Supplementary Materials

Subclonal diversification of primary breast cancer revealed by multiregion sequencing

Authors:

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) and Peter J Campbell(1)

Institutions:

(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, 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, United States of America
(9) Department of Pathology, Haukeland University Hospital, Bergen, Norway
(10) The Gade Laboratory for Pathology, Haukeland University Hospital, Bergen, Norway
(11) Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, USA
(12) Brigham and Women's Hospital, Harvard Medical School, Boston, USA

Address for correspondence:

Dr Peter J Campbell,

e-mail: pc8@sanger.ac.uk

Contents

Please visit ftp://ftp.sanger.ac.uk/pub/cancer/YatesEtAl/ for (i) Genome-wide substitution, indel and structural variant mutation calls; (ii) Targeted capture mutation calls; (iii) targeted capture biopsy plots and heatmaps for each individual case.

SUPPLEMENTARY TABLE LEGENDS

Table 1. Patient and sample characteristics (**a**) Primary diagnostic tumor and treatment information for all 50 patients. Data for grade, ER, PgR, HER2 status and ki67 estimates all relate to the pathological scores determined on the diagnostic, un-treated samples. (**b**) Sample types and numbers sequenced per patient. (**c**) Sample specific characteristics. (**e**) Independent ki67 assessment on the 4 core biopsy samples acquired from cohort 1 patients.

Table 2. Sequencing coverage. (**a**) Genome-wide sequence coverage and the aberrant cell fraction estimate derived from NGS data using Battenberg and ASCAT algorithms. (**b**) Targeted capture sequence coverage, is presented as the average target coverage and the percentage of all targets at each minimum level of coverage.

Table 3. Annotation of potential driver genes. (**a**) Genes within the cancer gene panel version 2 are annotated as of high, medium or low confidence driver genes in breast cancer based upon previous reports of recurrent point mutations. (**b**) Frequent arm level events and likely drivers of breast cancer recurrently altered through amplification or homozygous deletion that were specifically examined in our dataset.

Table 4. Validation data. (**a**) Validation of substitutions and indels within the targeted capture experiment and (**b**) absolute allele counts fro validation experiments including genes in regions that failed to pull down. (**c**) Validation of copy number calls using multiplex ligation dependent probe amplification. (**d**) Summary of validation rates for substitutions and indels in whole genome data using a custom capture pulldown and targeted re-sequencing. (**e**) Rearrangement breakpoint validation using

PCR and gel electrophoresis and visual inspection for breakpoint associated copy number changes. All individual variants with validation raw data counts may be downloaded from ftp://ftp.sanger.ac.uk/pub/cancer/YatesEtAl/.

Table 5. Mutation clusters, relates to **Figure 4** and **Supplementary Figure 1**. (**a–j**) Multi-dimensional dirichlet clustering of somatic point mutation calls from discovery and validation experiments is presented for each individual cancer and each sample within it.

Table 6. Heterogeneity scores. (**a**) For each cancer a heterogeneity score was calculated. Generalized linear models (glm's) with an overdispersed binomial family were used to test whether the observed differences in variant allele frequencies between genes and biopsies in a given patient can be explained by sampling fluctuations and differences in tumor cellularity alone. In subsequent sheets, for each cancer a heterogeneity score is presented for each individual mutation.

Table 7. Mutation and copy number calls from targeted capture data. (a) Frequency of individual mutations, amplifications and arm level copy number gains and losses across the cohort. (**b**) For each cancer, the proportion of samples containing each detected mutation. (**c**) Details of each point mutation detected within the scope of the cancer gene panel. (**d**) Breakdown of mutation types within each gene in the cancer gene panel. (**e**) Summary of variant allele fractions of each substitution and indel call (number of sequence reads reporting the mutant allele/ total number of reads covering that locus) for all related samples from individual cancers are presented on a case-by-case basis in subsequent sheets by name. Also

see heatmaps and biopsy plot visual representations of data at ftp://ftp.sanger.ac.uk/pub/cancer/YatesEtAl/.

Table 8. Coding mutations and oncogenic copy number events from whole genome data. (**a**) Somatic substitutions and (**b**) indels in coding regions and essential splice sites identified in the whole genome data from primary breast cancers are presented. Whether the mutation is likely to represent a driver mutation is annotated in the 'Driver Mutation' column. (**c**) Amplification and homozygous deletions in probable cancer genes (red text=amplification, blue text=homozygous deletion). (**d**) Possible driver events arising as a consequence of structural variant breakpoints falling within driver genes. (**e**) Likely deleterious mutations in *BRCA1* and *BRCA2*. See **Figure 4** where likely driver mutations are assigned retrospectively to reconstructed phylogenetic trees.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Phylogenetic tree construction. For each cancer subjected to multi-region whole genome sequencing (n = 10) the process of phylogenetic tree construction is presented. (**a**) Consensus phylogenetic trees as presented in **Figure 4a** and alternative tree solutions (*A*). Branch numbers relate to high confidence clusters labeled in (**b**). The underlying methodology uses a multi-dimensional Bayesian Dirichlet process to cluster somatic substitution data as presented in **Supplementary Table 5** (alongside estimates of the number of mutations in each cluster and 95% credible intervals). For each cancer, density plot(s) of cancer cell fractions are presented (**b**). Each cluster represents a group of mutations at a similar variant allele frequency, corrected for locus specific copy number and tumor cellularity, that are likely to co-exist in the same subclone. The number of mutations in each cluster estimates the length of the relevant branch. The position of the cluster reflects the fraction of cancer cells that contain that cluster of mutations – see oval plots in **Figure 4b**. Trees are predominantly constructed from the clustering of discovery substitution data and where indicated using clusters from the validation data (*v*). We only attempt to include high confidence clusters in tree construction; >= 2% of all mutations or >= 150 mutations for genome wide experiments (or >= 5% of mutations in the validation experiments). (**c**) Validation substitution and indel data for case PD9694: Colored rings identify clusters that relate to the phylogenetic tree branches. Numbers relate to the total number of mutations. Branch lengths are not to scale for validation data that is enriched for heterogeneous mutations. Variations in tree construction are described in the **Supplementary Note**.

Supplementary Figure 2. FFPE and fresh-frozen tissue sequencing. A comparison of mutant allele percentages in multiple fresh frozen and formalin fixed paraffin embedded (FFPE) (grey box) samples from PD9694 (**a**) and PD9193 (**b**) is presented. All but one mutation that is private to the FFPE samples was detected in more than one sample (including whole genome sample PD9694c), which is inconsistent with sporadic fixative related artifacts.

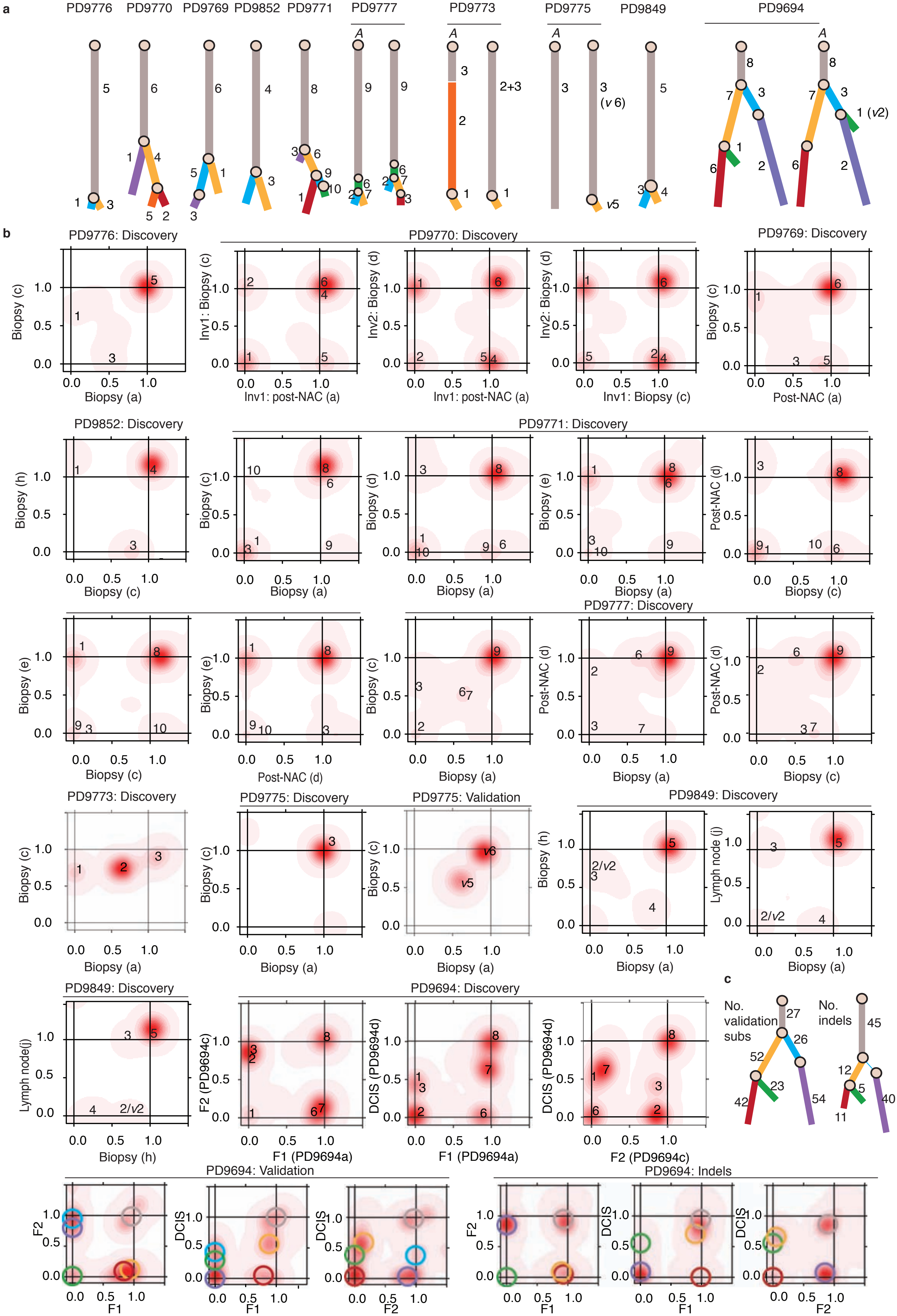
Supplementary Figure 3. Geographical patterns of heterogeneity, relates to **Figure 2.** Somatic mutational genotypes (as described in **Figure 2** legend**)** of individual biopsies are overlaid on the sample schema (as described in **Figure 1b** legend). Significant heterogeneity within individual cancers is identified as described in **Figure 2** legend. Notably q-values are derived from point mutation data only and do not take into account copy number changes that are clearly heterogeneous in some cases (*C*). '*' denotes q < 0.0001; '**' indicates q > 0.05; '!' indicates insufficient number of point mutations to calculate heterogeneity score. Genotypes are presented as coxcomb plots where for individual biopsies the extent to which the variant is present (variant allele fraction (VAF) or LogR for copy number) is represented by the lateral extension of black-outlined wedges. Transparent wedges represent a one-sided 95% confidence interval calculated based upon variant locus coverage as described in **Online Methods**.

Supplementary Figure 4. Study-wide driver mutational heterogeneity. For each cancer the variant allele fraction (VAF) and copy number level (LogR) for each potential driver mutation in each sample is reported. In cases where copy number was unreliable the entire copy number section is grey. RD; residual disease, pCR; pathological complete response, pS; primary surgery, TN, triple negative.

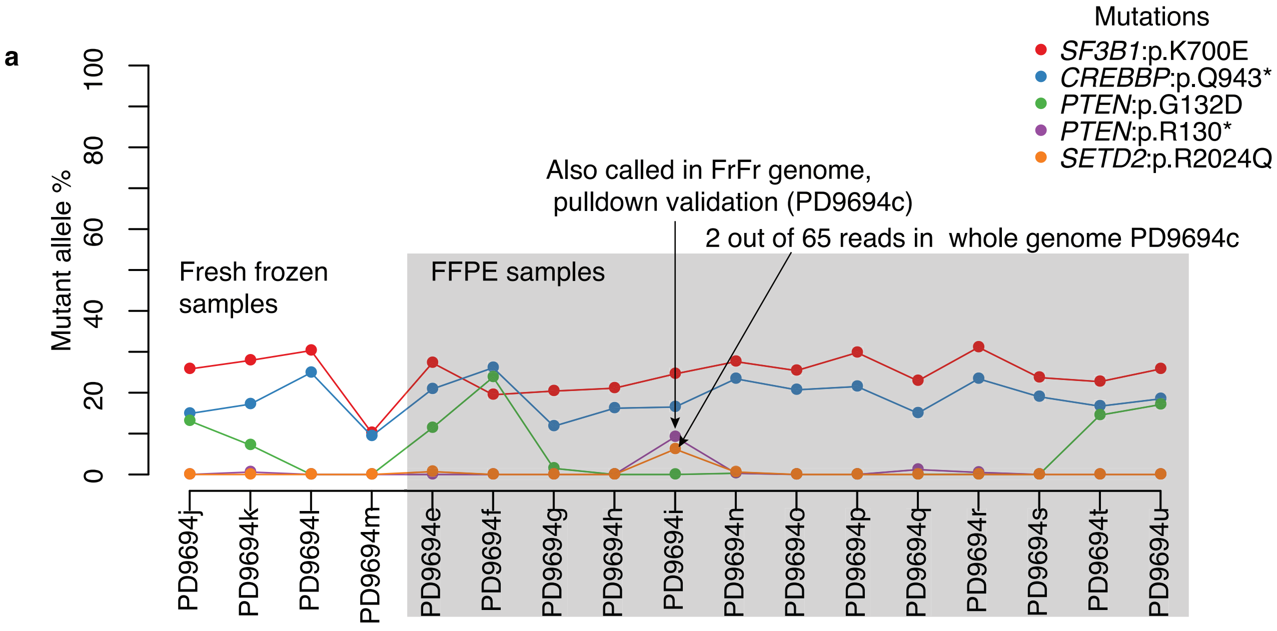
Supplementary Figure 5. Heterogeneity and clinico-pathological features. (**a–h**) Relationship between the heterogeneity score (calculated as described in **Online Methods**) and number of samples examined and various clinical and pathological factors. Individual p values for triple negative and ER$^+$ tumors is presented in blue and red respectively. Possible associations between clinical or pathological factors and genetic heterogeneity as a response were fitted using R's lm() function. F-tests for overall association were then computed using the anova() command. (**i**) Comparison of variability in ki67 with the heterogeneity index amongst the 12 systematically sampled cohort 1 cancers, association was tested as above.

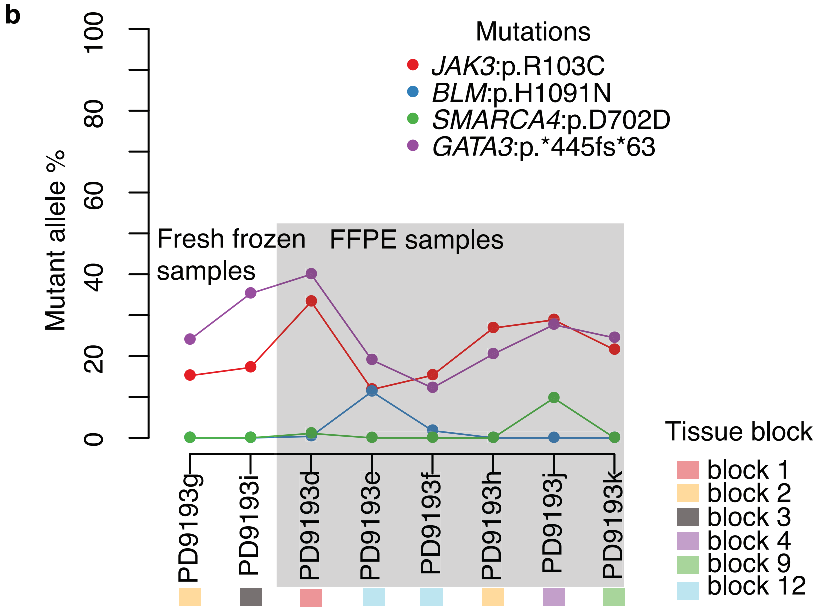
Supplementary Figure 6. Subclonal, targetable mutations and mutation signature evolution, relates to **Figure 5.** (**a**) Point mutational signatures derived from whole genome data (see **Supplementary Note**) operating early (clonal) and late (subclonal events) in evolution as determined by mutational clustering approaches. ‘Early’ mutations are those in the trunk and ‘Late’ mutations are in the branches for all samples except PD9694 – where we compare events in the ductal carcinoma in-situ (DCIS) with those in invasive components of the disease. (**b**) Genes affected by subclonal driver events and potential targeted therapeutic agents. (**c–e**) First hit in parallel evolution in 3 cases. (**c**) Case PD9694: Loss of heterozygosity (LOH) across the PTEN locus in all related samples. The format ‘1+0’ refers to the major and minor allele copy number while the overall copy number is indicated by the black dots. The fact that 100% of cells in the DCIS have LOH indicates that this is an early, truncal event even though additional later events span the locus. Colored lines reflect reconstructed rearrangement breakpoints. (**d**) Case PD9850: 17p LOH is an early event – LogR and B-allele frequency (BAF) data derived from ASCAT applied to the targeted capture sequenced data. (**e**) Case PD9769: ASCAT derived segmented copy number data from whole genome sequence data (blue = minor allele copy number, purple = total copy number) confirms chromosome 21 loss of heterozygosity (LOH), the first hit in *RUNX1* inactivation.

Supplementary Figure 1

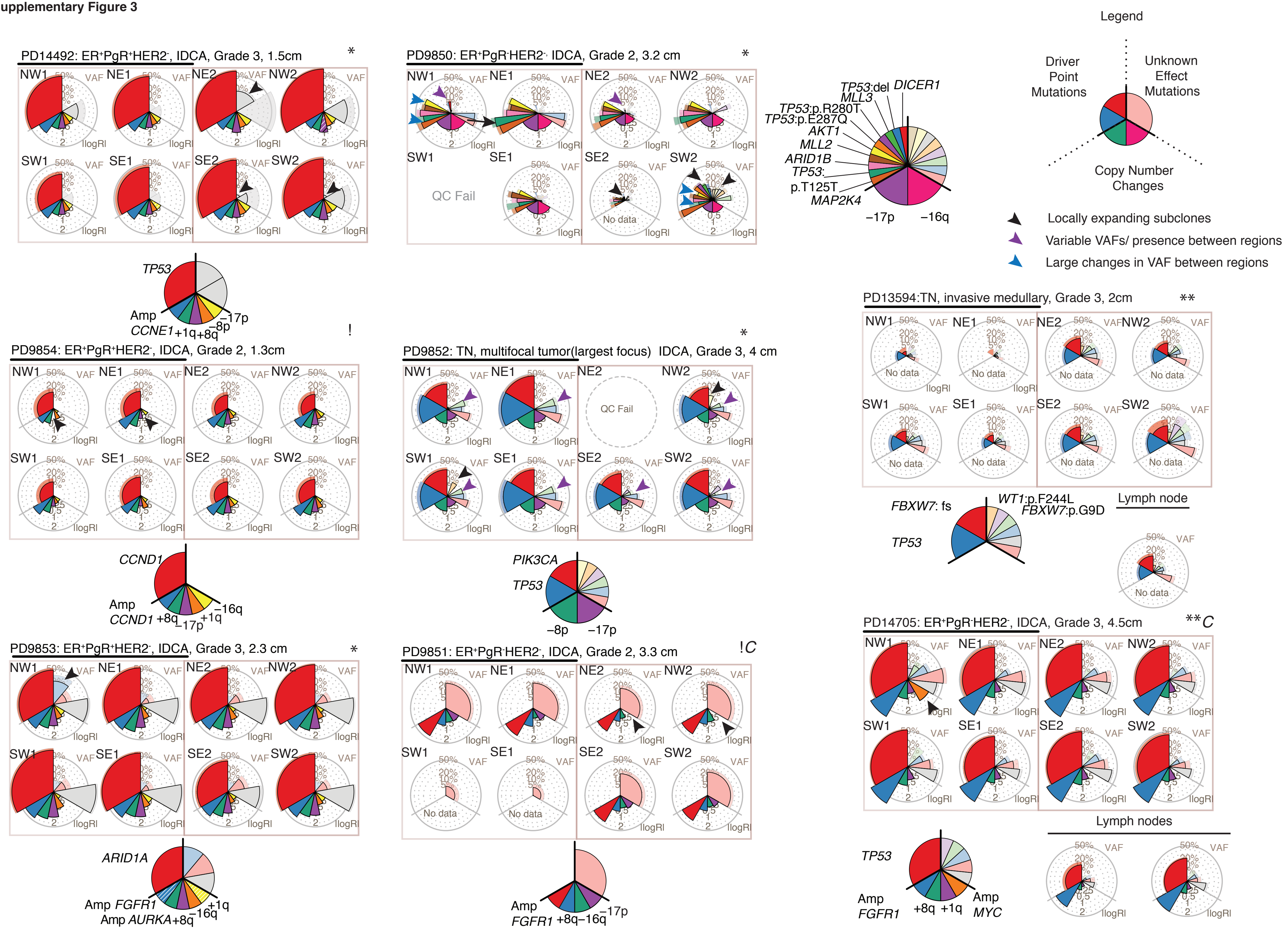


Supplementary Figure 2

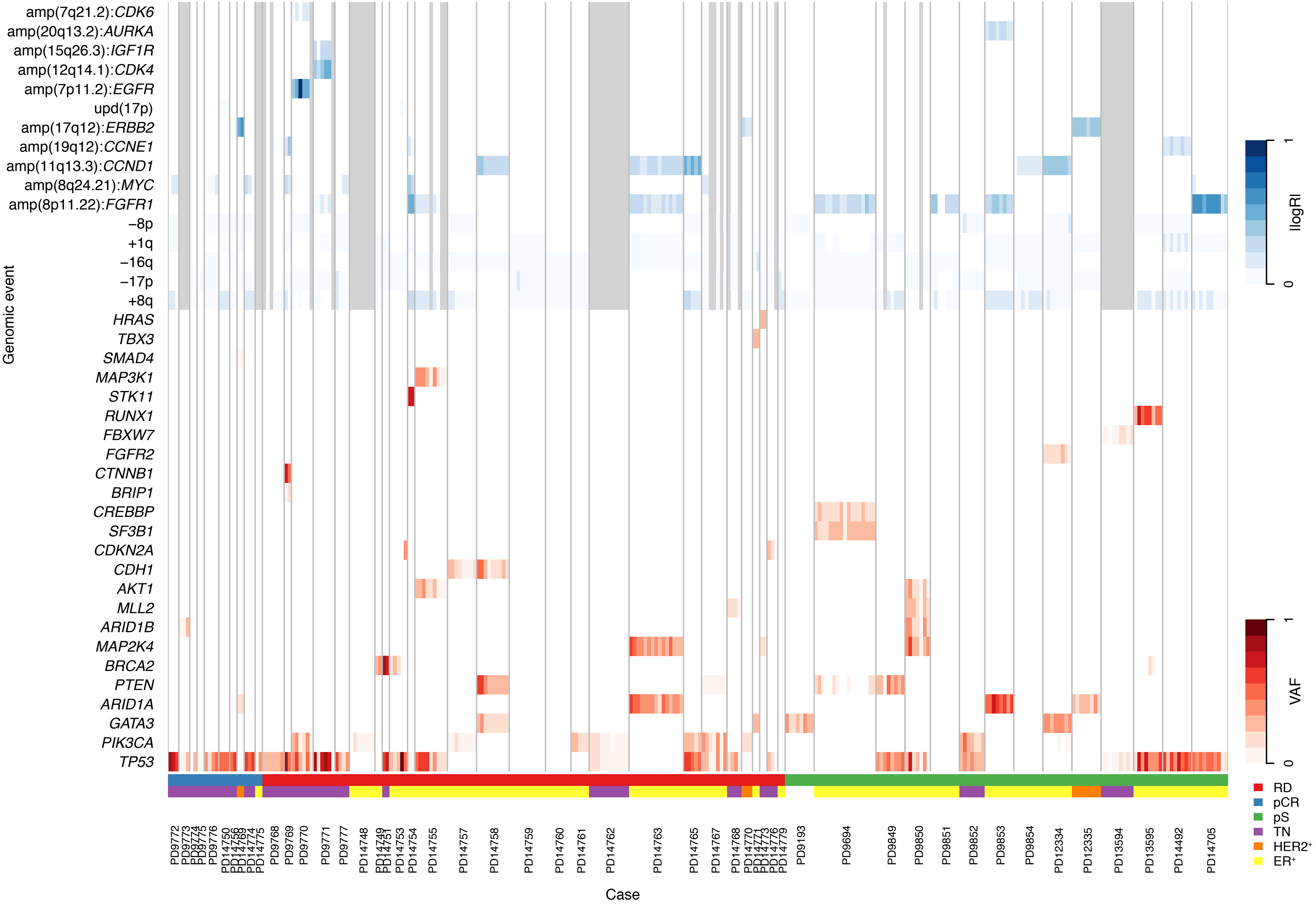




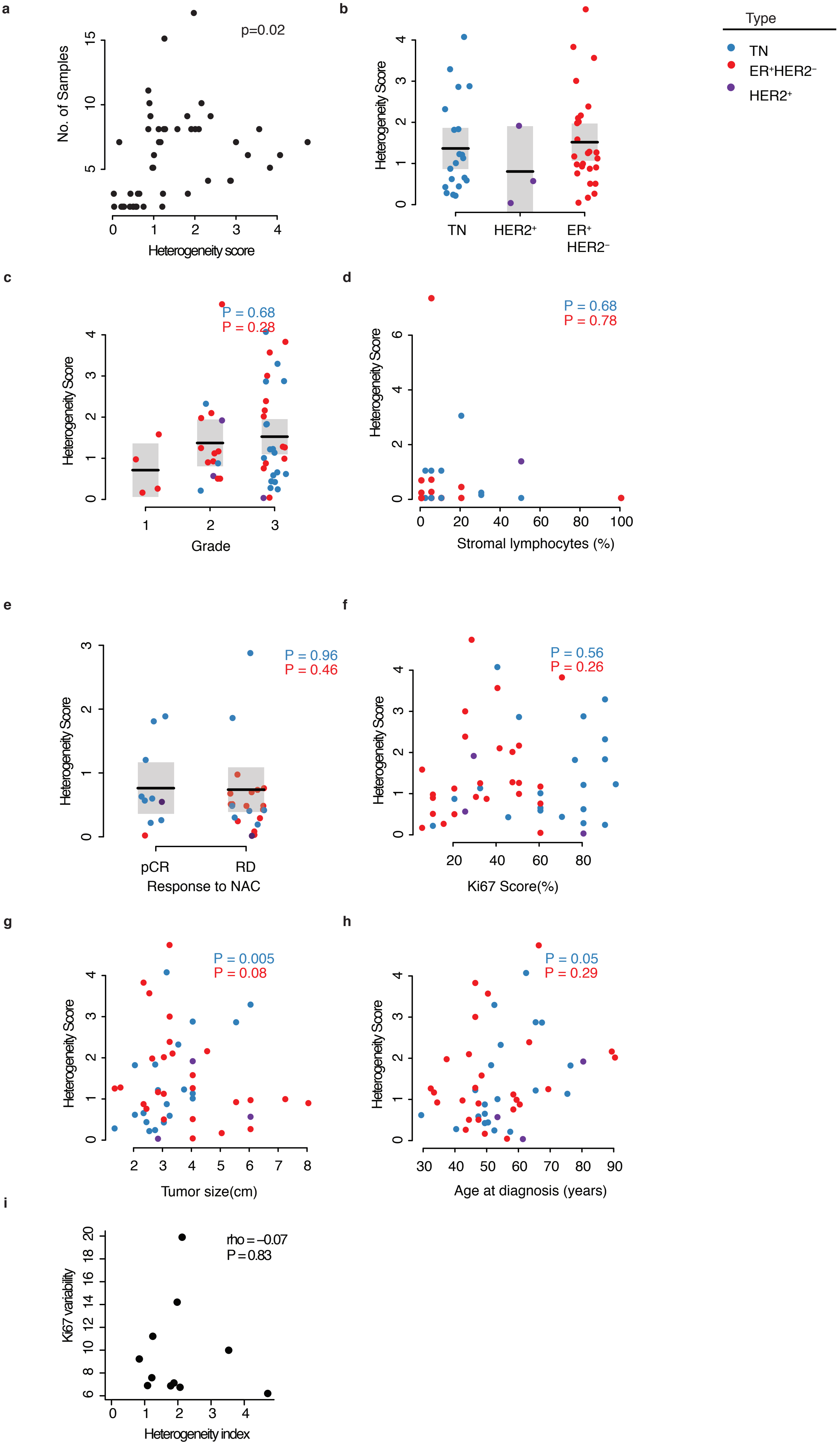
Supplementary Figure 3



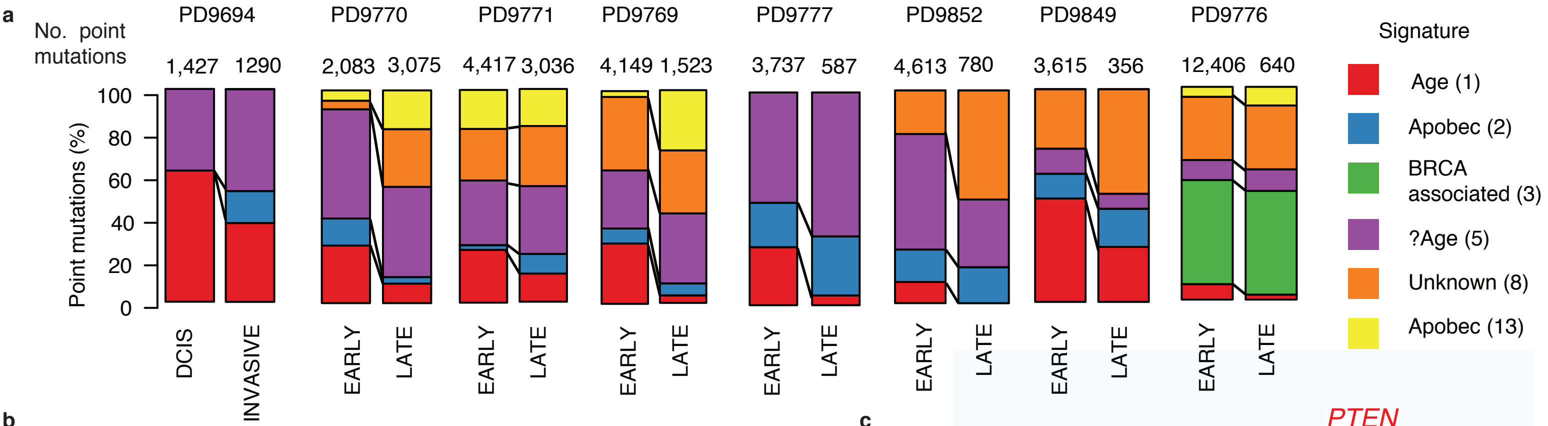
Supplementary Figure 4



Supplementary Figure 5

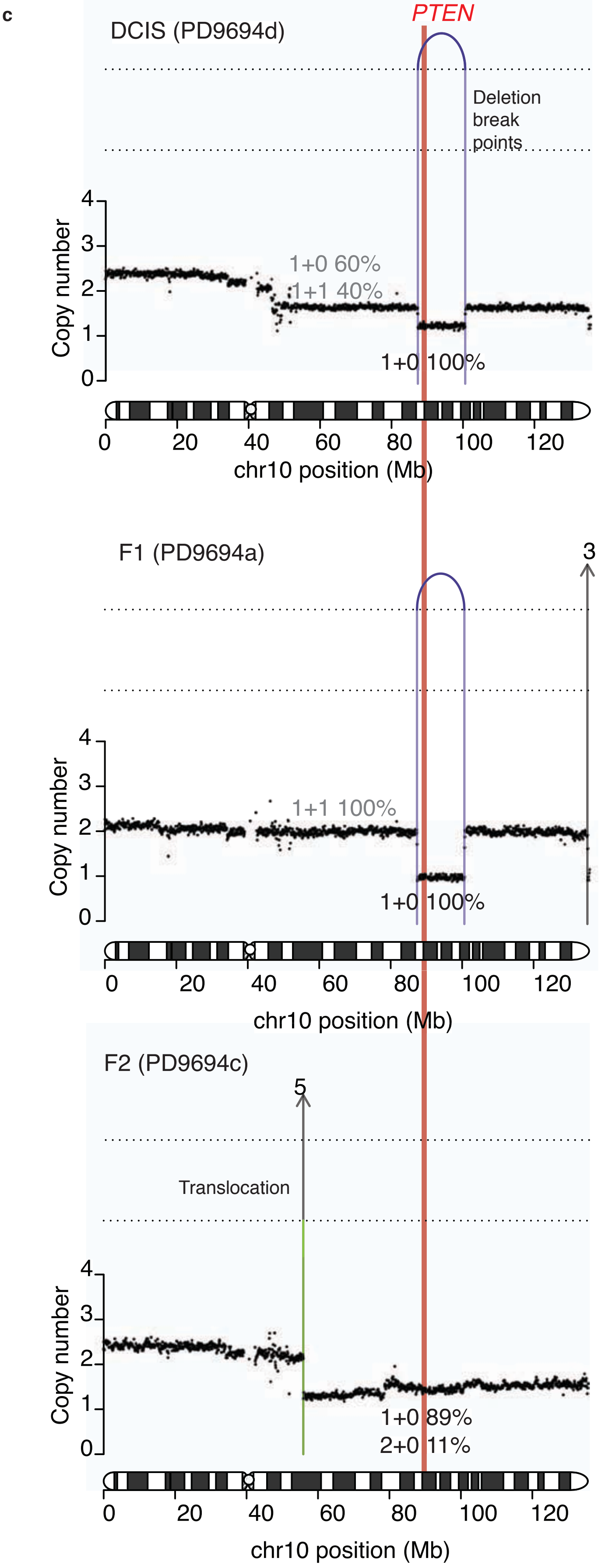


Supplementary Figure 6

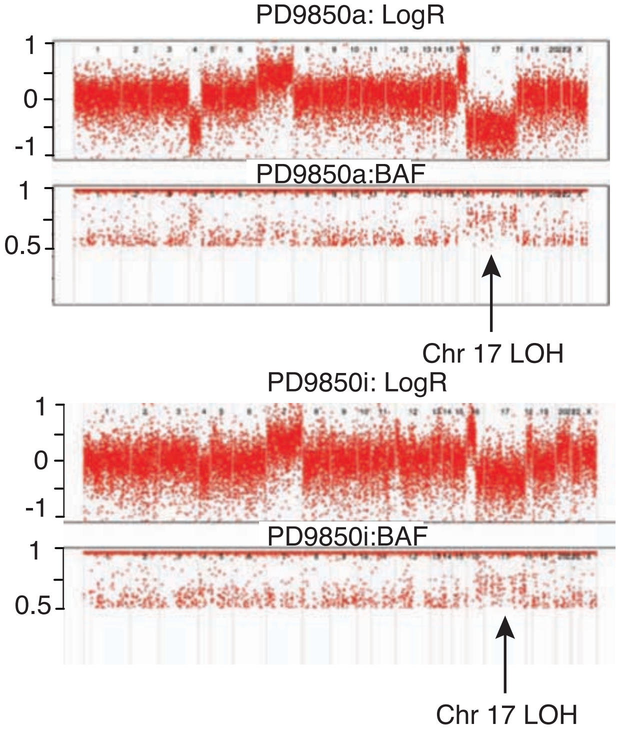


b

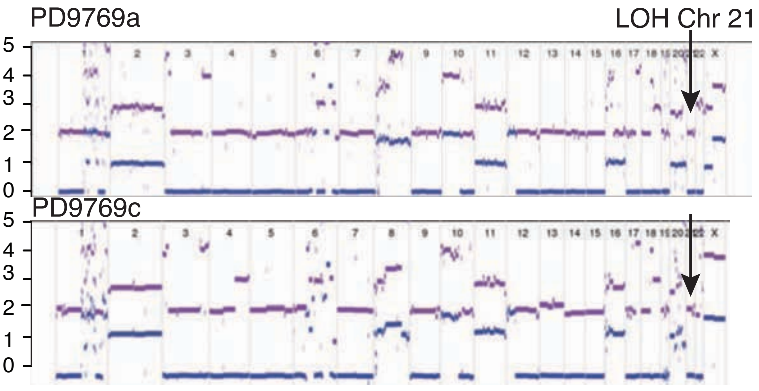
Gene	Samples with subclonal mutations	Example drug types in clinic or development
AKT1	PD9850	mTOR pathway/AKT inhibitors
AKT3	PD9694	mTOR pathway/AKT inhibitors
BRCA2	PD14753, PD13595	PARP inhibitors
CDK6	PD9770	CDK4/6 inhibitors
CDKN2A	PD14753	CDK2/4 inhibitors
FGFR1	PD9694	RTK inhibitors/ FGFR antibodies
FGFR2	PD9777	RTK inhibitors/ FGFR antibodies
MYC	PD14767,PD14705,PD9777, PD9776, PD9772, PD14774	CDK2/AURA inhibitors
PIK3CA	PD12334	Kinase inhibitors
PTEN	PD9694	Kinase inhibitors
RUNX1	PD9769	Methacholine chloride
TP53	PD9850	Vaccines



d



e



Supplementary Note: Methodological details

1. SAMPLE COLLECTION AND SEQUENCING

Sample collection: Cohort 1

Twelve consecutive patients from The Department of Surgery, Haukeland University Hospital, Bergen, Norway were included in the study (**Fig. 1a** and **Supplementary Table 1**). Samples and data were obtained and managed in accordance with the declaration of Helsinki under protocol 2011/2281, approved by the regional ethics committee of the Western Norwegian health authorities.

We adhered to the following procedure for collecting samples from the twelve subjects in cohort 1. Either the day before, or the morning of, the day of surgery, blood samples are collected by venipuncture in blood collection vials containing citrate and ficoll (CPT). The blood is separated by centrifugation at 1500 rcf for 30 minutes, before the lymphocyte layer is washed in 1XPBS and cells are harvested by centrifugation at 600 rcf for 6 minutes. The supernatant is removed and the cells are stored at -80^{o}C until further processing.

Next, multiple (n = 12) tissue samples (15-20 mg) are collected from each primary breast cancer mastectomy specimen immediately upon removal. The tissue samples are collected using 14G Tru-cut needles and each specimen is snap-frozen in liquid nitrogen in the surgical theatre immediately upon removal. The tissue samples are collected according a detailed map as depicted in **Figure 1b**. For two multifocal tumors, biopsies are collected from the main tumor mass. After removal of the tumor, the surgeon (T.A.) cuts the tumor into two halves along the equatorial line. From each hemisphere, 4 Tru-cut samples are collected and named after their relative position in the tumor specimen (NW = North-West, NE; North East and so on). Tru-cut samples from the two halves of the tumor (according to **Fig. 1b**) are named NW1 and NW2 respectively. As a result, we obtained samples from each tumor hemisphere located relatively close to each other (relatively short distance between the two "Northern" samples as well as between the two "Southern" samples) and we obtained samples further apart (relatively long distance in-between "Northern" and "Southern" samples). In addition to the samples described above we collected lymph node metastases from 3 of the patients (1 lymph node from each of PD9849 and PD13594, and 2 lymph nodes from PD14705). In addition, from the central region of each tumor hemisphere another 2 samples are collected as indicated in **Figure 1b**, fixed in formaldehyde and subsequently paraffin-embedding for histology and ki67 scoring.

Details of sample collection: Cohort 2

This cohort included a total of 38 patients' cancers and for each cancer we analyzed multiple samples (average of 5.4 per cancer, range = 2–21) (**Fig. 1a**). For 36 patients who received neo-adjuvant chemotherapy at the Jules Bordet Institute, Belgium, we selected samples for inclusion in the study from archived tissue stores or from prospectively recruited patients. Samples and data were obtained and managed in accordance with the declaration of Helsinki under protocol 1698 and 1634, approved by the Institut Jules Bordet local ethics committee.

All patients had diagnostic, pre-treatment 14G Tru-cut biopsies from the primary tumor – consisting of a minimum of two biopsies fixed in formalin and subsequently embedded in paraffin and another two immediately embedded in OCT and snap frozen using dry ice (**Fig. 1c**). For the surgical specimens, priority is given to adequate sampling for diagnostic purposes. Once sufficient samples are formalin fixed and paraffin embedded (FFPE) and if there is residual tumor tissue left, additional samples are embedded in OCT and snap frozen using dry ice. In addition to the samples described above, we collected lung metastasis samples from patient PD9771. To attain tissue for DNA extraction serial thick sections are cut from FFPE or frozen blocks. For samples estimated to consist of less than 50% tumor cells macro-dissection is performed guided by a pathologist from the same tumor block.

For 2 patients with multi-focal disease (PD9193, PD9694) we collected multiple samples from fresh frozen and pathological blocks taken from treatment-naïve surgical specimens. For case PD9694 several regions contained small foci of invasive disease within DCIS, distant from the 2 major tumor lesions – we performed micro-dissection of these regions to maintain genomic-pathological correlations. Samples and data were obtained and managed under protocol “project SHARE” #93-085, approved by the Dana-Farber Harvard Cancer Center institutional review board.

All samples used in this project were handled and managed within the wider framework and approval for the Breast Cancer Genome Analyses for the International Cancer Genome Consortium Working Group led by the Wellcome Trust Sanger Institute, Cambridgeshire, UK, REC reference: 09/H0306/36. Routine QC procedures resulted in rejection of four samples – as a consequence of genotype mismatch (PD9850c, PD9852f, PD9768a-whole genome sample aliquot only) or inadequate aberrant cell fraction (less than 10% – PD9774a (whole genome sample)).

Ki67 Assessment

We performed Ki67 immunohistochemistry on 5 μm slides from each of the 4 FFPE central cores from the systematically sampled tumors (cohort 1, **Figure 1a–b**) (**Supplementary Table 1**). We adhered to the following protocol whereby de-waxing is performed with xylene/ethanol before target retrieval in a pressure cooker (Decloaking Chamber Plus, Biocare Medical). Staining is performed on a DAKO autostainer using the K4061/Envision Dual Link System (rabbit+mouse). Sections are incubated for 30 minutes at room temperature with a monoclonal rabbit antibody (M 7240, clone MIB-1, DAKO) at a 1:100 dilution. Finally, diaminobenzidine (DAB) as chromogen for 10 minutes is followed by haematoxylin as counterstain for 3 minutes. Sections from tonsils act as positive controls; primary antibody replaced with Tris-buffered saline acts as the negative control. Controls are included in all staining runs.

Two pathologists, both with more than 25 years experience of surgical histopathology and immunohistochemistry examined and scored each slide separately. Slide evaluation is performed using light microscopy (Leica DMLB) with an eye-piece graticule for counting at x400 magnification. Care is taken to avoid areas of intense inflammation, fibrosis, necrosis, low cellularity or poor fixation. 500 tumor cells in each slide are counted in hot-spots; defined as the area containing the highest density of Ki67-labelled tumor cells by visual impression, according to the standards of the Norwegian Breast Cancer Group (www.NBCG.no).

Library preparation and sequencing details

Within this study we generated targeted capture (custom capture pulldown) sequence data for tumor and matched normal DNA using the following approach. Firstly, DNA is fragmented using Covaris® (average insert size ~150bp) and then subjected to Illumina® DNA sequencing library preparation using Agilent's® Bravo

Automated liquid handling platform. Tumor and normal samples are indexed with unique barcodes using PCR. Libraries are then hybridized to custom RNA baits (Cancer Gene Panel v1 and 2, **Online Methods**) according to the Agilent® SureSelect® protocol. Samples are multiplexed on average 16 samples per lane and flow-cell clusters created. Paired-end, 75bp sequence reads are generated using Illumina HiSeq 2000® with approximately 28 Gbp sequence data generated per lane.

We derived genomic (whole genome) libraries with insert sizes of 300bp-600bp from native DNA for 29 tumor and 13 matched normal samples using Illumina® paired end sample preparation kits according to manufacturers instructions. Following cluster generation, 100bp paired-end sequence data is generated using Illumina HiSeq 2000®.

All sequence data is re-aligned to the human genome (NCBI build 37) using BWA[48]. Unmapped reads, PCR duplicates and for targeted capture data, those outside of the target region are excluded from analysis.

2. SOMATIC MUTATION CALLING AND ANNOTATION

Somatic substitutions

An in-house algorithm that identifies Cancer Variants through Expectation Maximisation (CaVEMan) identifies single base, somatic substitutions independently in each sample. The algorithm compares sequence data from each tumor sample to its own matched normal sample and calculates a mutation probability at each genomic locus. Copy number and aberrant cell fraction information derived from ASCAT[32] (performed on SNP loci within the NGS data) informs the locus specific mutation probability calculation of the algorithm. Routine post-processing filters improved the specificity of the data as previously described[1]. We use extensive visual inspection using visualisation software (Gbrowse®) of all coding substitutions and selected non-coding substitutions within whole genome data specifically to assess the validity of branches of the phylogenetic tree. For all substitutions that pass post-processing filters we determined the number of wild-type and mutant non-duplicate reads (with a minimum base quality of 11) covering that genomic position in all related tumor and matched normal samples.

Somatic small insertions and deletions

We identify small somatic insertions and deletions (indels) using a modified version of Pindel[54]. We apply post-processing filters as previously described[27] and use additional steps to improve specificity including visualization of surrounding alignments and interrogation of hundreds of other samples to identify recurrently rejected calls that are likely to represent technical artifacts.

Somatic structural variant detection

Structural rearrangements are detected by an in house algorithm BRASS (**Br**eakpoints via **ass**embly) that first groups discordant read pairs that appear to span the same breakpoint and then assembles reads within the vicinity to reconstruct the breakpoint with nucleotide precision (BRASSII). In-silico reconstruction is attempted for all events supported by 4 or more reads in the tumor and not reported by any reads in the matched normal sample or in an unmatched normal panel. We report only events that are reconstructed by BRASSII. To determine heterogeneity of structural variants we interrogate all related samples' data for any discordant reads or copy number changes to determine if it is truly heterogeneous.

Mutation annotation

Within the cancer gene panels used for the targeted capture experiment we categorized genes as high, medium or low confidence drivers of breast cancer. A total of 45 genes (**Supplementary Table 3**) fulfilled the following criteria for high confidence:

i. Recurrence in the Cancer5000 breast cancer specific analysis ($q < 0.05$)[25];

ii. Recurrence in the Cancer5000 analysis across all cancers ($q < 0.05$) and either amongst the top 20 most frequently mutated genes in breast cancer from the COSMIC database (as of March 2014) or evidence for activity in breast cancer from literature review[3,27-30];

iii. Genes identified in other breast series but not meeting (1) and (2)(*AKT2, ARID1B, MAP3K13* [27], *ATR* [28] and *MAP4K3*);

iv. Additional breast cancer susceptibility genes but not meeting (1) and (2)(*BRIP1, CHEK2, PALB2*).

A further 105 genes where there is evidence for a driver role in other cancer types (significant in the Cancer5000 series and included in the Cancer Gene Census) – are treated as medium confidence for the purposes of mutation curation. Notable omissions from the CGP are the breast cancer driver genes *CBFB* and *NCOR1*[25,28,29]. Next, we assessed the features of each mutation that fell in high or medium confidence genes to determine if it is a likely driver of cancer as described in **Online Methods**.

3. MUTATION VALIDATION

Point mutation validation

We validated somatic point mutation calls using a combination of custom pulldown capture and MiSeq® or HiSeq® sequencing as described above for the targeted capture discovery experiment. We access sequence bam files from all related cancers and count the number of wild-type and mutant reads at each validation locus, in each related sample. A p-binomial test of statistical significance, using a conservative error rate of 1 in 200, calculates the probability of observing that number of mutant alleles at each locus given the coverage depth.

A validation call is made as follows:

i. True positive (validated somatic): $p <= 0.05$ in the tumor and $p > 0.05$ in the matched normal AND the matched normal sample contains <= 2% mutant allele reads AND <= 2 mutant reads in total;

ii. True negative: $p > 0.05$ in tumor and normal at validation and discovery;

iii. False positive: $p <= 0.05$ in the tumor at discovery but $p > 0.05$ at validation;

iv. False negative: $p > 0.05$ in the tumor at discovery but $p <= 0.05$ at validation;

v. Present in germ-line: mutant is present with a $p <= 0.05$ in the germ-line or > 2% mutant reads or > 2 mutant reads are reported in the germ-line.

Whole genome validation data is described in the **Online Methods.** Within the targeted capture we aimed to assess how reliably our experiment identified mutations as present or absent within multiple related samples from the same tumor. To do this we repeated the custom pull-down experiment for 38 tumor samples and 5 matched normal samples from 5 patients within the prospective, systematically sampled cohort.

Due to limited material from individual core biopsy samples we used whole genome amplification (WGA) derived DNA libraries. A total of 7 out of 34 genomic locations completely failed to pull-down in the validation experiment. All of these regions successfully pulled down using the same bait design in all 8 related native DNA libraries, indicating that the failure to pulldown is related to the WGA process. WGA is associated with allele drop-out and preferential allelic amplification[55]. Consistent with this, in the whole genome validation experiment, the highest rate of apparent ‘false positives' occurred amongst cases where validation material is WGA derived. In many whole genome cases apparent false-positive mutations validated somatic in other samples from the same cancer. This supports the notion that technical failures in the validation experiment will under-estimate the specificity of our data. Unfortunately, a limited amount of tissue is frequently encountered when working with real clinical samples and sometimes WGA is necessary.

To assess sensitivity and specificity of the discovery experiment approach we only used locations where the validation coverage was 25X in tumor and matched normal samples. We defined indels as present in both discovery and validation experiments if the indel is seen in tumor but not the matched normal in any number of BWA or Pindel reads in both experiments. Within the discovery experiment across the 38 samples we identified a total of 27 mutations (23 substitutions & 4 indels). All but one of 110 possible sample-mutation combinations validated as somatic (99% validation rate). Similarly, 80 out of 81 sample-mutation-absent combinations validated as absent (99% sensitivity). Overall we made a consistent call at validation and discovery for 189 out of 191 sample-mutation combinations (99% concordance). We assessed for differences amongst mutation calls made in FFPE and fresh frozen samples from the same individual and did not find any differences to suggest that artifacts are more frequent in the latter as long as our post-processing approaches are followed (**Supplementary Fig. 2**).

Validation of structural variants

For each structural variant (SV) reconstructed *in silico* in one or more related sample, we interrogate all related samples for any discordant reads that would suggest its shared existence. For over 2000 SVs, including all cases where discordant reads suggested that the variant is heterogeneous, we sought to confirm or discount this by visually inspecting for a copy number transition at the breakpoints. SVs with an associated copy number change are annotated as 'high confidence' and presented in the analysis in **Figures 6a–c**. For a minority of SVs a copy number change was not observed - this can reflect a balanced event in the case of inversions or translocations, or can be below the threshold of copy number resolution (for small deletions and insertions of <10kbp). To reflect the lack of independent validation we annotate these as 'medium confidence'. In high-level amplification regions (>20 copies), identifying copy number changes can be compromised, and often represent very complex events and we did not attempt to report all associated copy number changes in these regions but given the robustness of the amplification event we report all events as 'high-confidence'. For high confidence events if the copy number change is seen in all samples then it is reported as an early event, while if present in a subset of samples it is reported as a late event. We expect that a proportion of medium confidence events are real but in the absence of independent validation we do not attempt to assign them to the tree.

As further validation of our SV calling we used PCR/ gel electrophoresis for 7 tumor and matched normal samples from 2 patients (**Supplementary Table 4**). We designed primers to 162 breakpoints that successfully reconstructed in silica. We performed primer based PCR in each related tumor and normal from the relevant patient with PCR-amplicons from each sample then run on an agarose gel in tumor-normal pairs. We performed duplicate experiments. The presence of a band in the tumor and not the normal in duplicate experiments represents a somatic event, while a band in the normal indicates a germ-line event. Failure of bands in both tumor and normal represent a false positive call or a technical failure.

Overall a validation call could be made for 54% of rearrangements by PCR and gel electrophoresis. Validation failure can represent either a false positive or a technical failure. Most of the PCR failures included SVs that were associated with clear, and appropriate copy number changes, which suggested that at least some validation failures are likely to represent, PCR failure. We found that failure in one sample is usually associated with failure in all related samples. We used the PCR data to assess for concordance with BRASS1 discordant read calls. For both patients >80% of calls were concordant (**Supplementary Table 4**).

4. COPY NUMBER ANALYSIS

Copy number calling in targeted capture data

Within the targeted capture experiment we evaluated copy number using libraries from the ASCAT algorithm and used LogR and BAF values to identify common arm level copy number changes and amplified genes frequently identified as amplified in breast as described in **Online Methods**. Some frequently amplified breast cancer genes (eg. *MCL1, TERC* and *TERT*) were not included in the analysis because >=5% of normal samples had LogR ratios that approached or exceeded the thresholds used to identify copy number gains (LogR >~ 0.4). In most of these samples there were < 15 SNPs within the regions explaining the high signal to noise ratio.

To screen for copy number changes within these regions of interest we search for deviations in the average LogR and BAF in the region of interest. LogR values >= 0.3 and < -0.3 are used to screen for arm level gains and losses respectively. Assuming an aberrant cell fraction (a) of 0.5 and a diploid (2 allele copies) background (c_n): a logR = 0.3 equates to 3 or more copies of the tumor allele (c_t):

$\log2(((a * c_t)+((1 - a) * c_n))/2) = \log R$

$\log2(((0.5*3)+(0.5*2))/2) = 0.3$

For samples that meet these thresholds the BAF & LogR are then assessed in all related samples including the matched normal. A copy number call is only made if the LogR is approximately zero in the matched normal sample.

To identify likely driver amplification events we assessed the average LogR across each gene of interest in each indivudal sample (**Supplementary Table 3**). An average LogR >= 1.2 equates to 7 or more copies of the gene and is reported as an amplification if the LogR in the matched normal is approximately zero. As fluctuations in aberrant cell fraction and complex subclonal mixtures may alter the apparent contribution of an event we look visually inspect for deviation in LogR and BAF in all related samples, across all altered genes and compared to the matched normal to determine if there is support for the event across the tumor. As a consequence in some cases t an amplification is called when the logR is less than 1.2. This approach was adopted to minimize the risk of over-identifying heterogeneity.

If we identify overlapping events (e.g. *MYC* locus amplification on 8q gain) we only identify the focal amplification if it is > 0.3 higher than the arm level event (i.e it would qualify as a gain event in isolation). We attempted to identify focal deletions using a similar approach but were unable to consistently identify these in targeted capture data – for example, we were unable to differentiate between loss of an allele in a tetraploid tumor from LOH in a diploid tumor.

Validation of copy number changes in targeted capture data

For 96 tumor samples and matched normals from the 12 patients included in cohort 1 we performed Multiplex Ligation-Probe Amplification (MLPA) of 41 exons from 22 genes implicated in breast cancer. Exons targeted by the MLPA panel overlap 7 out of 12 of our regions of interest: *MYC, CCND1, FGFR1, CCNE1, AURKA, EGFR and ERBB2* (**Supplementary Table 4**).

We performed MLPA analysis using the SALSA MLPA P078-Breast tumor probemix-kit (MRC-Holland, Amsterdam, The Netherlands). Probe hybridisation, ligation and amplification are performed according to the manufacturer's instructions. Capillary gel electrophoresis and data collection are performed using an automated DNA sequencer (ABI 3700, Perkin-Elmer Biosystems). Copy numbers are estimated by assessing the peak areas representing each individual amplification product, using the Coffalyser v.7 software: The peak areas of all MLPA products resulting from amplification of the breast cancer relevant genes are first normalized by the average of peak areas resulting from control probes specific for a set of regions outside the genes of interest. A ratio is then calculated where this normalized value is divided by the corresponding value from a reference sample consisting of pooled DNA from >6 healthy individuals. We determine a sample as having a clear amplification at a specific location if this ratio is above 2 (equating to approximately 6 or more copies where the aberrant cell fraction is assumed to be ~50%).

Across the 12 subjects' cancers in cohort 1 we identified 9 amplification events within the scope of the MLPA panel. Using the above thresholds we confirmed all 9 events as somatic in one or more sample from each cancer (**Supplementary Table 4**). Across individual samples, of 663 validation calls, we identified 97% concordance. In

2 cases (PD9851 and PD14705) amplification of *FGFR1* and *MYC* respectively were identified as heterogeneous at discovery and were consistently identified in the relevant samples by the MLPA experiment. For 3 patients we made an inconsistent call and each involved a *MYC* amplification called by the MLPA but not the NGS approach. The reason for this is likely to be underlying widespread 8q gain (which includes the *MYC* locus) that we were able to normalize for in the NGS but not MLPA data.

Identification of driver copy number events in whole genome data

We derived segmental copy number information for each of the 29 tumor samples and matched normal samples for which we had whole genome NGS data using the ASCAT algorithm (allele-specific copy number analysis) of tumors as previously described [32]. The algorithm simultaneously determines and utilizes aberrant cell fraction and ploidy estimates to determine allele specific copy number 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 total copy number equals zero.

Potential driver amplification or homozygous deletion events in our data are identified by cross referencing the genes that we identified in amplified regions with the genes in recurrent regions identified by a recent analysis of copy number change in 4,934 cancers from The Cancer Genome Atlas (TCGA) Pan-Cancer data set[33]. We specifically look for amplifications involving the 39 genes detailed in **Supplementary Table 3**. These include 13 genes recurrently amplified specifically in breast cancer, 18 genes from the pan-cancer analysis and 5 additional genes reported as driver amplifications by the cancer gene census[24]. We also included in our analysis *FGFR1* which is amplified and/or overexpressed in ~10% of breast cancers and is associated with worse prognosis[56], *FGFR2* which is amplified in a subset of triple negative breast cancers[57], JAK2 and the cyclin *AURKA* both recently reported as recurrently amplified in triple negative breast cancers treated with neo-adjuvant chemotherapy[34]. We also sought evidence for amplification of oncogenes not classically activated through amplification. For regions containing these genes at or above the amplification threshold we determined its copy number in each related sample. As segmented data can miss fluctuations in copy number, occurring as a consequence of complex rearrangements, the copy number was calculated and visualized across each individual gene's locus in association with the surrounding breakpoints such that excessive noise was not mistaken for amplification and to avoid inaccuracies resulting form segmentation failures.

We specifically sought homozygous deletions in 12 recurrently deleted regions in the TCGA analysis that contained tumor suppressor genes implicated in breast cancer and we checked additional tumor suppressor genes implicated in breast cancer (**Supplementary Table 3**). Amplifications and homozygous deletions involving these likely driver regions are summarized in **Supplementary Table 8**.

5. MUTATIONAL SIGNATURE ANALYSIS

We detected mutational signatures in two independent ways: (i) *de novo* extraction based on somatic substitutions and their immediate sequence context and (ii) refitting of previously identified consensus signatures of mutational processes. The *de novo* extraction is performed using a previously developed theoretical model and its corresponding computational framework [53]. Briefly, the algorithm deciphers the minimal set of mutational signatures that optimally explains the proportion of each mutation type in each mutational catalogue and then estimates the contribution of each signature to each sample. The computational framework identified five reproducible mutational signatures, termed Signatures A through E. These signatures closely resemble previously identified breast cancer signatures[58]. Signature A corresponds to previously termed Signature 1B, Signature B to Signature 2, Signature C to Signature 5, Signature D to Signature 13, and Signature E to Signature 3.

Recently, we identified 27 distinct consensus mutational signatures from 7,042 samples across 30 different cancer types[58]. We evaluated all possible combinations of at least seven mutational signatures for each sample by minimizing the constrained linear function:

$$\min_{Exposures_i \geq 0} ||SampleMutations - \sum_{i=1}^{N} \left(\overrightarrow{Signature_i} * Exposure_i\right)||$$

Here, $\overrightarrow{Signature_i}$ represents a vector with 96 components (corresponding to the six somatic substitutions and their immediate sequencing context) and $Exposure_i$ is a nonnegative scalar reflecting the number of mutations contributed by this signature. N reflects the number of signatures found in the sample and all possible combinations of consensus mutational signatures for N between 1 and 7 are examined for each sample. This resulted in 1,285,623 solutions per sample and a model selection was applied to select the optimal solution. The model selection framework excludes any solution in which a mutational signature contributes less that 2% of the somatic mutations or less than 50 somatic mutations. Exceptions are made for Signatures 1A and 5 as these are believed to reflect on-going endogenous mutational processes that continuously contribute very low numbers of somatic mutations[58]. Further, the model selection framework selects the solution that optimizes the Pearson correlation between the original pattern of somatic mutations and the one based on refitting the sample with consensus mutational signatures such that each additional signature should improve the Pearson correlation with at least 0.02. The final solution for each sample contained not more than 6 mutational signatures and these signatures were mostly consistent with the ones previously identified by the *de novo* analysis: Signature 1A, Signature 2, Signature 3, Signature 5, Signature 8, and Signature 13. In the de novo analysis signature 8 was not identified, and appears to have been intermixed with signatures 2, 5 and 13, with its resolution limited by the small sample size.

We assessed the relative activity of mutational processes over time by allocating somatic mutations to their specific branch of the phylogenetic tree. Similar to what we previously reported [1,59], we found that base substitutions acquired early in breast cancer evolution are enriched for C>T mutations at CpG dinucleotides relative to all other base substitution signatures (**Supplementary Fig. 6e**). This is a universal, age-related mutational process thought to be driven by spontaneous deamination of methylated cytosines [58]. Later in breast cancer evolution, a signature attributed to APOBEC activity, with cytosine mutations occurring in a TpC context, becomes more pronounced. The neo-adjuvant therapies used (taxanes and anthracyclines) have not previously been associated with specific mutational signatures. No novel mutational signatures, or enrichment for specific signatures were observed in the post-chemotherapy samples.

6. PHYLOGENETIC TREE CONSTRUCTION: CASES WITH ALTERNATIVE SOLUTIONS

For the 4 cases where we were able to derive feasible alternative tree solutions (**Supplementary Fig. 1a**) beyond those presented in **Figure 4a** we describe tree construction below.

Case PD9694

We sequenced to whole genome level 4 samples from this cancer – a sample from an area of DCIS (PD9694d), one sample from each of 2 invasive foci F1 (PD9694a) and F2 (PD9694c) and a matched, blood derived normal sample. The general principle of somatic substitution clustering and phylogenetic tree construction is as described in the **Online Methods**. For this case, we identified 2 possible solutions that differ by the position of a terminal branch (cluster 1) that are compatible with the cancer cell fraction estimates from the genome-wide (discovery) substitution data (**Supplementary Fig. 1a–b, Supplementary Table 5**).

Somatic substitutions that are fully clonal in all three lesions represent events on the trunk of the phylogenetic tree (**Supplementary Fig. 1a–b**, cluster 8). The earliest detectable subclonal divergence gave rise to two branches that comprise approximately 60% and 40% of cells respectively in the DCIS sample (cluster 7 and 3). Interestingly, the mutations on the former branch were near or fully clonal in the major invasive focus (PD9694a), suggesting that a cell from this lineage seeded the focus. The minor invasive focus showed contributions from both major lineages of the DCIS lesion, contributing ~10% and ~90% of cells (cluster 7 and 3) respectively. Within all three lesions, there was evidence of on-going evolution and partial selective sweeps, manifesting as clusters of subclonal mutations not seen in any other lesion (PD9694c; cluster 2, PD9694d; cluster 1, PD9694a; cluster 6).

All clusters identified in the discovery data were independently identified in the validation data reiterating the tree structures inferred from whole genome data (**Supplementary Fig. 1c**). The cancer cell fraction of cluster 1 mutations within the validation experiment is lower than that in the discovery data further confirming that cluster 1 could arise in series from either cluster 3 or 7 (i.e. the alternative tree is as plausible as the consensus). All but one cluster was independently identified by the clustering of indels resulting in similar tree structures (**Supplementary Fig. 1c**). All branches were supported by copy number associated SV breakpoints across the 3 samples.

Case PD9773

We sequenced two samples from this subject's tumor. The 'most compatible' solution based upon the clustering alone indicates very early branching (**Supplementary Fig. 1a–b**, **Supplementary Table 5**) with just 7% of mutations being fully clonal in both samples (cluster 3) and 92% of mutations being subclonal and at similar levels in the two samples (cluster 2). This indicates very early branching and a stable subclone contribution in different regions. Although this is plausible, this would make this sample an outlier in our cohort and of all of the samples that we sequenced we found this to be technically the most challenging as ploidy estimates and aberrant cell fraction estimates were inconclusive. Battenberg and ASCAT algorithms were both employed and made estimates of ACF of 23% and 32%, respectively, in a tetraploid background respectively and less than 10% in a diploid background. Further hindering the analysis of this sample set was inadequate coverage in sample 'a' (15X) relative to the 'c' sample (45X). The imbalance in coverage is due to exhaustion of this library early in sequencing and unfortunately additional DNA was not available from this sample. There is no doubt that these factors produce large uncertainty in the true position of the major cluster which, based upon our experience, is most likely to lie at 100% in both samples.

We report the second solution in **Supplementary Figure 1a** as the most likely, with a clonal cluster comprised of both cluster 2 and 3 mutations accounting for 99% of mutations. Validation clustering placed the major cluster close to 100% (validation cluster 3, **Supplementary Table 5**), supporting this solution. A small subclone that is private to PD9773c was identified in whole genome data but this accounts for just 1.2% (n=23) of mutations and none of these were included in the validation. Additional subclonality may exist and has gone undetected on account of technical limitations in this sample. This sample set illustrates the challenges of working with clinical samples and unfortunately in this case little could be concluded biologically and so the uncertainty is represented in the faded out branches of the tree depicted in **Figure 4a**.

Case PD9777

We sequenced two pre-treatment diagnostic biopsies from this subject's triple negative tumor (PD9777a and PD9777c; 5cm, Grade 3, Ki67 80%) and a sample from the surgical resection (PD9777d; 1.8cm, Grade 3, Ki67 60%) after neo-adjuvant chemotherapy with epirubicin.

A total of 4,506 mutations identified genome-wide were clustered by our algorithms with 80% identified as clonal in all related samples. Four additional subclonal clusters (containing >= 2% of cells) were also identified in the discovery data. A solution was arrived at for 2 samples – PD9777a and PD9777d (left tree, **Supplementary Fig.**

1a). However, including the third sample 'PD9777c' was not consistent with an acceptable solution. The cancer cell fraction of a private cluster in the 'c' sample was estimated to be present in 100% of cells (**Supplementary Table 5**, cluster 3) while a cluster shared by all samples (cluster 6) was present in around 55% of cells in this sample – a situation that is incompatible with the basic principles of phylogenetic relationships. We validated a number of mutations from each of the clusters and there was compelling evidence that the estimated fractions of clusters in PD9777c were inaccurate. A strongly supported validation cluster (46 high depth mutation calls vs 166 low depth calls in discovery data) suggested that the earliest cluster (cluster 6) is present at a higher level in PD9777c (80% cells) than suggested by the discovery data. A small validation cluster containing 10 mutations also suggested that the PD9777c private cluster that appeared to be in 100% of cells in the discovery data was actually likely to be present in around 37% of cells. Adopting the refined cancer cell fraction estimates allowed a tree to be constructed that satisfied the basic principles outlined in the **Online Methods.** This is the right hand tree in **Supplementary Figure 1a** and **Figure 4a**.

Copy number associated SV breakpoints confirmed the overall branching pattern of the tree with both shared and private events within the relevant samples.

Case PD9775

We sequenced two pre-treatment diagnostic biopsies from this subject's triple negative tumor (4cm, ductal, Grade 3, ER-,PgR-,HER2-, Ki67=32). The patient went on to achieve a complete pathological response to epirubicin and taxane NAC so a post-treatment sample was not available.

Our algorithms identified in excess of 3,000 somatic mutations and 99% of these are clonal in both samples. The tree derived from whole genome data therefore consisted of a trunk and no branches (**Supplementary Fig. 1a**). The validation experiment, which was heavily enriched for private mutations or those with the greatest distance between allele fractions at discovery, identified a subclonal cluster that is estimated to be present in 60% of cells in each sample (**Supplementary Table 5**, **Supplementary Fig. 1b**). We report this cluster with some caution, because there was considerable dispersion of the mutations and clusters on the leading diagonal can reflect technical artifacts. This may occur as a consequence of missed copy number changes, although, this is unlikely since the mutations were dispersed across 19 chromosomes. Similar artifacts may also occur as a consequence of pull-down inefficiency for some locations. However, a large proportion of the mutations (51/150) were within this cluster and 8 of them were in coding sequence so unlikely to be in highly repetitive regions. With these caveats, we report that subclonality exists in PD9775, but the total number of mutations involved is likely to represent <1% of the total mutation burden as it was not seen in the discovery phase.

Consistent with the very low level of subclonality detected in the substitution clustering in this sample, only a single copy number associated SV breakpoint was identified as heterogeneous.

Supplementary references

54. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).

55. Pinard, R., *et al.* Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC genomics* **7**, 216 (2006).

56. Elbauomy Elsheikh, S., *et al.* FGFR1 amplification in breast carcinomas: a chromogenic in situ hybridisation analysis. *Breast cancer research : BCR* **9**, R23 (2007).

57. Turner, N., *et al.* Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. *Oncogene* **29**, 2013-2023 (2010).

58. Alexandrov, L.B., *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).

59. Nik-Zainal, S., *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993 (2012).