nuc-seq generates low allelic dropout rates (9.73 ± 2.19%) compared to previous studies (7–46%)[14]. We also achieved low false positive error rates for point mutations (FPR = 1.24×10^{-6}), equivalent to 1–2 errors per million bases, which represents a major technical improvement over previous methods[14,19] (FPR = 2.52×10^{-5} and 4×10^{-5}).

Population and single nuclei sequencing of an ERBC

We selected an invasive ductal carcinoma from an oestrogen-receptor positive ($ER^+/PR^+/Her2^-$) breast cancer patient for population and single cell sequencing (Fig. 2a, Methods). We flow-sorted millions of nuclei from the aneuploid G2/M peak (6N) and from matched normal tissue for population sequencing (46× and 54×) (Fig. 2b). We also flow-sorted 50 single nuclei for copy number profiling, 4 nuclei for whole-genome sequencing and 59 nuclei for exome sequencing. After filtering germline variants, we identified a total of 4,162 somatic SNVs in the aneuploid tumour cell population. Among these SNVs we identified 12 non-synonymous mutations, which we validated by exome sequencing (66×). Several non-synonymous mutations occurred in cancer genes, including *PIK3CA, CASP3, FBN2* and *PPP2R5E* (Fig. 2c, Supplementary Information). *PIK3CA* is the most common driver mutation in luminal A breast cancers[7,9].

To investigate copy number diversity, we performed single nucleus sequencing[11,18] on 50 single nuclei. We constructed a neighbour-joining tree, which showed that single tumour cells shared highly similar CNAs (mean $R^2 = 0.89$), representing a monoclonal population (Fig. 2d, Extended Data Fig. 3a). Next, we performed whole-genome sequencing of four single tumour nuclei at high coverage breadth (80.79 ± 3.31% s.e.m., $n = 4$) and depth (mean 46.75× ± 5.06 s.e.m., $n = 4$). From this data we identified three classes of mutations: (1) clonal mutations, detected in the population sample and in the majority of single tumour cells; (2) subclonal mutations, detected in two or more single cells, but not in the bulk tumour; and (3) *de novo* mutations, found in only one tumour cell. The *de novo* mutations are difficult to distinguish from technical errors and were therefore excluded from our initial analysis. In total we detected 12 clonal non-synonymous mutations and 32 subclonal mutations (Fig. 2e). Many subclonal mutations occurred in intergenic regions; however, two mutations (*MARCH11* and *CABP2*) were found in coding regions (Supplementary Table 4).

To identify additional subclonal mutations, we performed single nuclei exome sequencing on a larger set of cells (47 tumour cells and 12 normal cells). Each nucleus was sequenced at 46.78× (46.78 ± 4.95, s.e.m., $n = 59$) coverage depth and 92.77% (92.77 ± 4.85, s.e.m., $n = 59$) coverage breadth, from which somatic mutations were detected (Supplementary Table 5). The mutations were clustered and sorted by frequency to construct a heatmap (Fig. 2f). As expected, the 17 clonal mutations identified by population sequencing were present in many of the single tumour cells, however, we also identified 22 new subclonal mutations. In contrast, only a single subclonal mutation was detected in the 12 normal cells (Fig. 2f, right panel).

Population and single nuclei sequencing of a TNBC

We then proceeded to analyse a triple-negative ($ER^-/PR^-/Her2^-$) breast cancer (TNBC) (Fig. 3a). We performed population sequencing of the bulk tumour (72×) and matched normal tissue (74×), and identified 374 non-synonymous mutations. A number of mutations occurred in cancer genes, including *PTEN, TBX3, NOTCH2, JAK1, ARAF, NOTCH3,*

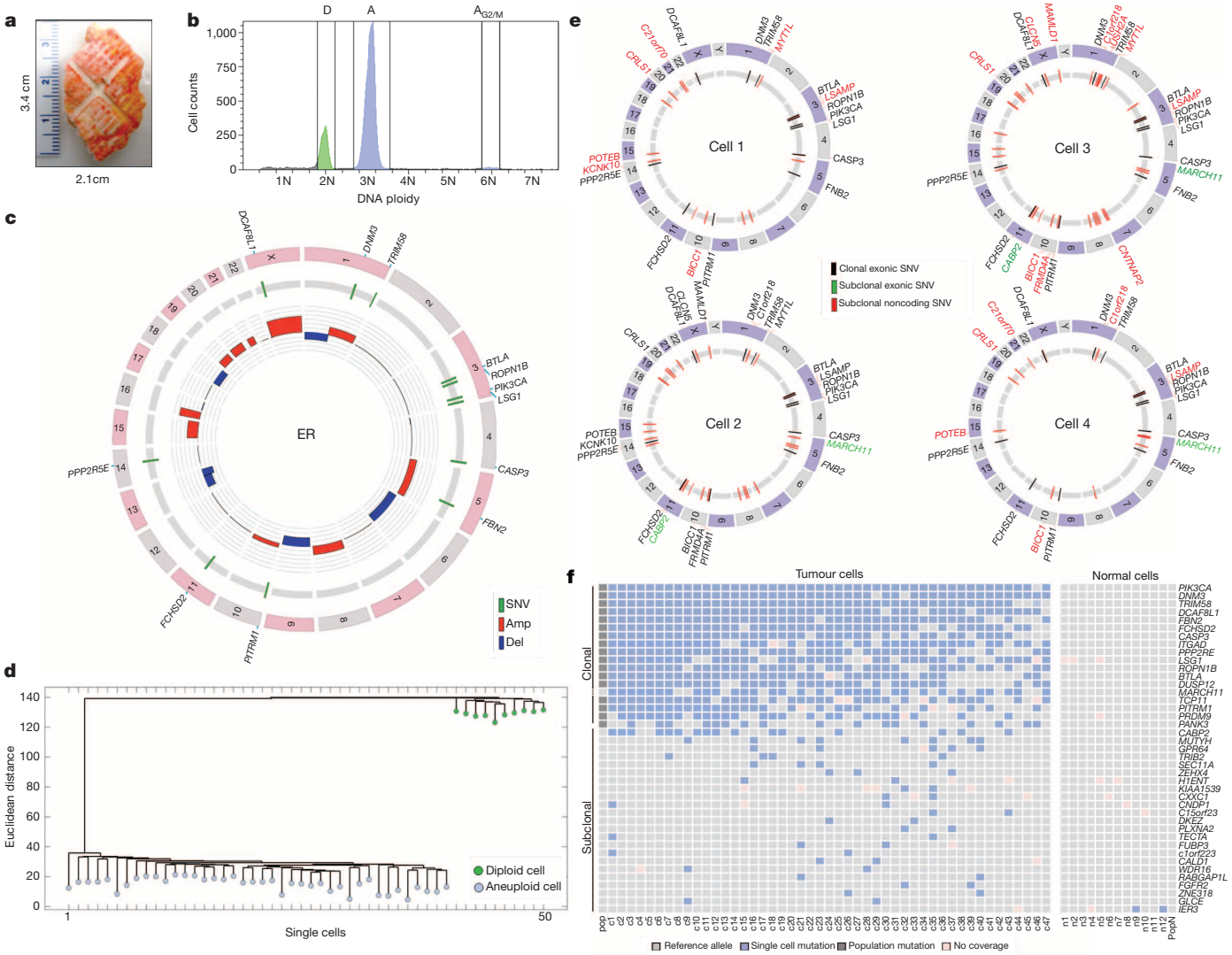


Figure 2 | Single cell and population sequencing of an ER tumour. **a**, Frozen ER tumour specimen. **b**, Flow-sorting histogram of ploidy distributions. **c**, Circos plot of mutations and CNAs detected in the population of aneuploid tumour cells. Cancer genes are on the outer ring. **d**, Neighbour-joining tree of integer copy number profiles from single diploid and aneuploid cells, rooted by the diploid node. **e**, Circos plots of whole-genome single cell sequencing data showing mutations detected in two or more cells. **f**, Heatmap of coding mutations detected by single-nuclei exome sequencing. Mutations detected by whole-genome sequencing (pop) and exome sequencing (ex) are also displayed.

MAP3K4, NTRK1, AFF4, CDH6, SETBP1, AKAP9, MAP2K7, ECM2 and *ECM1* (Supplementary Table 6) (Fig. 3b). Many of these mutations were previously reported in the TCGA breast cancer cohort[7]. Pathway analysis revealed two major pathways that were disrupted during tumour evolution: TGF-β ($P = 9.9 \times 10^{-2}$) and extracellular matrix-receptor signalling ($P = 2.7 \times 10^{-2}$). Copy number profiling identified many chromosomal deletions, in addition to a focal amplification on chromosome 19p13.2 (Fig. 3b).

To investigate genomic diversity at single cell resolution, we performed copy number profiling and exome sequencing. We flow-sorted 50 single nuclei from the hypodiploid (H), diploid (D) and aneuploid (A) ploidy distributions for copy number profiling using SNS (Fig. 3c). Neighbour-joining revealed two distinct subpopulations of tumour cells (A and H) in addition to the normal diploid cells (Fig. 3d). The single cell copy number profiles were analysed using clustered heatmaps, which showed highly similar rearrangements within each subpopulation (A mean $R^2 = 0.91$, H mean $R^2 = 0.88$), but were distinguished by two large deletions on chromosome 9 and 15 (Extended Data Fig. 3b).

Next, we flow-sorted 16 single tumour nuclei from the G2/M peaks (H and A) and 16 single normal nuclei for exome sequencing using nucseq (Fig. 3e). Non-synonymous point mutations were used to perform hierarchical clustering and multi-dimensional scaling (MDS). As expected, the 374 clonal non-synonymous mutations detected by bulk sequencing were found in the majority of the single tumour cells, however, we also identified 145 additional subclonal non-synonymous mutations that were not detected in the bulk tumour (Supplementary Table 7). MDS identified 4 distinct clusters, corresponding to three tumour subpopulations (H, A_1 and A_2) and the normal cells (Extended Data Fig. 5a). Hierarchical clustering showed that many of the subclonal mutations occurred exclusively in one subpopulation (H, A_1 or A_2) (Fig. 3e). The A_1 subpopulation contained 66 unique subclonal non-synonymous mutations, including *AURKA, SYNE2* and *PPP2R1A*. The A_2 subpopulation contained 52 unique subclonal non-synonymous mutations including *TGFB2* and *CHRM5*. In contrast only two subclonal mutations were shared between the normal cells (Fig. 3e, right panel). Many of the subclonal mutations (23.44%) were predicted to damage protein function by both POLYPHEN[21] and SIFT[22] (Extended Data Fig. 5b).

Single-molecule targeted deep sequencing

To validate the mutations detected by single cell sequencing and determine their frequencies in the bulk tumour, we performed targeted single-molecule deep-sequencing. Duplex libraries were constructed

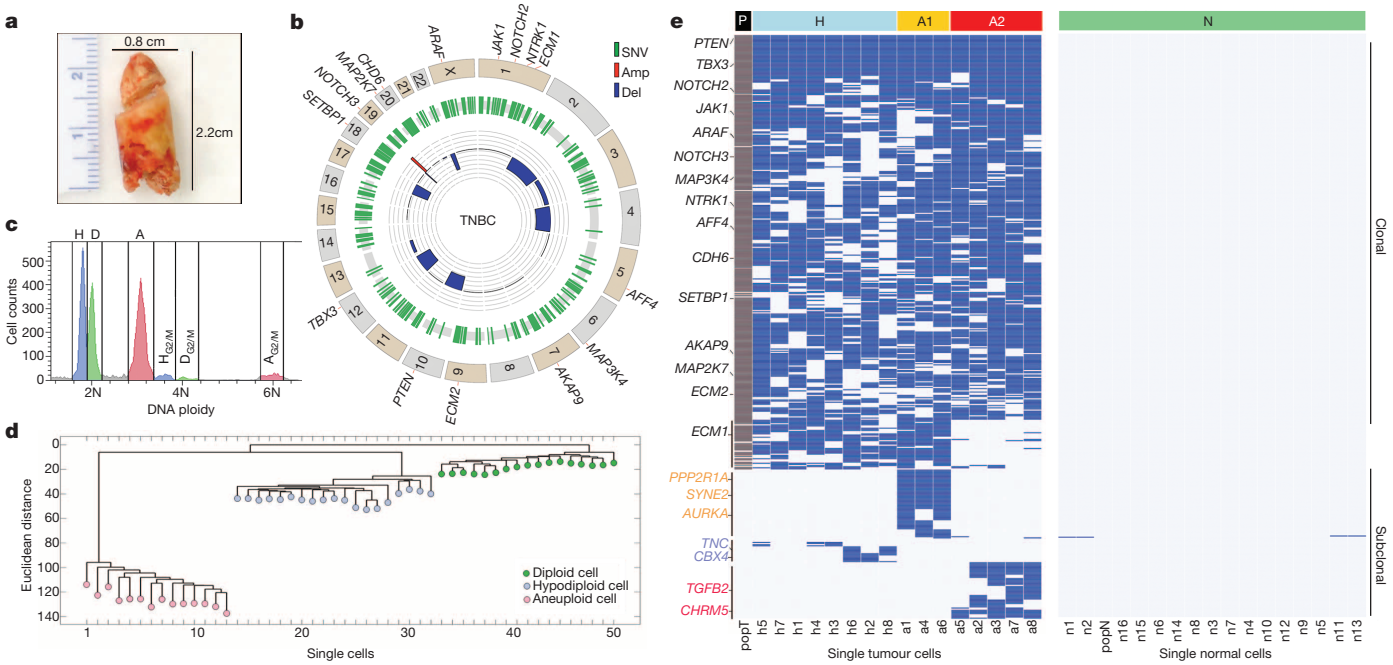


Figure 3 | Single cell and population sequencing of a triple-negative breast cancer. **a**, Frozen TNBC specimen. **b**, Circos plot of mutations and CNAs detected by population sequencing of the TNBC, with cancer genes on the outer ring. **c**, Flow-sorting histogram of ploidy distributions, showing three major subpopulations: diploid (D), hypodiploid (H) and aneuploid (A). **d**, Neighbour-joining tree of 50 single cell integer copy number profiles, rooted by the diploid node. **e**, Clustered heatmap of the nonsynonymous point mutations detected by single nuclei exome sequencing and population sequencing (P). Mutations detected in one cell are excluded.

from bulk tissue to reduce the error rate of next-generation sequencing[10]. Custom capture platforms were designed to target mutations detected in the single cells of the ERBC and TNBC tumours (Methods). Targeted deep-sequencing (116,952×) was performed in the ER tumour resulting in a single-molecule coverage depth of 5,695× using single-strand consensus sequences (SSCS). Deep-sequencing of the TNBC (118,743×) resulted in a single-molecule coverage depth of 6,634× using SSCS (Extended Data Fig. 4). We found that 61.5% of the reads were in the target regions in the ERBC and 80.2% in the TNBC.

The ERBC duplex data validated 94.44% (17/18) of the clonal mutations, 90.47% (19/21) of the subclonal mutations, and 19.40% (26/134) of the *de novo* mutations detected by single cell sequencing ($P < 0.01$) (Methods). The clonal mutations occurred at high frequencies in the tumour mass, whereas the subclonal mutations (0.0895 mean) and *de novo* mutations (0.0195 mean) were very rare (Fig. 4a). Similarly, in the TNBC we validated 99.73% (374/375) of the clonal mutations, 64.83% (94/145) of the subclonal mutations and 26.99% (152/563) of the *de novo* mutations ($P < 0.01$) (Methods). Similarly, we found that the clonal mutations in the TNBC showed high frequencies (0.4457 mean), however, the subclonal mutations were less prevalent (0.050 mean) and the *de novo* mutations were very rare (0.00047 mean) (Fig. 4b). This data suggests that many of the subclonal and *de novo* mutations are likely to be real biological variants that occur at low frequencies in the tumour mass.

Mathematical modelling of the mutation rates

To estimate the mutation rates in each tumour, we used the single cell mutation frequencies and designed a mathematical stochastic birth-and-death process model that uses experimentally derived parameters for cell birth rates (Ki-67 staining), cell death rates (caspase-3 staining), total tumour cell numbers (flow-sorting cell counts) and the tumour mass doubling time for invasive carcinomas (mean = 168 days)[23–25] (Methods). We modelled data for a series of mutation rates and compared the data to the empirical single cell mutation frequency distributions (Supplementary Table 8). Our data suggest that the ERBC had a mutation rate of $M_R = 0.6$ mutations per cell division for the exome data (Fig. 4c) and $M_R = 0.9$ for the single cell whole-genome data (Fig. 4d). These data are similar to the error rates reported for normal cells, which are approximately 0.6 mutations per cell division (error rate = 1×10^{-10})[26–28]. In contrast, our modelling suggests a mutation rate of $M_R = 8$ for the TNBC, suggesting a 13.3× fold increase relative to normal cells (Fig. 4e).

Discussion

In this study we report the development of a novel single cell genome sequencing method that utilizes G2/M nuclei to achieve high-coverage data with low error rates. Although G2/M nuclei were used in this study, the experimental protocol can also be used to sequence nuclei at any stage of the cell cycle. We applied nuc-seq to delineate clonal diversity and investigate mutational evolution in two breast cancer patients. Our data clearly show that no two single tumour cells are genetically identical, calling into question the strict definition of a clone. In both patients we observed a large number of subclonal and *de novo* mutations. These data suggest that point mutations evolved gradually over long periods of time, generating extensive clonal diversity (Fig. 4f, g). In contrast, the single cell copy number profiles were highly similar, suggesting that chromosome rearrangements occurred early, in punctuated bursts of evolution, followed by stable clonal expansions to form the tumour mass (Fig. 4h, i).

We previously reported punctuated copy number evolution by sequencing single cells from a TNBC patient[11]. This model has also been supported by bulk sequencing data in prostate cancer[29] and in rearrangement patterns called firestorms[30] or chromothripsis[31]. A punctuated model is consistent with the mechanisms that underlie CNAs, including chromosome missegregation[32], cytokinesis defects and breakage-fusion-bridge[33], which can generate complex rearrangements in just a few cell divisions. In contrast, point mutations occur through defects in DNA repair or replication machinery[34], which accumulate more gradually over many cell divisions. Our data are consistent with these mechanisms, and further show that two distinct molecular clocks were operating at different stages of tumour growth (Extended Data Fig. 6).

A pervasive problem in the field of single cell genomics is the inability to validate mutations that are detected in single cells. To address this problem, we combined single cell sequencing with targeted single-molecule

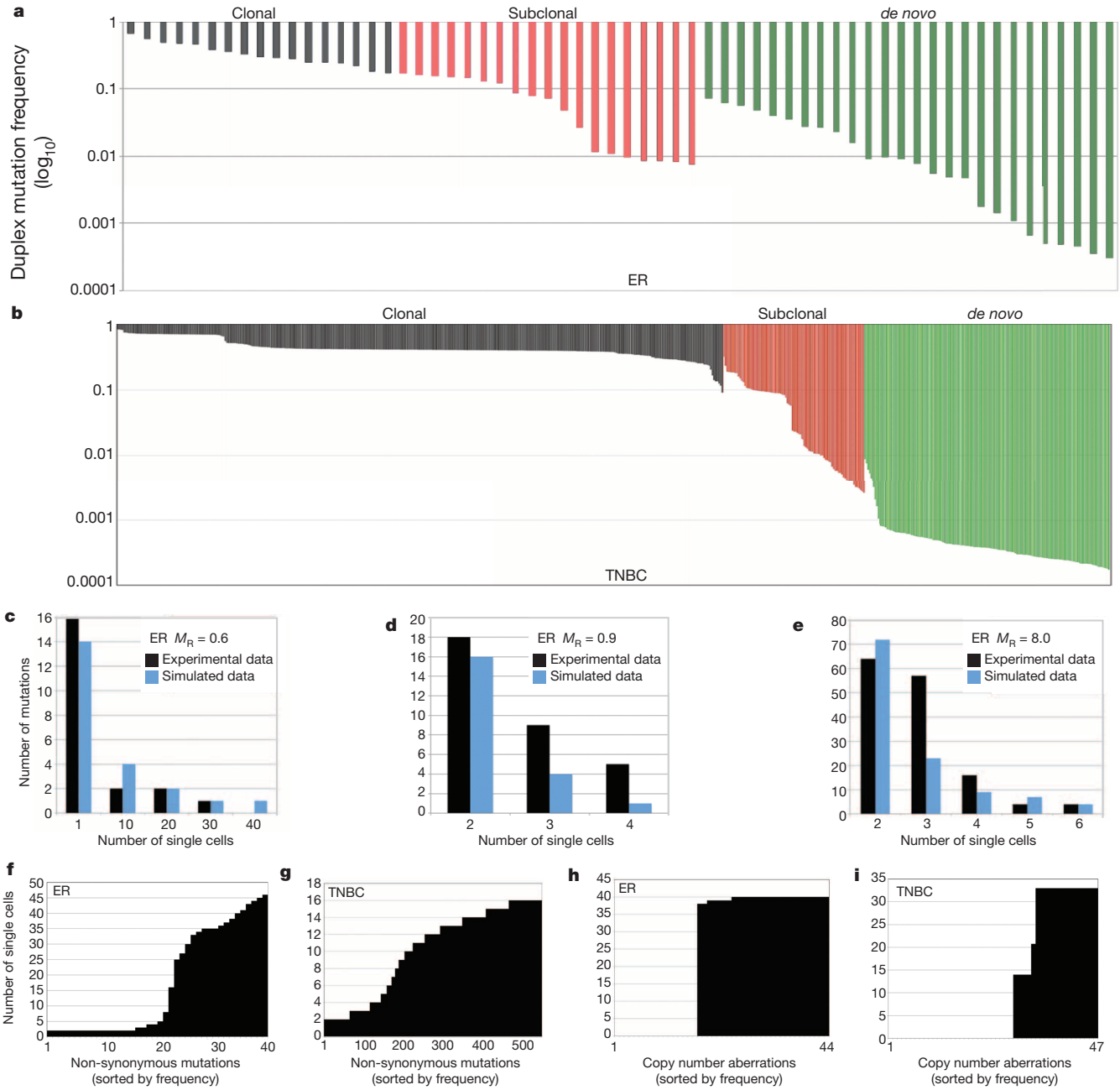


Figure 4 | Duplex mutation frequencies and mutation rates. **a**, ERBC duplex mutation frequencies from targeted deep-sequencing of the bulk tumour tissue. **b**, TNBC duplex mutation frequencies from deep-sequencing of the bulk tumour tissue. **c**–**e**, Mathematical modelling of mutation rates compared to experimental data. **c**, ERBC single-nuclei exome and modelling data at 0.6 mutation rate. **d**, ERBC whole-genome single nuclei and modelling data at 0.9 mutation rate. **e**, TNBC single nuclei exome and modelling data at a mutation rate of 8. **f**, Mutation frequencies shared by 2 or more cells in the ERBC. **g**, Mutation frequencies shared by 2 or more cells in the TNBC. **h**, CNAs shared by two or more cells in the ERBC. **i**, CNAs shared by two or more cells in the TNBC.

deep-sequencing. This approach not only validates mutations, but also measures the precise mutation frequencies in the bulk population. Using this approach, we identified hundreds of subclonal and *de novo* mutations that were present at low frequencies (<10%) in the tumour mass. These rare mutations may have an important role in diversifying the phenotypes of cancer cells, allowing them to survive selective pressures in the tumour microenvironment, including the immune system, hypoxia and chemotherapy[35,36].

A salient question in the field of chemotherapy is whether resistance mutations are pre-existing in rare cells in the tumour, or alternatively, emerge spontaneously in response to being challenged by the therapeutic agent. Although this question has been studied for decades in bacteria[37], it remains poorly understood in human cancers. Our data suggest that a large number of diverse mutations are likely to be pre-existing in the tumour mass before chemotherapy. Our data also has important implications for the mutator phenotype, which posits that tumour evolution is driven by increased mutation rates[34,38]. Although TCGA studies[39–41] report increased mutation frequencies, it remains unclear whether these mutations accumulate over many cell divisions (at a normal error rate) or through an increased mutation rate. Our TNBC data suggest an increased mutation rate (13.3×) relative to the normal cells, supporting this model.

We expect that single cell genome sequencing will open up new avenues of investigation in many diverse fields of biology. In cancer research there will be immediate applications for studying cancer stem cells and circulating tumour cells. In the clinic, these tools will have important

applications in early detection and non-invasive monitoring. Beyond cancer, these tools will have utility in microbiology, development, immunology and neuroscience and will lead to substantial improvements in our fundamental understanding of human diseases.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 July 2012; accepted 23 June 2014.
Published online 30 July 2014.

1. Torres, L. *et al.* Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast Cancer Res. Treat.* **102,** 143–155 (2007).
2. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20,** 68–80 (2010).
3. Park, S. Y., Gonen, 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).
4. Sørlie, T. *et al.* Gene expression patterns of carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98,** 10869–10874 (2001).
5. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (2012).
6. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).
7. The Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
8. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
9. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486,** 353–360 (2012).
10. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109,** 14508–14513 (2012).
11. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).
12. Woyke, T. *et al.* One bacterial cell, one complete genome. *PLoS ONE* **5,** e10314 (2010).
13. Dichosa, A. E. *et al.* Artificial polyploidy improves bacterial single cell genome recovery. *PLoS ONE* **7,** e37387 (2012).
14. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a *JAK2*-negative myeloproliferative neoplasm. *Cell* **148,** 873–885 (2012).
15. Klein, C. A. *et al.* Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. *Proc. Natl Acad. Sci. USA* **96,** 4494–4499 (1999).
16. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* **11,** R119 (2010).
17. Kytola, S. *et al.* Chromosomal alterations in 15 breast cancer cell lines by comparative genomic hybridization and spectral karyotyping. *Genes Chromosomes Cancer* **28,** 308–317 (2000).
18. Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nature Protocols* **7,** 1024–1041 (2012).
19. 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).
20. Lorenz, M. O. Methods of measuring the concentration of wealth. *J. Am. Stat. Assoc.* **9,** 209–219 (1905).
21. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
22. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).
23. Kuroishi, T. *et al.* Tumor growth rate and prognosis of breast cancer mainly detected by mass screening. *Jpn. J. Cancer Res.* **81,** 454–462 (1990).
24. Peer, P. G., van Dijck, J. A., Hendriks, J. H., Holland, R. & Verbeek, A. L. Age-dependent growth rate of primary breast cancer. *Cancer* **71,** 3547–3551 (1993).
25. Michaelson, J. *et al.* Estimates of breast cancer growth rate and sojourn time from screening database information. *J. Women's Imaging* **5,** 11–19 (2003).
26. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
27. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* **148,** 1667–1686 (1998).
28. Preston, B. D., Albertson, T. M. & Herr, A. J. DNA replication fidelity and cancer. *Semin. Cancer Biol.* **20,** 281–293 (2010).
29. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153,** 666–677 (2013).
30. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16,** 1465–1479 (2006).
31. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
32. Pellman, D. Cell biology: aneuploidy and cancer. *Nature* **446,** 38–39 (2007).
33. McClintock, B. The stability of broken ends of chromosomes in *Zea mays*. *Genetics* **26,** 234–282 (1941).
34. Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Rev. Cancer* **11,** 450–457 (2011).
35. Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Rev. Cancer* **6,** 924–935 (2006).
36. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).
37. Luria, S. E. & Delbruck, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28,** 491–511 (1943).
38. Bielas, J. H., Loeb, K. R., Rubin, B. P., True, L. D. & Loeb, L. A. Human cancers express a mutator phenotype. *Proc. Natl Acad. Sci. USA* **103,** 18238–18242 (2006).
39. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
40. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
41. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Ramagli, H. Tang, E. Thompson, K. Khanna, W. Schober and J. Tyler. We are grateful to S. Kennedy and L. Loeb for help with the duplex protocols. We thank M. Edgerton, J. Hicks, M. Wigler and J. Kendall for discussions. We thank R. Krahe and M. Rui for reviewing the manuscript. N.E.N. is a Nadia's Gift Foundation Damon Runyon-Rachleff Innovator (DRR-25-13). This research was supported by grants to N.E.N. from NIH (R21CA174397-01) and NCI (1RO1CA169244-01). N.E.N. was supported by T.C. Hsu and the Alice-Reynolds Kleberg Foundation. N.E.N. and P.S. were supported by the Center for Genetics & Genomics. F.M.-B was supported by an NIH UL1 (TR000371) and Susan Komen (SAC10006). K.C. was supported by the NCI (RO1CA172652). H.L. was supported by the NIH (U24CA143883). F.M. was supported by PS-OC (U54CA143798). K.C. and H.L. were supported by the Dell Foundation. M.L.L. is a CPRIT scholar and is supported by ALA. This work was also supported by an NCI center grant (CA016672). A.U. is a Rosalie B. Hite Fellow.

Author Contributions Y.W. performed experiments and data analysis. M.L.L., J.W., A.M. and X.S. performed experiments. A.U., W.R., K.C., H.L., P.S. and S.V. performed data and statistical analyses. H.Z. and F.M.-B. obtained clinical samples. R.Z. and F.M. performed modelling. N.E.N. performed experiments, analysed data and wrote the manuscript.

Author Information The data from this study has been deposited into the Sequence Read Archive (SRA053195). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.E.N. (nnavin@mdanderson.org).

METHODS

Tumour and cell line samples. SK-BR-3 is a Her2 positive ($ER^-/PR^-/Her2^+$) breast cancer cell line that was previously used for single cell copy number profiling[11]. The oestrogen receptor positive breast cancer (ERBC) and triple-negative breast tumour (TNBC) samples used in this study were obtained from the MD Anderson Cancer Center Breast Tissue bank as frozen tumour specimens. Histopathology classified both breast tumours as invasive ductal carcinomas. The ERBC was also reported to have mixed invasive lobular carcinoma. Both tumours were excised by lumpectomy before any chemotherapy or radiation therapy. The ERBC tumour grade was scored as Nottingham histological grade 2, whereas the TNBC tumour was scored as grade 3. Receptor staining showed that the ER tumour was positive for oestrogen receptor (80%), positive for progesterone receptor (90%) and negative for the Her2 receptor (FISH Her2/CEP17, ratio 1.1). The TNBC was negative for oestrogen receptor (2%), negative for progesterone receptor (3%) and negative for the Her2 receptor (FISH Her2/CEP17, ratio 1.3). This study was approved by the Internal Review Board (IRB) at MD Anderson Cancer Center.

Spectral karyotyping. Exponentially growing SK-BR-3 cells were exposed to Colcemid (0.04 μg ml^{-1}) for one hour at 37 °C and to hypotonic treatment (0.075 M KCl) for 20 min at room temperature. Cells were fixed in a methanol and acetic acid (3:1 by volume) mixture for 15 min and washed three times in the fixative. Slides were prepared by dropping the cell suspension on wet slides and air drying. SKY was performed according to the manufacturer's protocol using Human Paint probes (ASI, Vista, CA). Images were captured using Nikon 80i microscope equipped with Spectral Karyotyping software from ASI, Vista, CA. 100 metaphases from each sample were analysed in detail.

Isolation of single nuclei by flow-sorting. Nuclei of cell lines and frozen tumours 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, 0.2% Nonidet P-40)), 200 ml of 106 mM $MgCl_2$, 10 mg of DAPI and 0.1% DNase-free RNase A. Cultured cells were trypsinized and lysed directly in NST/DAPI buffer. Sectioned tumours were cut and minced using surgical blades in a Petri dish in NST/DAPI buffer in the dark. Samples were filtered through a 37-μm plastic mesh to a 5-ml polystyerene tube. Nuclei were then sorted using FACS Aria II (BD Biosciences) and single nuclei were deposited into individual wells on a 96-well plate. Single nuclei were gated from the G2/M distribution of cells.

Limited multiple-displacement-amplification. Multiple-displacement-amplification was performed on individual sorted nuclei using the REPLI-G UltraFast Mini Kit (Qiagen, #150035). The protocol was modified by heating the lysed DNA at 65 °C for 10 min and incubating the DNA with the ϕ29 polymerase at 30 °C for exactly 80 min. DNA was purified using the QIAamp DNA blood mini kit and quantified using the Qubit 2.0 fluorometer (Invitrogen, Q32866).

Nextera library preparation. The WGA DNA was incubated with the Nextera transposome (Epicentre, Inc.) to perform a tagmentation reaction in HMW buffer according to manufacturer's instruction. The libraries were purified using MinElute PCR purification kits (Qiagen, #28106), followed by 4 cycles of PCR. After PCR, the libraries were run on 2% agarose gels and size-selected in the 200–300 bp range (SK-BR-3) or 400–500 bp range (ER, TNBC). The excised gel blocks were purified using MinElute purification columns (#28606). The size distribution and concentration of the libraries were determined using the Bioanalyzer 2100 system (Agilent) using high sensitivity DNA microcapillary chips. The final concentration of the library was determined using quantitative PCR with the KAPA Library Quantification Kit (KAPA Biosystems, KK4835) and fluorescence was measured using the Qubit 2.0 system (Invitrogen, Q32866).

NEB library preparation. To prepare sequencing libraries by ligation cloning, 100 ng to 1μg of DNA was acoustically sonicated to 300 bp or 500 bp using the Covaris Sonicator S220. Libraries were constructed using NEBNext DNA library Prep Master Mix Set for Illumina (New England Laboratory, #F6040L) for end-repair, 3′ adenylation and ligation according to the manufacturer's instructions. MinElute PCR Purification Kit (Qiagen, #28006) is used for the purification step during library prep. Agarose electrophoresis is run for excision at 300 or 400 bp for size selection. We then performed 8 cycles of PCR following the manufacturer's instructions, using PE5/7 primers (Illumina Inc.). Agencourt AMPure XP (Beckman Coulter, #A63881) was used for final purification. Final concentration was measured by quantitative PCR using KAPA Library Quantification Kit (KAPA Biosystems, KK4835) and ABI PRISM real-time machine (Applied Biosystem 7900HT), as well as 2100 Bioanalyzer (Agilent).

Chromosome PCR panel. To evaluate the single cell WGA amplification efficiency we designed 22 pairs of primers (Sigma Aldrich) to target 22 loci on different chromosomes for PCR or qPCR amplification: chr1F (TATGGCTGCCCACTCCTTAG); chr1R(GACCTCGGCCTGGACTACTA); chr2F (CTGGGGCTCTTCAAACTGAG); chr2R (GGTGGCCGTAGTGGTAGATG); chr3F (CTTGTGGGTGTGGTCAGTTG); chr3R (CAGTACAAGGGTGGGAGGAA); chr4F (GTCAGAGGGTGAGGGCAGTA); chr4R (TCAAAATAATGGGCCTGGAA); chr5F (GGGGGACAGGACCAGTTATT); chr5R (TCAAAAGAAGTGGGAGGATTG); chr6F (CACCACTCCACAGGGAGAAT); chr6R (CAGAGACCAAGGGAGAAACG); chr7F (TCGTCTACCTCCTCCCTCCT); chr7R (GGACACGCAGTGCTCATAGA); chr8F (GGGTTTTGGTGTTGAGAAAAA); chr8R (GGAGGAGCAAGTTGATTGGTT); chr9F (CCACCTGCAAAGGGACATAC); chr9R (AGCAAGGAGTTGCCAGGTTA); chr10F (ACTTGCAGACCACTGGGATT); chr10R(GAGAGCATTGGCCTCCATAG); chr11F (GATGCAGGGAGGGTATGTGT); chr11R (CCTTGCCAGTAGGTTTCCTG); chr12F(ACCCTTCCACTGGACCTCTT); chr12R(CATTTGCTGCCTCACTTGTC); chr13F(TCTCCAGTTTTGAGGGGCTA); chr13R(TTGGCCTCCACTTCATTTCT); chr14F (GGATGGAACTAGCCATGCAG); chr14R (GGAGGATCACTGCACACCTT); chr15F (GCGAAAGGAGCTGAAAACAC); chr15R(TTGACTTCCTCCCTTCTCCA); chr16F (ATGGCCAATAGAATGCCAAA); chr16R (AAATTGCCTGAAACCCAGTC); chr17F (ATAGCCACACCTCCTGCATC); chr17R(CCCCGGAATAGACCACTTTC); chr18F (TACTACAGGCCAGCCATTTTG); chr18R (CTTGTGCTGCTGCTGGTGTA); chr19F (ATGTGAGACGTCATGGGTGT); chr19R (GGGCGTCTAGGAGCACTG); chr20F (CCCAAAGAAACAAGGGGAGT); chr20R (AAGCCTACAGCTGGGACTGA); chr21F(CCATGACTGGAATGACGATG); chr21R(CTTCCCAAAGAATGCCAAAC); chr22F (GCACCATTCAACCAATCTGA); chr22R(TGCCATTCCCTCTAATCCTG). The WGA amplification time was extended to generate 100 ng of DNA before PCR amplification. We used 1 ng of DNA for each PCR amplification reaction using the KAPA Taq PCR kit (Kapa #BK1001). The PCR conditions used are: 95 °C for 30 s, followed by 30 cycles (95 °C for 30 s, 60 °C for 60 s and 68 °C for 60 s) and a final extension at 68 °C for 5 min. DNA was separated on a 2% agarose gel.

Immunohistochemistry. For immunohistochemistry analysis frozen tissue sections (6 μm) were fixed in methanol, and allowed to air dry. Tissue sections were subjected to peroxidase quenching for 5 min, 10 min blocking, antibody incubation, 20 min biotin/streptavidin peroxidase binding, and revealed through a 3 min DAB chromagen detection system, according to standard Invitrogen protocol (Invitrogen, Frederick, MD). For cleaved caspase-3, a rabbit polyclonal anti-caspase3 antibody (BioCare Medical, Concord, CA) was diluted 1:200 in diluent (DAKO, Los Angeles, CA), and incubated for 1 h at room temperature. The rabbit monoclonal anti-Ki67 (Abcam, Cambridge, MA) was used to stain for Ki-67. The Ki-67 antibody was diluted 1:400 and allowed to incubate for 1 h at room temperature. Tissue samples were counterstained using haematoxylin nuclear counterstain, by applying haematoxylin for 2.5 min, dipping 10 times in acid rinse (2 ml glacial acetic acid + 98 ml of diH20), and incubating in bluing solution (1.5 ml NH_4OH (30%) stock + 98.5 ml of 70% EtOH) for 1 min.

Single cell exome capture. Exome capture was performed on single cell sequencing libraries using the TruSeq Exome Enrichment Kit (Illumina,15013230) following manufacturer's instructions with one modification: Nextera PCR primers (Epicentre) are used in place of the TruSeq PCR primers for library amplification. The capture platform targeted a 64 Mb region including exons, promoters and UTRs. Final samples were purified using the AMPpure XP beads (Beckman Coulter, #A63881).

Next-generation Illumina sequencing. We first performed pre-sequencing runs of the single cell libraries at low-coverage depth (1×). Libraries were multiplexed and run at 100 single-end cycles on the Illumina HiSeq2000 system. The pre-sequencing data was aligned to the human genome (HG18) to determine the % PCR duplicates and % reads mapping uniquely. Libraries that showed >50% coverage, > 60% reads mapping and <40% PCR duplicates were selected for full genome or exome sequencing. Nextera libraries were sequenced using Epicentre Sequencing primers (Epicentre, Inc.). NEB libraries were sequenced using TruSeq V2 Sequencing primers (Illumina Inc.). Data was processed using the CASAVA 1.8.1 pipeline (Illumina Inc.) and sequence reads were converted to FASTQ files.

Duplex targeted ultra-deep sequencing. Duplex sequencing libraries were prepared from frozen bulk tumour tissues using the experimental protocol described by Schmitt *et al.* 2012. We isolated genomic DNA from bulk tumour tissues using the DNAeasy Blood & Tissue Kit (Qiagen, cat #69504). The DNA concentration was quantified using the QuBit DNA fluorometer (Life Technologies) and 1 microgram of DNA was used as input material for each Duplex library construction, and 4 libraries were constructed in parallel. To generate duplex libraries we synthesized the following adapters at 100 micromolar scale with HPLC purification (Integrated DNA Technologies): DX1 –AATGATACGGCGACCACCGAATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT; DX2 5-phos-ACTGNNNNNNNNNNNNAGATCGGAAGAGACACACGTCTGAACTCCAGTCAC. We generated double-stranded adapters by diluting the oligonucleotides to 100 μM and combining 10 μl of DX1 and DX2 together for hybridization. The solution was heated to 95 °C for 5 min and cooled to room temperature for 1 h. Magnetic beads (Agencourt AMPure XP, Beckman Coulter, #A63881) were used in all purification steps to recover optimal concentrations of DNA. Genomic DNA was quantified using a fluorometer (Qubit, Life Sciences) after acoustic sonication (Covaris) at 400 bp and the size distribution was determined using microchip

capillary electrophoresis (Bioanalyzer, Agilent) using the High Sensitivity DNA microchip. Exactly 1 microgram of DNA was used as input material for each reaction and 4 libraries were performed in parallel. Following the TA cloning procedure, we quantified the ligated products before PCR amplification, which was performed at 13 cycles. This step is critical, as PCR will amplify the unique duplex tags before sequencing and it is necessary to amplify 10–20 duplicate read tags from each original molecule. After PCR enrichment, we measure the concentration of the libraries using the QuBit fluorometer (QuBit) and qPCR (Applied Biosciences), which resulted in a final concentration of approximately 500 ng per reaction. We measured the duplex library insert size by microcapillary gel electrophoresis on the Bioanalyzer system using the 'high-sensitivity' DNA chips (Agilent). We then pooled together 4 separate duplex libraries to generate approximately 2 micrograms of DNA as input material for the custom capture reaction. The custom capture platforms (Nimblegen, Roche) were designed to target regions containing mutations that were identified from the single cell sequencing data. In the ER tumour we synthesized probes to target 173 regions of 200 bp in length. In the TNBC tumour we synthesized probes to target 1,083 regions of 200 bp in length. Hybrid capture was performed following manufacturer's instructions (Nimblegen, Roche, SeqCap EZ Choice Protocol) with 8 cycles of final PCR amplification. The duplex libraries were sequenced at 100 cycles using paired-end reads on the HiSeq2000 system (Illumina) to generate approximately 100,000× target coverage depth. The duplex sequencing data was processed and analysed as described in the section: Analysis of duplex sequencing data.

Sequence alignment and processing. Image processing and base calling was performed using the CASAVA 1.8.1 pipeline (Illumina, Inc.). Sequence reads in FASTQ format were mapped to the human assembly US National Center for Biotechnology Information (NCBI) build 36 (hg18) using the Burrows-Wheeler alignment tool[42] (BWA version 0.6.0) with default parameters and sampe option to create SAM files with correct mate pair information, and read group tag that includes sample name. Samtools (0.1.16) was used to convert SAM files to compressed BAM files and sort the BAM files by chromosome coordinates[43]. The Genome Analysis Toolkit[44] (GATK v1.4-37) was used to locally realign the BAM files at intervals that have indel mismatches before PCR duplicate marking with Picard (version 1.56) (http://picard.sourceforge.net/). Reads with mapping quality score less than 40 were removed from the BAM files.

Single nucleotide variant detection. The GATK UnifiedGenotyper was used to detect single nucleotide variants (SNVs)[44]. All single cells and population samples were processed together to generate a single VCF4 file. Variants detected in the matched normal samples from the ER^+ and TNBC were filtered from the somatic variants to eliminate germline mutations using matched normal tissue samples, the diploid fraction and the normal single cells. We required a minimum base quality (mbq) of 20 for the base to be considered during variant detection. Coverage depth at a given locus of greater than 2,500 reads was down sampled to expedite analysis processing. We then used the GATK variant recalibrator to filter the output at default sensitivity level. Recalibration training databases include hapmap 3.3, dbSNP build 132, Omni 2.5M chip and Mills. Annotations used for training include variant quality score by depth (QD), mapping quality rank sum score, read position rank sum score, mapping quality (MQ), coverage depth (DP) and strand bias (FS). After recalibration, SNVs within 10bp of another SNV or Indel were excluded to avoid false positives caused by misalignment. A minimum coverage depth of 10 and at least 3 variant reads were required for the detection of SNVs. GATK SelectVariants was used to separate SNVs into VCF4 files for downstream annotation.

Structural variant detection in population samples. Structural variants were detected using CREST[45] and filtered using Perl scripts that required a minimum of 3 split-reads to detect an event. In the population sample structural variants were detected including intrachromosomal translocations, interchromosomal translocations, inversions, insertions and deletions. Structural variants were intersected with BED files from the cancer gene census[46] and RefSeq[47] in order to identify rearrangements in normal and cancer genes.

Copy number detection in single cell and population samples. Copy number was detected from sequence read density using the variable binning method[11,18]. Briefly, copy number is calculated from read density by dividing the genome into 'bins' and counting the number of unique reads in each interval. To determine interval sizes we simulated sequence reads by sampling 200 million sequences of length 48 from the human reference genome (HG19/NCBI37) and introduced single nucleotide errors with a frequency encountered during Illumina sequencing. These sequences were mapped back to the human reference genome using BWA and filtered for unique mappings. We assigned a number of bins to each chromosome based on the proportion of simulated reads mapped. We then divided each chromosome into bins with an equal number of simulated reads. This resulted in 12,508 genomic bins with no bins crossing chromosome boundaries. The median genomic length spanned by each bin is 220 kb. This variable binning efficiently reduces false deletion events when compared to uniform length-fixed bins. Large bins were filtered to remove false-positive amplifications in the centromeric and telomeric regions. We then applied Loess normalization to correct for GC bias[18]. The copy number profiles were segmented using the Kolmogorov–Smirnov (KS) statistical test[48].

Databases filtering and annotation. Single nucleotide variants and indels were annotated using Annovar (version 2013 Nov20, Aug 23rd)[49]. We downloaded databases dbSNP build 135, 1000Genomes, ployphen and avsift using the Annovar perl scripts. Results for SK-BR-3 were annotated with dbSNP and 1KG filtering, while variants for BC10 and TNBC were annotated without dbSNP and 1KG filtering as we were able to detect germline mutations for both tumours. Mutations in the COSMIC database were downloaded separately[50] as well as the cancer gene census database[46]. BEDtools (v2.14.2)[51] was used to annotate both COSMIC mutations and cancer genes. A Perl script was developed to run all of the annotation steps automatically and pool annotation results into one final file.

Calculation of coverage uniformity. Lorenz curves were calculated to determine coverage uniformity in the single cell and population samples. Briefly, sequence reads were aligned with BWA using unique mappings and PCR duplicates were removed with Picard. From the BAM files we ran samtools mpileup with the following parameters: "-A –B –d1000000000" to determine the read counts for every base in the human genome reference assembly HG18. The depth values were sorted using Unix sort with "-n" parameter and a custom perl script was used to read the sorted depth values and calculate the cumulative fraction of the genome that was covered and the cumulative fraction of reads. The curves for each cells and population samples were plotted in Matlab (Mathworks).

Calculation of neighbour-joining trees. Exome data from single cells were aligned to HG18 and variants were detected using GATK (as described above). VCF4 files were generated and a binary distance matrix was calculated using point mutations at sites with coverage $\geq 10\times$. The neighbour-joining trees[52] were calculated using Matlab (Mathworks) using one of two distance metrics: Hamming distance or Euclidean distance. The neighbour-joining trees were plotted as linear trees or circular trees, and were re-rooted by the matched normal population sample.

Analysis of duplex sequencing data. The analysis of duplex sequencing data was performed as described in Schmitt *et al.* 2012. We trimmed the 12 nucleotides tags from each paired-end reads, 5 nucleotide anchor sequence following the fixed adaptor sequence and 4 nucleotides after the anchor sequences from all reads using the script "tag_to_header.py". The python script combined the 12 nucleotides tags from both the forward and reverse reads to form a 24 nucleotides combined tag for each molecule. Trimmed sequence reads were then aligned to the human genome assembly HG18 using BWA. The resulting SAM files were converted to BAM files using Samtools and sorted by chromosome position. Unmapped reads were removed using "Samtools view –F4" command. We used a Python script "consensusMaker.py" to organize all reads with identical tags into one group to extract a single strand consensus sequence (SSCS). We used default parameters provided by the script with at least 3 reads required to form an SSCS and at least 70% of nucleotides at a position must be identical to form a consensus nucleotide in the SSCS. Resulting SSCSs were mapped using BWA to HG18. SSCS SAM files were converted to BAM files, which were sorted and merged together. A custom Perl script was used to extract data from base sites that overlap with targeted mutations. The following mpileup parameters were used: " samtools mpileup -A -BQ0 -d1000000000 -q 0 –f. For each nucleotide, we filtered by a minimum base quality of 20 and computed the number of reads supporting the base. Regions with less than 500 molecule depth were excluded from analysis. To determine if a duplex variant is validated we calculate the probability of error using a binomial model (next section).

Calculation of duplex mutation probabilities. We calculate the probability that a variant is an error in the single-molecule duplex sequencing data using a discrete binomial probability distribution model. In this model we calculate the probability that a base is due to chance by considering random errors in alternative bases and multiplying this probability by the probability of errors based on sequence read depth, by incorporating the error rate of duplex sequencing.

$$p(e) = \left[\binom{n}{x} p^x (1-p)^{(n-x)}\right]\left[\binom{k}{n} q^n (1-q)^{(k-n)}\right]$$

In this model p is the probability that a base is the variant base ($P = 0.33$) excluding probability of the reference base. We set x equal to the number of non-reference reads that support the variant allele, and n equal to the total number of non-reference reads. This is the probability that the reference base is not due to chance, which we multiply by the probability that a base occurs due to random chance at a given molecular read depth based on the duplex error rate q. The latter probability q is calculated using error rate of duplex sequencing, the single molecule

depth at each variant site k and the number of non-reference variant reads n. We consider variants with $p(e) < 0.01$ to be validated in our data.

Multi-dimensional-scaling analysis. Non-synonymous mutations were parsed from the VCF4 files containing single cell exome variant data to construct a binary distance matrix at sites where coverage depth was $\geq 6\times$. Distance was calculated using the hamming method and missing values with no coverage were converted to 0. The resulting binary matrix was used to perform multi-dimensional scaling in R (http://www.r-project.org). The MDS coordinates 1 and 2 were plotted against each other to identify clusters of cells with similar mutations.

Calculation of technical error rates. The allelic dropout rate (ADR) is defined as the percent of homozygous sites in the single cell samples (C_i) where the population reference sample (P) is heterozygous at the same nucleotide site. These calculations were made using the SK-BR-3 single cell samples ($n = 2$) and the population sample at nucleotide sites where read depth is $\geq 6\times$ in both samples and bases that have passed variant quality score recalibration[44].

$$ADR = \frac{1}{n}\sum_{i=1}^{n}\frac{C_i}{P}$$

The false positive rate (FPR) is defined as the number of heterozygous sites in the single cell sample (C_i) is divided by the number of sites in the population reference sample (P) that are homozygous for the reference allele at the same nucleotide site. From the single cell samples we subtract the number of validated mutations (v), which are not technical errors.

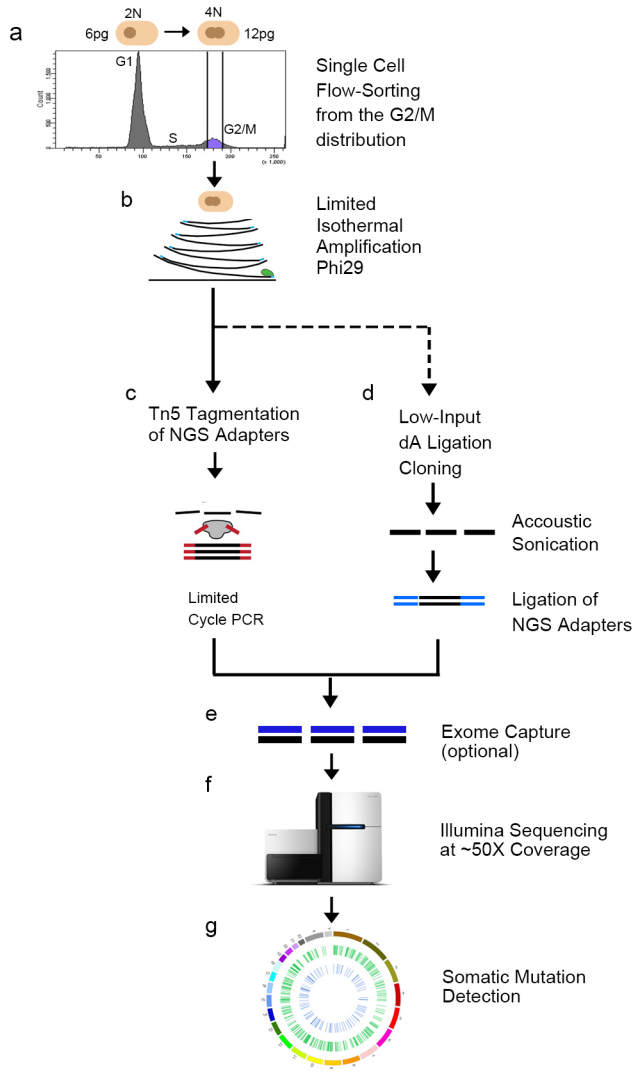
$$FPR = \frac{1}{n}\sum_{i=1}^{n}\frac{(C-v)_i}{P}$$

False negative coverage (FNC) is defined as the percent of bases in the human genome with $\geq 1\times$ unique reads in the population sample (P) minus the mean percent of bases in the single cell samples (S_i) where depth is $\geq 1\times$ using uniquely mapped reads.

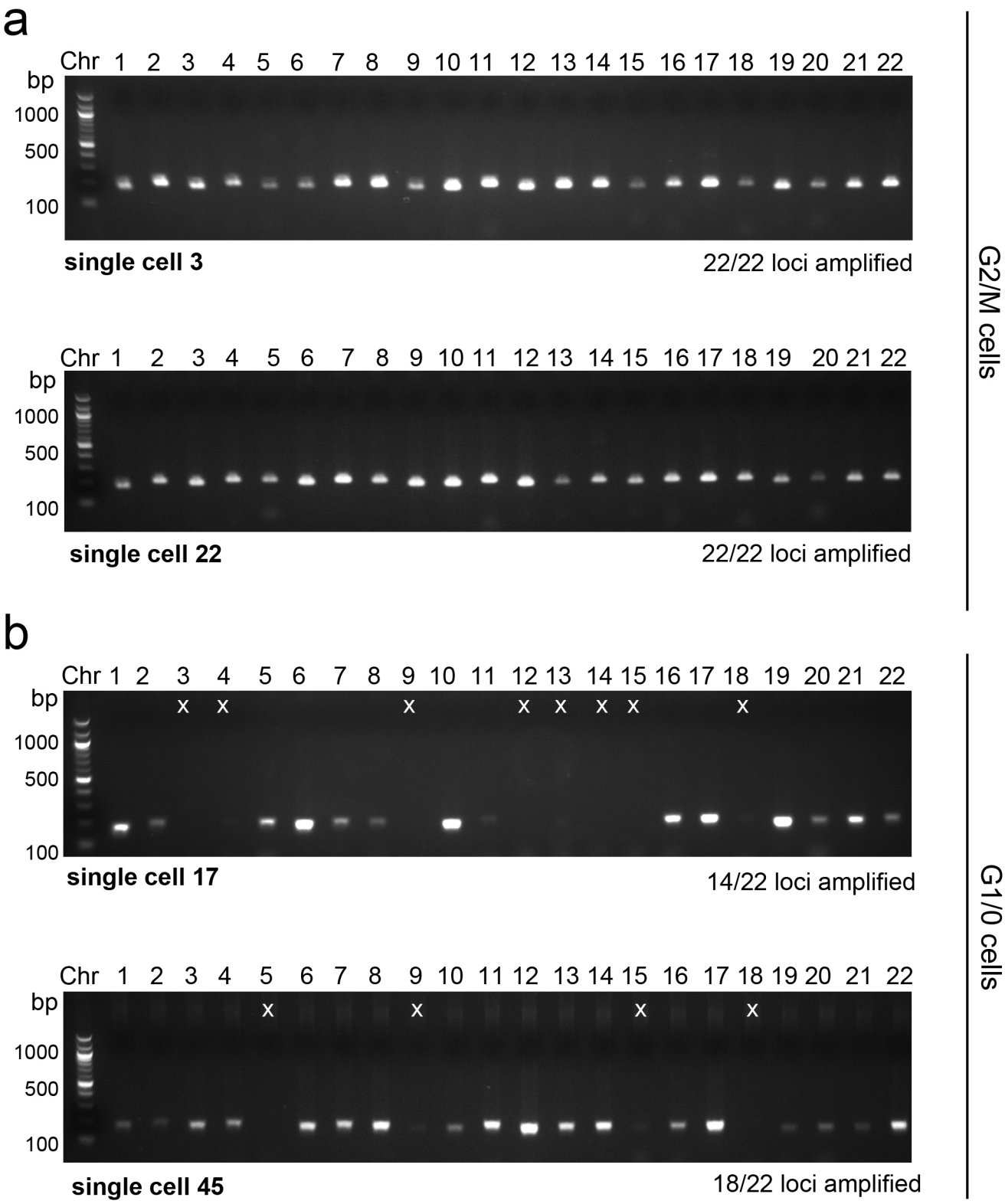
$$FNC = \frac{1}{n}\sum_{i=1}^{n}P - S_i$$

Calculation of mutation rates. Our model follows a branching process. During each elementary time step, birth and death may occur depending on their respective rates; if a cell division takes place, mutations may occur. The number of mutations that occur during each time step follows a Poisson distribution with a mean parameter, which is derived from the observed data for a particular cell type. Each mutation in a cell is assigned a unique identifier and mutations can be passed to new generations. The simulation starts with one cell and terminates when the total cell population reaches the total number in the tumour cell population (flow-sorting tumour cell counts). After completing the simulation, we sample a matching number of cells (matched to the number of cells sequenced) from the millions of tumour cells. For these sampled cells, we tabulate mutations that are shared between at least two single cells. We exclude mutation frequencies that occur in >6 single cells to focus our analysis on random mutations, excluding mutations that are likely to be influenced by positive selection. We repeat this process 1,000 times and average these results. We repeat this modelling for a large series of mutation rates and compare the distribution to the empirical distributions measured from single cell frequencies. We then compute the sum of square difference for each mutation rate, and selecting the distribution with the minimum difference, to determine the mutation rate. The following experimentally derived parameters were used for a series of mutation rates. For ER^+ tumour: total number of tumour cells = 12,451,945 (flow-sorted tumour cell counts); cell birth rate = 0.004654777 (24.2% Ki-67 index); cell death rate = 0.000535032 (1.1% caspase-3 index) and tumour cell doubling time = 168 days. For TNBC: total number of tumour cells = 17,719,218; cell birth rate= 0.051202551 (43.5% Ki-67 index); cell death rate= 0.001229775 (10.4% caspase-3 index) and tumour cell doubling time = 168 days.

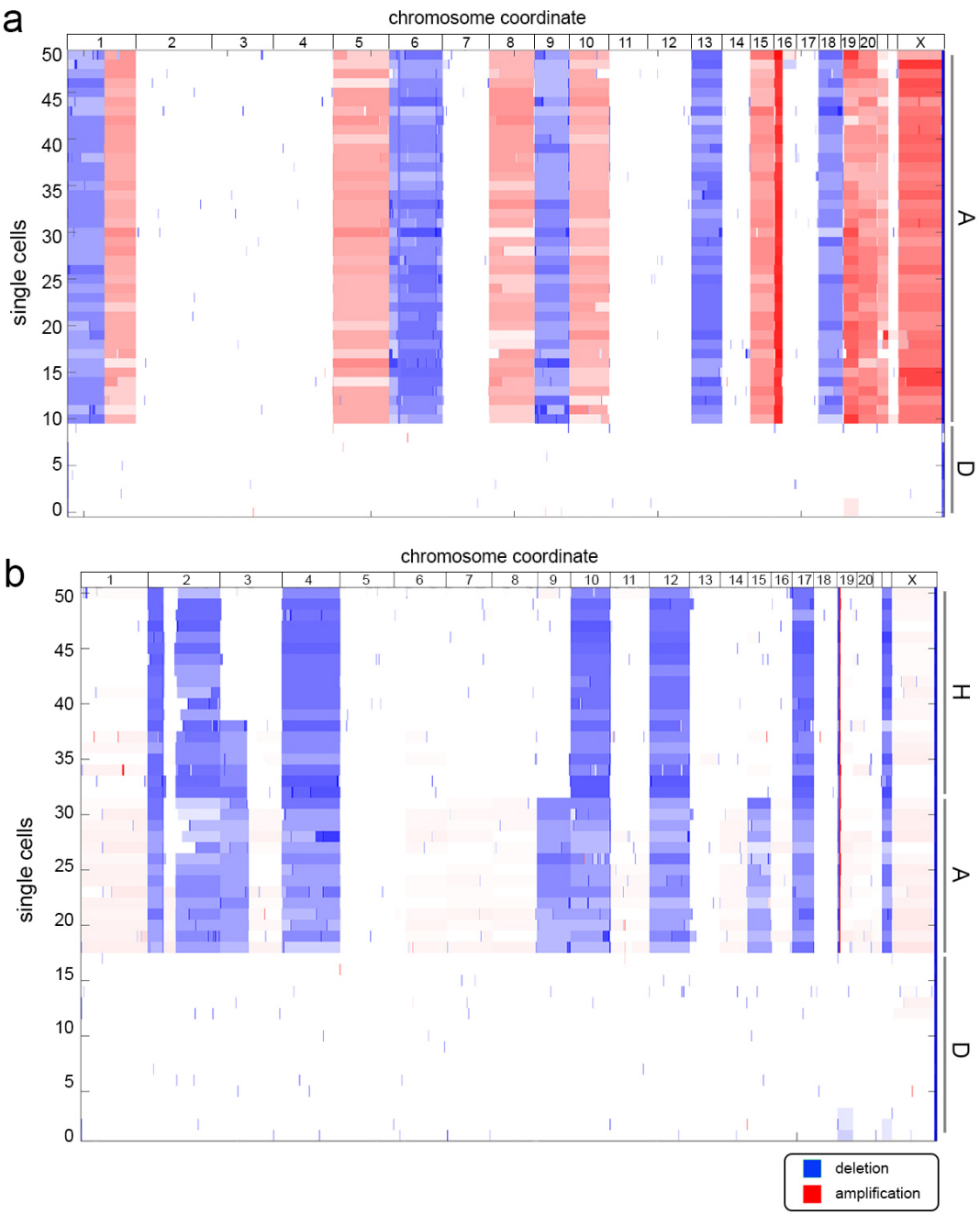
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
43. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
44. McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
45. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature Methods* **8,** 652–654 (2011).
46. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).
47. Hsu, F. *et al.* The UCSC known genes. *Bioinformatics* **22,** 1036–1046 (2006).
48. Grubor, V. *et al.* Novel genomic alterations and clonal evolution in chronic lymphocytic leukemia revealed by representational oligonucleotide microarray analysis (ROMA). *Blood* **113,** 1294–1303 (2009).
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. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **39,** D945–D950 (2011).
51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
52. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).



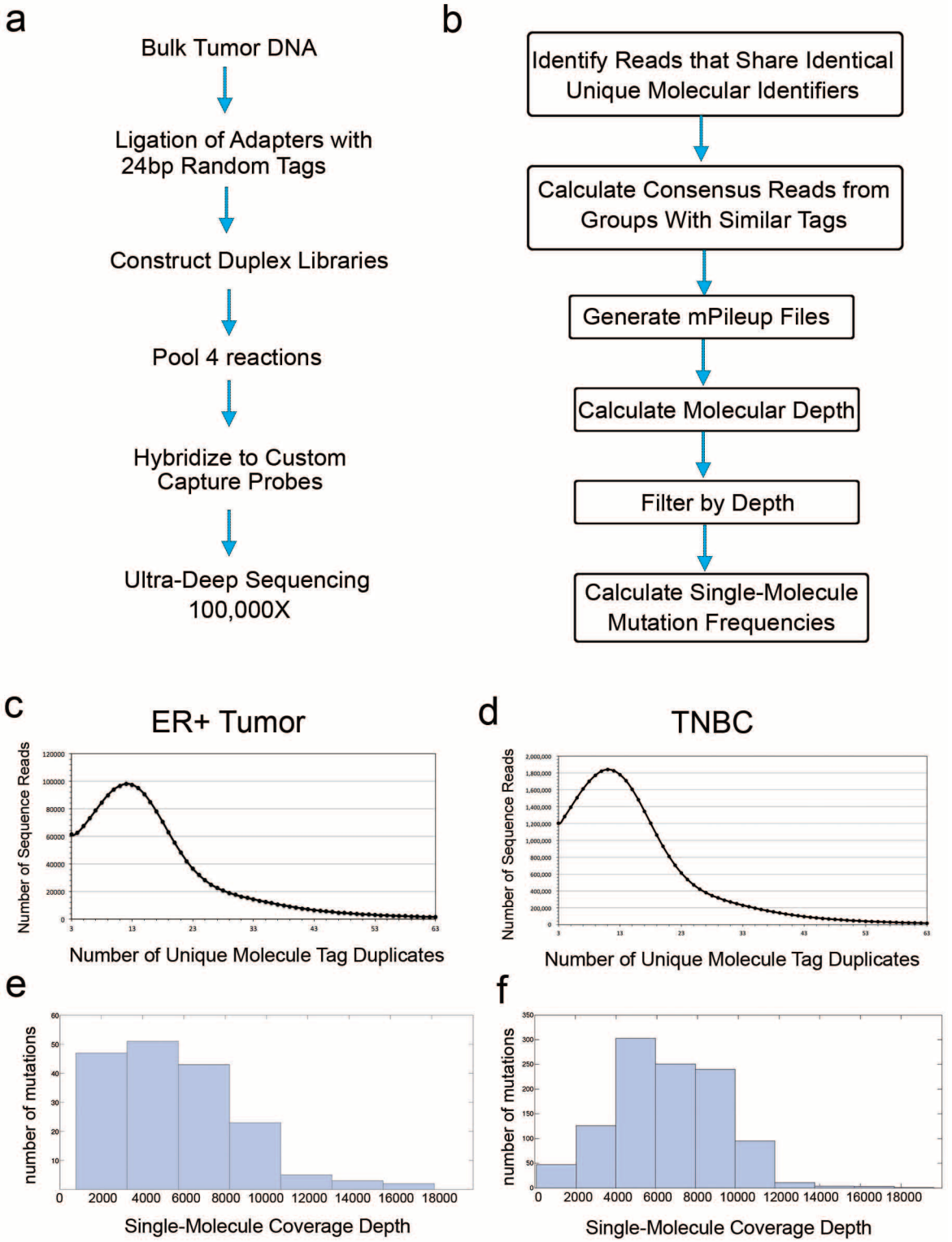
Extended Data Figure 1 | Nuc-seq method. **a**, Nuclear suspensions were prepared and stained with DAPI for flow-sorting, showing distributions of ploidy. The G2/M distribution was gated and single nuclei were deposited into wells. **b**, Cells were lysed and incubated with the Φ29 polymerase to perform multiple-displacement-amplification for a limited isothermal time-frame. **c**, **d**, Sequence libraries were prepared using one of two methods: Tn5 tagmentation (**c**), or low-input TA ligation cloning (**d**) (see Methods). **e**, Exome capture was optionally performed to isolate gDNA in exonic regions. **f**, Libraries were sequenced on the Illumina HiSeq2000 system. **g**, Somatic mutations were detected using a custom processing pipeline (Methods).



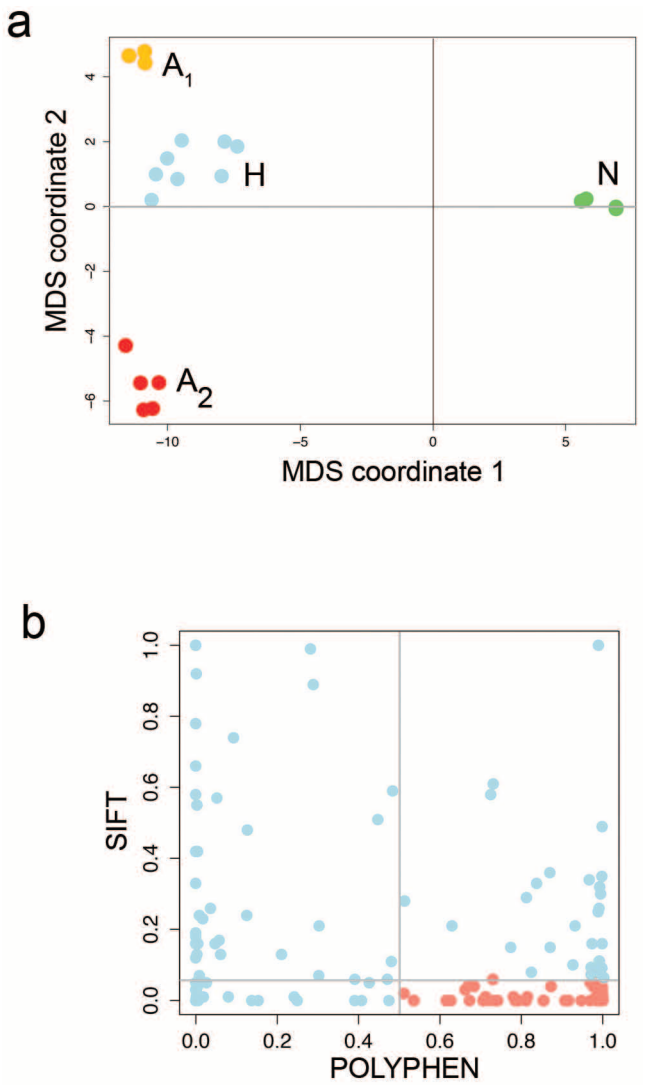
Extended Data Figure 2 | Evaluation of WGA efficiency using chromosome-specific primers. Whole genome amplified DNA from each single cell was used to perform PCR quality control experiments to determine WGA efficiency. For each cell, 22 reactions were performed using primer pairs that target each autosome and the resulting 200 bp PCR product were separated by gel electrophoresis (Methods). **a**, Two single nuclei were flow-sorted from the G2/M gate and amplified to WGA followed by PCR using 22 primer pairs. **b**, Two single nuclei were flow-sorted from the G1/0 gate and subject to WGA followed by PCR using 22 primer pairs. PCR products that failed to amplify are marked with an 'x' on the gel.



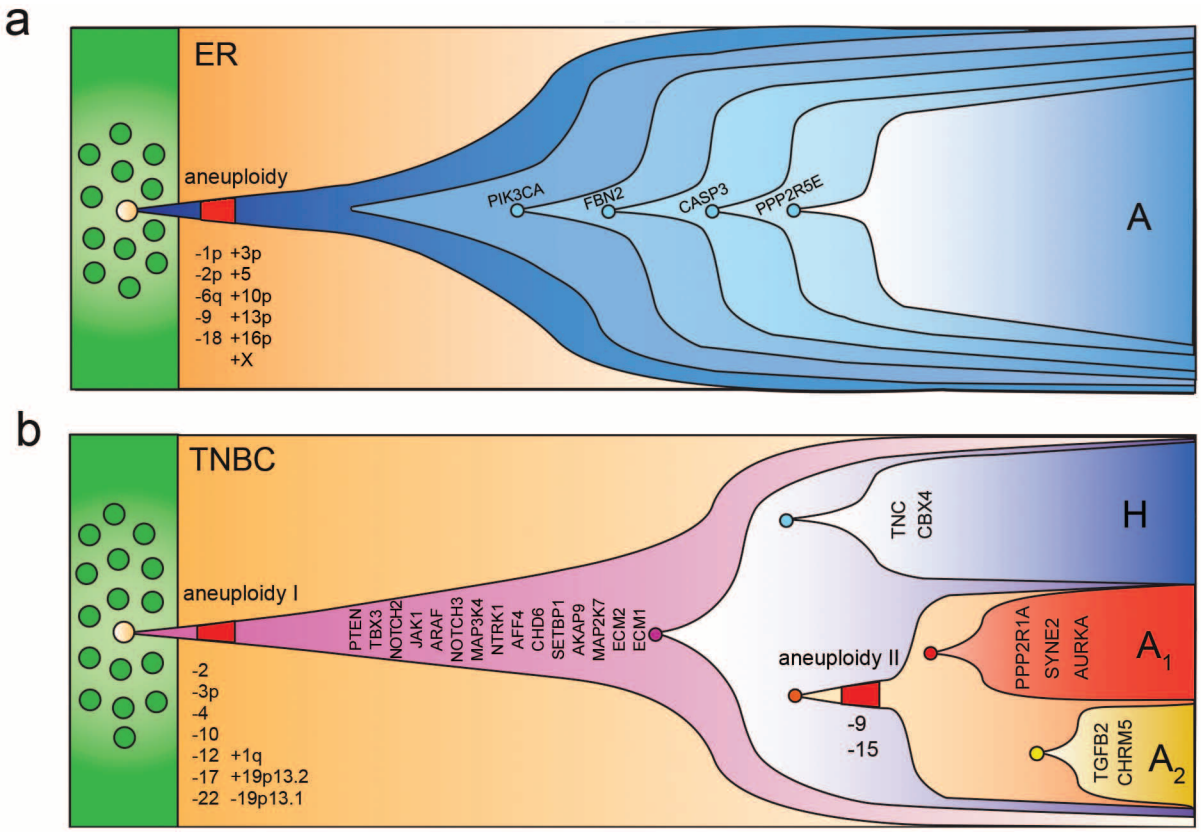
Extended Data Figure 3 | Clustered heatmaps of single cell copy number profiles. Single cell segmented copy number profiles were clustered and used to build heatmaps, showing amplifications in red and deletions in blue. **a**, Copy number profiles of 50 single cells from the ERBC. **b**, Copy number profiles of 50 single cells from the TNBC patient.



Extended Data Figure 4 | Duplex single-molecule targeted deep-sequencing. **a**, Experimental protocol for generating duplex libraries from bulk tumour DNA for custom capture and targeted ultra-deep sequencing. **b**, Data processing pipeline for duplex data to generate single-molecule data and detect mutation frequencies. **c**, Distribution of unique molecule tag duplicates for the ER breast cancer patient **d**, Distribution of unique molecule tag duplicates for the TNBC. **e**, Single-molecule coverage depth distribution for the ER^{+} tumour data. **f**, Single-molecule coverage depth distribution for the TNBC data.



Extended Data Figure 5 | TNBC Multi-dimensional scaling and protein prediction plots. **a**, Multi-dimensional scaling plot of the nonsynonymous mutations from the single-nuclei exome sequencing data in the TNBC **b**, Polyphen and SIFT protein impact prediction scores for the subclonal mutations in the TNBC patient.



Extended Data Figure 6 | Models of clonal evolution in breast cancer. **a**, Clonal evolution in the ERBC inferred from single cell exome and copy number data. **b**, Clonal evolution in the TNBC inferred from single cell exome and copy number data.