nature medicine

Subclonal diversification of primary breast cancer revealed by multiregion sequencing

Lucy R Yates[1,2], Moritz Gerstung[1], Stian Knappskog[3,4], Christine Desmedt[5], Gunes Gundem[1], Peter Van Loo[1,6], Turid Aas[7], Ludmil B Alexandrov[1,8], Denis Larsimont[5], Helen Davies[1], Yilong Li[1], Young Seok Ju[1], Manasa Ramakrishna[1], Hans Kristian Haugland[9], Peer Kaare Lilleng[9,10], Serena Nik-Zainal[1], Stuart McLaren[1], Adam Butler[1], Sancha Martin[1], Dominic Glodzik[1], Andrew Menzies[1], Keiran Raine[1], Jonathan Hinton[1], David Jones[1], Laura J Mudie[1], Bing Jiang[11], Delphine Vincent[5], April Greene-Colozzi[11], Pierre-Yves Adnet[5], Aquila Fatima[11], Marion Maetens[5], Michail Ignatiadis[5], Michael R Stratton[1], Christos Sotiriou[5], Andrea L Richardson[11,12], Per Eystein Lønning[3,4], David C Wedge[1] & Peter J Campbell[1]

The sequencing of cancer genomes may enable tailoring of therapeutics to the underlying biological abnormalities driving a particular patient's tumor. However, sequencing-based strategies rely heavily on representative sampling of tumors. To understand the subclonal structure of primary breast cancer, we applied whole-genome and targeted sequencing to multiple samples from each of 50 patients' tumors (303 samples in total). The extent of subclonal diversification varied among cases and followed spatial patterns. No strict temporal order was evident, with point mutations and rearrangements affecting the most common breast cancer genes, including *PIK3CA*, *TP53*, *PTEN*, *BRCA2* and *MYC*, occurring early in some tumors and late in others. In 13 out of 50 cancers, potentially targetable mutations were subclonal. Landmarks of disease progression, such as resistance to chemotherapy and the acquisition of invasive or metastatic potential, arose within detectable subclones of antecedent lesions. These findings highlight the importance of including analyses of subclonal structure and tumor evolution in clinical trials of primary breast cancer.

Driver mutations occur in single cells and are associated with subsequent clonal expansion. Consequently, a given patient's breast tumor comprises a complex patchwork of genetically related competing clones[1–3]. Genome sequencing has enabled analysis of clonal evolution in breast cancer through sequencing of primary tumor and metastasis pairs in a few cases[4,5], sequencing of single cells[2,6] and xenograft models[7], and deep sequencing for subclonal mutations[1,3]. These studies have revealed that subclonal evolution occurs in breast cancer, although the findings are based on relatively small sample sizes.

Most breast cancers are localized at first presentation and managed with curative intent by surgery, often in combination with radiotherapy and systemic therapies. Therapies targeting the estrogen and HER2 receptors improve survival, and benefit may extend to cases where the targetable alteration is subclonal[8,9]. Therapies directed against a wider range of biological targets are currently in early-phase trials, but heterogeneity could complicate study design and confound analysis[10,11]. The optimal therapy may be directed against mutations shared by all cells in a cancer, but subclonal mutations may become important later in therapy if they enable subclones to resist treatment or confer metastatic capacity. In colon, pancreatic and hematological cancers, preferred temporal orders of somatic mutation accumulation may predominate[12–15], but whether this applies to breast cancer has not been evaluated. In renal, pancreatic, colon and prostate tumors, geographical stratification of clonal structure is common, with subclones containing driver mutations expanding locally[16–21]. Whether early breast cancers show similar patterns is unknown.

RESULTS

Multiregion sequencing of breast cancer

To determine the patterns of spatial evolution in primary breast cancer, we sequenced multiregion samples from 50 invasive cancers (27 positive for estrogen receptor (ER) expression but negative for HER2 expression (ER^+HER2^-); 3 ER^+HER2^+; and 20 negative for expression of ER, progesterone receptor (PgR) and HER2 ('triple negative'; $ER^-PgR^-HER2^-$); **Supplementary Table 1**). We sequenced the cancers in two cohorts. Cohort 1 contained prospective, systematic needle biopsy samples of 12 primary, treatment-naive, surgically excised cancers (**Fig. 1a**,**b**). In cohort 2, we studied multiple treatment-naive

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK. [2]Department of Oncology, The University of Cambridge, Cambridge, UK. [3]Section of Oncology, Department of Clinical Science, University of Bergen, Bergen, Norway. [4]Department of Oncology, Haukeland University Hospital, Bergen, Norway. [5]Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium. [6]Department of Human Genetics, University of Leuven, Leuven, Belgium. [7]Department of Surgery, Haukeland University Hospital, Bergen, Norway. [8]Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA. [9]Department of Pathology, Haukeland University Hospital, Bergen, Norway. [10]The Gade Laboratory for Pathology, Department of Clinical Medicine, University of Bergen, Bergen, Norway. [11]Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [12]Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to P.J.C. (pc8@sanger.ac.uk).

Received 27 March; accepted 22 May; published online 22 June 2015; doi:10.1038/nm.3886

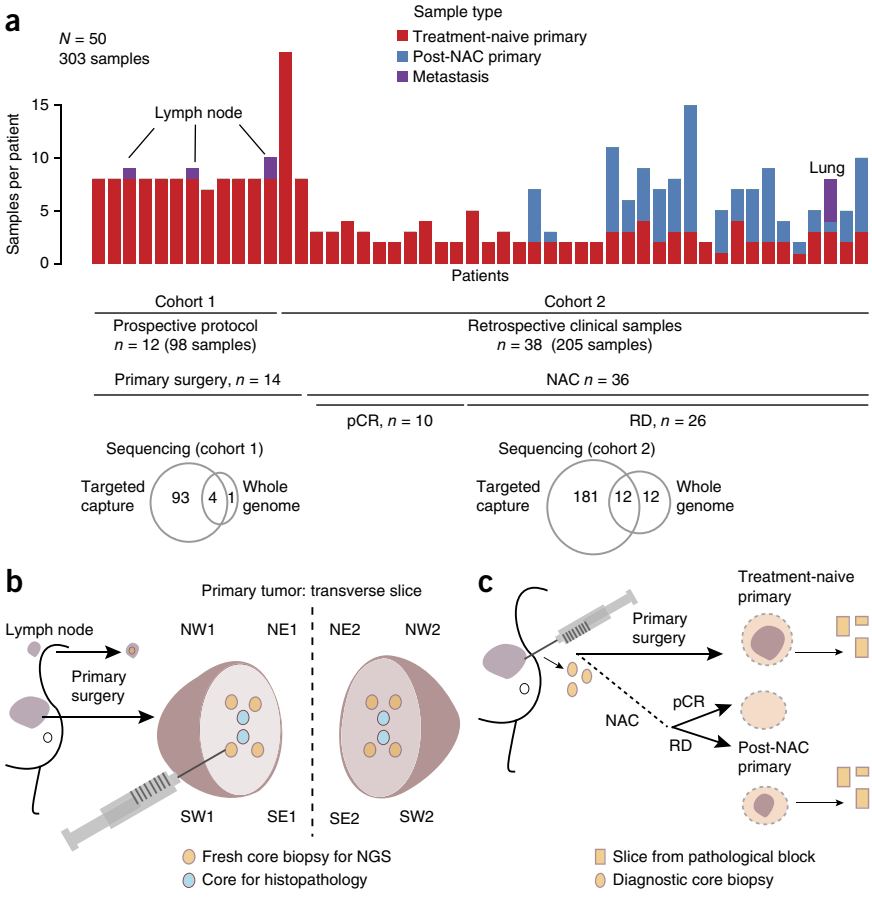

Figure 1 Study design. (**a**) Summary of samples within cohorts 1 and 2. *n* values denote the number of subjects. (**b**) Geographical sampling approach for tumor hemispheres 1 and 2, plus one or two involved lymph nodes in three cases. NW, northwest; NE, northeast; SW, southwest; SE, southeast; NGS, next-generation sequencing. For multifocal cancers, all samples were taken from the single largest focus. (**c**) Source of retrospective clinical samples in relation to primary tumor management. NAC, neo-adjuvant chemotherapy; pCR, pathological complete response; RD, residual disease.

needle biopsy or tissue-block samples from 38 cancers (**Fig. 1a**,**c**). Of the patients biopsied for cohort 2, all but 2 had undergone neoadjuvant chemotherapy, with 10 demonstrating a complete pathological response and 26 having histopathologically confirmed residual disease. For 18 of these 26 cases, we sequenced samples from both pretreatment and post-treatment residual invasive disease. For 290 samples from the 50 cancers studied in both cohorts, we carried out high-coverage sequencing (mean, 166×) (**Supplementary Table 2**) of 360 known cancer genes, chosen from a review of published literature[22–24] and including more than 40 genes recurrently mutated in breast cancer[3,25–30] (**Supplementary Table 3**). For 13 of these cancers we sequenced selected tumor samples ($n = 29$) and a matched constitutional DNA sample in each case to whole-genome level with an average depth of 40-fold (**Supplementary Table 2**).

We identified driver mutations as recurrent mutations in oncogenes or truncating mutations and recurrent missense substitutions in tumor-suppressor genes[3,14,25,27,28,30,31] (details of driver mutation annotation can be found in the Online Methods). Copy-number analysis focused on the five most frequent arm-level copy-number changes[32] and 12 frequently amplified genes in breast cancer[33,34] (**Supplementary Table 3b**).

False positive and false negative mutation calls in multisample studies can lead to the appearance of subclonal heterogeneity that is in fact artifactual. To validate our pipeline, we repeated the targeted capture experiment using independent libraries for 38 needle biopsy samples from five cancers. Positive predictive values and, critically, negative predictive values were on average >99%, confirming our ability to call both the presence and the absence of individual mutations across multiple samples from a single cancer (**Supplementary Table 4**). From whole-genome sequencing data, we successfully verified 2,217 of 2,235 (99%) substitutions and 18 of 19 (95%) insertion-deletions (indels) (**Supplementary Table 4**). We confirmed 1,567 of 1,778 (88%) structural variants using PCR or breakpoint-associated copy-number changes (**Supplementary Table 4**). We achieved 97% concordance between copy-number amplifications called by targeted gene sequencing and those called by multiplex ligation–dependent probe amplification (**Supplementary Table 4**). We validated phylogenetic trees reconstructed from whole-genome data (**Supplementary Fig. 1** and **Supplementary Table 5**) by targeted deep sequencing of mutations on each proposed branch. Cohort 2 contained fresh-frozen and formalin-fixed paraffin-embedded (FFPE) samples with no systematic differences in mutation calls (**Supplementary Fig. 2**).

Geographical patterns of subclonal growth

To assess the spatial distribution of subclones for 12 cancers, we sliced the tumors in half immediately after surgical resection and obtained six needle biopsy samples from the cut face of each half (**Fig. 1b**). We performed targeted gene sequencing of eight biopsies from each primary tumor, and for three of these cases we also sequenced an associated lymph node metastasis (**Fig. 1a**). We evaluated the remaining four biopsies from each primary tumor by histopathology to confirm the presence of invasive cancer and assess Ki-67 levels (**Supplementary Table 1**).

Eight of twelve tumors demonstrated statistically significant spatial heterogeneity of point mutations ($q < 0.05$), and an additional two samples displayed heterogeneity of copy-number changes alone (**Fig. 2a–d** and **Supplementary Table 6**). Layering mutational data onto the spatial arrangement of biopsies showed that local, geographically constrained expansion was the predominant pattern of heterogeneity, with 10 out of 12 cancers having at least one mutation confined to one to three adjacent regions.

Localized confinement of subclones was not always the case. In four cancers we found evidence of an admixture of clones (**Fig. 2d** and **Supplementary Figs. 3** and **4**). The subclonal mutations often had low, but variably distributed, allele fractions in the samples where they were detected, a pattern suggestive of extensive intermingling of subclones across wide geographical ranges. This pattern was common only among larger tumors (four of five tumors >3 cm in size). Similar findings have been observed in follicular lymphoma and colorectal cancer[15,17].

In all 12 cancers, we identified at least one clonal somatic driver mutation or copy-number event shared by all samples. In four cancers we identified subclonal driver mutations, including

recurrent *TP53* missense mutations, *MYC* amplification, a canonical mutation in *PIK3CA* and a nonsense mutation in *BRCA2*. In these four examples, the subclonal driver mutation was absent from five to seven of the eight samples sequenced, despite a collective coverage of about 1,000-fold. In 7 of 12 cases, some mutations were subclonal in the tumor as a whole but could be erroneously characterized as clonal if only a single biopsy were sequenced (**Supplementary Figs. 3** and **4**).

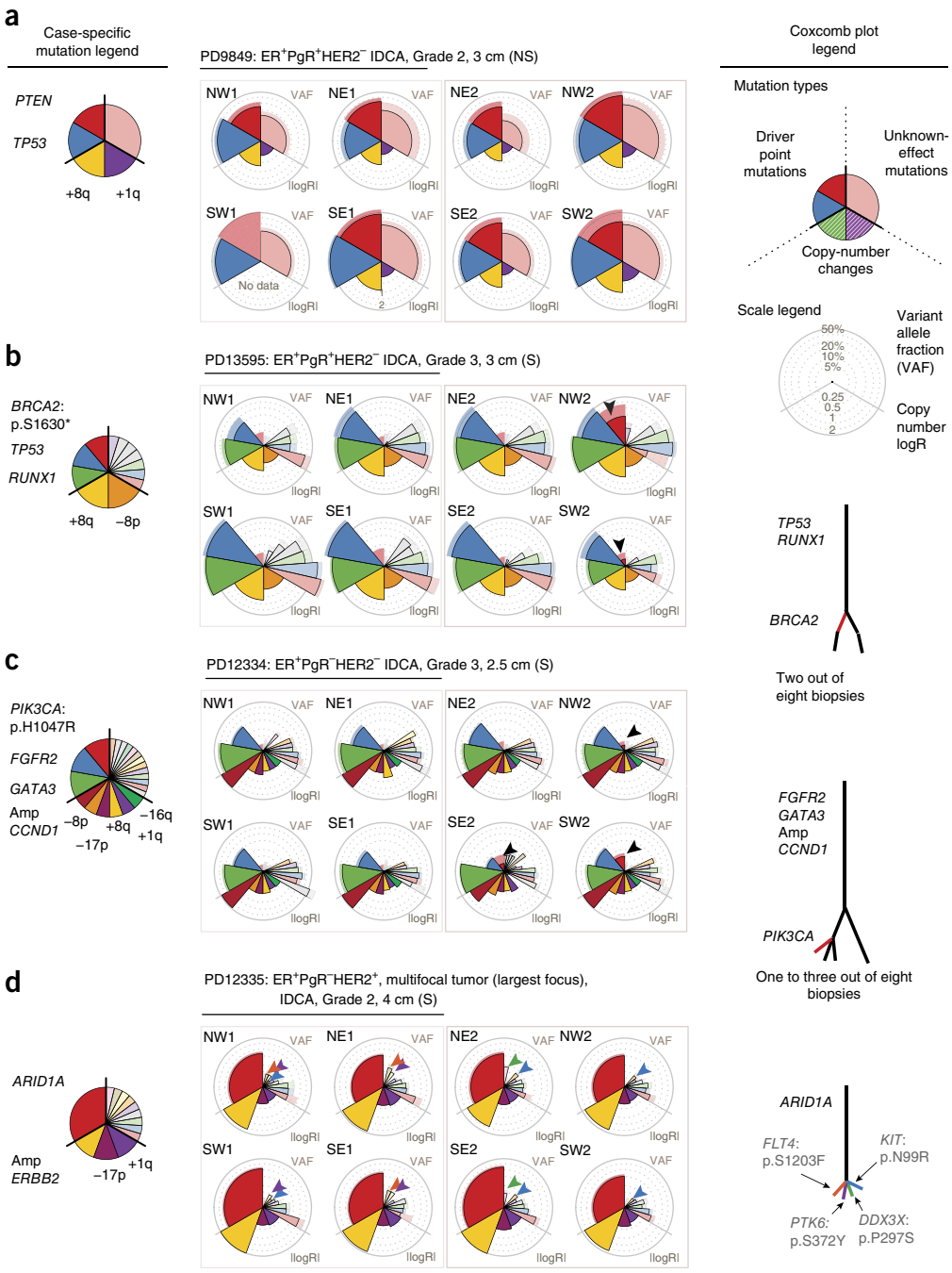

Figure 2 Systematic sampling revealed spatial and temporal tumor evolution. (**a**–**d**) Coxcomb plots presenting somatic mutation genotypes organized according to the sample schema described in **Figure 1b**. Point estimates of the variant-allele frequency (VAF) or copy number (logR) are represented by lateral extension of an outlined wedge. Pale wedges with no outline represent the 95% confidence interval; if coverage is low, the confidence of the VAF is reduced and the pale wedge appears beyond the point estimate. IDCA, invasive ductal carcinoma. Driver mutations and arm-level copy-number gains (+) and losses (–) detected in each cancer are annotated in the case-specific mutation legends. Significant heterogeneity among point mutations in individual cancers was determined using generalized linear models and Benjamini-Hochberg correction: $q < 0.05$ indicates significant point-mutational heterogeneity (S); NS, not significant. Mock phylogenetic trees are also shown. The presence and absence of mutations across related samples indicated distinct subclones and dictated the branching structure, and the number of mutations in each subclone determined the branch length. (**a**) No detected intratumoral heterogeneity ($q = 0.8$). (**b**,**c**) Local expansion of subclones (arrowheads). (**d**) Complex intermixing of subclones: individual mutations (each highlighted with a different-colored arrowhead, with colors corresponding to those in the tree to the right) appeared in different combinations of samples. Coxcomb plots and heat maps for every cancer in the cohort are available at ftp://ftp.sanger.ac.uk/pub/cancer/YatesEtAl/.

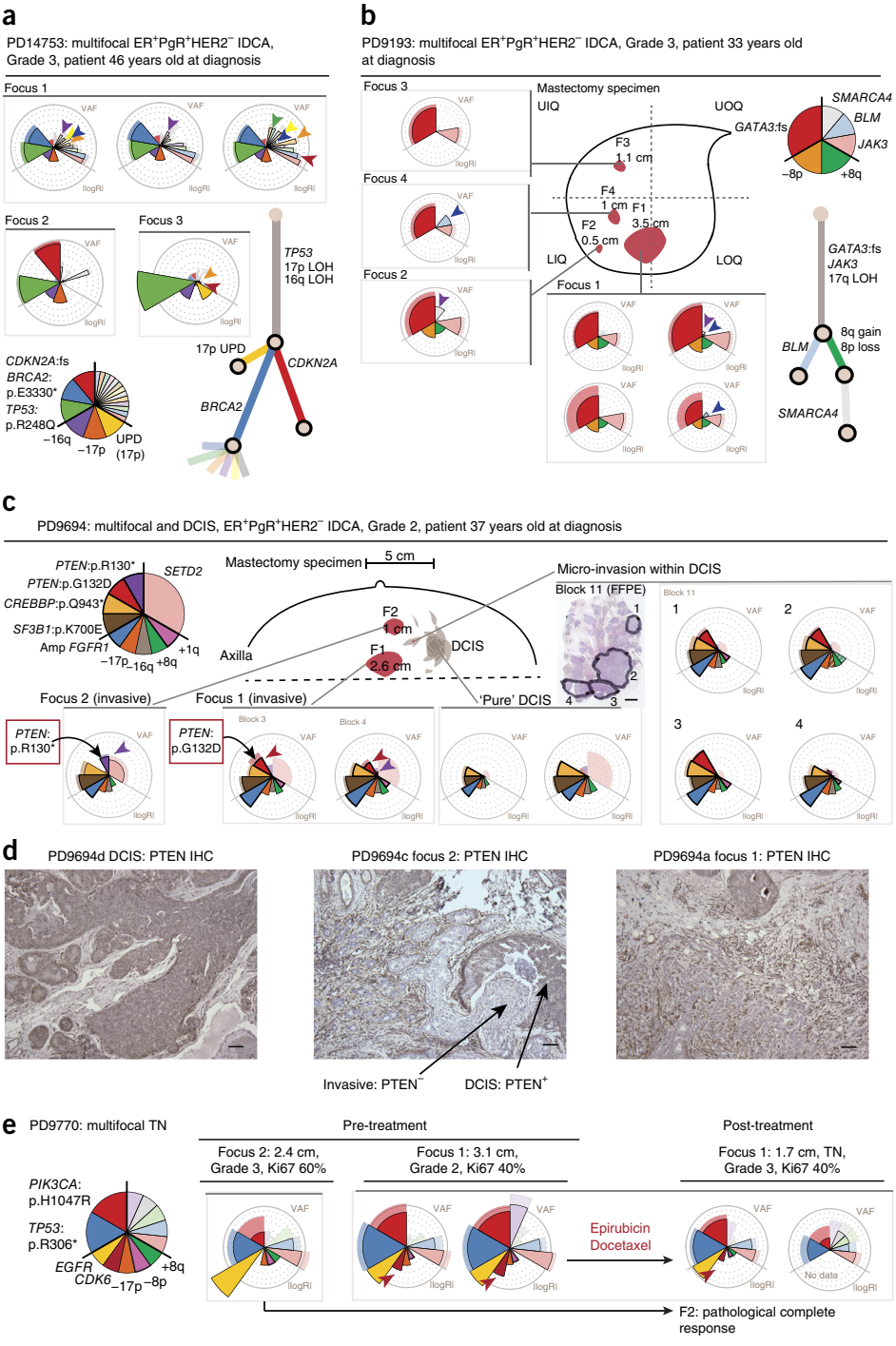

Figure 3 Subclonal patterns in multifocal breast cancers. (**a**–**e**) Targeted capture genomic analysis of subclonal structure (**a**–**c**,**e**) and immunohistochemistry (IHC; **d**) of multifocal cancers. Coxcomb plots and mock phylogenetic trees were generated as described for **Figure 2**, and the scale legend for that figure applies here. Plots from multiple samples from the same tumor focus are grouped together. Colored arrowheads identify subclones that were shared by fewer than all invasive foci. (**a**) Case PD14753: genotypes of five samples from three disease foci indicated deep branching of the tree, driver heterogeneity and subclone intermingling across foci. (**b**) Case PD9193: genotypes of seven samples from four disease foci demonstrated subclone intermingling. Orientation in mastectomy specimen: UIQ, upper inner quadrant; UOQ, upper outer quadrant; LIQ, lower inner quadrant; LOQ, lower outer quadrant. F1–F4, foci 1–4. (**c**) Case PD9694: parallel evolution with two unique *PTEN* driver mutations in different foci. The schematic representation of the mastectomy specimen (center) shows pathological features in the specimen; the dashed horizontal line represents the deep (chest wall) margin. Numbered foci are outlined in the image of a formalin-fixed paraffin-embedded (FFPE) tissue section. Scale bar, 3 mm. (**d**) Case PD9694: PTEN IHC showed that PTEN protein was present in DCIS but lost in invasive disease foci 1 and 2. Scale bars, 100 μm. (**e**) Genotypes of three samples from two disease foci in PD9770 before chemotherapy and two samples from focus 1 after neoadjuvant chemotherapy. Focus 2 exhibited a complete pathological response to three cycles of each chemotherapy agent. IDCA, invasive ductal carcinoma; LOH, loss of heterozygosity; TN, triple negative; UPD, uniparental disomy; VAF, variant-allele frequency.

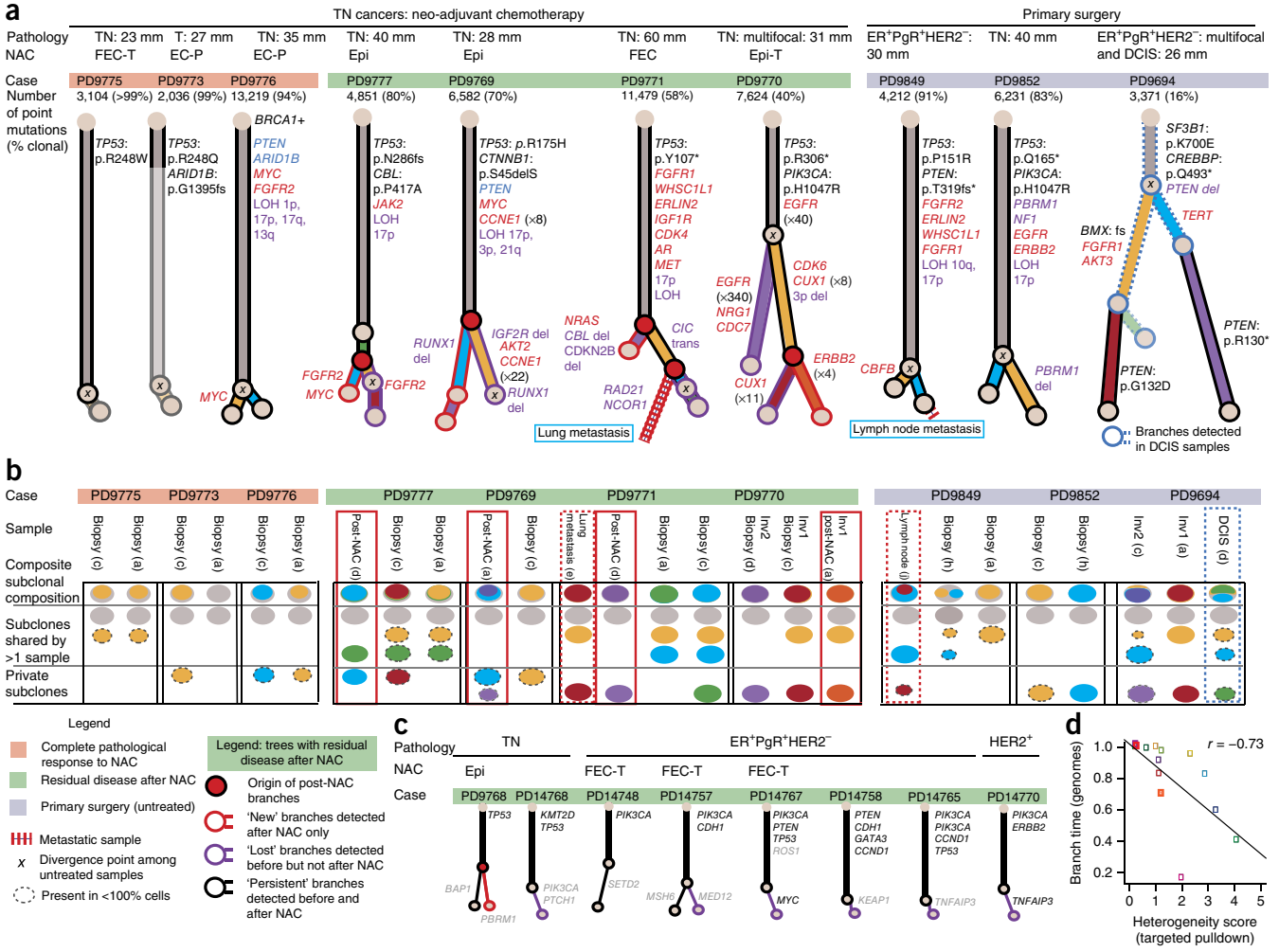

Figure 4 The genome-wide spectrum of branching evolution. (**a**–**c**) Phylogenetics (**a**,**c**) and subclonal composition (**b**) of primary cancers. (**a**) Phylogenetic trees generated by clustering genome-wide point mutation data from ten multiregion primary cancer samples. Relative branch lengths were determined from the proportion of mutations in each branch. An '*x*' indicates the most recent common ancestor inferred from treatment-naive samples alone. (**a**,**c**) Cases for which post-treatment samples were available (green bars above trees); red nodes indicate where subclones detected only after treatment (branches with red outlines) emerged in the tree. Branches detected only among pre-treatment samples are indicated by a purple outline; black branches indicate detection in both pre- and post-chemotherapy samples. Genes likely to be driver genes are colored according to mutation type: amplification, red text; homozygous deletion, blue text; point mutation, black text; and potentially relevant structural variants, purple text. Cancer type is specified: TN, triple negative; DCIS, ductal carcinoma *in situ*. Types of neo-adjuvant chemotherapy (NAC): Epi, epirubicin; T, docetaxel; P, paclitaxel; FEC, fluorouracil, epirubicin and cyclophosphamide. Panel **c** shows mock trees inferred from targeted capture data for samples with pre- and post-treatment samples. Six samples with no branching are not presented. In **b**, colors correspond to the tree branch directly above in **a**, and the area is proportional to the percentage of cells in that sample that contained the mutations in that branch. In **c**, branches are colored as stated above for genome data. (**d**) Pearson's correlation for heterogeneity estimates from whole-genome and targeted capture data.

Subclonal growth in multifocal cancer

For four cancers, we sequenced samples from two to five foci of a multifocal cancer. In each case, separate foci of disease were clonally related (**Fig. 3**). Within individual foci, we found that many private mutations had high variant allele fractions, indicating that during the growth of each focus complete 'clonal sweeps' had occurred in which a clone completely replaced all other tumor cells in that focus. In three of four cases, mutations private to a disease focus included known driver events: *BRCA2* and *CDKN2A* inactivation (**Fig. 3a**), *PTEN* point mutation (**Fig. 3c**,**d**) and *CDK6* amplification (**Fig. 3e**).

The complex intermixing of minor subclones seen in some unifocal tumors also existed within and between multiple foci of disease (**Fig. 3a**,**b**). By definition, lesions in multifocal breast cancer are separated by apparently normal breast tissue. Therefore, the fact that these distinct foci are clonally related shows that subclones in these developing tumors are capable of transiting considerable distances through normal breast tissue via the lymphatic, ductal or microcirculatory systems, as has been demonstrated in metastatic prostate cancers[35].

PD9694, a multifocal ER$^+$HER2$^-$ cancer with two macroscopic foci and several microscopic foci of invasive disease occurring within a large region of scattered ductal carcinoma *in situ* (DCIS), embodied a remarkable example of subclonal dissemination (**Fig. 3c**,**d**). Two distinct *PTEN* driver mutations appeared in the different regions; these mutations had evolved in parallel during the tumor's development (**Fig. 3c**) and were confined to disease with invasive potential (**Fig. 3d**). Critically, we detected one or the other of the *PTEN* mutations in discontiguous areas of microinvasive disease within predominant DCIS (**Fig. 3c**). The most plausible explanation for this is that the two

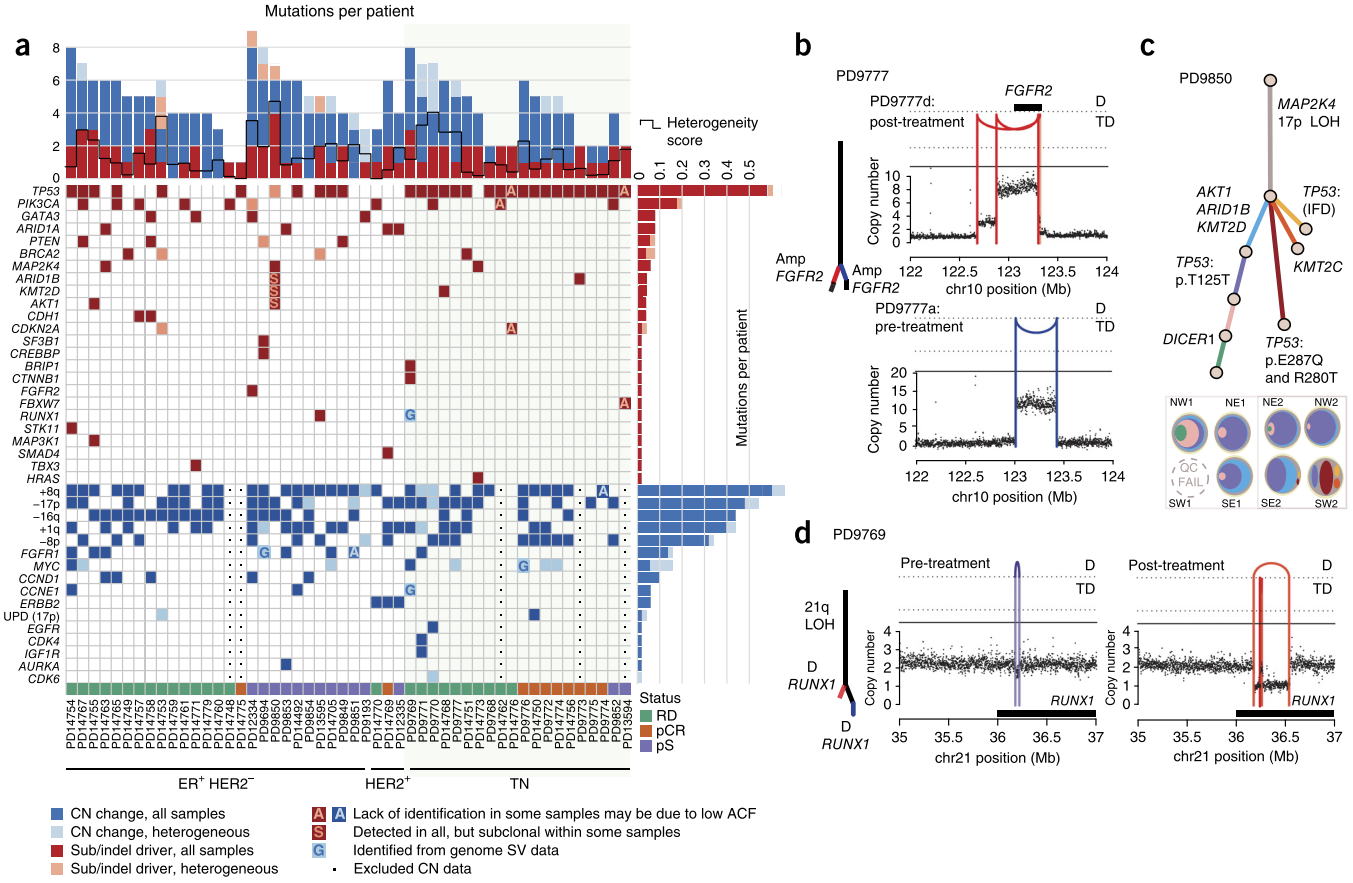

Figure 5 Subclonal driver mutations and parallel evolution. (**a**) Heat map of somatic driver mutations and copy-number (CN) changes identified from genomic sequencing of 50 tumors. Single-base substitutions (subs) and small insertions and deletions (indels) are denoted by dark red squares when detected in all associated samples from the tumor (omnipresent) and pink squares when present in fewer than all samples or when clearly subclonal. Omnipresent and heterogeneous copy-number changes are denoted by dark blue and light blue squares, respectively. TN, triple negative; SV, structural variant. (**b**–**d**) Three examples of parallel evolution; the fourth example is in **Figures 3c**,**d** and **4a** (PD9694). (**c**) One possible phylogenetic tree and sample subclonal compositions inferred from targeted capture data (as described in **Figs. 2** and **4c**) with *TP53* mutations arising on three branches. Coxcomb plots for PD9850 are in **Supplementary Figure 3**. Multiple independent episomal amplification events in *FGFR2* (**b**) and two independent deletions in *RUNX1* (**d**) were detected in two samples from the same cancer. In copy-number graphs (**b**,**d**) the black dots reflect the number of copies of genomic DNA from that specific locus, with a value greater than 2 reflecting a net gain and a value less than 2 reflecting a loss. Reconstructed rearrangement breakpoints are represented by colored lines according to whether they were detectable in pre- (purple) or post- (red) chemotherapy samples only. The type of event is indicated by the position of the arc joining the breakpoints. D, deletion; TD, tandem duplication; IFD, in-frame deletion; QC, quality control; LOH, loss of heterozygosity; RD, residual disease; pCR, pathological complete response; pS, primary surgery; ACF, aberrant cell fraction.

PTEN-null subclones disseminated intraductally within the DCIS, setting up several new, discrete foci of invasion.

Subclonal driver mutations in multifocal cancers were not restricted to point mutations; one sample showed a high-level *CDK6* amplification in one focus that was absent from the other focus (**Fig. 3e**). The *CDK6*-amplified focus showed only a partial response to neoadjuvant chemotherapy, whereas the other focus showed a complete pathological response.

Variable extent of subclonal heterogeneity in breast cancer

For the 50 cancers in cohorts 1 and 2, we assessed intratumoral heterogeneity in the targeted gene screen, taking into account fluctuations in normal cell contamination and sequence coverage. For 23 cancers, no significant difference in point mutations (**Supplementary Table 7**) existed across the different tumor subregions (**Fig. 1a** and **Supplementary Table 6**), although in 4 of these cancers there was heterogeneity in copy-number changes. For three cancers (PD14753, PD9850 and PD12334), we detected profound heterogeneity, exemplified by private mutations in most of the samples. Most cancers, however, had intermediate levels of intratumoral heterogeneity.

We created an index of heterogeneity on the basis of the discordance of mutation frequencies averaged across all possible pairs of samples from each cancer, after adjusting for normal cell contamination and differences in coverage (Online Methods and **Supplementary Fig. 5a**). Our data indicated no correlations among the level of heterogeneity and histology, ER expression status, grade, intratumoral lymphocyte infiltration and tumor Ki-67 score (**Supplementary Fig. 5b**–**h**). Heterogeneity in Ki-67 scores across samples did not correlate with our index of genomic heterogeneity (**Supplementary Fig. 5i**). We detected a trend toward a greater degree of heterogeneity with increasing patient age at diagnosis ($P = 0.05$, F-test) and larger tumor size ($P = 0.005$, F-test) among triple-negative cancers. Notably, the response to neoadjuvant chemotherapy (typically anthracycline-based regimens with or without a taxane) did not correlate with the extent of intratumoral heterogeneity among pretreatment samples in this cohort, albeit the sample size was limited ($P = 0.9$, F-test; **Supplementary Fig. 5e**).

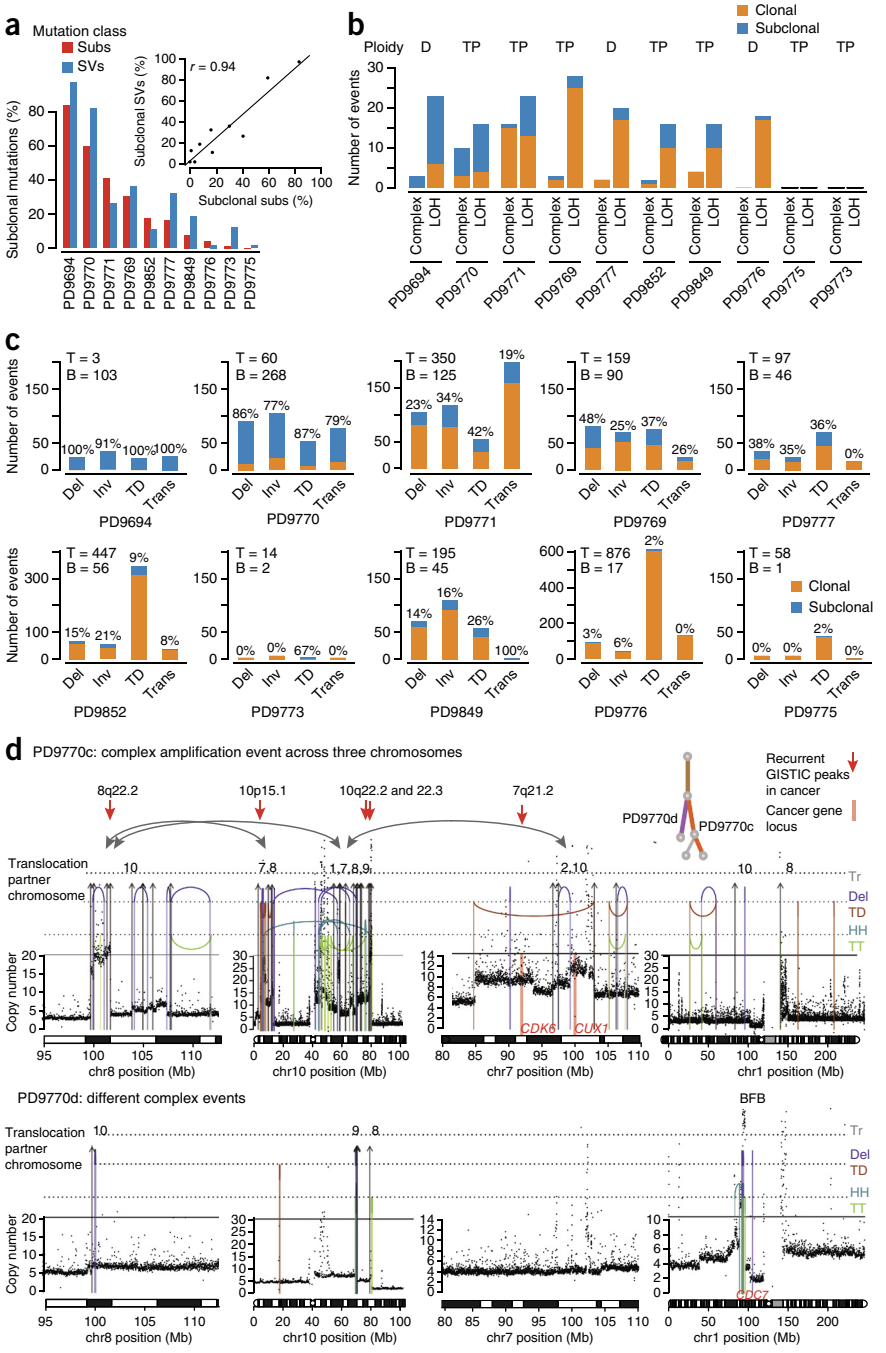

Figure 6 Structural variants shape cancer evolution. (**a**) Comparison of the proportions of substitutions (subs) and structural variants (SVs) that are subclonal in each cancer. Inset shows scatter plot and Pearson's correlation coefficient (*r*). (**b**) Clonal and subclonal complex rearrangements (as described in the **Supplementary Note**) and arm-level loss-of-heterozygosity (LOH) events. The average genome-wide ploidy is indicated. TP, tetraploid (four copies); D, diploid (two copies). (**c**) Breakdown of clonal and subclonal structural variants by category (Del, deletion; Inv, inversion; TD, tandem duplication; Trans, interchromosomal translocation). For each cancer, the total number of mutations assigned to the trunk (T) or branches (B) is indicated at the top left, and the proportion of each mutation type that was subclonal (i.e., within the branches) is presented above each bar. (**d**) Case PD9770: examples of two subclonal complex structural rearrangements arising on separate branches of the phylogenetic tree. In PD9770c, structural rearrangements link multiple regions of amplification across three chromosomes. Amplifications include multiple genomic regions that have been previously identified as recurrently amplified in cancers (red arrows); the locations of known oncogenes are marked by pink bars. In PD9770d these events are not seen, but a breakage fusion-bridge event (BFB) amplifies segments including the *CDC7* gene. Rearrangement types included interchromosomal translocations (Tr), deletions (Del; purple), tandem duplication (TD; brown), head-to-head inversions (HH; blue) and tail-to-tail inversions (TT; green).

To test whether genetic heterogeneity inferred from targeted capture data matches genome-wide distribution, we performed multiregion whole-genome sequencing on ten cancers (**Fig. 4**). For each, the thousands of somatic base substitutions enabled us to reconstruct phylogenetic trees and determine the subclonal composition of each sampled region (**Fig. 4** and **Supplementary Fig. 1**). In the targeted capture analysis, the extent of subclonal diversification varied markedly among tumors (**Fig. 4a**,**b**). We found good correlation between the branching time implied by whole-genome data and the heterogeneity score determined from targeted capture analysis of samples from the same cancer (**Fig. 4d**).

Resistant subclones may be unmasked by chemotherapy

For 18 cancers, we sequenced DNA from both diagnostic biopsies and residual invasive disease present after neoadjuvant chemotherapy. In six cancers (PD9768, PD9770, PD9771, PD9777, PD14748 and PD14757), mutations were subclonal (present in <100% of tumor cells) in both pre- and post-chemotherapy samples, indicating that some subclones persisted despite treatment (**Fig. 4a–c**). In five cancers (PD9768, PD9769, PD9770, PD9771 and PD9777), we identified a subclone in the post-chemotherapy residual tumor mass that was not evident in pre-chemotherapy samples (**Fig. 4a–c**). In these treatment-resistant subclones, potential driver mutations included amplifications of *CDK6* (PD9770), *FGFR2* and *MYC* (PD9777) and a deletion within *RUNX1* (PD9769).

Variants found only in post-chemotherapy samples could represent either mutations acquired during chemotherapy or mutations present in pre-existing subclones that were not sampled before therapy. For three cases, we had detailed phylogenies from samples before and after chemotherapy (PD9770, PD9777 and PD9771). In PD9777 and PD9771, the branching point of the post-treatment subclone predated the branching point inferred from pre-treatment samples only. Post-treatment and pre-treatment subclone branches were of similar lengths, suggesting a similar molecular age (**Fig. 4a**). Furthermore, similarity in mutational signature profiles (**Supplementary Fig. 6a**) in the pre- and post-treatment branches suggested a minimal contribution from chemotherapy-induced mutagenesis. Clones detected only in residual tumor mass after neoadjuvant chemotherapy are therefore likely to represent subclones in which most of the mutations were present before

treatment, a conclusion also reached on the basis of evolutionary simulations of breast cancers before and after chemotherapy[36].

Metastases can derive from subclones detectable in the primary tumor

For two cases, we studied whole genomes from primary tumor biopsies and a metastatic deposit. In the first cancer (PD9771), the lung metastasis and pre-chemotherapy biopsies all arose from a subclone that contained more than 800 base substitutions (**Fig. 4a**,**b**) and 43 structural variants. Notably, the residual disease sample, which also represented a chemotherapy-resistant population of cells, arose from a separate subclone. In the second cancer (PD9849), we found that an axillary lymph node metastasis arose from a defined subclonal lineage detected in the primary tumor (**Fig. 4a**,**b** and **Supplementary Fig. 1**), and not from the trunk of the phylogenetic tree.

This finding has clinical relevance: if metastatic disease arose from a very early branch of the phylogenetic tree, before all subclonal diversification within the primary tumor, treating actionable mutations that were subclonal in the primary tumor would not help prevent disease relapse. Although they need to be confirmed by larger studies, our results corroborate fluorescence *in situ* hybridization (FISH)-based studies of aneuploidy in metastatic breast cancer suggesting that metastases arise from subclones of the primary cancer[37].

Subclonal driver mutations and parallel evolution

Across the cohort, the majority of driver point mutations and copy-number changes were present in all lesions sequenced, which suggested that they occurred before the emergence of the cancer's most recent common ancestor (**Fig. 5a**). Although numbers of subclonal driver mutations are too low to allow for definitive conclusions about individual genes, and although phylogenetic reconstruction provides only the relative, not absolute, timing of driver mutations, it is clear that many of the common breast cancer genes can be mutated either early or late in disease. Driver mutations in *TP53*, *PIK3CA*, *PTEN*, *BRCA2* and *CDKN2A* were subclonal in some tumors in our study and fully clonal in others. Similarly, amplifications of *MYC*, *CDK6* and *FGFR1* sometimes occurred late in evolution. In 13 of 50 cancers, subclonal mutations affected genes that are potential targets of systemic therapies in clinical use or in development (**Supplementary Fig. 6b**).

We found four cancers (PD9694, PD9777, PD9769 and PD9850) with parallel evolution of driver mutations, including the case with convergent *PTEN* mutations discussed above (**Fig. 4a**,**b** and **Supplementary Fig. 6c**). In a triple-negative cancer (PD9777), we found three small, amplified episomal circles containing *FGFR2*, each present subclonally in the cancer and at variable proportions across the different samples. At least two of these circles must have arisen independently (**Fig. 5b**). In an $ER^{+}HER2^{-}$ cancer (PD9850), three separate subclonal lineages each carried different *TP53* driver mutations, including a recurrent silent mutation affecting *TP53* splicing (**Fig. 5c** and **Supplementary Fig. 6d**). In a triple-negative cancer (PD9769), we found distinct focal genomic rearrangements specifically deleting coding exons of *RUNX1* in two subclonal branches (**Fig. 5d** and **Supplementary Fig. 6e**). Three of the four examples of parallel evolution represent the second hit in a tumor-suppressor gene, with the first hit located on the trunk of the phylogenetic tree (**Supplementary Fig. 6c–e**).

Ongoing structural variation in subclonal diversification

We assessed the relative activity of mutational processes over time in the ten multiregion whole genomes (**Fig. 6b–d**). The proportion of structural variants that were subclonal broadly matched the proportion of subclonal substitutions ($r = 0.94$), although in some cancers, such as PD9777, late structural variants were the predominant driver of subclonal diversification (**Fig. 6b–e**).

Similar to point mutational signatures (**Supplementary Note** and **Supplementary Fig. 6b**), rearrangement processes active early in tumor evolution tended to continue later in disease (**Fig. 6c**,**d**). For some cancers, tandem duplications dominated the structural variant landscape, sometimes numbering in the hundreds, and these continued to accumulate late in disease (**Fig. 6d**). Complex chromosomal events were a frequent feature, being present in seven of ten cancers, and included four breakage-fusion-bridge cycles, a chromothripsis event followed by amplification, and complex amplification-associated rearrangements (**Fig. 6c**). In five tumors, complex events occurred both early and late in tumorigenesis and in some cases resulted in subclonal amplification of oncogenes (**Supplementary Table 8**), suggesting that catastrophic events can remodel the genome late in evolution and provide the phenotypic diversity upon which selection may operate.

DISCUSSION

Most breast cancers are diagnosed at an early stage and are considered curable. Once established, distant metastatic disease is incurable, meaning that the prevention of metastasis represents the best opportunity to improve breast cancer cure rates. We observed metastases that were derived from subclones in the primary cancer, a finding that emphasizes the importance of understanding the patterns, extent and nature of subclonal diversification in primary tumors.

We found variable degrees of genomic heterogeneity across breast cancers using targeted gene sequencing, which may underestimate subclones. This contrasts with the profound heterogeneity seen almost universally in clear cell renal cell carcinoma, where subclonal diversification occurs early after *VHL* mutation[18,19]. In non–small cell lung cancer, subclonal heterogeneity is less marked[38] and is minimal in early (stage IA–IIIA) tumors[31]. Kidney cancers are often large (>10 cm) when diagnosed, whereas breast and lung cancers are typically smaller. Indeed, we found a correlation between tumor size and degree of heterogeneity in triple-negative breast cancer. The direction of causality is unclear; it may be that tumors with profound heterogeneity grow to larger sizes, or it may be that once a tumor is beyond a certain size, complete clonal sweeps, where an especially fit clone expands to replace all other subclones in the tumor, become unlikely. In colon cancer, there is evidence that the latter theory can explain observed patterns of subclonal heterogeneity[17].

Transcriptome and histological studies have shown that breast cancer includes many subtypes[39–42], with distinct biological, prognostic and therapeutic implications. We found that subclonal heterogeneity can be present in all major immunohistological subgroups of breast cancer, but our 'all-comers' study design prevents us from drawing definitive conclusions about any particular subtype. Heterogeneity may explain cases of borderline ER and HER2 positivity[8,43], where survival benefits from anti-endocrine therapies may extend to cancers with nuclear ER staining in as few as 1% of tumor cells[8,44]. FISH-based studies have found that the heterogeneity of copy-number changes is predictive of the response to neoadjuvant chemotherapy[36], something we did not observe. Resolving this discrepancy will require studies focusing on specific molecular subtypes of breast cancer with larger sample sizes, potentially in the setting of clinical trials of neoadjuvant therapies.

Understanding subclonality is fundamental to improving cancer care but will require the prospective integration of genomics studies into clinical trials[45]. Important issues such as which subclones give rise to metastasis and the potential clinical benefits of treating subclonal actionable

mutations can be addressed, provided sample size, sample acquisition and sample analysis are carefully planned. Drug development is increasingly 'rational', based on an improved understanding of each tumor's individual biology; drug testing should follow this lead, incorporating the biology of cancer evolution into trial design and evaluation.

METHODS

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

Accession codes. The sequence data, aligned to the human reference genome (NCBI build37) using BWA, have been deposited in the European Genome-Phenome Archive with accession numbers EGAD00001000965 and EGAD00001000898.

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

ACKNOWLEDGMENTS

This work is supported by the Wellcome Trust. P.J.C. is a Wellcome Trust Senior Clinical Fellow (103858/Z/14/Z). L.R.Y., Y.L. and L.B.A. are funded by Wellcome Trust PhD fellowships. S.N.-Z. is funded by a Wellcome Trust Intermediate Clinical Research Fellowship (WT100183MA). P.V.L. is a postdoctoral researcher at the Research Foundation Flanders (FWO). Work within the project is supported by the Belgian Cancer Plan–Ministry of Health, the Breast Cancer Research Foundation, the Brussels Region, the Norwegian Cancer Society, the Norwegian Health Region West and the Bergen Research Foundation. Some samples referenced in this publication will be included in the Breast Cancer Genome Analyses for the International Cancer Genome Consortium (ICGC) Working Group led by the Wellcome Trust Sanger Institute. BASIS is a part of the ICGC working group and is funded by the European Community's Seventh Framework Programme (FP7/2010-2014) under grant agreement number 242006. This working group also encompasses a triple-negative breast cancer project funded by the Wellcome Trust (grant 077012/Z/05/Z) and a HER2$^+$ breast cancer project funded by Institut National du Cancer (INCa). We thank B. Leirvaag, D. Ekse, N.K. Duong and C. Eriksen for technical assistance. Research performed at Los Alamos National Laboratory was carried out under the auspices of the National Nuclear Security Administration of the US Department of Energy.

AUTHOR CONTRIBUTIONS

L.R.Y. and P.J.C. designed and directed the study and prepared the manuscript. L.R.Y. and M.G. performed analyses and prepared figures. S.K., T.A. and P.E.L. contributed to the study design and sample preparation for cohort 1. C.D., C.S., M.I. and M.M. contributed to the study design and sample preparation for cohort 2. D.C.W., P.V.L., G.G., H.D., Y.S.J., S. McLaren, M.R., S.N.-Z., A.B., D.G., A.M., K.R., J.H., D.J., M.R.S., Y.L. and L.B.A. contributed to analysis. S. Martin managed samples. A.L.R., D.L., H.K.H. and P.K.L. conducted histopathological assessment. P.-Y.A., D.V., B.J., A.G.-C. and A.F. performed DNA extraction. L.J.M. contributed to library preparation, PCR and gel electrophoresis.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
2. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
3. Shah, S.P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
4. Meric-Bernstam, F. *et al.* Concordance of genomic alterations between primary and recurrent breast cancer. *Mol. Cancer Ther.* **13**, 1382–1389 (2014).
5. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
6. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
7. Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* **4**, 1116–1130 (2013).
8. Hammond, M.E. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J. Clin. Oncol.* **28**, 2784–2795 (2010).
9. Seol, H. *et al.* Intratumoral heterogeneity of HER2 gene amplification in breast cancer: its clinicopathological significance. *Mod. Pathol.* **25**, 938–948 (2012).
10. Simon, R. & Roychowdhury, S. Implementing personalized cancer genomics in clinical trials. *Nat. Rev. Drug Discov.* **12**, 358–369 (2013).
11. Sleijfer, S., Bogaerts, J. & Siu, L.L. Designing transformative clinical trials in the cancer genome era. *J. Clin. Oncol.* **31**, 1834–1841 (2013).
12. Moskaluk, C.A., Hruban, R.H. & Kern, S.E. p16 and K-ras gene mutations in the intraductal precursors of human pancreatic adenocarcinoma. *Cancer Res.* **57**, 2140–2143 (1997).
13. Powell, S.M. *et al.* APC mutations occur early during colorectal tumorigenesis. *Nature* **359**, 235–237 (1992).
14. Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616–3627, 3699 (2013).
15. Green, M.R. *et al.* Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. *Blood* **121**, 1604–1611 (2013).
16. Yachida, S. & Iacobuzio-Donahue, C.A. Evolution and dynamics of pancreatic cancer progression. *Oncogene* **32**, 5253–5260 (2013).
17. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
18. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
19. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
20. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
21. Cooper, C.S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat. Genet.* **47**, 367–372 (2015).
22. Santarius, T., Shipley, J., Brewer, D., Stratton, M.R. & Cooper, C.S. A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010).
23. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
24. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
25. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
26. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
27. Stephens, P.J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
28. Ellis, M.J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353–360 (2012).
29. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
30. Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
31. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256–259 (2014).
32. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. USA* **107**, 16910–16915 (2010).
33. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
34. Balko, J.M. *et al.* Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets. *Cancer Discov.* **4**, 232–245 (2014).
35. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
36. Almendro, V. *et al.* Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Rep.* **6**, 514–527 (2014).
37. Almendro, V. *et al.* Genetic and phenotypic diversity in breast tumor metastases. *Cancer Res.* **74**, 1338–1348 (2014).
38. de Bruin, E.C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
39. Ali, H.R. *et al.* Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.* **15**, 431 (2014).
40. Nielsen, T.O. *et al.* Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin. Cancer Res.* **10**, 5367–5374 (2004).
41. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98**, 10869–10874 (2001).
42. Perou, C.M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
43. Rakha, E.A. & Ellis, I.O. Breast cancer: updated guideline recommendations for HER2 testing. *Nat. Rev. Clin. Oncol.* **11**, 8–9 (2014).
44. Early Breast Cancer Trialists' Collaborative Group. *et al.* Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet* **378**, 771–784 (2011).
45. Yuan, Y. *et al.* Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* **32**, 644–652 (2014).

ONLINE METHODS

Sample acquisition. In this exploratory study, we analyzed a total of 303 multiregion breast cancer tumor samples from 50 subjects and a matched normal sample derived from blood ($n = 49$) or adjacent normal breast tissue ($n = 1$) for each subject. Cohort 1 consisted of 98 samples from 12 cancers (average of 8.2 samples per cancer; range, 8–10). Cohort 2 comprised 205 samples from 38 cancers (average of 5.4 samples per cancer; range, 2–21) (**Supplementary Table 1**). All subjects were female, and in cohorts 1 and 2 the average ages at diagnosis were 67 years (range, 44–90 years) and 49 years (range, 29–67 years), respectively (**Supplementary Table 1**). Sample collection and management complied with local institutional review board approvals; details are provided in the **Supplementary Note**. Samples in cohort 1 were from 12 patients undergoing primary surgery who provided informed consent to participate in a prospective study. They encompassed 12 geographically predetermined tissue samples (15–20 mg) obtained with a 14G Tru-cut needle from fresh surgical specimens according to the map in **Figure 1b**. In cohort 2 we studied de-identified residual tissue samples collected during routine clinical care. Samples in cohort 1 were derived from fresh needle biopsy specimens ($n = 98$), whereas those in cohort 2 were from a combination of diagnostic tumor biopsies ($n = 95$) and surgical-specimen tissue blocks ($n = 110$), either FFPE ($n = 104$) or fresh-frozen at acquisition ($n = 101$) (**Supplementary Table 1**). Experienced local pathologists performed histopathological review of all primary tumors, including immunohistochemistry (IHC) for ER and PgR Allred scores and HER2 status, with FISH confirmation for HER2 IHC scores of 2+ or 3+ (**Supplementary Table 1**). Pathologists assessed intratumoral and stromal lymphocytes according to the criteria previously described[46] and Ki67 staining as described in the **Supplementary Note** and presented in **Supplementary Table 1**. Sample sizes were chosen to ensure that genes mutated in >10% of tumors were sampled five times in the cohort on average.

DNA extraction. We performed DNA extraction from serial thick sections cut from tumor tissue samples. Pathologist-guided macrodissection ensured tumor-cell enrichment in cases where the invasive tumor content was estimated to be less than 50% of cells. For two cases with multifocal disease (PD9193 and PD9694) we used histopathologically guided needle dissection of FFPE samples. We isolated tumor DNA from fresh or fresh-frozen tissues using the DNeasy Blood and Tissue Kit (QIAGEN) and from FFPE tissues using a QIAamp DNA FFPE Tissue Kit (QIAGEN) or a DNA nanoPurify kit (Argylla Technologies). We used a QIAamp DNA Blood Maxi Kit, QIAamp DNA Mini Kit or DNeasy Blood and Tissue Kit (all from QIAGEN) to isolate DNA from whole blood. In all cases we followed the manufacturer's recommended protocol.

Genomic sequencing. We created targeted capture pulldown (average insert size, 150 bp) and genome-wide shotgun (insert size, 300–600 bp) libraries from native DNA using previously described workflows[1,14,47] (details in **Supplementary Note**) and generated paired-end sequence data (75 bp and 100 bp, respectively) using Illumina HiSeq machines. The sequence data, aligned to the human reference genome (NCBI build 37) using BWA[48], have been deposited in the European Genome-Phenome Archive with accession numbers EGAD00001000965 and EGAD00001000898. In the custom targeted capture experiment, we sequenced 290 tumor samples from 50 subjects to a mean target coverage of 160×, with 63% of exonic regions achieving ≥100-fold coverage (**Supplementary Table 2**). We sequenced to whole-genome level 29 tumor and 13 matched normal samples with average sequence coverages of 40-fold and 31-fold, respectively (**Supplementary Table 2**). We used both sequencing approaches for 16 tumor samples.

We used two in-house cancer gene panels (CGP versions v1 and v2) designed to pull down a selection of genes (454 and 360 genes, respectively) that are known or suspected to have a role in cancer (**Supplementary Table 3**). The panel targets genes from the Cancer Gene Census (COSMIC)[24], genes recurrently amplified or overexpressed in cancer[22,23] and candidate cancer genes such as kinases from the MAP kinase signaling pathway. All genes in CGP v2 are also in CGP v1, and only genes present in both CGP v1 and CGP v2 are presented here. We used a custom RNA bait design according to the manufacturer's guidelines (SureSelect, Agilent, UK) to create designs of approximately 2 Mb in size. The data from 63 and 240 tumor samples were derived from CGP v1 and CGP v2, respectively (**Supplementary Table 1**).

Somatic-mutation calling. Comprehensive lists of all somatic substitutions, small indels and structural variants including variant allele frequencies from both whole-genome and targeted capture analysis are available for download at ftp://ftp.sanger.ac.uk/pub/cancer/YatesEtAl/. All high-confidence mutation calls within the scope of the cancer gene panel are presented in **Supplementary Table 7**. Coding substitution and indel calls and structural variants with potential oncogenic effects, identified in whole-genome data, are summarized in **Supplementary Table 8**. Mutation-calling algorithms used in the analysis are freely available at https://github.com/cancerit/ and are described in the **Supplementary Note**.

Validation approaches. In explorations of heterogeneity, determining when a mutation is absent is at least as important as determining when the mutation is present. To address this, we performed validation using custom pulldown and sequencing of mutations identified in any sample (Illumina HiSeq or MiSeq) in all related samples from the same cancer. We enriched the validation experiment with mutations that appeared to be heterogeneous (from the branches of phylogenetic trees) or that defied the consensus tree. Across the 39 whole-genome tumor samples, we selected more than 2,000 somatic substitution locations for validation and created a 473-kb custom capture probe design with Agilent Technologies' freely available online software Sure Select Design Wizard using high-stringency repeat masking, a tiling density of 2× and balanced boosting. We created DNA capture (paired-end; average insert size, 150 bp) libraries using native DNA where resources permitted or, if necessary, using whole-genome amplification. We sequenced multiplexed libraries to an average depth of 265× using the Illumina MiSeq platform. When a variant was called as present in the tumor sample and absent in the matched normal sample in both discovery and validation experiments for one or more related samples, we reported it as validated somatic (details of validation calls are in the **Supplementary Note**). On the basis of these criteria, 99% (2,217 out of 2,235) of substitutions were validated as somatic (**Supplementary Table 4**). The remaining calls were not detectable in any relevant sample's validation data (false positive, $n = 10$) or were detected in the matched normal at validation (germline, $n = 7$). We confirmed the absence of 1,301 out of 1,683 (77%) mutations. Overall concordance between the two experiments (true positives and true negatives versus all validation calls) was 90% (5,003 out of 5,527). The overall level of concordance for the targeted capture experiment was higher, with consistency between 189 out of 191 validation and discovery calls (99% concordance), which probably reflects the higher coverage in this experiment (**Supplementary Table 4** and **Supplementary Note**).

Variant annotation. To identify likely driver events, we used published literature to identify the genes most likely to contribute to breast cancer oncogenesis. For each individual mutation that occurred in one of 45 high-confidence breast cancer genes, we assigned a likely oncogenic status. Mutations presumed oncogenic were those that met any of the following criteria:

i. Canonical oncogenic mutations in recurrent hotspots.
ii. Recurrent mutations in a known oncogene: ≥2 confirmed non-synonymous or in-frame deletions; somatic mutations have previously been confirmed at this locus in COSMIC.
iii. Likely damaging events in a known tumor suppressor: truncating, frameshift, essential splice variant or in a mutation hotspot (≥2 somatic mutations) *or* synonymous mutation in a known recurrent splice-site hotspot[49].

Possible oncogenic mutations included previously unreported variants in a high-confidence breast cancer gene that occur within three amino acids of ≥2 confirmed somatic mutations or truncating events in medium-confidence tumor suppressors (defined as having a known tumor-suppressor role in cancers other than breast cancer). All other nonsynonymous mutations were assigned a status of 'unknown relevance'. On the basis of these criteria. the 260 mutations

doi:10.1038/nm.3886

identified across the data set were annotated as follows: 87 oncogenic, 8 possible oncogenic, 124 of unknown oncogenicity and 41 nononcogenic (synonymous) (**Supplementary Table 3**).

Copy-number analysis. Likely driver copy-number changes are reported for individual samples in the targeted gene capture experiment in **Supplementary Table 7** and for whole-genome samples in **Supplementary Table 8**. Segmental copy-number information was derived for each of the 29 tumor samples for which we had whole-genome next-generation sequencing (NGS) data using the ASCAT (allele-specific copy-number analysis) algorithm of tumors as previously described[32]. The algorithm simultaneously determines and utilizes aberrant cell fraction and ploidy estimates to determine allele-specific copy numbers from NGS data. A segment is considered amplified if it is present at more than twice the estimated average ploidy across the whole genome. Homozygous deletions were identified as segments where the total copy number was zero (subclonal homozygous deletions if the copy number was <1). Visual inspection of copy-number transitions and reconstructed associated rearrangement breakpoints were used to validate driver copy-number events as described in the **Supplementary Note**.

In the targeted capture experiment we evaluated copy number using libraries from the ASCAT algorithm and used LogR and B-allele frequency values to identify five of the most frequent arm-level copy-number changes in breast cancer—16p and 17p losses and 1q, 8q and 16p gains[32], and amplification of 12 genes frequently identified as amplified in breast cancers (*FGFR1*, *MYC*, *CCND1*, *CCND3*, *CCNE1*, *CDK4*, *CDK6*, *IGF1R*, *ZNF217*, *AURKA*, *EGFR* and *ERBB2*)[33,34]. Details are provided in the **Supplementary Note**, along with details of the targeted capture copy-number validation approach using multiplex ligation–dependent probe amplification (**Supplementary Table 4**).

Statistical and informatics approaches. We performed statistical analysis and produced graphics using R version 3.0.1 (http://www.R-project.org/). The "stars()" function was used to generate coxcomb plots. Other packages used were RColorBrewer, xlxs, lme4 and mgcv, as well as packages from bioconductor[50]. All hypothesis tests performed in the study were two-sided where appropriate.

Measuring heterogeneity in targeted capture data. *Estimating variant-allele frequencies and confidence intervals.* In a sequencing experiment with finite coverage, one is likely to completely miss mutations present at low variant-allele frequency (VAF). Our point estimate of the VAF is

$$\mathrm{VAF} = x/n$$

where x is the observed number of reads reporting the variant and n is the coverage. This is the maximum likelihood (ML) estimated under a simple binomial sampling model,

$$x \sim \mathrm{Bin}(n, \mathrm{VAF})$$

It is, however, also possible to observe x reads if the true VAF is greater than the ML estimate. To determine the maximum allele frequency compatible with our data, we defined a one-sided 95% confidence interval (CI) as that VAF beyond which the probability of observing x or fewer reads was less than 5%.

$$\mathrm{CI} = \arg\max_{\mathrm{VAF}} [F(x,n,\mathrm{VAF}) : F \le 0.95]$$

where F denotes the cumulative density function of the binomial distribution.

Testing for the presence or absence of mutations. For the purposes of determining whether an individual mutation was present or absent, as displayed in the driver mutation heat map (**Fig. 5**), we determined the presence of each mutation in a dichotomous fashion in each sample. A mutation was considered (i) present if found at a positive VAF (VAF > 0); (ii) indeterminate in cases with no detectable VAF = 0 but 95% CIs spanning >5% allele frequency, as the absence of such mutations could not be ruled out with sufficient certainty; or (iii) absent if undetectable (VAF = 0) and the 95% CI was <5%. Only in such cases were mutations reported as heterogeneous.

Measuring heterogeneity. In addition to being affected by sampling fluctuations, the observed VAF is confounded by the tumor-cell fraction T. Any comparison of VAF between samples should therefore normalize for T. A low T will rescale all observed VAFs by a factor of T. Thus, for the computation of the heterogeneity index, we used the average VAF in sample j as an estimate of T_j and rescaled all VAF values by $1/T_j$. To quantify and compare heterogeneity between cancers, we calculated a continuous index of heterogeneity across all data from the targeted gene screen. This index measures the average discordance of mutation frequencies between any two pairs of samples after adjusting for the tumor-cell content. The distance D in the T-adjusted VAF of gene i between samples j and k is computed as

$$D_{ijk} = \min(|\,\mathrm{VAF}_{ij}/T_j - \mathrm{VAF}_{ik}/T_k\,|, |\,\mathrm{CI}_{ij}/T_j - \mathrm{VAF}_{ik}/T_k\,|, |\,\mathrm{VAF}_{ij}/T_j - \mathrm{CI}_{ik}/T_k\,|)$$

Note that the above calculation uses the distance to the CI if that is closer to the observed VAF.

The heterogeneity index (HET) was then defined as the average distance between all genes and samples.

$$\mathrm{HET} = (\textstyle\sum_i \sum_j \sum_{k<j} D_{ijk})/(gb(b-1)/2)$$

where g is the number of genes and b is the number of samples. A heterogeneity value of 0 indicates perfect concordance of all samples, and a value of 1 corresponds to a situation in which one sample has one additional fully clonal mutation. The heterogeneity index shows a strong inverse correlation with the branch time derived from whole-genome sequencing data in the sense that late-branching tumors display higher levels of geographic heterogeneity ($\rho = -0.73$; **Supplementary Fig. 3j**).

Testing for heterogeneity. We used generalized linear models (GLMs) with an overdispersed binomial family to test whether the observed differences in VAFs between genes and samples in a given cancer can be explained by sampling fluctuations and differences in tumor cellularity alone. In a binomial GLM, the expected count of mutation i in sample j is given by

$$E[X_{ij}] = f(\alpha_i + \beta_j + \gamma_{ij})$$

where f is the inverse logit function. Here α_i sets the average frequency of each gene i, and β_j is the common factor by which the gene frequencies change in sample j as a result of changes in tumor cellularity across samples. The parameter γ_{ij} reflects the deviation of gene i in sample j from the trend imposed by the gene-specific allele frequency and cellularity in sample j. Note that there can be maximally $(g-1)\times(b-1)$ values of γ_{ij} because of the $g+b$ shared factors α_i and β_j; the total number of observations is $g \times b$.

An overall test for heterogeneity in a given cancer then involves testing whether all values of γ_{ij} are zero or whether there is variation in any gene. This can be achieved by means of a likelihood-ratio test (LRT) with $(g-1)\times(b-1)$ degrees of freedom. We used the following R commands to derive a P value for each sample:

```
# x is a vector of variant allele counts for all lesions and biopsies;
  the length of x is g*b
# n is a vector of the corresponding coverage
# genes is a factor() determining which gene x and n refer to
# biopsies is a factor() determining the biopsy
y <- cbind(x, n-x)
fit1 <- glm(y ~genes + biopsies -1 + genes:biopsies, family =
quasibinomial)
Ppatient <- anova(fit1, test = "LRT", dispersion = 1.5)[4,5]
```

P values for each patient were subsequently corrected for multiple testing using the Benjamini-Hochberg procedure. P and Q values for each patient are reported in **Supplementary Table 6.** Similarly, we tested for variation in a particular gene i using an interaction term for that gene only:

```
fit.gene < - glm(y ~ genes + biopsies − 1 + (genes = = gene): biopsies, family = quasibinomial)
glm.p.value < - anova(fit.gene, test = "LRT", dispersion = 1.5) [4.5]
```

Additionally, we used GLMs with random effects, implemented in mgcv:: gam(), to compute estimates of the variation of allele frequencies across samples. Gene-wise *P* values from GLMs and generalized additive models (GAMs) are also listed in **Supplementary Table 6** (glm.p.value, gam. stddev, gam.p.value).

Testing of clinical associations. Possible associations between clinical or pathological factors and genetic heterogeneity as a response are fitted using R's lm() function. *F*-tests for overall association are then computed using the anova() command.

Basic principles of phylogenetic-tree construction. To model the subclonal structure for ten patients with multisample whole-genome sequencing data, we employed a number of bioinformatic and deductive-reasoning approaches. The intellectual framework for our methods has been previously described[1], and this approach has been extended and reinterpreted by many others since, using the original data[51,52]. All the conclusions we derived followed from three basic principles that also underlie the 'mock' trees derived from targeted capture data: (i) cancer cells divide by asexual reproduction, (ii) the exact same mutation does not occur more than once during the evolution of the cancer (note that this 'infinitely many sites' assumption is potentially not true for hot-spot mutations in, say, *PIK3CA* but will be true for virtually all passenger mutations, given the size of the genome and the relative paucity of somatic mutations) and (iii) sequencing reads from massively parallel sequencing data are a random sample from the alleles present in the DNA.

The approach used in this study followed three main steps: (i) identification of large-scale subclonal copy-number changes using the Battenberg algorithm as previously described[1] (code is publically available at https://github.com/cancerit/cgpBattenberg), (ii) clustering of subclonal somatic substitutions in whole-genome data using a Bayesian Dirichlet process in multiple dimensions across related samples as previously described[47] and (iii) hierachical clustering across multiple samples by applying the 'pigeonhole principle' (PHP). Next, we performed validation of mutations in individual branches by targeted pulldown and validation of tree structures by independent clustering of indels and targeted pulldown substitutions following steps i–iii (**Supplementary Fig. 1**).

Each step is described in further detail below, and all potential solutions and reasoning for individual patients are represented in **Supplementary Figure 1**, with individual cases discussed in the **Supplementary Note**. Branch length, cluster sizes and poster CIs are provided in **Supplementary Table 5**.

Whole-genome data: mutation copy number and cancer-cell fraction. For each mutation, we calculated the mutation copy number as previously described[27], using the mutant allele burden, the aberrant cell fraction and the locus-specific copy number in the tumor and matched normal from ASCAT[32]. The mutation copy number reflects the percentage of tumor cells in a sample carrying that mutation and enables cross-comparison of the mutation in related samples despite differences in tumor purity and/or copy-number profiles, as previously demonstrated[47].

Mutations present on multiple copies of a chromosomal segment will have a mutation copy number greater than 1. In order for a mutation to be grouped according to the percentage of cells containing it, the number of chromosomes carrying the mutation must be determined. For all mutations in amplified regions with a major allele copy number of *C*, the observed fraction of mutated reads is compared to the expected fraction of mutated reads resulting from a mutation present on 1,2,3,…,*C* copies, assuming a binomial distribution. The fraction of cancer cells reporting the mutation, or the cancer-cell fraction, is then determined as the mutation copy number divided by the value of *C* with the maximum likelihood. Mutations are determined to be clonal if reported in ~100% of tumor cells and subclonal if present in significantly less than 100% of cells.

For the purpose of comparing multiple related samples, we excluded mutations from clustering analysis when they occurred in a region of different copy number between samples and where the absence or altered copy number explained the loss or different allele burden in the related samples. This approach is essential to reduce overestimation of intersample heterogeneity. Large-scale losses, including those at the arm or whole-chromosome level, are frequent during evolution (**Fig. 6b**). Allelic loss can therefore be accompanied by a loss of large numbers of point mutations and indels, which could be misinterpreted as gained events (i.e., ongoing evolution) in related samples. We placed the few individual driver mutations that occurred in regions of differential copy-number state on the reconstructed tree *post hoc*. This then allowed them to be included in the temporal ordering inference.

Mutational clustering. For individual samples, we inferred the number of subclones and the fraction of cells within each subclone using a previously described Bayesian Dirichlet process (DP) to cluster mutations according to their cancer-cell fraction[1,47]. We extended this process into multiple dimensions for the ten patients with multiple related samples, where the numbers of mutant reads obtained from multiple related samples were modeled as independent binomial distributions. Clusters were identified as local peaks in the posterior mutation density obtained from the DP. For each cluster, a region representing a 'basin of attraction' was defined by a set of planes running through the point of minimum density between each pair of cluster positions. Mutations were assigned to the cluster in the basin of attraction in which they were most likely to fall using posterior probabilities from the DP. The R code required to sample the clustering of mutations from a Dirichlet process and to make density plots of the clustering for each pair of samples is presented as a **Supplementary Source Code**.

Hierarchical ordering of mutation clusters using the PHP. To determine the most likely phylogenetic tree, we applied the PHP to determine the order in which mutational clusters arose in time and in relation to one another[1]. This principle operates upon the premise that if the fraction of cells reporting two different mutations adds up to >100%, then at least one tumor cell must contain both mutations. By the same principle, one can determine whether clusters of mutations are collinear (i.e., on the same branch of the phylogenetic tree), and often the temporal order in which they arose. For all clonally related samples, the same underlying phylogenetic tree must exist. This imparts greater stringency to the inferred ordering of subclonal clusters—the PHP must be fulfilled in all individual related samples, *and* the ordering of events cannot be contradictory across related samples.

We attempted to reconstruct phylogenetic trees from the whole-genome discovery substitution data using all clusters that were estimated to contain at least 150 substitutions or ≥2% of all clustered substitutions. To reflect the lower overall numbers in validation and indel data, this threshold requirement was set at 5%. In tree construction, the percentage of all mutations in a cluster determined the relative branch length. In an individual sample, the cancer-cell fraction of a given cluster *X* was the fraction of tumor cells reporting the mutations in cluster *X*. Credible intervals for the cancer-cell fraction are typically small, reflecting high numbers of mutations in most clusters. We allowed 5% variation in either direction for the assigned cluster sizes when determining ordering.

In seven out of ten cases we derived a single, unambiguous phylogenetic tree solution from the whole-genome discovery data (**Supplementary Fig. 1**). In two cases we identified one or more alternative trees using the discovery data (PD9773 and PD9694), whereas in another case (PD9777) a solution could be deconvoluted only with revised VAFs from high-depth validation data. The validation clustering data identified additional tree branches in two cases (PD9775 and PD9849). In cases where there was uncertainty regarding the position or size of a branch, the relevant branch(es) is 'faded out' in **Figure 4a**. Our approaches are described in detail in the **Supplementary Note**, where we focus on patient PD9694 and cases where solutions were less clear-cut.

Mutational signature analysis. Mutational signatures are detected in two independent ways: (i) *de novo* extraction on the basis of somatic substitutions and their immediate sequence context, and (ii) refitting of previously identified consensus signatures of mutational processes. We used a previously developed theoretical model and its corresponding computational framework to perform *de novo* extraction[53]. Details of mutational signature analysis can be found in the **Supplementary Note**.

46. Denkert, C. *et al.* Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. *J. Clin. Oncol.* **28**, 105–113 (2010).
47. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).

doi:10.1038/nm.3886

48. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
49. Supek, F., Minana, B., Valcarcel, J., Gabaldon, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324–1335 (2014).
50. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
51. Fischer, A., Vazquez-Garcia, I., Illingworth, C.J. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7**, 1740–1752 (2014).
52. Oesper, L., Mahmoody, A. & Raphael, B.J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14**, R80 (2013).
53. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).