Reactivation of multipotency by oncogenic PIK3CA induces breast tumour heterogeneity

Alexandra Van Keymeulen[1, *, #], May Yin Lee[1, *], Marielle Ousset[1], Sylvain Brohée[4], Sandrine Rorive[2, 3], Rajshekhar R. Giraddi[1], Aline Wuidart[1], Gaëlle Bouvencourt[1], Christine Dubois[1], Isabelle Salmon[2, 3], Christos Sotiriou[4], Wayne A. Phillips[5,6], Cédric Blanpain[1, 7, #]

[1] Université Libre de Bruxelles, IRIBHM, Brussels B-1070, Belgium.

[2] Department of Pathology, Erasme Hospital, Université Libre de Bruxelles, Brussels, Belgium

[3] DIAPATH - Center for Microscopy and Molecular Imaging (CMMI), Gosselies, Belgium

[4] Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium.

[5]Surgical Oncology Research Laboratory, Peter MacCallum Cancer Centre, Melbourne, Australia.

[6]Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Australia.

[7] WELBIO, Université Libre de Bruxelles, Brussels B-1070, Belgium

[#] Corresponding authors

*These authors contributed equally to this work

Breast cancer is the most frequent cancer in women and consists of heterogeneous types of tumours that are classified into different histological and molecular subtypes[1,2]. *Pik3ca* and *p53* are the two most frequently mutated genes and are associated with different types of human breast cancers[3]. The cellular origin and the mechanisms leading to *Pik3ca*-induced tumour heterogeneity remain unknown. Here, we used a genetic approach in mice to define the cellular origin of *Pik3ca*-derived tumours and its impact on tumour heterogeneity. Surprisingly, oncogenic *Pik3ca*-H1047R expression at physiological levels[4] in basal cells (BCs) using K5CREERT2 induced the formation of luminal ER+PR+ tumours, while its expression in luminal cells (LCs) using K8CREERT2 gave rise to luminal ER+PR+ tumours or basal-like ER-PR- tumours. Concomitant deletion of p53 and expression of *Pik3ca*-H1047R accelerated tumour development and induced more aggressive mammary tumours. Interestingly, expression of *Pik3ca*-H1047R in unipotent BCs gave rise to luminal-like cells, while its expression in unipotent LCs gave rise to basal-like cells before progressing into invasive tumours. Transcriptional profiling of cells that have undergone cell fate transition upon *Pik3ca*-H1047R expression in unipotent progenitors demonstrate a profound oncogene-induced reprogramming of these newly formed cells and identified gene signatures, characteristic of the different cell fate switches that occur upon *Pik3ca*-H1047R expression in BC and LCs, which correlated with the cell of origin, tumour type and different clinical outcomes. Altogether our study identifies the cellular origin of *Pik3ca*-induced tumours and reveals that oncogenic *Pik3ca*-H1047R activates a multipotent genetic program in normally lineage-restricted populations at the early stage of tumour initiation, setting the stage for future intratumoural heterogeneity. These results have important implications for our understanding of the mechanisms controlling tumour

heterogeneity and the development of new strategies to block *PIK3CA* breast cancer initiation.

Breast cancers can be classified into different histological and molecular subtypes including luminal (ER+ and/or PR+), Her2+ and basal-like (BL)/triple negative (TN) (ER-PR-Her2-) cancers, which are usually associated with different gene expression and mutation profiles, prognosis and response to therapies[5]. *PIK3CA* mutations are found in about 30% of breast cancers, more frequently in luminal tumours, although they are also found in BL/TN breast cancers[3,6-9]. Expression of oncogenic *Pik3ca*-H1047R in all mammary gland (MG) lineages using MMTV-CRE[4,10-12] or preferentially in luminal progenitors using WAP-CRE[13] induces heterogeneous mammary tumours[4,10-13]. The reason for this tumour heterogeneity upon expression *Pik3ca*-H1047R mutant in the MG is currently unknown.

To determine whether breast tumour heterogeneity is determined by the cancer cell of origin, we developed a genetic strategy allowing the expression of the oncogenic *Pik3ca* mutant at physiological levels using CRE-inducible *Pik3ca*-H1047R knockin mice[4], specifically in BCs using K5CREERT2 or in LCs using K8CREERT2 mice[14] and followed their fate and tumorigenic potential over time. Tamoxifen (TAM) was administered at a dose that does not impair the long-term MG development and homeostasis and resulting in the specific labelling of about 20% of BCs (Extended data Fig. 1), to 4-5 week old K5CREERT2/*Pik3ca*-H1047R mice (Fig. 1a). While it has been suggested that MG contains bipotent basal SCs[15,16], our data using K5CREERT2 knockin or K14rtTA/TetOCRE mice, despite the labelling of 20-50% of BCs showed no contribution of BCs to the luminal lineage (Extended data Fig. 1). Further lineage tracing studies that label all BCs or all LCs will be required to determine whether the discrepancy between the different studies arise from the unspecific and simultaneous labelling of BC and LCs. BC derived mammary tumours arose

with a latency of about 12±4 months (mean±SD) and were all luminal-like tumour cells expressing ER+, PR+, surrounded by BCs (Fig. 1c,e and Extended data Fig. 2a-e), classified by pathologists as adenomyoepithelioma[17] in mouse and humans[17] (Extended data Fig. 3a-d). Principal component analysis (PCA) and gene clustering analysis of gene expression profile from FACS isolated tumour cells using the PAM50 gene set showed that these BC derived-tumours clustered together with the luminal B breast cancer subtype (Extended data Fig. 3, 4).

The same dose of TAM was administered to 4-5 week old K8CREERT2/*Pik3ca*-H1047R mice resulting in the specific labelling of about 20-30% of LCs (Fig. 1b and Extended data Fig. 1). Mammary tumours arose with a similar latency (15±4 months) (Fig. 1d). Histological and immunofluorescence analysis revealed that these tumours were more heterogeneous, more aggressive and more proliferative than BC derived-tumours. These tumours comprised adenomyoepithelioma, mixed adenomyoepithelioma with myoepithelial carcinoma, invasive carcinoma of no special type (NST), as well as tumours which show features of metaplastic BL breast cancers similar to human breast cancers (Fig. 1f and Extended data Fig. 2, 3). Principal component and gene expression clustering analysis from cells isolated from 7 different luminal derived-tumours showed that ER+ tumour clustered together with luminal human breast cancers, NST clustered in between luminal B and Her2, and metaplastic carcinoma clustered with BL or Her2+ cancers depending on the clustering algorithm (Extended data Fig.3j, l), consistent with their phenotypic heterogeneity. These results revealed that *Pik3ca*-H1047R expression in the LC gives rise to distinct types of tumours that are generally more aggressive as compared to BCs derived-tumours. The greater tumour heterogeneity found in the luminal derived-tumours may arise from greater plasticity of LC and/or the heterogeneity of the luminal progenitor populations initially targeted by the K8CREERT2.

We then assessed whether concomitant *p53* deletion affects the phenotype of mammary tumours depending on their cellular origin. K5CREERT2/*Pik3ca*-H1047R/*p53*fl/fl mice treated with TAM rapidly developed skin and other cancers that required terminating the experiment before they developed mammary tumours (data not shown). To circumvent this problem, we used mice heterozygous for p53 K5CREERT2/*Pik3ca*-H1047R/*p53*fl/+ and another basal CRE driver (K14rtTA/TetOCre/*Pik3ca*-H1047R/p53fl/+) that alleviated the increased early mortality seen with the K5CREERT2/*Pik3ca*-H1047R/*p53*fl/fl. Basal cell derived-p53 heterozygous tumours arose with a latency of 9±2 months and consisted mostly of adenomyoepithelioma luminal-like tumours (42-75%), as well as myoepithelial carcinoma (0-16%), NST (0-12%) and metaplastic carcinoma (12-42%) (Fig. 2 and Extended data Fig. 5). As previously shown using MMTV-CRE[10], *Pik3ca*-H1047R expression together with *p53* deletion in LCs dramatically accelerates tumour formation, with a latency of 5±1 months for p53 homozygous and 9±3 months for p53 heterozygous mice (Fig. 2). In contrast to BCs, LCs-derived p53 deficient tumours always consisted of aggressive carcinomas consisting mostly of metaplastic carcinoma and high-grade myoepithelial carcinoma with characteristics of epithelial to mesenchymal transition (Fig. 2 and Extended data Fig. 5), as previously reported following *Pik3ca*-H1047R expression in all MG cells[10,12] and found in human BL breast cancers with activation of the PI3K pathway by somatic *PIK3CA* mutations and gene copy number amplification[3,7-9]. Gene expression clustering of these tumours using the PAM50 genes showed that they clustered together with human BL or Her2 subtypes depending on the clustering algorithm (Extended data 3k, l). These data demonstrate that concomitant *Pik3ca*-H1047R expression and *p53* deletion accelerates tumour development in basal and luminal lineages and induced more frequently very aggressive metaplastic tumours following oncogenic targeting of LCs.

To further define the cellular basis of intratumoral heterogeneity found in *Pik3ca*-H1047R-derived tumours, we combined Rosa-YFP lineage tracing and *Pik3ca*-H1047R expression specifically in LCs or BCs and assessed cell fate change over time. Very interestingly, as early as 5 weeks after *Pik3ca*-H1047R expression in LCs, YFP was also detected in basal-like cells clustered around LCs (Fig. 3a-e and Extended data Fig. 6), while, as previously described[14], K8CREERT2-targeted cells consist of a self-sustained unipotent population of LCs (Extended data Fig. 6a-e). Clonal analysis of LC expressing oncogenic PIK3CA revealed the presence of bipotent clones containing adjacent LCs and BCs, which was never observed in YFP control LCs (Fig. 3f and Extended data Fig. 6n-p). The relatively small proportion of K8+/K5+ BCs compared to K8-/K5+ BCs suggests that in the initial stage of LC to BC transition, these cells expressed markers of both lineages before maturing into basal-like cells and losing expression of LC markers (Fig. 3g and Extended data Fig. 6q-t), which are consistent with sequential gene expression shown by qRT-PCR analysis of FACS isolated BCs and LCs following *Pik3ca*-H1047R expression (Extended data Fig. 6u, v). The proportion of YFP expressing LCs increased over time, as well as the proportion of YFP+ BCs (Fig 3e), suggesting that *Pik3ca*-H1047R confers a competitive advantage to luminal targeted cells. To functionally determine whether LCs acquired multipotency upon PIK3CA expression, we tested the ability of *Pik3ca*-H1047R expressing LCs and their BC progeny to reconstitute mammary gland upon transplantation into mammary fat pads. FACS-isolated LCs expressing *Pik3ca*-H1047R were able to form outgrowths of mammary epithelium containing both BCs and LCs (observed in 6 out of 28 transplants), while, as previously described[14,18,19], WT LCs derived from K8CREERT2/Rosa-YFP were not able to form mammary outgrowths in the same conditions (in 0 out of 10 transplants)(Fig. 3h). Likewise, transplantation of newly formed BCs from K8CREERT2/*Pik3ca*-H1047R also generated mammary outgrowths containing BCs and LCs (7/11), as efficiently as control BCs (8/10) (Fig. 3i). Altogether these

data show that oncogenic *Pik3ca* promotes multilineage differentiation of LCs, inducing cellular heterogeneity at the early stage of the tumour initiation process.

In contrast to the early multilineage differentiation observed following oncogenic PIK3CA in LCs, during the first few months following TAM administration to K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice, BCs remained unipotent. Only around 7 months following oncogene expression, newly formed LCs became detectable and progressively increased over time (Fig. 3j-o and Extended data Fig.7). Immunostaining and qRT-PCR showed that these newly formed LCs expressed luminal markers to similar levels as WT LCs and no longer expressed high levels of basal markers (Extended data Fig. 7k, l). To determine whether LCs derived from BCs functionally correspond to luminal progenitors, we assessed the clonogenic potential of these cells using a colony-forming assay that only allows the growth of LCs[20]. The number of colonies derived from FACS-isolated LCs following *Pik3ca*-H1047R expression in BCs was significantly higher as compared to WT LCs, supporting the notion that oncogenic PIK3CA promotes the reprogramming of BCs into functional LCs (Fig. 3p). Altogether, these data show that *Pik3ca*-H1047R induces multipotency in otherwise lineage restricted basal and luminal unipotent progenitors, inducing cellular heterogeneity in oncogene targeted cells before progressing into more invasive tumours.

To define the molecular mechanisms by which *Pik3ca*-H1047R promotes multipotency and tumour heterogeneity, we performed transcriptional profiling of FACS-isolated basal and luminal-like cells following *Pik3ca*-H1047R induction in LCs (K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice) (B-K8PIK and L-K8PIK) and in BCs (K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice) (B-K5PIK and L-K5PIK) (Fig. 3d, m and Supplementary Table 1). Gene expression clustering analysis shows a profound reprogramming that occured in PIK3CA-H1047R-expressing cells as they underwent the

transition between basal and luminal lineages, becoming molecularly similar to the mammary lineages they have been converted into (Fig. 4a).

To unravel the molecular mechanisms by which oncogenic *Pik3ca* induced changes in cell fate and whether mechanisms are conserved or distinct across different cells of origin, we defined the gene signature induced by the expression of oncogenic *Pik3ca* in each population (i.e. B-K5PIK vs B-K5YFP, L-K5PIK vs L-K8YFP, etc…). Only 3 genes were upregulated in all conditions (*Serpina3n, Gdpd3, Zfp949*), and most of the genes upregulated by *Pik3ca*-H1047R expression were dependent on both the origin of the cell in which *Pik3ca*-H1047R was initially expressed and the cell lineage in which the oncogene was currently expressed (Fig. 4b Supplementary Table 2 and Extended data Fig. 8). While *Pik3ca*-H1047R induced the expression of specific genes according to their cellular origin and their basal or luminal phenotypes, 51 annotated genes were commonly upregulated in L-K5PIK and in B-K8PIK, including the long non-coding RNA *Neat1* and the transcription factor *Runx2* that both regulate mammary cell fate[21,22], genes regulating signal transduction (e.g. *Sfrp2*), cellular metabolism (e.g. *Tktl1*, *Bdh1*) and cell adhesion (eg: *EphB2, Trio*) (Fig. 4b, c and Supplementary Table 2), suggesting that common and distinct mechanisms induce cell fate changes upon oncogenic *Pik3ca* expression.

Basal to luminal cell fate transition induced by oncogenic *Pik3ca* induced the expression of a distinct set of genes (basal to luminal multipotent signature) (Fig 4). Some of these genes such as *Ntrk3*/TrkC were already upregulated in BCs following *Pik3ca*-H1047R expression (Supplementary Table 2), suggesting they represent the signature of their BC of origin. Following BC to LC fate transition, *Pik3ca*-H1047R expression induced genes that are specific for the newly formed LCs including *Nrtk2*/TrkB, a neurotrophin receptor, expressed in WT BCs and transiently expressed in LC during *Pik3ca*-H1047R-induced BC to LC fate transition (Fig. 4d and Extended data Fig. 7m). TrkB and TrkC have previously been shown

to be expressed in breast cancers and regulate survival of breast cancer stem cells in response to chemotherapy[23,24]. In addition, a translocation leading to a gene fusion between *Etv6* and *Ntrk3* causes breast tumours in both mouse and humans[25,26]. *Etv6-Ntrk3* expression in the MG induces predominantly the same type of bipotent mammary tumours[26] that arise from *Pik3ca*-H1047R expression in BCs, supporting a role for the *Pik3ca*/TrkB-TrkC axis in the establishment of bipotency during breast tumour initiation. The basal to luminal upregulated multipotent signature also contained genes commonly upregulated in L-K8PIK (Extended data Fig. 8k), reflecting the consequence of PIK3CA expression in LCs.

In contrast, the luminal to basal multipotent signature was characteristic of a wounding and proliferation response marked by the upregulation of *Il-33*/Alarmin (a cytokine that has been shown to be overexpressed in breast cancers and attenuates NK response against tumour cells[27,28]), *Il24a*, *Krt16*, *Itgb6*, *Itga2*, *Itga5*, *Tnc*, *CD109*, *Plau*, *Wnt10a*, *Timp3, Inhba, Ngf, Ereg, Ccdn1*, *Ccdn2* (Fig. 4e). As found during BC to LC transition, most of the luminal to basal multipotent signature genes were specific for the newly formed BCs (eg: *Ereg, Ccnd1, Wnt10a, Il33*), a significant fraction of these genes (eg: *Krt16, Il24a, Ccnd2, Inhba, Tnc*) were already upregulated in LCs targeted by oncogenic PIK3CA, suggesting they represent the signature of the LC of origin (Supplementary Table 2, 3 and Extended data Fig.8).

To define the relevance of these multipotency gene signatures to tumour progression, we assessed the expression of these genes in *Pik3ca*-H1047R-derived tumours. Some genes of luminal to basal multipotent signature such as *Il24a*, *Krt16*, and *Plau* were only upregulated during the initial stage of reprogramming and downregulated thereafter, while other genes such as *Col11a1*, EGFR ligand *Ereg*, *Inhba*, *Wnt10a* and *Tnc* continue to be expressed, or even further increased, in BL breast cancers arising from LCs (Fig. 4e, f). Similarly, *Ntrk2* and *Ntrk3* were expressed or even further upregulated in K5CREERT2-*Pik3ca*-H1047R derived luminal tumours (Fig. 4f). These data indicate that some of the genes associated with

cell fate transition during the early steps of tumour initiation increase with tumour progression.

To define the relevance of the *Pik3ca*-H1047R-induced multipotent gene signatures in human breast cancers, we assessed whether the different multipotent signatures correlated with a particular molecular breast cancer subtype[29] or disease free survival in a cohort of systemically untreated breast cancer patients[30]. Interestingly, the luminal to basal transition gene signature was strongly associated with basal-like breast cancers (Fig. 4g). Higher expression levels of this gene signature or individual genes such as *Ngf*, *Inhba*, *Itgb6* and *Wnt10a* were associated with poor clinical outcome (Fig 4h, and Extended data Fig. 9), consistent with the more aggressive tumour types induced by *Pik3ca*-H1047R expression in LCs. In contrast, the basal to luminal cell fate signature was associated with luminal A and normal-like human breast cancers (Fig. 4i). High gene expression levels of this gene signature were significantly associated with better prognosis (Fig 4j), consistent with the less aggressive tumours arising from BCs. These data indicate that the genetic program associated with *Pik3ca*-H1047R induced multipotency correlated with distinct molecular subtypes of human breast cancers and their levels of expression correlated with distinct clinical outcome.

In conclusion, our study shows that the cell of origin controls tumour heterogeneity in *Pik3ca*-H1047R-induced mammary tumours. *Pik3ca*-H1047R expression in LC gives rise to aggressive basal-like tumours while expression in BCs gives rise to less aggressive luminal-like tumours. We demonstrate that *Pik3ca*-H1047R-induced multipotency in unipotent progenitors. The promotion of multipotency induced by *Pik3ca*-H1047R is regulated by common and cell lineage specific molecular mechanisms that are influenced by the cellular origin in which the oncogene is initially expressed, setting the stage for future tumour heterogeneity and influencing clinical outcome in patients with breast cancers.

ACKNOWLEDGMENTS

C.B. is an investigator of WELBIO, A.V.K. is chercheur qualifié from the FNRS, C.S. is Maître de Recherche from the FNRS, M.Y.L. is supported by the Agency for Science, Technology and Research (A*STAR, Singapore) fellowship, M.O. and A.W. are supported by FNRS fellowships, R.R.G. is supported by a TELEVIE fellowship and S.B. is supported by the foundation "Amis de l'institut Jules Bordet". The CMMI is supported by the European Regional Development Fund and Wallonia. W.A.P. is supported by project grants from the National Health and Medical Research Council (NHMRC) of Australia. This work was supported by the FNRS, TELEVIE, a research grant from the Fondation Contre le Cancer, the ULB fondation, the Fond Yvonne Boël, the Fond Gaston Ithier, the foundation Bettencourt Schueller, the foundation Baillet Latour, and the European Research Council (ERC).

AUTHOR CONTRIBUTIONS

C.B., A.V.K. designed the experiments and performed data analysis. A.V.K., M.Y.L., M.O. performed all the experiments. S.B. and C.S. performed the bio-informatic analysis of gene expression and comparison with human breast cancers expression and gene amplification on human samples. S.R. and I.S helped to perform the histological classification of mouse mammary tumours with regard to their similarities to human breast cancers. G.B. provided technical support. C.D. provided technical support for cell sorting. A.W. and R.R.G. helped with some experiments. W.A.P provided animals and critically reviewed the manuscript. C.B. and AV.K. wrote the manuscript.

AUTHOR INFORMATION

Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be

addressed to C.B. (Cedric.Blanpain@ulb.ac.be) or A.V.K. (avkeymeu@ulb.ac.be). Microarray have been deposited in the Gene Expression Omnibus under accession number GSE69290 .

REFERENCES

1 Weigelt, B., Geyer, F. C. & Reis-Filho, J. S. Histological types of breast cancer: how special are they? *Mol Oncol* **4**, 192-208, doi:10.1016/j.molonc.2010.04.004 (2010).

2 Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752 (2000).

3 Network, C. G. A. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

4 Tikoo, A. *et al.* Physiological levels of Pik3ca(H1047R) mutation in the mouse mammary gland results in ductal hyperplasia and formation of ERalpha-positive tumors. *PLoS One* **7**, e36924, doi:10.1371/journal.pone.0036924 (2012).

5 Prat, A., Ellis, M. J. & Perou, C. M. Practical implications of gene-expression-based assays for breast oncologists. *Nature reviews. Clinical oncology* **9**, 48-57, doi:10.1038/nrclinonc.2011.178 (2012).

6 Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395-399, doi:10.1038/nature10933 (2012).

7 Stemke-Hale, K. *et al.* An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer. *Cancer Res* **68**, 6084-6091, doi:10.1158/0008-5472.can-07-6854 (2008).

8 Lopez-Knowles, E. *et al.* PI3K pathway activation in breast cancer is associated with the basal-like phenotype and cancer-specific mortality. *Int J Cancer* **126**, 1121-1131, doi:10.1002/ijc.24831 (2010).

9 Tilch, E. *et al.* Mutations in EGFR, BRAF and RAS are rare in triple-negative and basal-like breast cancers from Caucasian women. *Breast Cancer Res Treat* **143**, 385-392, doi:10.1007/s10549-013-2798-1 (2014).

10 Adams, J. R. *et al.* Cooperation between Pik3ca and p53 mutations in mouse mammary tumor formation. *Cancer Res* **71**, 2706-2717, doi:10.1158/0008-5472.can-10-0738 (2011).

11 Liu, P. *et al.* Oncogenic PIK3CA-driven mammary tumors frequently recur via PI3K pathway-dependent and PI3K pathway-independent mechanisms. *Nat Med* **17**, 1116-1120, doi:10.1038/nm.2402 (2011).

12 Yuan, W. *et al.* Conditional activation of Pik3ca(H1047R) in a knock-in mouse model promotes mammary tumorigenesis and emergence of mutations. *Oncogene* **32**, 318-326, doi:10.1038/onc.2012.53 (2013).

13 Meyer, D. S. *et al.* Luminal expression of PIK3CA mutant H1047R in the mammary gland induces heterogeneous tumors. *Cancer Res* **71**, 4344-4351, doi:10.1158/0008-5472.can-10-3827 (2011).

14 Van Keymeulen, A. *et al.* Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479**, 189-193, doi:10.1038/nature10573 (2011).

15 Rios, A. C., Fu, N. Y., Lindeman, G. J. & Visvader, J. E. In situ identification of bipotent stem cells in the mammary gland. *Nature* **506**, 322-327, doi:10.1038/nature12948 (2014).

16 Wang, D. *et al.* Identification of multipotent mammary stem cells by protein C receptor expression. *Nature* **517**, 81-84, doi:10.1038/nature13851 (2015).
17 Hayes, M. M. Adenomyoepithelioma of the breast: a review stressing its propensity for malignant transformation. *J Clin Pathol* **64**, 477-484, doi:10.1136/jcp.2010.087718 (2011).
18 Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439**, 84-88 (2006).
19 Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439**, 993-997 (2006).
20 Shehata, M. *et al.* Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res* **14**, R134, doi:10.1186/bcr3334 (2012).
21 Standaert, L. *et al.* The long noncoding RNA Neat1 is required for mammary gland development and lactation. *RNA (New York, N.Y.)* **20**, 1844-1849, doi:10.1261/rna.047332.114 (2014).
22 Owens, T. W. *et al.* Runx2 is a novel regulator of mammary epithelial cell fate in development and breast cancer. *Cancer Res* **74**, 5277-5286, doi:10.1158/0008-5472.can-14-0053 (2014).
23 Jin, W. *et al.* TrkC plays an essential role in breast tumor growth and metastasis. *Carcinogenesis* **31**, 1939-1947, doi:10.1093/carcin/bgq180 (2010).
24 Yin, B. *et al.* The TrkB+ cancer stem cells contribute to post-chemotherapy recurrence of triple-negative breast cancers in an orthotopic mouse model. *Oncogene*, doi:10.1038/onc.2014.8 (2014).
25 Euhus, D. M., Timmons, C. F. & Tomlinson, G. E. ETV6-NTRK3--Trk-ing the primary event in human secretory breast cancer. *Cancer Cell* **2**, 347-348 (2002).
26 Li, Z. *et al.* ETV6-NTRK3 fusion oncogene initiates breast cancer from committed mammary progenitors via activation of AP1 complex. *Cancer Cell* **12**, 542-558 (2007).
27 Milovanovic, M. *et al.* IL-33/ST2 axis in inflammation and immunopathology. *Immunologic research* **52**, 89-99, doi:10.1007/s12026-012-8283-9 (2012).
28 Wu, J. *et al.* Identification and functional analysis of 9p24 amplified genes in human breast cancer. *Oncogene* **31**, 333-341, doi:10.1038/onc.2011.227 (2012).
29 Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-1167, doi:10.1200/jco.2008.18.1370 (2009).
30 Haibe-Kains, B. *et al.* A three-gene model to robustly identify breast cancer molecular subtypes. *J Natl Cancer Inst* **104**, 311-325, doi:10.1093/jnci/djr545 (2012).

FIGURE LEGENDS

Figure 1 Oncogenic *Pik3ca* expression in BCs or LCs leads to distinct tumour phenotypes. Genetic strategy to target *Pik3ca*-H1047R in BCs (**a**) or LCs (**b**). **c, d**, Tumour free survival curves in K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice (*n*=23 mice) (latency of 12±4 months (mean±SD))(**c**) or K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice (*n*=11 mice) (latency of 15±4 months) (**d**). No tumour were observed in control K5CREERT2/Rosa-YFP mice (n=10 mice) (**c**) or in K8CREERT2/Rosa-YFP mice (n=10 mice) (**d**). **e, f**, Pie chart showing tumour classification in BC- (**e**) and LC- (**f**) derived tumours. BCs-derived tumours were all classified as adenomyoepitheliomas (n=36 tumours). LCs-derived tumours comprised adenomyoepithelioma (n=8 tumours), mixed adenomyoepithelioma with myoepithelial carcinoma (n=6 tumours), metaplastic carcinoma (n=2 tumours), invasive carcinoma of no special type (NST) (n=1 tumour) (**f**). Detailed histological characterization is presented in Extended data Fig. 2.

Figure 2. Oncogenic *Pik3ca* expression and p53 deletion in BCs or LC leads more frequently to highly invasive mammary tumours. Genetic strategy to target *Pik3ca*-H1047R expression and *p53* deletion in BCs (**a, b**) or in LCs (**c**). **d-f,** Tumour free survival curves in K5CREERT2/*Pik3ca*-H1047R/*p53*fl/+/Rosa-YFP (9±3 months latency; mean±SD) (*n*=6 mice) (**d**), K14rtTA/TetOCre/*Pik3ca*-H1047R/*p53*fl/+/Rosa-YFP (9±1 months latency) (*n*=14 mice) (**e**), K8CREERT2/*Pik3ca*-H1047R/*p53*fl/fl/Rosa-YFP mice (5±1 months latency) (*n*=20 mice) and K8CREERT2/*Pik3ca*-H1047R/*p53*fl/+/Rosa-YFP mice (9±3 months latency) (*n*=17 mice) (**f**). Control mice did not developed tumours (n=10 mice per condition). **g-i**, Pie charts depicting the classification of mammary tumours in K5CREERT2/*Pik3ca*-H1047R/*p53*fl/+/Rosa-YFP (n=8 tumours) (**g**),

K14rtTA/TetOCre/*Pik3ca*-H1047R/*p53*fl/+/Rosa-YFP (n=19 tumours) (**h**), K8CREERT2/*Pik3ca*-H1047R/*p53*fl/fl/Rosa-YFP (n=40 tumours) and K8CREERT2/*Pik3ca*-H1047R/*p53*fl/+/Rosa-YFP (n=22 tumours)(**i)**. Detailed histological characterization is presented in Extended data Fig. 5.

Figure 3. Oncogenic *Pik3ca* expression induces multipotency in unipotent luminal and basal progenitors. a, b, Immunofluorescence of K8/YFP at 1w (**a**) and K5/YFP at 5w (**b**) after TAM administration to K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice. **c, d**, FACS analysis of CD24 and CD29 expression in Lin-YFP+ cells 1w (**c**) or 8w (**d**) after TAM induction. **e,** Percentage of YFP+ cells within LCs ($CD29^{Lo}/CD24^{+}$) and BCs ($CD29^{Hi}/CD24^{+}$) at different time points following TAM administration (n= 3, 6, 3, 6, 4 mice respectively for 1w, 5w, 8w, 4m, 7m). **f**, Immunofluorescence of K8/K5/YFP 8w after clonal induction of K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice. Arrow points to newly K5/K8/*Pik3ca*-H1047R/YFP+ BC generated from a LC. **g**, Percentage of YFP+ cells expressing K5 and/or K8 at different time points following *Pik3ca*-H1047R expression in LCs (n=3 mice per condition). **h, i**, Immunofluorescence of K5/K8 of a mammary outgrowth derived from LCs (**h**) or BCs (**i**) from K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice. **j**, **k,** Immunofluorescence of K5/YFP 1w (**j**) or K8/YFP 7 months (**k**) after TAM administration to K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice. (**l, m**) FACS analysis of CD24 and CD29 expression in Lin-YFP+ cells 1w (**l**), or 7 months (**m**) after TAM induction. **n,** Percentage of YFP+ cells within LCs and BCs at different time points following TAM administration (n=3, 3, 5, 4, 3 mice respectively for 1w, 5w, 8w, 7m, 12m). **o**, Immunofluorescence for K8/K5/YFP 7 months after clonal *Pik3ca*-H1047R/YFP expression in BCs. Arrow points to newly formed K8+/YFP+ LC arising from a BC. **p**, Mean number of colonies per 1000 sorted luminal cells in in vitro colony forming assay of YFP+ LCs derived from

K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice induced for 12 months or WT LCs (n= 3 biologically independent experiments per condition). Error bars, SEM. Scale bars, 10 μm.

Figure 4. Molecular characterization of oncogenic *Pik3ca*-induced multipotency. a, Hierarchical gene expression clustering of BCs and LCs with or without *Pik3ca*-H1047R expression. Green and red correspond to high and low expressed genes respectively. The two major branches of the tree are supported by bootstrap values of 100. **b**, Venn diagram of upregulated genes (>1.5 fold) following *Pik3ca*-H1047R expression in BCs and LCs. **c-e,** qRT-PCR analysis of genes belonging to the common (**c**), basal to luminal (**d**), or luminal to basal multipotency signature (**e**) in B-K8PIK and L-K5PIK, 8w and 10-12 months respectively after *Pik3ca*-H1047R expression, compared to their age-matched controls. Gene expression was normalized to *Gapdh* housekeeping gene (n=4 biologically independent samples). **f**, qRT-PCR analysis of the multipotency signature genes in control cells, in BC-derived adenomyoepithelioma and in LC-derived metaplastic tumours. Data were normalized to gene expression in age-matched control LCs (L-K8YFPo) (n=4 biologically independent samples). **g, i**, Expression levels of the luminal to basal (**g**) or basal to luminal (**i**) multipotency signature in a large set of breast cancer patients according to their PAM50 subtype. **h, j**, Disease free survival in untreated patients according to the level of expression of the genes of the luminal to basal (**h**) or basal to luminal (**j**) multipotency signature. **k-n**, Summary of the role of the cancer cell of origin in regulating *Pik3ca*-H1047R-induced tumour heterogeneity. **k**, Expression of *Pik3ca*-H1047R in BCs give rise to luminal-like tumours, while in LCs *Pik3ca*-H1047R give rise to more heterogeneous and aggressive tumours. **l**, Additional p53 deletion promotes *Pik3ca*-H1047R induced tumour heterogeneity in BCs and leads to more aggressive metaplastic carcinoma in LCs. **m, n**, Model of *Pik3ca*-H1047R-induced multipotency in LCs and BCs. Genes shown are upregulated during cell fate change. Genes highlighted in blue belong to the common multipotency signature. Error bars, SEM.

METHODS

Mice - RosaYFP mice[31] were obtained from the Jackson laboratory. K5CREERT2, and K8CREERT2 mice were described previously[14]. K14rtTA mice[32]were provided by E. Fuchs. TetO-CRE mice[33] were provided by A. Nagy. Pik3ca-H1047R knock-in mice, in which wildtype exon 20 is replaced by H1047R mutant exon 20 upon CRE recombination, were described previously[4]. p53flox/flox mice[34] were obtained from the National Cancer Institute at Frederick.

All experimental mice used in this study were female, mixed strains and more than 6 weeks old. For all experiments presented in this study, the sample size was large enough to measure the effect size. No randomization and no blinding were performed in this study. Mice designated within tumour cohort were sacrificed when a palpable mass of maximum 1 cm^3 was detected. Mouse colonies were maintained in a certified animal facility in accordance with European guidelines. Ethical protocol was approved by the local ethical commitee for animal welfare (CEBEA) from Université Libre de Bruxelles (protocols #363 and 527).

Targeting expression of YFP and/or PIK3CA-H1047R and deletion of p53 –4 to 5w old K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP, K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP, K8CREERT2/*Pik3ca*-H1047R/p53fl/fl/Rosa-YFP, K8CREERT2/*Pik3ca*-H1047R/p53fl/+/Rosa-YFP, K5CREERT2/*Pik3ca*-H1047R/p53fl/+/Rosa-YFP, K5CREERT2/Rosa-YFP and K8CREERT2/Rosa-YFP mice were induced with 15 mg of tamoxifen (TAM) (Sigma, diluted in Sunflower seed oil, Sigma) by intraperitoneal injection (3 injections of 5 mg every three days). Tamoxifen administration induced a transient delay in MG during puberty but no long-term effect on MG development and homeostasis[35]. 5w old K14rtTA/TetOCre/*Pik3ca*-H1047R/p53fl/+/RosaYFP and K14rtTA/TetOCre/RosaYFP mice were induced by oral administration of doxycycline food diet (1 g/kg, BIO-SERV) for 5 days.

For clonal analyses, 4 to 5w old K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP and K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice were respectively induced with 0.05 mg or 2 mg TAM by intraperitoneal injection. For induction in adult mice, 8w old K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice were induced with 15 mg of tamoxifen by intraperitoneal injection (3 injections of 5 mg every three days).

Histology and immunostaining on sections– For immunofluorescence, dissected inguinal mammary glands or tumour samples were pre-fixed 2h in 4% paraformaldehyde at room temperature. Tissues were washed 3 times with PBS for 5 min and incubated overnight in 30% sucrose in PBS at 4°C. Tissues were embedded in OCT and kept at -80°C. Sections of 5 µm were cut using a HM560 Microm cryostat (Mikron Instruments Inc).

For immunofluorescence, tissue sections were incubated in blocking buffer (5% Horse serum/ 1% BSA/0.2% Triton-X in PBS) for 1h at room temperature. The different primary antibodies combinations were incubated overnight at 4°C. Sections were then rinsed three times for 5 minutes in PBS and incubated with corresponding secondary antibodies diluted at 1:400 in blocking buffer for 1h at room temperature. The following primary antibodies were used : anti-GFP (rabbit, 1:1000, A11122, Molecular Probes), anti-GFP (chicken, 1:1000, ab13970 abcam), anti-K8 (rat, 1:1000, Troma-I, Developmental Studies Hybridoma Bank, University of Iowa), anti-K14 (rabbit, 1:1000, PRB-155P-0100, Covance), anti-K14 (chicken, 1:1000, SIG-3476-0100, Covance), anti-K5 (rabbit, 1:1000, PRB-160P-0100, Covance), anti-K19 (rat, 1:500, Troma-III, Developmental Studies Hybridoma Bank, University of Iowa), anti-ER (rabbit, 1:300, sc-542, Santa Cruz), anti-PR (rabbit, 1:300, sc-7208, Santa Cruz), anti-Her2 (rabbit, 1:300, 2165, Cell Signalling), anti-Ki67 (rabbit, 1:500, ab15580, abcam), anti-Ecadherin (rat, 1:1000, 14-3249-82, ebioscience), anti-vimentin (rabbit, 1:400, ab92547, abcam), anti-TrkB (rabbit, 1:500, sc-12, Santa Cruz), anti-p63 (rabbit, 1:100, Mab306-05, Santa Cruz), anti-SMA-Cy3 (mouse, 1:500, C6198, Sigma-Aldrich), anti-claudin 3 (rabbit,

1:300, 34-1700, Invitrogen). The following secondary antibodies were used: anti-rabbit, anti-rat, anti-chicken conjugated to AlexaFluor488 (Molecular Probes), to Rhodamine Red-X or to Cy5 (JacksonImmunoResearch). Nuclei were stained with Hoechst solution (1:2000) and slides were mounted in DAKO mounting medium supplemented with 2.5% Dabco (Sigma).

For paraffin embedded tissues, dissected mammary glands were pre-fixed overnight, 4°C in paraformaldehyde 4%. Tissues were washed 3 times with PBS. Prior to automated paraffin processing, tissues were washed in tap water and kept in isopropanol 70%. 5 µm sections were made with a Leica RM2245 microtome.

Haematoxylin and eosin staining was performed as previously described[36]. p63 staining on tumour paraffin sections were performed on an automated IHC platform (Ventana Discovery XT). Briefly, paraffin sections were deparaffinized and rehydrated. The antigen unmasking procedure was performed for 36 min at 95°C in EDTA (pH 9). Slides were incubated with the anti-p63 (clone 7JUL, 1:100, Leica) for 3h, followed by a linker rabbit anti-mouse (clone M1gG51-4, abcam 1:750) for 16 min. Finally, slides were incubated with the OmniMap HRP-conjugated anti-rabbit antibody (Ventana) for 12 min. Standard ABC kit, and ImmPACT DAB (Vector Laboratories) were used for the detection of HRP activity. Nuclei staining was done with Mayer's Hematoxylin (Labonord), followed by dehydration and mounting with SafeMount (Labonord).

Whole mount mammary gland immunofluorescence- For clonal analyses and mammary transplant experiments, dissected inguinal mammary glands were incubated in 2ml HBSS + 30U/ml collagenase + 300 µg/ml hyaluronidase (Sigma) for 30 min at 37°C under agitation. Following three washes of 5 minutes with HBSS, mammary glands were fixed in 4% paraformaldehyde for 2h, room temperature, washed three times 10 min in PBS under agitation and incubated in blocking buffer (5% Horse serum/1% BSA/0.8% Triton-X in PBS)

3h at room temperature. The primary antibody combination, diluted in the blocking buffer, was incubated overnight at room temperature under agitation. Samples were washed 3 times for 10 minutes in PBS/0.2% Tween-20 and incubated in secondary antibodies diluted in the blocking buffer for 5h under agitation. Cell nuclei were stained with Hoechst for 30 min (1:1000 in PBS/0.2% Tween-20). Samples were mounted on slides in DAKO mounting medium supplemented with 2.5% Dabco (Sigma).

Staining on human breast cancer sections – Tissue samples were obtained retrospectively from archival formalin-fixed and paraffin-embedded samples in the Department of Pathology of the Erasme Hospital. Histopathological diagnoses were reviewed and assessed according to the 2012 World Health Organization Classification. Sections of 5 µm were subjected to standard immunochemistry (IHC) as previously described[36] using respectively monoclonal anti-CK8/18 (1:200; clone 5D3; BioGenex), anti-CK14 (1:100; clone LL002; Leica) and anti-P63 (1:200; clone 7JUL; Leica) antibodies. Saining was visualized with streptavidin-biotin-peroxydase complex kit reagents (BioGenex) using diaminobenzidine/H_2O_2 as the chromogenic substrate. Counterstaining with haematoxylin concluded the processing. Nuclei staining was done with Mayer's Haematoxylin (Labonord), followed by dehydration and mounting with SafeMount (Labornord).

Microscope image acquisition –Pictures were acquired on an Axio Observer Z1 Microscope using 10X and 40X Zeiss EC Plan-NEOFLUAR objectives, with an AxioCamMR3 camera and using the Axiovision software (Carl Zeiss Inc.). Confocal images in Fig. 3f, h, i, o and Extended data Fig. 6f, h-l, n, o, q-t and Extended data Fig. 7i, m, were acquired at room temperature using a Zeiss LSM780 multiphoton confocal microscope fitted on an Axiovert M200 inverted microscope equipped with C-Apochromat (40x=1.2 numerical aperture) water immersion objectives (Carl Zeiss Inc.). Optical sections of 1024 x 1024 pixels, were collected

sequentially for each fluorochrome. The data-sets generated were merged and displayed with the ZEN software.

Mammary gland and tumour cell dissociation – Mammary glands were dissected and lymph nodes removed. Tissues were briefly washed in HBSS, and chopped with a McIwain tissue chopper. Chopped tissues were placed in HBSS + 300 U/ml collagenase (Sigma) + 300 µg/ml hyaluronidase (Sigma) and digested 2h at 37°C under agitation. Physical dissociation using a P1000 pipette was done every 15 min throughout the enzymatic digestion duration. EDTA at a final concentration of 5 mM was added for 10 min to the resultant organoid suspension, followed by 0.25% Trypsin-EGTA for 2 min (only in the case of normal mammary glands) before filtration through a 70-µm mesh, 2 successive washes in 2% FBS/PBS and antibody labelling.

Cell labelling, flow cytometry and sorting – 2-5 millions of cells per condition were incubated in 250 µl of 2%FBS/PBS with flurochrome-conjugated primary antibodies for 30 min, vortexing every 10 min. Cells were washed with 2% FBS/PBS and were resuspended in 2.5 µg/ml DAPI (Invitrogen) before analysis. Primary antibodies used were: PE-Cy7-conjugated anti-CD24 (1:50, clone M1/69, BD Biosciences), APC-conjugated anti-CD29 (1:50, clone eBioHMb1-1, eBiosciences), PE-conjugated anti-CD45 (1:50, clone 30-F11, eBiosciences), PE-conjugated anti-CD31 (1:50, clone MEC 13.33, BD Biosciences), PE-conjugated anti-CD140a (1:50, clone APA5, eBiosciences). Data analysis and cell sorting were performed on a FACSAria sorter using the FACS DiVa software (BD Biosciences). Dead cells were excluded with DAPI; CD45, CD31 and CD140a positive cells were excluded (Lin-) before analysis of the YFP+ cells. For profile analysis, a minimum of 1000 YFP+ cells were analysed per sample.

Tumour harvesting and classification – Tumours were detected by mammary gland palpation. Mice were sacrificed when one tumour reached a maximum of 1 cm diameter. The K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice presented 1 tumour in 57%, 2 tumours in 26%, and 3 or more tumours in 17% of the cases at the time of analysis (a total of 36 tumours from 23 mice were analysed). The K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice presented 1 tumour in 64%, 2 tumours in 18%, and 3 or more tumours in 18% of the cases at the time of analysis (a total of 17 tumours from 11 mice were analysed). The K5CREERT2/*Pik3ca*-H1047R/p53fl/+/Rosa-YFP mice presented 1 tumour in 83%, and 3 tumours in 17% of the cases at the time of analysis (a total of 8 tumours from 6 mice were analysed). The K14rtTA/TetOCre/*Pik3ca*-H1047R/p53fl/+/Rosa-YFP mice presented 1 tumour in 57%, 2 tumours in 36%, and 3 or more tumours in 7% of the cases at the time of analysis (a total of 19 tumours from 14 mice were analysed). The K8CREERT2/*Pik3ca*-H1047R/p53fl/fl/Rosa-YFP mice presented 1 tumour in 20%, 2 tumours in 45%, and 3 or more tumours in 35% of the cases at the time of analysis (a total of 40 tumours from 20 mice were analysed). The K8CREERT2/*Pik3ca*-H1047R/p53fl/+/Rosa-YFP mice presented 1 tumour in 71%, and 2 tumours in 29% of the cases at the time of analysis (a total of 22 tumours from 17 mice were analysed). For each harvested tumour, the tumour was cut in three pieces, one for paraffin embedding, one for OCT embedding, and one for cell sorting and RNA extraction. Tumour classification was done based on histological features.

Mammary colony forming assay (Ma-CFC) assay - Luminal YFP+ cells from K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice induced for 12 months were flow-sorted as single cell suspension based on their Lin- $CD29^{lo}CD24^{+}YFP^{+}$ profile. Control YFP^{-} luminal cells from K5CREERT2/Rosa-YFP induced for 12 months were sorted based on their $CD29^{lo}CD24^{+}$ profile. Luminal cells were cultured with irradiated NIH 3T3 feeder cells in Mouse-Epicult B media (Stem Cell Technologies) supplemented with 10 ng/ml Epidermal

Growth Factor (Sigma-Aldrich), 10 ng/ml basic Fibroblast Growth Factor (R&D Systems), 4 µg/ml Heparin (Sigma-Aldrich), 1 mg/ml Bovine Serum Albumin (BSA, Sigma-Aldrich), 5% FBS (Life Technologies), 50 units/ml penicillin and 50 µg/ml streptomycin (Life Technologies), as previously described[20]. After 1w, colonies were fixed with methanol, stained with Giemsa stain (Sigma-Aldrich) and counted manually.

Mammary fat pad transplantation and analysis. – 8000 LCs from K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP or control K8CREERT2/Rosa-YFP or 1350 BCs from K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP or control K14rtTA/TetOCRE/rosa-YFP induced for 4 months were sorted based on their $Lin^-YFP^+CD29^{lo}CD24^+$ or $Lin^-YFP^+CD29^{hi}CD24^+$ profiles. LCs were resuspended in 10 µl DMEM+50% bovine serum. BCs were sorted in presence of 10 µM of Rock inhibitor (Y27632, Sigma) and resuspended in 75% DMEM-25% matrigel. Cell suspension was injected into the fourth mammary gland of 3 to 4w old NOD-SCID mice that had been cleared of endogeneous epithelium as previously described[18,19]. Recipient mice were mated 4 weeks after the transplantation, and sacrified 2 to 3w later, when fully pregnant. Recipient glands were dissected and stained for GFP, K8 and K5 as whole mounts. An outgrowth was defined as an epithelial structure comprising ducts and lobules and/or terminal end buds.

Quantification of keratin+ cells within YFP+ cells- A total of 1907, 1704 and 2391 of YFP+ cells from 3 different mice per condition were analysed respectively in K8CREERT2/Rosa-YFP induced 4w, K8CREERT2/*Pik3ca* -H1047R/Rosa-YFP induced 1w and 8w on 5µm cryosections stained for K5, K8 and GFP analysed with confocal microscope. Cells were scored as K8+K5- (K8), K8+K5+ (K5K8) or K8-K5+ (K5).

Quantification of Ki67 expression in tumours – A minimum of 6992 nuclei from minimum 3 tumours per condition were analysed manually for their Ki67 expression in the different

tumour subtypes derived from K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP, K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP, K8CREERT2/*Pik3ca*-H1047R/p53fl/fl/Rosa-YFP and K14rtTA/TetOCre/*Pik3ca*-H1047R/p53fl/+/Rosa-YFP mice, except for the NST carcinoma derived from K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP for which a total of 4622 nuclei were counted in the single NST carcinoma observed.

Quantification of clone composition – Mammary glands were processed as whole mount and stained for K8, K5 and GFP. Clones were analysed by confocal microscopy. A total of 822, 936, 714 and 360 clones from three independent mice per condition were analyzed in K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP induced for 1w, induced for 10w, and in K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP induced for 1w and induced for 7m respectively at clonal density. The clones were scored in three classes according to their keratin expression: luminal clones, composed only of K8+ cells, basal clones, composed only of K5+ cells, and mixed clones, composed of K8+ and K5+ cells.

Quantification of % YFP labelled cells- The percentage of YFP labelled cells within the luminal and basal populations was quantified by FACS. The luminal population was defined as the $CD29^{Lo}CD24^{+}$ population and the basal population was defined as the $CD29^{Hi}CD24^{+}$ population.

Whole-mount carmine staining – Whole-mount mammary fourth mammary gland were fixed in methanol Carnoy (60% methanol, 30% acetic acid, 10% chloroform) for at least 2h and rehydrated in 70% ethanol, followed by water. Staining in carmine alum (Sigma) was done overnight and excess dye was rinsed with water. This is followed by incubation in 70%, 95%, 100% ethanol (1h each) and fat-clearing in toluene overnight. All steps were carried out at room temperature.

Epithelial outgrowth measurement – Carmine-stained mammary glands were photographed with a Leica M80 stereomicroscope equipped with a Leica IC80 HD digital camera. The distance from the lymph node of the mammary epithelium was scored by measuring the distance between the distal edge of the lymph node and the most distal tip of the epithelium.

RNA extraction and quantitative real-time PCR- The protocol used for RNA extraction on FACS isolated cells has been previously described[37]. Briefly, RNA extraction was performed using the RNeasy micro kit (Qiagen) according to the manufacturer's recommendations and DNase treatment. After nanodrop RNA quantification and analysis of RNA integrity, purified RNA was used to synthesize the first strand cDNA in a 50-µl final volume, using Superscript II (Invitrogen) and random hexamers (Roche). Genomic contamination was detected by performing the same procedure without reverse transcriptase. Quantitative PCR analyses were performed with 1 ng of cDNA as template, using FastStart Essential DNA green master (Roche) and a Light Cycler 96 (Roche) for real-time PCR system.

Relative quantitative RNA was normalized using the housekeeping gene *Gapdh*. Primers were designed using PrimerBank database (http://pga.mgh.harvard.edu/primerbank/) and are listed in Supplementary Table 4. Analysis of the results was performed using Light Cycler 96 software (Roche) and relative quantification was performed using the ΔΔCt method using *Gapdh* as reference. The entire procedure was repeated in four biologically independent samples. For Extended data Fig. 6, 7, data are shown as fold change over luminal cells or basal cells derived from 3 months old wild-type mice (L-WT and B-WT).

Microarray analysis

Total RNA was analysed using Mouse whole genome MG-430 PM array from Affymetrix at the IRB Functional Genomics Core (Barcelona). All the results were normalized with RMA normalization using R-bioconductor package affy with standard parameters[38,39]. Two

biologically independent samples were analysed for each condition, except for tumours derived from K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP or K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP, for which 3 and 7 samples were analysed respectively. Sorted BCs from K5CREERT2/Rosa-YFP mice induced for 8w or 10-12 months, LCs from K8CREERT2/Rosa-YFP mice induced for 8w or 10-12 months, BCs and LCs from K5CREERT2/PIK3CA-H1047R/Rosa-YFP mice induced for 10-12 months, BCs and LCs from K8CREERT2/PIK3CA-H1047R/Rosa-YFP mice induced for 8w, Lin- cells from K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP derived tumours and from K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP derived tumours # 1, 2, 7, YFP+ cells from K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP derived tumours # 3, 4, 5, 6 and from K8CREERT2/*Pik3ca*-H1047R/p53fl/fl/Rosa-YFP derived tumours were analysed. The gene signature induced by the expression of PIK3CA-H1047R in each population by comparing their transcriptional profile with LCs arising from age-matched K8CREERT2/Rosa-YFP or BCs arising from K5CREERT2/Rosa-YFP mice. Only genes upregulated or downregulated by at least 1.5 fold were considered in the analysis.

Microarray data clustering

Clustering and bootstrap analyses were performed using the pvclust and gplots packages of the R statistical suite[40]. Clustering was performed with the default parameters of the R hclust function (Euclidean distance and complete linkage) considering only the top 500 most variant genes among all experiments.

Gene signature comparison

Venn diagrams were computed with the R statistical tool. The reported hypergeometric *P-values* for every comparison between two signatures correspond to the probability to observe an intersection of at least a given size by chance only, knowing the number of genes tested on a microarray chip.

Murine and human breast tumours gene expression profile comparison

To compare the murine tumour gene expression data to human tumour data, we used the METABRIC dataset composed of 1992 patients[41]. METABRIC expression data were downloaded from the EBI website (datasets EGAD00010000210 and EGAD00010000211). When multiple probes mapped to the same Entrez gene identifier, we kept the one with the highest variance in the dataset using the genefu package. The PAM50 subtypes were computed using the Bioconductor genefu package dedicated function[30] (1448 Basal, 1027 Her2, 2260 LumB, 2162 LumA and 323 Normal).

Boxplots and Kruskal-Wallis test p-values were computed using R. P-values reflect the probability that at least one of the cancer subtypes express the tested signature at a significantly different level.

Uni-directional Student t-test p-values reflect the probability that one signature is significantly more expressed (or repressed) in one subtype compared to all the others. For the t-tests, as they are more robust to the extreme values, median and interquartile ranges were chosen as estimators of the central tendency and of the dispersion (instead of the mean and the standard deviation).

We then merged the murine dataset with the METABRIC dataset by keeping those genes described as orthologous in the EnsEMBL database downloaded via Biomart[42] and having exactly the same identifier. Batch effect between murine and human data was corrected using the Combat function of the Bioconductor sva package[43]. The PCA and clustering analyses were performed using the R statistical software considering an expression matrix containing only the expression values of the 46 PAM50 orthologous genes between mouse and human. For clustering, we used the Euclidean distance combined to the complete hierarchical clustering method (default parameters). PAM50 subtypes were computed using the R/Bioconductor genefu package.

Survival analyses in humans

Mouse derived signatures were converted to human signatures by considering the orthologous genes in humans. Signatures score were then computed and re-scaled using the dedicated function of the R / Bioconductor genefu package. The scores were computed for each patients of the METABRIC together with those present in 33 other breast tumours reference datasets[30] (7220 patients). Survival curves were computed using the dedicated function of the genefu package only on untreated patients (1859 cases) with available survival data. Expression level categories correspond to the tertiles of the expression values in the untreated patients. P-values correspond to the log-rank p-value which reflects the probability that at least one of the class of signature expression present a significant differing outcome from the other classes.

44

Supplementary References

31 Srinivas, S. *et al.* Cre reporter strains produced by targeted insertion of EYFP and ECFP into the ROSA26 locus. *BMC Dev Biol* **1**, 4 (2001).

32 Nguyen, H., Rendl, M. & Fuchs, E. Tcf3 governs stem cell features and represses cell fate determination in skin. *Cell* **127**, 171-183 (2006).

33 Perl, A. K., Wert, S. E., Nagy, A., Lobe, C. G. & Whitsett, J. A. Early restriction of peripheral and proximal cell lineages during formation of the lung. *Proc Natl Acad Sci U S A* **99**, 10482-10487 (2002).

34 Jonkers, J. *et al.* Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat Genet* **29**, 418-425 (2001).

35 Shehata, M., van Amerongen, R., Zeeman, A. L., Giraddi, R. R. & Stingl, J. The influence of tamoxifen on normal mouse mammary gland homeostasis. *Breast Cancer Res* **16**, 411, doi:10.1186/s13058-014-0411-0 (2014).

36 Rorive, S. *et al.* TIMP-4 and CD63: new prognostic biomarkers in human astrocytomas. *Mod Pathol* **23**, 1418-1428, doi:10.1038/modpathol.2010.136 (2010).

37 Beck, B. *et al.* A vascular niche and a VEGF-Nrp1 loop regulate the initiation and stemness of skin tumours. *Nature* **478**, 399-403, doi:10.1038/nature10525 (2011).

38 McCall, M. N., Bolstad, B. M. & Irizarry, R. A. Frozen robust multiarray analysis (fRMA). *Biostatistics* **11**, 242-253, doi:10.1093/biostatistics/kxp059 (2010).

39 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).

40 Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540-1542, doi:10.1093/bioinformatics/btl117 (2006).

41 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352, doi:10.1038/nature10983 (2012).

42 Kasprzyk, A. BioMart: driving a paradigm change in biological data management. *Database : the journal of biological databases and curation* **2011**, bar049, doi:10.1093/database/bar049 (2011).
43 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883, doi:10.1093/bioinformatics/bts034 (2012).
44 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

Extended data Figure legends

Extended data Figure 1. Tamoxifen administration has no long-term effect on MG.

a, b, Effect of TAM on mammary epithelial postnatal growth. **a**, representative whole mount preparations of carmine alum-stained mammary epithelium from the fourth mammary gland, showing that TAM induces a delay in mammary epithelium growth at early time points, but no difference is observed 8w after TAM induction and **b**, mean distance from lymph node distal edge to the distal epithelial edge 1w, 5w and 8w following TAM injection or oil injection (n=6, 6, 4, 3, 5, 4 mice respectively for 1w ctr, 1w TAM, 5w ctr, 5w TAM, 8w ctr, 8w TAM). p value derived from two-sided student t-test is respectively 0.161, 0.035, 0.748 when comparing ctr and TAM conditions at 1w, 5w and 8w. **c,** Percentage of YFP+ cells in LCs ($CD29^{Lo}/CD24^{+}$) and in BCs ($CD29^{Hi}/CD24^{+}$) analysed by FACS 48h following TAM administration in K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP and K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP, or 1w following doxycycline administration to K14rtTA/TetOCre/*Pik3ca*-H1047R/p53fl/+/Rosa-YFP mice (n=5, 6, 3 mice respectively for K5CREERT2, K8CREERT2, K14rtTA/TetOCre). Scale bars, 100 µm. Error bars, sem.

Extended data Figure 2. Characterisation of tumours derived from basal or luminal cells upon oncogenic *Pik3ca* expression.

a-e, Characterisation of adenomyoepithelioma (adenomyo) tumours derived from K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice. **f-y,** Characterisation of tumours derived from K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice. **f-j**, Characterisation of adenomyoepithelioma. **k-o**, Characterisation of myoepithelial carcinoma (C). **p-t**, Characterisation of invasive carcinoma of no special type (NST C). **u-y**, Characterisation of mmetaplastic carcinoma. **a, f, k, p, u,** Haematoxylin eosin staining. **b, g, l, q, v,** p63 immunohistochemistry. **c, h, m, r, w,** Immunofluorescence of ER/K8. **d, i, n, s, x,** Immunofluorescence of K8/K14. **e, j, o, t, y**, Mean percentage of Ki67+ cells within tumours (n= 6, 3, 3, 1, 3 tumours respectively in **e, j, o, t, y**) . Error bars, SEM. Scale bars, 10 μm.

Extended data Figure 3. Similarities between mouse *Pik3ca*-derived mammary tumours and human breast cancers

a-d, Human breast tumour histologically classified as adenomyoepithelioma resembling to K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP derived tumours (K5PIK TU). **a**, Haematoxylin and eosin staining. **b-d**, p63 (**b**), K8/K18 (**c**) and K14 (**d**) immunohistochemistry in the human adenomyoepithelioma. **e-h**, Human breast tumour histologically classified as metaplastic carcinoma resembling to K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP derived tumours (K8PIK TU). **e**, Haematoxylin and eosin staining. **f-h**, p63 (**f**), K8/K18 (**g**) and K14 (**h**) immunohistochemistry in the human metaplastic carcinoma. **i-k**, Principal component analysis (PCA) of the METABRIC patients together with murine tumours according to the expression values of the PAM50 genes common to mice and humans. **i**, PCA of three K5CREERT2-*Pik3ca*-H1047R tumours (black dots) showing that these tumours cluster with human luminal B cancer subtype. **j**, PCA of seven K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP derived tumours (numbered black dots). Histological classification of each numbered tumour

is described below the figure. **k**, PCA of 2 K8CREERT2/*Pik3ca*-H1047R/*p53*fl/fl/Rosa-YFP derived tumours (K8PIKp53 TU)(black dots) showing that these tumours cluster together with human Her2+ subtype. **l**, Clustering of the murine tumours amongst human tumours of the METABRIC dataset. Clustering has been performed by grouping tumours presenting similar expression patterns of PAM50 genes. Colours on top of the heatmap represent the PAM50 subtypes attributed to the human tumours. The discrepancy between PCA and clustering analysis are due to the influence of Her2 low expression in these tumors, for which around 60% of PC2 relies on erbb2 expression. Scale bars, 10 μm.

Extended data Figure 4. Gating strategy to analyse and isolate tumour cells, LCs, and BCs according to their YFP, CD29 and CD24 profile.

a-e. Dot plot FACS analysis of unicellular suspension of mammary tumour cells (in this example from K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP tumour) stained for Lin (CD31, CD45, CD140a). Debris were eliminated from all events in P1 (**a**), doublets were discarded in P2 (**b**), the living cells were gated in P3 by DAPI dye exclusion (**c**), the non-epithelial Lin+ cells were discarded in P4 (**d**), and the YFP+ cells were gated in P5 (**e**). **f.** Gating strategy used for FACS analysis and cell sorting, showing the proportion of parent and total cells for each gate. Tumour cells were isolated based on their Lin- profile for YFP- tumours (P4 gate), or were isolated based in their YFP profile (P5 gate) for the YFP+ tumours, as described in methods. **g-m**. Dot plot FACS analysis of unicellular suspension of mammary cells (in this example from K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice 12 m after TAM induction) stained for CD24, CD29 and Lin (CD31, CD45, CD140a). Debris were eliminated from all events in P1 (**g**), doublets were discarded in P2 (**h**), the living cells were gated in P3 by DAPI dye exclusion (**i**), the non-epithelial Lin+ cells were discarded in P4 (**j**), and the YFP+ cells

were gated in P5 (**k**). **l, m.** CD29 and CD24 expression were used to gate the $CD29^{LO}CD24^{+}$ population, corresponding to LCs, and to gate the $CD29^{HI}CD24^{+}$ population, corresponding to BCs, either in YFP+ cells (**l**) or in Lin- cells (**m**). **n.** Gating tree showing the gating strategy used for FACS analysis and sorting, showing the proportion of parent and total cells for each gate.

Extended data Figure 5. Characterisation of tumours derived from BCs or LCs cells upon concomitant expression of oncogenic *Pik3ca* and deletion of p53.

a-o, Characterisation of tumours derived from K14rtTA/TetOCRE/*Pik3ca*-H1047R/*p53*fl/+/Rosa-YFP mice. **a-e**, Characterization of adenomyoepithelioma (adenomyo). **f-j**, Characterization of myoepithelial carcinoma. **k-o,** Characterization of metaplastic carcinoma. **p-aa,** Characterisation of tumours derived from K8CREERT2/*Pik3ca*-H1047R/*p53*fl/fl/Rosa-YFP mice. **p-t**, Characterization of myoepithelial carcinoma (C). **u-aa,** Characterization of metaplastic carcinoma. **a, f, k, p, u,** Haematoxylin eosin staining. **b, g, l, q, v,** p63 immunohistochemistry. **c, h, m, r, w,** Immunofluorescence of K8/ER. **d, i, n, s, x,** Immunofluorescence of K8/K14. **e, j, o, t, y**, Mean percentage of Ki67+ cells within tumours (n= 3, 3, 3, 3, 6 tumours respectively in **e, j, o, t, y**). **z,** Immunofluorescence of K8/Her2. **aa,** Immunofluorescence of Ecadherin/vimentin. Error bars, SEM. Scale bars, 10 μm.

Extended data Figure 6. Oncogenic *Pik3ca* expression induces multipotency in unipotent luminal progenitors

a-d, Immunofluorescence showing the expression of K8/YFP (**a, c**) or K5/YFP (**b, d**) 1w (**a, b**) and 7months (**c, d**) after TAM injection in control K8CREERT2/Rosa-YFP MG. **e**,

Percentage of YFP+ cells in LCs ($CD29^{Lo}/CD24^{+}$) and in BCs ($CD29^{Hi}/CD24^{+}$) at different time points following TAM administration to K8CREERT2/Rosa-YFP (n=3 mice per time point) showing that no YFP+ cells expressing $CD29^{Hi}/CD24^{+}$ was detected in control K8CREERT2/Rosa-YFP MGs at any time point. **f-h.** Immunofluorescence of K14/YFP (**f**), p63/YFP (**g**), SMA/YFP (**h**) 8w (**f, g**) or 10w (**h**) after TAM administration to K8CREERT2/*Pik3ca*-H1047/Rosa-YFP mice, shows that the BCs arising from LCs upon oncogenic *Pik3ca* targeting expressed these classical markers of BCs. **i-m**, Induction of *Pik3ca*-H1047R expression in LCs in adult mice. **i-l**, Immunofluorescence showing the expression of K8/YFP (**i, k**) or K5/YFP (**j, l**) 1w (**i, j**) and 8w (**k, l**) after TAM injection in K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice induced in adulthood. **m**, Percentage of YFP+ cells in LCs ($CD29^{Lo}/CD24^{+}$) and in BCs ($CD29^{Hi}/CD24^{+}$) at different time points following TAM administration to K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice induced in adulthood (n=4 mice per time point). **n, o.** Immunofluorescence of K5/YFP showing the clonal YFP expression in a single isolated LC 1w after TAM injection (**n**), and 8w after TAM injection showing a clone that gave rise to LC and BC (**o**) in K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP MG. Arrow in **n** points to the isolated LC, while arrow in **o** points to the newly arisen BC. **p**, distribution of clones 1w or 10w after TAM injection in K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP at clonal dose. Clones were scored as composed of only luminal cells (Luminal clones), composed of only basal cells (Basal clones) or composed of luminal and basal cells (Mixed clones) (n= 4 mice per time point). **q-t.** Immunofluorescence of K5/K8 (**q**), K5 (**r**), K8 (**s**) and K5/K8/YFP (**t**) shows that in WT MG K5 and K8 are not co-expressed (**q**), while K5/K8 double positive cells are observed in K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP MG 8w following oncogenic *Pik3ca* expression in LCs (**r-t**). Arrows in **r-t** point to K5+K8+YFP+ cells. **u, v** RT-PCR analysis of luminal (**u**) or basal (**v**) genes in YFP+ LCs and BCs sorted from K8CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice induced for 1w, 4w or 8w, in YFP+ LCs

derived from K8CREERT2/Rosa-YFP and in YFP+ BCs derived from K5CREERT2/Rosa-YFP mice induced for 8w. Data for luminal genes are compared to adult WT LCs (**u**) while data for basal genes are compared to adult WT BCs (**v**) (n=4 biologically independent samples per condition). Scale bars, 10 μm. Error bars, SEM.

Extended data Figure 7. Oncogenic *Pik3ca* expression induces multipotency in unipotent basal progenitors

a-c, Immunofluorescence showing the expression of K5/YFP (**a, b**) or K8/YFP (**c**) at 1w (**a**) and 7months (**b, c**) in control K5CREERT2/Rosa-YFP MG. **d**, Percentage of YFP+ cells in LCs ($CD29^{Lo}/CD24^{+}$) and in BCs ($CD29^{Hi}/CD24^{+}$) at different time points following TAM administration to K5CREERT2/Rosa-YFP (n= 5, 4, 4, 3 mice respectively for 1w, 8w, 7m, 12m) showing that no YFP+ cells expressing $CD29^{low}/CD24^{+}$ was detected in control K5CREERT2/Rosa-YFP MGs at any time point. **e-h.** Immunofluorescence of K19/YFP (**e**), ER/YFP (**f**), PR/YFP (**g**), claudin 3/YFP (**h**), 8 months after TAM administration to K5CREERT2/*Pik3ca*-H1047/Rosa-YFP mice, shows that LCs arising from BCs upon oncogenic *Pik3ca* targeting expressed these classical markers of LCs. **i**, Immunofluorescence of K5/YFP showing the YFP expression in a single isolated BC 1w after TAM injection at a clonal dose. Arrow points to the isolated BC. **j**, distribution of clones 1w or 7m after TAM injection in K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP at a clonal dose. Clones were scored as composed of only luminal cells (Luminal clones), composed of only basal cells (Basal clones) or composed of luminal and basal cells (Mixed clones) (n= 3, 4 mice respectively for 1w and 7m). **k, l** RT-PCR analysis of luminal (**k**) or basal (**l**) genes in YFP+ LCs and BCs sorted from K5CREERT2/*Pik3ca*-H1047R/Rosa-YFP mice induced for 10-12 months, in YFP+ LCs derived from K8CREERT2/Rosa-YFP and in YFP+ BCs derived from K5CREERT2/Rosa-

YFP mice induced for 10-12 months. Data for luminal genes are compared to adult WT LCs (**k**) while data for basal genes are compared to adult WT BCs (**l**) (n=4 biologically independent samples per condition). **m**, Confocal microscopy analysis of immunofluorescence of YFP, TrkB and K8 of MGs 7 months following *Pik3ca* expression in BCs, showing that the newly-formed LCs following *Pik3ca* expression in BCs co-expressed TrkB and K8. Arrow points to formed K8+/TrkB+/YFP+ cell. Scale bars, 10 μm. Error bars, SEM.

Extended data Figure 8. Molecular characterization of oncogenic *Pik3ca* induced multipotency

a, b, Venn diagram representing the common and distinct upregulated (**a**) and downregulated (**b**) genes in BCs and LCs following *Pik3ca* expression in BCs and LCs compared to age-matched control BCs and LCs respectively, with the name of the list of genes and number of genes in each section. The list of genes in each Venn section is provided in Supplementary Tables 2 and 3. (**c-l**). Venn diagrams representing the common genes upregulated (**c, e, g, i, k**) or downregulated (**d, f, h, j, l**) in the newly generated LCs or BCs following *Pik3ca*-H1047R expression in unipotent progenitors (**c, d**); in LCs and in BCs following *Pik3ca*-H1047R expression in LCs (genes regulated following the initial targeting of *Pik3ca*-H1047R in LCs, and thus reflecting the LC of origin) (**e, f**); in LCs and in BCs following *Pik3ca*-H1047R expression in BC (genes regulated by *Pik3ca*-H1047R in BCs, and thus reflecting the BC of origin) (**g, h**); in BCs following *Pik3ca*-H1047R expression in LCs and in BCs (genes regulated by *Pik3ca*-H1047R expression in BCs, irrespective of cell of origin) (**i, j**); in LCs following *Pik3ca*-H1047R expression in BC and in LCs (genes regulated by *Pik3ca*-H1047R expression in LCs, irrespective of cell of origin) (**k, l**). Diameter of the diagram is

proportional to the number of genes it contains. The reported hypergeometric *P-values* correspond to the probability to observe an intersection of this size by chance only, knowing the number of genes tested on a microarray chip.

Extended data Figure 9. Genes of luminal to basal multipotency signature correlate with patient outcome in untreated breast cancer patients.

a-d, Disease free survival in untreated patients according to the level of expression (low =blue, intermediate =green or high=red) of the genes in the luminal to basal multipotency signature namely *Ngf* (**a**), *inhba* (**b**), *Itgb6* (**c**), *Wnt10a* (**d**) showing that genes of luminal to basal multipotency signature predict disease free survival in untreated breast cancer patients. Patients expressing high levels of this signature are more prone to tumour relapse while those expressing lower levels of this signature show lower rate of relapse. The log rank p-values accounts for the significance of this difference.