LETTER

doi:10.1038/nature15742

Barcoding reveals complex clonal dynamics of *de novo* transformed human mammary cells

Long V. Nguyen[1]*, Davide Pellacani[1,2]*, Sylvain Lefort[1], Nagarajan Kannan[1,3], Tomo Osako[3,4], Maisam Makarem[1], Claire L. Cox[1], William Kennedy[1], Philip Beer[1], Annaick Carles[5], Michelle Moksa[5], Misha Bilenky[5,6], Sneha Balani[1], Sonja Babovic[1], Ivan Sun[7,8], Miriam Rosin[7,8], Samuel Aparicio[3,4], Martin Hirst[5,6] & Connie J. Eaves[1,2]

Most human breast cancers have diversified genomically and biologically by the time they become clinically evident[1–3]. Early events involved in their genesis and the cellular context in which these events occur have thus been difficult to characterize. Here we present the first formal evidence of the shared and independent ability of basal cells and luminal progenitors, isolated from normal human mammary tissue and transduced with a single oncogene (*KRAS*^{G12D}), to produce serially transplantable, polyclonal, invasive ductal carcinomas within 8 weeks of being introduced either subrenally or subcutaneously into immunodeficient mice[4]. DNA barcoding[5,6] of the initial cells revealed a dramatic change in the numbers and sizes of clones generated from them within 2 weeks, and the first appearance of many 'new' clones in tumours passaged into secondary recipients. Both primary and secondary tumours were phenotypically heterogeneous and primary tumours were categorized transcriptionally as 'normal-like'. This system challenges previous concepts that carcinogenesis in normal human epithelia is necessarily a slow process requiring the acquisition of multiple driver mutations. It also presents the first description of initial events that accompany the genesis and evolution of malignant human mammary cell populations, thereby contributing new understanding of the rapidity with which heterogeneity in their properties can develop.

To investigate the susceptibility of different normal human mammary cell types to transformation under the influence of known oncogenes, we isolated CD49f^{+}EpCAM^{low} basal cells (BCs), CD49f^{+}EpCAM^{+} luminal progenitors (LPs), CD49f^{-}EpCAM^{+} non-clonogenic luminal cells (LCs), and non-epithelial CD49f^{-}EpCAM^{-} stromal cells (SCs) at high purity (>97%) by fluorescence-activated cell sorting (FACS) from 37 normal human reduction mammoplasty samples depleted of endothelial and haematopoietic cells[4,7,8] (Fig. 1a and Extended Data Tables 1 and 2). We then exposed one or more of these subsets to one or more oncogene-encoding lentiviral preparations (encoding complementary DNAs (cDNAs) for TP53^{R273C} and green fluorescent protein (TP53^{R273C}–GFP), PIK3CA^{H1047R} and yellow fluorescent protein (PIK3CA^{H1047R}–YFP)), and KRAS^{G12D} and mCherry (KRAS^{G12D}–mCherry) and, in some experiments, to a library of biologically neutral, barcoded lentiviral GFP vectors to allow subsequent clonal tracking of their progeny using a DNA sequencing approach[5,6] (Extended Data Fig. 1a). The cells were then embedded in a collagen gel (0.3×10^5 to 16×10^5 cells per gel) and the gels transplanted into highly immunodeficient NOD-SCID *Il2rg*$^{-/-}$ (NSG) or NOD *Rag1*$^{-/-}$ *Il2rg*$^{-/-}$ (NRG) female mice.

In initial experiments, 2×10^5 irradiated C3H-10T1/2 fibroblasts were co-embedded in the gels which were then transplanted subrenally, followed by subcutaneous implantation of the recipients with a slow-release capsule containing 17-β-oestradiol and progestin (EP pellets)[4], recognizing this would limit follow-up to 8 weeks because of incurred bladder toxicity. BCs and/or LPs isolated from 17 of 27 normal donors and exposed to all three oncogenic vectors produced tumours within 8 weeks (Fig. 1b) at similar overall frequencies (46% of BC isolates and 61% of LP isolates, respectively, Extended Data Fig. 1b). Identical treatment of LCs and SCs isolated from three of these samples did not produce any tumours in the same 8-week period. Both the BC- and LP-derived tumours resembled invasive ductal carcinomas (Fig. 2a and Extended Data Table 3) and were histologically very different from the organized bilayered structures generated in analogous xenografts of unmanipulated or simply barcoded normal human mammary cells[4,6]. Secondary female immunodeficient mice transplanted subcutaneously with a small piece (~25–33% of the initial tumour mass)[5,9] from four of nine of these primary tumours (one BC- and eight LP-derived) developed palpable tumours within another 8 weeks (Extended Data Table 4).

FACS analysis showed most of the transduced cells in all primary tumours examined co-expressed all three fluorescent reporters, consistent with the high transduction efficiencies measured in separate cell aliquots maintained *in vitro* for 72 h after virus exposure, and a similar expression profile was maintained in the single secondary tumour similarly analysed (Extended Data Fig. 1c, d). Notably, when the three oncogene-encoding viruses were tested on their own, or in pairs, tumours were obtained with similar efficiency only when the *KRAS*^{G12D} vector was included, and even on its own (64/102 for all transductions that included *KRAS*^{G12D} compared with 1/12 when *KRAS*^{G12D} was not present; for more details, see Extended Data Fig. 1b and Extended Data Tables 1 and 2). PCR and Sanger sequencing confirmed the tumour cells contained the expected oncogene sequences including doubly and triply fluorescent cells isolated separately from tumours arising from cells initially exposed to three oncogenic vectors (Extended Data Fig. 2a–c).

Subsequent experiments demonstrated that invasive ductal carcinomas were also obtained at a similar frequency from both BCs and LPs (but not LCs or SCs from the same mammoplasty samples) when the transduced cells were transplanted subcutaneously without irradiated fibroblasts or EP pellets, even when the cells were exposed only to the *KRAS*^{G12D} vector (Extended Data Figs 1e and 2d and Extended Data Tables 1 and 2). These tumours could frequently also be serially passaged (Extended Data Fig. 2e and Extended Data Table 4) and their

[1]Terry Fox Laboratory, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. [2]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. [3]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. [4]Department of Molecular Oncology, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. [5]Centre for High-Throughput Biology, Department of Microbiology & Immunology, University of British Columbia, 2125 East Mall, Vancouver, British Columbia V6T 1Z4, Canada. [6]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. [7]Biomedical Physiology and Kinesiology, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada. [8]Cancer Control Unit, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada.
*These authors contributed equally to this work.

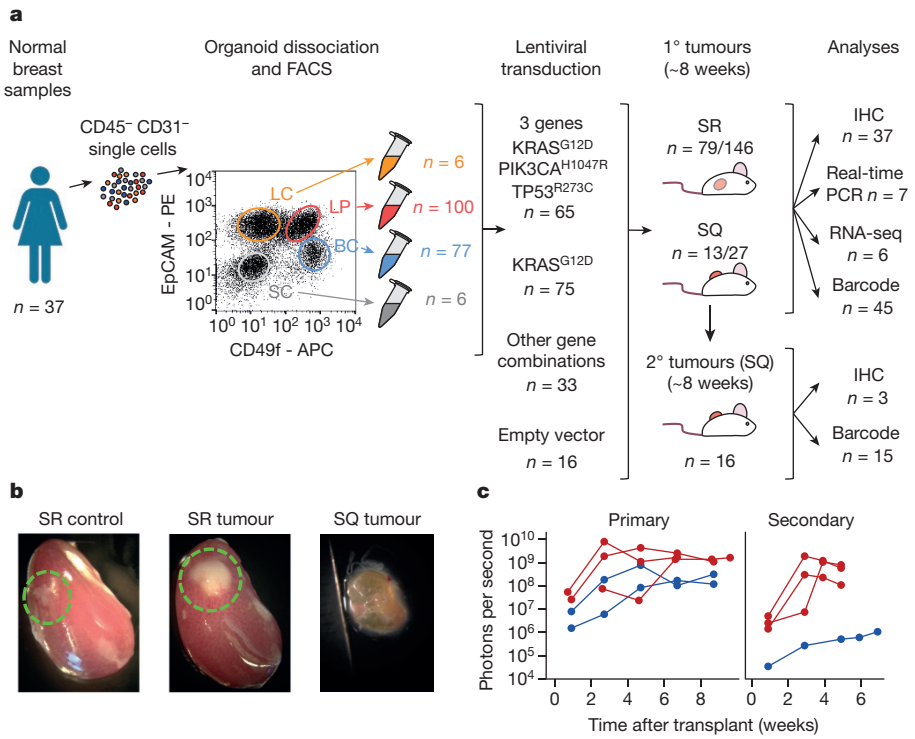


Figure 1 | *De novo* formation of tumours from normal human mammary BCs and LPs. a, Experimental design (n = number of experiments performed or tumours obtained). **b**, Examples of tumours compared with a control transplant which showed no evidence of a tumour when the mouse was killed. **c**, Bioluminescence measured over time in primary and secondary mice transplanted subcutaneously with oncogene- and luciferase-transduced BCs (blue) and LPs (red). SQ, subcutaneous; SR, subrenal.

growth more accurately monitored by luciferase bioluminescence (Fig. 1c and Supplementary Table 1).

FACS analysis of 15 tumours showed that 48 ± 5% of the cells were human EpCAM$^+$ and/or HLA$^+$, with similar results for BC- and LP-derived tumours (Extended Data Fig. 1c, f). Immunohistochemical (IHC) analyses of tumour sections (Fig. 2a, b) showed 88% and 55% of primary BC- and LP-derived tumours contained >5% ERα$^+$ cells (median = 58% and 8% ERα$^+$ cells, respectively), but none contained >2% PR$^+$ cells. HER2$^+$ cells were present at similar frequencies (in 88% and 52% of BC- and LP-derived tumours, respectively). Frequencies of Ki67$^+$ cells ranged from 2% to 30%, with only one secondary tumour containing as many as 70% Ki67$^+$ cells. In contrast, cells expressing EGFR, MUC1 and K8/18 were prevalent in almost all tumours examined. High K5 expression, normally exclusive to BCs, was prevalent (median = 90% K5$^+$ cells) in most LP-derived tumours, and less (median = 5% K5$^+$ cells) in BC-derived tumours. Expression of CD44, a marker associated with undifferentiated epithelial cells, was also less prevalent in BC- compared with LP-derived tumour cells (median = 2% and 50%, respectively). K14, another marker of normal human BCs, was also variably detected in both BC- and LP-derived tumours.

Gene expression analyses (Extended Data Fig. 2f) showed that transcripts for vimentin (*VIM*) and N-cadherin (*CDH2*), normally found exclusively in BCs, were present at high levels in both LP- and BC-derived tumours, with similar results for E-cadherin (*CDH1*) and *ELF5*, genes normally expressed exclusively in LPs and LCs. However, transcript levels of SLUG (*SNAI2*), another BC marker, were strongly decreased in BC-derived tumours whereas transcript levels of both *GATA3* and *NOTCH3*, two markers of LPs, showed little change. Cyclin-dependent kinase 1 (*CDK1*) was also highly expressed in all tumours, but other proliferation-associated genes, such as cyclin B1 (*CCNB1*) and *PCNA*, were highly expressed only in the LP-derived tumours. *TERT* transcripts remained at a similar level to that seen in the parental normal cell populations whereas those for *VEGFA*, *HIF1A* and *MAPK3* were more variable.

RNA sequencing analysis was conducted on FACS-purified human cells isolated from six primary tumours generated from triply transduced cells (three from BCs and three from LPs) and the matched starting cells. Unsupervised clustering showed a closer relation of coding gene transcript levels in all six tumour populations to each other than to the normal cells from which the tumours had arisen (Fig. 2c). This prominent sharing of transcriptome changes in tumours derived both from BCs and from LPs suggests a key role of their mode of creation on their resultant molecular features. Specific differences in the gene expression changes that distinguished BC- and LP-derived tumours and their respective starting cell populations showed shared increased and decreased expression of 146 and 22 genes, respectively in both, indicative of a common gene signature in the transformants (Fig. 2d, top). Further analyses using either PAM50 (ref. 10) or AIMS classifier methodologies[11] indicated the transcriptional profiles of the *de novo* tumours most closely resembled those of spontaneous human breast cancers classified as 'normal-like' (Fig. 2e).

However, the unsupervised clustering also indicated that the three BC- and three LP-derived tumours formed separate groups, suggesting some retained influence of their different origins. This was further supported by the finding that >20% (72) of the differentially expressed genes in the BC- and LP-derived tumours were similarly differentially expressed in the cells from which the tumours had arisen (Fig. 2d, bottom). Nevertheless, genes whose expression was upregulated in BC-derived tumours included several that are normally highly expressed in LPs and LCs but not BCs (for example, *AR*, *ESR1*, *FOXA1*, *TOX3*, *EPCAM*, *EHF* and *ELF5*). Conversely, the genes whose expression was upregulated in LP-derived tumours included several recognized BC-specific genes (for example, *VIM*, *TP63*, *ACTA2*, *THY1* and *CDH2*, Supplementary Table 2).

Clonal analyses were performed on primary tumours obtained from 45 isolates of BCs and LPs both from DNA extracted directly from tumour tissue and from FACS-purified human cells (Extended Data Fig. 3). The results showed a high variability in the clone content of different tumours (up to 1,700 using a threshold of 70 cells per

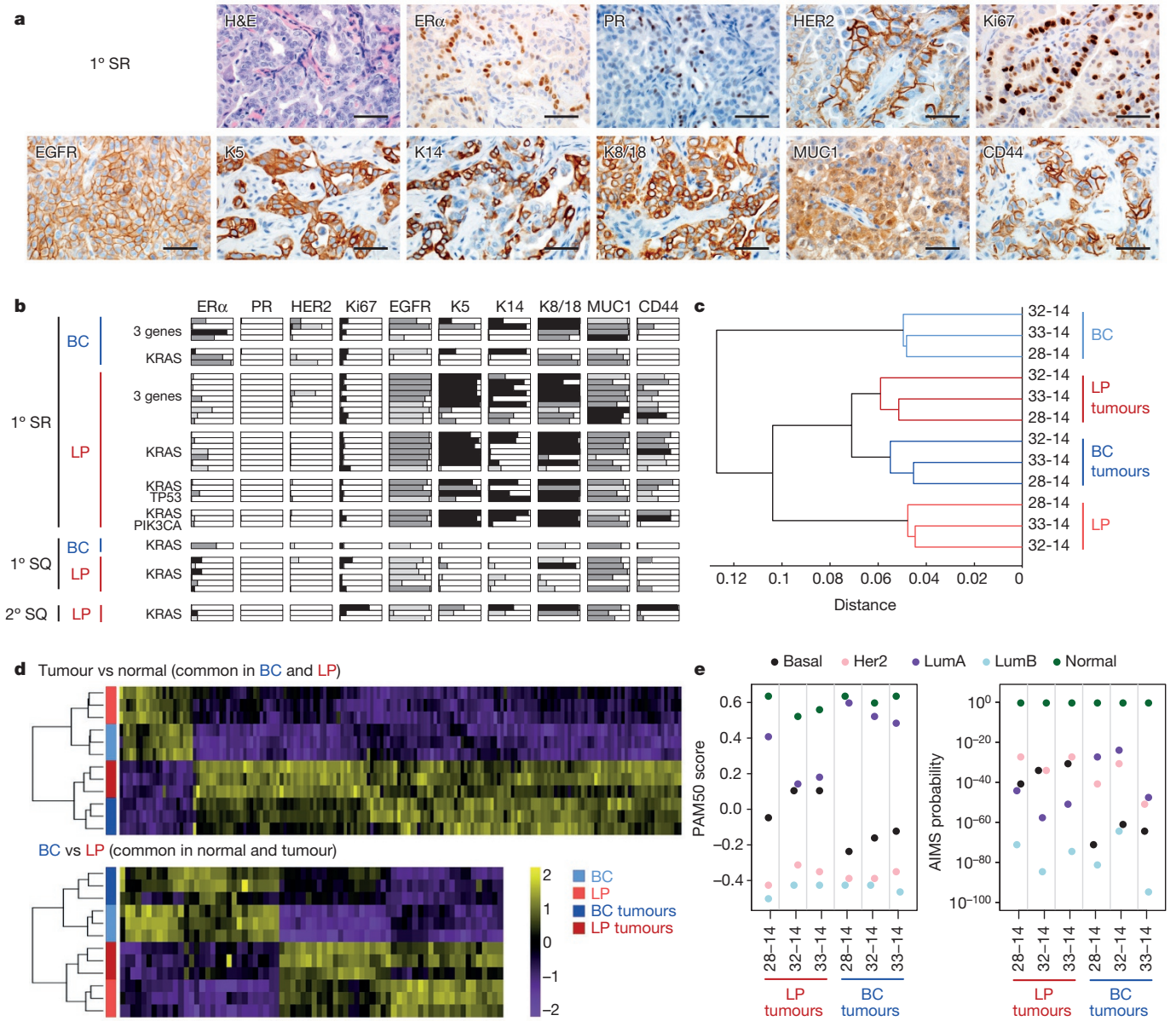


Figure 2 | Phenotypic heterogeneity of primary and secondary *de novo* tumours. a, Representative images of haematoxylin and eosin (H&E)- and antibody-stained sections from different primary BC- or LP- derived tumours arising from cells transplanted subrenally. Scale bar, 50 μm. **b**, Prevalence of cells in each tumour analysed expressing the indicated marker at different levels: white, negative; light grey, dim; dark grey, mid; black, bright. Each bar represents a single tumour. **c**, Unsupervised clustering (Spearman correlation and average linkage) of RNA sequencing data from *de novo* tumours compared with their matched cells of origin. **d**, Heat maps showing commonly differentially expressed genes between tumours and their matched cells of origin (top), and between BC- and LP-derived tumours in common with BCs and LPs (bottom). **e**, PAM50 and AIMS analyses of the RNA sequencing data from the *de novo* tumours.

clone), regardless of the protocol used for their generation (Extended Data Fig. 4a, b). Calculated (minimal) frequencies of tumorigenic clone-forming cells (T-CFCs) using the total number of initial cells transplanted as the denominator, ranged from 1/23,000 to 1/150. Paired comparisons for tumours produced from BCs and LPs from the same donor also did not reveal any effect on T-CFC frequency (Fig. 3a). To estimate clone sizes, we first derived 'relative' clone size values by normalizing each tumour to the sum of its absolute clone sizes. We then pooled the data for all tumours in each group being assessed. The overall distribution of relative clone sizes, like the clone frequencies, was very broad and showed no evidence of any effect of the cell of origin, oncogene(s) used or the transplantation site (Fig. 3b).

Analysis of 15 secondary tumours showed their clonal content was often high but again very variable, regardless of their origin (Fig. 3c and Extended Data Fig. 5a, b). Calculated frequencies of secondary clones (with respect to the number of cells initially transplanted into primary mice) also yielded highly variable secondary T-CFC values but with no consistent difference from the calculated primary T-CFC frequencies. However, >75% of the clones detected in each secondary tumour were 'new'; that is, not detected in the matching primary tumour (Fig. 3d). Moreover, most of the clones present in multiple sibling secondary tumours produced from a common primary tumour were also different from one another (two primary tumours analysed, Extended Data Fig. 5c). Overall the total measured T-CFC frequencies (calculated from the total number of different clones in the primary or secondary tumours combined) ranged from ~1/5,700 to ~1/120 (Extended Data Fig. 5d). The relative sizes of the clones present in secondary tumours were also highly variable (Fig. 3c). Interestingly, in secondary tumours, the median size of the 'continuing' clones (evident in both primary and secondary tumours) was significantly larger

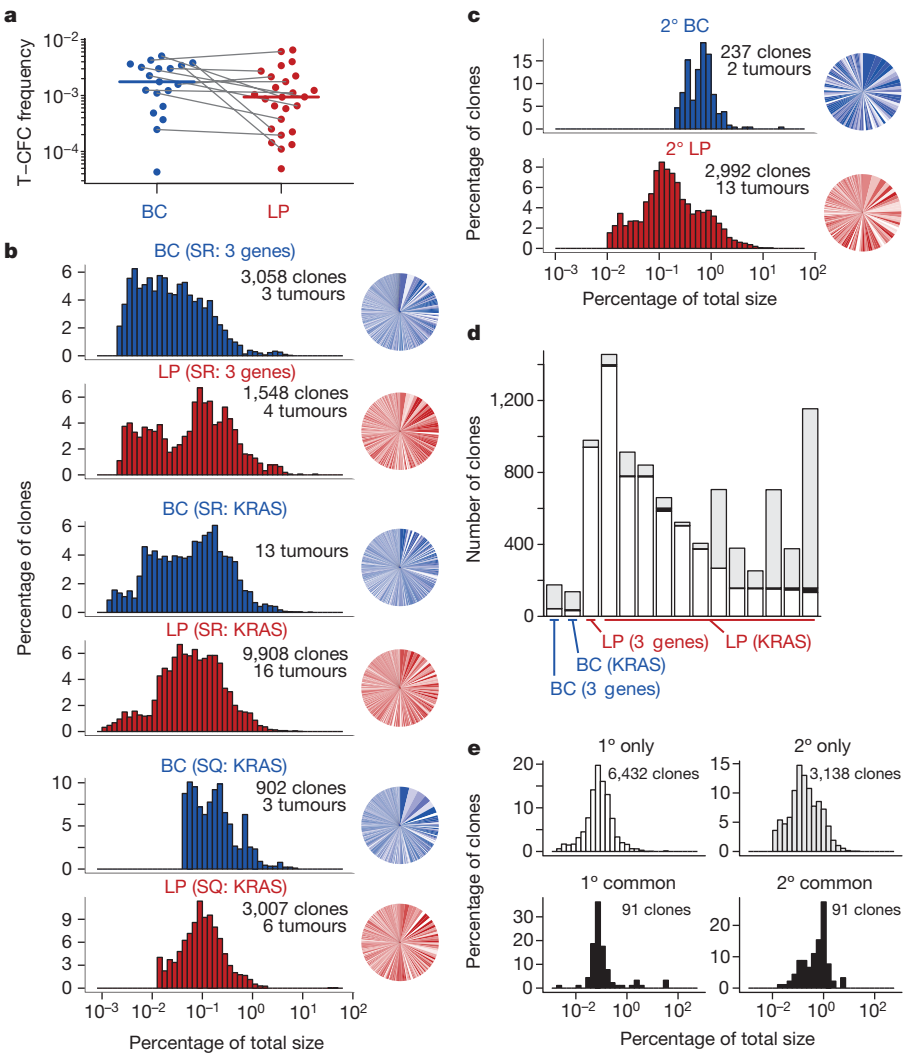


Figure 3 | Barcoding reveals a complex clonal landscape of primary and secondary tumours. a, Distribution of T-CFC frequencies (grey lines connect patient-matched BC- and LP-derived tumours, coloured lines show medians). **b**, Relative clone size distributions for all tumours in each group (left). Pie charts showing the clone size distributions in a representative tumour from each group (right). **c**, Distributions of the relative sizes of all clones pooled from all BC- and LP-derived secondary tumours (left). Pie charts showing the clone size distributions in a representative tumour from each group (right). **d**, Clones detected exclusively in primary (white) or secondary (grey) tumours, or in both (black). **e**, Clone size distributions of combined primary or secondary tumours subdivided into those detected exclusively in either primary (white) or secondary (grey) tumours or both (black).

than the clones that first became detectable upon tumour passaging ($P = 4 \times 10^{-12}$, Mann–Whitney *U*-test, Fig. 3e right panels).

We then analysed the clonal composition of the cells produced from oncogene-transduced BCs and LPs after just 2 weeks in subrenal transplants, before tumours become grossly evident. The results showed the sizes as well as the numbers of clones detected at this time to be similar to those detected 6 weeks later in tumours derived from the same input cells (Fig. 4a and Extended Data Fig. 6a, b). The distributions of the relative clone sizes measured in the 2-week transplants both of BCs and of LPs were also similar (Fig. 4b). However, after 2 weeks, the absolute sizes of the clones derived from the $KRAS^{G12D}$-transduced LPs were already significantly larger than the sizes of the clones produced by matching transplants of control vector-transduced cells (median = 206 and 93, respectively, $P = 3.3 \times 10^{-8}$, Mann–Whitney *U*-test), with a slightly smaller effect apparent in the progeny of BCs from the same two donors (median = 112 and 94, respectively, $P = 3.6 \times 10^{-7}$, Mann–Whitney *U*-test, Fig. 4c).

These studies provide new insights into the earliest phases of malignant transformation *in vivo* of cells isolated directly from normal human mammary tissue. Four findings are particularly noteworthy. The first is the rapidity and efficiency, albeit with high variability, with which this process can be induced in prospectively purified, biologically distinct types of normal human mammary epithelial cells using a single transducing oncogene ($KRAS^{G12D}$). This finding challenges previous assumptions of a requirement for a slow, multi-step selective process to accrue the genetic and/or epigenetic changes needed to obtain a continuously growing tumour. Interestingly, we did not obtain tumours from LCs or SCs subjected to the same protocols, in contrast to a recent report of highly ERα$^+$ tumours generated by transduction of EpCAM$^+$CD49f$^-$ LCs with SV40/Ras[12].

The second important finding was the considerable heterogeneity displayed in the numbers, phenotypes and growth behaviour of clonally tracked human cells with tumorigenic activity *in vivo* within 2–8 weeks. This result suggests that a similar range and speed of perturbations may accompany the spontaneous development of some breast cancers in patients.

A third and unexpected finding was the lack of a strong influence of the human mammary cell type initially transduced with the frequency of clones generated, the histopathology of the tumours produced or their loss of lineage-specific expression profiles. Taken together, this suggests a greater effect of the potent transforming role of the $KRAS^{G12D}$ oncogene in these cells.

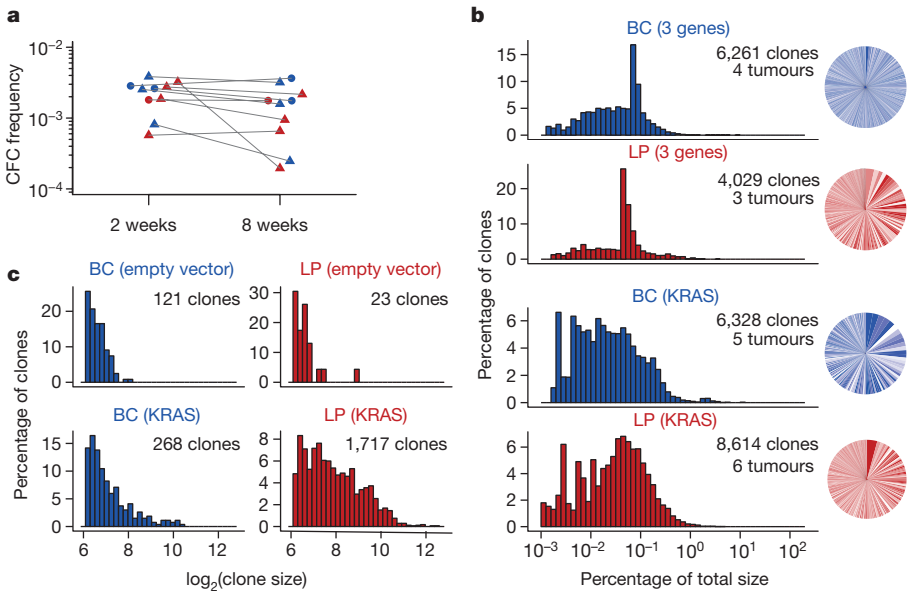


Figure 4 | Early changes in clone growth in cells transduced with *KRAS*^{G12D} only. a, Comparison of the frequencies of BCs (blue) and LPs (red) that made clones in 2-week grafts (CFCs) versus those that made clones found in tumours 6 weeks later (T-CFCs). Cells transduced with all three oncogenes are shown as triangles, and with *KRAS*^{G12D} only as circles (lines connect donor-matched samples). **b**, Distribution of relative clone sizes for grafts containing BCs and LPs transduced with three oncogenes or *KRAS*^{G12D} only (left). Pie charts showing the clone size distributions in a representative graft (right). **c**, Absolute clone size distributions of matched vector control-transduced BCs and LPs, and grafts containing BCs and LPs transduced with *KRAS*^{G12D} ($n = 2$ donors for each group).

The fourth finding was the frequent delayed activation of clonal growth observed in secondary tumours. This latency could either be biologically determined, reflecting an origin of these late appearing clones from their normal counterparts with similar features[5], or simply reflective of a stochastic process, as previously indicated for established human breast cancer cell lines passaged *in vivo*[4].

These results set the stage for examining the molecular basis of the biological heterogeneity now revealed that can occur during the earliest stages of breast cancer formation, the role of additional modifiers and how these may influence the acquisition of treatment response and resistance[13,14].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 31 July 2014; accepted 22 September 2015.
Published online 2 December 2015.

1. Stephens, P. J. *et al.* Oslo Breast Cancer Consortium (OSBREAC). The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486,** 400–404 (2012).
2. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98,** 10869–10874 (2001).
3. Curtis, C. *et al.* METABRIC Group. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (2012).
4. Eirew, P. *et al.* A method for quantifying normal human mammary epithelial stem cells with *in vivo* regenerative ability. *Nature Med.* **14,** 1384–1389 (2008).
5. Nguyen, L. V. *et al.* DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts. *Nature Commun.* **5,** 5871 (2014).
6. Nguyen, L. V. *et al.* Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells. *Cell Stem Cell* **14,** 253–263 (2014).
7. Kannan, N. *et al.* Glutathione-dependent and -independent oxidative stress-control mechanisms distinguish normal human mammary epithelial cell subsets. *Proc. Natl Acad. Sci. USA* **111,** 7789–7794 (2014).
8. Kannan, N. *et al.* The luminal progenitor compartment of the normal human mammary gland constitutes a unique site of telomere dysfunction. *Stem Cell Rep.* **1,** 28–37 (2013).
9. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518,** 422–426 (2015).
10. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).
11. Paquet, E. R. & Hallett, M. T. Absolute assignment of breast cancer intrinsic molecular subtype. *J. Natl Cancer Inst.* **107,** 357 (2015).
12. Keller, P. J. *et al.* Defining the cellular precursors to human breast cancer. *Proc. Natl Acad. Sci. USA* **109,** 2772–2777 (2012).
13. Kreso, A. *et al.* Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* **339,** 543–548 (2013).
14. Bhang, H. E. *et al.* Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature Med.* **21,** 440–448 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Wilkinson, G. Edin and M. Hale for technical support, E. Bovill, J. Boyle, S. Bristol, P. Gdalevitch, A. Seal, J. Sproul and N. van Laeken for access to discarded reduction mammoplasty tissue, T. Nielsen and N. Poulin for discussions, the Centre for Translational and Applied Genomics (BC Cancer Agency) for assistance with IHC, and T. MacDonald for assistance with rodent husbandry. This work was supported by grants from the Canadian Cancer Society Research Institute, the Canadian Breast Cancer Foundation and the Canadian Breast Cancer Research Alliance. L.V.N. received a Vanier Canada Graduate Scholarship from the Canadian Institutes of Health Research (CIHR), and N.K. was supported by a MITACS Elevate Fellowship. T.O. was supported by a Molecular Oncologic Pathology Fellowship from CIHR and the Terry Fox Foundation, and by grants from the Sumitomo Life Welfare and Culture Foundation, the Mochida Memorial Foundation for Medical and Pharmaceutical Research, and the Takashi Tsuruo Memorial Fund. S.A. is supported by a Canada Research Chair.

Author Contributions L.V.N., D.P. and C.J.E. designed the project, drafted the manuscript and were assisted by S.L., C.L.C., W.K. and S. Balani in performing the experiments. M.M. and M.H. oversaw the generation of sequence data, and L.V.N., D.P., A.C. and M.B. analysed it. All authors contributed to the interpretation of the results, and read and approved the manuscript.

Author Information Final transcriptome data has been deposited in the European Genome-phenome Archive (www.ebi.ac.uk/ega) under accession number EGAS00001001310. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.J.E. (ceaves@bccrc.ca).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cells. Reduction mammoplasty tissue was collected with informed consent, as approved by the University of British Columbia Research Ethics Board, and dissociated to obtain organoid-rich pellets which were then viably cryopreserved[4]. Thawed organoids were rinsed with Hank's Balanced Salt Solution supplemented with 2% FBS (HF), and the cells then dissociated in 2.5 mg ml^{-1} trypsin with 1 mM EDTA and 5 mg ml^{-1} dispase (STEMCELL Technologies) with 100 μg ml^{-1} DNaseI (Sigma), washing with HF between each step. The resulting cell suspension was filtered through a 40 μm mesh and BCs isolated by FACS according to their CD45$^-$CD31$^-$EpCAM^{lo}CD49f$^+$ (or CD45$^-$CD31$^-$CD10$^+$CD90$^+$CD49f$^+$) phenotype, LPs according to their CD45$^-$CD31$^-$EpCAM^{hi}CD49f$^+$ (or CD45$^-$CD31$^-$CD10$^-$CD90$^-$CD49f$^+$) phenotype, LCs according to their CD45$^-$CD31$^-$EpCAM^{hi}CD49f$^-$ phenotype and SCs according to their CD45$^-$CD31$^-$EpCAM$^-$CD49f$^-$ phenotype. In each case, a small aliquot of cells was immediately re-analysed to measure the purity of each sorted population (routinely >97%). Supplementary Table 3 lists the fluorochrome-labelled antibodies used.

Lentiviral constructs and transduction. Variations of the MNDU3-PGK–GFP lentiviral construct[15] encoding *YFP* or *mCherry* in place of the *GFP* reporter were generated and *KRAS^{G12D}*, *PIK3CA^{H1047R}* and *TP53^{R273C}* mutant cDNAs then cloned into these, using flanking AscI and PacI restriction sites downstream of the MNDU3 promoter. Human *KRAS* cDNA was cloned from a human cell line, and altered by site-directed mutagenesis to obtain the *G12D* mutant. The *TP53^{R273C}* mutant was cloned directly from a human cell line harbouring this mutation, and the human *PIK3CA^{H1047R}* cDNA was obtained from A. Weng (Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC, Canada). All cDNAs were sequence-verified before ligation into the lentiviral constructs. Clones confirmed to contain the mutant genes in the correct orientation were then selected for plasmid purification. Lentiviral supernatants containing ~10^9 infectious units per millilitre were produced[16] and added at a final dilution of 1:100 (for each) to cell suspensions containing 1×10^6 to 2×10^6 cells per 100 μl. The library of barcoded lentiviruses (titre of ~10^9 infectious units per millilitre)[6] was added to cells at a 1:200 dilution (~5×10^5 infectious units per 100 μl), to achieve an ~30% transduction efficiency.

Xenografts. Transduced human mammary epithelial cells were suspended in a neutralized rat tail collagen preparation[4,17] with 2×10^5 irradiated (15 Gy) C3H-10T1/2 mouse fibroblasts as indicated per 20 μl gel and the gels allowed to solidify at 37 °C for 30 min (refs 4, 17) before being implanted either subrenally or subcutaneously in 5- to 8-week-old virgin female NSG or NRG mice, that were then also implanted under the dorsal skin with a slow-release EP pellet, as indicated. Mice were bred, maintained and followed under specific-pathogen-free conditions in the Animal Resource Centre in the British Columbia Cancer Research Centre in accordance with protocols approved by the University of British Columbia Animal Care Committee. This included the monitoring of tumour growth for the periods indicated, or their removal earlier whenever a tumour reached a size of 1 cm^3. The length along the tumour's longest axis was measured at the time of removal, and was immediately categorized as small (<5 mm), medium (5–10 mm) or large (10–15 mm). No other measurements were recorded. For some of the primary tumours, a small mechanically dissociated fragment or an enzymatically dissociated cell suspension was prepared and transplanted subcutaneously with 50% (v/v) matrigel, with or without an EP pellet, as indicated, into secondary recipients. To measure tumour bioluminescence from luciferase expression, mice were injected intraperitoneally with 150 mg kg^{-1} body weight D-luciferin (Promega) and 10 min later imaged using a Xenogen IVIS Lumina system with Living Image version 3.0 software (Caliper Life Sciences). To prepare cell suspensions from tumours, the tissue was minced with a scalpel, incubated at 37 °C in DMEM/Ham's F12 media, supplemented with 5% FBS and 300 U ml^{-1} collagenase and 100 U ml^{-1} hyaluronidase for 2–4 h with periodic vortexing, washed with HF, and treated with 2.5 mg ml^{-1} trypsin with 1 mM EDTA and 5 mg ml^{-1} dispase with 100 μg ml^{-1} DNaseI. Human cells were sorted after staining with anti-human-specific antibodies directed against EpCAM and HLA with simultaneous depletion of mouse cells stained with anti-mouse-specific antibodies directed against BP-1, CD140a, CD45 and CD31 (Supplementary Table 3).

Histopathology and immunohistochemistry. Collagen gels or pieces of tumours obtained from mice were fixed in 10% buffered formalin (Fisher), washed in 70% ethanol and embedded in paraffin. A tissue microarray using 1.5 mm dual cores per sample was constructed for all primary tumours analysed. From this tissue microarray, 4 μm sections were obtained. All secondary tumours were embedded in paraffin and sections prepared directly from these blocks. The sections of both primary and secondary tumours were either stained directly with H&E, or were first treated with Target Retrieval solution (DAKO) and then a cytomation serum-free protein block (DAKO) followed by staining with either an anti-K14 antibody, an anti-MUC1 antibody, an anti-K5 antibody, an anti-K8/18 antibody, an anti-ER antibody, an anti-CD44 antibody, an anti-Ki67 antibody, an anti-HER2 antibody, an anti-PR antibody, or an anti-EGFR antibody. Use of a secondary rabbit antibody conjugated to horseradish peroxidase and treatment with 3,3′-diaminobenzidine (DAB, DAKO) was used to obtain a positive brown staining. Supplementary Table 3 provides details of the antibodies used and their sources. A negative control using one of the tumour samples, and a positive control using normal reduction mammoplasty tissue, was included for each marker analysed. Tumour sections stained with H&E were reviewed for histomorphological analysis. Each of the tumours was classified into one of the histological subtypes according to the World Health Organization human breast tumour classification[18]. Nuclear grade was scored on a three-point scale according to the National Surgical Adjuvant Study of Breast Cancer (NSAS-BC) grading system[19], and tubular formation, nuclear atypia, mitotic counts and histological grade were scored on a three-point scale according to the Nottingham grading system[20]. For each of the markers examined by IHC, intensity of staining (weak, intermediate or strong) and percentage of stained invasive tumour cells (0–100%) were scored. All of the histological and IHC parameters were scored blinded to the sample identity by a trained pathologist (T.O.).

Barcode analysis. Extracted genomic DNA was transferred to a 96-well plate in which researchers were blinded to their identity, although the order of the samples was not randomized. The samples were then treated identically, as previously described[6]. Defined numbers of control cells (10^2–10^6) containing a known barcode sequence at a single copy number per cell were analysed alone, and following their addition to each experimental sample, where they served as an internal normalization standard from which the relationship between fractional read value and cell numbers could be derived to calculate clone sizes (in absolute cell numbers). A threshold corresponding to a fractional read value equivalent of 70 cells was applied.

Real-time PCR. Total RNA was extracted from cryopreserved tumour samples using a mirVana miRNA isolation kit (Life Technologies) and cDNA then synthesized using SuperScript II Reverse Transcriptase (Life Technologies). Real-time PCR was performed using a SYBR Green master mix (Applied Biosystems) and samples were run in triplicate with the custom-designed primers listed in Supplementary Table 4. Human-specific primers were validated before use by testing for lack of reactivity with RNA from a panel of mouse tissues. Test gene cycle threshold (C_t) values were normalized by subtracting the geometric mean of ΔC_t values obtained for *GAPDH* and *EIF4A1* (control) genes.

RNA sequencing. RNA was extracted from snap-frozen cells using the mirVana miRNA Isolation Kit (Life Technologies, AM1560) followed by ethanol precipitation. RNA was quantified using an Agilent Bioanalyzer (Life Technologies) and 100 ng of total RNA was ribosomal RNA (rRNA) depleted using a NEBNext rRNA Depletion Kit (New England BioLabs, E6310L). First strand cDNA was generated using a Maxima H minus First Strand cDNA Synthesis Kit (Thermo Scientific, K1652) with added Actinomycin D (1 μg, Sigma, A9415). The product was purified using in-house-prepared 20% PEG in 1 M NaCl Sera-Mag bead solution at a 1.8× ratio and then eluted in 35 μl of Qiagen EB buffer. Second-strand cDNA was synthesized in a 50 μl volume using SuperScript Choice System for cDNA Synthesis (Life Technologies, 18090-019) with 12.5 mM GeneAmp dNTP Blend with dUTP. Double-stranded cDNA was purified with 20% PEG in 1 M NaCl Sera-Mag bead solution at a 1.8× ratio and eluted in 40 μl of Qiagen EB buffer, and fragmented using Covaris E220 (55 s, 20% duty factor, 200 cycles per burst). Sheared cDNA was end repaired/phosphorylated, single A-tailed, and adaptor ligated using custom reagent formulations (New England BioLabs, E6000B-10) and in-house-prepared Illumina forked adaptor. PEG (20%) in 1 M NaCl Sera-Mag bead solution was used to purify the template between each of the enzymatic steps. To complete the process of generating strand directionality, adaptor-ligated template was digested with 5 U of AmpErase Uracil N-Glycosylase (Life Technologies, N8080096). Libraries were then indexed and PCR amplified using Phusion Hot Start II High Fidelity Polymerase (Thermo Scientific, F 549-L). An equal molar pool was sequenced on an Illumina MiSeq platform, which produced between 3×10^6 and 4×10^6 aligned sequence reads.

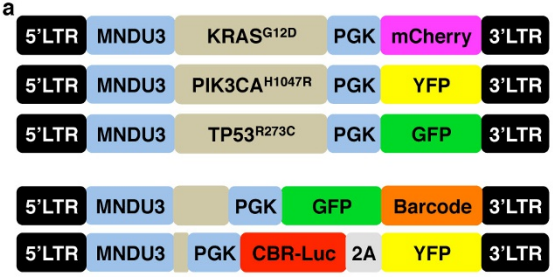
Adaptor sequences were stripped from the resulting 125-nucleotide sequence reads and the sequences uniformly trimmed to 75 nucleotides. Trimmed reads were aligned using BWA (version 0.5.7)[21] to a transcriptome reference[22] consisting of genomic sequence (GRCh37-lite July 2010) supplemented by read-length-specific exon–exon junction sequences. SAMtools (version 0.1.13)[23] was used to sort the alignment bam files. The sorted bam files were repositioned to GRCh37-lite using JAGuaR (version 2.0.3)[22] to assign sequences that aligned across exon–exon junctions to their correct 'split' genomic coordinates. An in-house RNA quality control and analysis pipeline[24] was used to generate a report (Supplementary Table 5) and calculate a normalization constant for computing RPKM values (reads per kilobase per million mapped reads). The normalization constant was inferred from the total

number of exonic reads (excluding mitochondrial reads, reads from ribosomal genes and reads from highest 0.5% expressed exons).

Pairwise comparisons between different sample types within the same donor were performed to identify differentially expressed genes using a custom DEfine matlab tool[24] (false discovery rate cutoff = 0.015, differentially expressed in at least two of three samples). PAM50 classification was performed following the methods described in ref. 10 using the R-Bioconductor script available at genome.unc.edu/pubsup/breastGEO/. AIMS classification[11] was performed using the Web tool available at www.bci.mcgill.ca/AIMS/.

15. Logan, A. C. *et al.* Factors influencing the titer and infectivity of lentiviral vectors. *Hum. Gene Ther.* **15,** 976–988 (2004).
16. Imren, S. *et al.* High-level β-globin expression and preferred intragenic integration after lentiviral transduction of human cord blood stem cells. *J. Clin. Invest.* **114,** 953–962 (2004).
17. Eirew, P., Stingl, J. & Eaves, C. J. Quantitation of human mammary epithelial stem cells with *in vivo* regenerative properties using a subrenal capsule xenotransplantation assay. *Nature Protocols* **5,** 1945–1956 (2010).
18. Lakhani, S. R., Ellis, I. O., Schnitt, S. J., Tan, P. H. & van de Vijver, M. J. *WHO Classification of Tumours of the Breast* 4th edn, Ch. 2 (World Health Organization, 2012).
19. Tsuda, H., Akiyama, F., Kurosumi, M., Sakamoto, G. & Watanabe, T.; Japan National Surgical Adjuvant Study of Breast Cancer (NSAS-BC) Pathology Section. Establishment of histological criteria for high-risk node-negative breast carcinoma for a multi-institutional randomized clinical trial of adjuvant therapy. *Jpn. J. Clin. Oncol.* **28,** 486–491 (1998).
20. Elston, C. W. & Ellis, I. O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* **19,** 403–410 (1991).
21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
22. Butterfield, Y. S. *et al.* JAGuaR: junction alignments to genome for RNA-seq reads. *PLoS One* **9,** e102398 (2014).
23. Li, H. *et al.* 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
24. Gascard, P. *et al.* Epigenetic and transcriptional determinants of the human breast. *Nature Commun.* **6,** 6351 (2015).

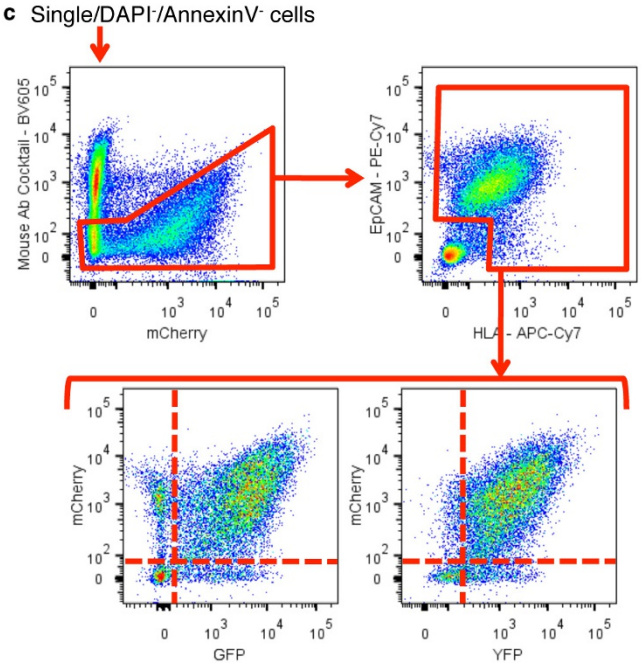
a



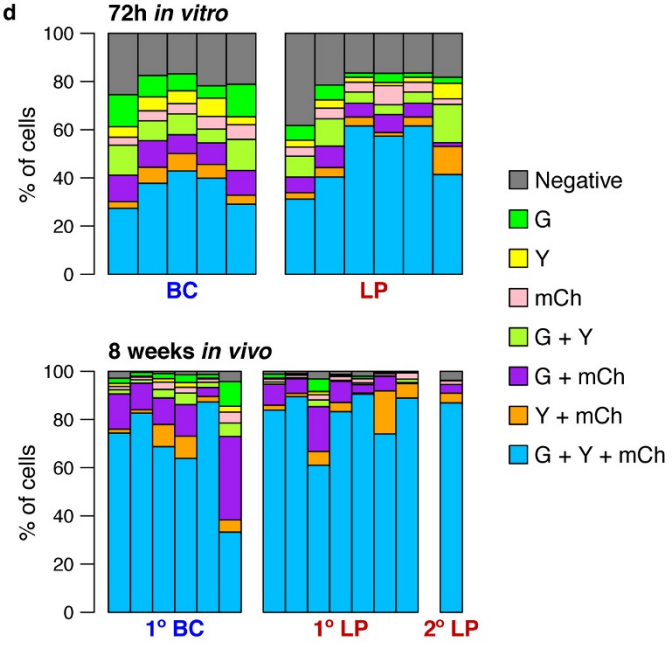
b

	KRAS TP53 PIK3CA	KRAS	KRAS TP53	KRAS PIK3CA	TP53	PIK3CA	Empty vector
BC	11/24 (46%)	11/18 (61%)	3/5 (60%)	2/5 (40%)	0/1 (0%)	0/1 (0%)	0/6 (0%)
LP	14/23 (61%)	14/16 (88%)	6/6 (100%)	3/5 (60%)	0/5 (0%)	1/5 (20%)	0/7 (0%)

c



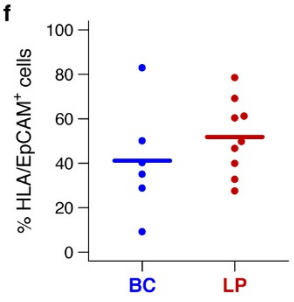
d



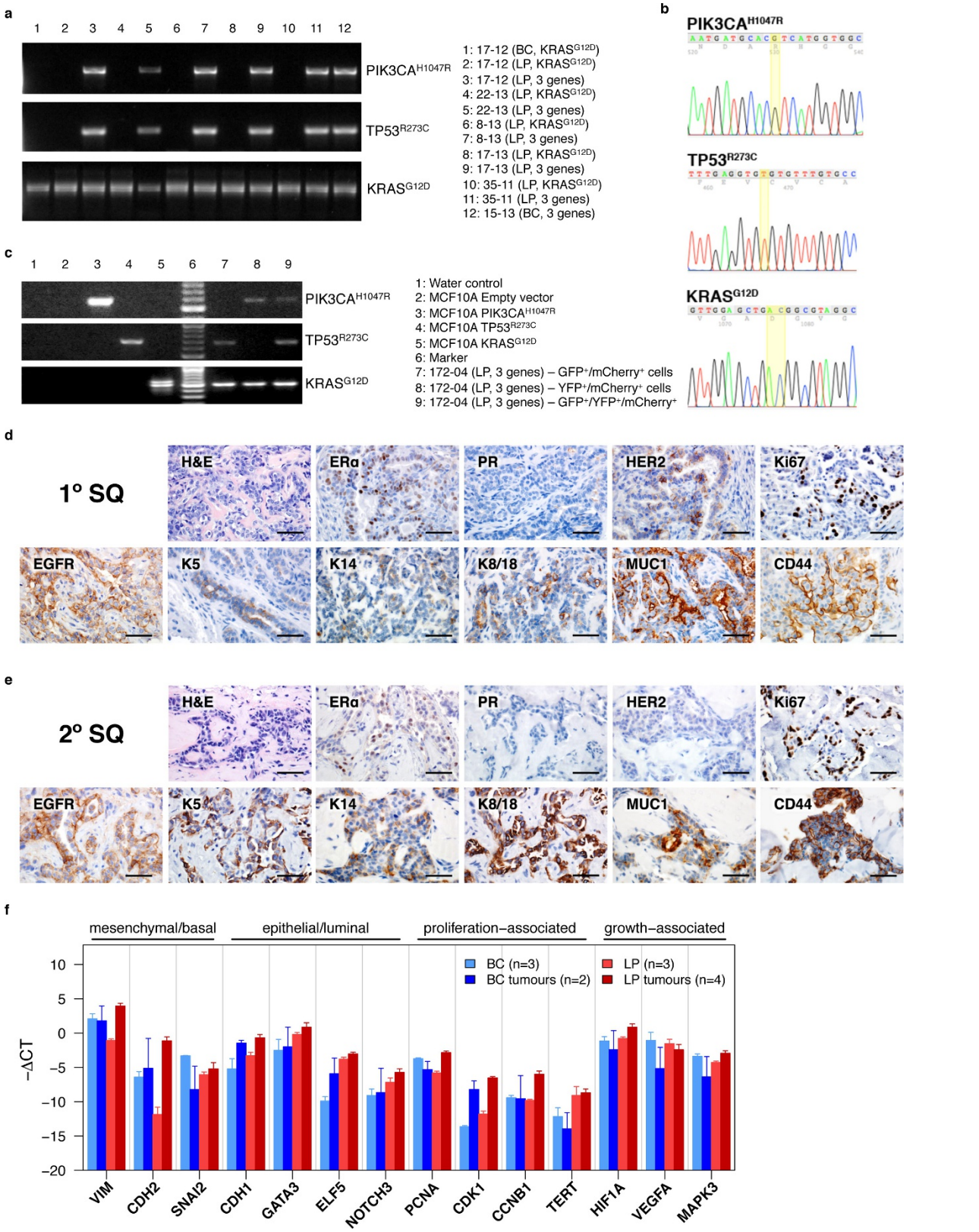
e

	KRAS			Empty vector
	SR -EP +fibroblasts	SQ +EP +fibroblasts	SQ -EP -fibroblasts	SQ -EP -fibroblasts
BC	2/3 (67%)	1/2 (50%)	4/6 (67%)	NA
LP	3/3 (100%)	3/3 (100%)	5/8 (63%)	0/3 (0%)

f

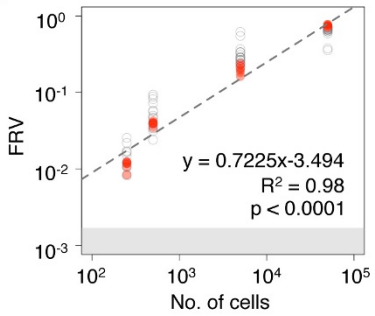


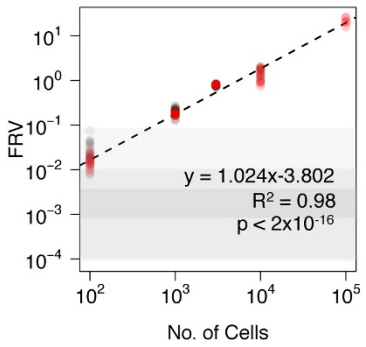
Extended Data Figure 1 | Quantification of human cells containing different vector reporters in tumours derived from triply transduced starting populations. a, Lentiviral constructs used. CBR-Luc, click beetle red luciferase. **b**, Frequencies of donor samples producing at least one tumour subrenally from BCs or LPs exposed to different combinations of oncogene-encoding vectors. **c**, Representative FACS profiles of a cell suspension prepared from a tumour produced from cells transduced with all three genes and sorted for human EpCAM and HLA using human-specific antibodies. **d**, Percentage of cells expressing different lentiviral reporters in cells maintained *in vitro* for 72 h after transduction, and in primary and secondary tumours (G, GFP; Y, YFP; mCh, mCherry). **e**, Frequencies of donor samples producing at least one tumour under various transplantation conditions using BCs or LPs transduced with $KRAS^{G12D}$. **f**, Percentages of human cells in BC- and LP-derived tumours detected by FACS on the basis of their expression of human EpCAM and/or HLA.



Extended Data Figure 2 | Molecular characterization of the tumours. **a**, Examples of PCR evidence of all three vectors in DNA extracts obtained from a subset of tumours analysed with vector-specific primers. **b**, A representative Sanger sequencing chromatograph showing the expected point mutations in the tumour cells analysed. **c**, PCR evidence of the three vectors in FACS-purified doubly and triply transduced cells. **d**, **e**, Representative images of H&E- and IHC-stained sections of primary tumours (**d**, arising from cells transplanted subcutaneously) and secondary tumours (**e**, all arising subcutaneously) derived from either BCs or LPs. Scale bar, 50 μm. **f**, Relative expression (negative ΔC_t values, mean ± s.e.m.) of gene transcripts typically associated with mesenchymal/basal or epithelial/luminal phenotypes, or associated with proliferation and cell growth.

a





b

No. of cells	Sensitivity		Specificity
	Controls	Samples	
$1x10^6$	6/6 (100%)	-	100% (486/486 known false positives excluded)
$5x10^4$	6/6 (100%)	40/40 (100%)	
$5x10^3$	6/6 (100%)	39/40 (98%)	
$5x10^2$	6/6 (100%)	28/40 (70%)	
$2.5x10^2$	6/6 (100%)	23/40 (58%)	

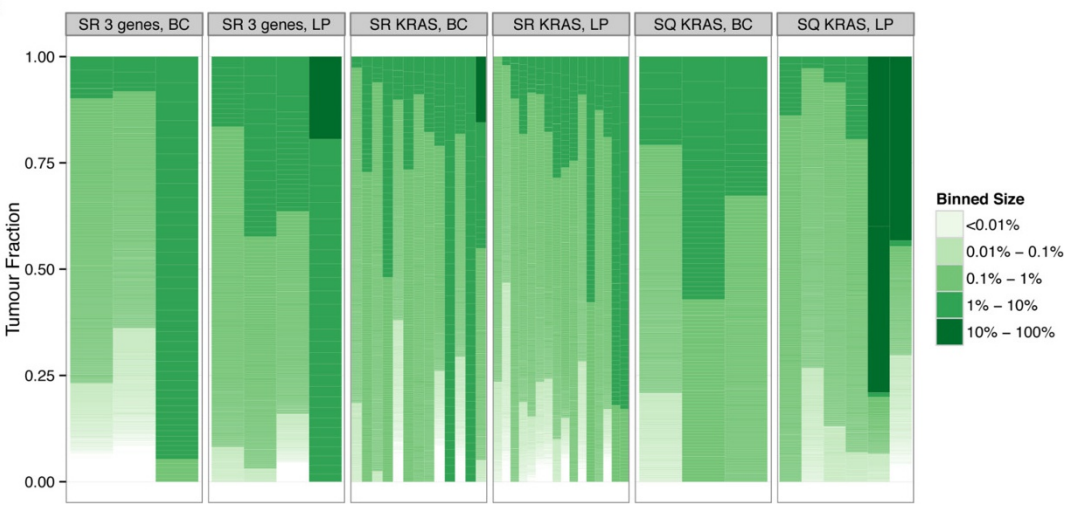
No. of cells	Sensitivity		Specificity
	Controls	Samples	
$1x10^6$	6/6 (100%)	-	97% (10116/10429 known false positives excluded)
$1x10^4$	12/12 (100%)	18/18 (100%)	
$3x10^3$	12/12 (100%)	48/48 (100%)	
10^3	12/12 (100%)	48/48 (100%)	
10^2	12/12 (100%)	37/48 (77%)	

Extended Data Figure 3 | Threshold set for detection of barcoded clones for the two sequencing runs from which barcode data were acquired. **a**, The relationship between the fractional read value (FRV) and the number of cells per clone. Spiked-in controls only and spiked-in controls added to experimental samples are shown as red and grey points, respectively. The shaded grey box represents distribution of false positive barcodes. **b**, Sensitivity and specificity data for controls compared with experimental samples for different sized clones.

a

Donor	BC (3 genes)		BC ($KRAS^{G12D}$)		LP (3 genes)		LP ($KRAS^{G12D}$)	
	No. of clones	T-CFC freq.	No. of clones	T-CFC freq.	No. of clones	T-CFC freq.	No. of clones	T-CFC freq.
SR +EP +Fibroblasts								
14-13	1,335	1/570	1,200	1/630	941	1/570	1,155	1/460
18-13	1,680	1/270	1,462	1/310	-	-	1,108	1/1,100
15-13	43	1/23,000	-	-	-	-	-	-
17-12	-	-	267	1/900	224	1/4,500	376	1/2,700
17-13	-	-	-	-	355	1/1,500	645	1/960
25-11	-	-	-	-	28	1/20,000	332	1/1,700
31-13	-	-	1,650	1/440	-	-	883	1/1,500
21-13	-	-	171	1/230	-	-	1,399	1/160
27-14	-	-	149	1/800	-	-	602	1/900
33-14	-	-	612	1/260	-	-	277	1/4,000
33-14	-	-	37	1/4,100	-	-	84	1/5,100
33-14	-	-	311	1/2,100	-	-	-	-
28-14	-	-	447	1/1,600	-	-	-	-
23-14	-	-	37	1/2,700	-	-	-	-
8-13	-	-	-	-	-	-	959	1/290
22-13	-	-	-	-	-	-	862	1/810
13-12	-	-	-	-	-	-	375	1/720
15-14	-	-	-	-	-	-	65	1/7,50
SR -EP +Fibroblasts								
13-12	-	-	115	1/330	-	-	735	1/370
15-14	-	-	292	1/290	-	-	54	1/9,100
SQ +EP +Fibroblasts								
27-14	-	-	609	1/200	-	-	505	1/1,100
21-13	-	-	-	-	-	-	923	1/250
SQ -EP -Fibroblasts								
27-14	-	-	88	1/820	-	-	-	-
15-14	-	-	205	1/330	-	-	-	-
21-13	-	-	-	-	-	-	372	1/1,200
23-14	-	-	-	-	-	-	268	1/450
28-14	-	-	-	-	-	-	158	1/7,000
28-14	-	-	-	-	-	-	781	1/150

b

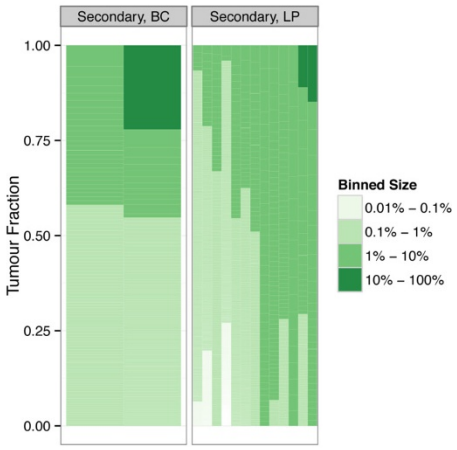


Extended Data Figure 4 | Clonal analyses of primary barcoded tumours. **a**, Numbers of clones and frequencies of T-CFCs in primary tumours. **b**, Relative clone size distributions for individual primary tumours grouped by the cell type initially manipulated and the oncogene(s) used. Each column represents a single tumour. Each rectangle represents one clone. Its relative clone size is indicated by the shade of green, and its proportional contribution within each tumour is indicated by its length on the *y* axis.

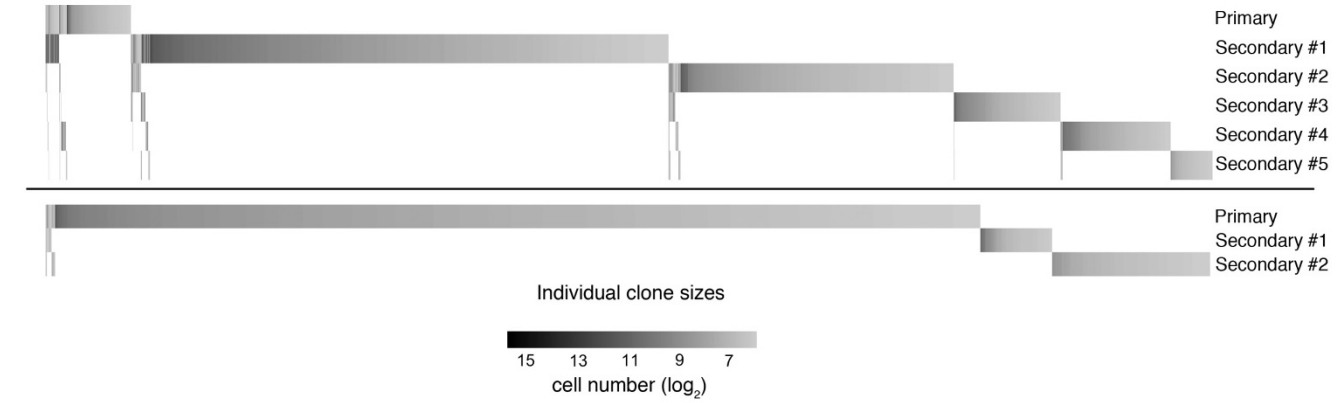
a

Donor	BC (3 genes)		BC (KRAS^{G12D})		LP (3 genes)		LP (KRAS^{G12D})	
	No. of clones	2° T-CFC freq.	No. of clones	2° T-CFC freq.	No. of clones	2° T-CFC freq.	No. of clones	2° T-CFC freq.
17-12	-	-	-	-	-	-	31	1/32,000
14-13	-	-	-	-	39	1/14,000	-	-
15-13	132	1/7,600	-	-	-	-	-	-
23-14	-	-	105	1/950	-	-	437	1/280
21-13	-	-	-	-	-	-	66	1/3,500
27-14	-	-	-	-	-	-	78	1/6,900
27-14	-	-	-	-	-	-	19	1/28,400
28-14	-	-	-	-	-	-	552	1/2,000
	-	-	-	-	-	-	99	1/11,000
	-	-	-	-	-	-	1021	1/1,100
	-	-	-	-	-	-	229	1/4,800
	-	-	-	-	-	-	224	1/4,900
28-14	-	-	-	-	-	-	136	1/880
	-	-	-	-	-	-	65	1/1,900

b



c



d

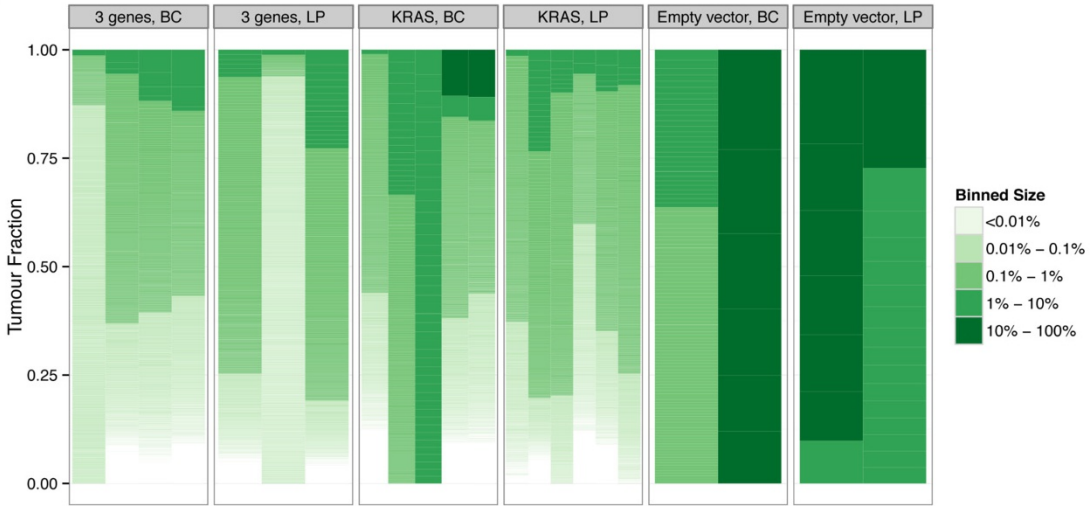
Donor	BC (3 genes)		BC (KRAS^{G12D})		LP (3 genes)		LP (KRAS^{G12D})	
	No. of clones	T-CFC freq.	No. of clones	T-CFC freq.	No. of clones	T-CFC freq.	No. of clones	T-CFC freq.
17-12	-	-	-	-	-	-	406	1/2,500
14-13	-	-	-	-	979	1/550	-	-
15-13	175	1/5,700	-	-	-	-	-	-
23-14	-	-	137	1/730	-	-	705	1/170
21-13	-	-	-	-	-	-	1456	1/160
27-14	-	-	-	-	-	-	660	1/820
27-14	-	-	-	-	-	-	523	1/1,030
28-14	-	-	-	-	-	-	2161	1/510
28-14	-	-	-	-	-	-	973	1/120

Extended Data Figure 5 | Clonal analyses of secondary barcoded tumours. a, Numbers of clones and frequencies of T-CFCs in secondary tumours. **b**, Relative clone size distributions for individual secondary tumours grouped by the cell type initially manipulated. Each column represents a single tumour. Each rectangle represents one clone. Its relative clone size is indicated by the shade of green, and its proportional contribution within each tumour is indicated by its length on the *y* axis. **c**, Clonal landscape of replicate secondary tumours generated from single primary tumours in two separate experiments. Clones present in sibling tumours are shown above one another and unique clones are shown in the same horizontal bar. Increasing clone sizes are indicated by a grey intensity scale. **d**, Numbers of clones and T-CFC frequencies of combined primary and secondary tumours.

a

Donor	BC (3 genes)		BC (KRAS^{G12D})		BC (Empty vector)		LP (3 genes)		LP (KRAS^{G12D})		LP (Empty vector)	
	No. of clones	CFC freq.	No. of clones	CFC freq.	No. of clones	CFC freq.	No. of clones	CFC freq.	No. of clones	CFC freq.	No. of clones	CFC freq.
24-13	1,218	1/170	2,365	1/90			1,759	1/280	2,471	1/200		
14-13	1,988	1/380	1,929	1/390			966	1/550	1,467	1/360		
18-13	1,307	1/350	1,766	1/260			1,304	1/900	2,187	1/540		
31-13	1,748	1/420							772	1/1,700		
28-14			226	1/160	6	1/5,700			1,112	1/100	16	1/5,900
33-14			42	1/1,200	115	1/560			605	1/300	7	1/46,000

b



Extended Data Figure 6 | Clonal analyses of transduced cells transplanted subrenally after 2 weeks *in vivo*. a, Number of clones and frequency of CFCs in xenografts of transduced cells assessed after 2 weeks *in vivo*. **b**, Relative clone size distributions of individual 2-week transplants grouped by the cell type initially manipulated and the oncogene(s) used. Each column represents a single transplant. Each rectangle represents one clone. Its relative clone size is indicated by the shade of green, and its proportional contribution within each tumour is indicated by its length on the *y* axis.

Extended Data Table 1 | Primary xenotransplant experiments performed subrenally with EP pellets and irradiated fibroblasts

3 Oncogenes, Primary SR, +EP +fibroblasts

Patient ID	BC				LP			
	Cells transplanted ($x10^3$)	Weeks	Size	Analyses	Cells transplanted ($x10^3$)	Weeks	Size	Analyses
14-13	1000	7	S	B/C	1000	7	S	IHC, B/C, RT-PCR
172-04	40	8.5	M	IHC, RT-PCR	700	8.5	M	IHC,FACS
32-14	300	8	S	IHC	770	8	M	IHC, FACS, RNA-Seq
	770	8	S	FACS, RNA-Seq	300	8	S	FACS
28-14	260	8	S	IHC, FACS, RNA-Seq	740	8	M	FACS, RNA-Seq
23-14	70	8	S	H&E	500	8	M	IHC, FACS
33-14	360	8	S	FACS	470	8	M	IHC, FACS, RNA-Seq
	-	-	-	-	470	8	M	FACS
34-13	460	6	S	FACS	-	-	-	-
18-13	30	4	S	None	60	4	S	None
	450	8	S	B/C	-	-	-	-
15-13	1000	7	M	IHC, B/C, RT-PCR	-	-	-	-
22-13	100	8	S	None	-	-	-	-
31-13	730	8	S	None	1400	8	S	None
17-12	240	8	-	No tumour	1000	8	S	IHC, B/C, RT-PCR
17-13	120	8	-	No tumour	620	8	S	IHC, B/C, RT-PCR
25-11	120	8	-	No tumour	570	8	M	IHC, B/C, RT-PCR
64-06	-	-	-	-	358	8	S	None
	-	-	-	-	358	8	S	None
51-09	-	-	-	-	700	8.5	M	IHC
74-03	170	8	-	No tumour	50	8	S	H&E, FACS
	170	8	-	No tumour	503	8	-	No tumour
199-04	-	-	-	-	500	8.5	-	No tumour
55-07	1000	6	-	No tumour	1000	6	-	No tumour
36-04	1000	6	-	No tumour	1000	6	-	No tumour
52-08	400	8.5	-	No tumour	540	8.5	-	No tumour
143-10	890	8	-	No tumour	1200	8	-	No tumour
	-	-	-	-	1200	8	-	No tumour
28-13	960	8	-	No tumour	1600	8	-	No tumour
	-	-	-	-	1600	8	-	No tumour
129-10	110	8	-	No tumour	120	8	-	No tumour
33-13	285	8	-	No tumour	413	8	-	No tumour
	-	-	-	-	413	8	-	No tumour
32-13	220	8	-	No tumour	700	8	-	No tumour
	-	-	-	-	700	8	-	No tumour
	-	-	-	-	700	8	-	No tumour
8-13	130	8	-	No tumour	-	-	-	-

KRAS(G12D) only, Primary SR, +EP +fibroblasts

17-12	240	8	S	B/C	1000	8	S	IHC, B/C
22-13	160	8	S	IHC	700	8	M	IHC, B/C
33-14	640	8	S	H&E, FACS, B/C	1100	8	S	FACS, B/C
	160	8	S	FACS, B/C	430	8	M	FACS, B/C
	150	8	S	FACS, B/C	-	-	-	-
27-14	120	8	S	IHC, FACS, B/C	540	8	S	IHC, FACS, B/C
21-13	40	8	S	H&E, FACS, B/C	230	8	M	H&E, FACS, B/C
14-13	760	8	S	B/C	540	8	S	B/C
31-13	730	8	S	B/C	1400	8	S	B/C
34-13	150	6	S	None	290	6	S	None
28-14	700	8	S	IHC, FACS, B/C	-	-	-	-
23-14	100	8	S	FACS, B/C	-	-	-	-
18-13	30	4	-	No tumour	60	8	S	None
	450	8	S	B/C	1000	8	M	IHC, B/C
17-13	120	8	-	No tumour	620	8	M	IHC, B/C
25-11	120	8	-	No tumour	570	8	S	IHC, B/C
13-12	38	8	-	No tumour	270	8	S	FACS, B/C
15-14	85	8	-	No tumour	490	8	S	FACS, B/C
8-13	-	-	-	-	280	8	M	IHC, B/C
29-14	65	8	-	No tumour	580	8	-	No tumour
74-06	200	8	-	No tumour	400	8	-	No tumour
	-	-	-	-	400	8	-	No tumour
	-	-	-	-	400	8	-	No tumour
	-	-	-	-	400	8	-	No tumour
33-13	63	9	-	No tumour	-	-	-	-

This table provides information for each experiment where transplants were performed subrenally with the addition of an EP pellet implant, and irradiated mouse fibroblasts were combined with the transduced human mammary cells in the transplanted collagen gel. Shown are the unique numbers assigned to donor-specific mammary samples, the subtypes of cells transduced and the oncogene(s) used for the transduction. Also shown are the sizes of the tumours obtained, and the subsequent analyses performed, where applicable. Small (S), medium (M) and large (L) refer to tumours for which the longest axis was measured as <5 mm, from 5 to 10 mm, or 10 to a maximum of 15 mm, respectively, to conform to our institutionally approved animal use protocols that stipulate tumours must be removed when they reach a size of 1 cm^3. When an experiment was not performed or a data field was not applicable, this is indicated with '-'. When a transplant was performed and no tumour was produced, this is indicated as 'no tumour'. The subsequent analyses performed are indicated, where applicable (B/C, barcode analysis).

Extended Data Table 2 | Primary xenotransplant experiments testing different variables

KRAS(G12D) only, Primary SR, no EP, +fibroblasts (continued)

Patient ID	BC				LP			
	Cells transplanted (x10^3)	Weeks	Size	Analyses	Cells transplanted (x10^3)	Weeks	Size	Analyses
13-12	38	16	S	FACS, B/C	270	9	M	FACS, B/C
15-14	85	16	S	FACS, B/C	490	9	S	FACS, B/C
33-13	63	9	-	No tumour	940	9	M	H&E
KRAS(G12D) only, Primary SQ, +EP +fibroblasts								
27-14	120	8	S	IHC, FACS, B/C	540	8	M	IHC, FACS, B/C
33-14	156	8	-	No tumour	1065	8	S	IHC
21-13	-	-	-	-	230	8	S	H&E, FACS, B/C
KRAS(G12D) only, Primary SQ, no EP, no fibroblasts								
38-14	10	7	S	FACS	135	9	-	No tumour
15-14	68	9	S	FACS, B/C	350	13	-	No tumour
27-14	72	9	M	H&E, FACS, B/C	405	13	-	No tumour
33-14	135	9	-	No tumour	210	7	S	FACS
	-	-	-	-	930	10	-	No tumour
42-14	280	9	-	No tumour	40	7	M	FACS
21-13	40	11	S	FACS	430	9	M	IHC, FACS, B/C
28-14	-	-	-	-	1100	9	M	IHC, FACS, B/C
	-	-	-	-	120	9	M	IHC, FACS, B/C
23-14	-	-	-	-	650	9	S	IHC, FACS, B/C
KRAS(G12D)+TP53(R273C), Primary SR, +EP +fibroblasts								
22-12	160	8	S	None	700	8	S	None
17-13	120	8	S	None	620	8	M	None
25-11	120	8	M	None	570	8	M	None
17-12	240	8	-	No tumour	1000	8	S	None
18-13	30	4	-	No tumour	60	4	S	None
8-13	-	-	-	-	280	8	L	None
KRAS(G12D)+PIK3CA(H1047R), Primary SR, +EP +fibroblasts								
17-12	-	-	-	-	1000	8	M	None
17-13	120	8	S	None	620	8	M	None
4-13	30	4	S	None	60	4	-	No tumour
22-13	160	8	-	No tumour	-	-	-	-
25-11	120	8	-	No tumour	570	8	M	None
8-13	130	8	-	No tumour	280	8	-	No tumour
TP53(R273C) only, Primary SR, +EP +fibroblasts								
17-12	240	8	-	No tumour	1000	8	-	No tumour
22-13	-	-	-	-	70	8	-	No tumour
17-13	-	-	-	-	620	8	-	No tumour
25-11	-	-	-	-	570	8	-	No tumour
8-13	-	-	-	-	280	8	-	No tumour
PIK3CA(H1047R) only, Primary SR, +EP +fibroblasts								
8-13	-	-	-	-	280	8	S	None
17-12	240	8	-	No tumour	1000	8	-	No tumour
22-13	-	-	-	-	70	8	-	No tumour
17-13	-	-	-	-	620	8	-	No tumour
25-11	-	-	-	-	570	8	-	No tumour

This table provides information for all other primary xenotransplant experiments not included in Extended Data Table 1. Shown are the unique numbers assigned to donor-specific mammary samples, the subtypes of cells transduced, the oncogene(s) used for the transduction, whether irradiated mouse fibroblasts were combined with transduced human mammary cells when a collagen gel was transplanted, the site of the transplant (subrenally or subcutaneously) and whether an EP pellet implant was used. Also indicated are the sizes of the tumours obtained, and the subsequent analyses performed, where applicable. Small (S), medium (M), and large (L) refer to tumours for which the longest axis was measured as <5 mm, from 5 to 10 mm, or 10 to a maximum of 15 mm, respectively, to conform to our institutionally approved animal use protocols that stipulate tumours must be removed when they reach a size of 1 cm^3. When an experiment was not performed or a data field was not applicable, this is indicated with '-'. When a transplant was performed and no tumour was produced, this is indicated as 'no tumour'. The subsequent analyses performed are indicated, where applicable (B/C, barcode analysis).

Extended Data Table 3 | Histopathological characterization of the *de novo* tumours

Patient ID	Starting cells	Histology	Fibrous Stroma	Nuclear grade	Histological grade	Tubular formation	Nuclear atypia	Mitotic index	Analyses
3 Oncogenes, Primary SR, +EP +fibroblasts									
15-13	BC	Invasive ductal	1	2	1	1	2	2	H&E, IHC
172-04	BC	Invasive ductal	-	-	-	-	-	-	H&E, IHC
	LP	Invasive ductal	2	1	1	2	2	1	H&E, IHC
28-14	BC	Invasive ductal	2	1	1	2	2	1	H&E, IHC
32-14	BC	Invasive ductal	2	2	1	1	2	2	H&E, IHC
	BC	Invasive ductal	2	1	1	1	1	1	H&E, IHC
	LP	Invasive ductal	2	1	1	2	1	1	H&E
	LP	Invasive ductal	2	1	1	1	2	1	H&E
23-14	BC	Invasive ductal	2	1	1	1	1	1	H&E
	LP	Invasive ductal	1	1	1	2	2	1	H&E, IHC
14-13	LP	Invasive ductal	3	2	1	2	2	1	H&E, IHC
17-12	LP	Invasive ductal	3	1	1	1	2	1	H&E, IHC
17-13	LP	Invasive ductal	2	1	1	2	2	1	H&E, IHC
33-14	LP	Invasive ductal	2	1	1	2	2	1	H&E, IHC
25-11	LP	Invasive ductal	2	1	1	1	2	1	H&E, IHC
51-09	LP	Invasive ductal	2	1	1	2	2	1	H&E, IHC
74-03	LP	Invasive ductal	-	1	1	1	1	1	H&E
KRASG12D only, Primary SR, +EP +fibroblasts									
21-13	BC	Invasive ductal	2	1	1	2	1	1	H&E
	LP	Invasive ductal	2	1	1	2	1	1	H&E
22-13	BC	Invasive ductal	2	1	1	2	2	1	H&E, IHC
	LP	Invasive ductal	3	1	1	1	1	1	H&E, IHC
27-14	BC	Invasive ductal	1	1	1	2	2	1	H&E, IHC
	LP	Invasive ductal	2	1	1	2	1	2	H&E, IHC
28-14	BC	Invasive ductal	2	1	1	1	2	1	H&E, IHC
33-14	BC	Invasive ductal	3	1	1	1	1	1	H&E
17-12	LP	Invasive ductal	2	1	1	1	2	1	H&E, IHC
17-13	LP	Invasive ductal	3	1	1	2	2	1	H&E, IHC
18-13	LP	Invasive ductal	3	1	2	3	2	1	H&E, IHC
35-11	LP	Invasive ductal	3	1	1	1	1	1	H&E, IHC
8-13	LP	Invasive ductal	2	1	1	1	1	1	H&E, IHC
KRASG12D only, Primary SR, no EP +fibroblasts									
33-13	LP	Invasive ductal	2	2	2	2	2	2	H&E
KRASG12D only, Primary SQ, +EP +fibroblasts									
27-14	BC	Invasive ductal	2	1	1	2	2	1	H&E, IHC
21-13	LP	Invasive ductal	2	1	1	2	1	1	H&E
27-14	LP	Invasive ductal	2	1	1	2	1	1	H&E, IHC
33-14	LP	Invasive ductal	1	2	2	2	2	2	H&E, IHC
KRASG12D only, Primary SQ, no EP, no fibroblasts									
27-14	BC	Invasive ductal	1	1	2	3	1	1	H&E
21-13	LP	Invasive ductal	1	1	1	2	1	1	H&E, IHC
23-14	LP	Invasive ductal	3	2	2	2	2	2	H&E, IHC
28-14	LP	Invasive ductal	3	3	2	2	2	3	H&E, IHC
	LP	Invasive ductal	3	2	2	2	2	2	H&E, IHC
Secondary tumors (all SQ)									
21-13	LP	Invasive ductal	2	1	1	2	2	1	H&E, IHC
28-14	LP	Invasive ductal	3	2	2	2	2	2	H&E, IHC
	LP	Invasive ductal	3	1	1	2	2	1	H&E, IHC
	LP	Invasive ductal	-	1	1	1	1	1	H&E
	LP	Invasive ductal	3	2	1	1	2	2	H&E

This table summarizes the results on all tumours analysed according to the cell type initially transduced, the oncogenes used, the site of the transplant and whether EP or irradiated fibroblasts were used. The amount of fibrous stroma present was scored as low (1), intermediate (2) or high (3). Nuclear and histological features were used to score tumours as low-grade or well-differentiated (1), intermediate-grade or moderately differentiated (2) or high-grade or poorly differentiated (3). Tubular formation was scored as >75% (1), 10–75% (2), or <10% (3) of the tumour area forming glandular/tubular structures. Nuclear atypia was scored as increasing from 1 to 3 on the basis of evidence of slight to marked variation compared with normal tissue. Increasing mitotic counts were scored as 0–4 (1), 5–10 (2), or ≥11 (3) mitotic figures per 400× high-power field. Tumours were also examined for evidence of necrosis, but none was found.

Extended Data Table 4 | Details of all secondary xenotransplant experiments

3 Oncogenes, Secondary tumours (all SQ)

Patient ID	Basal cells			Luminal Progenitor cells		
	Weeks	Size	Analyses	Weeks	Size	Analyses
15-13	8	M	B/C	-	-	-
14-13	-	-	-	8	M	B/C
18-13	-	-	-	8	M	None
64-06	-	-	-	7	S	FACS
17-13	-	-	-	8	-	No tumour
25-11	-	-	-	8	-	No tumour
23-14	-	-	-	9	-	No tumour
28-14	-	-	-	9	-	No tumour
33-14	-	-	-	9	-	No tumour
	-	-	-	9	-	No tumour

KRAS(G12D) only, Secondary tumours (all SQ)

23-14	7	S	FACS, B/C	5	S	FACS, B/C
27-14	9	-	No tumour	8	S	FACS, B/C
	-	-	-	8	S	FACS, B/C
17-12	-	-	-	8	M	B/C
28-14	9	-	No tumour	5	S	IHC, FACS, B/C
	9	-	No tumour	5	S	IHC, FACS, B/C
	-	-	-	5	S	H&E, FACS, B/C
	-	-	-	5	S	H&E, FACS, B/C
	-	-	-	5	S	FACS, B/C
	-	-	-	5	S	FACS, B/C
	-	-	-	5	S	FACS, B/C
	-	-	-	7	-	No tumour
21-13	9	-	No tumour	8	S	IHC, FACS, B/C
	-	-	-	9	-	No tumour
	-	-	-	9	-	No tumour
17-13	-	-	-	8	-	No tumour
13-12	9	-	No tumour	9	-	No tumour
42-14	-	-	-	9	-	No tumour
38-14	9	-	No tumour	-	-	-
33-14	9	-	No tumour	9	-	No tumour
	9	-	No tumour	-	-	-
	9	-	No tumour	-	-	-
15-14	9	-	No tumour	9	-	No tumour

The following information is indicated for each primary tumour tested for secondary tumour formation: the experimental conditions used to obtain the primary tumour, the size(s) of the secondary tumour(s) produced and the subsequent analyses performed, where applicable (B/C, barcode analysis). Small (S), medium (M), and large (L) refer to tumours for which the longest axis was measured as <5 mm, from 5 to 10 mm, or 10 to a maximum of 15 mm, respectively, to conform to our institutionally approved animal use protocols that stipulate tumours must be removed when they reach a size of 1 cm^3. When an experiment was not performed or a data field was not applicable, this was indicated with '-'. When a transplant was performed and no tumour was produced, this is indicated as 'no tumour'. The subsequent analyses performed are indicated, where applicable (B/C, barcode analysis).