Supplementary Material and Methods

Sample preparation for NGS

DNA samples were isolated using commercially available kits following manufacturer's instructions. In brief, whole peripheral blood was centrifuged to separate plasma and buffy coat fractions; DNA was isolated from the buffy coats using the DNeasy kit (Qiagen) while cell-free circulating tumor DNA was isolated from plasma using the QIAamp Circulating Nucleic Acid kit (Qiagen) DNA from fresh frozen specimens was extracted from thirty 10um cryosections using the AllPrep Qiagen kit. For every FFPE specimen, ten 5-10um sections were stained with nuclear fast red and scored by a pathologist. Manual needle micro-dissection was performed on the stained sections to separate distinct tumour areas highlighted by the pathologist and to isolate surrounding healthy tissue. DNA was further extracted using either the QuickDNA FFPE kit (Zymo) or the QiaAmp DNA FFPE tissue kit (Qiagen).

Bioinformatics Analysis

Purity estimation: purity estimates were calculated using histopathology scoring, the VAF of a potential clonal diploid mutation (PCDM) in the WES variants and estimated using ASCAT (1). Table S2 presents the purity estimates for all of the three approaches.

Copy number analysis: tumour sample copy number $\log_2$ ratios (LRR) and B-allele frequencies (BAF) were calculated using PureCN (2). LRRs were normalised using the global median and outliers smoothed using CGHcall (3). Mirrored BAFs (mBAF) (4) of heterozygous loci were segmented using piecewise constant fitting (PCF) (5). A two-component Gaussian mixture model was fitted to the BAF values of each segment utilising mixtools version 1.0.4 (6) to estimate the major allele distribution in each segment. However, if the segment BAF values were considered to be drawn from a normal distribution expected in allelic balance (BAF=0.5), the segment was considered balanced (Kolmogorov–Smirnov test, p>=0.05). The major allele BAF and LRR of each segment was used as input for ASCAT to estimate tumour purity and ploidy, limiting the minimum purity of the solution to the lower 95% binomial confidence limit (Wilson method) of a PCDM purity adjusted for a tetraploid solution. Ploidy range was restricted in tumours manually assessed to have a high ploidy state, and purity and ploidy was preset if a solution was not obtained, parameter ranges are provided in Table S2. Solution purity and ploidy were used to estimate copy number alteration (CNA) clonality of each segment using the Battenberg methodology (7), for segments considered subclonal, the copy number state with the highest prevalence was taken. Copy number calls are provided as Supplementary Data.
WGS data from Patients 3 and 4 were analysed for CNAs using Sequenza (8).

Identification and validation of somatic variants: adapter trimming was performed with Skewer v0.1.126 (9) with minimum read length after trimming 35 and mean quality value before trimming of 10. Trimmed reads were aligned to the full human reference genome hg19 with Burrows-Wheeler Aligner (BWA) v0.7.12 (10). PCR duplicates were marked using Picard tools. Joint mutation calling between multiple samples from the same patient was performed per patient using a combination of Platypus v0.8.1 (11) with biased prior ('source' option) for mutations called by Mutect2 (12) on single tumour-normal pairs. This allowed us to exploit the sensitivity of Mutect2 with the joint calling capability of Platypus. The

following filtering criteria were used to call somatic variants in WES samples: i) only variants with Platypus filter PASS, alleleBias, Q20, QD, SC and HapScore were kept, ii) minimum coverage and genotype quality of 10 was required iii) variants in segmental duplicated regions and centromeric regions were removed, iv) minimum of 3 reads covering the variant in at least one of the tumour samples per patient were considered, v) 0 number of reads covering the variant in the germline sample, and vi) genotype of the 0/0 in the germline sample. Only somatic alterations and indels with a Variant Allele Frequency (VAF) >5% were considered. Somatic variants were annotated both with CAVA (13) and VEP (14). Potential drivers such as PIK3CA, TP53, CTCF, ARID1A, FOXA1 were also reviewed manually using the Integrated Genomics Viewer (IGV) (15). Fig S2 shows the distribution of mutations in the WES samples. The same analysis was used for the analysis of WGS samples (pat. 3 and pat. 4). In total, 807 variants were selected from the WES samples for targeted sequencing. SNV calling on the targeted capture samples was performed using Platypus in genotyping mode. Somatic SNVs with minimum genotype quality of 10, minimum coverage of 300 and identified by a minimum of 10 reads were considered for further analysis. Somatic mutations that failed the validation in all samples per patient were removed otherwise VAF is indicated as NA in the failed sample. All validated variants (SNVs and indels) were used for the rest of the analysis. Indels were removed from ctDNA analysis to avoid calling false positives because of the low allele frequency variants in ctDNA. Variant allele frequencies of TES and WES data are available from Tables S3 and S5.

Driver genes: the complete set of SNVs was compared to a list of known potential driver genes in breast cancer. This list included genes found previous studies (16-19) and breast cancer driven genes found in COSMIC Cancer Gene Census(20) (downloaded 27/02/2017).

Cancer cell fraction estimation: for each variant the local total copy number state, VAF and sample purity was used to estimate cancer cell fraction(21). ASCAT purity was taken unless the purity was equal to 100%, in which case the PCDM purity estimate was used. The number of alleles mutated was assumed to be 1 to avoid overcalling subclonality. Cancer cell fractions for TES and WES data are available from Tables S4 and S6.

Mutational Signature analysis: Signatures of mutational processes were analysed with deconstructSigs (22) using the Wellcome Trust Sanger Institute mutational signatures framework (23). Mutational signatures were estimated in three groups per patient using i) clonal mutations, ii) private to the primary mutations iii) private to lymph node mutations initially for the WES samples and then, to validate for the WGS samples (pat.3 and pat.4). Differences in mutational signature analyses between WES and WGS samples were minimal, showing that number of mutations in WES samples were sufficient to identify the major mutational signatures. Lastly, analysis was repeated using the full list of mutations from all patients to show the overall trend in the three subgroups.

Phylogenetic reconstruction: validated indels and variants from the targeted sequencing (Fig. 2) were used for phylogenetic reconstruction with PAUP* maximum-parsimony (24) by binarising the CCF values to produce tables indicating the presence/absence of each mutation in each sample. To account for the potential for trees to be confounded by subclones or clonal mixing, we considered trees constructed from mutations both with CCF>0.1 (all mutations, Fig. 3 trees) and with CCF>0.8 (clonal mutations only, Fig. S6 trees). Excluding the tree constructed from TES sequencing for Patient 7 and those trees comprising only two non-normal samples, trees were determined by an enumeration and evaluation of all possible trees. For the Patient 7 TES sequencing tree an exhaustive search is infeasible owing to the

large number of samples and a heuristic search was used with the stepwise addition option to obtain the starting trees for branch swapping and 500 replications were performed. For all trees, the normal tissue (blood) sample was designated as the outgroup. Bootstrap analysis was carried out to assess the support of the phylogenetic tree nodes with 10000 replicates for each tree (Fig. S5). The same analysis was also applied to the variants from the WES samples (Fig. S5). For patients with only two samples, trees were drawn manually with the trunk and branch length to be proportional to the number of clonal and private mutations of each of the two samples respectively.

In situ genomic profiling

CISH data generation: CISH on a FFPE tissue was performed using BaseScope™ for PIK3CA mutation profiling and RNAscope® for APOBEC expression according to manufacturer's guidelines provided by Advanced Cell Diagnostics (ACD Bio, Newark, CA). Four μm sections were prepared and incubated at 60^0C for 1 hour before xylene and ethanol treatment for deparaffinisation and rehydration. Following this, endogenous peroxidase was blocked using Pretreat 1 (hydrogen peroxidase) for 10 minutes at RT. Antigen retrieval was performed using Pretreat 2 for 15 minutes at 100^0C and Pretreat 3 (protease) was applied for 30 minutes at 40^0C in a HybEZ™ oven. Distilled water was used to rinse slides between each Pretreat. Next, BaseSope™ probes (PIK3CA Wild Type and PIK3CA H1047R Mutation) were provided by ACD Bio. APOBEC3A and APOBEC3B RNAscope® probes were custom designed and purchased from ACD Bio. Probes were hybridised for 2 hours at 40^0C in HybEZ™ oven. Signal amplification steps were performed using AMP reagents (ACD Bio) in the following order: AMP0 is at 40^0C for 30 minutes, AMP1 is at 40^0C for 15 minutes, AMP2 is at 40^0C for 30 minutes, AMP3 is at 40^0C for 30 minutes, AMP4 is at 40^0C for 15 minutes, AMP5 is at RT for 30 minutes and AMP6 is at RT for 15 minutes. Wash buffer was used two times in between each AMP reagents. Finally, slides were incubated with fast red and counterstained with Gill's haematoxylin.

CISH image analysis and signal quantification: in order to identify regions of over- or under-expression, detection was performed across the entire slide. Chromogen spots are commonly under 5 microns in size, and thus will only be visible in visual fields with sufficiently high resolution. For this reason, detection was performed on slide images with a pixel resolution of 0.22μm/pixel, comparable to a 20× optical magnification. Each whole slide image (WSI) was divided into 2048×2048 blocks and chromogen spot detection was performed on each block individually. Results from each block were then combined to produce the complete set of detections for the WSI. Detections that occurred outside of regions identified as tumour were not of interest in this work and were excluded from further analysis.

The intensity of the chromogen staining was extracted from the image using the matrix-based colour deconvolution approach introduced by Ruifrok and Johnston (25). The haematoxylin-chromogen stain matrix, *M*, given below, was determined empirically by sampling 500 pixels of each stain.

$$M = \begin{bmatrix} 0.741 & 0.607 & 0.286 \\ 0.463 & 0.784 & 0.414 \\ 0.027 & -0.174 & 0.300 \end{bmatrix}$$

It is common for the extracted chromogen intensity channel to contain variable background noise, and thus the signal was enhanced by the use of a Laplacian of Gaussian (LoG) filter (26). For a visual field with a resolution of 0.22μm/pixel, we recommend a LoG filter with $\sigma = 5$. The LoG filtered stain channel was thresholded to isolate the chromogen spots, each connected component above the threshold is considered a separate detection. For this work, a threshold of 0.007 was chosen.

To eliminate many common types of false detection, the pixel region defined by each connected component of the thresholded image was checked against the following criteria:

- Mean saturation > 0.2
- Mean luminance > 0.2
- Mean background channel intensity > 0.8

Detections failing to meet each criterion were excluded from the final result.

Spot density was calculated for sections marked with APOBEC3A or APOBEC3B. The spot density of a single detection was computed over a circular region, with a radius $r = 429.44$μm, centred on the detection. Thus, for a spot, s, the density, $D(s)$, was calculated as:

$$D(s) = \frac{N(s)}{\pi r^2}$$

where $N(s)$ is the number of spots within the circular region surrounding s. The density for the section was computed as the mean of individual spot densities.

The procedure for the detection of lymphocytes follows a similar principal to that of spot detection. The intensity of the Haematoxylin stain is extracted from the image using stain matrix M. A LoG filter is then applied to reduce background noise. For a visual field with a resolution of 0.22μm/pixel, we recommend a LoG filter with $\sigma = 10$. The LoG filtered Haematoxylin channel is thresholded to isolate individual nuclei, each connected component above the threshold is considered a separate detection. A threshold of 0.001 was chosen for this experiment. This process produces the initial set of detected cells. In order to exclude other cell types, the following criteria are applied:

- Mean saturation > 0.2
- 0.2 < Mean luminance < 0.4
- Detection eccentricity < 0.7

Only detections meeting all criteria are selected as lymphocytes.

Supplementary Figures and Tables

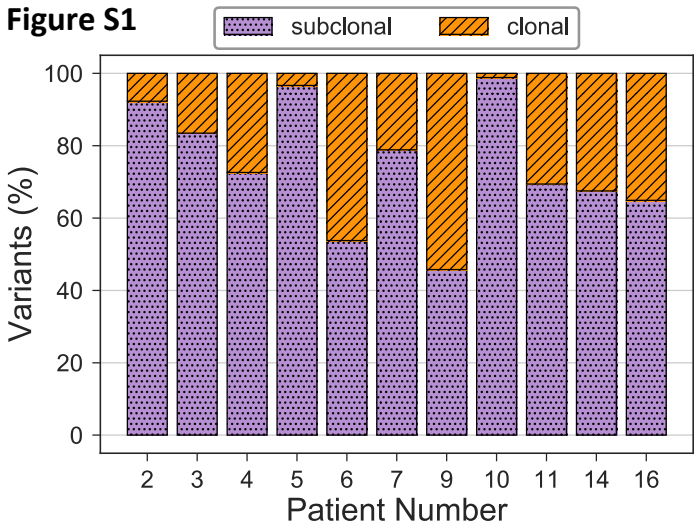

Figure S1. Proportion of clonal and subclonal variants identified by whole exome sequencing. Clonal variants are defined as mutations exhibiting CCF>0.5 in each of the patient samples. 73.5% of all mutations identified were subclonal**.** We note that for tumours with very low mutational burden, like 5 and 10, the number of clonal mutations is very small, although all samples do share similar copy number profiles, as illustrated in Figure 2B.

Figure S2

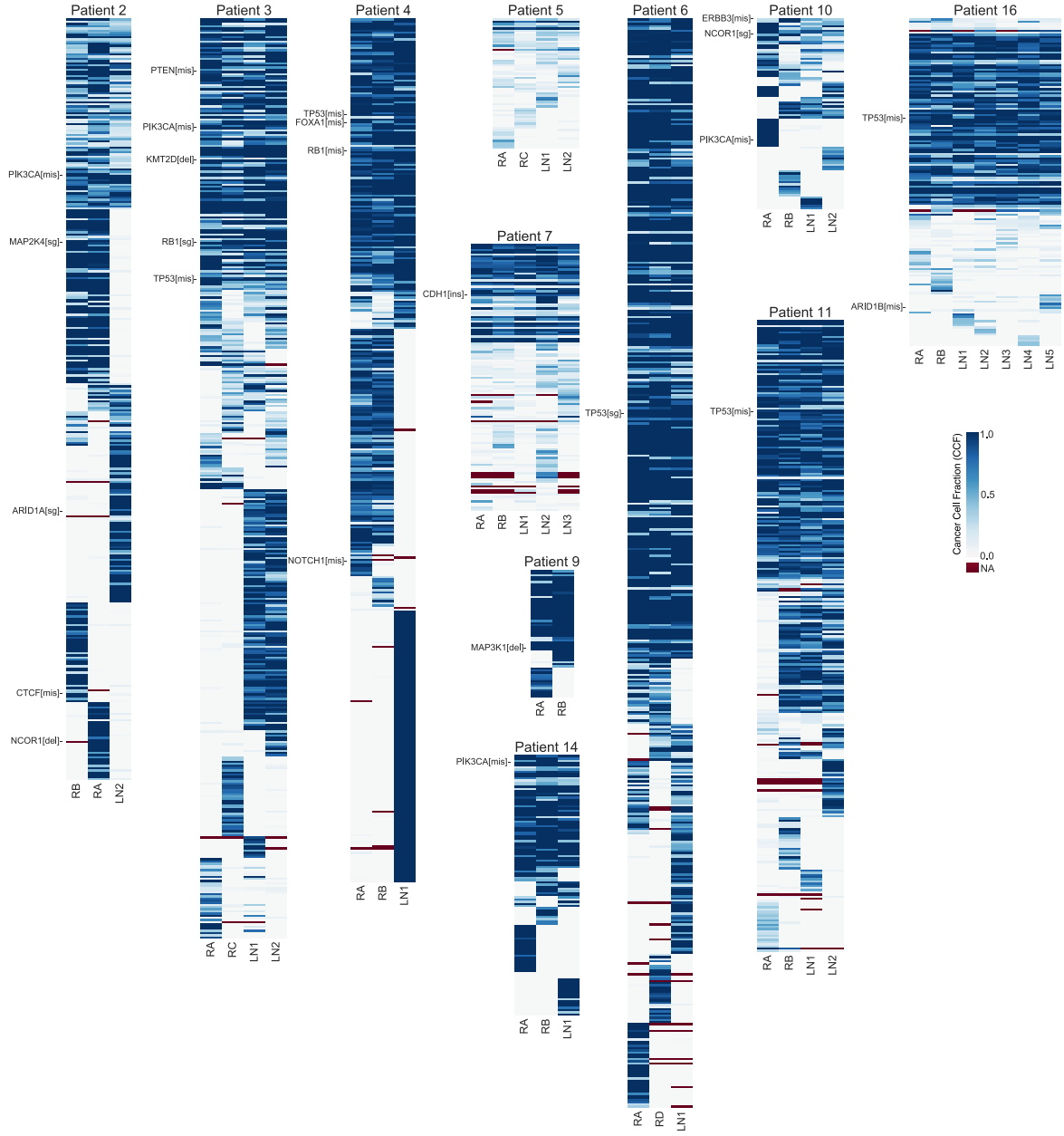

Figure S2. Mutation cancer cell fractions from whole-exome sequencing.
Heatmaps indicate cancer cell fractions of mutations in all whole-exome sequencing samples per patient. Putative drivers are annotated at the left of each heatmap.

(attached separately due to size)
Figure S3. Log-R-ratios and B-allele-frequency values.
Log-R-ratios (left) and B-Allele-Frequency (BAF-right) values for each whole exome sequencing samples estimated by PureCN. ASCAT estimated purity and ploidy is shown on the top of each graph.

Figure S4

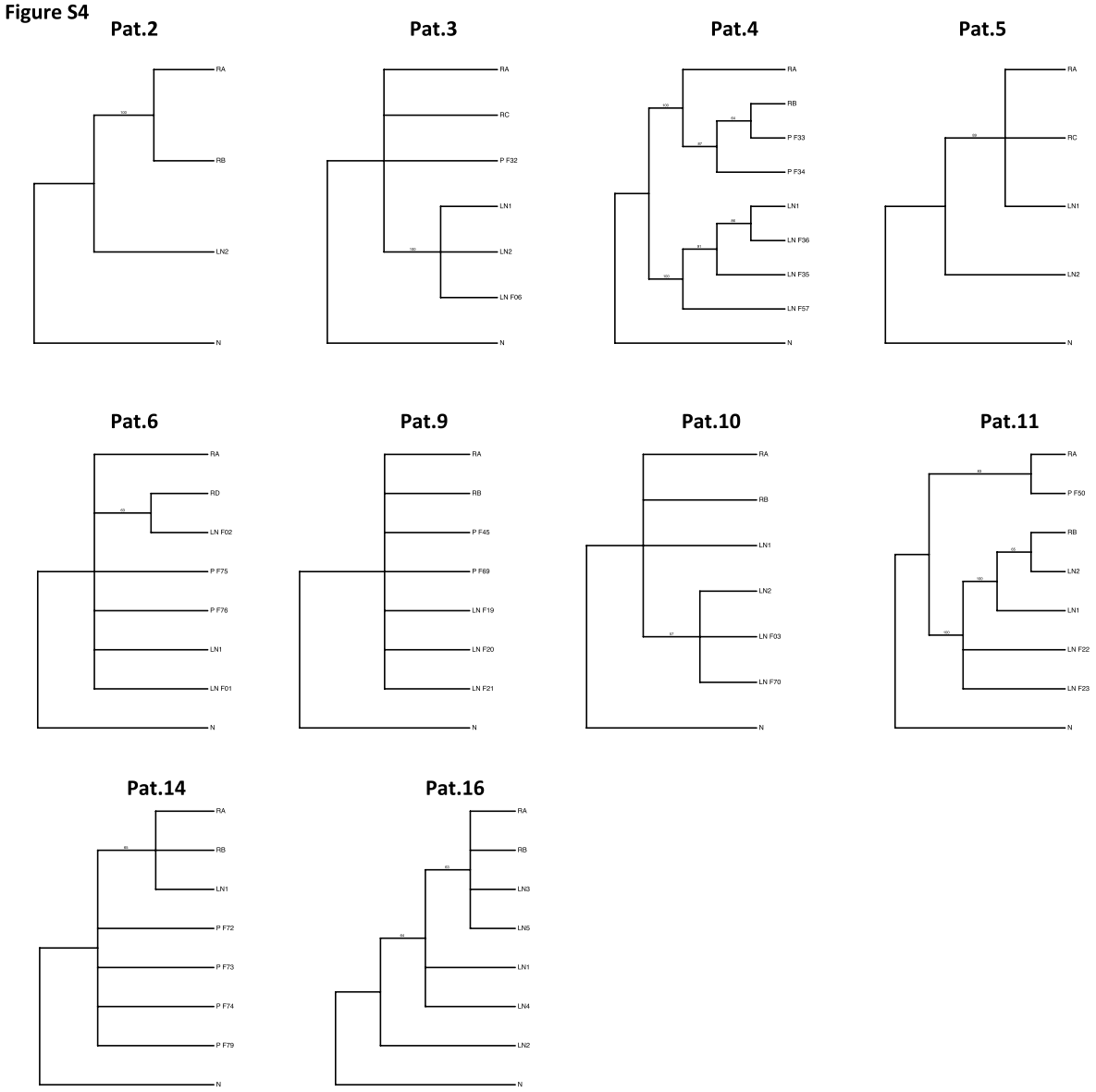

Figure S4. Bootstrap values for targeted sequencing phylogenetic trees. Bootstrap values of phylogenetic trees reconstructed with maximum parsimony for each patient from mutations validated with TES.

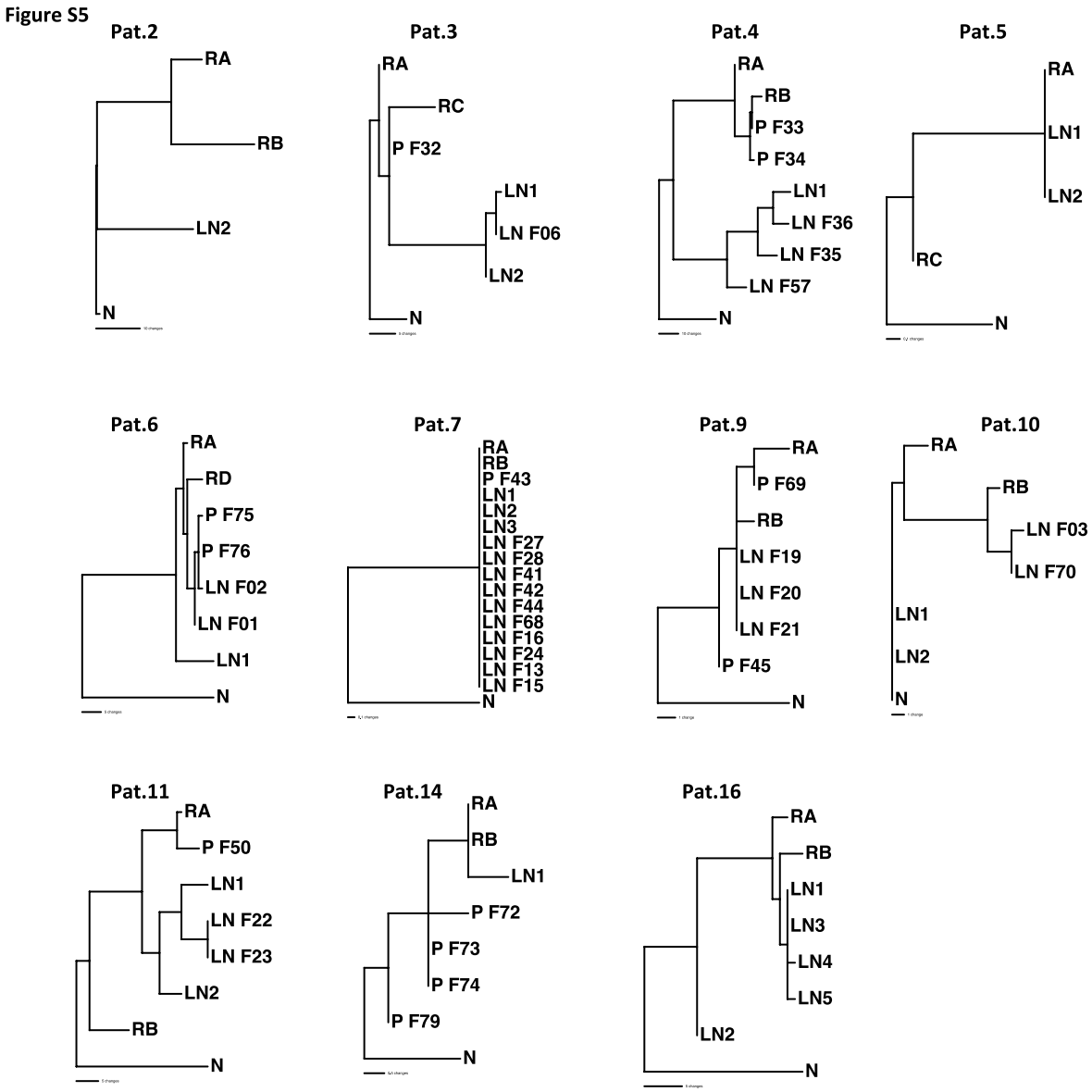

Figure S5. Phylogenetic trees reconstructed from WES data.
Phylogenetic trees reconstructed with maximum parsimony for each patient from the whole-exome sequencing samples with the corresponding bootstrapping values at each branch. The tree topologies were recapitulated using targeted sequencing (Fig. 3A), confirming that early divergence was not due to sampling bias.

Figure S6

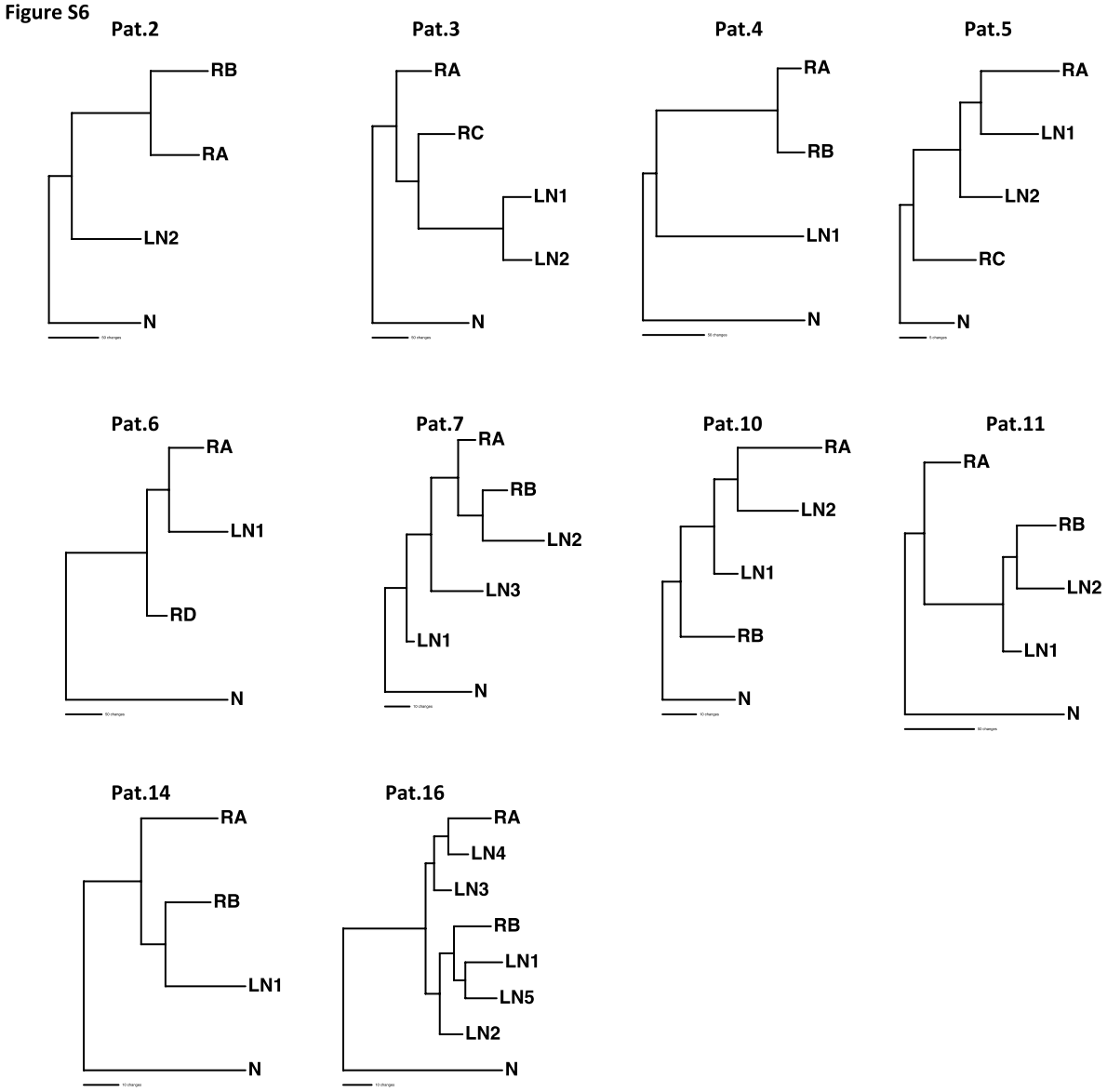

Figure S6. Phylogenetic trees reconstructed from TES data using only clonal mutations in each sample (CCF>80%).

Phylogenetic trees reconstructed with maximum parsimony for each patient from the targeted sequencing samples by selecting CCF>80% to assess confounding factors of subclones within samples. All trees recapitulate targeted sequencing trees in Fig. 3A.

Figure S7

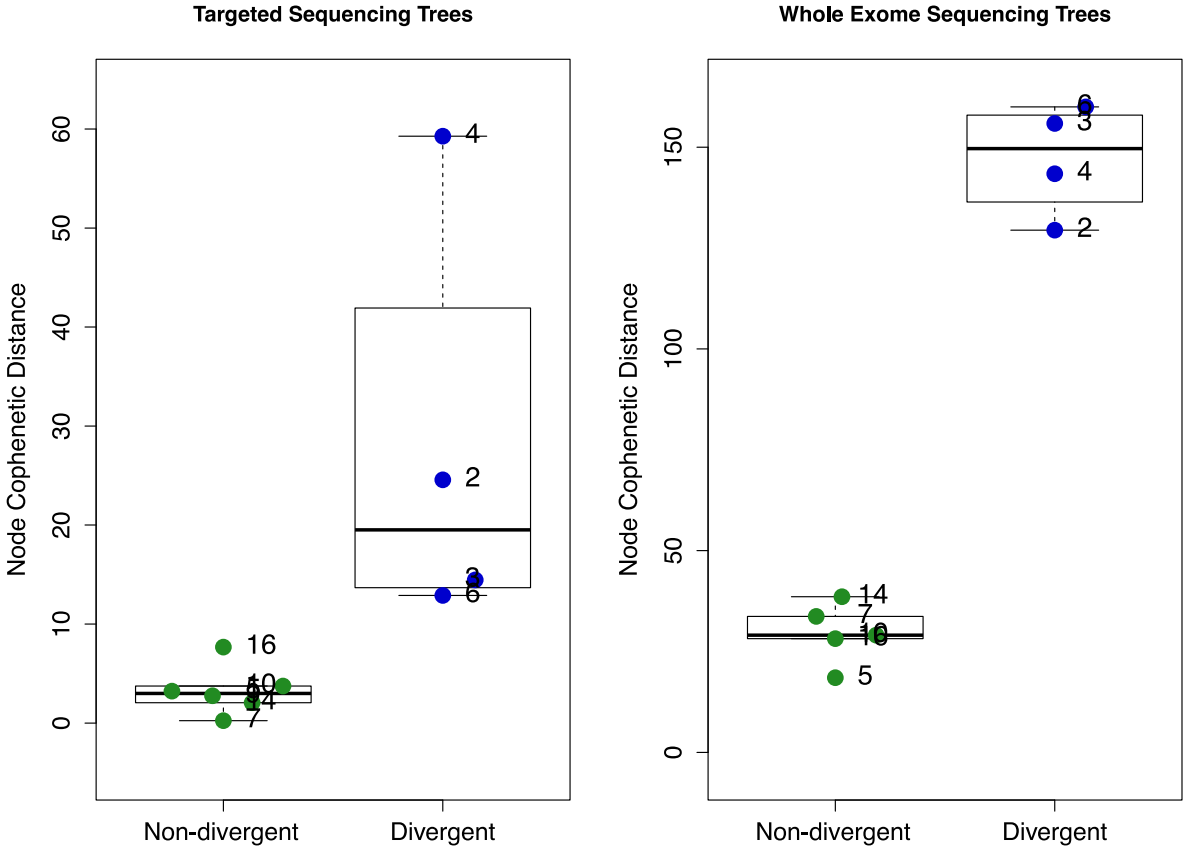

Figure S7. Measure of phylogenetic divergence. The level of phylogenetic divergence quantified using the Node Cophenetic Distance confirmed the high level of divergence in the divergent cases, both for TES as well as WES data.

(attached separately due to size)
Figure S8. CCF comparisons between multiple tumour samples. CCF comparisons, points coloured in red are present in a CNA segment called as subclonal in at least one tumour sample.

Table S1: Clinical information.

Table S2: Purity and ploidy estimates per sample. Three estimations of purity per samples are included: (1) purity estimation by pathologist review; (2) purity estimation using a PCDM; (3) purity and ploidy estimates derived from ASCAT. Additionally, purity and ploidy parameter limits for ASCAT are also provided.

Table S3: Targeted sequencing mutations variant allele frequencies used to calculate cancer cell fractions.

Table S4: Targeted sequencing mutations cancer cell fractions used in Figure 2A and 4A.

Table S5: Whole-exome sequencing mutations variant allele frequencies used to calculate cancer cell fractions.

Table S6: Whole-exome sequencing mutations cancer cell fractions used in Figure S2.

References

1. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci USA. National Academy of Sciences; 2010;107:16910–5.

2. Riester M, Singh AP, Brannon AR, Yu K, Campbell CD, Chiang DY, et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med. 2016;11:13.

3. van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B. CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics. Oxford University Press; 2007;23:892–4.

4. Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Göransson H, Juliusson G, et al. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol. BioMed Central; 2008;9:R136.

5. Nilsen G, Liestøl K, Van Loo P, Vollan HKM, Eide MB, Rueda OM, et al. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics 2012 13:1. BioMed Central; 2012;13:591.

6. Benaglia T, Chauveau D, Hunter DR, Young D. mixtools: An RPackage for Analyzing Finite Mixture Models. J Stat Soft. 2009;32.

7. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell. 2012;149:994–1007.

8. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Annals of Oncology. 2014;26:64–70.

9. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014;15:182.

10. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

11. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics. Nature Publishing Group; 2014;46:912–8.

12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. 2013;31:213–9.

13. Münz M, Ruark E, Renwick A, Ramsay E, Clarke M, Mahamdallie S, et al. CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting. Genome Med. 2015;7:76.

14. McLaren W, Gil L, Hunt SE, Riat HS. The Ensembl Variant Effect Predictor. Genome …. 2016.

15. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011;29:24–6.

16. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.

17. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486:395–9.

18. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534:47–54.

19. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med. 2015;21:751–9.

20. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.

21. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. PNAS. 2016;:201522203.

22. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016;17:31.

23. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. Nature Publishing Group; 2013;500:415–21.

24. Swofford DL. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0 Beta [Internet]. Sinauer, Sunderland, MA. 2005 [cited 2015 Mar 9]. Available from: http://www.sinauer.com/paup-phylogenetic-analysis-using-parsimony-and-other-methods-4-0-beta.html

25. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol. 2001;23:291–9.

26. Smal I, Loog M, Niessen W, Meijering E. Quantitative comparison of spot detection methods in fluorescence microscopy. IEEE Trans Med Imaging. 2010;29:282–301.