Precision modeling of breast cancer in the CRISPR era

Annunziato, S.

4

Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular carcinoma

Sjors M. Kas[a*], Julian R. de Ruiter[a,b*], Koen Schipper[a*], **Stefano Annunziato**[a], Eva Schut[a], Sjoerd Klarenbeek[c], Anne Paulien Drenth[a], Eline van der Burg[a], Christiaan Klijn[a], Jelle J. ten Hoeve[b], David J. Adams[d], Marco J. Koudijs[a], Jelle Wesseling[a,e], Micha Nethe[a], Lodewyk F. A. Wessels[b,f,g] and Jos Jonkers[a,f]

[a] Division of Molecular Pathology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
[b] Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
[c] Experimental Animal Pathology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
[d] Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton, UK
[e] Department of Pathology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
[f] Cancer Genomics Netherlands, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
[g] Department of EEMCS, Delft University of Technology, Delft, the Netherlands
* The first three authors contributed equally to this work

Published in *Nature Genetics*, 2017 Jun 26.

Abstract

Invasive lobular carcinoma (ILC) is the second most common breast cancer subtype and accounts for 8–14% of all cases. Although the majority of human ILCs are characterized by the functional loss of E-cadherin (encoded by *CDH1*), inactivation of *Cdh1* does not predispose mice to develop mammary tumors, implying that mutations in additional genes are required for ILC formation in mice. To identify these genes, we performed an insertional mutagenesis screen using the *Sleeping Beauty* transposon system in mice with mammary-specific inactivation of *Cdh1*. These mice developed multiple independent mammary tumors of which the majority resembled human ILC in terms of morphology and gene expression. Recurrent and mutually exclusive transposon insertions were identified in *Myh9, Ppp1r12a, Ppp1r12b* and *Trp53bp2*, whose products have been implicated in the regulation of the actin cytoskeleton. Notably, *MYH9, PPP1R12B* and *TP53BP2* were also frequently aberrated in human ILC, highlighting these genes as drivers of a novel oncogenic pathway underlying ILC development.

Introduction

ILC belongs to the luminal subtype of breast cancer and accounts for 8–14% of all breast cancer cases (Martinez *et al.,* 1979; Borst *et al.,* 1993; Wong *et al.,* 2014). The majority of human ILCs (hILCs) are characterized by functional loss of E-cadherin (CDH1), a cell–cell adhesion molecule that is a key component of adherens junctions, where it associates with actin and the microtubule cytoskeleton to maintain epithelial integrity (Niessen *et al.,* 2008). Functional loss of E-cadherin in ILC generally results from mutational inactivation, loss of heterozygosity (LOH), or impaired integrity of the components of the E-cadherin–catenin complex (Moll *et al.*, 1993; Vos *et al.*, 1997; Ciriello *et al.*, 2015; Rakha *et al.,* 2010). Of note, female mice with mammary-specific inactivation of E-cadherin are not prone to developing mammary tumors (Boussadia *et al.*, 2002; Derksen *et al.*, 2006; Derksen *et al.*, 2011), indicating that additional mutations are required for ILC development.

Several studies have shed light on genetic alterations that are thought to be driver events in hILC, such as chromosomal gains of chromosomes 1q and 16p (Stange *et al.,* 2006), loss of chromosome 16q (Simpson *et al.,* 2008), activating mutations in PIK3CA (Buttitta *et al.,* 2006; Christgen *et al.,* 2013) and inactivating mutations in *TP53* (Ercan *et al.,* 2012). Molecular characterization of hILCs has further identified multiple aberrations in genes encoding components of the PI3K–AKT signaling pathway and increased AKT phosphorylation as compared to those in other breast cancer subtypes, underscoring the importance of PI3K–AKT signaling in hILC (Ciriello *et al.,* 2015; Michaut *et al.,* 2016; Desmedt *et al.,* 2016). However, only 50–60% of hILCs can be explained by PI3K–AKT activation and mutations in TP53, and relatively little is known about the roles of other genes and signaling pathways in hILC. To identify novel genes and pathways that drive ILC development, we performed a *Sleeping Beauty* (SB) insertional mutagenesis screen in mice that also had mammary-specific inactivation of *Cdh1*.

Results

Sleeping Beauty–induced mammary tumors in Wap–Cre;Cdh1F/F;SB mice

To generate mice with mammary-specific inactivation of E-cadherin and concomitant activation of the *Sleeping Beauty* (SB) insertional mutagenesis system, *Wap–Cre;Cdh1$^{F/F}$* mice were crossed with *T2/Onc;Rosa26^{Lox66SBLox71}* mice, which contain the transgenic SB transposon concatemer (*T2/Onc*) and the conditional SB11 transposase (*Rosa26^{Lox66SBLox71}*), resulting in *Wap–Cre;Cdh1$^{F/F}$;T2/Onc;Rosa26$^{Lox66SBLox71/+}$* (hereafter referred to as *Wap–Cre;Cdh1$^{F/F}$;SB*) mice (Figure 1A; Derksen *et al.*, 2011; Collier *et al.,* 2005; March *et al.,* 2011). In these mice, the transgenic Cre recombinase was expressed from the promoter of the mammary-specific gene *Wap*, resulting in the combined inactivation of *Cdh1* and the mobilization of transposons in mammary epithelial cells. To account for a potential bias toward transposition events occurring *in cis* on the chromosome containing the transgenic SB transposon concatemer, we used two different *T2/Onc* transgenic lines carrying the transposon donor loci on chromosomes 1 and 15, respectively. Mice that lacked at least one of the two SB components and mice that retained one wild-type allele of *Cdh1* were used as SB-inactive (*Wap–Cre;Cdh1$^{F/F}$*) and Cdh1-proficient (*Wap–Cre;Cdh1$^{F/+}$;SB*) control mice, respectively.

Wap–Cre;Cdh1$^{F/F}$;SB female mice developed multiple independent mammary tumors, with a significantly decreased median mammary tumor-specific survival (537 d) than in *Wap–Cre;Cdh1$^{F/F}$* female mice (Figure 1b). No difference in median survival was observed between *Wap–Cre;Cdh1$^{F/F}$;SB* mice carrying the *T2/Onc* transposon donor locus on chromosome 1 or 15 (Supplementary Figure S1A). Taken together, these data indicate that mammary-specific SB transposition accelerates mammary tumor formation in *Wap–Cre;Cdh1$^{F/F}$;SB* mice, underscoring the idea that additional mutations are required for malignant transformation of E-cadherin-deficient mammary epithelial cells.

SB-induced mammary tumors reflect human ILC

Histopathological analysis of 123 mammary tumors from 89 *Wap–Cre;Cdh1$^{F/F}$;SB* mice showed that 80% of the tumors (99/123) showed an infiltrative growth pattern with noncohesive E-cadherin-negative and cytokeratin 8 (CK8)-positive cells invading the surrounding tissue in single-cell strands, thus resembling hILC (Figure 1C-D and Supplementary Figure S1B-C). Growth patterns that were reminiscent of the alveolar or solid variants of ILC were also occasionally observed, with nests and sheets of tumor cells, respectively. As such, these tumors were classified as mouse ILC (mILC). Squamous metaplasia and tumors with a spindle cell morphology were observed in 24% and 44% of all tumors, respectively. Microscopic analysis showed metastasis in 34% of all tumor-bearing *Wap–Cre;Cdh1$^{F/F}$;SB* mice with predominant colonization of the lungs, lymph nodes, kidneys, spleen and liver (Figure 1E). In conclusion, SB-mediated insertional mutagenesis in *Wap–Cre;Cdh1*$^{F/F}$*;SB* female mice results in an accelerated development of mammary tumors, with the majority of tumors closely resembling hILC.

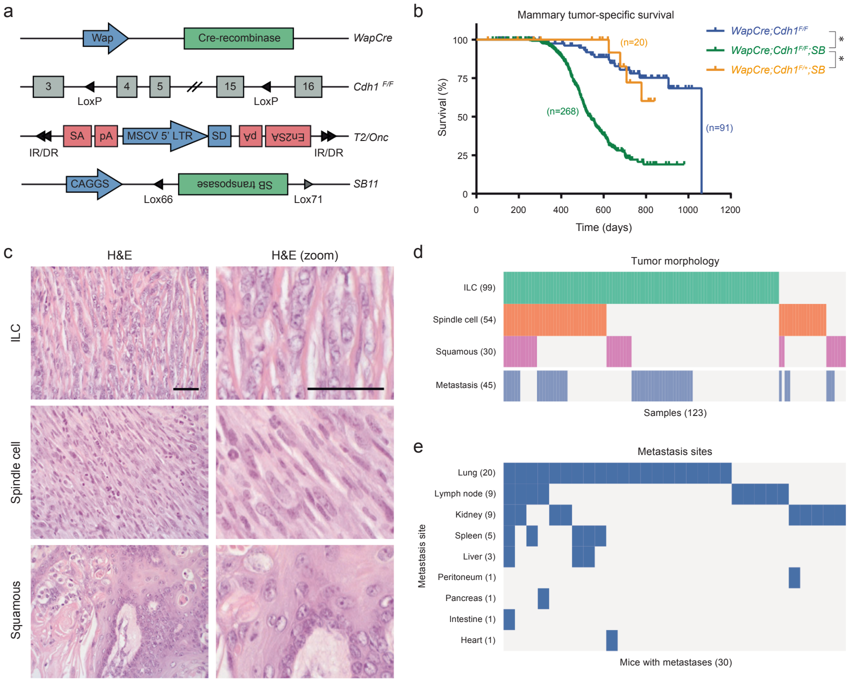


Figure 1 SB insertional mutagenesis induces tumorigenesis in female mice with mammary-gland-specific inactivation of E-cadherin. (A) Overview of the engineered alleles in *Wap–Cre;Cdh1*$^{F/F}$*;SB* mice. In this SB mutagenesis system, genetically engineered transposons, which contain a 5′ long terminal repeat (LTR) from the murine stem cell virus (MSCV) and two splice acceptor sites (SA/En2SA) in opposite orientations, are excised from a transgene concatemer by the SB transposase through indirect and direct repeats (IR/DR) and randomly reintegrated elsewhere in the genome (Collier *et al.,* 2005). Depending on the location and orientation of their insertion, these transposons can activate neighboring genes by inducing expression from the MSCV LTR or truncate gene transcripts using either of the splice acceptor sites. Numbered boxes represent exons of the canonical gene transcript. (B) Kaplan–Meier curve showing mammary tumor-specific survival (as defined in the Online Methods) for the indicated genotypes. *Wap–Cre;Cdh1*$^{F/F}$*;SB* (n = 268) females show reduced survival as compared to *Wap–Cre;Cdh1*$^{F/F}$ (n = 91) (537 d versus >1,000 d; P < 0.0001, Mantel–Cox test) and *Wap–Cre;Cdh1*$^{F/+}$*;SB* (n = 20) (537 d versus >1,000 d; P = 0.0002) females. *P < 0.05 by Mantel–Cox test. (C) Representative low- (left) and high-magnification (right) hematoxylin and eosin (H&E)-stained images of cells with the different morphologies (ILC, n = 99; spindle cell, n = 54; squamous metaplasia, n = 30). Scale bars, 50 µm. (D) Histological classification of 123 tumors from 89 *Wap–Cre;Cdh1*$^{F/F}$*;SB* females and the overlap with metastasis formation. (E) Overview of metastases to distant organs in metastasis-bearing *Wap–Cre;Cdh1*$^{F/F}$*;SB* females (30/89 mice).

To establish whether the SB-induced mammary tumors modeled the luminal breast cancer subtype of hILC, we used the PAM50 gene signature, which distinguishes intrinsic breast cancer subtypes, to cluster mouse tumors with human tumors from the Cancer Genome Atlas (TCGA; Parker *et al.,* 2009; Cancer Genome Atlas Network, 2012). For additional reference, two existing mouse models of luminal breast cancer (*Wap–Cre;Cdh1*$^{F/F}$*;Pten*$^{F/F}$; Boelens *et al.,* 2016) and basal-like breast cancer (*K14–Cre;Brca1*$^{F/F}$*;Trp53*$^{F/F}$; Liu *et al.,* 2007) were included in the clustering analysis. The resulting unsupervised hierarchical clustering showed that the majority of the SB-induced tumors coclustered with luminal breast cancers, confirming that these tumors reflected the luminal subtype (Figure 2A and Supplementary Figure S2A).

SB-induced tumors comprise distinct molecular subtypes

To determine whether the SB-induced mammary tumors consisted of distinct molecular subtypes, we used a non-negative matrix factorization (NMF) procedure to cluster tumors by their gene expression profiles. This analysis identified four subtypes (Figure 2B), which were not associated with a specific *T2/Onc* transgenic line (Supplementary Figure S3). Two of these subtypes (spindle-cell-like and squamous-like) were associated with a spindle cell morphology and squamous metaplasia, respectively (one-sided Fisher's exact test with Benjamini–Hochberg correction, false discovery rate (FDR) < 0.05). These morphological associations were supported by the expression of corresponding marker genes (Supplementary Figure S2B-C). The remaining two molecular subtypes consisted mainly of mILCs (FDR < 0.05), suggesting that the *Wap–Cre;Cdh1*$^{F/F}$*;SB* females developed two distinct subtypes of mILC (which we refer to as mILC-1 and mILC-2).
By projecting the gene expression profiles of these subtypes onto the PAM50 gene signature, we found that mILC-1 tumors were characterized by high expression of *Esr1* (which encodes estrogen receptor (ER)-α) and the ER transcriptional modulator *Foxa1*, as well as low expression of the proliferation marker *Mki67* (Figure 2C-D and Supplementary Figure S2D). Consequently, we found that mILC-1 tumors most closely reflect the luminal A subtype of tumors (Carroll *et al.,* 2005; Hurtado *et al.,* 2011; Goldhirsch *et al.,* 2011). As compared to mILC-1 tumors, mILC-2 and spindle-cell-like tumors generally showed lower expression of *Esr1* and higher expression of *Mki67*, indicating that these tumors are more proliferative. Squamous-like tumors were mainly distinguished by the high expression of keratin-encoding genes, such as *Krt5*.

To explore the potential links between our mILC subtypes and the three subtypes (reactive-like, immune-related and proliferative) that were identified in hILC (Ciriello *et al.,* 2015), we compared our mILCs with hILCs from the TCGA ILC study using the TCGA 60-gene subtype classifier. After translating this 60-gene signature into a mouse signature using 49 orthologous mouse genes, we combined the two data sets and compared the

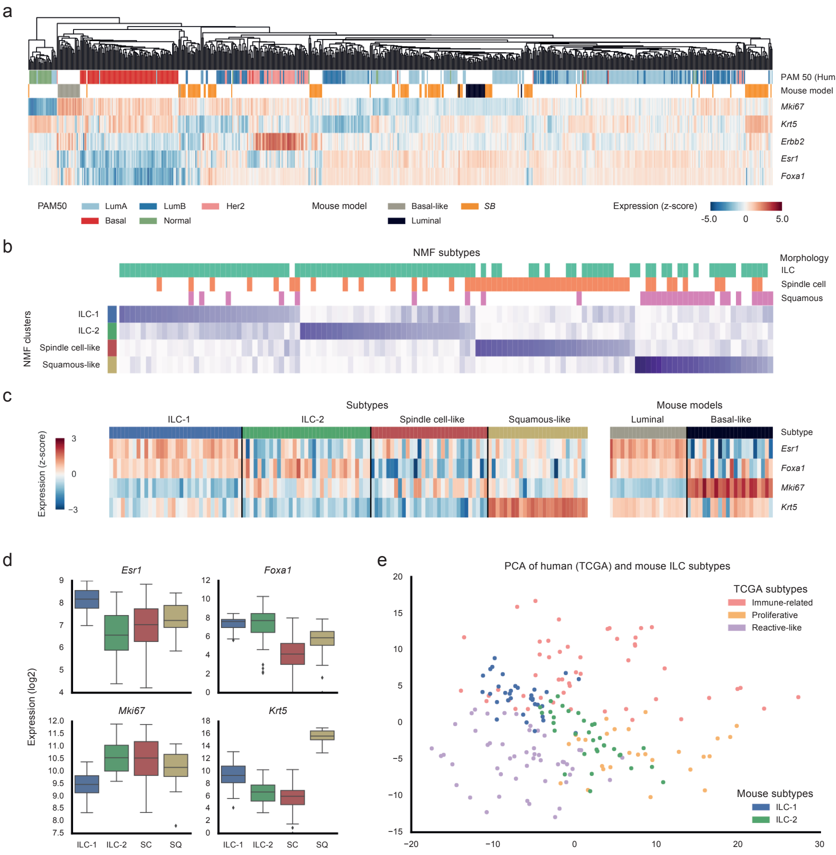


Figure 2 Gene expression analysis of SB-induced tumors. (A) Unsupervised clustering analysis (Euclidean distance, average linkage) of the SB-induced tumors (n = 123) with human breast cancer samples from TCGA (LumA, n = 231; LumB, n = 127; basal-like, n = 95; HER2-enriched, n = 57 and normal-like, n = 29) and tumors derived from mouse models of luminal (n = 20) and basal-like (n = 22) breast cancer using the PAM50 gene signature. The clustering was performed using 46 orthologous mouse genes from the PAM50 signature, but only a representative subset of genes is shown. (B) Coefficient matrix from the non-negative matrix factorization (NMF) analysis of the SB-induced tumors, indicating the membership of each sample to each of the four subtypes (ILC-1, n = 34; ILC-2, n = 33; spindle-cell-like, n = 30; squamous-like, n = 26). The matrix is annotated with the morphological characteristics of samples and shows a clear association between the clusters and the different morphologies. (C-D) Heat map (C) and quantification (D) of the expression of four key genes from the PAM50 gene signature for the different SB-induced subtypes and the mouse reference models, highlighting differences in expression between the different subtypes described in b. SC, spindle-cell-like; SQ, squamous-like. Boxes extend from the third (Q3) to the first (Q1) quartile (interquartile range, IQR), with the line at the median; whiskers extend to Q3 + 1.5 × IQR and to Q1 − 1.5 × IQR. Points beyond the ends of the whiskers are outliers. (E) Principal component analysis (PCA) plot comparing the two mILC subtypes to the hILC subtypes from TCGA (immune-related, n = 50; reactive-like, n = 50; proliferative, n = 27) using orthologous genes from TCGA's 60-gene subtype classifier. a.u., arbitrary units.

expression of the genes using principle component analysis (PCA). This analysis showed that mILC-2 tumors are more similar to the proliferative human subtype, which was also supported by the relatively higher expression of *Mki67* in mILC-2, whereas mILC-1 tumors reflected the immune-related human subtype (Figure 2E).

Identification of candidate genes involved in ILC development via SB insertional mutagenesis

To identify the genes that were involved in ILC development, we sequenced the SB transposon insertion sites of the 99 tumors with an ILC morphology by using the ShearSplink protocol, which permits semiquantitative high-throughput analysis of insertion sites (Koudijs *et al.,* 2011). This allowed us to determine both the location and the relative clonality of the insertions within each tumor. We then used Gaussian kernel convolution (GKC) to identify common insertion sites (CISs; de Ridder *et al.,* 2006), which represented genomic loci that were more frequently occupied by SB insertions than those expected by chance, and assigned CISs to putative target genes using a rule-based mapping (RBM) approach (Figure 3A; de Jong *et al.,* 2011).

This analysis identified 3,230 insertions with a median of 29 insertions per tumor (Supplementary Figure S4). From these insertions, we identified 58 CISs, which could be assigned to 30 candidate genes that were potentially involved in ILC development (hereafter referred to as candidate genes) (Figure 3B). A comparison between the *T2/Onc* lines showed that, although line-specific biases were evident for four candidate genes that were located *in cis* with the donor locus (*Myh9, Ppp1r12b, Trps1* and *Trp53bp2*), only *Trp53bp2* showed significant bias toward one of the lines. Furthermore, separate analyses on the individual *T2/Onc* lines independently identified these genes as CISs, demonstrating that none of these CIS-associated genes were unique to either line. We therefore decided to include the chromosomes that contained the donor loci in the CIS analysis to increase the power of the screen.

To prioritize candidate genes, we ranked the genes by their frequency and the median value of the clonality of their insertions (Supplementary Figure S5A). Using this approach, we selected 19 main candidate genes that were mutated in at least six samples, four of which were mutated in more than 25 samples (*Fgfr2, Trps1, Ppp1r12a* and *Myh9*). The majority of these genes had a high median clonality (≥0.5), which supported their role as drivers of ILC (Supplementary Figure S5B). In contrast, a subset of genes (for example, *Rasa1, Setd5* and *Ywhae*) had a lower clonality, which indicated that these may represent later events in tumorigenesis.

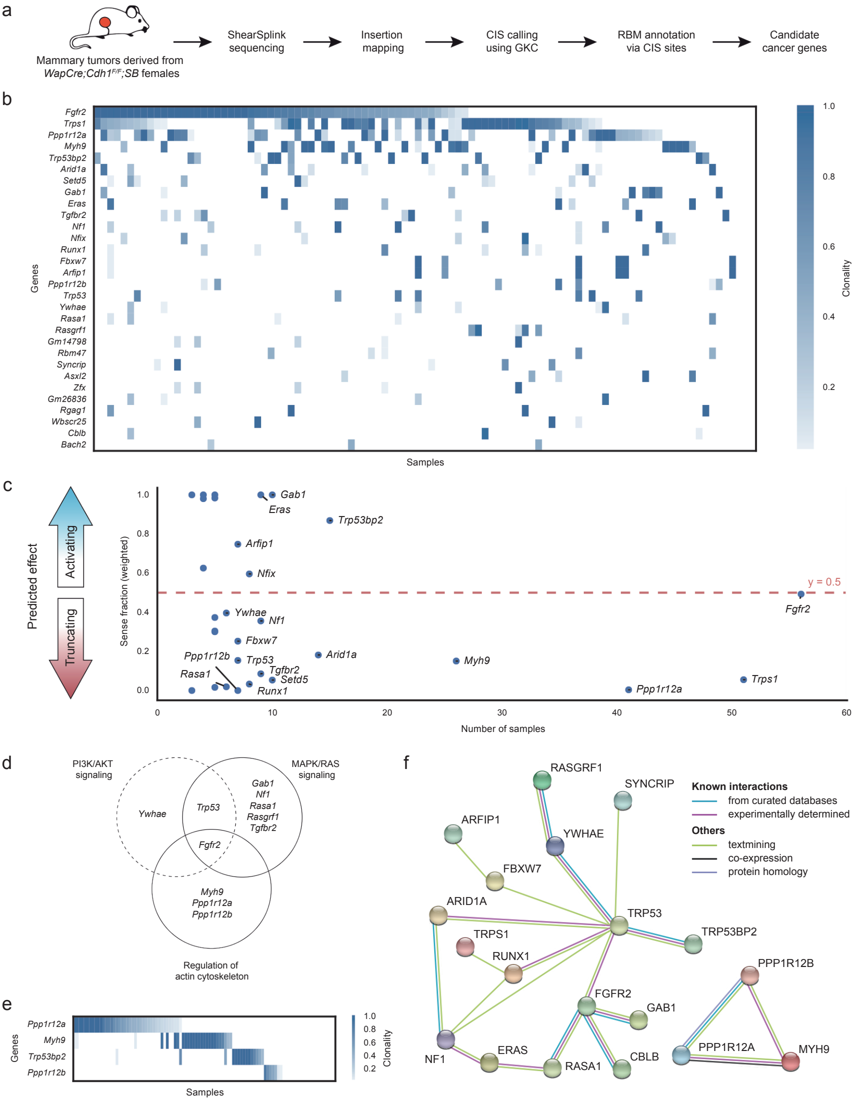


Figure 3 Insertion analysis of tumors from Wap–Cre;Cdh1$^{F/F}$;SB females. (A) Overview of the pipeline used to identify candidate genes. (B) Overview of the insertions in candidate genes across all samples with an ILC morphology (n = 99). The relative clonality of the insertions within each sample is depicted in blue. (C) Orientation bias of the candidate genes, indicated by their fraction of sense insertions. Genes with a strong bias toward sense insertions are expected to be activated, whereas those biased toward antisense insertions are predicted to be inactivated or to yield truncated products. The dashed red line (y = 0.5) indicates an equal ratio of sense and antisense insertions. For clarity, only the main candidates (which occur in six or more samples) are labeled. (D) Venn diagram depicting the candidate genes (according to KEGG, dashed circle) involved in PI3K–AKT signaling, which is known to be associated with hILC, and two significant pathways from the KEGG analysis. (E) Overview of insertions in the four genes that were identified to be significantly mutually exclusive (P < 1 × 10−3) using the DISCOVER algorithm. The relative clonality of the insertions within each sample is depicted in blue. (F) Projection of all candidate genes onto the STRING protein–protein interaction network (version 10). Only connected nodes are shown.

With regard to their associations with subtypes, insertions in *Trps1* were enriched in the combined mILC-1 and mILC-2 subtypes, whereas insertions in *Eras* and *Tgfbr2* were enriched in the mILC-2 and squamous-like subtypes, respectively (one-sided Fisher's exact test with Benjamini–Hochberg correction, FDR < 0.1; Supplementary Figure S6).

SB insertional mutagenesis identifies known ILC drivers

To determine their biological relevance, we compared our candidate genes with known drivers of ILC formation. This analysis showed that the SB screen was able to identify known cancer genes such as *Trp53*, which has been shown to collaborate with E-cadherin loss in the formation of mouse mammary tumors that resemble human pleomorphic ILC (Derksen *et al.,* 2006; Derksen *et al.,* 2011; Ercan *et al.,* 2012). Similarly, the screen identified several genes involved in the PI3K–AKT signaling pathway (for example, *Fgfr2* and *Eras*), which is mutated in approximately 50% of hILC (Ciriello *et al.,* 2015; Michaut *et al.,* 2016; Desmedt *et al.,* 2016). These results demonstrated that our screen identified cancer driver genes and pathways that are known to be involved in hILC.

Candidate genes are biased toward inactivating insertions

To determine how the SB insertions affected expression of the candidate genes, we investigated orientation biases of the SB insertions in each candidate gene. This analysis (Figure 3C) showed that four of the candidates (*Trp53bp2, Gab1, Arfip1* and *Eras*) mainly contained insertions in the sense orientation, which indicated that these genes were likely activated by their insertions (for example, *Gab1*; Supplementary Figure S7A). In support of this hypothesis, *Trp53bp2, Gab1* and *Eras* showed significantly (P < 1 × 10−3) increased expression of exons downstream of the insertion site (Supplementary Figure S7B). In contrast, most of the candidate genes either showed no orientation bias or were biased toward antisense insertions (for example, *Trps1*; Supplementary Figure S7C), and their products were, therefore, likely inactivated or truncated by the insertions. As expected, these genes typically showed substantially decreased mRNA expression of exons downstream of the insertion site (Supplementary Figure S7B).

SB insertion patterns identify oncogenic pathways in mILC

To determine which processes or pathways were affected by the SB insertions, we performed pathway enrichment analysis with all of the candidate genes using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Figure 3D and Table 1). This analysis identified several significantly enriched pathways with an FDR of < 0.1, including the RAS–MAPK signaling pathway and that involved in the regulation of the actin cytoskeleton. Consistent with this, several tumors showed positive

Table 1 Overview of the significantly enriched pathways (hypergeometric test with Benjamini–Hochberg correction, FDR < 0.1) according to KEGG pathway enrichment analysis using all candidate genes

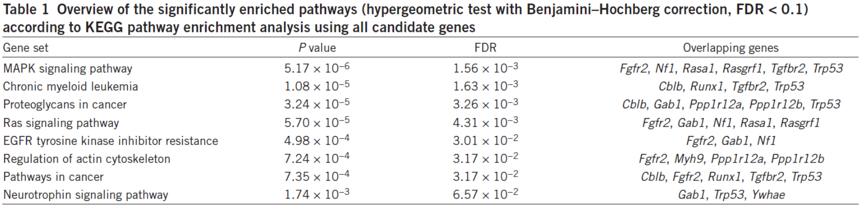


Table 1 Overview of the significantly enriched pathways (hypergeometric test with Benjamini–Hochberg correction, FDR < 0.1) according to KEGG pathway enrichment analysis using all candidate genes

Gene set	*P* value	FDR	Overlapping genes
MAPK signaling pathway	5.17×10^{-6}	1.56×10^{-3}	*Fgfr2*, *Nf1*, *Rasa1*, *Rasgrf1*, *Tgfbr2*, *Trp53*
Chronic myeloid leukemia	1.08×10^{-5}	1.63×10^{-3}	*Cblb*, *Runx1*, *Tgfbr2*, *Trp53*
Proteoglycans in cancer	3.24×10^{-5}	3.26×10^{-3}	*Cblb*, *Gab1*, *Ppp1r12a*, *Ppp1r12b*, *Trp53*
Ras signaling pathway	5.70×10^{-5}	4.31×10^{-3}	*Fgfr2*, *Gab1*, *Nf1*, *Rasa1*, *Rasgrf1*
EGFR tyrosine kinase inhibitor resistance	4.98×10^{-4}	3.01×10^{-2}	*Fgfr2*, *Gab1*, *Nf1*
Regulation of actin cytoskeleton	7.24×10^{-4}	3.17×10^{-2}	*Fgfr2*, *Myh9*, *Ppp1r12a*, *Ppp1r12b*
Pathways in cancer	7.35×10^{-4}	3.17×10^{-2}	*Cblb*, *Fgfr2*, *Runx1*, *Tgfbr2*, *Trp53*
Neurotrophin signaling pathway	1.74×10^{-3}	6.57×10^{-2}	*Gab1*, *Trp53*, *Ywhae*

immunohistochemical staining for phosphorylated ERK1–ERK2, which are downstream effectors of RAS–MAPK signaling (Supplementary Figure S8). In contrast to that seen in hILC, we did not find a significant enrichment for genes that encoded the canonical components of the PI3K–AKT pathway (FDR = 0.44).

To identify further evidence that the insertions may be targeting a common biological process or pathway, we used the DISCOVER (Canisius *et al.,* 2016) algorithm to test for associations of co-occurrence and mutual exclusivity between candidate genes. Although this analysis did not identify any significant co-occurrences, it did identify a subgroup of four genes (*Myh9, Trp53bp2, Ppp1r12a* and *Ppp1r12b*) that showed strong mutual exclusivity (P < 1 × 10−3), suggesting that these genes were likely involved in a common pathway (Figure 3E). This hypothesis was supported by a projection of the candidate genes onto the STRING protein–protein interaction network (Figure 3F), which showed that three of these genes (*Ppp1r12a, Ppp1r12b* and *Myh9*) are in fact known interactors in the STRING network.

Taken together, these analyses identified *Myh9* (which encodes nonmuscle myosin IIa heavy chain 9), *Ppp1r12a* and *Ppp1r12b* (also known as myosin phosphatase-targeting subunit family members *Mypt1* and *Mypt2*, respectively), and *Trp53bp2* (also known as *Aspp2*) as potential drivers of a novel oncogenic pathway in ILC. The mutual exclusivity, combined with the observation that *Ppp1r12a*, *Ppp1r12b* and *Trp53bp2* encode protein phosphatase 1 (PP1) targeting subunits (Grassie *et al.,* 2011; Zhang *et al.,* 2015; Zhang *et al.,* 2015), supports the idea that these genes function in a common pathway. According to the KEGG analysis, this novel pathway may be involved in the regulation of the actin cytoskeleton, suggesting that the disruption of this regulatory process could have a role in the malignant transformation of E-cadherin-deficient mammary epithelial cells.

TP53BP2, PPP1R12B and MYH9 are frequently aberrated in hILC

To establish the human relevance of the identified candidate genes, we assessed their mutational status in human breast cancers from TCGA (Figure 4A and Supplementary Figure S9; Ciriello *et al.,* 2015). This analysis showed that *TP53BP2*, *PPP1R12B* and *MYH9* are commonly aberrated in the 127 hILCs. In particular, *TP53BP2* and *PPP1R12B* are both located within the human chromosome 1q locus, which is frequently gained or amplified in hILC, and in breast cancer in general. In the breast cancer samples in TCGA, expression of these genes was significantly correlated with their copy-number level (Figure 4B-C), indicating that gain or amplification of *TP53BP2* and *PPP1R12B* generally results in increased mRNA expression. In contrast, *MYH9* was mainly affected by truncating or missense mutations and heterozygous copy-number loss, the latter of which was correlated with reduced expression of MYH9 mRNA (Figure 4D), which supported a haploinsufficient tumor suppressive role of *MYH9*. Collectively, these data indicate that three of four mutually exclusive genes are frequently mutated in hILC and that these aberrations result in altered gene expression, supporting their role as potential drivers of hILC.

SB insertions show haploinsufficiency of Myh9 in ILC

SB insertions in *Myh9* were mainly heterozygous and did not show any clustering, indicating that they likely resulted in heterozygous loss of *Myh9* (Figure 5A and Supplementary Figure S10A). To assess the effects of SB insertions on *Myh9* expression, we derived tumor cells from SB-induced tumors with or without insertions in Myh9. PCR amplification of the transposon–*Myh9* junction fragments confirmed the presence of heterozygous *Myh9* insertions in the isolated tumor cells, which coincided with decreased levels of MYH9 protein (Figure 5B and Supplementary Figure S10B). Notably, MYH9 expression was never completely lost, suggesting that it may function as a haploinsufficient tumor suppressor in ILC development. To rule out the possibility of a mixed cell population, heterozygous *Myh9* insertions were also confirmed by PCR in clones that were derived from the tumor cell lines (Supplementary Figure S10C).

SB insertions cause truncation of PP1-targeting subunits

In contrast to *Myh9*, SB insertions in the genes encoding PP1 targeting subunits (*Trp53bp2*, *Ppp1r12a* and *Ppp1r12b*) were strongly clustered, which suggested the expression of truncated transcripts (Figure 5C-E). To test this hypothesis, we visualized the expression of samples with insertions in these genes at the exon level to identify biases in read coverage before and after the insertion sites. This analysis showed a relative increase in expression of the exons 5′ of the SB insertions in *Ppp1r12a* and *Ppp1r12b* and the exons

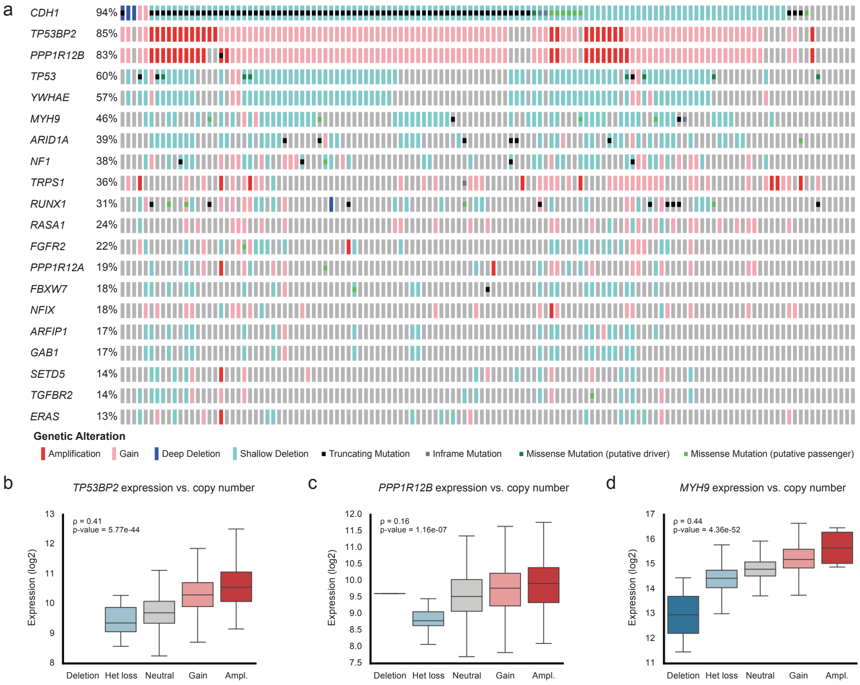


Figure 4 Overview of the candidate genes in hILC. (A) Overview of the mutations and copy-number events in 127 TCGA ILC samples for each of the main candidate genes. Percentages indicate the fraction of tumors with alterations in the respective genes. (B-D) Correlation between the expression of *TP53BP2* (B), *PPP1R12B* (C) and *MYH9* (D) and their respective copy-number levels, using the entire TCGA breast cancer data set (n = 1,068) to ensure sufficient numbers for each copy-number level. Boxes extend from the third (Q3) to the first (Q1) quartile (IQR), with the line at the median; whiskers extend to Q3 + 1.5 × IQR and to Q1 − 1.5 × IQR. Correlation scores (ρ) and P values were calculated using Spearman's rank correlation. Het. loss, heterozygous loss; Ampl., amplification.

3′ of the insertions in *Trp53bp2*, as compared to expression levels of the full-length transcripts. Overexpression of the sequences encoding the truncated PP1 targeting subunits was confirmed by northern blot analysis (Supplementary Figure S11A-C) and by western blotting for PPP1R12A (Supplementary Figure S11D).

Analysis of the predicted proteins showed that the truncated PP1 targeting subunits lacked various regulatory domains but retained their PP1-binding domains (Figure 5F). To test whether the truncated proteins were still able to bind PP1, we performed immunoprecipitation with a Flag-specific antibody followed by liquid chromatography–tandem mass spectrometry (LC-MS/MS) analysis in mouse mammary epithelial HC11 cells expressing a Flag-tagged truncated PPP1R12A protein (encoded by *Ppp1r12a* exons 1–9) or TRP53BP2 protein (encoded by *Trp53bp2* exons 13–18). This showed that both

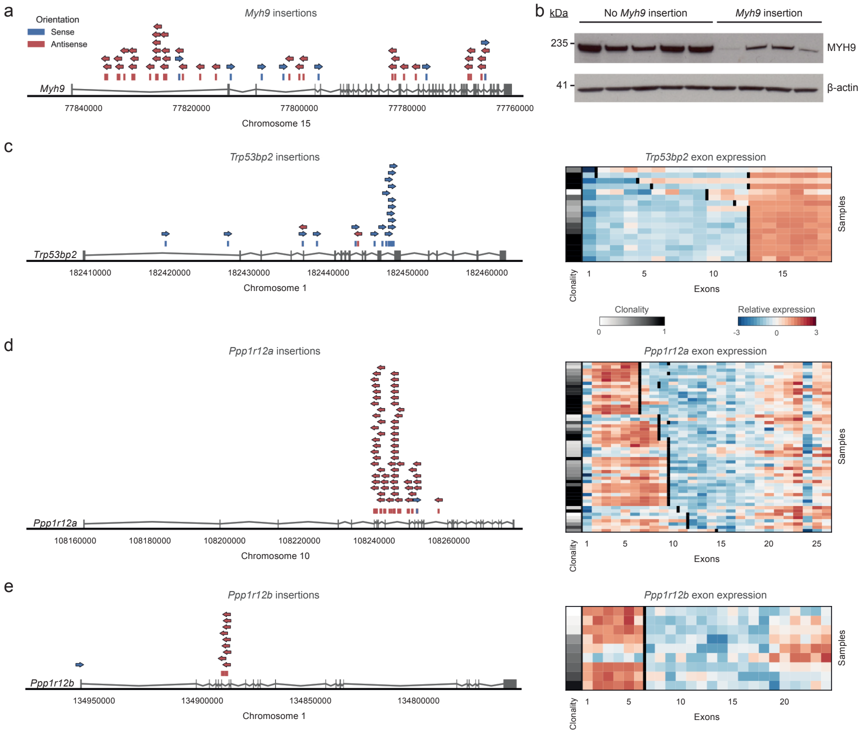


Figure 5 Overview of the insertions and corresponding gene expression of the mutually exclusive genes. (A) Visualization of SB insertions (arrows) in *Myh9* (n = 33 tumors). Bars represent the exact genomic locations of the insertions. (B) Immunoblot for MYH9 levels in SB-induced tumor-derived cells without (n = 5) or with (n = 4) insertions in *Myh9*. β-actin was used as a loading control. (c–e) Left, schematic representation of insertions in *Trp53bp2* (C), *Ppp1r12a* (D) and *Ppp1r12b* (E) (from 17, 52 and 9 tumors, respectively) showing strong clustering of insertions within the genes. Right, heat maps of the exon-level expression of the indicated genes in samples with an insertion, using a z-score measure to normalize for overall expression differences between samples. The positions of the insertions in each sample are indicated by black lines. Red indicates relatively increased expression of an exon; blue signifies relatively decreased expression. Increased expression toward the end of *Ppp1r12a* and *Ppp1r12b* is due to the use poly(A) tail selection in the RNA sequencing analysis, which has well-documented 3′ bias.

truncated proteins were still able to bind specific PP1 isoforms, with PPP1R12A able to bind both PPP1CA and PPP1CB, and TRP53BP2 preferentially able to bind PPP1CA (Figure 5G-H and Supplementary Figure S11E). Taken together, these data suggest that truncated PPP1R12A and TRP53BP2 are able to bind PP1 and that the loss of other regulatory domains could affect their function.

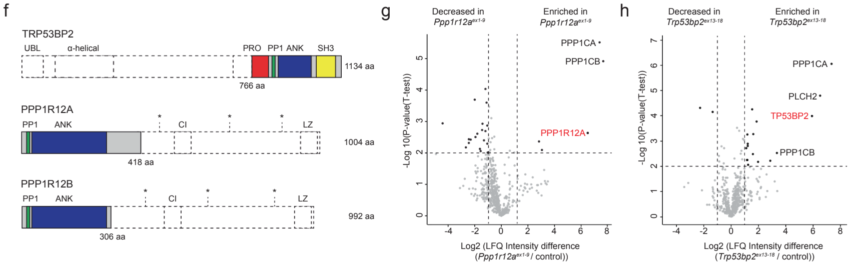


Figure 5 **Continued.** (F) Overview of the binding domains of mouse TRP53BP2, PPP1R12A and PPP1R12B, based on previously published work (Grassie *et al.,* 2011; Rotem *et al.,* 2008). Colors indicate the predicted proteins from the truncated genes. UBL, ubiquitin-like domain; PRO, proline-rich domain; PP1, PP1-binding domain; ANK, ankyrin repeats; SH3, Src homology 3 domain; CI, central insert; LZ, leucine zipper; aa, amino acid. Asterisks indicate inhibitory or regulatory phosphorylation sites. (G-H) Volcano plots showing protein interactors of truncated PPP1R12A (G) and TRP53BP2 (H) in HC11 cells that were transduced with pBABE-*Ppp1r12a*$^{ex1–9}$ or pBABE-*Trp53bp2*$^{ex13–18}$, respectively, as compared to that in cells that were transduced with the pBABE empty vector control. P values were calculated using a permutation-based FDR-corrected t-test. Proteins were considered interactors if P < 0.01 and log2(abundance difference) > 1. LFQ, label-free quantification.

Candidate ILC drivers enhance survival of Cdh1$^{Δ/Δ}$ mouse mammary epithelial cells

To study the consequences of E-cadherin loss in primary mouse mammary epithelial cells (MMECs), we used *Cdh1*$^{F/F}$*;Rosa26*$^{ACTB\text{-}tdTomato\text{-}EGFP}$ MMECs, which contain, in addition to floxed *Cdh1* alleles, a *Rosa26*$^{ACTB\text{-}tdTomato\text{-}EGFP}$ reporter allele (termed *mT/mG*) that expresses membrane-targeted mTomato before, and mGFP after, Cre switching (Boelens *et al.,* 2016; Muzumdar *et al.,* 2007). Transduction of *Cdh1*$^{F/F}$*;mT/mG* MMECs with a Cre-encoding adenovirus (AdCre) resulted in reduced proliferation and clonogenic survival, indicating that E-cadherin loss alone is not sufficient for cellular transformation *in vitro* (Figure 6A-C). To test the effects of truncated PPP1R12A and TRP53BP2 in E-cadherin-deficient MMECs, we transduced *Cdh1*$^{F/F}$*;mT/mG* MMECs with lentiviruses encoding Ppp1r12a$^{ex1–9}$ or Trp53bp2$^{ex13–18}$ (Figure 6D). Simultaneous transduction of these cells with AdCre showed that expression of truncated TRP53BP2 or PPP1R12A decreased cell death and increased clonogenic survival of E-cadherin-deficient MMECs, without affecting canonical PI3K–AKT signaling (Figure 6A-C and Supplementary Figure S12A-C). Similar results were obtained after reduction of MYH9 levels by short hairpin RNA (shRNA)-mediated knockdown of *Myh9* expression (Figure 6E-H and Supplementary Figure S12D).

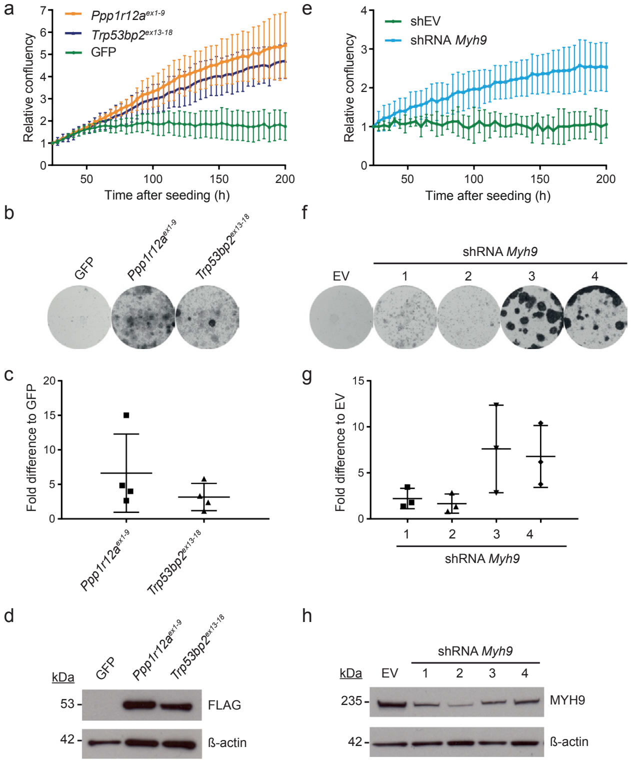


Figure 6 Limited proliferation and survival of AdCre-transduced $Cdh1^{F/F}$*;mT/mG* mouse mammary epithelial cells (MMECs) that were rescued by expression of truncated PPP1R12A and TRP53BP2 or by dosage reduction of MYH9. (A) Cell survival analysis of AdCre-transduced $Cdh1^{F/F}$*;mT/mG* MMECs that were also transduced with lentiviruses encoding $Ppp1r12a^{ex1-9}$ or $Trp53bp2^{ex13-18}$, quantified using real-time IncuCyte imaging for 200 h. AdCre-transduced $Cdh1^{F/F}$*;mT/mG* MMECs also transduced with a GFP-expressing lentivirus (Lenti-GFP) is shown as control. Data are mean ± s.d. of four independent experiments. (B-C) Representative images (B) and quantification (C) of clonogenic assays (14 d after seeding the cells) of AdCre-transduced $Cdh1^{F/F}$*;mT/mG* MMECs that were also transduced with lentiviruses expressing the indicated constructs. Fold difference is relative to the GFP control. Data are mean ± s.d. of four independent experiments. Scale bar, 1 cm. (D) Representative immunoblot (n = 3) for expression of Flag-tagged and truncated PPP1R12A and TRP53BP2 in AdCre-transduced $Cdh1^{F/F}$*;mT/mG* MMECs 7 d after transduction. β-actin was used as a loading control. (E) Cell survival analysis of AdCre-transduced $Cdh1^{F/F}$*;mT/mG* MMECs with simultaneous shRNA-mediated knockdown of *Myh9* expression, as quantified by real-time IncuCyte imaging for 200 h. Average survival of AdCre-transduced $Cdh1^{F/F}$*;mT/mG* MMECs of all shRNAs is shown. Independent survival curves are depicted in Supplementary Figure S12D. Data are mean ± s.d. of three independent experiments. EV, empty vector. (F-G) Representative images (F) and quantification (G) of clonogenic assays of AdCre-transduced $Cdh1^{F/F}$*;mT/mG* MMECs with simultaneous shRNA-mediated knockdown of *Myh9* expression 14 d after seeding the cells. Fold difference is relative to the value observed in the EV control. Data are mean ± s.d. of three independent experiments. Scale bar, 1 cm. (H) Representative immunoblot (n = 3) for the expression of MYH9 in AdCre-transduced $Cdh1^{F/F}$*;mT/mG* MMECs that also had simultaneous shRNA-mediated knockdown of *Myh9* expression (7 d after transduction). β-actin was used as a loading control.

Previous work has shown that MYH9 is involved in regulating post-transcriptional stabilization of the tumor suppressor p53, suggesting that an altered p53 response in MYH9-deficient keratinocytes induces squamous cell carcinoma (SSC) in *Tgfbr2* conditional-knockout mice (Schramek *et al.,* 2014). In contrast, we and others (Conti *et al.,* 2015) have observed an intact p53 response after DNA damage in cells with reduced MYH9 levels (Supplementary Figure S12E-G), suggesting that an alternative mechanism of cellular transformation may be involved. Taken together, these data show that dosage reduction of MYH9 or overexpression of truncated PP1 targeting subunits enhances survival of E-cadherin-deficient MMECs and indicate deregulation of conventional actin-related processes rather than loss of nuclear p53 retention or activation of canonical PI3K–AKT signaling as the underlying mechanism.

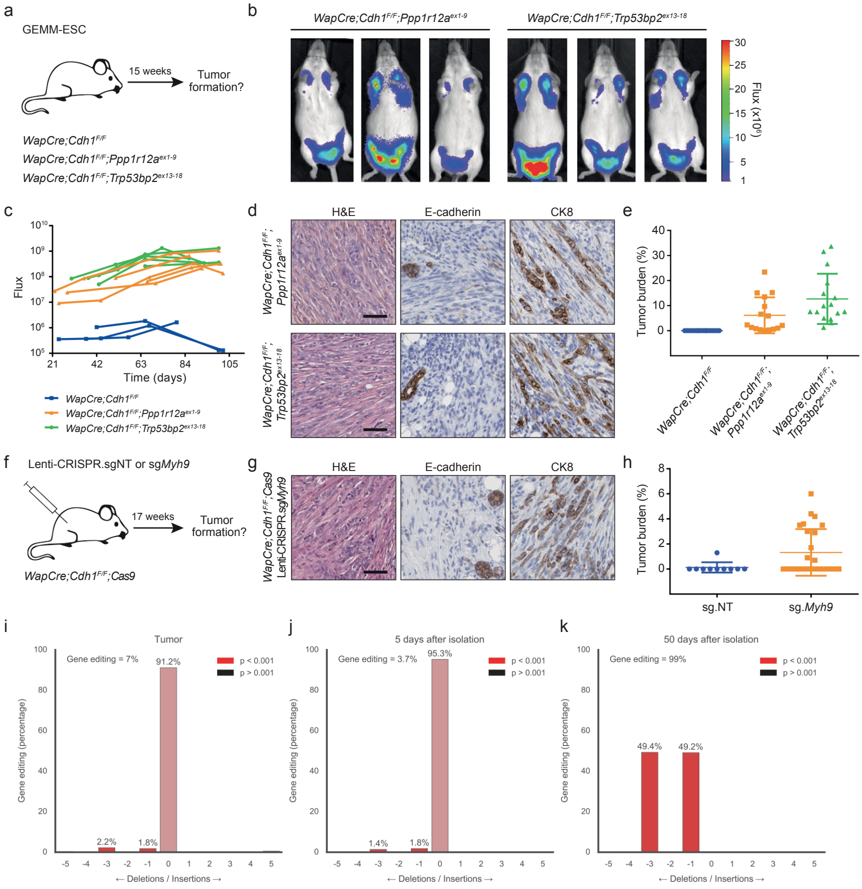
Truncated PPP1R12A and TRP53BP2 induce ILC formation

Next we investigated whether expression of Ppp1r12a$^{ex1–9}$ and Trp53bp2$^{ex13–18}$ in *Wap–Cre;Cdh1*$^{F/F}$ mice could induce mammary tumor formation *in vivo*. To this end, we introduced *invCAG-Ppp1r12a*$^{ex1–9}$*-IRES-Luc* and invCAG-Trp53bp2$^{ex13–18}$-IRES-Luc alleles for Cre-inducible expression of firefly luciferase and *Ppp1r12a* exons 1–9 or *Trp53bp2* exons 13–18, respectively, into the *Col1a1* locus of *Wap–Cre;Cdh1*$^{F/F}$ embryonic stem cells (ESCs) and subsequently generated chimeric mice by blastocyst injection of the modified ESCs (Supplementary Figure S13A; Huijbers *et al.,* 2015). Male chimeras were mated with *Cdh1*$^{F/F}$ females to generate *Wap–Cre;Cdh1*$^{F/F}$*;Col1a1*$^{invCAG-Ppp1r12a-ex1-9-IRES-Luc/+}$ (hereafter referred to as *Wap–Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1–9}$) and *Wap–Cre;Cdh1*$^{F/F}$*;Col1a1*$^{invCAG-Trp53bp2-ex13-18-IRES-Luc/+}$ (hereafter referred to as *Wap–Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13–18}$) mice, which showed mammary-specific loss of E-cadherin expression and concomitant expression of luciferase and truncated PPP1R12A or TRP53BP2, respectively (Figure 7A).

Wap–Cre;Cdh1$^{F/F}$*;Ppp1r12a*$^{ex1–9}$ and *Wap–Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13–18}$ female mice showed mammary-specific bioluminescence signals that increased over time, which indicated the development of mammary tumors (Figure 7B-C). Analysis of the mammary glands from 15-week-old *Wap–Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1–9}$ and *Wap–Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13–18}$ females showed multifocal tumor formation (in 27/29 and 30/30 of the analyzed glands, respectively), whereas no tumors were detected in mammary glands from age-matched *Wap–Cre;Cdh1*$^{F/F}$ females (Figure 7D-E). All of the tumors showed recombination of the conditional alleles (Supplementary Figure S13B), which confirmed the inactivation of E-cadherin expression and the activation of the expression of the truncated PP1 targeting subunits. Morphologically, most of the tumors were CDH1^{-}CK8^{+} and strongly invaded the surrounding tissue (Supplementary Figure S13C). Taken together, these data confirm that loss of expression of E-cadherin and concomitant expression of truncated PPP1R12A and TRP53BP2 results in the development of mILCs that closely resemble hILC.

Myh9 haploinsufficiency induces ILC formation

To determine whether loss of *Myh9* also results in tumor formation *in vivo*, we performed in situ gene editing of *Myh9* in mammary epithelial cells using CRISPR–Cas9 genome editing technology. We intraductally injected CRISPR–Cas9-encoding lentiviruses that targeted the second exon of *Myh9* (Lenti-CRISPR-sg*Myh9*) in *Wap–Cre;Cdh1*$^{F/F}$*;Col1a1*$^{invCAG-Cas9-IRES-Luc/+}$ (*Wap–Cre;Cdh1*$^{F/F}$*;Cas9*) female mice (Figure 7F). Analysis of the mammary glands 17 weeks after injection showed that 11 of 26 successfully injected glands contained mammary tumors that were CDH1^{-}CK8^{+} and closely resembled hILC (Figure 7G-H and Supplementary Figure S13D).



Analysis of *Myh9* target modification in these tumors showed equal levels of in-frame and out-of-frame alterations (Figure 7I), suggesting that functional inactivation of only one *Myh9* allele had occurred, resulting in hemizygous expression of *Myh9*. To substantiate this possibility, tumor cells from a successfully injected mammary gland were isolated and cultured *in vitro* to remove stromal contaminants. Indeed, DNA isolated from these tumor cells showed enrichment for in-frame and out-of-frame genetic alterations in *Myh9*, supporting the idea of heterozygous loss of one functional allele (Figure 7J-K). Collectively, these data provide additional support for *Myh9* as a haploinsufficient tumor suppressor in ILC development.

Figure 7 Validation and characterization of candidate genes in genetically engineered mice (GEMM). (A) Overview of the genetically engineered mice. ESC, embryonic stem cell. (B) Representative images of *in vivo* bioluminescence imaging of luciferase expression in *Wap–Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1–9}$ (n = 3) and *Wap–Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13–18}$ (n = 3) females (at 100 d). Scale bar, 1 cm. (C) Quantification of bioluminescence imaging of luciferase expression over time in *Wap–Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1–9}$ (n = 5) and *Wap–Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13–18}$ (n = 5) females. *Wap–Cre;Cdh1*$^{F/F}$ females (n = 3) were used to show background luminescence. (D) Representative images for H&E staining (left) and for expression of E-cadherin (middle) and CK8 (right) by immunohistochemistry in tumors of 15-week-old *Wap–Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1–9}$ (n = 27 tumors) and *Wap–Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13–18}$ (n = 30 tumors) females. Scale bars, 50 µm. (E) Tumor burden in mammary glands of 15-week-old *Wap–Cre;Cdh1*$^{F/F}$ (n = 34 glands), *Wap–Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1–9}$ (n = 29 glands) and *Wap–Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13–18}$ (n = 30 glands) females. Data are mean ± s.d. (F) Overview of intraductal injections performed in *Wap–Cre;Cdh1*$^{F/F}$*;Cas9* females with high-titer lentivirus containing a vector encoding Cas9, GFP and either a non-targeting (NT) single-guide RNA (sgRNA) (Lenti-CRISPR-sgNT) or a sgRNA targeting the second exon of *Myh9* (Lenti-CRISPR-sg*Myh9*). (G) Representative images for H&E staining (left) and for expression of E-cadherin (middle) and CK8 (right) in tumors (n = 11 tumors) of *Wap–Cre;Cdh1*$^{F/F}$*;Cas9* females at 17 weeks after injection of Lenti-CRISPR-sg*Myh9*. Scale bar, 50 µm. (H) Tumor burden of *Wap–Cre;Cdh1*$^{F/F}$*;Cas9* females 17 weeks after injection of Lenti-CRISPR-sgNT (n = 10 glands) or Lenti-CRISPR-sg*Myh9* (n = 26 glands). Data are mean ± s.d. (I) Representative spectrum of insertions and deletions (indels) of targeted *Myh9* alleles depicting the CRISPR–Cas9-induced editing efficacy in the tumors (n = 3) as quantified with the TIDE algorithm (Brinkman *et al.,* 2014). Fraction of unmodified alleles are depicted in pink; red ($P < 0.001$) and black ($P > 0.001$) bars represent fractions of modified alleles. (J-K) TIDE analysis of tumor cells cultured *in vitro* 5 d (J) and 50 d (K) after isolation from a single tumor-bearing gland injected with Lenti-CRISPR-sg*Myh9*.

Discussion

Here we performed an SB-based insertional mutagenesis screen in mice to identify novel genes and signaling pathways that are involved in ILC formation. SB-mediated mutagenesis in *Wap–Cre*;*Cdh1*$^{F/F}$;*SB* female mice resulted in the development of multiple independent mammary tumors, which showed similarities to hILC in terms of morphology and gene expression. CIS analysis identified multiple candidate cancer genes, several of which have previously been implicated in hILC (such as *Arid1a*, *Nf1* and *Runx1*) and have also been shown to induce ILC formation in mice (such as *Trp53*; Ciriello *et al.,* 2015; Ercan *et al.,* 2012; Michaut *et al.,* 2016; Desmedt *et al.,* 2016), demonstrating the relevance of the genes we identified.

In contrast to previous analyses of human data sets, we identified only a few components of the canonical PI3K–AKT signaling pathway. Nevertheless, insertions in *Fgfr2* were observed in over half of the tumors, which could result in PI3K–AKT pathway activation in the same manner as previously identified *FGFR1* amplifications in hILC (Reis-Filho *et al.,* 2006; Xian *et al.,* 2009; Turner *et al.,* 2009). Notably, mutations that can be linked to activated PI3K–AKT signaling are found in only 50% of hILCs (Ciriello *et al.,* 2015; Michaut *et al.,* 2016; Desmedt *et al.,* 2016), indicating that additional pathways are involved in ILC formation. Consistent with this, our KEGG analysis of the candidate genes in the SB-induced mILCs indicated a potential role for RAS–MAPK signaling, which was further supported by the presence of *KRAS*, *NF1*, *MAP2K4*, *MAP3K1* and *ERBB2* mutations in hILC (Ciriello *et al.,* 2015; Michaut *et al.,* 2016; Desmedt *et al.,* 2016; Ross *et al.,* 2013). Additionally, we provided strong evidence that genes involved in the regulation of the actin cytoskeleton (*Ppp1r12a*, *Ppp1r12b*, *Trp53bp2* and *Myh9*) effectively collaborated with E-cadherin loss in mILC formation and were frequently mutated in hILCs. We therefore conclude that this process constitutes a novel oncogenic pathway in ILC development.

An important advantage of cancer gene discovery using insertional mutagenesis screens in mice is that these approaches can identify driver mutations that are not readily apparent in humans. For example, *MYH9* has not been identified as a tumor suppressor in hILC because it is rarely mutated and mainly characterized by shallow deletions. In contrast, our SB screen identified *Myh9* as a haploinsufficient tumor suppressor in ILC, as it was mainly affected by heterozygous inactivating insertions that resulted in dosage reduction of MYH9. Haploinsufficiency of *Myh9* has also been observed for platelet development, resulting in macrothrombocytopenia in human patients with heterozygous germline mutations in *MYH9* (Vicente-Manzanares *et al.,* 2009; Pecci *et al.,* 2005).

A second example involves the two candidate cancer genes *Trp53bp2* and *Ppp1r12b*, whose orthologs in humans are both located on the 1q locus. Although this locus has already been described to be frequently amplified or gained in human breast cancer (Ciriello *et al.,* 2016; Stange *et al.,* 2006), the size of the amplicon makes it difficult to identify the relevant driver gene(s). Notably, der(1;16)(q10;p10) unbalanced translocations, which result in a chromosome 1q gain and chromosome 16q loss, are frequently observed in ILC (Flagiello *et al.,* 1998). Although chromosome 16q loss is associated with LOH of *CDH1*, candidate driver genes on chromosome 1q have remained elusive. Our screen pinpoints *TP53BP2* and *PPP1R12B* as potential drivers on this locus. Furthermore, we identified these genes as part of a mutually exclusive group of four genes (*Myh9*, *Ppp1r12a*, *Ppp1r12b* and *Trp53bp2*), indicating that these genes are targeting the same biological process. Three of these genes (*Myh9*, *Ppp1r12a* and *Ppp1r12b*) are involved in the regulation of the actin cytoskeleton, which implicates *Trp53bp2* as an additional player in this process.

Finally, three of the four genes in our mutually exclusive subgroup (*Ppp1r12a*, *Ppp1r12b* and *Trp53bp2*) encode targeting subunits of PP1. Our data indicate that strongly clustered SB insertions in these genes may be targeting specific domains, thereby affecting PP1 functionality. Consistent with this, we observed that the truncated PPP1R12A and TRP53BP2 proteins were still able to bind PP1 but lacked a number of regulatory and/or inhibitory domains. Because these mutants rescued cell survival of E-cadherin-deficient cells *in vitro* and collaborated with E-cadherin loss in mammary tumorigenesis, our combined data suggest that rewiring of PP1 signaling by dosage reduction of nonmuscle myosin IIa (MYH9) or by expression of truncated PPP1R12A or TRP53BP2 promotes malignant transformation of E-cadherin-deficient mammary epithelial cells.

Although *in vivo* transposon mutagenesis is a powerful tool for cancer gene discovery, there are a number of limitations of our screen. First, insertional mutagenesis does not capture the full spectrum of hILC driver mutations (for example, *PIK3CA* point mutations). Additionally, the bias toward gene inactivation or protein truncation indicates that our screen might underestimate the number of oncogenes that are involved in hILC. Complementary chemical or genetic mutagenesis strategies may be used to identify additional ILC drivers. Second, *Trp53bp2* and *Ppp1r12b* are affected by mutually exclusive truncations in mILC, whereas both genes are co-amplified in human breast cancers. It is, however, possible that these individual truncations may have a similar effect on cellular transformation as the combined amplification of both genes. Moreover, human breast cancers have been shown to express an N-terminally truncated TP53BP2 isoform (ΔN-ASPP2) that is similar to the truncated TRP53BP2 observed in mILCs (van Hook *et al.,* 2017), warranting further investigation in future studies.

In summary, we show that our SB screen in mammary-specific E-cadherin-deficient mice uncovers previously unidentified cancer genes whose orthologs are frequently altered in hILC. This emphasizes the utility of *in vivo* insertional mutagenesis screens in mice as a powerful genetic tool for unraveling biological processes underlying human cancer development. Finally, our results identify a novel oncogenic pathway involved in ILC formation, providing new opportunities for the development of targeted therapies for hILC.

Acknowledgments

We are grateful to E. Wientjens, J. Houthuijzen, M. van Miltenburg, O. Bleijerveld, B. van Gerwen, B. Siteur, R. de Korte-Grimmerink, N. Proost, U. Boon, C. Brambillasca, R. Mezzadra, S. Cornelissen and L. Braaf for providing technical suggestions and/or help with the experiments, S. Canisius for critical reading of the manuscript and N. Hynes (Friederich Miescher Institute, Basel) for the HC11 cells. We thank the NKI animal facility, the animal pathology facility, the mouse clinic transgenic and intervention unit, the core facility molecular pathology and biobanking (CFMPB), and the genomics core facility for their expert technical support. Financial support was provided by the Netherlands Organization for Scientific Research (NWO: Cancer Genomics Netherlands (CGCNL), Cancer Systems Biology Center (CSBC), the Netherlands Genomics Initiative (NGI) grants Zenith 93512009 (J.J.), VENI 016156012 (M.N.) and VICI 91814643 (J.J.), the EU Seventh Framework Program (EurocanPlatform project 260791 (J.J.) and Infrafrontier-I3 project 312325 (J.J.)), the European Research Council (ERC Synergy project CombatCancer) (J.J.), and a National Roadmap grant for Large-Scale Research Facilities from NWO (J.J.).

Author contributions

S.M.K., K.S., S.A., E.S., A.P.D. and E.v.d.B. performed laboratory experiments; J.R.d.R. identified the insertion sites and CISs, analyzed the RNA sequencing data sets and performed the other bioinformatic analyses; C.K. and J.J.t.H. assisted in the initial bioinformatic analysis for the identification of insertion sites and CISs; S.K. and J.W. assessed the histology of mouse tumors and quantified the immunohistochemically stained images; D.J.A. was responsible for sequencing the transposon insertions; M.J.K. initiated the breeding of the mouse lines; M.N., L.F.A.W. and J.J. supervised the experiments; and J.R.d.R., S.M.K., L.F.A.W. and J.J. wrote the manuscript.

References

Boelens, M.C. et al. PTEN loss in E-cadherin-deficient mouse mammary epithelial cells rescues apoptosis and results in development of classical invasive lobular carcinoma. Cell Rep. 16, 2087–2101 (2016).

Borst, M.J. & Ingold, J.A. Metastatic patterns of invasive lobular versus invasive ductal carcinoma of the breast. Surgery 114, 637–641 (1993).

Boussadia, O., Kutsch, S., Hierholzer, A., Delmas, V. & Kemler, R. E-cadherin is a survival factor for the lactating mouse mammary gland. Mech. Dev. 115, 53–62 (2002).

Brinkman, E.K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence-trace decomposition. Nucleic Acids Res. 42, e168 (2014).

Bureau, A. et al. Identifying SNPs predictive of phenotype using random forests. Genet. Epidemiol. 28, 171–182 (2005).

Buttitta, F. et al. PIK3CA mutation and histological type in breast carcinoma: high frequency of mutations in lobular carcinoma. J. Pathol. 208, 350–355 (2006).

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. Nature 490, 61–70 (2012).

Canisius, S., Martens, J.W.M. & Wessels, L.F.A. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. Genome Biol. 17, 261 (2016).

Carroll, J.S. et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell 122, 33–43 (2005).

Christgen, M. et al. Oncogenic PIK3CA mutations in lobular breast cancer progression. Genes Chromosom. Cancer 52, 69–80 (2013).

Ciriello, G. et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 163, 506–519 (2015).

Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J. & Largaespada, D.A. Cancer gene discovery in solid tumors using transposon-based somatic mutagenesis in the mouse. Nature 436, 272–276 (2005).

Conti, M.A. et al. Conditional deletion of nonmuscle myosin II-A in mouse tongue epithelium results in squamous cell carcinoma. Sci. Rep. 5, 14068 (2015).

de Jong, J. et al. Computational identification of insertional mutagenesis targets for cancer gene discovery. Nucleic Acids Res. 39, e105 (2011).

de Ridder, J., Uren, A., Kool, J., Reinders, M. & Wessels, L. Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. PLOS Comput. Biol. 2, e166 (2006).

Derksen, P.W.B. et al. Somatic inactivation of E-cadherin and p53 in mice leads to metastatic lobular mammary carcinoma through induction of anoikis resistance and angiogenesis. Cancer Cell 10, 437–449 (2006).

Derksen, P.W.B. et al. Mammary-specific inactivation of E-cadherin and p53 impairs functional gland development and leads to pleomorphic invasive lobular carcinoma in mice. Dis. Model. Mech. 4, 347–358 (2011).

Desmedt, C. et al. Genomic characterization of primary invasive lobular breast cancer. J. Clin. Oncol. 34, 1872–1881 (2016).

Ercan, C. et al. p53 mutations in classic and pleomorphic invasive lobular carcinoma of the breast. Cell. Oncol. (Dordr.) 35, 111–118 (2012).

Flagiello, D. et al. Highly recurrent der(1;16)(q10;p10) and other 16q arm alterations in lobular breast cancer. Genes Chromosom. Cancer 23, 300–306 (1998).

Hurtado, A., Holmes, K.A., Ross-Innes, C.S., Schmidt, D. & Carroll, J.S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat. Genet. 43, 27–33 (2011).

Goldhirsch, A. et al. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann. Oncol. 22, 1736–1747 (2011).

Grassie, M.E., Moffat, L.D., Walsh, M.P. & MacDonald, J.A. The myosin phosphatase targeting protein (MYPT) family: a regulated mechanism for achieving substrate specificity of the catalytic subunit of protein phosphatase type 1⍰. Arch. Biochem. Biophys. 510, 147–159 (2011).

Huijbers, I.J. et al. Using the GEMM–ESC strategy to study gene function in mouse models. Nat. Protoc. 10, 1755–1785 (2015).

Koudijs, M.J. et al. High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors. Genome Res. 21, 2181–2189 (2011).

Liu, X. et al. Somatic loss of BRCA1 and p53 in mice induces mammary tumors with features of human BRCA1-mutated basal-like breast cancer. Proc. Natl. Acad. Sci. USA 104, 12111–12116 (2007).

March, H.N. et al. Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. Nat. Genet. 43, 1202–1209 (2011).

Martinez, V. & Azzopardi, J.G. Invasive lobular carcinoma of the breast: incidence and variants. Histopathology 3, 467–488 (1979).

Michaut, M. et al. Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. Sci. Rep. 6, 18517 (2016).

Moll, R., Mitze, M., Frixen, U.H. & Birchmeier, W. Differential loss of E-cadherin expression in infiltrating ductal and lobular breast carcinomas. Am. J. Pathol. 143, 1731–1742 (1993).

Muzumdar, M.D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. Genesis 45, 593–605 (2007).

Niessen, C.M. & Gottardi, C.J. Molecular components of the adherens junction. Biochim. Biophys. Acta Biomembr. 1778, 562–571 (2008).

Parker, J.S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. 27, 1160–1167 (2009).

Pecci, A. et al. Pathogenetic mechanisms of hematological abnormalities of patients with MYH9 mutations. Hum. Mol. Genet. 14, 3169–3178 (2005).

Rakha, E.A. et al. Clinical and biological significance of E-cadherin protein expression in invasive lobular carcinoma of the breast. Am. J. Surg. Pathol. 34, 1472–1479 (2010).

Reis-Filho, J.S. et al. FGFR1 emerges as a potential therapeutic target for lobular breast carcinomas. Clin. Cancer Res. 12, 6652–6662 (2006).

Ross, J.S. et al. Relapsed classic E-cadherin (CDH1)-mutated invasive lobular breast cancer shows a high frequency of HER2 (ERBB2) gene mutations. Clin. Cancer Res. 19, 2668–2676 (2013).

Rotem, S. et al. The structure and interactions of the proline-rich domain of ASPP2. J. Biol. Chem. 283, 18990–18999 (2008).

Schramek, D. et al. Direct in vivo RNAi screen unveils myosin IIa as a tumor suppressor of squamous cell carcinomas. Science 343, 309–313 (2014).

Simpson, P.T. et al. Molecular profiling pleomorphic lobular carcinomas of the breast: evidence for a common molecular genetic pathway with classic lobular carcinomas. J. Pathol. 215, 231–244 (2008).

Stange, D.E. et al. High-resolution genomic profiling reveals association of chromosomal aberrations on 1q and 16p with histologic and genetic subgroups of invasive breast cancer. Clin. Cancer Res. 12, 345–352 (2006).

Turner, N. & Grose, R. Fibroblast growth factor signaling: from development to cancer. Nat. Rev. Cancer 10, 116–129 (2010).

Van Hook, K. et al. ⍰N-ASPP2, a novel isoform of the ASPP2 tumor suppressor, promotes cellular survival. Biochem. Biophys. Res. Commun. 482, 1271–1277 (2017).

Vicente-Manzanares, M., Ma, X., Adelstein, R.S. & Horwitz, A.R. Nonmuscle myosin II takes center stage in cell adhesion and migration. Nat. Rev. Mol. Cell Biol. 10, 778–790 (2009).

Vos, C.B. et al. E-cadherin inactivation in lobular carcinoma in situ of the breast: an early event in tumorigenesis. Br. J. Cancer 76, 1131–1133 (1997).

Wong, H. et al. Lobular breast cancers lack the inverse relationship between ER/PR status and cell growth rate characteristic of ductal cancers in two independent patient cohorts: implications for tumor biology and adjuvant therapy. BMC Cancer 14, 826 (2014).

Xian, W. et al. Fibroblast growth factor receptor 1–transformed mammary epithelial cells are dependent on RSK activity for growth and survival. Cancer Res. 69, 2244–2251 (2009).

Zhang, P. et al. ASPP1/2–PP1 complexes are required for chromosome segregation and kinetochore–microtubule attachments. Oncotarget 6, 41550–41565 (2015).

Zhang, Y. et al. The tumor suppressor proteins ASPP1 and ASPP2 interact with C-Nap1 and regulate centrosome linker reassembly. Biochem. Biophys. Res. Commun. 458, 494–500 (2015).

Materials and Methods

Mouse models

Generation of mice. The generation of *Wap–Cre*;*Cdh1*$^{F/F}$ mice has previously been described by Derksen *et al*., 2011. We used the transgenic SB transposon concatemer (*T2/Onc*) mouse lines 68 and 76 with the transposon donor locus located on chromosomes 15 and 1, respectively (Collier *et al.,* 2005). The mice containing the conditional HSB5 variant of the SB transposase in the *Rosa26* locus has been described by March *et al*., 2011. *Wap–Cre*;*Cdh1*$^{F/F}$ mice were crossed with *T2/Onc* and *Rosa26*^{Lox66SBLox71} mice to generate *Wap–Cre*;*Cdh1*$^{F/F}$;*T2/Onc*;*Rosa26*$^{Lox66SBLox71/+}$ (*Wap–Cre*;*Cdh1*$^{F/F}$;*SB)* mice. All *Wap–Cre*;*Cdh1*$^{F/F}$;*SB* mice were from a mixed genetic background (C57BL/6J and FVB/N). The experimental cohort was monitored, and mice were sacrificed (mammary tumor-specific survival) when the (total) mammary tumor burden reached a size of ∼1,500 mm3 (tumor volume: length × width2 × 0.5) or the mice suffered from clinical signs of distress caused by the mammary tumor burden or metastatic disease (such as respiratory distress, ascites, distended abdomen, rapid weight loss and severe anemia). Lungs, heart, liver, spleen, intestines, mesenterium, kidneys, mammary glands, pancreas and tumor-draining lymph nodes were collected and analyzed microscopically for the presence of metastatic foci.

The *Ppp1r12a*$^{ex1–9}$ sequence (NM_027892.2) was isolated from the cDNA of mouse tumor tissue using two subsequent Phusion Flash High-Fidelity (ThermoFisher Scientific) PCRs and subsequently cloned in a zero TOPO blunt vector (ThermoFisher Scientific). The *Trp53bp2*$^{ex13–18}$ sequence (NM_173378.2) was commercially synthesized by Genscript, and FseI-PmeI overhangs were added using Phusion Flash High-Fidelity DNA Polymerase. The cDNAs were sequence-verified and inserted as FseI-PmeI fragments into the *Frt-invCag-IRES-Luc* vector, resulting in *Frt-invCag-Ppp1r12a*$^{ex1–9}$*-IRES-Luc* and *Frt-invCag-Trp53bp2*$^{ex13–18}$*-IRES-Luc*, respectively. Flp-mediated integration of the shuttle vectors in *Wap–Cre*;*Cdh1*$^{F/F}$;*Col1a1*$^{frt/+}$ ESC clones and subsequent blastocyst injections of the modified ESCs were performed as previously described (Huijbers *et al.,* 2015). The resulting chimeric males were crossed with *Cdh1*$^{F/F}$ (FVB) females to generate *Wap–Cre*;*Cdh1*$^{F/F}$;*Col1a1*$^{invCAG-Ppp1r12a-ex1-9-IRES-Luc/+}$ (*Wap–Cre*;*Cdh1*$^{F/F}$;*Ppp1r12a*$^{ex1–9}$) and *Wap–Cre*;*Cdh1*$^{F/F}$;*Col1a1*$^{invCAG-Trp53bp2-ex13-18-IRES-Luc/+}$ (*Wap–Cre*;*Cdh1*$^{F/F}$;*Trp53bp2*$^{ex13–18}$) mice, respectively.

The generation of *Wap–Cre*;*Cdh1*$^{F/F}$;*Col1a1*$^{invCAG-Cas9-IRES-Luc/+}$ (*Wap–Cre*;*Cdh1F/F*;*Cas9)* mice has been described by Annunziato *et al*., 2016. The *Wap–Cre*, *Cdh1*^{F}, *T2/Onc*, *Rosa26*^{Lox66SBLox71} and *Col1a1*$^{invCAG-Cas9-IRES-Luc}$ alleles were confirmed by PCR as previously described (Derksen *et al.,* 2006; Derksen *et al.,* 2011; Collier *et al.,* 2005; Dupuy *et al.,* 2005; Huijbers *et al.,* 2014). *Col1a1*$^{invCAG\text{-}Ppp1r12a\text{-}ex1\text{-}9\text{-}IRES\text{-}Luc}$, *Col1a1*$^{invCAG\text{-}Trp53bp2\text{-}ex13\text{-}18\text{-}IRES\text{-}Luc}$ and wild-type alleles were detected by standard PCR with an annealing temperature of 58 °C (product sizes of 433 bp, 351 bp and 234 bp, respectively).

All mouse experiments were approved by the Animal Ethics Committee of the Netherlands Cancer Institute and performed in accordance with institutional, national and European guidelines for animal care and use.

In vivo bioluminescence imaging. In vivo bioluminescence imaging was performed as previously described (Henneman *et al.,* 2015). Signal intensity was measured over the region of interest and quantified as flux (photons per s per cm^2 per sr).

Intraductal injection. Intraductal injections were performed as previously described (Annunziato *et al.,* 2016; Krause *et al.,* 2013). Lentiviral titers ranging from 2 × 10^8 transducing units (TU)/ml to 2.5 × 10^9 TU/ml were used.

Histology and immunohistochemistry. Tissues were formalin-fixed and paraffin-embedded (FFPE) by routine procedures. Hematoxylin and eosin (H&E) staining was performed as previously described (Doornebal *et al.,* 2013). All of the tissues were stained with H&E. Mammary tumors were also stained for expression of E-cadherin, CK1, CK8 and/or vimentin and were reviewed by a European College of Veterinary Pathologists (ECVP) certified veterinary pathologist (S.K.) in a blinded manner, according to international consensus of mammary pathology (Cardiff *et al.,* 2000), and by a consultant clinical pathologist with expertise in breast cancer (J.W.). ILCs were characterized by small to moderately sized neoplastic epithelial cells with a dominant 'Indian file' growth pattern in tumor-associated stroma with moderately polymorphic nuclei, sporadic mitoses and occasional intracytoplasmic vacuolization. The neoplastic cells were CK8$^+$ and lacked expression of E-cadherin (CDH1$^-$), as confirmed by immunohistochemistry. Growth patterns reminiscent of the alveolar or solid variants of ILC were also occasionally observed with nests and sheets of tumor cells, respectively. Tumors with a spindle cell morphology were classified based on CK8$^-$VIM$^+$ neoplastic cells. Squamous metaplasia was classified based on morphology as well as expression of CK1. Tumor burden was calculated as the ratio between the total tumor area and the area of the whole mammary gland, using ImageJ software version 1.4.3.67. Immunohistochemical staining was processed as previously described (Henneman *et al.,* 2015; Doornebal *et al.,* 2013). The images on the slides were captured using an Axioskop 40 microscope and an AxioCam MRc5 camera (Zeiss) and analyzed using the ZEN lite 2012 (Blue edition) software, or the slides were digitally processed using the Aperio ScanScope (Aperio, Vista, CA, USA) and captured using ImageScope software version 12.0.0 (Aperio).

Cell culture

Cell lines. Purified primary MMECs were isolated from 10- to 15-week-old females as previously described (Boelens *et al.,* 2016; Ewald *et al.,* 2008) and cultured in Dulbecco's modified Eagle's medium (DMEM)-F12 containing 10% fetal bovine serum (FBS), 100 IU/ml penicillin, 100 µg/ml streptomycin, 5 ng/ml insulin, 5 ng/ml epidermal growth factor (EGF) (all from Life Technologies) and 5 ng/ml cholera toxin (Gentaur). HC11 cells (Ball *et al.,* 1988) were cultured in DMEM/F12-Glutamax medium containing 10% FBS, 100 IU/ml penicillin, 100 µg/ml streptomycin, 5 ng/ml insulin, 5 ng/ml EGF (all from Life Technologies). All cell lines were routinely tested for mycoplasma contamination using the MycoAlert mycoplasma detection kit (Lonza).
Additional experimental details regarding the cell lines, retroviral vectors and virus production, clonogenic assays, DNA damage response in SB-induced tumor derived cell lines and flow cytometry are described in the Supplementary Note (online).

Nucleic acid isolation and analysis

Clustering with human TCGA breast cancers based on the PAM50 gene signature. To compare the expression of the mILCs to the human intrinsic breast cancer subtypes, we integrated the SB tumor gene expression data with expression data from mammary tumors from *Wap–Cre*;*Cdh1*$^{F/F}$;*Pten*$^{F/F}$ and *K14–Cre*;*Brca1*$^{F/F}$;*Trp53*$^{F/F}$ mice and with expression data from the TCGA BRCA data set. For the mouse models, sequence reads were downloaded from the European Nucleotide Archive (accession number PRJEB14147) and processed in the same manner as described for the SB-induced tumor samples to obtain gene counts. For the TCGA data, normalized gene expression counts were downloaded from Firehose (data set version 2016_01_28), and the PAM50 subtype assignment of the TCGA breast tumors was obtained from the TCGA BRCA publication (Cancer Genome Atlas, 2012). The mouse read counts were normalized by correcting for sequencing depth using the DESeq's (Love *et al.,* 2014) median-of-ratios approach and then applying a log-transformation. The human expression data were log-transformed and combined with the normalized mouse counts by concatenating the two data sets on orthologous genes and then normalizing for batch effects using ComBat (Jonhson *et al.,* 2007), as implemented in the sva R package (version 3.15.0). Unsupervised hierarchical clustering (Euclidean distance, average linkage) was performed using the 46 orthologous mouse genes from the PAM50 signature (Parker *et al.,* 2009).

Identification of molecular mILC subtypes. The molecular mILC subtypes were identified using the NMF package (Gaujoux *et al.,* 2010) (version 0.20.6) in R (version 3.3.1). The number of subtypes was determined by performing the NMF analysis for 2–5 clusters (using 30 iterations) and selecting the optimal number of clusters according to the

consensus silhouette statistic and the consensus clustering results (Supplementary Fig. 14). The final clustering was determined by performing the NMF factorization for four clusters and using 200 iterations. We tested for associations between subtypes and morphological characteristics by using the Fisher's exact test (one-sided), correcting for multiple testing by using Benjamini–Hochberg correction. Associations with an FDR < 0.05 were considered to be significant. The PAM50 expression of the different subtypes was compared by performing unsupervised hierarchical clustering (Euclidean distance, average linkage) on the normalized mouse expression data using the PAM50 genes.
The mouse molecular subtypes were compared with the hILC subtypes by combining our mouse data set and the human TCGA ILC expression data set (Ciriello *et al.,* 2015) using ComBat in the same manner as for the PAM50 analysis. The distribution of human/mouse subtypes in the combined data set was visualized by performing a PCA analysis on 49 orthologous genes from TCGA's 60-gene subtype classifier and plotting the samples using the first two principal components.

Mapping of SB insertion sequences and identification of common insertion sites. Genomic DNA from the SB-containing tumors was processed using the ShearSplink protocol, and transposon insertions were amplified as previously described (Koudijs *et al.,* 2011). Samples were sequenced in four sequencing runs on the 454-Titanium platform according to the manufacturer's protocol (Roche). The resulting reads were filtered for contaminant sequences, which represent non-hopped transposons, and were trimmed using Cutadapt (Martin *et al.,* 2011; version 1.12) to remove the splinkerette adaptor and transposon sequences (which require a minimum overlap of 10 bp with both sequences). Reads without a valid adaptor and transposon sequence were discarded from the analysis, as were reads shorter than 15 bp after trimming. The trimmed reads were aligned to the mouse reference genome (mm10) using Bowtie2 (Langmead *et al.,* 2012; version 2.2.8). After the alignment, redundant sequences that belonged to the same tumor and that mapped to the same genomic location were collapsed into a single insertion. To avoid issues with slight variations in the alignment, insertions from the same sample that occurred within 5 bp of each other were collapsed into a single insertion.

To identify CISs, we analyzed the insertions using CIMPL (version 1.1), which uses a Gaussian kernel convolution (GKC)-based approach to identify CISs (de Ridder *et al.,* 2006). The CIMPL analysis was performed with the following options: scales of 10 kb and 30 kb, 10,000 iterations, the SB preset for correction of insertion bias and exclusion of local hopping. The identified CISs were assigned to putative target genes using rule-based mapping (RBM; de Jong *et al.,* 2011), using the SB preset and restricting assignments to the closest gene in the event that multiple putative targets were identified. The final set of CIS insertions was obtained by removing insertions that did not belong to at least one CIS. CIS insertions were assigned to their putative target genes via their respective CIS(s).

Insertion clonality and association with subtypes. Clonality scores of insertions were calculated using ShearSplink's unique ligation point (ULP) score, which counts the number of unique positions in the ligation point (LP) between the genomic DNA and the splinkerette adaptor for a given insertion. These differing ligation points are the result of ShearSplink's stochastic shearing process, in which the splinkerette adaptor is ligated to randomly fragmented DNA, effectively barcoding tumor cells with a unique identifier. Using this ULP score we calculated a relative clonality score by normalizing the ULP of each insertion by the highest ULP within the corresponding tumor. This ensured that each insertion was assigned a score between 0 (no insertion) and 1 (clonal insertion).

Identifying CIS insertion biases. The orientation insertion bias of a gene was calculated as the ratio of sense/antisense insertions in the gene, weighted by the clonality of the insertions to give more weight to relatively clonal insertions:

$$b_g = \sum_i I(i)w_i \; with \; i \in \mathrm{T}_g$$

where *bg* is the bias of gene *g*, *Tg* is the collection of insertions in gene *g*, *wi* is the clonality of insertion *i* and *I* is an indicator function that returns 1 if insertion *i* is sense and 0 otherwise. Genes with *bg* close to 1 are biased toward sense insertions, whereas genes with *bg* close to 0 are biased toward antisense insertions.

CIS differential expression analysis. Genes were tested for differential expression over their insertion sites by using the group-wise differential expression test implemented in IM-Fusion (de Ruiter *et al.,* 2017; version 0.3.0). This test essentially divides the exons of a given gene into two groups: exons before the insertion sites in the gene and exons after the insertion sites. The expression counts of exons before the insertion sites are then used to normalize expression differences between samples, as the expression of these exons is not affected by the presence of insertions. The normalized expression counts of exons after the insertion sites are then compared between samples with and without an insertion to test for differential expression. Genes with $P < 0.05$ were considered to show a significant increase or decrease in expression as a result of their insertions.

KEGG enrichment analysis. To test whether insertions were enriched in genes involved in specific pathways, we first downloaded KEGG pathway gene sets using KEGGs REST API. We then tested for enrichment using the hypergeometric test, applying Benjamini–Hochberg correction to correct for multiple testing. Pathways with an FDR < 0.1 were considered to be significantly enriched.

Mutual exclusivity and co-occurrence analysis of CISs. Mutual exclusivities and co-occurrences between CISs were identified using DISCOVER (Canisius *et al.,* 2016). Pairs of mutually exclusive or co-occurring genes were identified by testing for significant pairwise gene associations (FDR < 0.25). Co-occurrences were filtered to remove trivial associations between overlapping genes. Pairwise mutual exclusivities were used to assemble larger groups of genes, which were tested for significance using DISCOVER's group-wise test.

Candidate genes in hILC and human breast cancer. Mutations and copy-number events in the TCGA ILC and BRCA data sets (Ciriello *et al.,* 2015) were visualized using cBioPortal (Gao *et al.,* 2013). The correlation between copy-number events and gene expression levels was determined by using copy-number and expression data from the TCGA BRCA provisional data set (as downloaded from Firebrowse, version 2016_01_28; Cancer Genome Atlas, 2012). After removing normal samples and duplicate patient samples, correlation scores and *P* values were calculated using Spearman's rank correlation statistic.

Exon-level expression of Ppp1r12a, Ppp1r12b and Trp53bp2. Exon-level expression was calculated as the average depth of coverage of the exons in each respective sample to avoid biases due to differences in exon sizes. To obtain a relative measure of exon expression within the gene for a given sample, these expression values were transformed to *z*-scores for each combination of sample and gene. This effectively provided a measure of the degree to which the expression of an exon is increased or decreased relative to the mean expression of the gene in the corresponding sample. The resulting *z*-scores were visualized as a heat map per gene, with black lines indicating the position of insertions in the respective samples.

Experimental details regarding genomic DNA isolation, PCR amplification and TIDE analysis, RNA preparation and sequencing, and northern blotting are described in the Supplementary Note (online).

Protein isolation and analysis

Antibodies. Primary antibodies to the following proteins were used: MYH9 (1:5,000, Sigma HPA 001644), Flag (1:1,000, Sigma F7425), PPP1R12A (1:1,000, Cell Signaling Technology (CST) 2634), PP1 (1:200, Santa Cruz sc-7482), P21 (1:1,000, BD Bioscience 556430), p53 (1:1,000, Monosan MONX10194), ribosomal protein S6 (1:1,000, CST 2217), phospho-S6 ribosomal protein (Ser235,Ser236) (1:1,000, CST 2211), p44/42 MAP kinase (1:1,000, CST 4695), phospho-p44/42 MAPK ERK1/ERK2(Thr202/Tyr204) (1:1,000 CST 9101), AKT1 (1:1,000, CST 2938), phospho-AKT(Ser473) (1:1,000, CST 4060) and β-actin (1:50,000, Sigma A5441).

Additional experimental details regarding immunoblotting, immunoprecipitation and LC/MS-MS analysis are described in the Supplementary Note (online).

Statistical analysis

For the mouse studies, no statistical tests were performed to determine the sample size, and no blinding of investigators was performed. Survival probabilities were estimated using the Kaplan–Meier method and compared using the Mantel–Cox test. Associations with the expression subtypes (concerning tumor morphology and candidate genes) were identified by using a one-sided Fisher's exact test (testing for co-occurrence) with Benjamini–Hochberg correction. Biases of morphologies or subtypes and candidate genes toward either of the *T2/Onc* lines were identified by using a two-sided Fisher's exact test with Benjamini–Hochberg correction. The KEGG pathway enrichment analysis was performed using a hypergeometric test with Benjamini–Hochberg correction. Correlation between copy number and expression in the TCGA data set was calculated using Spearman's rank correlation. Volcano plots of the immunoprecipitations followed by LC-MS/MS analysis were constructed using a permutation-based FDR-corrected *t*-test. Proteins were considered interactors when $P < 0.01$ and log2(abundance difference) > 1. In the cell death assay, groups were compared using Welch's *t*-test. The investigators were not blinded to the sample groups for all experiments. Graphs and error bars represent means ± s.d. Python 3.5, R 3.3.1 and GraphPad Prism 7.0 were used for the statistical analyses. *P* values < 0.05 and FDR values < 0.1 were considered significant, unless stated otherwise.

Code availability

Jupyter notebooks containing the code and results of the various computational analyses are available on GitHub. The software used for the insertion and CIS analysis has been implemented in a Python package called PyIM, which is also available on GitHub. For these analyses, version 0.2.0 of this package was used.

Data availability

All sequence data that support the findings of this study are available in the European Nucleotide Archive under accession number PRJEB14134. The ShearSplink sequencing data, together with additional files containing sample barcodes and other materials, are available in Figshare under the identifier https://doi.org/10.6084/M9.FIGSHARE.4765111. Processed expression and insertion data have also been deposited in Figshare under the identifier https://doi.org/10.6084/M9.FIGSHARE.4929866. All other data are available within this paper and its supplementary information files, the Jupyter notebooks, or are available from the corresponding author on request.

Methods-only references

Annunziato, S. et al. Modeling invasive lobular breast carcinoma by CRISPR–Cas9-mediated somatic genome editing of the mammary gland. Genes Dev. 30, 1470–1480 (2016).

Dupuy, A.J., Akagi, K., Largaespada, D.A., Copeland, N.G. & Jenkins, N.A. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. Nature 436, 221–226 (2005).

Huijbers, I.J. et al. Rapid target gene validation in complex cancer mouse models using re-derived embryonic stem cells. EMBO Mol. Med. 6, 212–225 (2014).

Henneman, L. et al. Selective resistance to the PARP inhibitor olaparib in a mouse model for BRCA1-deficient metaplastic breast cancer. Proc. Natl. Acad. Sci. USA 112, 8409–8414 (2015).

Krause, S., Brock, A. & Ingber, D.E. Intraductal injection for localized drug delivery to the mouse mammary gland. J. Vis. Exp. 80, e50692 (2013).

Doornebal, C.W. et al. A preclinical mouse model of invasive lobular breast cancer metastasis. Cancer Res. 73, 353–363 (2013).

Cardiff, R.D. et al. The mammary pathology of genetically engineered mice: the consensus report and recommendations from the Annapolis meeting. Oncogene 19, 968–988 (2000).

Ewald, A.J., Brenot, A., Duong, M., Chan, B.S. & Werb, Z. Collective epithelial migration and cell rearrangements drive mammary branching morphogenesis. Dev. Cell 14, 570–581 (2008).

Ball, R.K., Friis, R.R., Schoenenberger, C.A., Doppler, W. & Groner, B. Prolactin regulation of ⍰-casein gene expression and of a cytosolic 120-kDa protein in a cloned mouse mammary epithelial cell line. EMBO J. 7, 2089–2095 (1988).

Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA–seq data with DESeq2. Genome Biol. 15, 550 (2014).

Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127 (2007).

Gaujoux, R. & Seoighe, C. A flexible R package for non-negative matrix factorization. BMC Bioinformatics 11, 367 (2010).

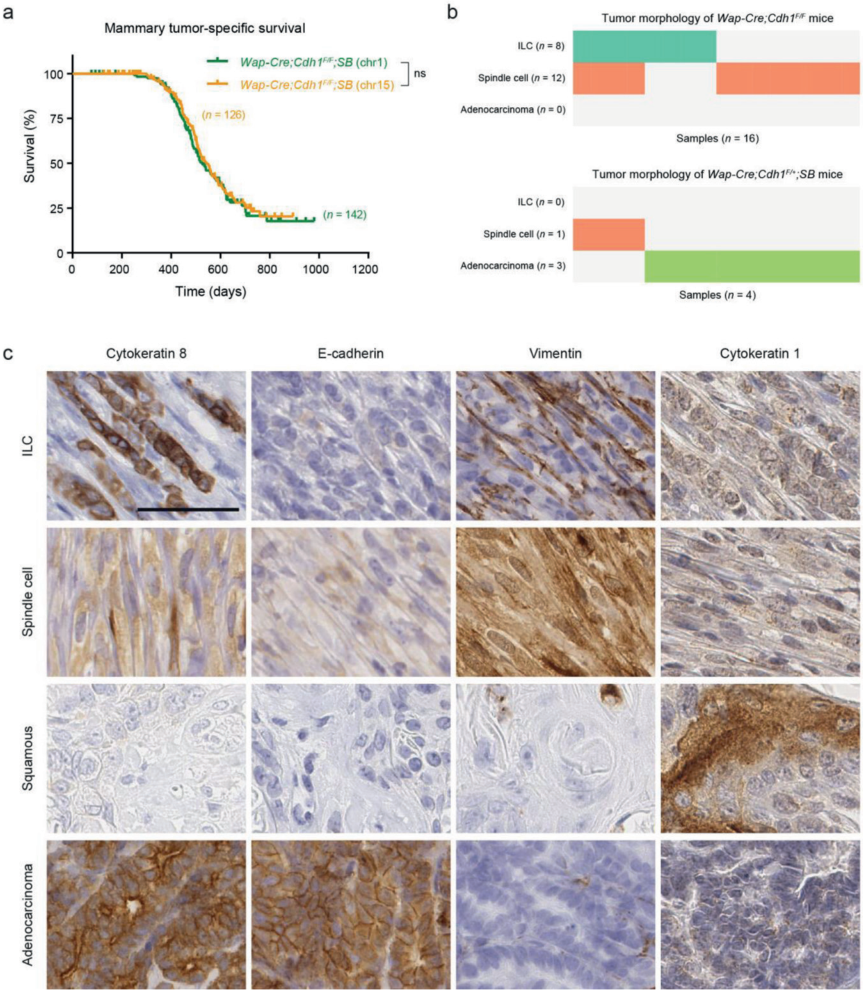
Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17, 10 (2011).

Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

de Ruiter, J.R. et al. Identifying transposon insertions and their effects from RNA-sequencing data. Nucleic Acids Res. [Epub ahead of print]

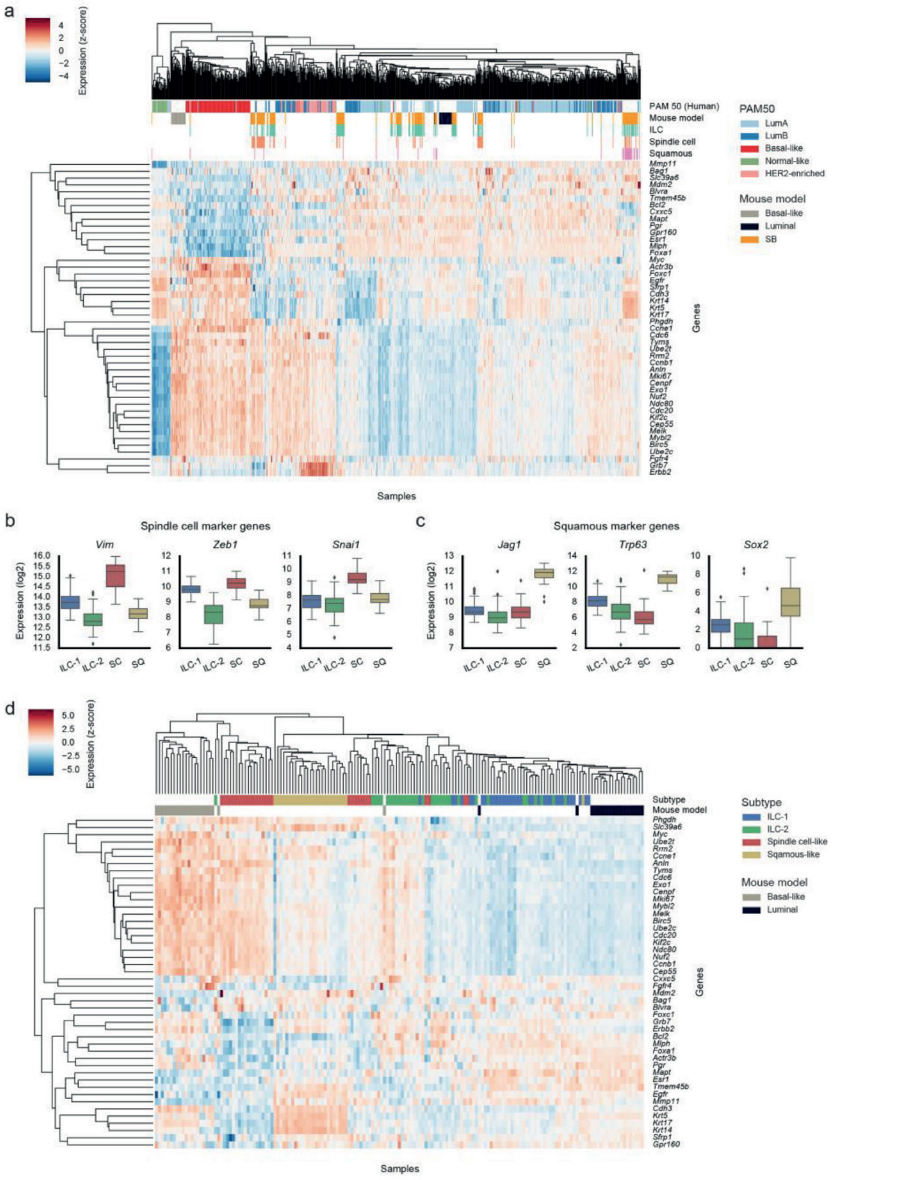
Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6, pl1 (2013).

Supplementary Figure legends



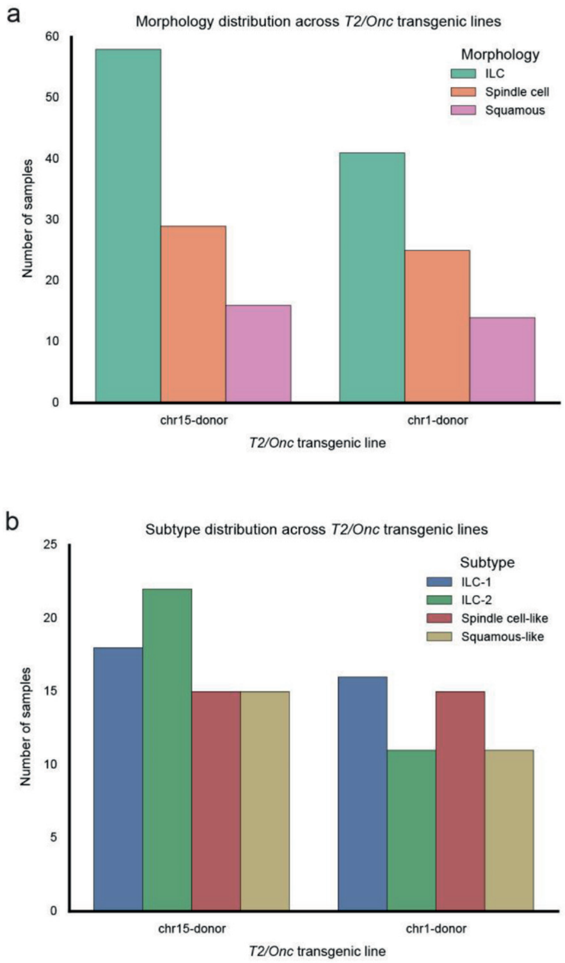
Supplementary Figure S1

Survival curves for the individual *T2/Onc* transgenic mouse lines, distribution and immunohistochemical stainings of the different tumor morphologies. (A) Survival curves *Wap-Cre;Cdh1$^{F/F}$;SB* mice carrying the *T2/Onc* transposon donor loci on chromosomes 1 and 15 (n = 142 and n = 126 mice, respectively). ns, no significant difference (Mantel–Cox test). (B) Tumor morphologies of *Wap-Cre;Cdh1$^{F/F}$* (n = 16) and *Wap-Cre;Cdh1$^{F/+}$;SB* (n = 4) female mice. (C) Immunohistochemical detection of E-cadherin, cytokeratin 1, cytokeratin 8 and vimentin in the different tumor morphologies. Scale bar, 50 µm.



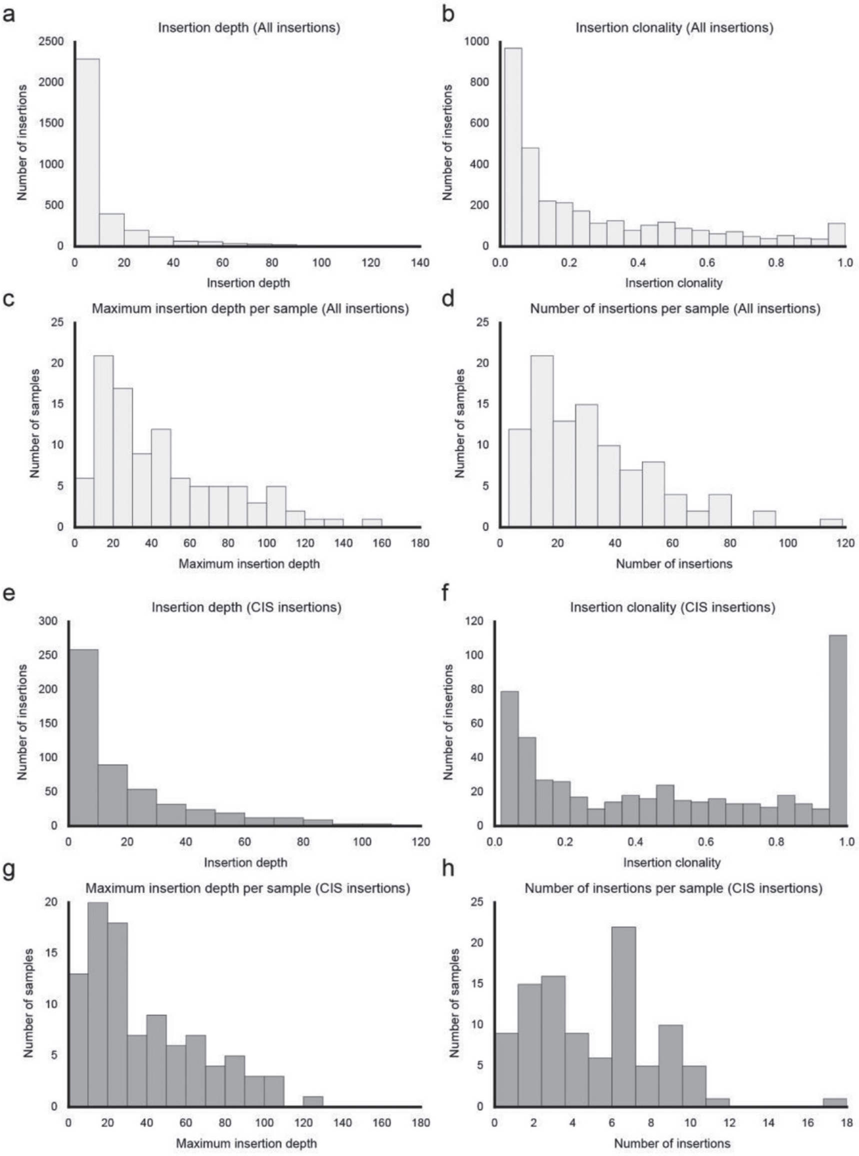
Supplementary Figure S2

Additional plots accompanying the PAM50 gene expression analyses. (A) Unsupervised clustering (Euclidean distance, average linkage) of the SB-induced tumors with reference mouse models and human breast cancer samples from TCGA, showing the expression of all 46 orthologous mouse genes from the PAM50 gene signature. (B-C) Expression of known marker genes associated with squamous and spindle cell tumors, showing high expression of the corresponding markers in the associated molecular subtype. Boxes extend from the third (Q3) to the first (Q1) quartile (IQR), with the line at the median; whiskers extend to Q3 + 1.5 IQR and Q1 − 1.5 IQR. Points beyond the ends of the whiskers are outliers. SC, spindle cell-like; SQ, squamous-like. (D) Unsupervised clustering (Euclidean distance, average linkage) of the SB-induced molecular subtypes with the reference luminal and basal-like mouse models.



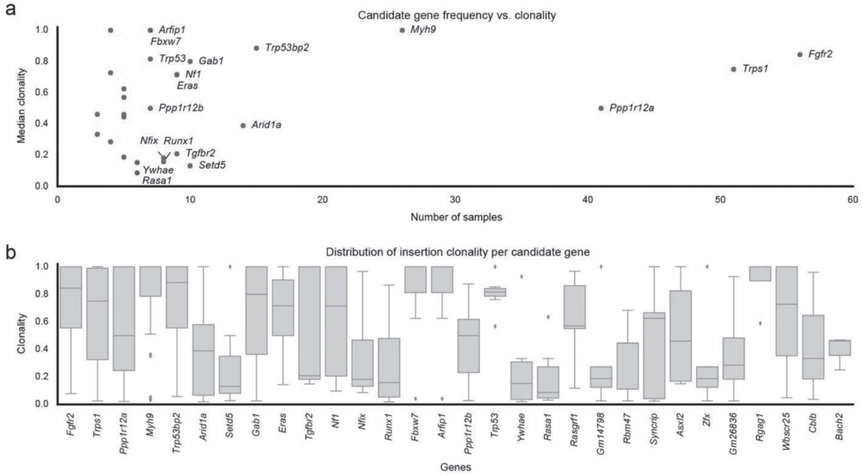
Supplementary Figure S3

Distribution of morphology and expression subtypes across the *T2/Onc* transgenic mouse lines. (A) Distribution of morphology across the two *T2/Onc* transgenic lines, showing that there is no significant bias between the two lines. Lack of any significant bias was confirmed using pairwise Fisher’s exact tests with Benjamini–Hochberg correction. (B) Distribution of subtypes across the two *T2/Onc* transgenic lines, also showing no significant bias between the two lines.



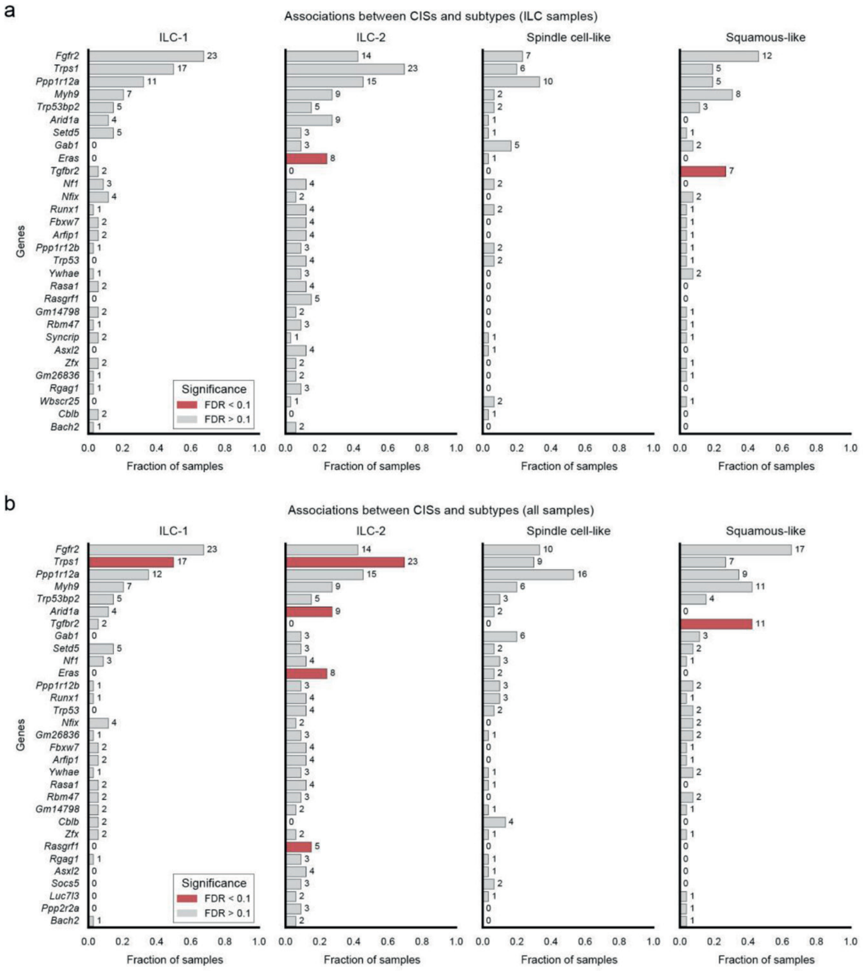
Supplementary Figure S4

Histograms detailing various statistics of the identified insertions, both for all insertions and for insertions within CISs. (A) Distribution of the insertion depths (as indicated by ShearSplink's ULP score) over all insertions. (B) Distribution of insertion clonality scores. (C) Distribution of the maximum insertion depth (maximum ULP score) per sample. (D) Distribution of the number of insertions per sample, with a median of 29 insertions per sample. (E-H) Same statistics as in a–d, but calculated for the CIS insertions, with a median of 5 CIS insertions per sample.



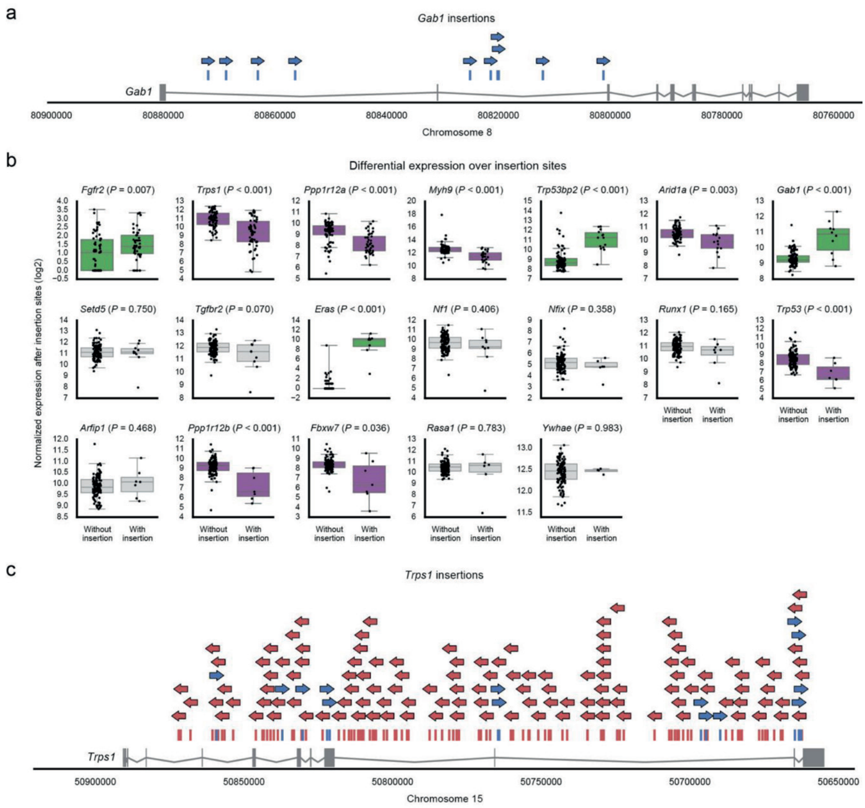
Supplementary Figure S5

Additional details on the clonality of the candidate genes. (A) Ranking of genes by their overall frequency and the median clonality of their insertions. For samples with multiple insertions in the same gene, we used the clonality of the strongest insertion to avoid underestimating the median clonality due to local hopping. For clarity, only the main candidates (occurring in six or more samples) are labeled. (B) Clonality distribution of insertions in all candidate genes, ranked (from left to right) by the frequency of each gene.



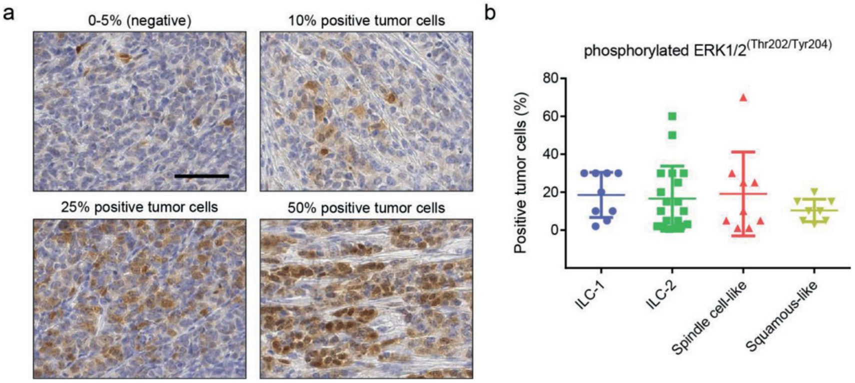
Supplementary Figure S6

Distribution of the candidate genes over the identified molecular subtypes. Bar plots indicating the distribution of all candidate genes over the molecular subtypes. Red bars indicate significant associations between candidate genes and the corresponding molecular subtype (FDR < 0.1, one-sided Fisher's exact test with Benjamini–Hochberg correction). (A-B) This analysis was performed for both the set of 99 tumors with an ILC morphology (A) and the full set of 123 tumor samples (to increase statistical power) (B). The latter analysis identified an additional enrichment for *Trps1* in the combined mILC-1 and mILC-2 subtypes, suggesting that *Trps1* may play a role in the ILC morphology of these tumors. Additionally, the mILC-2 subtype was further enriched for insertions in *Arid1a* and *Rasgrf1*, indicating that expression differences between mILC-1 and mILC-2 may in part be driven by insertions in these genes.



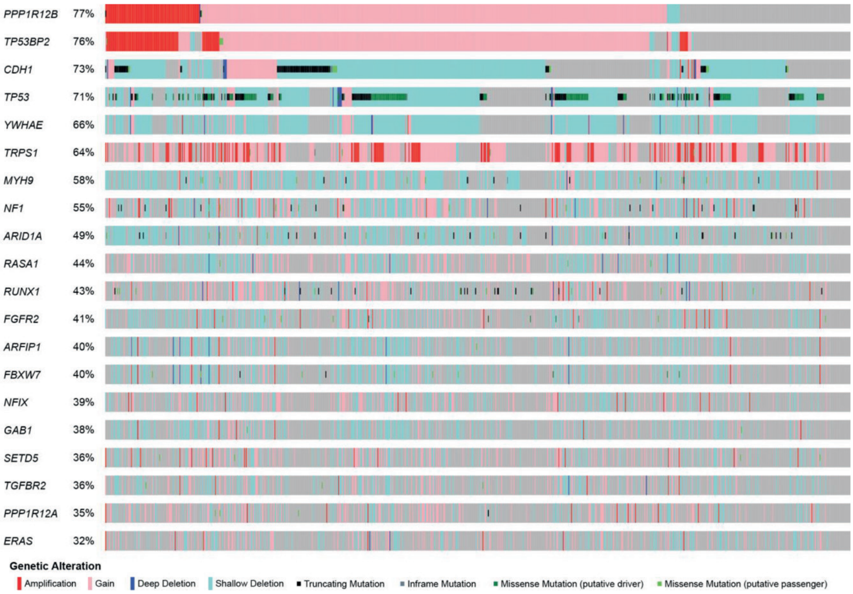
Supplementary Figure S7

Additional details on the insertion patterns and differential expression of the candidate genes. (A) Insertion pattern in *Gab1*, showing bias toward activating insertions. (B) Boxplots for the main candidate genes showing the difference in expression after the insertion sites in the corresponding gene between samples with and without an insertion. The boxplots and P values were calculated using IM-Fusion's differential expression test, which essentially compares the expression of exons after the insertion sites in each gene between samples with an insertion and samples without an insertion, after normalizing for differences in overall expression between samples (see the Online Methods for more details). Boxes extend from the third (Q3) to the first (Q1) quartile (IQR), with the line at the median; whiskers extend to Q3 + 1.5 IQR and Q1 − 1.5 IQR. P values were calculated using the non-parametric Mann–Whitney U test. Green/purple boxplots indicate significant increases and decreases in expression (P < 0.05), respectively. (C) Insertion pattern in *Trps1*, showing bias toward truncating/inactivating insertions.



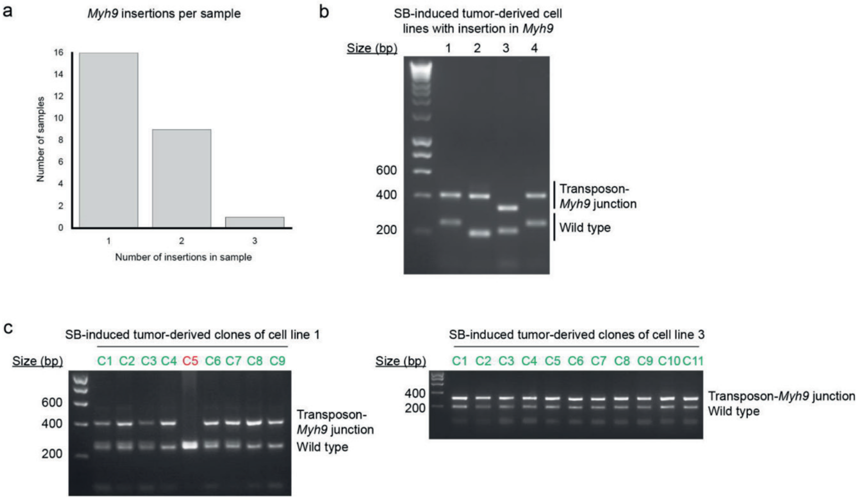
Supplementary Figure S8

Representative images and quantification of immunohistochemical staining for phosphorylated ERK1/2, a downstream protein of the RAS/MAPK signaling pathway. (A) Representative images of different percentages of phosphorylated ERK1/2$^{(Thr202/Tyr204)}$ staining in SB-induced tumors. Scale bar, 50 µm. (B) Quantification of phosphorylated ERK1/2$^{(Thr202/Tyr204)}$ staining in the different molecular subtypes. Mean ± s.d.

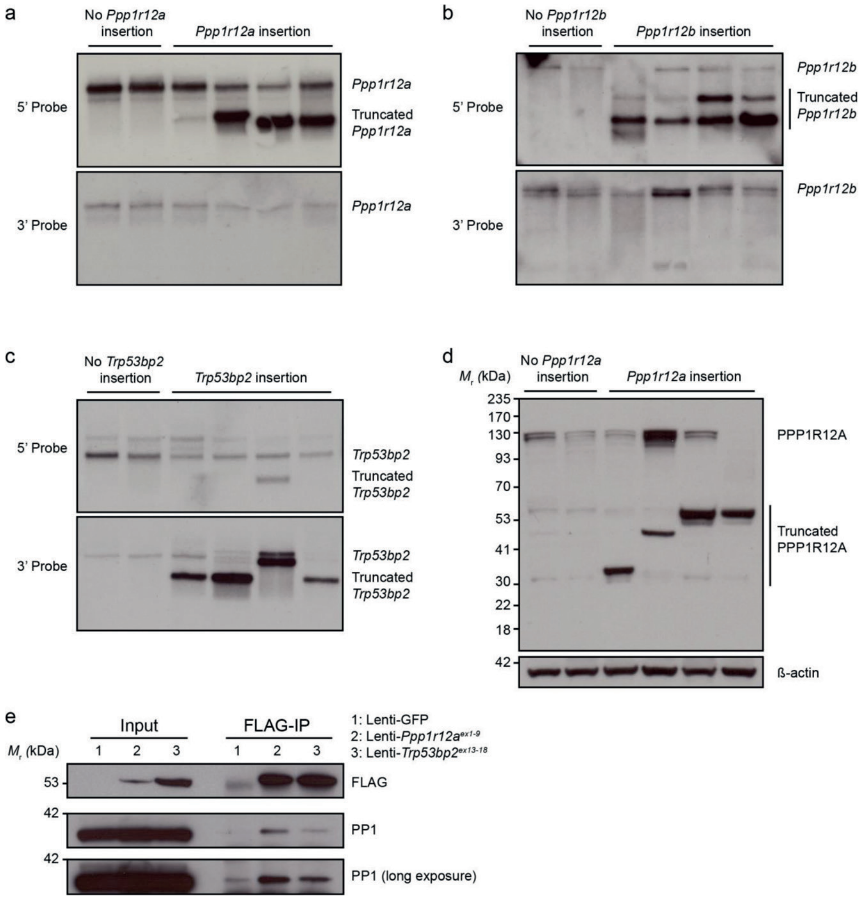


Supplementary Figure S9

Overview of the mutations and copy-number events in the TGCA breast cancer data set (816 samples) for each of the main candidate genes and *CDH1*. Percentages indicate the fraction of tumors with alterations in the respective genes.

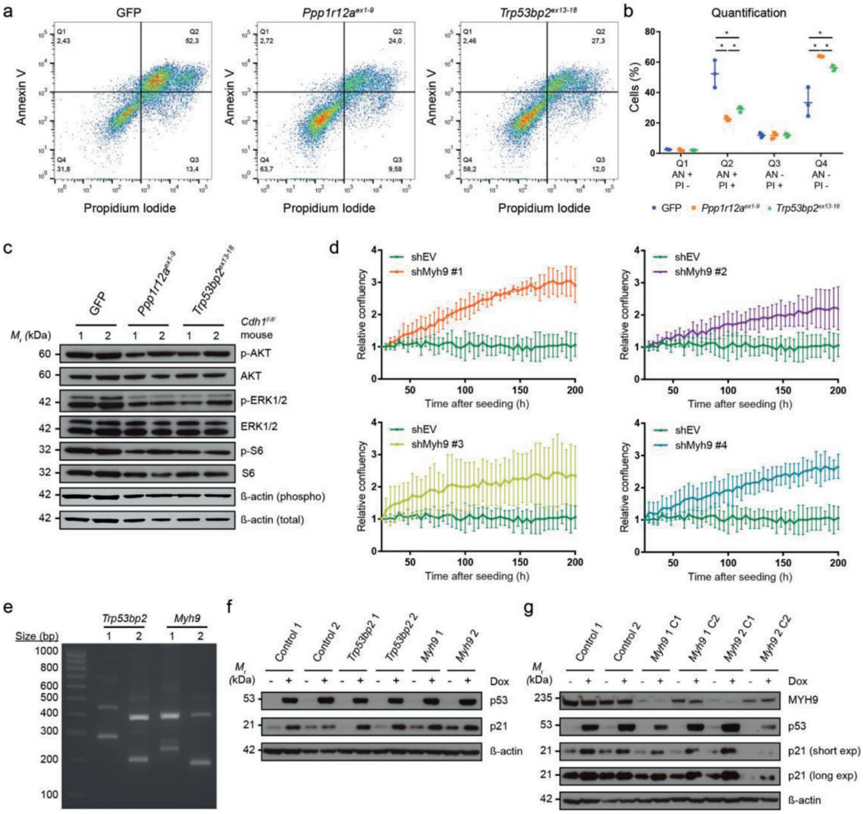


Supplementary Figure S10
Myh9 haploinsufficiency in ILC formation in SB-induced tumor-derived cell lines with or without *Myh9* insertion. (A) Distribution of the number of insertions per sample in *Myh9*, showing that the majority of the tumors show single insertions in the gene. (B) PCR amplification of the transposon–*Myh9* junction fragments in polyclonal SB-induced tumor-derived cell lines. (C) Heterozygous *Myh9* insertions by PCR in clones derived from two SB-induced tumor cell lines. Green color indicates a correct clone; red color indicates an incorrect clone.



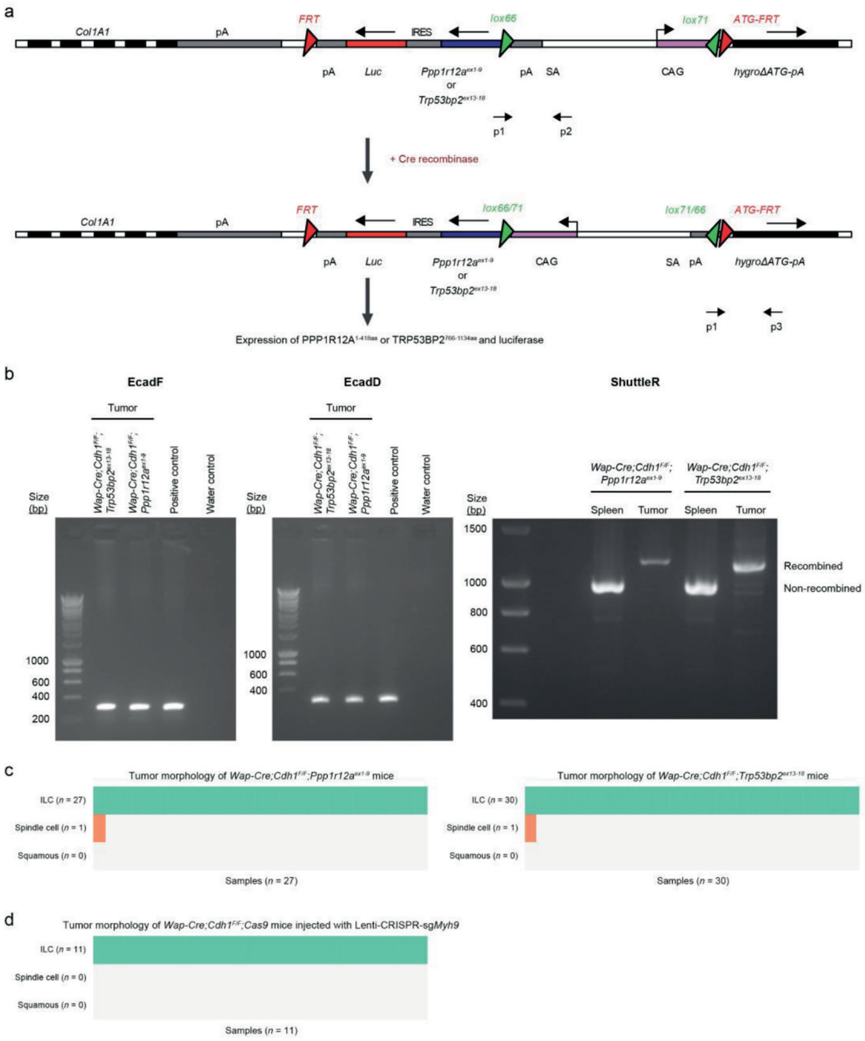
Supplementary Figure S11

Expression of truncated *Ppp1r12a/b* and *Trp53bp2* and PPP1R12A protein expression in SB-induced tumors. (A-C) Truncated *Ppp1r12a/b* and *Trp53bp2* in SB-induced tumors, as visualized by northern blot analysis. (D) Expression of PPP1R12A in SB-induced tumors with and without insertions in *Ppp1r12a*, as visualized by immunoblotting using an anti-PPP1R12A antibody. β-actin is shown as a loading control. (E) Coimmunoprecipitation of PP1 in HC11 cells expressing GFP, truncated PPP1R12A or TRP53BP2, as visualized by immunoblotting using anti-FLAG and anti-PP1 antibodies.



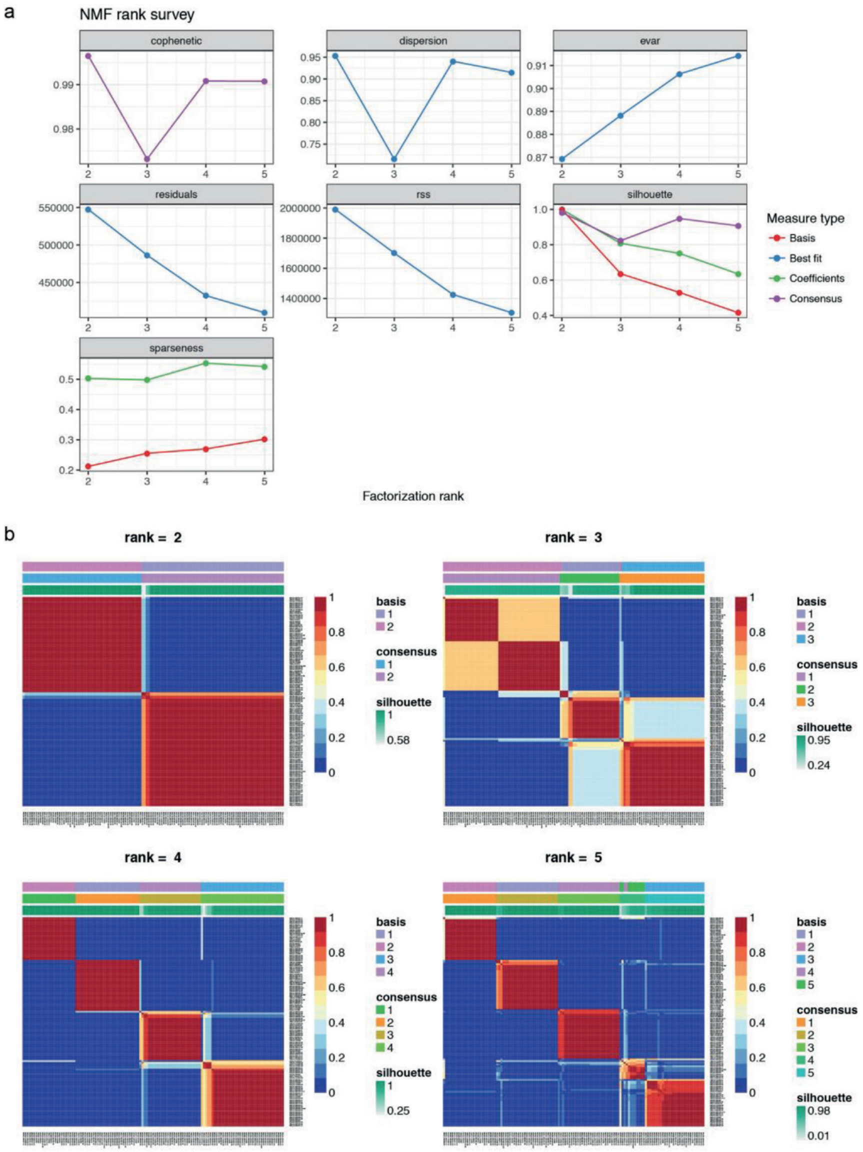
Supplementary Figure S12

Characterization of cell death, cell survival markers and activation of a p53 response in E-cadherin-deficient cells expressing truncated PPP1R12A or TRP53BP2 or showing reduced levels of MYH9. (A) Representative dot plots depicting the proportions of Annexin V (AN)- and/or propidium iodide (PI)-positive cells 72 h after seeding of AdCre-transduced *Cdh1*$^{F/F}$ MMECs with simultaneous transduction of Lenti-GFP, Lenti-*Ppp1r12a*$^{ex1-9}$ or Lenti-*Trp53bp2*$^{ex13-18}$. (B) Quantification of Annexin V– and PI-positive cells from AdCre-transduced *Cdh1*$^{F/F}$ MMECs with simultaneous transduction of Lenti-GFP, Lenti-*Ppp1r12a*$^{ex1-9}$ or Lenti-*Trp53bp2*$^{ex13-18}$. Asterisks indicate P < 0.05 (Welch's t-test). Mean ± s.d. of three independent experiments. (C) Expression of total and phosphorylated AKT, ERK1/2 and S6 72 h after seeding of AdCre-transduced *Cdh1*$^{F/F}$ MMECs with simultaneous transduction of Lenti-GFP, Lenti-*Ppp1r12a*$^{ex1-9}$ or Lenti-*Trp53bp2*$^{ex13-18}$, as visualized by immunoblotting. β-actin is shown as a loading control. (D) Cell survival of AdCre-transduced *Cdh1*$^{F/F}$*;mT/mG* MMECs with simultaneous shRNAmediated knockdown of *Myh9* with different shRNAs, as quantified by using real-time IncuCyte imaging for 200 h. Mean ± s.d. of three independent experiments. (E) PCR amplification of the transposon–*Trp53bp2* and transposon–*Myh9* junction fragments in polyclonal SB-induced tumor-derived cell lines. (F) Expression of p53 and p21 in SB-induced tumor-derived cell lines with insertions in *Trp53bp2* or *Myh9* as compared to SB-induced tumor-derived cell line controls, as visualized by immunoblotting. Cells were treated for 6 h with water (control) or doxorubicin (Dox; 1 µM). β-actin is shown as a loading control. (G) Expression of p53 and p21 in SB-induced tumorderived clones with insertions in *Myh9* as compared to SB-induced tumor-derived control clones, as visualized by immunoblotting. Cells were treated for 6 h with water (control) or doxorubicin (Dox; 1 µM). -actin is shown as a loading control. Exp, exposure.



Supplementary Figure S13

Overview of Cre-conditional alleles and distribution of the tumor morphology in the different genetically engineered mouse models. (A) Depiction of Cre-conditional *invCAG-Ppp1r12a*$^{ex1\text{-}9}$*-IRES-Luc* and *invCAG-Trp53bp2*$^{ex13\text{-}18}$*-IRES-Luc* alleles in the *Col1a1* locus. Cre-mediated recombination allows inversion of the CAG promoter, resulting in expression of PPP1R12A$^{1\text{-}418aa}$ or TRP53BP2$^{766\text{-}1134aa}$ accompanied by luciferase expression. (B) Recombination status of the Cre-conditional alleles in tumors from *Wap-Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1\text{-}9}$ and *Wap-Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13\text{-}18}$ mice, as visualized by PCR. EcadF and EcadD are PCRs to detect the *Cdh1*^{F} and *Cdh1*$^{\Delta}$ alleles, respectively. ShuttleR detects the recombined (p1 and p3; 1,054 bp) and non-recombined (p1 and p2; 897 bp) Cre-conditional alleles (primer positions are shown in a). (C) Histological classification of tumors from *Wap-Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1\text{-}9}$ (n = 27) and *Wap-Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13\text{-}18}$ (n = 30) mice. (D) Histological classification of tumors (n = 11) from *Wap-Cre;Cdh1*$^{F/F}$*;Cas9* mice injected with Lenti-CRISPR-sg*Myh9*.



Supplementary Figure S14
Clustering statistics for different numbers of clusters in the NMF subtype analysis. (A) Overview of various clustering statistics for different numbers of clusters (2–5) in the NMF analysis of the SB-induced tumors. (B) Consensus maps for the same cluster sizes, showing the consistency of cluster assignments for the different numbers of clusters.