**Supplementary Information**

**Lobular carcinomas *in situ* display intra-lesion genetic heterogeneity and clonal evolution in the progression to invasive lobular carcinoma**

Lee et al.

**Supplementary Methods**

**Legends for Supplementary Tables S1-S2**

**Legends for Supplementary Figures S1-S4**

**Legends for Supplementary Data File 1**

**SUPPLEMENTARY METHODS**

**Subjects and samples**

This study is based on the 24 cases with available WES out of the 30 cases previously subjected to microarray-based comparative genomic hybridization and/or WES by Begg *et al* (**Table 1**)*.* Eight out of the 24 cases included in this study were also included in the targeted sequencing analysis previously reported by Sakr *et al*. (**Table 1**). These samples were collected through a Memorial Sloan Kettering Cancer Center (MSKCC) institutional review board-approved protocol (IRB 01-135). Owing to the challenges for the procurement of fresh/ frozen samples of LCIS and related synchronous lesions, in 2005 we established this prospective protocol for the systematic collection of frozen samples from patients with a pre-operative documented history of LCIS, presenting for prophylactic or therapeutic mastectomy. The cases subjected to WES were from patients enrolled in the study from 2008-2014 and include 43 LCIS and 27 synchronous more clinically-advanced lesions (**Table 1**). Briefly, following standard clinical sampling, mastectomy specimens were subjected to random sampling, and up to 10 fresh frozen blocks per quadrant were harvested and stored at -80°C for subsequent analysis. Five-micron hematoxylin-and-eosin frozen sections were reviewed by two study pathologists (DDG and JVSC) to identify LCIS, DCIS and/or invasive cancer, according to the World Health Organization Classification of Tumors of the Breast (1). The methods for DNA extraction and WES are described in Begg *et al.* (2).

**Immunohistochemistry**

The immunohistochemical profile of the invasive lesions was assessed on 4μm-thick sections, using antibodies against estrogen receptor (ER), progesterone receptor (PR) and HER2 essentially as previously described (3). E-cadherin slides were also reviewed when available. Positive and negative controls were included in each slide run. The results of ER, PR, HER2 and E-cadherin immunohistochemistry were evaluated by two pathologists according to the American Society of Clinical Oncology (ASCO)/ College of American Pathologists (CAP) guidelines (4, 5).

**Whole-exome sequencing (WES) data analysis**

WES data were retrieved from Begg et *al.* (2) and analyzed as described in Ng *et al.* (6). In brief, reads were aligned to the reference human genome GRCh37 using the Burrows-Wheeler Aligner (BWA, v0.6.2) (7). Local realignment and quality score recalibration were performed using the Genome Analysis Toolkit (GATK, v3.1.1). Deduplication was performed using Picard (v1.92) (8). Somatic single nucleotide variants (SNVs) were identified using MuTect (v.1.1.4) (9), and small somatic insertions and deletions (indels) were defined using VarScan2 (v2.3.6) (10) and Strelka (v3.1.1) (11). In order to define long somatic indels, Lancet, Pindel (12), Platypus (13), and Scalpel (14) were employed. Variants found with >5% global minor allele frequency in dbSNP (Build 138) or that were covered by <10 reads were disregarded. Variants for which the tumor variant allele fraction was <5 times that of the normal variant allele fraction were disregarded. All indels were manually inspected using the Integrative Genomics Viewer (IGV) (15). In addition to the identification of SNVs and indels described above, for samples from a given patient, mutations that were identified in at least one sample were subsequently interrogated in all related samples by manual inspection of BAM files using mpileup files generated from SAMtools mpileup (version 1.2 htslib 1.2.1)(8), and all variants supported by at least three reads or variant allele frequency with more than 2% were considered present at this stage.

Given that *CDH1* germline mutations have been shown to be causative of familial gastric and breast cancer syndrome (16), the germline DNA samples from each patient were evaluated for the presence of *CDH1* germline mutations. In brief, germline variants were identified using HaplotypeCaller (GATK v3.1.1) using standard settings. Germline variants found with >5% global minor allele frequency in dbSNP (Build 138) or that were covered by <10 reads were disregarded. All putative missense, truncating and frameshift *CDH1* germline variants were manually inspected using the IGV (15) and were further investigated as pathogenicity in the ClinVar database (17). Mutation function prediction of the somatic mutations identified was performed using a combination of MutationTaster (18), FATHMM (19), PROVEAN (20) and CHASM (breast classifier) (21), as previously described (22, 23), and the genes affected were annotated according to their presence in three cancer gene datasets, Kandoth *et al*. (24), the Cancer Gene Census (25) and Lawrence *et al*. (26).

Copy number analysis was performed using FACETS (27), as previously described (23), and purity and ploidy estimations as well as modal copy number were calculated using ABSOLUTE (28), based on segmented copy number data from FACETS and mutant allelic fractions identified on the basis of the methods outlined above. In order to determine whether a gene harboring a somatic mutation was also targeted by loss of heterozygosity, we used FACETS, as previously described (29), followed by manual curation of the results.

**Clonality analysis**

To infer the clonal relatedness between synchronous lesions, we defined the “clonality index” (CI) as the probability of two lesions sharing mutations not expected to have co-occurred by chance based on a previously reported approach (30). Briefly, all somatic mutations were included and at least one non-synonymous mutation was required to be shared in two lesions compared. Given that this analysis is potentially confounded by the presence of highly recurrent somatic mutations, such as *PIK3CA* hotspot mutations, the CI provides an adjustment for the presence and frequency of a given somatic mutation, including synonymous and non-synonymous, in ILCs dataset (n=127) from TCGA (31). Adopting previous approach (30), we defined $CI=-log\_{10}\prod\_{m=1}^{M}P(X)\_{m}. $ Given the repertoire of mutations of two samples, the probability of observing a given mutation in both samples is defined by the binomial probability , *n*=2, *k*=2, where *p* is the percentage of ILCs harboring a given mutation and *n* is the number of shared mutations between a pair of lesions or the average number of mutations found in the two samples in the target regions divided by the size of the target regions. Thus, the probability of observing a given set of *M* identical mutations in the two samples is given by $\prod\_{m=1}^{M}P(X)\_{m}$. To objectively define a cut-off for clonal relatedness, we used the mutational data from the 127 unrelated from TCGA (31). As the positive control (i.e. clonally related), we randomly selected 40%, 60% and 80% of the set of mutations from the 127 unrelated cases in duplicate to simulate heterogeneity between biologically related samples. As the negative control (i.e. unrelated), we randomly selected an equivalent number of pairs (i.e. 3x127=381) of unrelated from TCGA. To define the optimum cut-offs, the R package ‘ROCR’ was used to maximize accuracy. To avoid over-fitting of data, the above procedures were repeated 100 times to define the median and 95% confidence interval of the optimum cut-off. The median optimum cut-off was 19.98 (95% confidence interval 18.54-20.53), and with median accuracy of 85.75% (95% confidence interval 83.07%-90.5%), the median sensitivity was 88.50% (95% confidence interval 83.64%-93.18%), and the median specificity 95.14% (95% confidence interval 87.83%-99.1%).

**Clonal frequencies**

To estimate the clonal architecture and composition of the lesions from each patient, mutant allelic fractions from all somatic mutations were adjusted for tumor cell content, ploidy, local copy number and sequencing errors using PyClone (32). To convert the mutant allelic fraction measurements into estimates of clonal frequencies jointly for all lesions from a given patient, we employed the Dirichlet process clustering model described by Roth *et al.* (32) that simultaneously estimates the genotype and clonal frequency given a list of somatic mutations and their local copy number. Purity and ploidy estimates, as well as modal copy number from ABSOLUTE (28) were employed as the input data for PyClone analysis (32). 10,000 iterations of Markov Chain Monte Carlo (MCMC) were performed with the first 1,000 iterations discarded as “burn-in”. Clustering was performed using the “mpear” algorithm as implemented in PyClone. Clusters whose clonal frequencies were <5% and clusters that were composed of single mutation were discarded.

**Truncal and branch mutations**

For each patient displaying at least one LCIS sample clonally-related to DCIS or ILC, we categorized the mutations into truncal and branch. Truncal mutations were defined as those concurrently present in the modal populations of all LCIS and their clonally-related DCIS/ILC from a given patient. Three classes of branch mutations were defined: i) all branch mutations, which comprise all non-truncal mutations, ii) branch mutations in DCIS/ILC, which encompass the subset of branch mutations present in the subsequent tumor clones found in the DCIS/ILC, and iii) LCIS branch mutations, which comprise the subset of branch mutations present in subsequent tumor clones found in the LCIS. The definition of truncal mutations included a modal clonal frequency (33) as defined by PyClone (32) in all clonally-related lesions within the set of samples from a given patient, and branch mutations were defined as mutations that were not truncal (i.e. mutations that were not found in at least one of the clonally-related lesions, and/or were present at subclonal frequencies in at least one of the clonally-related lesions within the set of samples from a given patient as previously described (33)).

**Measure of diversity**

To quantitate the intra-lesion genetic heterogeneity of each sample analyzed, we used the Shannon diversity index (34) and Gini-Simpson index (35), as previously described (36-38). Shannon index is borrowed from information theory, where it specifies the information content of a message, and can be used to summarize the diversity of a population by a single number. It is defined as H = $-\sum\_{i=1}^{n}p\_{i}x ln(p\_{i})$, where *H* is the Shannon index metric, *pi* is the percentage of a subpopulation in the overall population and *n* is the number of subpopulations. Simpson index is defined as the probability that two entities taken randomly from the dataset of interest represent different types as shown by the formula D = $1-\sum\_{i=1}^{n}p\_{i}^{2}$ where D is the Gini-Simpson index metric, *pi* is the percentage of a subpopulation in the overall population and *n* is the number of subpopulations. In this study for both the Shannon and Gini-Simpson indices, *p*i and *n* were defined as the percentage of a genetically distinct subclone within a lesion and the number of subclones, respectively, derived from the tumor clone structure inferred using Pyclone (32). For the Shannon and Gini-Simpson indices, values ranged from 0 to 1.05 and 0 to 0.64, respectively, in our dataset, where 0 (both indices) represents a sample with no heterogeneity (presence of only one clone), and 1.05 (Shannon index) and 0.64 (Gini-Simpson index) represent the most heterogeneous sample in this study (sample 10LCIS-A, which was composed of 3 subclones).

**Phylogenetic tree construction**

Maximum parsimony trees were built using binary presence/absence matrices built from the somatic genetic alterations, including synonymous and non-synonymous SNVs, indels and CNAs, within the clonally-related lesions from each patient, essentially as described by Murugaesu et al. (39). Tumors with subclones that have undergone parallel evolution were treated independently to enable the construction of maximum parsimony trees. A starting tree was constructed using the Neighbor-joining method and Hamming distance and optimized using the parsimony ratchet method (40) implemented in the R package Phangorn (41). Trees were rooted at the normal lobule. Branch lengths were determined according to the ACCTRAN criterion as implemented in the Phangorn package and were drawn to scale in all figures. We have also employed TreeOmics as an alternative approach for the reconstruction of phylogenetic tree (42). Treeomics reconstructs phylogenies using a Bayesian inference model and determines the probability that a variant is either present or absent in a given sample.

**Reverse transcription quantitative PCR (RT-qPCR)**

Total RNA was extracted with the TRIZOL method and reverse transcribed using SuperScript VILO Master Mix (Life Technologies) according to the manufacturer’s instructions from cases where representative frozen tissue samples were available. Quantitative RT-PCR (qRT-PCR) was performed to analyze the expression levels of *APOBEC3B*, *APOBEC3H* and *REV1* genes using TaqMan Assay-on-Demand (IDs: Hs00358981\_m1; Hs00419665\_m1 and Hs00249411\_m1, respectively; Applied Biosystems) on the StepOnePlus Real-Time PCR System (Applied Biosystems) as previously described (43) and quantified using the comparative ΔΔCt method (44). *GAPDH* (ID: Hs99999905\_m1) was employed for assay normalization. All reactions were performed in triplicate. Statistical comparisons were performed using a Mann–Whitney U test (GraphPad Prism 6).

**SUPPLEMENTARY REFERENCES**

1. Lakhani SR, Ellis IO, Schnitt SJ, Tan PH, van de Vijver MJ. WHO classification of tumours of the breast. Lyon: IARC Press; 2012.

2. Begg CB, Ostrovnaya I, Carniello JV, Sakr RA, Giri D, Towers R, et al. Clonal relationships between lobular carcinoma in situ and other breast malignancies. Breast cancer research : BCR. 2016;18:66.

3. Andrade VP, Morrogh M, Qin LX, Olvera N, Giri D, Muhsen S, et al. Gene expression profiling of lobular carcinoma in situ reveals candidate precursor genes for invasion. Mol Oncol. 2014:10.1016/j.molonc.2014.12.005.

4. Hammond ME, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, 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. 2010;28:2784-95.

5. Wolff AC, Hammond ME, Hicks DG, Dowsett M, McShane LM, Allison KH, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol. 2013;31:3997-4013.

6. Ng CKY, Bidard FC, Piscuoglio S, Geyer FC, Lim RS, de Bruijn I, et al. Genetic Heterogeneity in Therapy-Naive Synchronous Primary Breast Cancers and Their Metastases. Clin Cancer Res. 2017;23:4402-15.

7. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589-95.

8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-9.

9. 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.

10. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research. 2012;22:568-76.

11. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20:1297-303.

12. Li S, Li R, Li H, Lu J, Li Y, Bolund L, et al. SOAPindel: efficient identification of indels from short paired reads. Genome research. 2013;23:195-200.

13. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Consortium WGS, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature genetics. 2014;46:912-8.

14. Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee YH, Wang Z, et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nature methods. 2014;11:1033-6.

15. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013;14:178-92.

16. Richards FM, McKee SA, Rajpar MH, Cole TR, Evans DG, Jankowski JA, et al. Germline E-cadherin gene (CDH1) mutations predispose to familial gastric cancer and colorectal cancer. Human molecular genetics. 1999;8:607-10.

17. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2014;42:D980-5.

18. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nature methods. 2010;7:575-6.

19. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human mutation. 2013;34:57-65.

20. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31:2745-7.

21. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer research. 2009;69:6660-7.

22. Martelotto LG, Ng C, De Filippo MR, Zhang Y, Piscuoglio S, Lim R, et al. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol. 2014;15:484.

23. Ng CKY, Piscuoglio S, Geyer FC, Burke KA, Pareja F, Eberle CA, et al. The Landscape of Somatic Genetic Alterations in Metaplastic Breast Carcinomas. Clin Cancer Res. 2017;23:3859-70.

24. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502:333-9.

25. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nature reviews Cancer. 2004;4:177-83.

26. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505:495-501.

27. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44:e131.

28. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology. 2012;30:413-21.

29. Weigelt B, Bi R, Kumar R, Blecua P, Mandelker DL, Geyer FC, et al. The Landscape of Somatic Genetic Alterations in Breast Cancers From ATM Germline Mutation Carriers. J Natl Cancer Inst. 2018.

30. Schultheis AM, Ng CK, De Filippo MR, Piscuoglio S, Macedo GS, Gatius S, et al. Massively Parallel Sequencing-Based Clonality Analysis of Synchronous Endometrioid Endometrial and Ovarian Carcinomas. J Natl Cancer Inst. 2016;108:djv427.

31. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015;163:506-19.

32. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nature methods. 2014;11:396-8.

33. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014;346:251-6.

34. Shannon CE. The mathematical theory of communication. 1963. MD computing : computers in medical practice. 1997;14:306-17.

35. Simpson EH. Measurement of diversity. Nature. 1949.

36. Almendro V, Kim HJ, Cheng YK, Gonen M, Itzkovitz S, Argani P, et al. Genetic and phenotypic diversity in breast tumor metastases. Cancer research. 2014;74:1338-48.

37. Begg CB, Eng KH, Hummer AJ. Statistical tests for clonality. Biometrics. 2007;63:522-30.

38. Ostrovnaya I, Seshan VE, Begg CB. Comparison of properties of tests for assessing tumor clonality. Biometrics. 2008;64:1018-22.

39. Murugaesu N, Wilson GA, Birkbak NJ, Watkins T, McGranahan N, Kumar S, et al. Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. Cancer discovery. 2015.

40. Nixon KC. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics. 1999;15:407-14.

41. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27:592-3.

42. Reiter JG, Makohon-Moore AP, Gerold JM, Bozic I, Chatterjee K, Iacobuzio-Donahue CA, et al. Reconstructing metastatic seeding patterns of human cancers. Nat Commun. 2017;8:14114.

43. Piscuoglio S, Ng CK, Martelotto LG, Eberle CA, Cowell CF, Natrajan R, et al. Integrative genomic and transcriptomic characterization of papillary carcinomas of the breast. Mol Oncol. 2014;8:1588-602.

44. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25:402-8.

45. Sakr RA, Schizas M, Carniello JV, Ng CK, Piscuoglio S, Giri D, et al. Targeted capture massively parallel sequencing analysis of LCIS and invasive lobular cancer: Repertoire of somatic genetic alterations and clonal relationships. Mol Oncol. 2016;10:360-70.

**SUPPLEMENTARY FIGURES LEGENDS**

**Supplementary Fig. S1: Repertoire of somatic mutations, clonal composition and phylogenetic tree analysis of cases 1-3 and cases 5-9.**

On the top left of each subfigure, a schematic representation illustrating the quadrants from which each sequenced lesion was sampled is shown, and on the top right the clonal frequency heatmap of mutations within the lesions of each case, grouped by their inferred clonal/ subclonal structure (clusters). The clusters inferred by PyClone are indicated below the clonal frequency heatmap. Only genes affected by non-synonymous somatic mutations are shown in the clonal frequency heatmap. The Shannon index measuring intra-lesion genetic heterogeneity for each lesion is specified within parentheses following the sample name in the clonal frequency heatmap. On the bottom of each subfigure, the parallel\_coordinates plot generated by PyClone is displayed on the left, and the matching phylogenetic tree based on the clusters identified by PyClone on the right, where the color of the trunk and branches matches the color of their respective clusters shown in the parallel\_corrdinates plot. Below the PyClone cluster-based tree, a histologic lesion-based phylogenetic tree constructed using Treeomics is shown (42). In both phylogenetic trees, the mutations affecting cancer genes (colored in orange) and the hotspot mutations (colored in blue) that define a given clone are illustrated alongside the branches. The length of the branches is proportional to the number of mutations that distinguish a given clone from its ancestor. The numbers alongside the branches represent the total number of somatic mutations. **A)** Case 1. **B)** Case 2. **C)** Case 3. **D)** Case 5. **E)** Case 6. **F)** Case 7. **G)** Case 8. **H)** Case 9

**Supplementary Fig. S2: Repertoire of somatic mutations, clonal composition and phylogenetic tree analysis of cases 11-18.**

On the top left of each subfigure, a schematic representation illustrating the quadrants from which each sequenced lesion was sampled is shown, and on the top right the clonal frequency heatmap of mutations within the lesions of each case, grouped by their inferred clonal/ subclonal structure (clusters). The clusters inferred by PyClone are indicated below the clonal frequency heatmap. Only genes affected by non-synonymous somatic mutations are shown in the clonal frequency heatmap. The Shannon index measuring intra-lesion genetic heterogeneity for each lesion is specified within parentheses following the sample name in the clonal frequency heatmap. On the bottom of each subfigure, the parallel\_coordinates plot generated by PyClone is displayed on the left, and the matching phylogenetic tree based on the clusters identified by PyClone on the right, where the color of the trunk and branches matches the color of their respective clusters shown in the parallel\_corrdinates plot. Below the PyClone cluster-based tree, a histologic lesion-based phylogenetic tree constructed using Treeomics is shown (42). In both phylogenetic trees, the mutations affecting cancer genes (colored in orange) and the hotspot mutations (colored in blue) that define a given clone are illustrated alongside the branches. The length of the branches is proportional to the number of mutations that distinguish a given clone from its ancestor. The numbers alongside the branches represent the total number of somatic mutations **A)** Case 11. **B)** Case 12. **C)** Case 13. Note that no phylogenetic trees could be constructed as the lesions are not clonally related. **D)** Case 14. **E)** Case 15. **F)** Case 16. **G)** Case 17. **H)** Case 18.

**Supplementary Fig. S3: Repertoire of somatic mutations, clonal composition and phylogenetic tree analysis of cases 19-24.**

On the top left of each subfigure, a schematic representation illustrating the quadrants from which each sequenced lesion was sampled is shown, and on the top right the clonal frequency heatmap of mutations within the lesions of each case, grouped by their inferred clonal/ subclonal structure (clusters). The clusters inferred by PyClone are indicated below the clonal frequency heatmap. Only genes affected by non-synonymous somatic mutations are shown in the clonal frequency heatmap. The Shannon index measuring intra-lesion genetic heterogeneity for each lesion is specified within parentheses following the sample name in the clonal frequency heatmap. On the bottom of each subfigure, the parallel\_coordinates plot generated by PyClone is displayed on the left, and the matching phylogenetic tree based on the clusters identified by PyClone on the right, where the color of the trunk and branches matches the color of their respective clusters shown in the parallel\_corrdinates plot. Below the PyClone cluster-based tree, a histologic lesion-based phylogenetic tree constructed using Treeomics is shown (42). In both phylogenetic trees, the mutations affecting cancer genes (colored in orange) and the hotspot mutations (colored in blue) that define a given clone are illustrated alongside the branches. The length of the branches is proportional to the number of mutations that distinguish a given clone from its ancestor. The numbers alongside the branches represent the total number of somatic mutations. **A)** Case 19. Note that no phylogenetic trees could be constructed as the lesions are not clonally related. **B)** Case 20. **C)** Case 21. **D)** Case 22. **E)** Case 23. **F)** Case 24.

**Supplementary Fig. S4: Clonal relatedness between lobular carcinoma *in situ* (LCIS), invasive lobular carcinoma (ILC), ductal carcinoma *in situ* (DCIS) and invasive ductal carcinoma (IDC), Shannon and Gini-Simpson Indices measure of intra-lesion genetic heterogeneity, and truncal and branch mutations affecting cancer genes.**

**A)** Pairwise comparison of the clonality index between samples based on the presence of synonymous and non-synonymous somatic mutations. The color gradient represents the *CI (Clonal Index)* of the probability that a given pair of samples is clonally-related based on the number of shared mutations and the frequency of each shared mutation in ILCs breast cancers from The Cancer Genome Atlas breast cancer study (31). **B)** Barplot showing the proportion of LCIS harboring intra-lesion genetic heterogeneity in the LCIS clonally-related to DCIS/ILC (red bar) and LCIS not clonally-related to DCIS/ILC (gray bar) (p-value = 0.0016, Fisher’s exact test). Boxplots illustrating the distribution of the Shannon diversity index **(C)** and the Gini-Simpson diversity index **(D)** for each lesion type (LCIS, DICS, ILC and IDC). **E-F)** Truncal and branch mutations affecting cancer genes as defined by Kandoth *et al*. (24), the Cancer Gene Census (25) and/ or Lawrence *et al*. (26). Venn diagrams representing the overlap between genes affected by truncal or branch mutations and cancer genes. Hypergeometric test *P-value* and representation factors are shown below the respective Venn diagrams.

**SUPPLEMENTARY TABLES**

**Supplementary Table S1:** Sequencing statistics.

**Supplementary Table S2:** Comparison of clonal relatedness of the lesions analyzed between Begg *et al.* (2), Sakr *et al.* (45) and this study.

**SUPPLEMENTARY DATA FILE:**

**Supplementary Data File 1:** Non-synonymoussomatic mutations identified by whole-exome sequencing and amplicon sequencing in the LCIS, DCIS, ILCs and IDCs studied.