LETTER

doi:10.1038/nature09807

Tumour evolution inferred by single-cell sequencing

Nicholas Navin[1,2], Jude Kendall[1], Jennifer Troge[1], Peter Andrews[1], Linda Rodgers[1], Jeanne McIndoo[1], Kerry Cook[1], Asya Stepansky[1], Dan Levy[1], Diane Esposito[1], Lakshmi Muthuswamy[3], Alex Krasnitz[1], W. Richard McCombie[1], James Hicks[1] & Michael Wigler[1]

Genomic analysis provides insights into the role of copy number variation in disease, but most methods are not designed to resolve mixed populations of cells. In tumours, where genetic heterogeneity is common[1–3], very important information may be lost that would be useful for reconstructing evolutionary history. Here we show that with flow-sorted nuclei, whole genome amplification and next generation sequencing we can accurately quantify genomic copy number within an individual nucleus. We apply single-nucleus sequencing to investigate tumour population structure and evolution in two human breast cancer cases. Analysis of 100 single cells from a polygenomic tumour revealed three distinct clonal subpopulations that probably represent sequential clonal expansions. Additional analysis of 100 single cells from a monogenomic primary tumour and its liver metastasis indicated that a single clonal expansion formed the primary tumour and seeded the metastasis. In both primary tumours, we also identified an unexpectedly abundant subpopulation of genetically diverse 'pseudodiploid' cells that do not travel to the metastatic site. In contrast to gradual models of tumour progression, our data indicate that tumours grow by punctuated clonal expansions with few persistent intermediates.

In single-nucleus sequencing (SNS), we isolate nuclei by flow-sorting and amplify DNA using whole genome amplification (WGA) for massively parallel sequencing (Supplementary Fig. 1). We achieve low coverage (~6%) of the genome of a single cell, sufficient to quantify copy number from sequence read depth. Several features of our data analysis were designed for SNS and differ from previous methods[4–6] for measuring copy number from sequencing data. In contrast to using fixed intervals to calculate copy number, we use variable length bins but with uniform expected unique counts, which correct for biases that have been reported[7–9] in WGA (Supplementary Fig. 2; see Methods). For each single cell, we typically achieve a mean read density of 138 per bin (standard error of the mean (s.e.m.) ± 5.55, $n = 200$). Over-replicated loci called 'pileups', which have been previously reported in WGA[10–12], do occur in our data but not at recurrent locations in different cells (Supplementary Fig. 3). Pileups are sufficiently randomly distributed and sparse so as not to affect counting at the resolution we

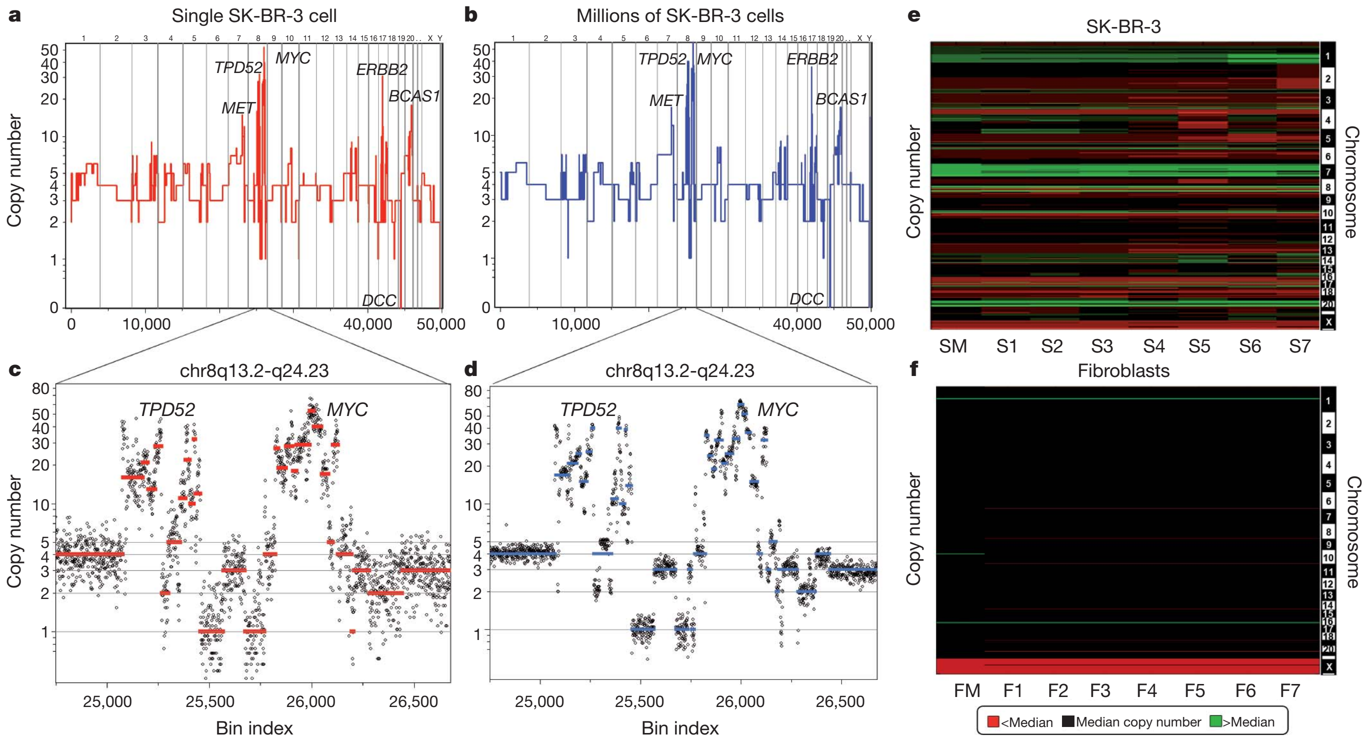

Figure 1 | Comparison of SK-BR-3 single cells to millions. **a**, **b**, The integer copy number profile for a single SK-BR-3 cell is shown (**a**) compared to a sequence count profile using millions of cells (**b**). **c**, **d**, A region on chromosome 8q13.2-q24.23 is plotted showing the integer copy number profile (in red or blue) and a ratio of raw bin counts in grey for a single cell (**c**), and a million cells (**d**). **e**, A heatmap of SK-BR-3 copy number profiles comparing a million-cell sample (SM) to seven single cells (S1–S7). **f**, A heatmap of SKN1 normal fibroblast profiles comparing a million-cell sample (FM) to seven single cells (F1–F7).

---

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [2]Department of Genetics, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [3]Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada.

have chosen (54 kb). Assuming that single cells will have discrete copy number states, we segment the variable bins and calculate integer copy number profiles (Supplementary Fig. 4; see Methods).

To validate our method, we compared the sequence counting profile of DNA from a single SK-BR-3 cell (Fig. 1a) with DNA from one million cells (Fig. 1b). The major amplifications (*MET*, *TPD52*, *ERBB2*, *BCAS1*) and deletions (*DCC*) are detected in both profiles, as are much more abundant but less marked small changes in copy number. To demonstrate how reproducible small differences are, we assessed data for a complex region on chromosome 8q13.2-q24.23 that contains more than thirty segments with differing copy number. These data were reproducible in both a single-cell (Fig. 1c) and a million-cell sample (Fig. 1d). We also compared the sequence read profiles from several single cells and from a million cells to each other and to the profile measured by microarray comparative genomic hybridization (CGH) from bulk DNA (Supplementary Fig. 5). In all instances the profiles showed very high ($r^2 > 0.85$) correlation. The reproducibility and variation between single-cell copy number profiles was also investigated by comparing seven single cells from a culture of SK-BR-3 and seven from normal human fibroblasts. These data are shown as heat maps (Fig. 1e–f), which show that some genomic variation exists between cells. The diploid fibroblast cultures showed no random events; we observed only a few consistent events at levels expected for heritable copy number variations.

We selected next two high-grade (III), triple-negative (ER$^-$, PR$^-$, HER2$^-$) ductal carcinomas (T10, T16P) and a paired metastatic liver carcinoma (T16M) to study tumour population structure and infer tumour evolution by single-cell analysis. T10 was selected to study primary tumour growth because it was previously shown[13] to be genetically heterogeneous (polygenomic), and T16P was selected because it was classified as genetically homogeneous (monogenomic).

T10 was macrodissected into 12 sectors to preserve anatomical information, and nuclei were flow-sorted from six sectors (S1–S6) for SNS (Fig. 2a). Fluorescence-activated cell sorting (FACS) analysis

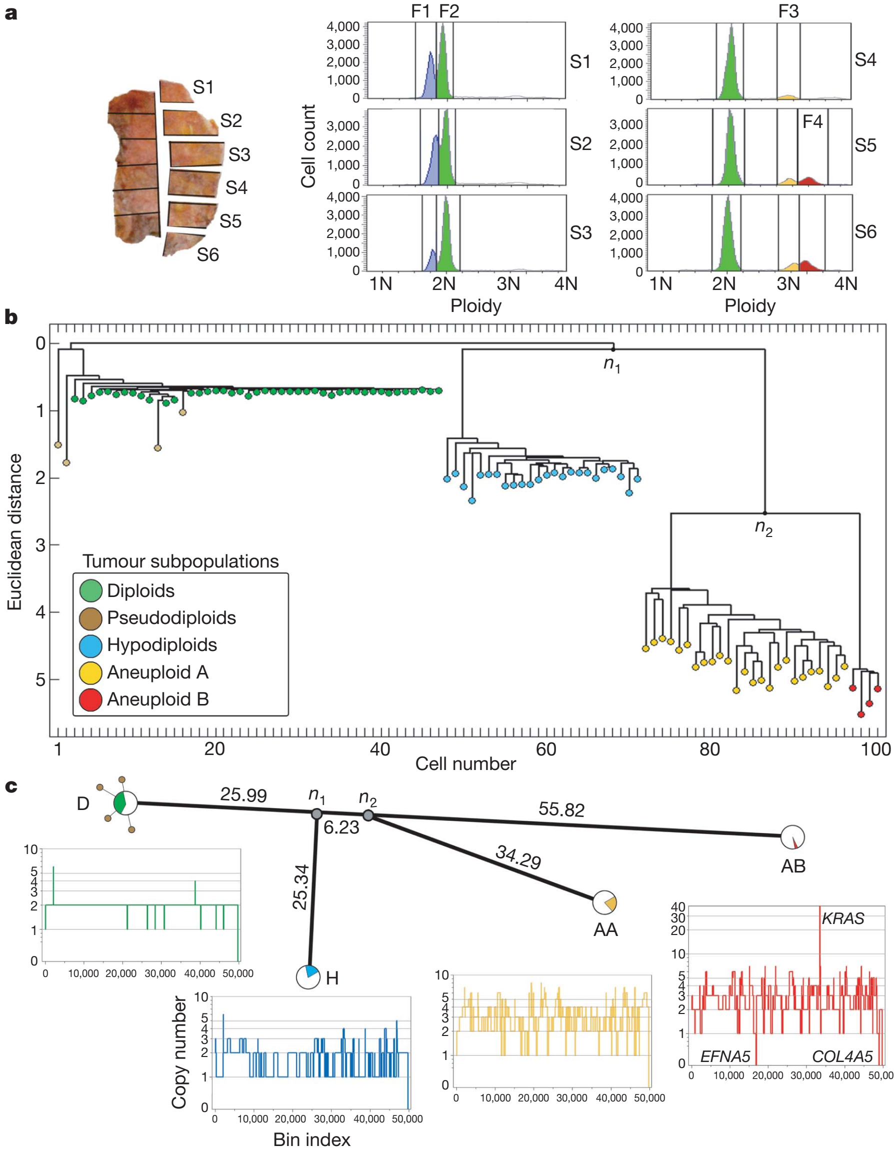

Figure 2 | Analysis of 100 single cells from a polygenomic breast tumour. **a**, T10 was macrodissected into 12 sectors, and nuclei were isolated from six sectors and flow-sorted by ploidy. FACS profiles show four distributions of ploidy (F1–F4), which were gated to isolate 100 single cells. **b**, Neighbour-joining tree of integer copy number profiles showing four major branches of evolution. **c**, Phylogenetic tree of consensus profiles show the common ancestors and evolutionary distance between subpopulations. Integer copy number profiles from single cells are displayed below, and pie charts indicate the percentage of cells that constitute each subpopulation.

showed four major distributions of ploidy: a hypodiploid fraction (F1) exclusive to sectors 1–3; a diploid 2N fraction (F2) in all sectors; and two subtetraploid fractions (F3 and F4) in sectors 4–6. We selected 100 single cells from multiple sectors and ploidy fractions for sequencing and calculation of integer copy number profiles (Supplementary Table 1).

Breast tumours are typically mixtures of cancer cells with normal tissue, stroma and infiltrating leukocytes. By histopathology, T10 was assessed to contain 63% normal and 37% tumour cells and noted to be heavily infiltrated with leukocytes. Most of the diploid nuclei from F2 had flat genome profiles, characteristic of normal cells. Nearly two-thirds (31/47) of these diploid profiles showed narrow deletions in the T-cell receptor loci or one or more immunoglobulin variable region loci, consistent with infiltration by immunocytes (data not shown). Of the remaining sixteen nuclei from F2, twelve showed no discernable aberrations, but four nuclei showed aberrant profiles with diverse chromosome gains and losses. Each of these 'pseudodiploid' nuclei profiles seemed unrelated to the others or to those of the major tumour cell populations found in fractions F1, F3 and F4.

To determine population substructure we calculated pair-wise distances between the 100 integer copy number profiles, and built a tree using neighbour joining[14] (Fig. 2b). The 100 profiles clustered into four subpopulations (D+P, H, AA and AB) regardless of their sector of origin. The D+P subpopulation contains predominantly flat diploid (D) profiles, but also pseudodiploid (P) cells that have diverged by varying degrees from the diploids. The three major 'advanced' tumour subpopulations (H, AA and AB) are highly clonal with complex genomic rearrangements, and together comprise slightly less than half the cells of the tumour. These cells were isolated from the hypodiploid (F1) and two subtetraploid (F3 and F4) ploidy fractions, respectively. We had previously identified these subpopulations by profiling millions of cells by array CGH[13], but we could not determine if they were composite mixtures of different tumour clones. By SNS we can now see that each subpopulation is composed of cells that share highly similar copy number profiles, probably representing three clonal expansions. Each subpopulation (H, AA and AB) is clearly related to the others by many shared genomic alterations, but they have also diverged and developed distinct attributes (for example, a massive 50-fold amplification of the *KRAS* oncogene in AB). The H cells display the characteristic 'sawtooth' pattern[15] comprising broad chromosomal deletions (Fig. 2c). They are anatomically segregated in sectors S1–S3 of the tumour, whereas the AA and AB clones are intermixed and occupy sectors S4–S6.

To understand the relationship between subpopulations, we clustered profiles by chromosome breakpoints (which are directly related to the steps by which tumour cells diverge). We identified 657 copy number breakpoints and used them to build a phylogenetic tree, which closely resembles the structure of the neighbour-joining tree based on copy number (Supplementary Fig. 6). We also applied biclustering[16] to construct a heat map of breakpoints, and ordered it on the basis of the copy number tree to show which breakpoints were common or divergent between the major subpopulations (Supplementary Fig. 7a). Although there is considerable variation within each subpopulation, no obvious further population substructure was evident. To estimate the common ancestors, we constructed a phylogenetic lineage using the consensus breakpoint patterns from the

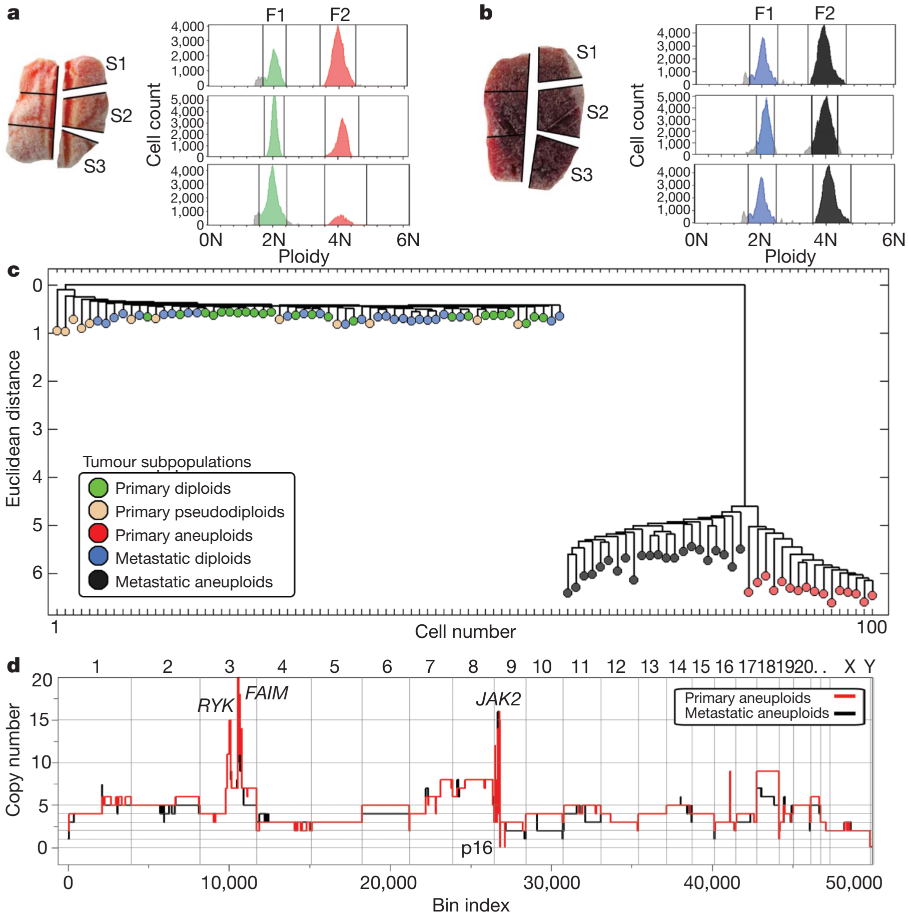

Figure 3 | Analysis of 100 single cells from a monogenomic breast tumour and its liver metastasis. **a**, **b**, Primary breast tumour T16P was macrodissected and 52 nuclei were isolated from three sectors for FACS, showing two distributions of ploidy (F1 and F2). **b**, Liver metastasis T16M was macrodissected and 48 nuclei were isolated from three sectors for FACS also showing two ploidy distributions (F1 and F2). **c**, Neighbour-joining tree of combined integer copy number profiles from the primary and metastatic tumours. **d**, Comparison of primary and metastatic aneuploid consensus copy number profiles.

major tumour subpopulations (Fig. 2c). This lineage shows that the n_1 common ancestor diverged a significant distance from the diploid cells, but that the distance between n_1 and n_2 is very small. By contrast, the divergence of the subpopulations after n_1 and n_2 is very large, with AB showing the greatest phylogenetic distance from the diploids. Thus we infer that the three subpopulations emerged when the tumour was much smaller.

We investigated a second tumour to determine whether these findings extend. We isolated 52 cells from a primary breast tumour (T16P) and 48 cells from its associated liver metastasis (T16M). Each tumour was macrodissected into six sectors, three of which were flow-sorted (Fig. 3a, b). Both T16M and T16P showed diploid peaks (F1) and a single aneuploid tetraploid peak (F2) of roughly equal cell count in all sectors (Supplementary Table 2), consistent with histological sections showing approximately 50% tumour and 50% normal (stromal) cells with low leukocyte infiltration in both samples. To explore population substructure we again constructed neighbour-joining trees from the integer copy number profiles, combining the primary and metastasis cells (Fig. 3c). We observed again numerous pseudodiploid cells, but a single subpopulation of aneuploid cells very diverged from the diploid population. As for T10, the 12 pseudodiploid cells from T16P showed diverse genomic lesions with no clear relationships to each other or to the main tumour lineage. Of the 24 normal diploids in the primary, two had deletions of the T-cell receptor. There were no pseudodiploid cells among the 26 diploid cells from the metastasis.

These data indicate that the primary tumour mass formed by a single clonal expansion of an aneuploid cell, and that one of the cells from this expansion subsequently seeded the metastatic tumour with little further evolution. There are no branches of the tree corresponding to cells intermediate between the aneuploid subpopulation and the diploid root. Although closely related, the primary and metastatic aneuploid cells cleanly separate using the Euclidean metric (Fig. 3c), indicating that the two populations have not mixed since seeding the metastasis. The differences in the profiles that distinguish the primary and metastatic tumour populations are in the degree of copy number change rather than breakpoints (Fig. 3d). In a hierarchical tree created from breakpoints alone, we cannot cleanly separate primary from metastatic aneuploid cells (Supplementary Fig. 6b). Moreover, when we calculate common breakpoints in the single-cell profiles and apply biclustering to ordered samples (Supplementary Fig. 7b), a large number of breakpoints are common to both populations and no breakpoints cleanly distinguish them. By these analyses, no further population substructure is evident.

In contrast to the clear clonal relationships among aneuploid subpopulations, pseudodiploid cells are unusual in showing remarkable genomic heterogeneity (Fig. 4). Pseudodiploid profiles are characterized by nonrecurring copy number changes (including whole chromosome arms) that are not shared between any two pseudodiploid cells, nor with the corresponding tumour profiles (Fig. 4e). These data indicate that unlike the aneuploid cells, pseudodiploids do not undergo clonal expansions in the tumour. Nevertheless, they comprise a substantial proportion of the diploid gated cells: 8% in T10 (4/47) and 33% in T16P (12/36), or approximately 4% and 24% of the tumour mass, respectively. In contrast, the 18 profiles from single nuclei of normal adjacent breast tissue are all flat (Fig. 4a). The relative abundance of pseudodiploid cells in primary tumours indicates that they may emerge from an ongoing aberrant process that generates genomic diversity in the tumour.

In principle, we can learn about DNA sequence mutations from SNS data. However, the sparse sequence coverage makes this analysis problematic. By combining data from multiple cells, belonging to well-defined subpopulations, we can perform global and regional analysis at the many nucleotide positions where sufficient numbers of sequence

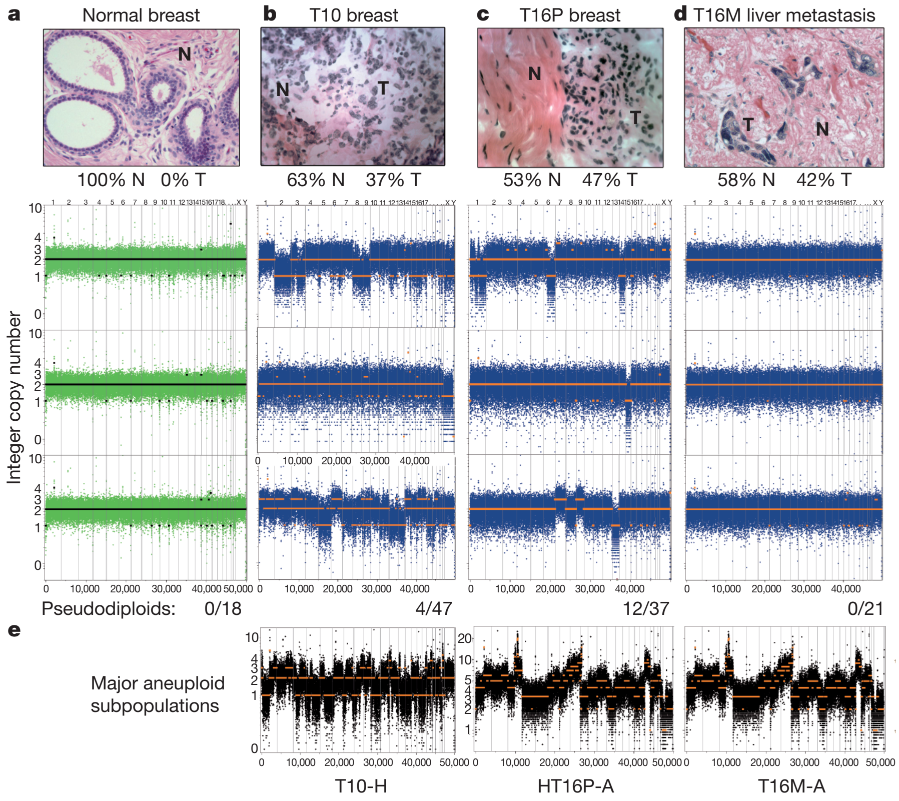

Figure 4 | Genetically diverse pseudodiploid cells in the diploid fractions of tumours. a–d, Haematoxylin and eosin stained tissues sections are shown in the upper panels with normal (N) and tumour (T) cell percentages indicated. Lower rows show bin counts and copy number profiles of single cells isolated from the 2N gated ploidy distributions, and the total number of cells analysed is indicated below each column. The columns are: normal breast tissue cells (**a**); pseudodiploid cells in T10 (**b**); pseudodiploid cells in T16P (**c**); and diploid-gated nuclei from T16M (**d**). **e**, Bin counts and copy number profiles of single cells from the major aneuploid tumour subpopulations.

reads overlap. When examined this way, losses of heterozygosity are unequivocally significant, and map in large contiguous genomic blocks that correlate well with copy number loss (Supplementary Fig. 8 and Supplementary Table 3). The extensive loss of heterozygosity detected in all of the T10 subpopulations and in T16 indicates that both cancers passed through a hypodiploid stage.

Our study demonstrates that we can obtain robust high-resolution copy number profiles by sequencing a single cell and that by examining multiple cells from the same cancer we can make inferences about the evolution and spread of cancer. Moreover, the identification of pseudodiploid cells shows that these methods can identify cell types previously undetectable by other methods. Our findings are consistent with previous findings[17] using bulk DNA, which indicate that copy number profiles in primary tumours are highly similar to the metastases. Thus, the metastatic cells emerge from a main advanced expansion, and not from an earlier intermediate or a completely different subpopulation. This is consistent with recent deep-sequencing studies of primary–metastatic pairs, all indicating that metastatic cells arise late in tumour development[18,19].

There are many gradual models for tumour progression, including clonal evolution[20], the mutator phenotype[21,22] and stochastic progression[23]. Although we have examined only two cancers in depth, both show a pattern of tumour growth that we call ‘punctuated clonal evolution’, borrowing a term from species evolution used to explain gaps in the fossil record[24]. Explicitly, the tumour subpopulations are each distant from their root, without observable intermediate branching. In contrast to gradual models, this pattern reflects the sudden emergence of a tumour cell whose rate of effective population growth markedly exceeds its rate of genomic evolution.

METHODS SUMMARY

To perform SNS, nuclei are isolated either from cells in culture or frozen tumour sections and stained with 4′,6-diamidino-2-phenylindole (DAPI). We use FACS to gate a desired population of nuclei by total DNA content and to deposit nuclei singly into 96-well plates. After WGA using Sigma GenomePlex, we sonicate to create free DNA ends without WGA adapters, and then construct libraries for 76 bp, single-end sequencing using one lane of an Illumina GA2 flowcell per nucleus. For each nucleus we typically achieve 9 million (mean = 9.042 million, s.e.m. ± 0.328, $n = 200$) uniquely mapping reads using the Bowtie[25] alignment software. These sequences cover about 6% (mean = 5.95%, s.e.m. ± 0.229, $n = 200$) of the genome, and are used to count sequence reads in 50,000 variable bins. The bin counts are segmented using a KS statistic and used to calculate integer copy number profiles. Neighbour-joining trees are constructed from the integer profiles and from the chromosome breakpoint patterns of each cell to infer evolution.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 May 2010; accepted 7 January 2011.

Published online 13 March 2011.

1. 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).
2. 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).
3. Farabegoli, F. *et al.* Clone heterogeneity in diploid and aneuploid breast carcinomas as detected by FISH. *Cytometry* **46,** 50–56 (2001).
4. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* **6,** 99–103 (2009).
5. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19,** 1586–1592 (2009).
6. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41,** 1061–1067 (2009).
7. Geigl, J. B. *et al.* Identification of small gains and losses in single cells after whole genome amplification on tiling oligo arrays. *Nucleic Acids Res.* **37,** e105 (2009).
8. Fuhrmann, C. *et al.* High-resolution array comparative genomic hybridization of single micrometastatic tumor cells. *Nucleic Acids Res.* **36,** e39 (2008).
9. Pugh, T. J. *et al.* Impact of whole genome amplification on analysis of copy number variants. *Nucleic Acids Res.* **36,** e80 (2008).
10. Talseth-Palmer, B. A., Bowden, N. A., Hill, A., Meldrum, C. & Scott, R. J. Whole genome amplification and its impact on CGH array profiles. *BMC Res. Notes* **1,** 56 (2008).
11. Hughes, S. *et al.* Use of whole genome amplification and comparative genomic hybridisation to detect chromosomal copy number alterations in cell line material and tumour tissue. *Cytogenet. Genome Res.* **105,** 18–24 (2004).
12. Huang, J., Pang, J., Watanabe, T., Ng, H. K. & Ohgaki, H. Whole genome amplification for array comparative genomic hybridization using DNA extracted from formalin-fixed, paraffin-embedded histological sections. *J. Mol. Diagn.* **11,** 109–116 (2009).
13. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20,** 68–80 (2010).
14. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).
15. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16,** 1465–1479 (2006).
16. Prelic, A. *et al.* A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22,** 1122–1129 (2006).
17. Liu, W. *et al.* Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nature Med.* **15,** 559–565 (2009).
18. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
19. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).
20. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
21. Loeb, L. A., Springgate, C. F. & Battula, N. Errors in DNA replication as a basis of malignant changes. *Cancer Res.* **34,** 2311–2321 (1974).
22. 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).
23. Heng, H. H. *et al.* Stochastic cancer progression driven by non-clonal chromosome aberrations. *J. Cell. Physiol.* **208,** 461–472 (2006).
24. Gould, S. J. & Eldredge, N. Punctuated equilibria comes of age. *Nature* **366,** 223–227 (1993).
25. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Ronemus, T. Spencer, A. Leotta, J. Meth, M. Kramer, L. Gelley, E. Ghiban. We also thank P. Blake and N. Navin at Sophic Systems Alliance. This work was supported by the NCI T32 Fellowship to N.N., and grants to M.W. and J.H. from the Department of the Army (W81XWH04-1-0477), the Breast Cancer Research Foundation, and the Simons Foundation. M.W. is an American Cancer Society Research Professor.

Author Contributions N.N. designed and performed experiments and analysis, and wrote the manuscript. J.K., A.K., L.M., D.L. and P.A. developed analysis programs. J.T., L.R., K.C., J.M., D.E. and A.S. performed experiments. W.R.M. designed experiments. J.H. and M.W. designed experiments, performed analysis and wrote manuscript.

Author Information All data has been deposited into the NCBI Sequence Read Archive under accession number SRA018951.105. 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 this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.W. (wigler@cshl.edu).

METHODS

Samples. The frozen ductal carcinoma T10 (CHTN0173) was obtained from the Cooperative Human Tissue Network, and T16P and T16M were obtained from Asterand. Pathology shows that both tumours were poorly differentiated and high grade (III) as determined by the Bloom–Richardson score, and triple-negative (ER$^-$, PR$^-$ and HER2/NEU$^-$) as determined by immunohistochemistry. The cell lines used in this study include a normal male immortalized skin fibroblast (SKN1) and a breast cancer cell line (SK-BR-3). Normal breast tissue was obtained from H. Hibshoosh from Columbia University.

SNS. Nuclei were isolated from cell lines and from the frozen tumour using an 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. The frozen tumour was first macrodissected into 12 sectors of equal size using surgical scalpels and nuclei were isolated from six sectors for FACS by finely mincing a tumour sector in a Petri dish in 1.0–2.0 ml of NST-DAPI buffer using two no. 11 scalpels in a cross-hatching motion. The cell lines were lysed directly in a culture plate using the NST-DAPI buffer, after first removing the cell culture media. All nuclei suspensions were filtered through 37-µm plastic mesh before flow-sorting.

Single nuclei were sorted by FACS using the BD Biosystems Aria II flow cytometer by gating cellular distributions with differences in their total genomic DNA content (or ploidy) according to DAPI intensity. First, a small amount of prepared nuclei from each tumour sample was mixed with a diploid control sample (derived from a lymphoblastoid cell line of a normal person) to accurately determine the diploid peak position within the tumour and establish FACS collection gates. Before sorting single nuclei, a few thousand cells were sorted to determine the DNA content distributions for gating. A 96-well plate was prepared with 10 µl of lysis solution in each well from the Sigma-Aldrich GenomePlex WGA4 kit. Single nuclei were deposited into individual wells in the 96-well plate along with several negative controls in which no nuclei were deposited.

WGA was performed on single flow-sorted nuclei as described in the Sigma-Aldrich GenomePlex WGA4 kit (catalogue no. WGA4-50RXN) protocol. WGA fragments from the frozen breast tumour and SK-BR-3 single cells were used directly for single-read library construction using the Illumina Genomic DNA Sample Prep Kit (catalogue no. FC-102-1001) and following standard protocol with a gel purification size range of 300–250 bp. WGA fragments from the fibroblast cell line were first sonicated using the Diagenode Bioruptor using the following program: 2 times, 7 min with 30 s high on/off mode in ice-cold water. Sonication removes a specific 28 bp adaptor sequence that is added on during WGA, and improves the total number of sequencing reads per lane.

Single-read libraries from single nuclei were sequenced on individual flow-cell lanes using the Illumina GA2 analyser for 76 cycles. Data was processed using the Illumina GAPipeline-1.3.2 to 1.6.0. Sequence reads were aligned to the human genome (HG18/NCBI36) using the Bowtie alignment software[25] with the following parameters: 'bowtie –S –t –m 1 –best –strata –p16' to report only top scoring unique mappings for each sequence read. For each nucleus we typically achieve 9 million (mean = 9.042 million, s.e.m. ± 0.328, $n = 200$) uniquely mapping reads. These sequences cover about 6% (mean = 5.95%, s.e.m. ± 0.229, $n = 200$) of the genome uniquely. To eliminate PCR duplicates, we removed sequences with identical start coordinates.

Read depth counting in variable bins. Copy number is calculated from read density, by dividing the genome into 'bins' and counting the number of unique reads in each bin. In previous copy number studies read density was calculated using bins with uniform fixed length[16–19]. In contrast, we use bins of variable length that adjust size depending on the mappability of sequences to regions of the human genome. In regions of repetitive elements, lower numbers of reads are expected and thus the bin size is increased. To determine interval sizes we simulated sequence reads by sampling 200 million sequences of length 48 from the human reference genome (HG18/NCBI36) and introduced single nucleotide errors with a frequency encountered during Illumina sequencing. These sequences were mapped back to the human reference genome using Bowtie[25] with unique parameters as described earlier. 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 50,009 genomic bins with no bins crossing chromosome boundaries. The median genomic length spanned by each bin is 54 kb. For each cell the number of reads mapped to each variable length bin was counted. This variable binning efficiently reduces false deletion events when compared to uniform length-fixed bins as shown in Supplementary Fig. 2b and c. For a single cell we typically measure 138 sequence reads per bin.

Integer copy number quantification. Single cells will have integer copy number states that we can infer from sequence read counts, as follows. Unique sequence reads are counted in variable bins (Supplementary Fig. 4a) and segmented using the Kolmogorov–Smirnov (KS) statistic (Supplementary Fig. 4b). To estimate the integer differences of copy number states, we calculate Gaussian kernel smoothed density plots using Splus (MathSoft), showing the difference between median bin counts for all pair-wise combinations of different segments (Supplementary Fig. 4c–e). The uniform steps between groups are very apparent, and are a general property of single-cell data. We then convert our KS-segmented data into profiles of integer copy number as follows. We take the differential bin count of the second peak, denoted by an asterisk in Supplementary Fig. 4a, to represent a copy number 'increment' of 1. We then divide every bin count in the profile by the increment and round to infer the integer copy number. We show in Supplementary Fig. 4f–g how closely the segmentation profile agrees with the integer copy number profile. However, for diploid or near diploid cells there are few to no steps from which to observe the increment, and we use a different method, taking the increment as the median bin count on the autosomes divided by two.

Gene annotations. Amplifications and deletions identified in the single-cell copy number profiles were annotated to identify UCSC genes. Cancer genes were identified using a compiled database from the cancer gene consensus and the NCI cancer gene index (Sophic Systems Alliance, Biomax Informatics AG).

Neighbour-joining trees of copy number profiles. Integer copy number profiles of single cells were used to calculate neighbour-joining trees using a Euclidean distance metric with Matlab (Mathworks). Branches were flipped to orient nodes within subpopulations and trees were rooted using the last common diploid node.

Common breakpoint detection. Breakpoints are defined as bins with a copy number different than the previous bin in genome order. A transition from a lower copy number to a higher copy number (in genome order) is considered to be a different event than the opposite transition. To find breakpoint regions we count each breakpoint in each cell and the immediately neighbouring bins. A contiguous set of bins with counts greater than 1 is designated a breakpoint region. This results in a set of common breakpoint regions. Each cell is then scored for the occurrence of each of these events, a one meaning the cell has a copy number transition of that type (low to high or high to low) in that genomic region and a zero meaning no copy number transition of that type in that region.

Hierarchical tree of chromosome breakpoints. We used chromosome breakpoints patterns to build a neighbour-joining tree. To eliminate breakpoint events with a high standard deviation, we limited our analysis to breakpoint regions covering no more than seven adjacent bins ($N = 657$). Using a Euclidean metric, we calculated a distance matrix from the binary chromosome breakpoint patterns identified in the single cells using Matlab (Mathworks). From this distance matrix we constructed a tree using average linkage.

Heatmap of chromosome breakpoints. The biclustering heatmap is based on the same set of breakpoints used to build the neighbour-joining tree. Colour indicates the presence of an event, and white means no event. The columns are ordered as in the tree. The rows are events ordered to show clearly which of the subsets of the four main groups share which events. The groups are ordered by subpopulation. A four-dimensional binary vector represents each of the 16 possible subsets of these groups (subset vector). Each breakpoint is represented by a four-dimensional vector of the per cent of cells in each group having an event at that breakpoint (the 'breakpoint vector'). The angle from each breakpoint vector to each subset vector is computed as well as the length of each projection vector. If the length of the projection vector is less than 0.05 the breakpoint vector is assigned to the empty (0,0,0,0) subset, otherwise it is assigned to the subset vector with the smallest angle to the breakpoint vector. The rows are ordered by subset vector in the following order: (1,1,1,1), (0,0,0,1), (0,0,1,0), (0,1,0,0), (1,0,0,0), (0,0,1,1), (0,1,0,1), (1,0,0,1), (0,1,1,0), (1,0,1,0), (1,1,0,0), (0,1,1,1), (1,0,1,1), (1,1,0,1), (1,1,1,0), (0,0,0,0). Within each subset the rows are in descending order by the number of cells in that subset having that event and then in ascending order by the number of cells outside of that subset that do not have that same event.

Analysis of loss of heterozygosity using sequence mutations. PCR duplicates were removed from mapped sequence reads and bases with a quality score below 30 were excluded from analysis. We then determined the set of observed nucleotide types for each cell sequenced from the T10 and T16P and T16M tumours and every position in the genome. For each subpopulation we classified a position as the observed nucleotides only if one or two nucleotide types were each observed in five or more cells in the subpopulation. For each grouping of subpopulations DH, DA, if a classification was made in every subpopulation in the group, we translated the classifications into the generic nucleotides (a,b) based upon the order in which they were seen in the group, from left to right. We counted the resulting classifications of positions for each group by class, and determined whether long blocks of identical classifications along a chromosome were expected by chance. To establish the significance of our classification counts, we repeated our analysis 100 times with randomly permuted cell labels within each group of subpopulations. We eliminated any effects from differing subpopulation size in a separate set of runs of the same analysis, each with 24 randomly selected cells in every subpopulation.

Copyright of Nature is the property of Nature Publishing Group and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.