

Universiteit
Leiden
The Netherlands

Into the blue...Using mouse models to uncover genes driving tumorigenesis and therapy resistance in human breast cancer

Ruiter, J.R. de

Citation
Ruiter, J. R. de. (2019, May 22). *Into the blue..Using mouse models to uncover genes driving tumorigenesis and therapy resistance in human breast cancer*. Retrieved from https://hdl.handle.net/1887/73551

Version:	Not Applicable (or Unknown)
License:	Leiden University Non-exclusive license
Downloaded from:	https://hdl.handle.net/1887/73551

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle http://hdl.handle.net/1887/73551 holds various files of this Leiden University dissertation.

Author: Ruiter, J.R. de
Title: Into the blue...Using mouse models to uncover genes driving tumorigenesis and therapy resistance in human breast cancer
Issue Date: 2019-05-22

Using mouse models to uncover genes driving tumorigenesis and therapy resistance in human breast cancer

Julian Rutger de Ruiter

About the cover:

Scuba diving is one of the major hobbies that my wonderful wife managed to impart on me. One thing that has always amazed me about diving is how many beautiful and complex interactions happen under the surface of the water, but sadly remain hidden to most of us. Even with our technological advancements, the majority of our ocean's depths are still a mystery — 'alien' worlds which remain to be explored. Personally, I feel that this reflects our current understanding of (cancer) cell biology, as we have learned much but are still only scratching the surface of the complex processes that occur within our cells. Fortunately, new technologies are providing us with more and more tools to venture out into the unknown 'deep blue', so that we may one day truly understand the underpinnings of our being.

Cover design: Julian R. de Ruiter
Thesis layout: Julian R. de Ruiter
Printed by: Gildeprint, Enschede
ISBN: 978-94-6323-579-2

The layout design was based on the Clean Thesis style by Ricardo Langner. The cover photos were taken during the LSD 2018 Lustrum Diving Trip to Pico in the Azores.

The research described in this thesis was performed in the divisions of Molecular Pathology and Molecular Carcinogenesis of the Netherlands Cancer institute (Amsterdam, the Netherlands). Financial support was provided by the Netherlands Organization for Scientific Research (NWO: Cancer Genomics Netherlands) and the European Research Council (ERC Synergy project CombatCancer). The printing of this thesis was financially supported by the Netherlands Cancer Institute (Amsterdam, the Netherlands).

©Julian de Ruiter, 2019. All rights reserved. No part of this thesis may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without prior permission of the author and the publisher holding the copyright of the articles.

Using mouse models to uncover genes driving tumorigenesis and therapy resistance in human breast cancer

Het gebruik van muismodellen om de sturende krachten achter tumor ontwikkeling en therapie resistentie in humane borstkanker te ontdekken

Proefschrift

Ter verkrijging van de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties,
te verdedigen op woensdag 22 mei 2019
klokke 13:45 uur

door
Julian Rutger de Ruiter

Geboren te Farnborough, Groot Britannië
in 1988

Promotor: Prof. dr. Jos Jonkers
Copromotor: Prof. dr. Lodewyk Wessels

Promotiecommissie:
Prof. dr. Hubertus Irth (voorzitter)
Prof. dr. Joke Bouwstra (secretaris)
Prof. dr. Peter Devilee (Universiteit Leiden)
Prof. dr. Bas van Steensel (Erasmus Universiteit)
Dr. Jeroen de Ridder (UMC Utrecht)

1 Mouse models in the era of large human tumour sequencing studies 1

2 Scope of this thesis 33

3 Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular breast carcinoma 43

4 Identifying transposon insertions and their effects from RNA-sequencing data 105

5 Comparative oncogenomics and iterative mouse modeling identifies combinations of driver genes and drug targets in *BRCA1*-mutated breast cancer 149

6 Transcriptomics and transposon mutagenesis identify multiple *in vivo* mechanisms of resistance to the FGFR inhibitor AZD4547 189

7 Selective loss of PARG restores PARylation and counteracts PARP inhibitor-mediated synthetic lethality 229

8 General discussion 285

A Appendices 301

1

Mouse models in the era of large human tumour sequencing studies

Julian R. de Ruiter[1,2], Lodewyk F. A. Wessels[2,3,4], Jos Jonkers[1,3]

[1] Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[2] Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[3] Cancer Genomics Netherlands, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[4] Department of EEMCS, Delft University of Technology, Delft, the Netherlands

Published in Open Biology, 15 August 2018
Volume 8, Issue 8

1.1 Abstract

Cancer is a complex disease, in which cells progressively accumulate mutations disrupting their cellular processes. A fraction of these mutations drive tumourigenesis by affecting oncogenes or tumour suppressor genes, but many mutations are passengers that do not actively contribute to tumour development. The advancement of DNA/RNA sequencing technologies has enabled in-depth analysis of thousands of human tumours from various tissues to perform systematic characterization of their genomes/transcriptomes and to identify (epi)genetic changes associated with cancer. Combined with appreciable progress in algorithmic development, this expansion in scale has resulted in the identification of many cancer-associated mutations, genes and pathways that are considered to be (potential) drivers of tumour development. However, it remains challenging to systematically identify drivers in complex amplicons or deletions of copy number driven tumours, in non-coding regions of the genome and drivers affected by complex structural rearrangements. Furthermore, functional characterization of drivers, including their underlying biological mechanism and their effects on tumour evolution, is challenging in the human context due to the lack of genetically tractable experimental model systems in which the effects of drivers can be studied in the context of their tumour microenvironment. In this respect, mouse models of human cancer provide unique opportunities for pinpointing novel driver genes and detailed characterization of these genes. In this review, we provide an overview of approaches for complementing human studies with data from mouse models. We also discuss technological developments for cancer gene discovery and validation in mice.

1.2 Introduction

Cancer is a disease in which normal cells are deregulated by disruption of their cellular processes, resulting in increased proliferation, survival and invasion of surrounding tissues. This disruption is generally attributed to mutations in so-called 'driver' genes, which provide cells with a selective growth advantage and thereby drive their malignant transformation. Broadly speaking, driver genes can be divided into two classes of genes: oncogenes and tumour suppressor genes (TSGs)[1]. Oncogenes drive tumour development when activated by their mutations and are typically involved in processes related to proliferation[2], whereas TSGs are inactivated during tumourigenesis and are generally involved in processes protecting cells from DNA damage and malignant transformation.

Tumours are however mostly not the product of single mutations, but develop progressively over time through the accumulation of multiple mutations. Depending

on the affected gene(s), these mutations can increase the fitness or tumourigenic potential of cells (additional driver mutations), or have no effect on tumourigenesis (passenger mutations). Over time, this accumulation gives rise to subpopulations of cells (subclones) harbouring distinct sets of mutations, which are subject to Darwinian competition (clonal evolution) within the tumour lesion[3] (Figure 1.1). This competition pressures cells to acquire further mutations, resulting in increased fitness and the continued evolution of competing subclones within the tumour. External influences such as the immune micro-environment or drug treatment can also strongly influence the evolutionary process, either by selecting for specific subclones that are intrinsically resistant to immune surveillance or treatment, or by applying additional evolutionary pressure to acquire new mutations that confer resistance[4].

To effectively treat patients using precision medicine, in which treatments are tailored to target specific mutations that are present in a patient's tumour, it is crucial to identify exactly which mutations contribute to tumourigenesis and how they do so. However, although human sequencing studies have identified many genes contributing to cancer development, they do not provide evidence for causality or detailed insight into the biological mechanisms by which these genes drive tumour development. These studies also do not reveal whether drivers are essential for tumour maintenance and may therefore be of limited use for designing effective therapeutic strategies. In contrast, preclinical model systems such as genetically engineered mouse models (GEMMs) provide an experimentally tractable approach, in which the biological effects of specific mutations can be studied in more detail in a controlled genetic background. In this review, we describe several aspects of mouse models and how these can ultimately be applied to improve treatment of cancer patients. To this end, we first highlight several challenges in translating findings from human sequencing studies to the clinical setting, before explaining how some of these challenges can be addressed using complementary approaches in mouse model systems.

1.3 Challenges in human tumour sequencing studies

Several major human sequencing studies have been undertaken over the past years, aiming to identify and catalogue potential driver mutations across many different cancer types[5–8]. One of the key challenges in analyzing data from these efforts has been the separation of driver mutations from passenger mutations. To address this issue, many computational approaches have been developed to select driver genes using signals of positive selection in the pattern of somatic mutations in genes across

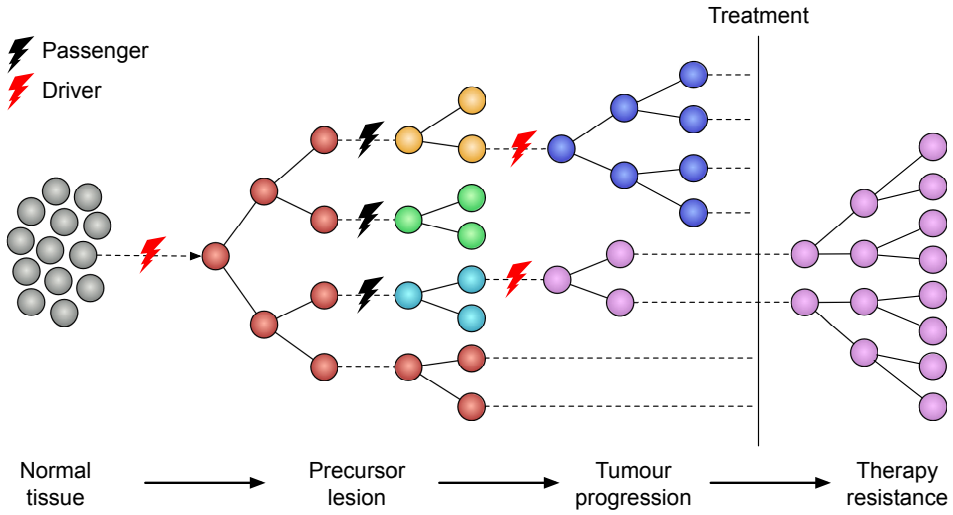


Fig. 1.1. Darwinian evolution during tumour development. Tumourigenesis is a multi-step process in which initially healthy cells progressively acquire multiple mutations disrupting their cellular processes and increasing their tumourigenic potential. Although many of these mutations are passengers that do not affect the ability of cells to survive and proliferate, a few rare mutations are drivers that will strongly increase the fitness of individual cells, allowing them to outcompete other cells. Over time, this stochastic process gives rise to Darwinian competition between subclones of cells harbouring different sets of mutations, driving selection towards subclones with increasing tumourigenic potential. External interventions such as drug treatments can influence this process by eradicating subclones that are sensitive to the given treatment. However, they can also drive selection towards subclones that are resistant to the treatment, leading to the emergence of therapy resistance that is seen in many cancer patients.

tumour samples. Examples include approaches based on mutation frequency[9,10], biases in the functional consequences of mutations[11–14] and clustering of mutations within genes[15].

Although these approaches have proven successful in identifying many driver genes affected by hotspot mutations, other types of mutations have proven more challenging. This is especially the case for DNA copy number driven diseases such as breast cancer, in which approaches aimed at identifying recurrent copy number gains/losses typically identify regions harbouring many genes[16,17]. Similarly, complex structural rearrangements[18] and mutations in non-coding regions[19] can be difficult to pinpoint specific target genes, requiring further prioritization of candidate genes using complementary approaches and/or exhaustive validation of potential drivers.

Another important challenge is the validation of putative driver genes and the further characterization of their biological mechanism(s), as this insight is crucial for understanding a driver's role in tumourigenesis and identifying potential therapeutic opportunities. Currently, most large-scale studies perform no or limited *in vivo* validation of candidate genes[20–24], as this additional follow-up is typically time- and

labour-intensive. Furthermore, although some studies do perform *in vitro* validation of candidate genes in human cell line models, this is likely to be of limited relevance[25] as cancer cell lines harbour many additional mutations and are grown in a highly artificial environment. Other *in vitro* models such as 3D tumour organoids[26,27] may provide an interesting alternative, but also lack a tumour microenvironment and need to be grown in specific media[28], which may limit the clinical translatability of findings in these models.

Finally, many promising targeted therapies fail in the clinic due to the emergence of treatment resistance. To understand why this is the case, it is important to determine how different therapies impact the clonal evolution of a tumour and how this leads to the development of treatment resistance. These insights can then be used to develop new strategies that aim to prevent or overcome resistance. However, detailed studies of clonal evolution and treatment resistance are challenging as the development of resistance is often a stochastic process, as is evident from the observation that patients can often develop multiple mechanisms of resistance to the same treatment[29–31]. Combined with the limited availability of pre- and post-treatment tumour samples from patients, this makes human studies less than ideal for determining the range of potential resistance mechanisms. To identify and predict these potential resistance mechanisms, we will therefore need experimental systems that allow us to quantify the range of expected resistance mechanisms for a given tumour and determine how these are impacted by different treatments or other factors such as diverse genetic backgrounds.

1.4 Experimental models of human cancer

1.4.1 Patient-derived models

Experimental models of human cancer should be easy to manipulate and recapitulate the genetic features and microenvironment of the original patient tumour as much as possible. Human cancer cell lines have often been used for this purpose, as these represent cell lines derived from patient tumours that are easy to manipulate *in vitro*. However, cell lines are grown in a highly artificial environment and therefore undergo a stringent selection process when being established, resulting in homogeneous populations that no longer fully represent the genetic heterogeneity of the original tumour[32]. More recently, 3D organoid models have been developed to overcome this limitation by growing cells in three-dimensional media, which allows the formation of more realistic organ-like structures[33]. This technique has enabled the development of *in vitro* models for tissues that could not be established as cell lines[28], suggesting

that organoids are subject to less evolutionary pressure and are therefore more likely to reflect the heterogeneity of the original patient tumours.

Human cell lines have been very popular in cancer research, as they remain relatively close to the human setting, whilst providing a convenient platform for studying cancer cell biology. As such, these *in vitro* models have proved instrumental in delineating key biological signalling pathways and in preclinical drug screening[34,35]. A drawback of human cell lines and organoids is however that they do not model interactions with the tumour microenvironment and the effects of angiogenesis and drug metabolism. To address these limitations, cell lines and organoids can be injected into immune-deficient mice (Figure 1.2A) to create *in vivo* xenograft models. However, although these xenograft models do capture interactions between tumour cells and the (mouse) microenvironment, they do not recapitulate interactions with the immune system due to the use of immunocompromised mice.

An alternative approach is to directly transplant tissue from human tissue into immunodeficient host mice, thereby creating patient-derived tumour xenograft (PDTX) models. Compared to cell line- or organoid-based transplantation models, PDTX tumours more faithfully retain the molecular, genetic, and histological heterogeneity observed in the respective cancer patients, even after serial passaging in mice[36,37]. As such, PDTX models have been a popular textitin vivo platform for preclinical drug screening in a large variety of cancer types, such as breast cancer[38,39], melanoma[29,40] and colorectal cancer[41–45]. However, drawbacks of PDTX models are that certain tumour types are much harder to establish in mice than others, and xenografts may undergo mouse-specific tumour evolution[46]. Moreover, PDTX models still lack an active immune system. Humanized mouse models aim to address this gap, by introducing functional human immune systems into otherwise immunocompromised mice[47].

1.4.2 Genetically engineered mouse models (GEMMs)

A significant limitation of patient-derived models, is that they are typically established using heavily mutated, end-stage tumours and can therefore not be used to study the effects of individual mutations on tumour initiation and progression. In contrast, genetically engineered mouse models (GEMMs) can be used to introduce individual mutations identified from human sequencing projects into a clean genetic background, allowing detailed characterization of these mutations and their effects on cancer susceptibility, tumour formation, progression and maintenance.

The first GEMMs were developed by introducing cloned cancer genes into the genome of transgenic mice (Figure 1.2B), providing the first conclusive evidence that mice

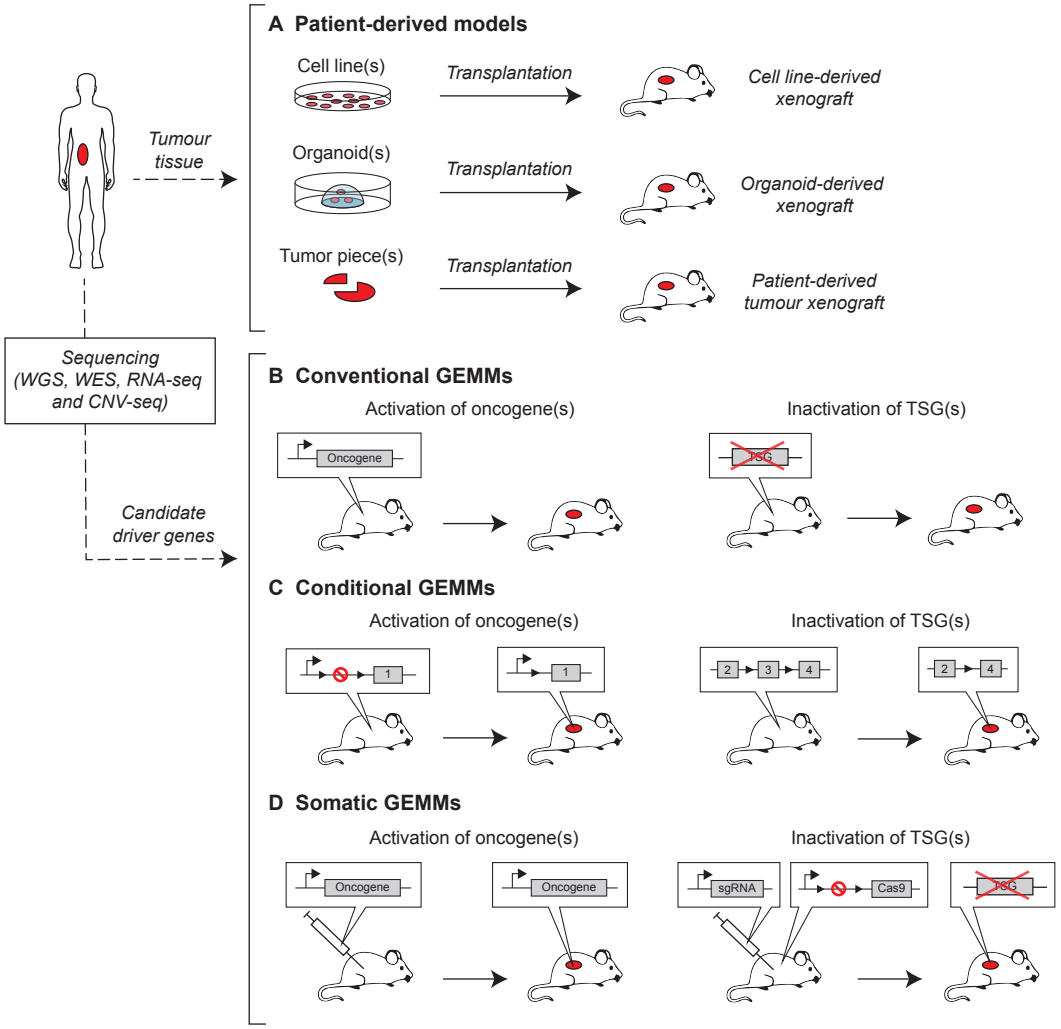


Fig. 1.2. Schematic overview of different types of mouse models of human cancer. **(A)** Patient-derived models are created by transplanting human material into immune-deficient mice. This can be done by either injecting cell lines or organoids which were previously derived from human tissue, or by directly grafting human tumour pieces into mice. **(B)** In conventional GEMMs, *de novo* tumourigenesis is induced either by tissue-specific expression of an oncogene or by germline inactivation of a TSG. The engineered genes are typically selected based on pre-existing data from (human) sequencing studies. **(C)** In conditional GEMMs, *de novo* tumourigenesis is induced by tissue-specific inactivation of conditional TSG alleles and/or activation of conditional oncogenes via Cre-loxP-mediated recombination. **(D)** In somatic GEMMs, tumourigenesis is induced either by injecting lentiviral vectors expressing specific oncogene(s) into the tissue of interest, or by injecting Cas9 together with sgRNAs targeting specific TSGs. In the latter approach, Cas9 can also be expressed conditionally in the host mouse.

could be made prone to developing tumours in a specific tissue by introducing transgenic expression of an oncogene under control of a tissue-specific promoter[48]. Later, with the rise of gene-targeting technology, the effects of mutations in tumour suppressor genes (TSGs) on tumour formation could be studied in knockout mice[49]. However, a significant limitation of these conventional GEMMs is that oncogenes are expressed in all cells of a particular tissue in transgenic mice, whilst TSGs in

knockout mice are inactivated in all cells. In this respect, conventional GEMMs fail to mimic sporadic cancers, in which the accumulation of genetic events in a single cell results in tumourigenesis in an otherwise healthy organ.

To address this issue, conditional GEMMs were developed by employing somatic activation of oncogenes and somatic inactivation of tumour suppressors[50] (Figure 1.2C). One of the most frequently used conditional strategies is the Cre/loxP recombinase system[51], in which parts of target genes are flanked by loxP recombinase recognition sites that recombine in the presence of Cre-recombinase to delete intervening DNA sequences. Using this system, oncogenes can be activated by removing engineered stop sequences that prevent gene expression in the absence of the recombinase, whereas TSGs can be inactivated by deleting exons that are crucial for gene function. Conditional GEMMs have been developed for a large variety of different cancers, generating a wealth of models that closely mimic the histopathological, molecular, and clinical features of human tumours[52,53].

A limitation of conditional GEMMs is that generating new models is still time-consuming and expensive. Recent developments in somatic gene-editing techniques provide incredible potential to speed-up this process by allowing mutations to be introduced somatically into existing mouse models (Figure 1.2D). Using these approaches, oncogenes can be introduced by injecting (viral) vectors expressing these gene(s) into the tissue of interest[54]. Similarly, TSGs can be inactivated using CRISPR-Cas9-mediated gene editing by injecting constructs containing Cas9 and guide RNAs (sgRNAs) targeting the TSGs into Cas9-proficient mice[55]. Further developments such as CRISPR interference/activation[56] and CRISPR-mutagenesis approaches[57,58] promise to further expand this toolkit, enabling the use of somatic engineering to rapidly model a wide variety of cancer-associated mutations.

1.5 Identifying cancer drivers in mouse models

1.5.1 Mouse tumour sequencing

Besides studying the effects of candidate cancer drivers, GEMMs can also be used to identify additional driver mutations by sequencing mouse tumours and identifying additional genes that are frequently mutated across tumours (Figure 1.3A). Following this approach, driver mutations can be detected using various computational approaches, in the same fashion as previously described for human tumours. One of the main advantages of using mouse tumour sequencing to identify additional drivers is that, by sequencing tumours from a specific and controlled genetic background, we can specifically identify drivers that collaborate with the pre-existing

driver mutations. As such, whole-exome and whole-genome sequencing approaches have been used to characterize the mutational landscapes of *Kras*-mutant mouse skin squamous cell carcinoma[59] and *Egfr*-, *Myc*- and *Kras*-driven lung cancers[60,61]. Similarly, copy number-based approaches have identified several driver genes in mouse models of *Brca1*- and *Brca2*-deficient breast cancer[62]. A potential drawback of using mouse tumour sequencing to identify driver mutations is that (depending on the pre-existing drivers) mouse tumours may harbour as many aberrations as human tumours, complicating the identification of driver genes. This is, for example, particularly the case for tumours with high copy number instability, such as *Brca1/Brca2*-deficient breast cancer models[62]. Combined with the relatively low numbers of tumours involved in mouse studies, this means that mouse sequencing studies may be relatively underpowered for identifying candidate driver genes.

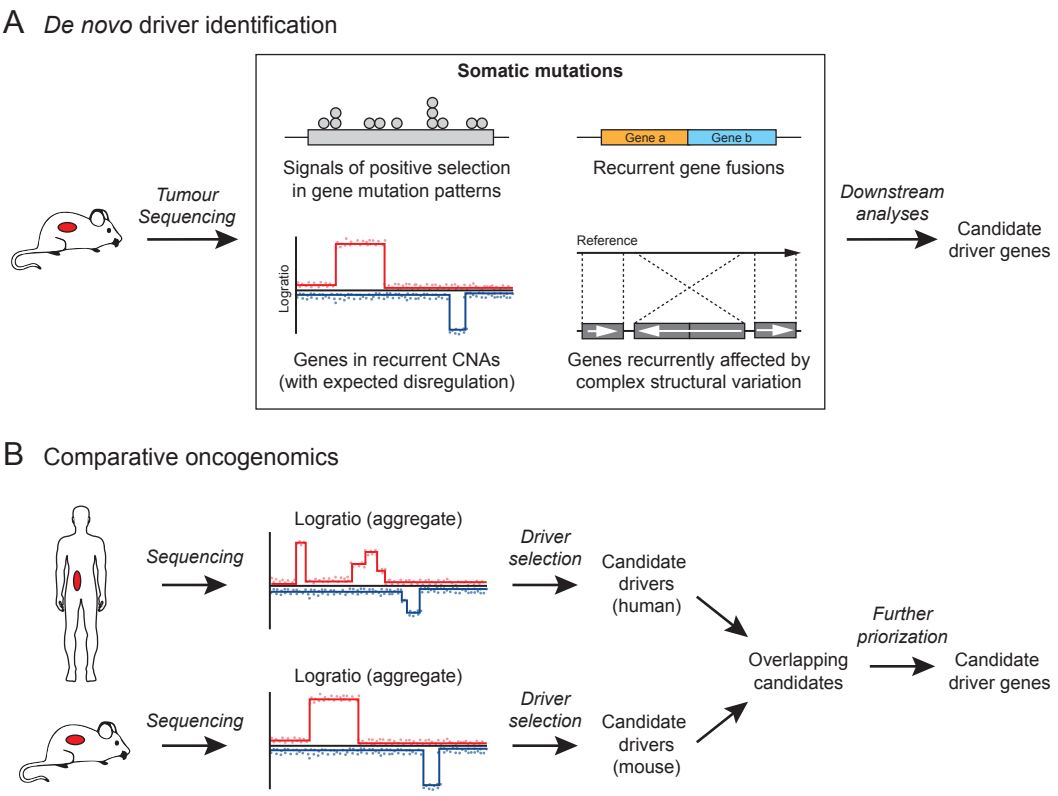


Fig. 1.3. ***De novo* driver gene identification in mice. (A)** Driver genes can be identified in mice in the same fashion as for human tumours by using DNA/RNA-sequencing approaches aimed at identifying recurrent mutations, copy number aberrations, gene fusions and complex structural rearrangements. **(B)** Comparative oncogenomics approaches allow refinement of candidate driver gene lists by focusing on genes that are recurrently mutated in both mouse and human tumours. In such an approach, candidates are typically first identified for both species individually, after which shared (orthologous) genes are selected and further prioritized based on existing knowledge or other data sources.

1.5.2 Comparative oncogenomics

A more powerful approach for identifying driver genes using sequencing approaches is to combine insights from mouse and human datasets and prioritize genes that are mutated in both species, as these are most likely to represent true driver genes. This can be done in an ad hoc setting, by identifying drivers of mouse tumours and comparing these with known mutations in human datasets, or as a deliberate strategy using comparative oncogenomics. In the latter approach, sequencing data from mouse and human tumours are typically first analyzed to identify candidate driver genes for both species individually. These species-specific candidates are then integrated by only selecting genes and/or networks that are aberrated in both species (Figure 1.3B). Remaining candidates can optionally be filtered using additional criteria, such as correlation with gene expression or prior knowledge from literature.

This comparative strategy has proven particularly effective for distinguishing driver genes from passengers in chromosomally unstable tumours. For example, in a mouse model of hepatocellular carcinoma (HCC), Zender *et al.* identified a focal amplicon on mouse chromosome 9qA1, which was syntenic with amplifications in human HCCs on 11q22[63]. Further filtering based on expression identified two drivers on this locus, *cIAP1* and *Yap*, which were shown to act synergistically in tumourigenesis. Similarly, copy number sequencing of metastases from an inducible *Hras* model of a traditionally non-metastatic melanoma identified a focal amplification on mouse chromosome 16, which contained only eight candidate driver genes[64]. Further comparison with human *RAS*- and *MET*-driven melanomas identified a single gene, *NEDD9*, as the driver of these metastases.

1.6 Identifying drivers using forward genetic screening

Although additional driver genes can be identified by mouse tumour sequencing, this approach is not always optimal as many mouse models have a long tumour latency and may be more prone to acquire other types of mutations than those of interest (i.e., copy number aberrations rather than point mutations). Forward genetic screening approaches can address these problems by using various mutagenesis strategies to induce additional mutations and accelerate tumour formation, after which any tumours that developed can be studied to identify new drivers. The type of mutations that occur depends strongly on the used type of mutagenesis, meaning that different mutagenesis strategies can used to specifically model different kinds of mutations.

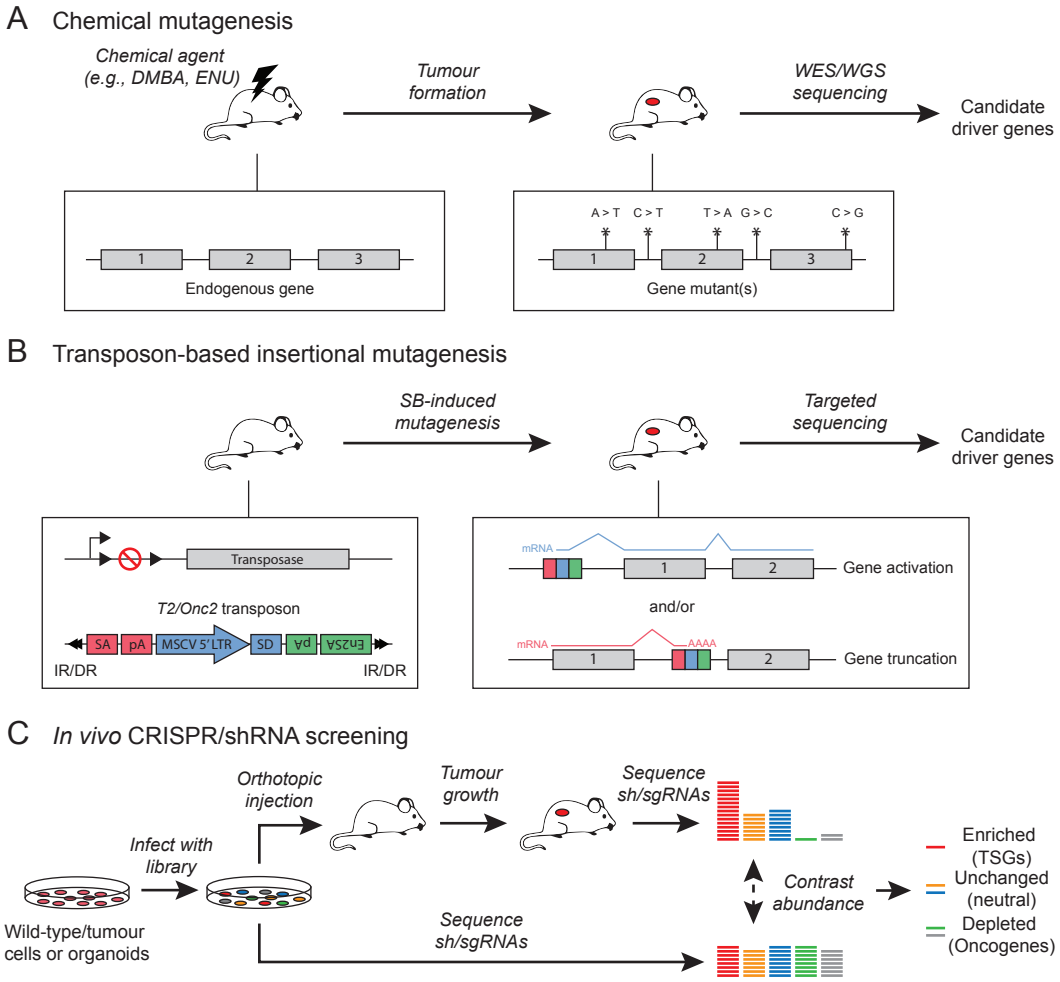


Fig. 1.4. Schematic overview of forward genetic screening approaches. (A) In chemical mutagenesis approaches, mice are treated with a mutagenic compound and subsequently monitored for tumour formation. The resulting tumours can then be sequenced using whole-genome or targeted sequencing to identify mutations driving tumourigenesis. **(B)** In transposon-based insertional mutagenesis (TIM), tissue-specific expression of a transposase enzyme induces the mobilization of mobile elements called transposons, which can be re-integrated elsewhere in the genome. By doing so, transposons can result in the activation of oncogenes or inactivation of TSGs. **(C)** With *in vivo* CRISPR/shRNA screening, cells or organoids are transduced with a library of sh/sgRNAs targeting a set of genes. The transduced cells/organoids are injected *in vivo*, after which the mice are monitored for tumour formation. Developed tumours are sequenced to determine the abundance of individual sh/sgRNAs, which is contrasted to the starting population to identify if sh/sgRNAs targeting specific genes are enriched (potential TSGs) or depleted (potential oncogenes and/or drug targets).

1.6.1 Screening using chemical mutagenesis

Chemical-based mutagenesis is one of the oldest mutagenesis strategies, in which cells or animals are treated with a chemical substance that damages the DNA and thereby induces mutations. The induced mutations are typically single nucleotide changes; however, the spectrum of these mutations depends on the used substance and can vary greatly between different chemicals. During *in vivo* chemical mutagenesis screens, animals (typically zebrafish or mice) are treated with a controlled dose of a chemical mutagen and subsequently monitored for tumour formation (Figure 1.4A). Developed tumours can be sequenced using whole-genome or targeted sequencing strategies to identify mutations that may be driving tumourigenesis or metastasis[60,65].

An advantage of chemical mutagenesis compared to other mutagenesis approaches, is that its bias towards single nucleotide changes makes it suitable for modelling the effects of human variants, which are often single point mutations that result in changes in the levels of expression or activity of a gene product. For this reason, chemical mutagenesis approaches have been used to mimic human mutational processes and to characterize the genomic landscapes of the mutational landscapes of mouse skin squamous cell carcinoma[59] and *Kras*-driven lung cancers[60]. However, when designing screens, it is important to take the inherent mutational bias of chemicals into account. Currently, N-ethyl-N-nitrosourea (ENU) is a popular choice for chemical mutagenesis strategies aiming to model single-nucleotide variants (SNVs), as it results in a range of point mutations that mirrors the range of mutations observed in human tumours[66].

1.6.2 Screening using insertional mutagenesis

A drawback of chemical mutagenesis strategies is that their bias towards point mutations limits their utility for modelling other types of mutations, such as increased gene expression resulting from gene amplifications. Insertional mutagenesis (IM) strategies provide an alternative approach, in which viral or transposon sequences are stochastically inserted into the genome, disrupting the expression of nearby genes (Figure 1.4B). In transposon insertional mutagenesis (TIM) strategies, this process is mediated by a transposase enzyme, which excises transposons from a concatemer located in the genome of the mice and reintegrates them stochastically elsewhere. By placing the expression of this transposase under a tissue-specific promoter, mutagenesis can be restricted to specific tissues in the mouse.

The effects of insertions depend on the used transposon, but typically involve the activation of oncogenes using promoter sequences and/or inactivation of tumour

suppressors by truncating genes. For example, the *T2Onc/2* transposon, which is frequently used in *Sleeping Beauty* IM screens[67,68], contains enhancer/promoter (MSCV) and splice donor (SD) sequences that allow the transposon to initiate transcription and drive the (over)expression of nearby genes. The transposon also contains two splice acceptor sites (SA/En2SA) combined with a bi-directional polyA sequence, which allow the transposon to truncate transcripts if integrated within a gene. Depending on the gene and the relative location of the insertion, these truncations can inactivate genes by resulting in an unstable transcript or inactive protein, or activate genes by removing inhibitory protein domains[69,70].

A considerable advantage of TIM strategies is that insertions can be specifically targeted via PCR amplification before sequencing, enabling cheap and efficient retrieval of the insertion sites compared to genome-wide sequencing. However, a drawback of transposon-based systems is that they generally show bias in terms of their insertion patterns, either due to sequence integration biases or biases towards specific gene features (e.g., gene bodies or promoters)[71]. For this reason, screens using different transposon systems (such as the *Sleeping Beauty*[67,68] or *piggyBac*[72,73] systems) may identify different candidate genes, even if screens are performed in the same genetic background. Moreover, target genes are typically identified using windows around the insertion sites[71,74], which may lead to the identification of many false positive candidate genes.

Despite these drawbacks, TIM has been valuable for identifying cancer-associated driver genes in mouse models of a large variety of cancer types, including breast cancer[75–77], melanoma[78], hepatocellular carcinoma[79] and gastric cancer[80]. Additionally, as mutagenesis remains constitutively active in these models, TIM has also been used to identify drivers of metastasis formation[81] and acquired resistance to drug treatments[82,83]. Finally, new computational approaches based on RNA-sequencing data have been developed to improve target gene prediction and offer additional insight into the effects of insertions on the expression of the affected gene[84,85].

1.6.3 shRNA screening

In contrast to the previously described genome-wide screening approaches, library-based screening approaches, such as loss-of-function screens based on RNA interference (RNAi) technology, can be used to target specific sets of genes. In pooled RNAi screening approaches, cells are transfected with short hairpin RNAs (shRNAs) targeting specific genes that, when integrated into the genome, result in heritable and long-term suppression of the corresponding gene[86]. To perform an *in vivo* RNAi screen, the transfected cells can be injected orthotopically in animals[87], which are monitored for tumour growth (Figure 1.4C). Once developed, tumours are harvested

and sequenced to quantify the frequency of each shRNA in the tumour cell population. By contrasting these frequencies to those of the initial starting population, this approach can identify which shRNAs are enriched in the tumour and are therefore likely targeting TSGs whose loss is beneficial for tumourigenesis. Conversely, depleted shRNAs may identify potential oncogenes and/or genes that are crucial for tumour maintenance.

The scope of shRNA screens depends entirely on the used library, meaning that screens can be designed to target all genes in a genome-wide fashion or to test a small number of pre-selected candidate genes. As such, shRNA screens can not only be used to identify novel driver genes, but can also be used for narrowing down lists of potential drivers or to validate putative driver genes. Compared to *in vitro* approaches, *in vivo* shRNA screens provide the opportunity to expose the vulnerabilities of tumour cells in the context of their microenvironment and can be used to study drivers of metastasis and therapy resistance[88–90]. Drawbacks of shRNA screening include variable efficiency between shRNAs in the knockdown of their respective target genes[91] and off-target effects[92]. Successful *in vivo* shRNA screens have been reported for a variety of cancer types, including xenograft models of hepatocellular carcinoma[93], lymphoma[94,95], leukemia[96,97] and glioma[98,99].

1.6.4 CRISPR screening

With the development of CRISPR–based technologies, it has also become possible to perform pooled loss-of-function screens by knocking-out genes using CRISPR–Cas9-mediated genome editing. CRISPR–Cas9 loss-of-function screens are generally performed in the same fashion as shRNA screens by transducing cells with pools of single guide RNAs (sgRNAs) targeting different genes, injecting the transduced cells *in vivo* and contrasting the abundance of sgRNAs in tumours with their abundance in the starting population[100–102]. However, in contrast to shRNA screening, gene editing via CRISPR-Cas9 disrupts the genes by DNA cleavage and thereby introducing insertions/deletions in their genomic sequence, resulting in frameshifts that induce heterozygous or homozygous knockout of genes rather than a reduction in expression.

Compared to shRNA screens, CRISPR-based screens have been reported to be remarkably efficient and suffer less from off-target effects than shRNA screens[103,104]. The approach is less amenable to studying dosage-dependent effects as genes are knocked-out rather than knocked-down, although dosage reduction can be achieved if Cas9 only induces heterozygous loss of the gene, as we have previously observed in the validation of candidates from a *Sleeping Beauty* IM screen[77]. With the development of new technologies, CRISPR-based screening approaches are extending

beyond loss-of-function screens by enabling gene activation using CRISPRa[105], gene inhibition using CRISPRi[106] and the introduction of mutations with CAS9-AID or CAS9-APOBEC base editors[57,58].

1.7 Validating and characterizing candidate driver genes

After identifying putative driver genes, it remains important to verify that these genes actually contribute to tumourigenesis. In many studies, human tumour cell lines are used for this purpose by either knocking-out/down candidate TSGs or (over)expressing candidate oncogenes and studying the effects of these perturbations on tumour cell growth. However, as previously described, these *in vitro* models suffer from several drawbacks, including the presence of additional mutations and/or the lack of a tumour microenvironment. An alternative approach is to use GEMMs to validate driver genes by introducing the corresponding mutations into a mouse model with a similar genetic background as in which the mutations were identified, and studying their effect on tumourigenesis (Figure 1.5A). These additional mutations can be introduced either using germline approaches (such as the GEMM-ESC strategy)[107] or somatic approaches (e.g., via injection of viral vectors), as described in previous sections.

Once established, the resulting mouse model(s) can be studied to determine how the added mutation(s) affect tumour formation compared to the baseline mouse model. For validating putative driver genes, the key question is whether the new model containing the additional mutation develops tumours with a higher incidence and/or shorter latency, indicating that the extra mutation does indeed contribute towards tumourigenesis. The new model can also be used to study the effects of the extra mutation on tumour metastasis and therapy response. Any developed tumours can be studied in detail to determine how the added mutation affects various tumour characteristics, such as the morphology of the tumours and interactions with the microenvironment and immune system. Additionally, by using sequencing strategies or screening approaches to identify additional driver genes in these more complex mouse models, mutational landscapes can be compared between models with different drivers to determine how additional drivers affect the evolution of tumours initiated by the engineered mutations, and if this provides clues to any driver-specific vulnerabilities.

A key advantage of using mouse models to validate candidate driver genes is that they can provide unambiguous proof of whether candidates are bona fide driver genes in a given genetic background. Especially CRISPR-CAS9-based somatic cancer modelling

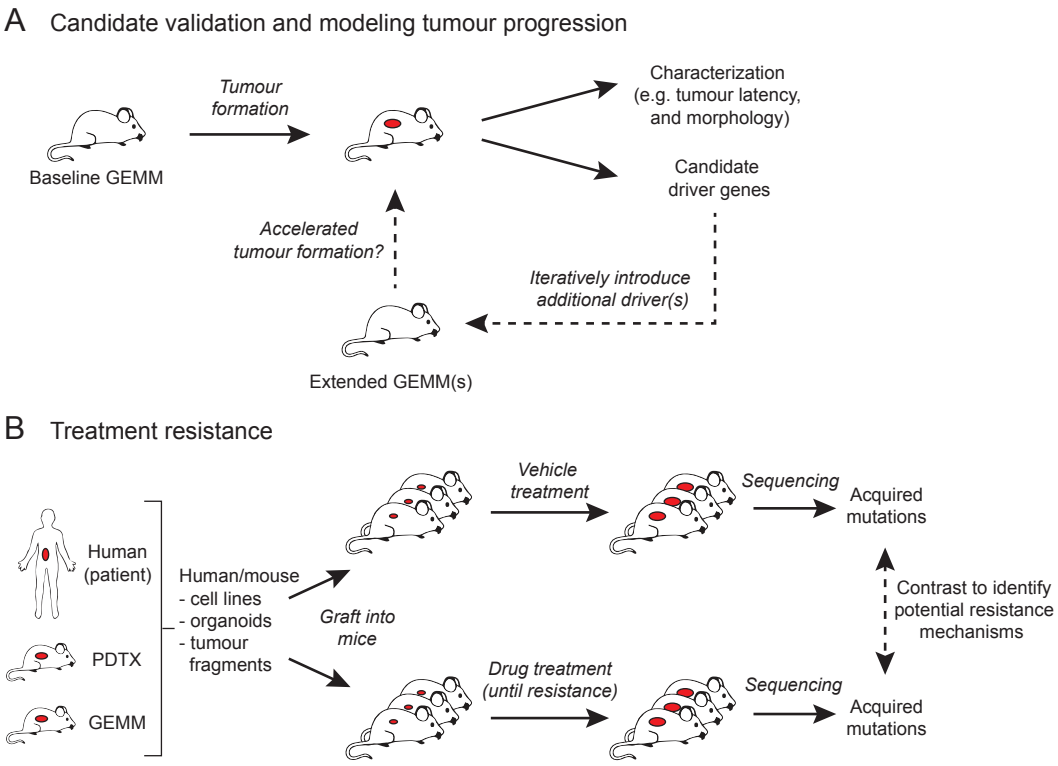


Fig. 1.5. Validating candidate genes and studying treatment resistance.
(A) Candidate cancer driver genes can be validated *in vivo* by developing mouse models containing the observed mutations and monitoring the mice for tumour formation. Developed tumours can be characterized to determine if they reflect the expected phenotype(s) and sequenced to identify additional candidate driver genes. These additional candidates can be introduced into the same baseline GEMM model to determine their effect on tumourigenesis. By applying this process iteratively, this type of approach can be used to unravel the contribution of different cancer genes during various stages of tumourigenesis. **(B)** Mouse models can also be used to identify mechanisms of acquired therapy resistance by transplanting cell lines, organoids or tumour pieces into multiple recipient mice and subjecting these mice to different treatments. Tumours that become resistant to treatment can then be sequenced and contrasted to vehicle-treated tumours to identify acquired mutations driving the resistance phenotype.

enables rapid *in vivo* testing of (combinations of) candidate cancer genes and has been used to validate driver genes for a wide range of cancer types, including breast cancer[55], colorectal cancer[108], pancreatic cancer[109] and liver cancer[110]. Additionally, multiplexed somatic engineering approaches provide the opportunity to rapidly validate multiple candidate genes at the same time, whilst simultaneously studying Darwinian selection between the different candidates and how this selection is influenced by cellular/tissue contexts and pre-existing mutations[110,111]. Finally, using iterative approaches, drivers can be identified and introduced progressively into mouse models of increasing complexity. This type of approach can be used to study tumour formation and progression in detail and establish the contributions of different driver genes at various stages of tumourigenesis[112].

1.8 Studying drug response and treatment resistance

Ultimately, knowledge of driver genes and their effects is used to develop novel therapeutic strategies that target specific vulnerabilities of the tumour, enabling effective treatments with minimal side effects. Following this premise, personalized therapies are generally designed by either targeting the identified driver genes directly (if possible) or by targeting other genes in the same signalling pathway. A well-known example is *BRAF*-mutant melanoma, which is targeted by inhibiting the mutant BRAF kinase and/or MEK, which is downstream of BRAF in the RAS/MAPK signalling pathway. Alternatively, tumours can be targeted therapeutically by exploiting a synthetic lethality resulting from the driver mutation(s). A classic example of such a synthetic lethality is PARP inhibition in BRCA-deficient tumours, which specifically targets cells with defects in their homologous recombination (HR) pathway due to loss HR-associated proteins such as BRCA1 and BRCA2.

Before moving to the clinic, drugs are generally first tested for efficacy in a preclinical setting, either using *in vitro* models (cell lines, organoids) or *in vitro* models (xenograft models, GEMMs, PDTXs). To identify which treatments are most effective in different cancer types or tumours with different genetic backgrounds, several efforts have been made to setup large biobanks of PDTX models for high-throughput drug screening purposes[39,42,113,114]. By correlating treatment sensitivity with sequencing data from the same tumours, these approaches can also be used to identify genetic markers of intrinsic (pre-existing) therapy resistance. A nice example of this approach has been given by Bertotti *et al.*, who identified HER2 amplification to be driving resistance in a subset of cetuximab-resistant colorectal PDTX tumours and showed that combined inhibition of HER2 and EGFR induced overt, long-lasting tumour regression[42].

Besides intrinsic therapy resistance, many targeted therapies fail in the clinic due to the emergence of drug resistance which is acquired during treatment. As such, a key challenge for improving the efficacy of these therapies is to identify (and ideally pre-empt) (epi)genetic changes that underlie this acquired treatment resistance. Both GEMMs and PDTXs can be used to identify potential *in vivo* resistance mechanisms by grafting cell lines, organoids or tumour fragments into multiple recipient mice; which are then subjected to different treatments (Figure 1.5B). Upon relapse, resistant tumours can be sequenced and compared with tumours from vehicle-treated mice to identify possible resistance mechanisms. Using this approach, our lab identified several resistance mechanisms to PARP-inhibitor treatment in mouse models of *BRCA1*- and *BRCA2*-deficient breast cancer[115–117]. This type of

approach can also be combined with various mutagenesis strategies, for example by using insertional mutagenesis to induce resistance and identify potential resistance mechanisms[82,83,118].

Besides identifying potential resistance mechanisms, an important challenge is to determine how resistance actually arises and to design therapeutic strategies accordingly. For example, in cases where tumours acquire additional (epi)genetic changes during treatment (as described above), therapies should ideally be designed to pre-empt and prevent the most likely paths of resistance. On the other hand, cases where resistance is driven by pre-existing sub-populations of intrinsically resistant cells[119] will require different treatment strategies. Traditionally, studying intra-tumour heterogeneity has been challenging with bulk-sequencing technologies. Single-cell sequencing approaches[120,121] promise to revolutionize these analyses by providing detailed insight into the (transcriptional) heterogeneity of tumour cells, enabling the identification of sub-populations of cells that may be driving resistance[122]. Furthermore, approaches such as lineage tracing can be used to track the dynamics of tumour evolution, providing detailed insight into which cell populations expand and contract during treatment. As such, lineage tracing-based approaches have been used to identify origins of resistance in mouse models of squamous cell carcinoma[123], prostate cancer[124] and mouse intestinal adenomas[125].

1.9 Conclusions and future perspectives

The success of personalized anti-cancer therapies hinges on how accurately we can predict whether a given patient tumour will respond to a given treatment, allowing clinicians to select the most effective therapeutic strategy for treating a patient. Ideally such an approach would be implemented by feeding omics data and other data types (e.g. imaging, pathology) from patient tumours into (computational) models that predict which therapies are most likely to succeed based on specific tumour biomarkers (Figure 1.6). Creating such models requires detailed insight into which mutations are driving tumour development and how these affect therapy response. Combined with high-throughput drug screening approaches, *in vitro* and *in vivo* model systems provide crucial platforms for assessing sensitivity to different therapies across multiple cell lines or tumours, enabling the construction of correlative models that predict the efficacy of these treatments for new tumours. More detailed genetic modelling in (mouse) model systems can further refine these correlative models by providing causative evidence that (combinations of) mutations drive cancer development and/or affect therapy response, and by enabling detailed characterization of the effects of drivers on other tumour phenotypes, such as tumour latency, morphology, mutational landscape and interactions with the (immune)

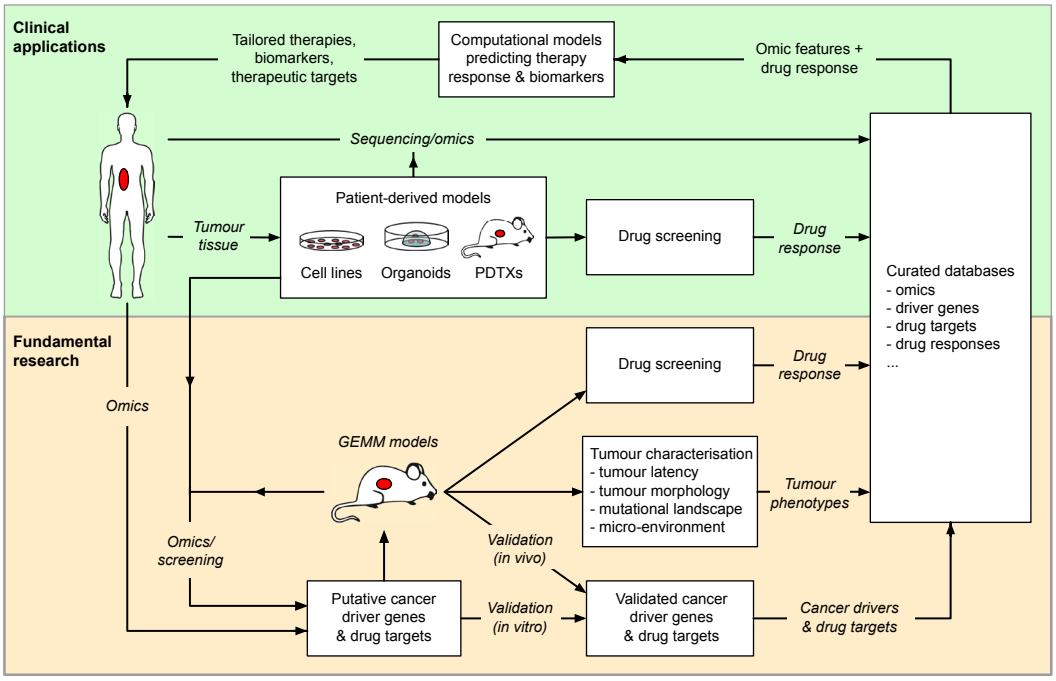


Fig. 1.6. The roles of model systems in designing personalized combination therapies for effective cancer treatment. This figure illustrates how human/mouse model systems are ultimately used to identify and characterize cancer driver genes in different types of cancer, and how these insights can be used to inform patient treatment. This process can essentially be divided into two parts: clinical application and fundamental research. On the clinical side, approaches to personalized medicine sequence patient tumours to identify mutations and potential cancer drivers. In parallel, tumour material can also be used to establish patient-derived model systems, which can be used to screen for drug efficacy and study treatment resistance. By combining this drug response data with mutations identified through sequencing, this approach can be used to train computational models predicting optimal therapies and identify biological features explaining the observed drug response. On the fundamental side, mouse models can also be used to identify cancer drivers (through sequencing and genetic screening) and for drug screening. In contrast to patient-derived models, GEMMs can also be used to conclusively validate cancer driver genes *in vivo* and to perform detailed dissection of how different driver genes affect tumour development and progression. This information can be used for predictive models in the clinic, but can also be used to formulate new hypotheses and additional experiments, fueling further fundamental research into the molecular underpinnings of cancer.

microenvironment. Genetically engineered mouse models also provide a powerful platform to critically evalutate new candidate drug targets[126] and thereby improve the robustness of preclinical cancer target identification[127].

Modeling of human cancer using genetically engineered mice is complicated by the observation that tumours generally contain multiple driver lesions, which can strongly influence their sensitivity to treatments targeting specific drivers. As a result, accurately assessing therapy response may require the generation of complex mouse models containing multiple driver genes that are frequently encountered together in a given type of cancer. Using germline engineering approaches, generating models with

multiple driver lesions has been challenging due to the extensive breeding and animal husbandry involved. Somatic approaches using CRISPR-CAS9 based gene editing and overexpression vectors alleviate this bottleneck, by providing the technology to quickly create new mouse models by introducing different combinations of mutations into a pre-existing baseline mouse model. The rapidly ongoing refinement of these tools will further expand the repetoire of mutations that can be modelled in this manner, enabling the rapid creation of new models reflecting the mutations observed in patient tumours, which can be used to test the effects of novel therapeutic strategies targeting these mutations.

Besides designing novel therapies, other important clinical challenges include identifying which patients are most at risk of developing cancer and should be screened for preventative treatments. Although factors such as genetic background and lifestyle have been shown to have a profound influence on cancer risk and survival, our insights into how these factors influence cancer development is still limited. Due to their tightly controlled genetics, mouse models are uniquely suited for examining the effects of genetic backgrounds and how these interact with specific driver genes in an *in vivo* setting. Similarly, mouse models can also be used to model the effects of specific lifestyles (e.g. diet, gut microbiome, circadian rithm, exposure to mutagens) on cancer risk and development[128]. Combined, insights from such models will hopefully allow us to incorporate knowledge of genetic modifiers and lifestyle influences into clinical tests, improving the identification of individuals at high risk for cancer and matching preventive therapies.

Improved screening imposes its own challenges, as population screening programmes identify many lesions that will not necessarily progress to cancer and do therefore not actually require treatment. Unfortunately, in many cases it is currently not possible to distinguish which lesions are indolent and which will progress to invasive cancer, leading to overdiagnosis and overtreatment[129]. Identifying which tumour cell-intrinsic and -extrinsic factors contribute to tumour progression will hopefully provide important insight into the tumourigenic process and allow us to develop tests that distinguish between high- and low-risk lesions. These studies will require models that allow us to study the early stages of cancer, which is not possible using end-stage tumour material from patients. GEMMs can provide a powerful platform for this type of research, as lesions in these models can be studied at any stage during tumour progression. Moreover, by introducing mutations identified in pre-malignant lesions from patients, mouse models can be used to determine the contributions of these mutations to tumour progression and to screen for additional factors that may be required for malignant transformation.

Tumour progression, metastasis and escape from therapy are phenomena which are driven by intra-tumour heterogeneity[130]. Single-cell sequencing approaches provide

a particularly promising approach for studying tumour progression by enabling detailed characterization of distinct cell populations within a given tumour[120,121]. For example, combined with (CRISPR-based) lineage tracing[131,132], single-cell approaches in GEMM tumours can be used to study the early dynamics of pre-malignant lesions and determine which cell populations play a role in driving eventual tumourigenesis. Similarly, longtitudinal sampling of mouse tumours may be used to determine how cell populations within tumours evolve during tumour progression and under therapeutic pressure, providing insight into how certain (epi)genetic changes may drive tumour evolution and the development of therapy resistance. Finally, detailed characterization of non-tumour cells within the tumour can be used to explore the interactions with the (immune) microenvironment[133,134] and how these interactions change during tumour progression or during therapy stress.

It is important to keep in mind is that tumours arising in GEMMs of human cancer may not necessarily reflect all characteristics of human tumours. For example, GEMM tumours may contain lower numbers of somatic mutations compared to the cognate human tumours[61] and fewer mutations seem to be required for cancer formation in mice compared to humans[135]. As such, it remains important to establish whether mouse tumours accurately reflect the aspects of the (sub)type of human cancer that we wish to study, both in terms of morphology, genetic landscape and transcriptional profile. Additionally, due to their limited genetic heterogeneity, it is unrealistic to expect that mouse models will sufficiently represent the heterogeneity of patient populations. It will therefore remain crucial to combine findings from mouse models with information from other sources, including sequencing data from patient populations and experimental data from other model systems, such as (human) tumour organoids and PDTX models.

Finally, efforts to collect catalogue mouse sequencing data have been relatively limited compared to efforts involving human sequencing studies. To fully exploit the large compendium of mouse sequencing and screening data, it will be important to collect these data in centralized repositories and create portals to query these data, allowing researchers to quickly explore existing datasets and compare tumour characteristics across different mouse models. Fortunately, several efforts are already underway to collect data in application-specific databases[136], to create portals visualizing data from PDTX models and to adapt software like cBioPortal[137] for visualizing tumour data from non-human organisms. We expect that these initiatives will play an important role in disseminating insights from mouse models and improve accessibility for cross-pollination with human sequencing efforts.

1.10 References

[1] Robert Weinberg. *The Biology of Cancer*. Garland science, 2013 (cit. on p. 2).

[2] Carlo M Croce. "Oncogenes and cancer". In: *New England Journal of Medicine* 358.5 (2008), pp. 502–511 (cit. on p. 2).

[3] Mel Greaves and Carlo C Maley. "Clonal evolution in cancer". In: *Nature* 481.7381 (2012), p. 306 (cit. on p. 3).

[4] Deborah R Caswell and Charles Swanton. "The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome". In: *BMC Medicine* 15.1 (2017), p. 133 (cit. on p. 3).

[5] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge". In: *Contemporary Oncology* 19.1A (2015), A68 (cit. on p. 3).

[6] John N Weinstein, Eric A Collisson, Gordon B Mills, et al. "The cancer genome atlas pan-cancer analysis project". In: *Nature Genetics* 45.10 (2013), p. 1113 (cit. on p. 3).

[7] International Cancer Genome Consortium et al. "International network of cancer genome projects". In: *Nature* 464.7291 (2010), p. 993 (cit. on p. 3).

[8] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, et al. "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups". In: *Nature* 486.7403 (2012), p. 346 (cit. on p. 3).

[9] Michael S Lawrence, Petar Stojanov, Paz Polak, et al. "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499.7457 (2013), p. 214 (cit. on p. 4).

[10] Nathan D Dees, Qunyuan Zhang, Cyriac Kandoth, et al. "MuSiC: identifying mutational significance in cancer genomes". In: *Genome Research* 22.8 (2012), pp. 1589–1598 (cit. on p. 4).

[11] Pauline C Ng and Steven Henikoff. "SIFT: Predicting amino acid changes that affect protein function". In: *Nucleic Acids Research* 31.13 (2003), pp. 3812–3814 (cit. on p. 4).

[12] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. "Predicting functional effect of human missense mutations using PolyPhen-2". In: *Current Protocols in Human Genetics* 76.1 (2013), pp. 7–20 (cit. on p. 4).

[13] Chengliang Dong, Peng Wei, Xueqiu Jian, et al. "Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies". In: *Human Molecular Genetics* 24.8 (2014), pp. 2125–2137 (cit. on p. 4).

[14] Abel Gonzalez-Perez and Nuria Lopez-Bigas. "Functional impact bias reveals cancer drivers". In: *Nucleic Acids Research* 40.21 (2012), e169–e169 (cit. on p. 4).

[15] David Tamborero, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. "OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes". In: *Bioinformatics* 29.18 (2013), pp. 2238–2244 (cit. on p. 4).

[16] Rameen Beroukhim, Craig H Mermel, Dale Porter, et al. “The landscape of somatic copy number alteration across human cancers”. In: *Nature* 463.7283 (2010), p. 899 (cit. on p. 4).

[17] Travis I Zack, Steven E Schumacher, Scott L Carter, et al. “Pan-cancer patterns of somatic copy number alteration”. In: *Nature Genetics* 45.10 (2013), p. 1134 (cit. on p. 4).

[18] Cheng-Zhong Zhang, Mitchell L Leibowitz, and David Pellman. “Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements”. In: *Genes & Development* 27.23 (2013), pp. 2513–2530 (cit. on p. 4).

[19] Ekta Khurana, Yao Fu, Dimple Chakravarty, et al. “Role of non-coding sequence variants in cancer”. In: *Nature Reviews Genetics* 17.2 (2016), p. 93 (cit. on p. 4).

[20] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, et al. “Comprehensive molecular portraits of invasive lobular breast cancer”. In: *Cell* 163.2 (2015), pp. 506–519 (cit. on p. 4).

[21] Cancer Genome Atlas Network et al. “Comprehensive genomic characterization of head and neck squamous cell carcinomas”. In: *Nature* 517.7536 (2015), p. 576 (cit. on p. 4).

[22] Cancer Genome Atlas Network et al. “Comprehensive molecular portraits of human breast tumours”. In: *Nature* 490.7418 (2012), p. 61 (cit. on p. 4).

[23] Cancer Genome Atlas Research Network et al. “Comprehensive genomic characterization of squamous cell lung cancers”. In: *Nature* 489.7417 (2012), p. 519 (cit. on p. 4).

[24] Magali Michaut, Suet-Feung Chin, Ian Majewski, et al. “Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer”. In: *Scientific Reports* 6 (2016), p. 18517 (cit. on p. 4).

[25] Jean-Pierre Gillet, Sudhir Varma, and Michael M Gottesman. “The clinical relevance of cancer cell lines”. In: *Journal of the National Cancer Institute* 105.7 (2013), pp. 452–458 (cit. on p. 5).

[26] Toshiro Sato, Robert G Vries, Hugo J Snippert, et al. “Single Lgr5 stem cells build crypt villus structures in vitro without a mesenchymal niche”. In: *Nature* 459.7244 (2009), pp. 262–265 (cit. on p. 5).

[27] Aliya Fatehullah, Si Hui Tan, and Nick Barker. “Organoids as an in vitro model of human development and disease”. In: *Nature Cell Biology* 18.3 (2016), p. 246 (cit. on p. 5).

[28] Jarno Drost, Wouter R Karthaus, Dong Gao, et al. “Organoid culture systems for prostate epithelial and cancer tissue”. In: *Nature Protocols* 11.2 (2016), p. 347 (cit. on p. 5).

[29] Kristel Kemper, Oscar Krijgsman, Paulien Cornelissen-Steijger, et al. “Intra-and inter-tumor heterogeneity in a vemurafenib-resistant melanoma patient and derived xenografts”. In: *EMBO Molecular Medicine* 7.9 (2015), pp. 1104–1118 (cit. on pp. 5, 6).

[30] David Quigley, Joshi J Alumkal, Alexander W Wyatt, et al. “Analysis of circulating cell-free DNA identifies multiclonal heterogeneity of BRCA2 reversion mutations associated with resistance to PARP inhibitors”. In: *Cancer Discovery* 7.9 (2017), pp. 999–1005 (cit. on p. 5).

[31] Jane Goodall, Joaquin Mateo, Wei Yuan, et al. “Circulating cell-free DNA to guide prostate cancer treatment with PARP inhibition”. In: *Cancer Discovery* 7.9 (2017), pp. 1006–1017 (cit. on p. 5).

[32] Jennifer L Wilding and Walter F Bodmer. “Cancer cell lines for drug discovery and development”. In: *Cancer Research* 74.9 (2014), pp. 2377–2384 (cit. on p. 5).

[33] Norman Sachs and Hans Clevers. “Organoid cultures for the analysis of cancer phenotypes”. In: *Current opinion in genetics & development* 24 (2014), pp. 68–73 (cit. on p. 5).

[34] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, et al. “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. In: *Nature* 483.7391 (2012), p. 603 (cit. on p. 6).

[35] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, et al. “Systematic identification of genomic markers of drug sensitivity in cancer cells”. In: *Nature* 483.7391 (2012), p. 570 (cit. on p. 6).

[36] Manuel Hidalgo, Frederic Amant, Andrew V Biankin, et al. “Patient-derived xenograft models: an emerging platform for translational cancer research”. In: *Cancer Discovery* 4.9 (2014), pp. 998–1013 (cit. on p. 6).

[37] Peter Eirew, Adi Steif, Jaswinder Khattra, et al. “Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution”. In: *Nature* 518.7539 (2015), p. 422 (cit. on p. 6).

[38] Elisabetta Marangoni, Anne Vincent-Salomon, Nathalie Auger, et al. “A new model of patient tumor-derived breast cancer xenografts for preclinical assays”. In: *Clinical Cancer Research* 13.13 (2007), pp. 3989–3998 (cit. on p. 6).

[39] Alejandra Bruna, Oscar M Rueda, Wendy Greenwood, et al. “A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds”. In: *Cell* 167.1 (2016), pp. 260–274 (cit. on pp. 6, 17).

[40] Maria Romina Girotti, Gabriela Gremel, Rebecca Lee, et al. “Application of sequencing, liquid biopsies, and patient-derived xenografts for personalized medicine in melanoma”. In: *Cancer Discovery* 6.3 (2016), pp. 286–299 (cit. on p. 6).

[41] Andrea Bertotti, Eniko Papp, Siân Jones, et al. “The genomic landscape of response to EGFR blockade in colorectal cancer”. In: *Nature* 526.7572 (2015), p. 263 (cit. on p. 6).

[42] Andrea Bertotti, Giorgia Migliardi, Francesco Galimi, et al. “A molecularly annotated platform of patient-derived xenografts (“xenopatients”) identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer”. In: *Cancer Discovery* 1.6 (2011), pp. 508–523 (cit. on pp. 6, 17).

[43] Shyam M Kavuri, Naveen Jain, Francesco Galimi, et al. “HER2 activating mutations are targets for colorectal cancer treatment”. In: *Cancer Discovery* 5.8 (2015), pp. 832–841 (cit. on p. 6).

[44] Sandra Misale, Ivana Bozic, Jingshan Tong, et al. “Vertical suppression of the EGFR pathway prevents onset of resistance in colorectal cancers”. In: *Nature Communications* 6 (2015), p. 8305 (cit. on p. 6).

[45] Eugenia R Zanella, Francesco Galimi, Francesco Sassi, et al. “IGF2 is an actionable target that identifies a distinct subpopulation of colorectal cancer patients with marginal response to anti-EGFR therapies”. In: *Science Translational Medicine* 7.272 (2015), 272ra12–272ra12 (cit. on p. 6).

[46] Uri Ben-David, Gavin Ha, Yuen-Yi Tseng, et al. “Patient-derived xenografts undergo mouse-specific tumor evolution”. In: *Nature Genetics* 49.11 (2017), p. 1567 (cit. on p. 6).

[47] Ryoji Ito, Takeshi Takahashi, Ikumi Katano, and Mamoru Ito. “Current advances in humanized mouse models”. In: *Cellular & Molecular Immunology* 9.3 (2012), p. 208 (cit. on p. 6).

[48] Douglas Hanahan, Erwin F Wagner, and Richard D Palmiter. “The origins of oncomice: a history of the first transgenic mice genetically engineered to develop cancer”. In: *Genes & Development* 21.18 (2007), pp. 2258–2270 (cit. on p. 7).

[49] Cathy A Finlay. “What the papers say: p53 Loss of Function: Implications for the Processes of Immortalization and Tumorigenesis”. In: *Bioessays* 14.8 (1992), pp. 557–560 (cit. on p. 7).

[50] Jos Jonkers and Anton Berns. “Conditional mouse models of sporadic cancer”. In: *Nature Reviews Cancer* 2.4 (2002), p. 251 (cit. on p. 8).

[51] Brian Sauer. “Inducible gene targeting in mice using the Cre/lox System”. In: *Methods* 14.4 (1998), pp. 381–392 (cit. on p. 8).

[52] Jessica C Walrath, Jessica J Hawes, Terry van Dyke, and Karlyne M Reilly. “Genetically engineered mouse models in cancer research”. In: *Advances in Cancer Research*. Vol. 106. Elsevier, 2010, pp. 113–164 (cit. on p. 8).

[53] Kristopher K Frese and David A Tuveson. “Maximizing mouse cancer models”. In: *Nature Reviews Cancer* 7.9 (2007), p. 654 (cit. on p. 8).

[54] Wen Bu, Li Xin, Michael Toneff, Lei Li, and Yi Li. “Lentivirus vectors for stably introducing genes into mammary epithelial cells in vivo”. In: *Journal of Mammary Gland Biology and Neoplasia* 14.4 (2009), p. 401 (cit. on p. 8).

[55] Stefano Annunziato, Sjors M Kas, Micha Nethe, et al. “Modeling invasive lobular breast carcinoma by CRISPR/Cas9-mediated somatic genome editing of the mammary gland”. In: *Genes & Development* 30 (2016), pp. 1470–1480 (cit. on pp. 8, 16).

[56] Antonia A Dominguez, Wendell A Lim, and Lei S Qi. “Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation”. In: *Nature Reviews Molecular Cell Biology* 17.1 (2016), p. 5 (cit. on p. 8).

[57] Alexis C Komor, Yongjoo B Kim, Michael S Packer, John A Zuris, and David R Liu. “Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage”. In: *Nature* 533.7603 (2016), p. 420 (cit. on pp. 8, 15).

[58] Yunqing Ma, Jiayuan Zhang, Weijie Yin, et al. “Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells”. In: *Nature Methods* 13.12 (2016), p. 1029 (cit. on pp. 8, 15).

[59] Dany Nassar, Mathilde Latil, Bram Boeckx, Diether Lambrechts, and Cédric Blanpain. "Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma". In: *Nature Medicine* 21.8 (2015), p. 946 (cit. on pp. 9, 12).

[60] Peter MK Westcott, Kyle D Halliwill, Minh D To, et al. "The mutational landscapes of genetic and chemical models of Kras-driven lung cancer". In: *Nature* 517.7535 (2015), p. 489 (cit. on pp. 9, 12).

[61] David G McFadden, Katerina Politi, Arjun Bhutkar, et al. "Mutational landscape of EGFR-, MYC-, and Kras-driven genetically engineered mouse models of lung adenocarcinoma". In: *Proceedings of the National Academy of Sciences* 113.42 (2016), E6409–E6417 (cit. on pp. 9, 21).

[62] Henne Holstege, Erik van Beers, Arno Velds, et al. "Cross-species comparison of aCGH data from mouse and human BRCA1-and BRCA2-mutated breast cancers". In: *BMC Cancer* 10.1 (2010), p. 455 (cit. on p. 9).

[63] Lars Zender, Mona S Spector, Wen Xue, et al. "Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach". In: *Cell* 125.7 (2006), pp. 1253–1267 (cit. on p. 10).

[64] Minjung Kim, Joseph D Gans, Cristina Nogueira, et al. "Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene". In: *Cell* 125.7 (2006), pp. 1269–1281 (cit. on p. 10).

[65] Melissa Q McCreery, Kyle D Halliwill, Douglas Chin, et al. "Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers". In: *Nature Medicine* 21.12 (2015), p. 1514 (cit. on p. 12).

[66] Jonathan S Brammeld, Mia Petljak, Inigo Martincorena, et al. "Genome-wide chemical mutagenesis screens allow unbiased saturation of the cancer genome and identification of drug resistance mutations". In: *Genome Research* 27.4 (2017), pp. 613–625 (cit. on p. 12).

[67] Adam J Dupuy, Keiko Akagi, David A Largaespada, Neal G Copeland, and Nancy A Jenkins. "Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system". In: *Nature* 436.7048 (2005), p. 221 (cit. on p. 13).

[68] Lara S Collier, Corey M Carlson, Shruthi Ravimohan, Adam J Dupuy, and David A Largaespada. "Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse". In: *Nature* 436.7048 (2005), pp. 272–276 (cit. on p. 13).

[69] Neal G Copeland and Nancy A Jenkins. "Harnessing transposons for cancer gene discovery". In: *Nature Reviews Cancer* 10.10 (2010), pp. 696–706 (cit. on p. 13).

[70] Michael B Mann, Nancy A Jenkins, Neal G Copeland, and Karen M Mann. "Sleeping Beauty mutagenesis: exploiting forward genetic screens for cancer gene discovery". In: *Current Opinion in Genetics & Development* 24 (2014), pp. 16–22 (cit. on p. 13).

[71] Johann de Jong, Jeroen de Ridder, Louise van der Weyden, et al. "Computational identification of insertional mutagenesis targets for cancer gene discovery". In: *Nucleic Acids Research* 39.15 (2011), e105–e105 (cit. on p. 13).

[72] Sheng Ding, Xiaohui Wu, Gang Li, et al. “Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice”. In: *Cell* 122.3 (2005), pp. 473–483 (cit. on p. 13).

[73] Roland Rad, Lena Rad, Wei Wang, et al. “PiggyBac transposon mutagenesis: a tool for cancer gene discovery in mice”. In: *Science* 330.6007 (2010), pp. 1104–1107 (cit. on p. 13).

[74] Aaron L Sarver, Jesse Erdman, Tim Starr, David A Largaespada, and Kevin AT Silverstein. “TAPDANCE: An automated tool to identify and annotate transposon insertion CISs and associations between CISs from next generation sequence data”. In: *BMC Bioinformatics* 13.1 (2012), p. 1 (cit. on p. 13).

[75] Roberto Rangel, Song-Choon Lee, Kenneth Hon-Kim Ban, et al. “Transposon mutagenesis identifies genes that cooperate with mutant Pten in breast cancer progression”. In: *Proceedings of the National Academy of Sciences* 113.48 (2016), E7749–E7758 (cit. on p. 13).

[76] Roberto Rangel, Liliana Guzman-Rojas, Takahiro Kodama, et al. “Identification of new tumor suppressor genes in triple-negative breast cancer”. In: *Cancer Research* 77.15 (2017), pp. 4089–4101 (cit. on p. 13).

[77] Sjors M Kas, Julian R de Ruiter, Koen Schipper, et al. “Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular breast carcinoma”. In: *Nature Genetics* 49.8 (2017), pp. 1219–1230 (cit. on pp. 13, 14).

[78] Michael B Mann, Michael A Black, Devin J Jones, et al. “Transposon mutagenesis identifies genetic drivers of Braf V600E melanoma”. In: *Nature Genetics* 47.5 (2015), p. 486 (cit. on p. 13).

[79] Takahiro Kodama, Justin Y Newberg, Michiko Kodama, et al. “Transposon mutagenesis identifies genes and cellular processes driving epithelial-mesenchymal transition in hepatocellular carcinoma”. In: *Proceedings of the National Academy of Sciences* 113.24 (2016), E3384–E3393 (cit. on p. 13).

[80] Haruna Takeda, Alistair G Rust, Jerrold M Ward, et al. “Sleeping Beauty transposon mutagenesis identifies genes that cooperate with mutant Smad4 in gastric cancer development”. In: *Proceedings of the National Academy of Sciences* 113.14 (2016), E2057–E2065 (cit. on p. 13).

[81] Branden S Moriarity, George M Otto, Eric P Rahrmann, et al. “A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and metastasis”. In: *Nature Genetics* 47.6 (2015), p. 615 (cit. on p. 13).

[82] Emilie A Chapeau, Agnieszka Gembarska, Eric Y Durand, et al. “Resistance mechanisms to TP53-MDM2 inhibition identified by in vivo piggyBac transposon mutagenesis screen in an Arf-/- mouse model”. In: *Proceedings of the National Academy of Sciences* 114.12 (2017), pp. 3151–3156 (cit. on pp. 13, 18).

[83] Daniele Perna, Florian A Karreth, Alistair G Rust, et al. “BRAF inhibitor resistance mediated by the AKT pathway in an oncogenic BRAF mouse melanoma model”. In: *Proceedings of the National Academy of Sciences* 112.6 (2015), E536–E545 (cit. on pp. 13, 18).

[84] Julian R de Ruiter, Sjors M Kas, Eva Schut, et al. “Identifying transposon insertions and their effects from RNA-sequencing data”. In: *Nucleic Acids Research* 45.12 (2017), pp. 7064–7077 (cit. on p. 13).

[85] Nuri A Temiz, Branden S Moriarity, Natalie K Wolf, et al. “RNA sequencing of Sleeping Beauty transposon-induced tumors detects transposon-RNA fusions in forward genetic cancer screens”. In: *Genome Research* 26.1 (2016), pp. 119–129 (cit. on p. 13).

[86] David Sims, Ana M Mendes-Pereira, Jessica Frankum, et al. “High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing”. In: *Genome Biology* 12.10 (2011), R104 (cit. on p. 13).

[87] Gaetano Gargiulo, Michela Serresi, Matteo Cesaroni, Danielle Hulsman, and Maarten Van Lohuizen. “In vivo shRNA screens in solid tumors”. In: *Nature Protocols* 9.12 (2014), p. 2880 (cit. on p. 13).

[88] Nirupa Murugaesu, Marjan Iravani, Antoinette Van Weverwijk, et al. “An in vivo functional screen identifies ST6GalNAc2 sialyltransferase as a breast cancer metastasis suppressor”. In: *Cancer Discovery* 4.3 (2014), pp. 304–317 (cit. on p. 14).

[89] Arnaud Duquet, Alice Melotti, Sonakshi Mishra, et al. “A novel genome-wide in vivo screen for metastatic suppressors in human colon cancer identifies the positive WNT-TCF pathway modulators TMED3 and SOX12”. In: *EMBO Molecular Medicine* 6.7 (2014), pp. 882–901 (cit. on p. 14).

[90] Ramona Rudalska, Daniel Dauch, Thomas Longerich, et al. “In vivo RNAi screening identifies a mechanism of sorafenib resistance in liver cancer”. In: *Nature Medicine* 20.10 (2014), p. 1138 (cit. on p. 14).

[91] Yu Shao, Chi Yu Chan, Anil Maliyekkel, et al. “Effect of target secondary structure on RNAi efficiency”. In: *RNA* 13.10 (2007), pp. 1631–1640 (cit. on p. 14).

[92] Shuo Gu, Lan Jin, Yue Zhang, et al. “The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing in vivo”. In: *Cell* 151.4 (2012), pp. 900–911 (cit. on p. 14).

[93] Lars Zender, Wen Xue, Johannes Zuber, et al. “An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer”. In: *Cell* 135.5 (2008), pp. 852–864 (cit. on p. 14).

[94] Anka Bric, Cornelius Miething, Carl Uli Bialucha, et al. “Functional identification of tumor-suppressor genes through an in vivo RNA interference screen in a mouse lymphoma model”. In: *Cancer Cell* 16.4 (2009), pp. 324–335 (cit. on p. 14).

[95] Claudio Scuoppo, Cornelius Miething, Lisa Lindqvist, et al. “A tumour suppressor network relying on the polyamine–hypusine axis”. In: *Nature* 487.7406 (2012), p. 244 (cit. on p. 14).

[96] Peter G Miller, Fatima Al-Shahrour, Kimberly A Hartwell, et al. “In vivo RNAi screening identifies a leukemia-specific dependence on integrin beta 3 signaling”. In: *Cancer Cell* 24.1 (2013), pp. 45–58 (cit. on p. 14).

[97] Corbin E Meacham, Lee N Lawton, Yadira M Soto-Feliciano, et al. “A genome-scale in vivo loss-of-function screen identifies Phf6 as a lineage-specific regulator of leukemia cell growth”. In: *Genes & Development* 29.5 (2015), pp. 483–488 (cit. on p. 14).

[98] Tyler E Miller, Brian B Liau, Lisa C Wallace, et al. “Transcription elongation factors represent in vivo cancer dependencies in glioblastoma”. In: *Nature* 547.7663 (2017), p. 355 (cit. on p. 14).

[99] Gaetano Gargiulo, Matteo Cesaroni, Michela Serresi, et al. “In vivo RNAi screen for BMI1 targets identifies TGF-β/BMP-ER stress pathways as key regulators of neural-and malignant glioma-stem cell homeostasis”. In: *Cancer Cell* 23.5 (2013), pp. 660–676 (cit. on p. 14).

[100] Sidi Chen, Neville E Sanjana, Kaijie Zheng, et al. “Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis”. In: *Cell* 160.6 (2015), pp. 1246–1260 (cit. on p. 14).

[101] Ophir Shalem, Neville E Sanjana, and Feng Zhang. “High-throughput functional genomics using CRISPR–Cas9”. In: *Nature Reviews Genetics* 16.5 (2015), p. 299 (cit. on p. 14).

[102] Hiroko Koike-Yusa, Yilong Li, E-Pien Tan, Martin Del Castillo Velasco-Herrera, and Kosuke Yusa. “Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library”. In: *Nature Biotechnology* 32.3 (2014), p. 267 (cit. on p. 14).

[103] Bastiaan Evers, Katarzyna Jastrzebski, Jeroen PM Heijmans, et al. “CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes”. In: *Nature Biotechnology* 34.6 (2016), p. 631 (cit. on p. 14).

[104] Ian Smith, Peyton G Greenside, Ted Natoli, et al. “Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map”. In: *PLoS Biology* 15.11 (2017), e2003213 (cit. on p. 14).

[105] Jeffry D Sander and J Keith Joung. “CRISPR-Cas systems for editing, regulating and targeting genomes”. In: *Nature Biotechnology* 32.4 (2014), p. 347 (cit. on p. 15).

[106] Luke A Gilbert, Matthew H Larson, Leonardo Morsut, et al. “CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes”. In: *Cell* 154.2 (2013), pp. 442–451 (cit. on p. 15).

[107] Ivo J Huijbers, Jessica Del Bravo, Rahmen Bin Ali, et al. “Using the GEMM-ESC strategy to study gene function in mouse models”. In: *Nature protocols* 10.11 (2015), p. 1755 (cit. on p. 15).

[108] Jatin Roper, Tuomas Tammela, Naniye Malli Cetinbas, et al. “In vivo genome editing and organoid transplantation models of colorectal cancer and metastasis”. In: *Nature Biotechnology* 35.6 (2017), p. 569 (cit. on p. 16).

[109] Roman Maresch, Sebastian Mueller, Christian Veltkamp, et al. “Multiplexed pancreatic genome engineering and cancer induction by transfection-based CRISPR/Cas9 delivery in mice”. In: *Nature Communications* 7 (2016), p. 10770 (cit. on p. 16).

[110] Julia Weber, Rupert Öllinger, Mathias Friedrich, et al. “CRISPR/Cas9 somatic multiplex-mutagenesis for high-throughput functional cancer genomics in mice”. In: *Proceedings of the National Academy of Sciences* 112.45 (2015), pp. 13982–13987 (cit. on p. 16).

[111] Zoë N Rogers, Christopher D McFarland, Ian P Winters, et al. “A quantitative and multiplexed approach to uncover the fitness landscape of tumor suppression in vivo”. In: *Nature Methods* 14.7 (2017), p. 737 (cit. on p. 16).

[112] Stefano Annunziato, Julian R de Ruiter, Linda Henneman, et al. “Comparative oncogenomics identifies combinations of driver genes and drug targets in BRCA1-mutated breast cancer”. In: *Nature Communications* 10 (2019), p. 397 (cit. on p. 16).

[113] Alberto Bardelli, Simona Corso, Andrea Bertotti, et al. “Amplification of the MET receptor drives resistance to anti-EGFR therapies in colorectal cancer”. In: *Cancer Discovery* 3.6 (2013), pp. 658–673 (cit. on p. 17).

[114] Hui Gao, Joshua M Korn, Stéphane Ferretti, et al. “High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response”. In: *Nature Medicine* 21.11 (2015), p. 1318 (cit. on p. 17).

[115] Janneke E Jaspers, Ariena Kersbergen, Ute Boon, et al. “Loss of 53BP1 causes PARP inhibitor resistance in Brca1-mutated mouse mammary tumors”. In: *Cancer Discovery* 3.1 (2013), pp. 68–81 (cit. on p. 17).

[116] Guotai Xu, J Ross Chapman, Inger Brandsma, et al. “REV7 counteracts DNA double-strand break resection and affects PARP inhibition”. In: *Nature* 521.7553 (2015), pp. 541–544 (cit. on p. 17).

[117] Ewa Gogola, Alexandra A Duarte, Julian R de Ruiter, et al. “Selective Loss of PARG Restores PARylation and Counteracts PARP Inhibitor-Mediated Synthetic Lethality”. In: *Cancer Cell* 33.6 (2018), pp. 1078–1093 (cit. on p. 17).

[118] Sjors M Kas, Julian R de Ruiter, Koen Schipper, et al. “Transcriptomics and transposon mutagenesis identify multiple mechanisms of resistance to the FGFR inhibitor AZD4547”. In: *Cancer Research* 78.19 (2018), pp. 5668–5679 (cit. on p. 18).

[119] Peter Valent, Dominique Bonnet, Ruggero De Maria, et al. “Cancer stem cell definitions and terminology: the devil is in the details”. In: *Nature Reviews Cancer* 12.11 (2012), p. 767 (cit. on p. 18).

[120] Charles Gawad, Winston Koh, and Stephen R Quake. “Single-cell genome sequencing: current state of the science”. In: *Nature Reviews Genetics* 17.3 (2016), p. 175 (cit. on pp. 18, 21).

[121] Serena Liu and Cole Trapnell. “Single-cell transcriptome sequencing: recent advances and remaining challenges”. In: *F1000 Research* 5 (2016) (cit. on pp. 18, 21).

[122] Alice Giustacchini, Supat Thongjuea, Nikolaos Barkas, et al. “Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia”. In: *Nature Medicine* 23.6 (2017), p. 692 (cit. on p. 18).

[123] Naoki Oshimori, Daniel Oristian, and Elaine Fuchs. “TGF-β promotes heterogeneity and drug resistance in squamous cell carcinoma”. In: *Cell* 160.5 (2015), pp. 963–976 (cit. on p. 18).

[124] Min Zou, Roxanne Toivanen, Antonina Mitrofanova, et al. “Transdifferentiation as a mechanism of treatment resistance in a mouse model of castration-resistant prostate cancer”. In: *Cancer Discovery* 7.7 (2017), pp. 736–749 (cit. on p. 18).

[125] Arnout G Schepers, Hugo J Snippert, Daniel E Stange, et al. “Lineage tracing reveals Lgr5+ stem cell activity in mouse intestinal adenomas”. In: *Science* 337.6095 (2012), pp. 730–735 (cit. on p. 18).

[126] Kelly Kersten, Karin E de Visser, Martine H van Miltenburg, and Jos Jonkers. “Genetically engineered mouse models in oncology research and cancer medicine”. In: *EMBO Molecular Medicine* 9.2 (2017), pp. 137–153 (cit. on p. 19).

[127] William G Kaelin Jr. “Common pitfalls in preclinical cancer target validation”. In: *Nature Reviews Cancer* 17.7 (2017), p. 425 (cit. on p. 19).

[128] Alexandra E Tammariello and John A Milner. “Mouse models for unraveling the importance of diet in colon cancer prevention”. In: *The Journal of Nutritional Biochemistry* 21.2 (2010), p. 77 (cit. on p. 20).

[129] Laura J Esserman, Ian M Thompson, Brian Reid, et al. “Addressing overdiagnosis and overtreatment in cancer: a prescription for change”. In: *The Lancet Oncology* 15.6 (2014), e234–e242 (cit. on p. 20).

[130] Nicholas McGranahan and Charles Swanton. “Clonal heterogeneity and tumor evolution: past, present, and the future”. In: *Cell* 168.4 (2017), pp. 613–628 (cit. on p. 20).

[131] Reza Kalhor, Prashant Mali, and George M Church. “Rapidly evolving homing CRISPR barcodes”. In: *Nature Methods* 14.2 (2017), p. 195 (cit. on p. 21).

[132] Mollie B Woodworth, Kelly M Girskis, and Christopher A Walsh. “Building a lineage from single cells: genetic techniques for cell lineage tracking”. In: *Nature Reviews Genetics* 18.4 (2017), p. 230 (cit. on p. 21).

[133] Yonit Lavin, Soma Kobayashi, Andrew Leader, et al. “Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses”. In: *Cell* 169.4 (2017), pp. 750–765 (cit. on p. 21).

[134] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, et al. “Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer”. In: *Nature Communications* 8 (2017), p. 15081 (cit. on p. 21).

[135] William C Hahn and Robert A Weinberg. “Modelling the molecular circuitry of cancer”. In: *Nature Reviews Cancer* 2.5 (2002), p. 331 (cit. on p. 21).

[136] Justin Y Newberg, Karen M Mann, Michael B Mann, Nancy A Jenkins, and Neal G Copeland. “SBCDDB: Sleeping Beauty Cancer Driver Database for gene discovery in mouse models of human cancers”. In: *Nucleic Acids Research* 46.D1 (2017), pp. D1011–D1017 (cit. on p. 21).

[137] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, et al. “The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data”. In: *Cancer Discovery* 2.5 (2012), pp. 401–404 (cit. on p. 21).

2

Scope of this thesis

2.1 Introduction

Breast cancer is the most common malignancy affecting women in the Western world, with more than 17,000 cases being diagnosed in the Netherlands alone each year*. Overall treatment of breast cancer is relatively successful, however recurrence of the disease remains a significant problem in clinical practice[1]. Breast cancer is known to be a heterogeneous disease and has therefore been subdivided into different subtypes, based on histological characteristics[2], gene expression patterns[3] and expression of different markers such as estrogen receptor-α (ER-α), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2, aka ERBB2), which are known to drive expression of downstream signaling pathways and increase cellular proliferation[4–6]. Genetic analyses have identified several common genetic alterations associated with specific subtypes[7–9], indicating that biological differences between the subtypes strongly shape tumor development.

2.2 A tale of two breast cancer subtypes

In the remainder of this thesis, we focus on two particular subtypes of human breast cancer: invasive lobular carcinoma and triple-negative breast cancer.

Invasive lobular carcinoma (ILC) is a histological subtype of breast cancer that represents 8-14% of all breast cancer cases. The classical form of ILC is characterized by rows of small discohesive cells, which invade into the surrounding stroma in a single-file pattern[10]. This invasive phenotype is generally attributed to the functional loss of E-cadherin (encoded by the *CDH1* gene), a cell-cell adhesion molecule that forms a key component of adherens junctions and plays an important role in maintaining epithelial integrity[11]. Functional loss of E-cadherin occurs in 90% of all ILCs and is mainly due to mutational inactivation, loss of heterozygosity (LOH) or impaired integrity of the components of the E-cadherin–catenin complex[9,12–14]. Besides this, ILCs are generally ER- and PR-positive and rarely show amplification of HER2. However, long-term outcome of ILC is generally worse than stage-matched invasive ductal carcinoma (IDC)[15], suggesting that biological differences between the two subtypes may be influencing treatment efficacy.

Triple-negative breast cancer (TNBC) is a heterogeneous subtype of breast cancer that is characterized by low expression of ER, PR and HER2. Altogether, TNBC accounts for 10-17% of all breast cancer cases, depending on the methods and thresholds used to assess the status of the three receptors[16]. At the mutational level, TNBCs are enriched for mutations in *TP53*[7] and *BRCA1*[17], which plays a key role in the

*www.cijfersoverkanker.nl

repair of DNA double-strand breaks (DSBs) via homologous recombination (HR). As such, BRCA-deficient TNBCs generally show high levels of chromosomal instability, which is attributed to the HR deficiency of these tumors[18]. Compared to other subtypes, TNBCs occur more frequently in younger patients (<50 years) and are significantly more aggressive[19,20]. Moreover, due to the lack of specialized therapies, chemotherapy currently remains standard-of-care for patients with TNBC[21], resulting in a relatively poor prognosis.

2.3 Identifying drivers of human (breast) cancer

A common feature of ILC and TNBC, is that they respond more poorly to existing therapies than other breast cancer subtypes. As such, both breast cancer subtypes would be benefited by the development of novel therapies targeting specific vulnerabilities in these tumors. Efforts to identify such vulnerabilities have generally focused on identifying genes driving tumor development and determining if these drivers can be exploited to develop novel therapies, either by targeting the drivers themselves or by exploiting other vulnerabilities stemming from the drivers.

Recently, several human sequencing studies have been undertaken to identify drivers of human ILC besides functional loss of E-cadherin[9,22]. Together, these studies have shed light on additional genetic alterations that are thought to be driver events, including chromosomal gains of 1q and 16p[23], loss of 16q[24], activating mutations in *PIK3CA*[25,26] and inactivating mutations in *TP53*[27]. Further molecular characterization has identified multiple aberrations in other components of the PI3K-AKT pathway, indicating that PI3K-AKT signaling plays an important role in ILC development[9,22,28]. However, a large fraction of human ILCs cannot be explained by activated PI3K-AKT signaling and *TP53* mutations, indicating that other aberrations are likely to play additional roles in tumorigenesis. Therefore, to identify novel genes and pathways driving ILC development, we used the *Sleeping Beauty* (SB) transposon system to perform an insertional mutagenesis (IM) screen in female mice with mammary-gland specific inactivation of *Cdh1*. The results of this screen are described in **Chapter 3**.

One of the main challenges of identifying candidate cancer drivers using an IM-based forward genetic screen, is that these screens can detect many potential cancer genes, of which only a fraction is actually involved in driving tumorigenesis. Besides this, it can be challenging to identify how genes are affected by their transposon insertions, as the targeted DNA-sequencing approaches that are typically employed for detecting insertion sites[29–31] do not provide any evidence of how insertions affect gene expression. We reasoned RNA-sequencing-based insertion site detection approaches could

alleviate these issues by focusing on detecting insertions that are actually expressed (and therefore more likely to have an actual effect), whilst simultaneously providing insight into how the expression of candidate genes is affected. To demonstrate this, we developed a computational approach and accompanying software package called IM-Fusion, which identifies transposon insertions from gene-transposon fusions in RNA-sequencing data. Details of the approach, including a comparison with targeted DNA-sequencing-based approaches, are described in **Chapter 4**.

As a result of their chromosomal instability, BRCA-deficient TNBCs develop characteristic patterns of copy number aberrations[32], suggesting that these aberrations harbor additional genes driving tumorigenesis. Unfortunately, these aberrations generally harbor tens-to-hundreds of genes, complicating the search for the true driver genes in these regions. To address this issue, computational approaches (e.g. RUBIC, GISTIC) have been developed to identify minimal recurrently aberrated regions and thereby narrow down lists of potential drivers[33,34]. Besides this, comparative oncogenomics approaches have also been used to restrict lists of candidate driver genes, by focusing on genes that are recurrently aberrated in tumors from both mouse models and human patients[35]. In **Chapter 5**, we explore the copy number landscape of *BRCA1*-mutated TNBC using several mouse models containing previously identified drivers such as *Myc*, *Met* and *Rb1*. By applying RUBIC in a comparative analysis between both mouse and human tumors, we show that engineered MYC overexpression in BRCA1-deficient TNBC dramatically reshapes the copy number landscape and identify MCL1 as a druggable driver in these tumors.

2.4 Preventing therapy resistance

Besides identifying druggable target genes, a significant challenge in the development of targeted therapies is the emergence of (acquired) therapy resistance, which is unfortunately frequently observed in patients after prolonged treatment with several targeted therapies[36]. To prevent the development of therapy resistance, it is crucial to gain an understanding of how tumors become resistant to therapies and use these insights to develop new (combination) treatments that aim to prevent or overcome resistance. Besides this, it is important to identify which patients are likely to be intrinsically resistant to treatment, so that these patients can be treated accordingly.

As part of the insertional mutagenesis screen described in Chapter 3, we identified FGFR2 as a key driver of ILC, suggesting that FGFR inhibition would be a suitable therapeutic strategy for treating FGFR-driven ILC. Although no FGFR-targeting therapies are currently approved for the treatment of human cancers, several thera-

peutics are currently being evaluated in phase I/II clinical trials for different types of cancers[37,38]). Unfortunately, studies with some of these inhibitors have shown that tumors can develop resistance to treatment, mainly via secondary mutations in FGFRs[39–41] and activation of alternative RTKs[42–45]. In **Chapter 6**, we explore the effectiveness of FGFR inhibition in FGFR-driven ILCs by transplanting tumor fragments into multiple recipient mice and treating them with the FGFR inhibitor AZD4547. Besides this, we exploit the ongoing transposon mutagenesis in these tumors to identify potential resistance mechanisms to AZD4547, which may be used for developing novel (combination) therapies that prevent or overcome resistance.

In BRCA-deficient TNBC, the most promising targeted treatments have aimed to exploit vulnerabilities resulting from the HR-deficiency incurred by BRCA1/BRCA2 loss[18]. This has led to the development of several PARP inhibitors, which indirectly induce the accumulation of DSBs in the DNA[46,47]. These DSBs cannot be repaired in an error-free fashion without BRCA1/2, leading to extensive DNA damage and cell death in BRCA-mutant cells[48,49]. Unfortunately, the clinical effectiveness of PARP inhibitors is limited by the emergence of therapy resistance[50,51], typically due to restoration of HR function via secondary mutations in *BRCA1/2*[52,53] or mutations in the 53BP1-RIF1-REV7 pathway (reviewed by Annunziato *et al.*[54]). However, for *BRCA2*-mutant tumors, there is no evidence that HR can be restored in the absence of BRCA2, suggesting that other mechanisms must be driving therapy resistance. To identify these additional resistance mechanisms, we combined *in vitro* screens in BRCA2-deficient mammary tumor cells with multi-omics analysis of BRCA2-deficient mouse mammary tumors that acquired PARPi resistance *in vivo*. The results of these analyses are described in **Chapter 7**.

2.5 Future perspectives

Coming to the end of this thesis, in **Chapter 8** we reflect on the methods and results presented in this work and how they may be applied or extended in future endeavours. Besides this, we also consider several technological advances and important challenges that remain to be addressed in the field. Finally, we discuss several limitations of mouse models in the light of expanding human datasets and development of three-dimensional cell culture models, and what this means for the future role of mouse models in cancer research.

2.6 References

[1] Collin Kent, Janet Horton, Rachel Blitzblau, and Bridget F Koontz. "Whose disease will recur after mastectomy for early stage, node-negative breast cancer? A systematic review". In: *Clinical Breast Cancer* 15.6 (2015), pp. 403–412 (cit. on p. 34).

[2] Britta Weigelt, Felipe C Geyer, and Jorge S Reis-Filho. "Histological types of breast cancer: how special are they?" In: *Molecular Oncology* 4.3 (2010), pp. 192–208 (cit. on p. 34).

[3] Joel S Parker, Michael Mullins, Maggie C U Cheang, et al. "Supervised risk predictor of breast cancer based on intrinsic subtypes". In: *Journal of Clinical Oncology* 27.8 (2009), pp. 1160–1167 (cit. on p. 34).

[4] Julie M Hall, John F Couse, and Kenneth S Korach. "The multifaceted mechanisms of estradiol and estrogen receptor signaling". In: *Journal of Biological Chemistry* 276.40 (2001), pp. 36869–36872 (cit. on p. 34).

[5] Manfred Beleut, Renuga Devi Rajaram, Marian Caikovski, et al. "Two distinct mechanisms underlie progesterone-induced proliferation in the mammary gland". In: *Proceedings of the National Academy of Sciences* 107.7 (2010), pp. 2989–2994 (cit. on p. 34).

[6] Alejandro Wolf-Yadlin, Neil Kumar, Yi Zhang, et al. "Effects of HER2 overexpression on cell signaling networks governing proliferation and migration". In: *Molecular Systems Biology* 2.1 (2006), p. 54 (cit. on p. 34).

[7] Sohrab P Shah, Andrew Roth, Rodrigo Goya, et al. "The clonal and mutational evolution spectrum of primary triple-negative breast cancers". In: *Nature* 486.7403 (2012), p. 395 (cit. on p. 34).

[8] Cancer Genome Atlas Network et al. "Comprehensive molecular portraits of human breast tumours". In: *Nature* 490.7418 (2012), p. 61 (cit. on p. 34).

[9] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, et al. "Comprehensive molecular portraits of invasive lobular breast cancer". In: *Cell* 163.2 (2015), pp. 506–519 (cit. on pp. 34, 35).

[10] Frank W Foote and Fred W Stewart. "A histologic classification of carcinoma of the breast". In: *Surgery* 19.1 (1946), pp. 74–99 (cit. on p. 34).

[11] Carien M Niessen and Cara J Gottardi. "Molecular components of the adherens junction". In: *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1778.3 (2008), pp. 562–571 (cit. on p. 34).

[12] CB Vos, AM Cleton-Jansen, Geert Berx, et al. "E-cadherin inactivation in lobular carcinoma in situ of the breast: an early event in tumorigenesis". In: *British Journal of Cancer* 76.9 (1997), p. 1131 (cit. on p. 34).

[13] Roland Moll, Margarete Mitze, Uwe H Frixen, and Walter Birchmeier. "Differential loss of E-cadherin expression in infiltrating ductal and lobular breast carcinomas." In: *The American Journal of Pathology* 143.6 (1993), p. 1731 (cit. on p. 34).

[14] Emad A Rakha, Arjun Patel, Des G Powe, et al. “Clinical and biological significance of E-cadherin protein expression in invasive lobular carcinoma of the breast”. In: *The American Journal of Surgical Pathology* 34.10 (2010), pp. 1472–1479 (cit. on p. 34).

[15] Bernhard C Pestalozzi, David Zahrieh, Elizabeth Mallon, et al. “Distinct clinical and prognostic features of infiltrating lobular carcinoma of the breast: combined results of 15 International Breast Cancer Study Group clinical trials”. In: *Journal of Clinical Oncology* 26.18 (2008), pp. 3006–3014 (cit. on p. 34).

[16] Sunil Badve, David J Dabbs, Stuart J Schnitt, et al. “Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists”. In: *Modern Pathology* 24.2 (2011), pp. 157–167 (cit. on p. 34).

[17] Beth N Peshkin, Michelle L Alabek, and Claudine Isaacs. “BRCA1/2 mutations and triple negative breast cancers”. In: *Breast Disease* 32.1-2 (2011), pp. 25–33 (cit. on p. 34).

[18] Christine S Walsh. “Two decades beyond BRCA1/2: homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy”. In: *Gynecologic Oncology* 137.2 (2015), pp. 343–350 (cit. on pp. 35, 37).

[19] Rebecca Dent, Maureen Trudeau, Kathleen I Pritchard, et al. “Triple-negative breast cancer: clinical features and patterns of recurrence”. In: *Clinical Cancer Research* 13.15 (2007), pp. 4429–4434 (cit. on p. 35).

[20] Katrina R Bauer, Monica Brown, Rosemary D Cress, Carol A Parise, and Vincent Caggiano. “Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype”. In: *Cancer* 109.9 (2007), pp. 1721–1728 (cit. on p. 35).

[21] Hanan Ahmed Wahba and Hend Ahmed El-Hadaad. “Current approaches in treatment of triple-negative breast cancer”. In: *Cancer Biology & Medicine* 12.2 (2015), p. 106 (cit. on p. 35).

[22] Magali Michaut, Suet-Feung Chin, Ian Majewski, et al. “Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer”. In: *Scientific Reports* 6 (2016), p. 18517 (cit. on p. 35).

[23] Daniel E Stange, Bernhard Radlwimmer, Falk Schubert, et al. “High-resolution genomic profiling reveals association of chromosomal aberrations on 1q and 16p with histologic and genetic subgroups of invasive breast cancer”. In: *Clinical Cancer Research* 12.2 (2006), pp. 345–352 (cit. on p. 35).

[24] P T Simpson, J S Reis-Filho, M B K Lambros, et al. “Molecular profiling pleomorphic lobular carcinomas of the breast: evidence for a common molecular genetic pathway with classic lobular carcinomas”. In: *The Journal of Pathology* 215.3 (2008), pp. 231–244 (cit. on p. 35).

[25] Fiamma Buttitta, Lara Felicioni, Fabio Barassi, et al. “PIK3CA mutation and histological type in breast carcinoma: high frequency of mutations in lobular carcinoma”. In: *The Journal of Pathology* 208.3 (2006), pp. 350–355 (cit. on p. 35).

[26] Matthias Christgen, Monika Noskowicz, Elisa Schipper, et al. “Oncogenic PIK3CA mutations in lobular breast cancer progression”. In: *Genes, Chromosomes and Cancer* 52.1 (2013), pp. 69–80 (cit. on p. 35).

[27] Cigdem Ercan, Paul J Van Diest, Bram Van Der Ende, et al. “p53 mutations in classic and pleomorphic invasive lobular carcinoma of the breast”. In: *Cellular Oncology* 35.2 (2012), pp. 111–118 (cit. on p. 35).

[28] Christine Desmedt, Gabriele Zoppoli, Gunes Gundem, et al. “Genomic characterization of primary invasive lobular breast cancer”. In: *Journal of Clinical Oncology* 34.16 (2016), pp. 1872–1881 (cit. on p. 35).

[29] Marco J Koudijs, Christiaan Klijn, Louise van der Weyden, et al. “High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors”. In: *Genome Research* 21.12 (2011), pp. 2181–2189 (cit. on p. 35).

[30] Aaron L Sarver, Jesse Erdman, Tim Starr, David A Largaespada, and Kevin AT Silverstein. “TAPDANCE: An automated tool to identify and annotate transposon insertion CISs and associations between CISs from next generation sequence data”. In: *BMC Bioinformatics* 13.1 (2012), p. 1 (cit. on p. 35).

[31] Karen M Mann, Justin Y Newberg, Michael A Black, et al. “Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq”. In: *Nature Biotechnology* 34.9 (2016), pp. 962–972 (cit. on p. 35).

[32] Tesa M Severson, Justine Peeters, Ian Majewski, et al. “BRCA1-like signature in triple negative breast cancer: Molecular and clinical characterization reveals subgroups with therapeutic potential”. In: *Molecular Oncology* 9.8 (2015), pp. 1528–1538 (cit. on p. 36).

[33] Ewald van Dyk, Marlous Hoogstraat, Jelle ten Hoeve, Marcel JT Reinders, and Lodewyk FA Wessels. “RUBIC identifies driver genes by detecting recurrent DNA copy number breaks”. In: *Nature Communications* 7 (2016), p. 12159 (cit. on p. 36).

[34] Craig H Mermel, Steven E Schumacher, Barbara Hill, et al. “GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy number alteration in human cancers”. In: *Genome Biology* 12.4 (2011), R41 (cit. on p. 36).

[35] Daniel Peeper and Anton Berns. “Cross-species oncogenomics in cancer gene identification”. In: *Cell* 125.7 (2006), pp. 1230–1233 (cit. on p. 36).

[36] Caitriona Holohan, Sandra Van Schaeybroeck, Daniel B Longley, and Patrick G Johnston. “Cancer drug resistance: an evolving paradigm”. In: *Nature Reviews Cancer* 13.10 (2013), p. 714 (cit. on p. 36).

[37] Irina S Babina and Nicholas C Turner. “Advances and challenges in targeting FGFR signalling in cancer”. In: *Nature Reviews Cancer* 17.5 (2017), pp. 318–332 (cit. on p. 37).

[38] Rut Porta, Roberto Borea, Andreia Coelho, et al. “FGFR a promising druggable target in cancer: Molecular biology and new drugs”. In: *Critical reviews in oncology/hematology* 113 (2017), pp. 256–267 (cit. on p. 37).

[39] Lipika Goyal, Supriya K Saha, Leah Y Liu, et al. “Polyclonal Secondary FGFR2 Mutations Drive Acquired Resistance to FGFR Inhibition in Patients with FGFR2 Fusion–Positive Cholangiocarcinoma”. In: *Cancer Discovery* 7.3 (2017), pp. 252–263 (cit. on p. 37).

[40] V Chell, K Balmanno, AS Little, et al. “Tumour cell responses to new fibroblast growth factor receptor tyrosine kinase inhibitors and identification of a gatekeeper mutation in FGFR3 as a mechanism of acquired resistance”. In: *Oncogene* 32.25 (2013), pp. 3059–3070 (cit. on p. 37).

[41] Sara A Byron, Huaibin Chen, Andreas Wortmann, et al. “The N550K/H mutations in FGFR2 confer differential resistance to PD173074, dovitinib, and ponatinib ATP-competitive inhibitors”. In: *Neoplasia* 15.8 (2013), 975IN28–988IN30 (cit. on p. 37).

[42] Maria Teresa Herrera-Abreu, Alex Pearson, James Campbell, et al. “Parallel RNA interference screens identify EGFR activation as an escape mechanism in FGFR3-mutant cancer”. In: *Cancer Discovery* 3.9 (2013), pp. 1058–1071 (cit. on p. 37).

[43] SM Kim, H Kim, MR Yun, et al. “Activation of the Met kinase confers acquired drug resistance in FGFR-targeted lung cancer therapy”. In: *Oncogenesis* 5.7 (2016), e241 (cit. on p. 37).

[44] Jinjia Chang, Shanshan Wang, Zhe Zhang, et al. “Multiple receptor tyrosine kinase activation attenuates therapeutic efficacy of the fibroblast growth factor receptor 2 inhibitor AZD4547 in FGFR2 amplified gastric cancer”. In: *Oncotarget* 6.4 (2015), p. 2009 (cit. on p. 37).

[45] Jun Wang, Oliver Mikse, Rachel G Liao, et al. “Ligand-associated ERBB2/3 activation confers acquired resistance to FGFR inhibition in FGFR3-dependent cancer cells”. In: *Oncogene* 34.17 (2015), pp. 2167–2177 (cit. on p. 37).

[46] Guotai Xu, Jos Jonkers, and Sven Rottenberg. “PARP Inhibitor Resistance—What Is Beyond BRCA1 or BRCA2 Restoration?” In: *PARP Inhibitors for Cancer Therapy*. Springer, 2015, pp. 453–471 (cit. on p. 37).

[47] Junko Murai, N Huang Shar-yin, Benu Brata Das, et al. “Trapping of PARP1 and PARP2 by clinical PARP inhibitors”. In: *Cancer Research* 72.21 (2012), pp. 5588–5599 (cit. on p. 37).

[48] Helen E Bryant, Niklas Schultz, Huw D Thomas, et al. “Specific killing of BRCA2-deficient tumours with inhibitors of poly (ADP-ribose) polymerase”. In: *Nature* 434.7035 (2005), pp. 913–917 (cit. on p. 37).

[49] Hannah Farmer, Nuala McCabe, Christopher J Lord, et al. “Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy”. In: *Nature* 434.7035 (2005), pp. 917–921 (cit. on p. 37).

[50] M William Audeh, James Carmichael, Richard T Penson, et al. “Oral poly (ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial”. In: *The Lancet* 376.9737 (2010), pp. 245–251 (cit. on p. 37).

[51] Peter C Fong, David S Boss, Timothy A Yap, et al. “Inhibition of poly (ADP-ribose) polymerase in tumors from BRCA mutation carriers”. In: *New England Journal of Medicine* 361.2 (2009), pp. 123–134 (cit. on p. 37).

[52] Stacey L Edwards, Rachel Brough, Christopher J Lord, et al. “Resistance to therapy caused by intragenic deletion in BRCA2”. In: *Nature* 451.7182 (2008), pp. 1111–1115 (cit. on p. 37).

[53] Elizabeth M Swisher, Wataru Sakai, Beth Y Karlan, et al. “Secondary BRCA1 mutations in BRCA1-mutated ovarian carcinomas with platinum resistance”. In: *Cancer Research* 68.8 (2008), pp. 2581–2586 (cit. on p. 37).

[54] Stefano Annunziato, Marco Barazas, Sven Rottenberg, and Jos Jonkers. “Genetic dissection of cancer development, therapy response, and resistance in mouse models of breast cancer”. In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 81. Cold Spring Harbor Laboratory Press. 2016, pp. 141–150 (cit. on p. 37).

3

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

Julian R. de Ruiter[1,2,*], Sjors M. Kas[1,*], Koen Schipper[1,*], Stefano Annunziato[1], Eva Schut[1], Sjoerd Klarenbeek[3], Anne Paulien Drenth[1], Eline van der Burg[1], Christiaan Klijn[1], Jelle J. ten Hoeve[2], David J. Adams[4], Marco J. Koudijs[1], Jelle Wesseling[1,5], Micha Nethe[1], Lodewyk F. A. Wessels[2,6,7], Jos Jonkers[1,6]

[1] Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[2] Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[3] Experimental Animal Pathology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[4] Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton, UK
[5] Department of Pathology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[6] Cancer Genomics Netherlands, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[7] Department of EEMCS, Delft University of Technology, Delft, the Netherlands
[*] These authors contributed equally to this work

Published in Nature Genetics, 26 June 2017
Volume 49, Issue 8, Pages 1219–1230

3.1 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.

3.2 Introduction

ILC belongs to the luminal subtype of breast cancer and accounts for 8-14% of all breast cancer cases[1–3]. 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[4]. 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[5–8]. Of note, female mice with mammary-specific inactivation of E-cadherin are not prone to developing mammary tumors[9–11], 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[12], loss of chromosome 16q[13], activating mutations in *PIK3CA*[14,15] and inactivating mutations in *TP53*[16]. 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[7,17,18]. 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*.

3.3 Results

3.3.1 *Sleeping Beauty*-induced mammary tumors in *Wap-Cre;Cdh1$^{F/F}$;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 (*Rosa26Lox66SBLox71*), resulting in *Wap-Cre;Cdh1$^{F/F}$;T2/Onc;Rosa26$^{Lox66SBLox71/+}$* (hereafter referred to as *Wap-Cre;Cdh1$^{F/F}$;SB)* mice (Figure 3.1A)[11,19,20]. 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 3.1B). 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 S3.1A). 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.

3.3.2 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 pat-

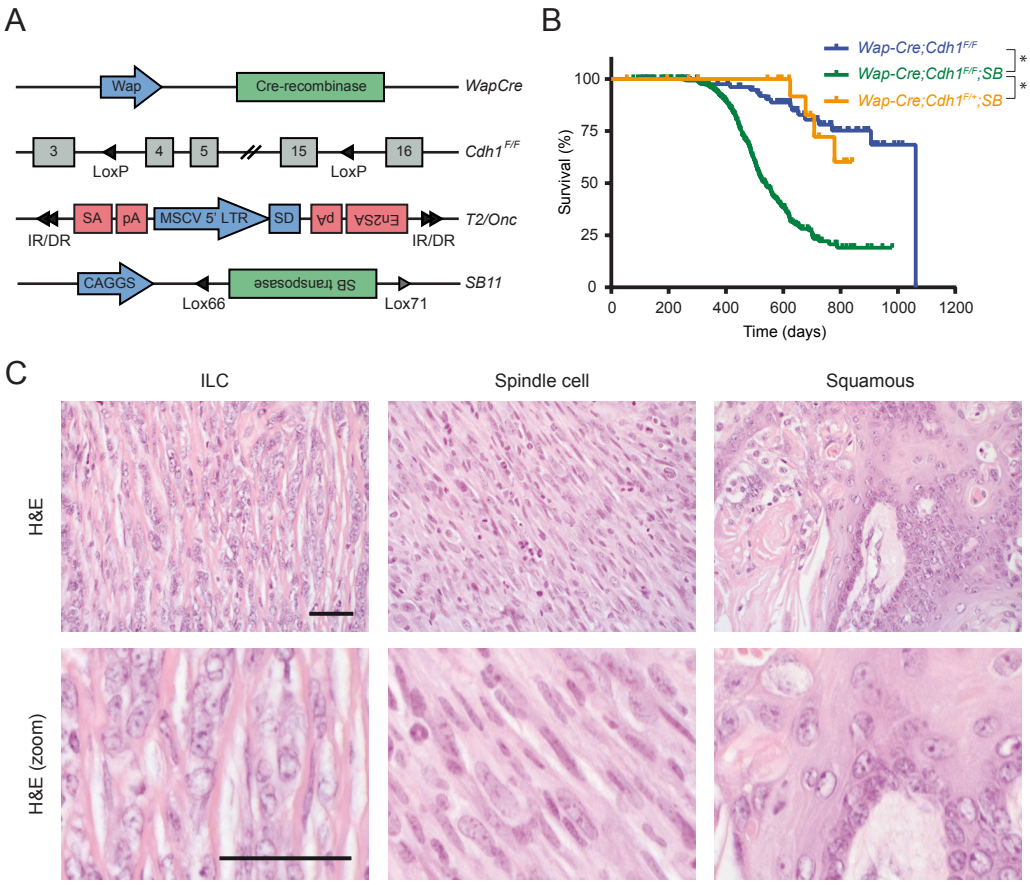


Fig. 3.1. SB insertional mutagenesis induces tumorigenesis in female mice with mammary-gland-specific inactivation of E-cadherin. (A) Overview of the engineered alleles in *Wap–Cre;Cdh1$^{F/F}$;SB* mice. In this SB mutagenesis system, genetically engineered transposons, which contain a 5' long terminal 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[19]. 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.

tern with noncohesive E-cadherin-negative and cytokeratin 8 (CK8)-positive cells invading the surrounding tissue in single-cell strands, thus resembling hILC (Figure 3.1C-D, Supplementary Figure S3.1B-C and Supplementary Table S3.1). 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 tu-

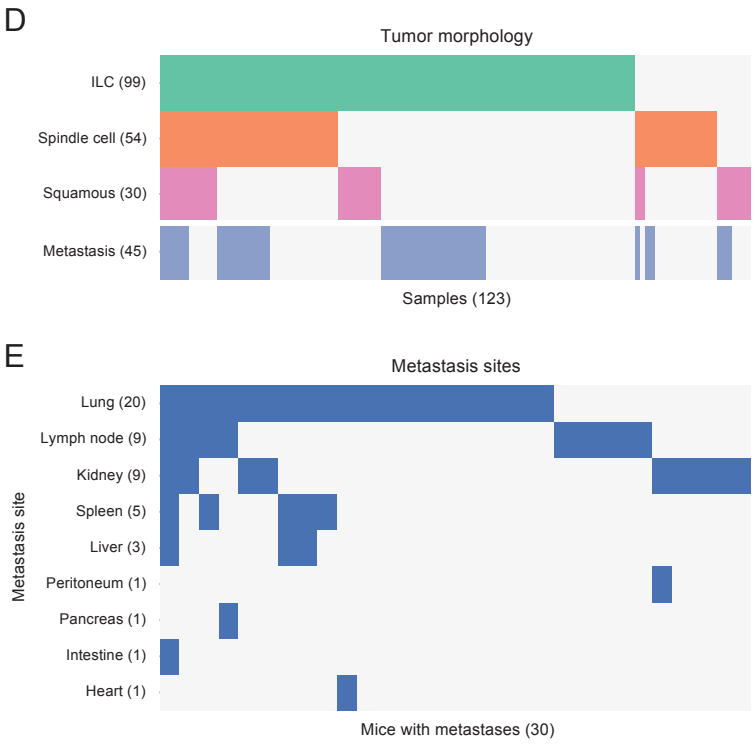


Fig. 3.1. **Continued.** **(D)** Histological classification of 123 tumors from 89 *Wap–Cre;Cdh1$^{F/F}$;SB* females and the overlap with metastasis formation. **(E)** Overview of metastases to distant organs in metastasis-bearing *Wap–Cre;Cdh1$^{F/F}$;SB* females (30/89 mice).

mors 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 3.1E). 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.

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)[21,22]. For additional reference, two existing mouse models of luminal breast cancer (*Wap-Cre;Cdh1$^{F/F}$;Pten$^{F/F}$*)[23] and basal-like breast cancer (*K14-Cre;Brca1$^{F/F}$;Trp53$^{F/F}$*)[24] 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 3.2A and Supplementary Figure S3.2A).

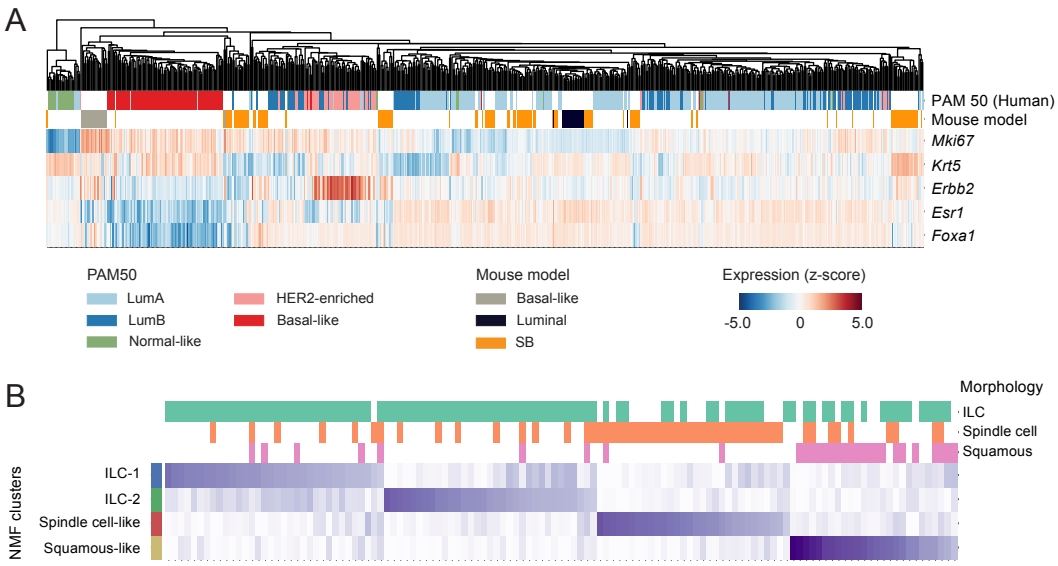


Fig. 3.2. **Gene expression analysis of SB-induced tumors. (A)** Unsupervised clustering analysis (Euclidean distance, average linkage) of the SB-induced tumors ($n = 123$) with human breast cancer samples from TCGA (LumA, $n = 231$; LumB, $n = 127$; basal-like, $n = 95$; HER2-enriched, $n = 57$ and 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.

3.3.3 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 3.2B), which were not associated with a specific *T2/Onc* transgenic line (Supplementary Figure S3.3). 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 S3.2B-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).

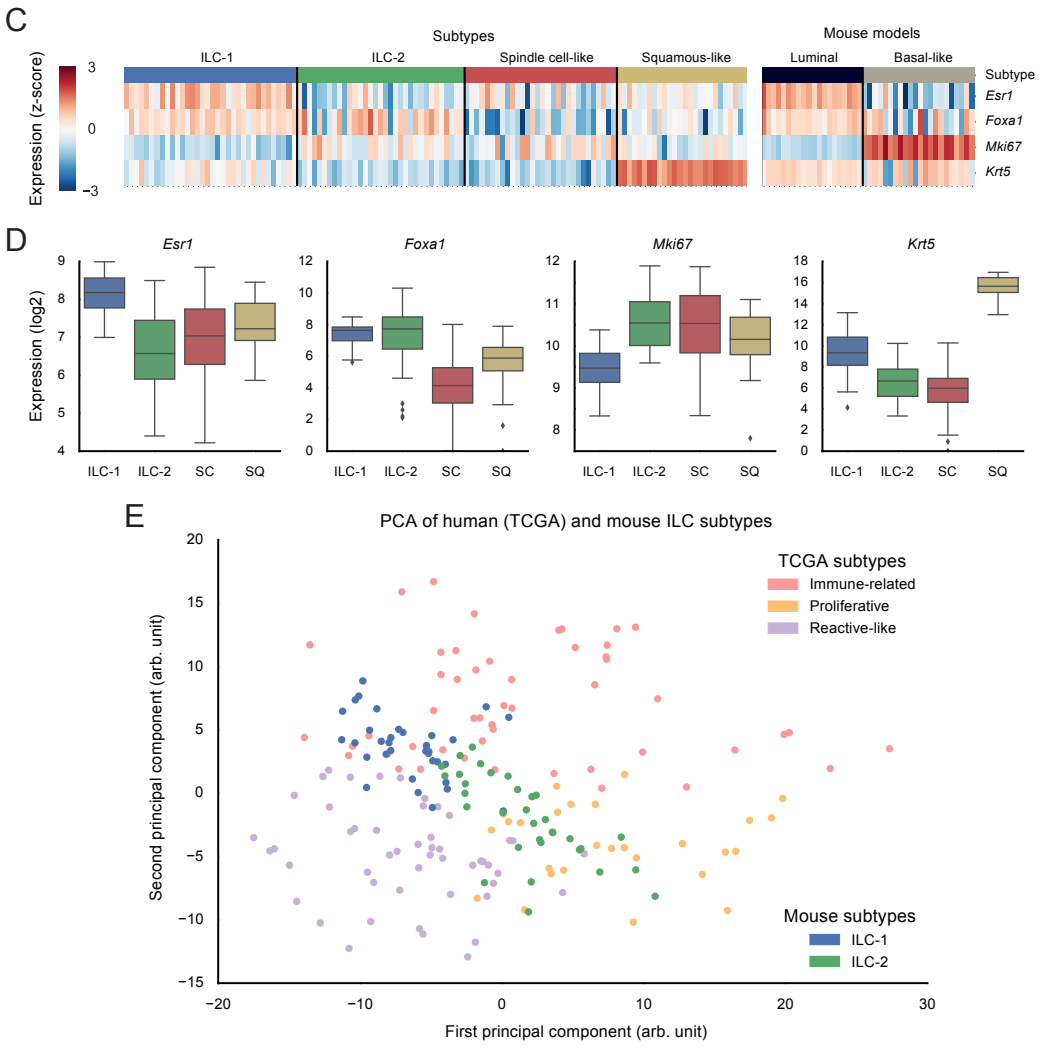


Fig. 3.2. **Continued. (C,D)** Heat map (C) and quantification (D) of the expression of four key genes from the PAM50 gene signature for the different SB-induced subtypes and the mouse reference models, highlighting differences in expression between the different subtypes described in B. SC, spindle-cell-like; 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.

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 3.2C-D and Supplementary Figure S3.2D). Consequently, we found that mILC-1 tumors most closely reflect the luminal A subtype of tumors[25–27]. 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 hILC7, 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 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 3.2E).

3.3.4 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[28]. 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)[29], 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[30] (Figure 3.3A).

This analysis identified 3,230 insertions with a median of 29 insertions per tumor (Supplementary Figure S3.4). 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 3.3B). 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 (Supplementary Table S3.2). 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.

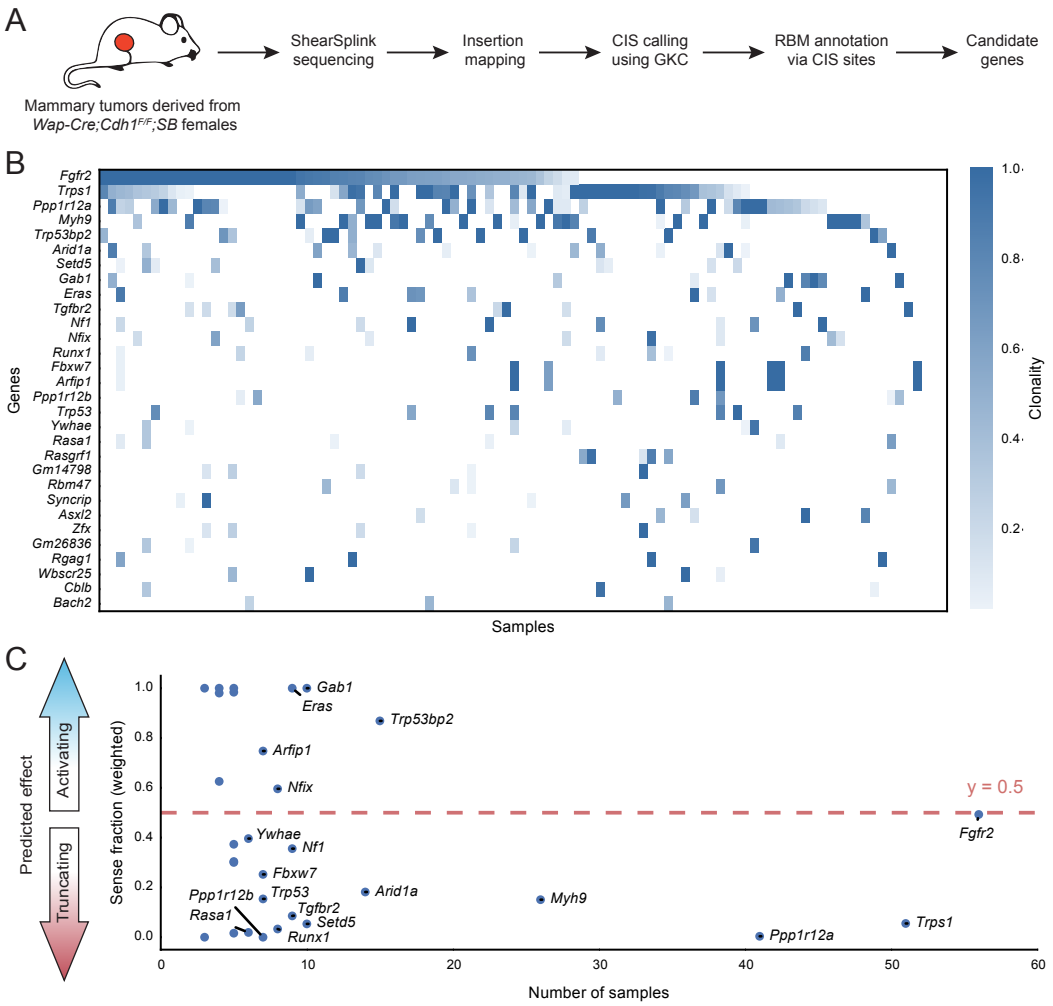


Fig. 3.3. Insertion analysis of tumors from *Wap–Cre;Cdh1$^{F/F}$;SB* females. **(A)** Overview of the pipeline used to identify candidate genes. **(B)** Overview of the insertions in candidate genes across all samples with an ILC morphology ($n = 99$). The relative clonality of the insertions within 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.

To prioritize candidate genes, we ranked the genes by their frequency and the median value of the clonality of their insertions (Supplementary Figure S3.5A). 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 S3.5B). 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.

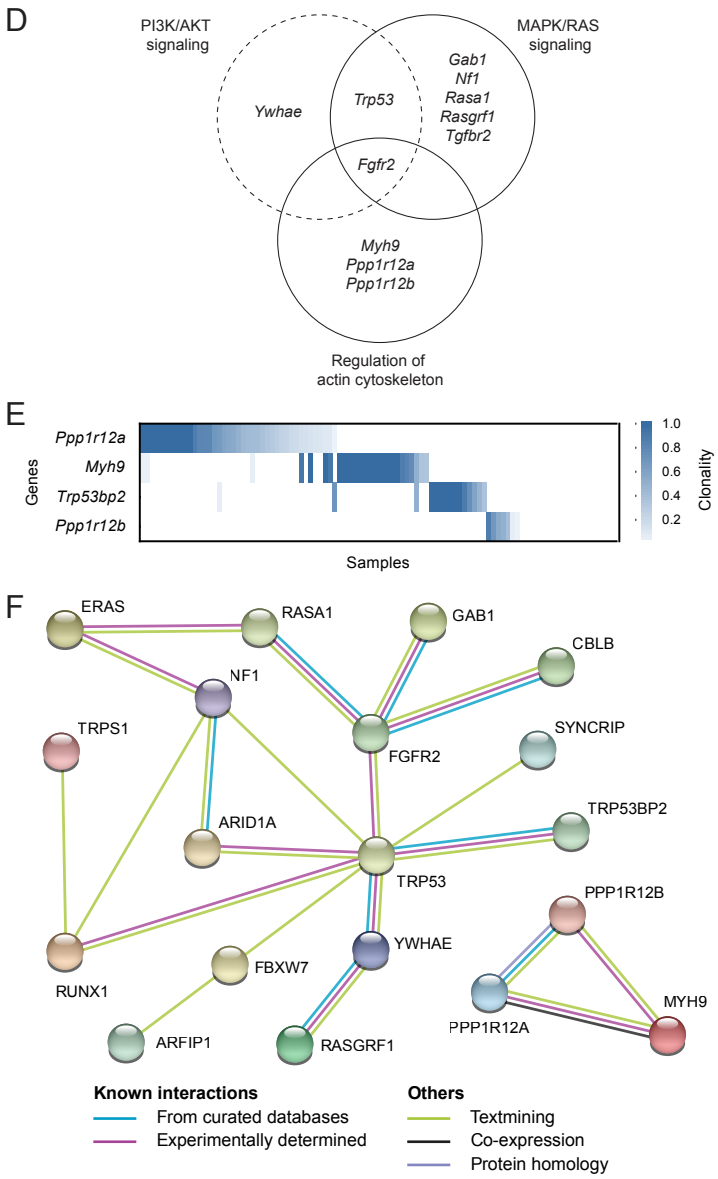


Fig. 3.3. **Continued. (D)** Venn diagram depicting the candidate genes involved in PI3K-AKT signaling (according to KEGG, dashed circle), which is known to be associated with hILC, together with two significant pathways from the KEGG analysis. **(E)** Overview of insertions in the four genes that were identified to be significantly mutually exclusive ($p < 10^{-3}$) 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 S3.6).

3.3.5 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[10,11,16]. 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[8,17,18]. These results demonstrated that our screen identified cancer driver genes and pathways that are known to be involved in hILC.

3.3.6 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 3.3C) 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 S3.7A). In support of this hypothesis, *Trp53bp2*, *Gab1* and *Eras* showed significantly ($p < 10^{-3}$) increased expression of exons downstream of the insertion site (Supplementary Figure S3.7B). In contrast, most of the candidate genes either showed no orientation bias or were biased toward antisense insertions (for example, *Trps1*; Supplementary Figure S3.7C), 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 S3.7B).

3.3.7 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 3.3D and Table 3.1). 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 immunohistochemical staining for phosphorylated ERK1-ERK2, which are

downstream effectors of RAS-MAPK signaling (Supplementary Figure S3.8). 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[31] 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 ($p < 10^{-3}$), suggesting that these genes were likely involved in a common pathway (Figure 3.3E). This hypothesis was supported by a projection of the candidate genes onto the STRING protein-protein interaction network (Figure 3.3F), 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[32–34], 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.

3.3.8 *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 TCGA7 (Figure 3.4A and

Tab. 3.1. 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*

Supplementary Figure S3.9). 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 3.4B-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 3.4D), 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.

3.3.9 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 3.5A and Supplementary Figure S3.10A). 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 3.5B and Supplementary Figure S3.10B). 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 S3.10C).

3.3.10 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 3.5C-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 3' of the insertions in *Trp53bp2*, as compared to expression levels of the full-length transcripts. Overexpression of the sequences

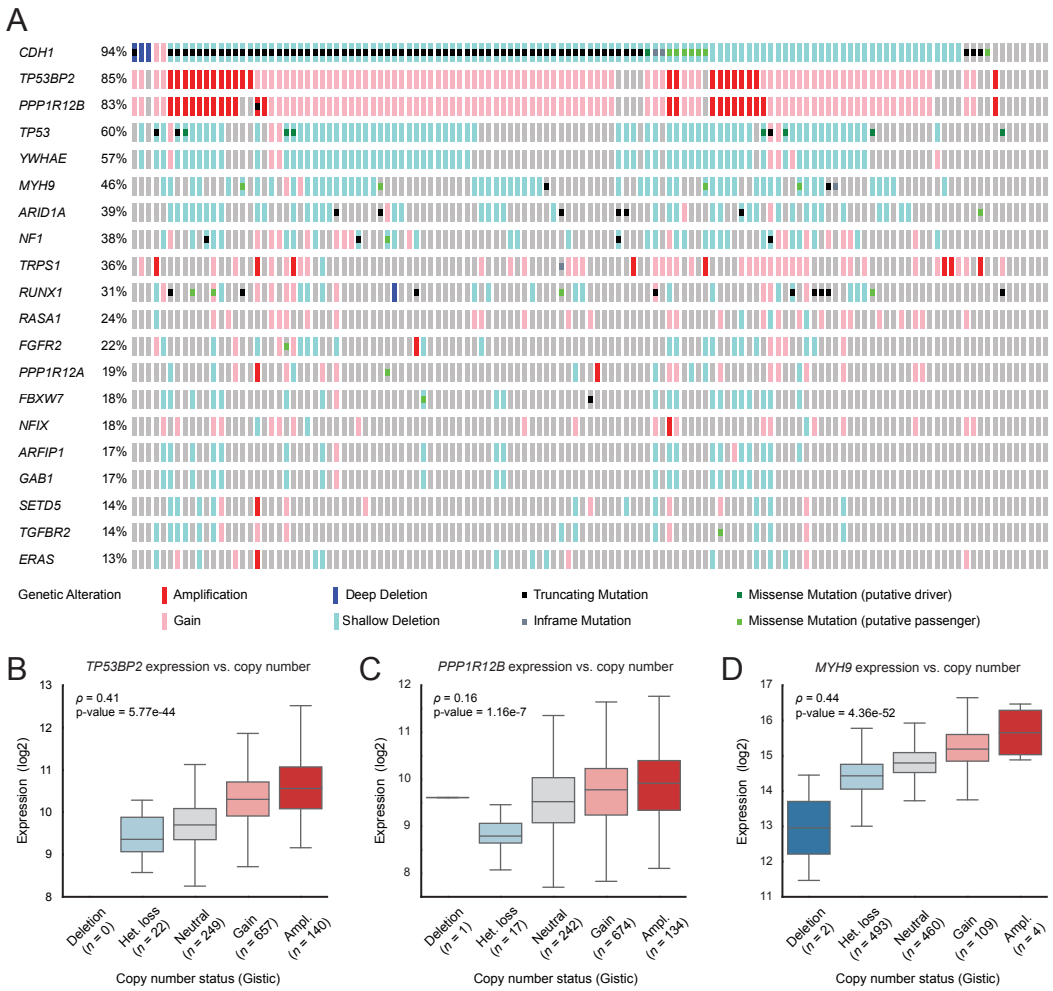


Fig. 3.4. Overview of the candidate genes in hILC. (A) Overview of the mutations and copy number events in 127 TCGA ILC samples for each of the main candidate genes. Percentages indicate the fraction of tumors with alterations in the respective genes. **(B–D)** Correlation between the expression of *TP53BP2* (B), *PPP1R12B* (C) and *MYH9* (D) and their respective copy number levels, using the entire TCGA breast cancer data set ($n = 1068$) 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.

encoding the truncated PP1 targeting subunits was confirmed by northern blot analysis (Supplementary Figure S3.11A-C) and by western blotting for PPP1R12A (Supplementary Figure S3.11D).

Analysis of the predicted proteins showed that the truncated PP1 targeting subunits lacked various regulatory domains but retained their PP1-binding domains (Figure 3.5F). 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

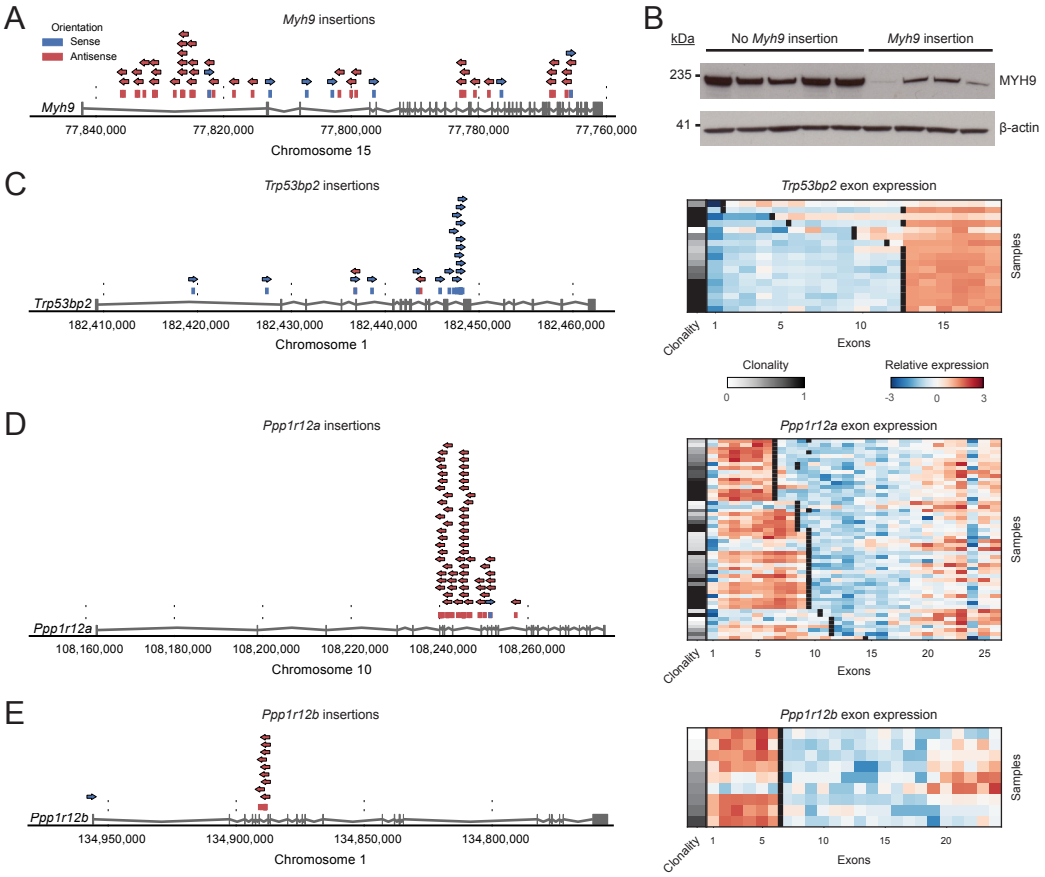


Fig. 3.5. Overview of the insertions and corresponding gene expression of the mutually exclusive genes. **(A)** Visualization of SB insertions (arrows) in *Myh9* ($n = 33$ tumors). Bars represent the exact genomic locations of the insertions. **(B)** Immunoblot for MYH9 levels in SB-induced tumor-derived 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 polyA tail selection in the RNA sequencing analysis, which has well-documented 3' bias.

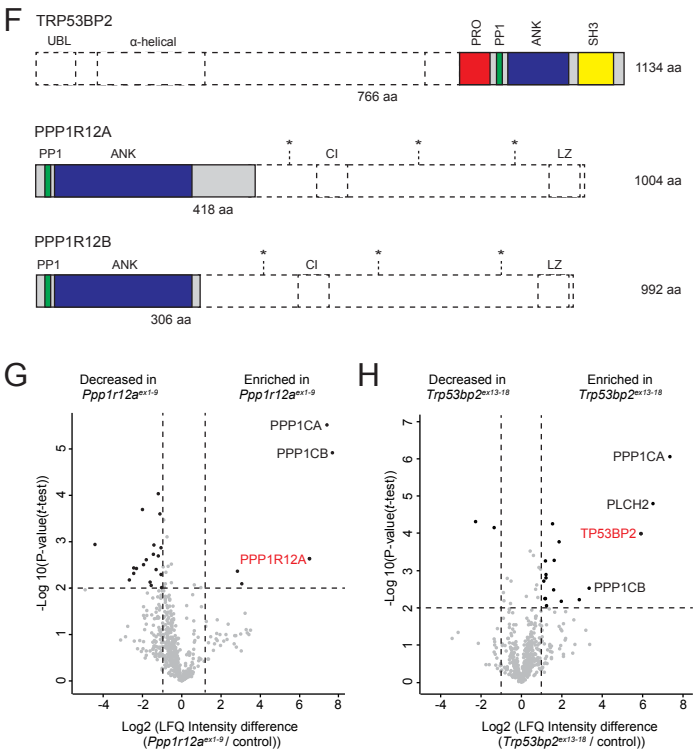


Fig. 3.5. Continued. **(F)** Overview of the binding domains of mouse TRP53BP2, PPP1R12A and PPP1R12B, based on previously published work[32,35]. Colors indicate the predicted proteins from the truncated genes. UBL, ubiquitin-like domain; PRO, proline-rich domain; PP1, PP1-binding domain; ANK, ankyrin 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*$^{ex1\text{-}9}$ or *pBABE-Trp53bp2*$^{ex13\text{–}18}$, 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.

(encoded by *Ppp1r12a* exons 1-9) or TRP53BP2 protein (encoded by *Trp53bp2* exons 13-18). This showed that both 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 3.5G-H and Supplementary Figure S3.11E). 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.

3.3.11 Candidate ILC drivers enhance survival of *Cdh1*$^{\Delta/\Delta}$ 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[23,36]. 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 3.6A-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*$^{ex1\text{-}9}$ or *Trp53bp2*$^{ex13\text{-}18}$ (Figure 3.6D). 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 3.6A-C and Supplementary Figure S3.12A-C). Similar results were obtained after reduction of MYH9 levels by short hairpin RNA (shRNA)-mediated knockdown of *Myh9* expression (Figure 3.6E-H and Supplementary Figure S3.12D).

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[37]. In contrast, we and others[38] have observed an intact p53 response after DNA damage in cells with reduced MYH9 levels (Supplementary Figure S3.12E-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.

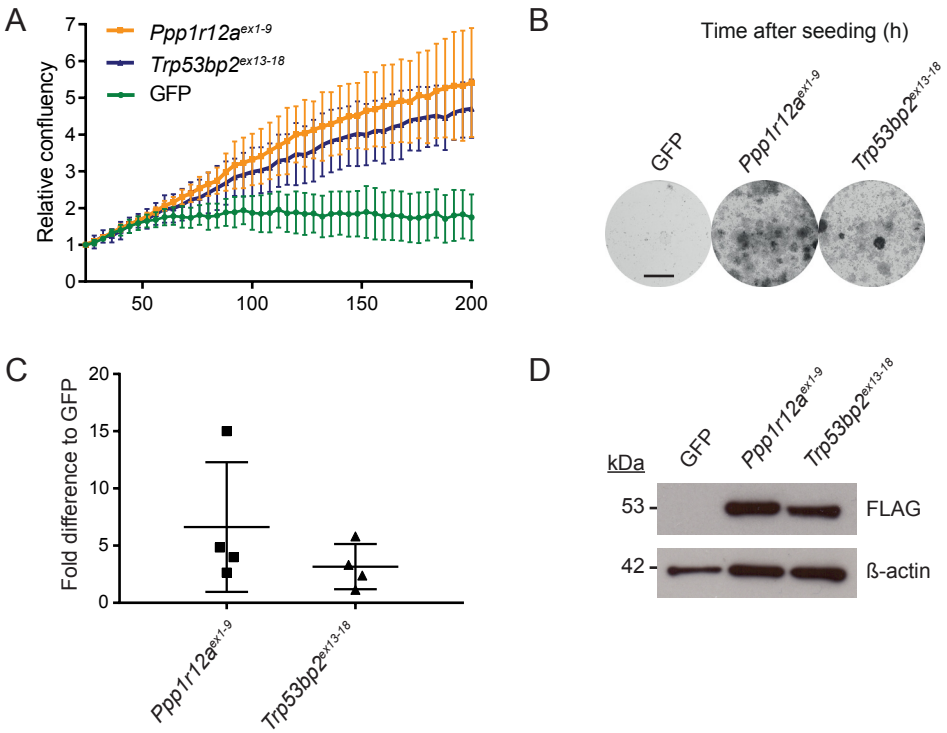


Fig. 3.6. **Limited proliferation and survival of AdCre-transduced *Cdh1$^{F/F}$;mT/mG* mouse mammary epithelial cells (MMECs) that were rescued by expression of truncated PPP1R12A and TRP53BP2 or by dosage reduction of MYH9. (A)** Cell survival analysis of AdCre-transduced *Cdh1$^{F/F}$;mT/mG* MMECs 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.

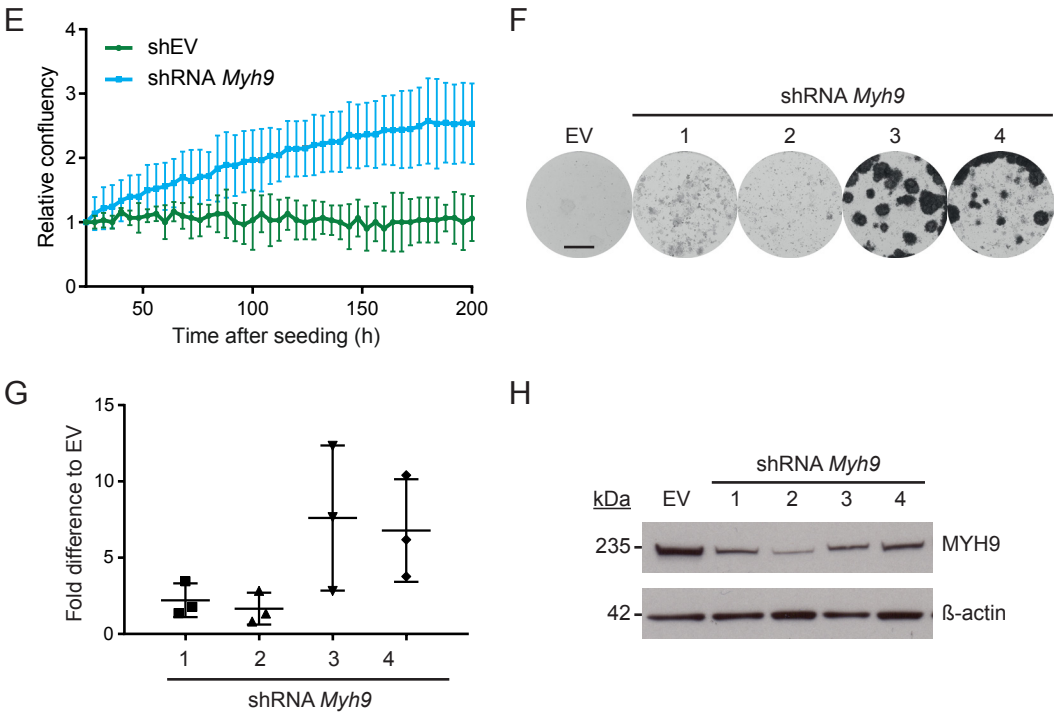


Fig. 3.6. **Continued.** **(E)** Cell survival analysis of AdCre-transduced *Cdh1*$^{F/F}$*;mT/mG* MMECs with simultaneous shRNA-mediated knockdown of Myh9 expression, as quantified by real-time IncuCyte imaging for 200 h. Average survival of AdCre-transduced *Cdh1*$^{F/F}$*;mT/mG* MMECs of all shRNAs is shown. Independent survival curves are depicted in Supplementary Figure S3.12D. 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.

3.3.12 Truncated PPP1R12A and TRP53BP2 induce ILC formation

Next we investigated whether expression of *Ppp1r12a*$^{ex1\text{-}9}$ and *Trp53bp2*$^{ex13\text{-}18}$ in *Wap-Cre;Cdh1*$^{F/F}$ mice could induce mammary tumor formation *in vivo*. To this end, we introduced *invCAG-Ppp1r12a*$^{ex1\text{-}9}$*-IRES-Luc* and *invCAG-Trp53bp2*$^{ex13\text{-}18}$*-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[39] (Supplementary Figure S3.13A). Male chimeras were mated with *Cdh1*$^{F/F}$ females to generate *Wap-Cre;Cdh1*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Ppp1r12a\text{-}ex1\text{-}9\text{-}IRES\text{-}Luc/+}$ (hereafter referred to as *Wap-*

Cre;Cdh1$^{F/F}$*;Ppp1r12a*$^{ex1-9}$) and *Wap-Cre;Cdh1*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Trp53bp2\text{-}ex13\text{-}18\text{-}IRES\text{-}Luc/+}$ (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 3.7A).

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 3.7B-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 3.7D-E). All of the tumors showed recombination of the conditional alleles (Supplementary Figure S3.13B), 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 S3.13C). 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.

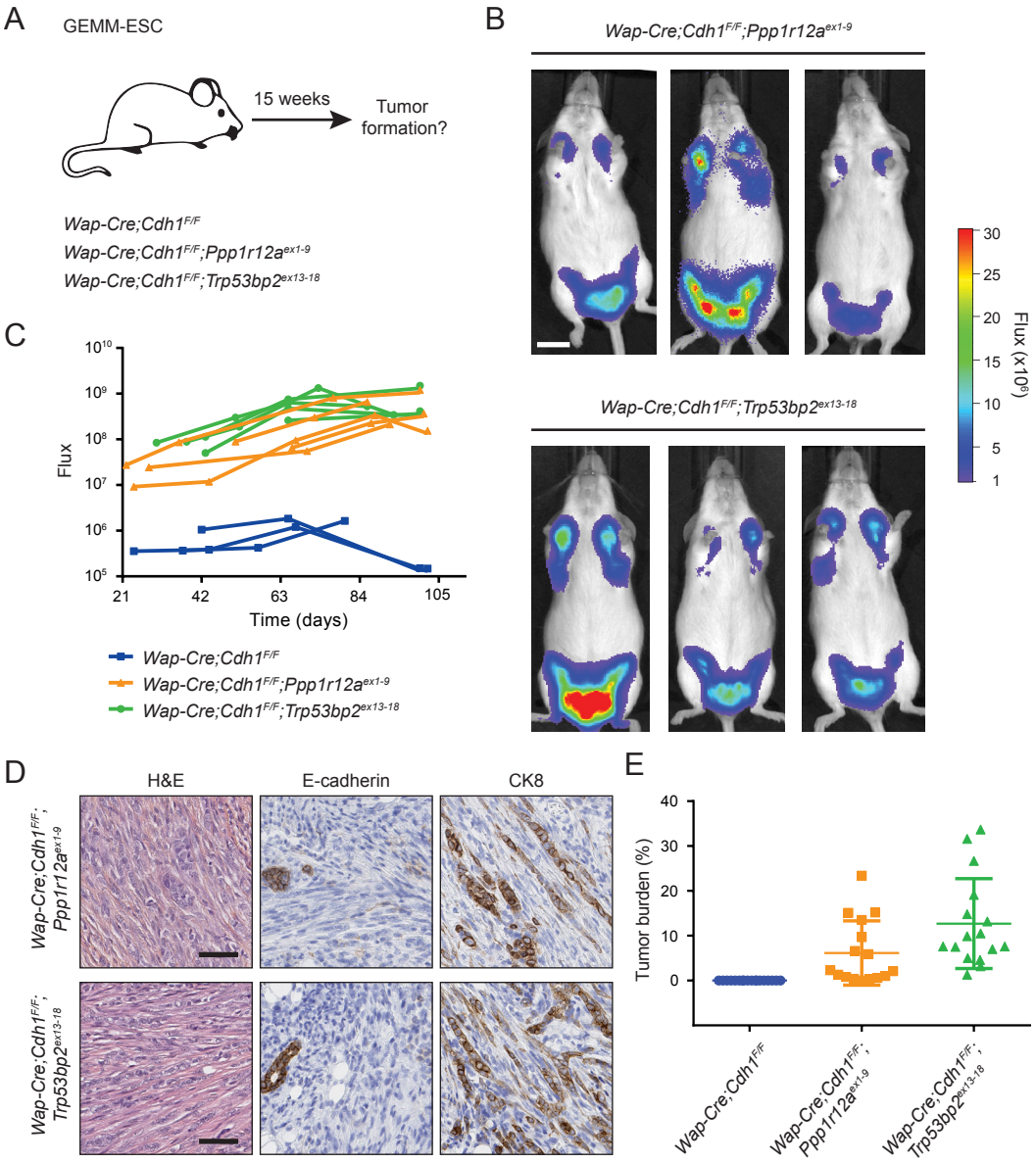


Fig. 3.7. Validation and characterization of candidate genes in genetically engineered mice (GEMM). (A) Overview of the genetically engineered mice. ESC, embryonic stem cell. **(B)** Representative images of *in vivo* bioluminescence imaging of luciferase expression in *Wap–Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1–9}$ ($n = 3$) 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 $\pm$ s.d.

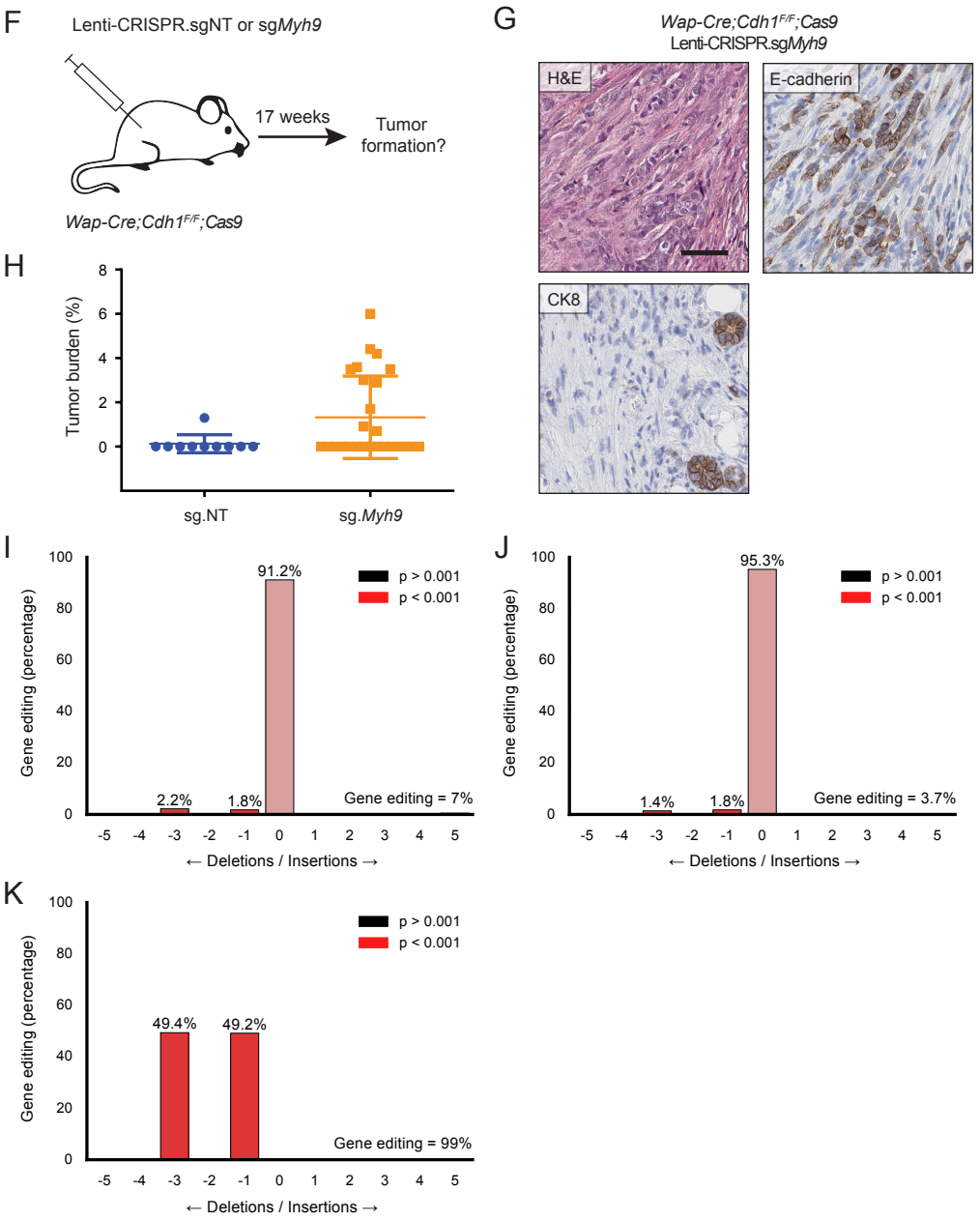


Fig. 3.7. Continued. **(F)** Overview of intraductal injections performed in *Wap–Cre;Cdh1$^{F/F}$;Cas9* females with high-titer lentivirus containing a vector encoding Cas9, GFP and either a non-targeting (NT) single-guide RNA (sgRNA) (Lenti-CRISPR-sgNT) or a sgRNA targeting the second exon of *Myh9* (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 $\pm$ 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[40]. 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*.

3.3.13 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*$^{invCAG\text{-}Cas9\text{-}IRES\text{-}Luc/+}$ (*Wap-Cre;Cdh1*$^{F/F}$*;Cas9*) female mice (Figure 3.7F). 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 3.7G-H and Supplementary Figure S3.13D).

Analysis of *Myh9* target modification in these tumors showed equal levels of in-frame and out-of-frame alterations (Figure 3.7I), 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 3.7J-K). Collectively, these data provide additional support for *Myh9* as a haploinsufficient tumor suppressor in ILC development.

3.4 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*)[7,16–18], 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[41–43]. Notably, mutations that can be linked to activated PI3K-AKT signaling are found in only 50% of hILCs[7,17,18], 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[7,17,18,44]. 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*[45,46].

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[7,12], 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[47]. 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[48], 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.

Competing financial interests The authors declare no competing financial interests.

3.5 Methods

3.5.1 Mouse models

Generation of mice The generation of *Wap-Cre;Cdh1*$^{F/F}$ mice has previously been described by Derksen et *al.*[11]. 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[19]. The mice containing the conditional HSB5 variant of the SB transposase in the Rosa26 locus has been described by March *et al.*[20]. *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 1500 mm^3 (tumor volume $=$ length $*$ width^2 $*$ 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[39]. The resulting chimeric males were crossed with *Cdh1*$^{F/F}$ (FVB) females to generate *Wap–Cre;Cdh1*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Ppp1r12a\text{-}ex1\text{-}9\text{-}IRES\text{-}Luc/+}$ (*Wap–Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1–9}$) and *Wap–Cre;Cdh1*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Trp53bp2\text{-}ex13\text{-}18\text{-}IRES\text{-}Luc/+}$ (*Wap–Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13–18}$) mice, respectively.

The generation of *Wap–Cre;Cdh1*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Cas9\text{-}IRES\text{-}Luc/+}$ (*Wap–Cre;Cdh1*$^{F/F}$*;Cas9*) mice has been described by Annunziato *et al.*[49]. The *Wap–Cre*, *Cdh1*^{F}, *T2/Onc*, *Rosa26*^{Lox66SBLox71} and *Col1a1*$^{invCAG\text{-}Cas9\text{-}IRES\text{-}Luc}$ alleles were confirmed by PCR as previously described[10,11,19,39,50]. *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 of the primers are listed in Supplementary Table S3.3.

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[51]. 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[49,52]. Lentiviral titers ranging from $2 * 10^8$ transducing units (TU) per milliliter 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[53]. 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[54], 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[51,53]. Antibody details and antigen retrieval methods are described in Supplementary Table S3.4. 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).

3.5.2 Cell culture

Cell lines Purified primary MMECs were isolated from 10-to-15 week-old females as previously described[23,55] and cultured in Dulbecco's modified Eagle's medium (DMEM)-F12 containing 10 % fetal bovine serum (FBS), 100 IU/ml penicillin, 100 μg ml^{-1} streptomycin, 5 ng ml^{-1} insulin, 5 ng ml^{-1} epidermal growth factor (EGF) (all from Life Technologies) and 5 ng ml^{-1} cholera toxin (Gentaur). HC11 cells[56] were cultured in DMEM/F12-Glutamax medium containing 10 % FBS, 100 IU/ml penicillin, 100 μg ml^{-1} streptomycin, 5 ng ml^{-1} insulin, 5 ng ml^{-1} EGF (all from Life Technologies). All cell lines were routinely tested for mycoplasma contamination using the MycoAlert mycoplasma detection kit (Lonza).

For the isolation of primary tumor cells, whole mammary tissue or 50 mm^3 to 100 mm^3 tumor sample was finely chopped using a McIlwain tissue chopper (The Mickle Laboratory Engineering Co.) and digested for 30 minutes shaking at 37 °C with 3 mg ml^{-1} collagenase, 0.1 % trypsin and 0.5 μg ml^{-1} Fungizone in DMEM/F12-glutamax medium containing 2 % FBS and 100 IU/ml penicillin (all Life Technologies). Enzyme activity was neutralized by addition of DMEM/F12-glutamax with 2 % FBS and suspension was dispersed through a 40 μm cell strainer. Cells were washed twice and cultured in DMEM/F12-glutamax medium containing 10 % FBS, 100 IU/ml penicillin, 100 μg ml^{-1} streptomycin, 5 ng ml^{-1} insulin, 5 ng ml^{-1} EGF (all Life Technologies), and 5 ng ml^{-1} cholera toxin (Gentaur). Single clones were derived by seeding the polyclonal cell lines, which were first dispersed through a 40 μm cell strainer, at a concentration of 0.5 cells per 96-well. 293T and Phoenix packaging cells were cultured in Iscove's medium (Life Technologies) containing 10 % FBS, 100 IU/ml penicillin, and 100 μg ml^{-1} streptomycin.

Retroviral vectors and virus production The *Ppp1r12a*$^{ex1\text{-}9}$ and *Trp53bp2*$^{ex13\text{-}18}$ sequences were isolated with BamHI-SalI overhangs and a 3' or 5' FLAG-tag, respectively using Phusion Flash High-Fidelity DNA Polymerase from *Frt-invCag-Ppp1r12a*$^{ex1\text{-}9}$*-IRES-Luc* or *Frt-invCag-Trp53bp2*$^{ex13\text{-}18}$*-IRES-Luc* and subsequently cloned in zero TOPO blunt vectors. These cDNAs were inserted as BamHI-SalI fragments into the pBABE-puro retroviral vector. Lenti-GFP, Lenti-*Ppp1r12a*$^{ex1\text{-}9}$ and Lenti-*Trp53bp2*$^{ex13\text{-}18}$ were constructed as follows. AgeI and SalI were used to remove GFP from the SIN.LV.SF vector[57], which was a gift of Eugenio Montini. T2A-puro was isolated with AgeI-SalI overhangs using standard PCR from pSpCas9(BB)-2A-Puro (PX459) V2.0, which was a gift from Feng Zhang (Addgene plasmid #62988)[58]. T2A-puro was inserted as AgeI-SalI fragment into the SIN.LV.SF vector, resulting in SIN.LV.SF-T2A-puro. GFP was isolated with BamHI-AgeI overhangs from the original SIN.LV.SF vector. *Ppp1r12a*$^{ex1\text{-}9}$ and *Trp53bp2*$^{ex13\text{-}18}$ were isolated with BamHI-AgeI overhangs and a 3' and 5' FLAG-tag, respectively using Phusion Flash High-Fidelity DNA Polymerase from pBABE-*Ppp1r12a*$^{ex1\text{-}9}$ and pBABE-*Trp53bp2*$^{ex13\text{-}18}$. These cDNAs were inserted as BamHI-AgeI fragments into SIN.LV.SF- T2A-puro, resulting in SIN.LV.SF-GFP-T2A-puro (Lenti-GFP), SIN.LV.SF-*Ppp1r12a*$^{ex1\text{-}9}$-T2A-puro (Lenti-*Ppp1r12a*$^{ex1\text{-}9}$) and SIN.LV.SF-*Trp53bp2*$^{ex13\text{-}18}$-T2A-puro (Lenti-*Trp53bp2*$^{ex13\text{-}18}$). All primers are listed in Supplementary Table S3.3.

The following Broad Institute's Mission TRC-1 mouse library shRNAs targeting *Myh9* are used #1: TRCN0000071503, #2: TRCN0000071504, #3: TRCN0000071506 and #4: TRCN0000071507. pL-CRISPR.EFS.GFP was a gift from Benjamin Ebert (Addgene plasmid #57818). The non-targeting sgRNA (TGATTGGGGGTCGTTCGCCA) and sgRNA targeting *Myh9* (CTCAAGGAGCGATACTACTC) were cloned in the pL-CRISPR.EFS.GFP vector as previously described[59]. All vectors were validated by Sanger sequencing. Concentrated lentiviral stocks were produced by transient co-transfection of four plasmids in 293T cells as previously described[60]. Viral titers were determined using the qPCR lentivirus titration kit from Abm (LV900).

Clonogenic assay MMECs were seeded in 6-well plates and 70 % to 90 % confluent wells were transduced with viral Ad5-CMV-Cre particles ($1 * 109$ transducing units [TU]; Gene Transfer Vector Core, University of Iowa) and Lenti-GFP, Lenti-*Ppp1r12a*$^{ex1\text{-}9}$, Lenti-*Trp53bp2*$^{ex13\text{-}18}$ or shRNAs (MOI>40) in the presence of $8\,\mu g\,ml^{-1}$ polybrene (Sigma). Medium was refreshed 24 hours after transduction and cells were selected with puromycin ($1.8\,\mu g\,ml^{-1}$) for 48 hrs. Confluent wells were trypsinized and 50,000 cells were seeded in 6-well plates. Medium was refreshed 24 hours after seeding and the well confluency was recorded every 4 hours with 36 images per 6-well for 200 hrs using an IncuCyte Zoom Live-Cell Analysis System (Essen Bioscience). The images were analyzed using IncuCyte Zoom software. The relative confluency was calculated as a ratio between the confluency at

the indicated time point and $t = 0$. After 14 days, cells were fixed with 4 % formalin in PBS and stained with 0.1 % crystal violet in demineralized water. Quantification was performed by dissolving the crystal violet in 10 % acetic acid in demineralized water and determining the absorbance at 590 nm. The experiments with the MMECs were all performed at least three times, using independent MMECs isolations and transductions of the cells.

DNA damage response in SB-induced tumor derived cell lines SB-induced tumor-derived cell lines were seeded and allowed to grow till 80 % to 90 % confluency. The cells were subsequently treated with vehicle (water) or doxorubicin (1 μmol; sc-200923) for 6 hours and protein lysates were made using lysis buffer complemented with protease inhibitors (Roche).

Flow cytometry MMECs were transduced as described previously and collected 72 hours after seeding 200.000 MMECs and stained using propidium iodide (556463, BD Biosciences) and APC Annexin V (550474, BD Biosciences) according to manufacturer's protocols. For the analysis 20.000 cells were measured and single cells were gated on size and shape using forward and side scatter. All experiments were performed using a BD LSRFortessa Cell Analyzer with BD FACSDiva Software. Data analysis was performed using FlowJo Software version 10. These experiments were performed three times, using independent MMECs isolations.

3.5.3 Nucleic acid isolation and analysis

Genomic DNA isolation, PCR amplification and TIDE analysis Genomic DNA was isolated from frozen mammary tumor pieces by proteinase K lysis and organic extraction with phenol-chloroform. Genomic DNA from cell pellets was isolated using the Gentra Puregene genomic DNA isolation kit (Qiagen 158767). Genomic DNA from FFPE mammary tissue was isolated using the Allprep DNA/RNA FFPE kit (Qiagen) according to the manufacturer's protocol. PCR amplification of *Myh9* exon 2 was performed with specific primers spanning the target site (FW_*Myh9*: CCCGGATGATGCACTTACTT; RV_*Myh9*: GCAGGGAACAGAGGGTAATG, product size 655 bp) and 1 μg DNA template, using the Q5 High-Fidelity PCR kit (NEB M0492). Amplicons were run on 1.5 % agarose gel and gel purification was performed using the Isolate II PCR and Gel kit (Bioline). PCR products were Sanger sequenced using the FW primer and CRISPR/Cas9-induced editing efficacy was quantified with the TIDE algorithm[40,*]. Cells isolated from non-injected glands were used as a negative control in all genomic DNA amplifications and only TIDE outputs with $R^2 > 0.9$ were considered. Transposon-gene

[*] https://tide.nki.nl

(*Myh9* and *Trp53bp2*) fusions were detected by standard PCR with two cell line specific primers and a transposon specific primer. The following primer sequences were used: Tumor 1 (11KOU35, *Myh9*) forward 5'-AAAGCCAAGTGACTCATATCC-3', reverse 5'-CAGTAGGCTTTCCCAGACC-3' (wildtype band [WT] 244 bp and transposon-gene junction band [TR] 408 bp); Tumor 2 (11KOU58, *Myh9*) forward 5'-CTTGGCTATGTAAGAGTGCG-3', reverse 5'-TCCTTTGCAACTTGTTCTGTC-3' (WT 190 bp and TR 402 bp); Tumor 3 (12SKA18, *Myh9*) forward 5'-CTACACCCAAGACCCCTTC-3', reverse 5'-ATCAATGCTGGGTATGGTG-3' (WT 204 bp and TR 327 bp); Tumor 4 (13SKA1, *Myh9*) forward 5'-TAGACCTAGCTACAGACACAG-3', reverse 5'- TACAGATCTGGGTACATGTTTG-3' (WT 242 bp and TR 421 bp); Tumor 1 (11KOU31-R5, *Trp53bp2*) forward 5;-CCTTCTCACTGTTTCATAGCAC-3', reverse 5'-GGTGCTGTTGTGGATTCTG-3' (WT 295 bp and TR 452 bp); Tumor 2 (13SKA10-R2, *Trp53bp2*) forward 5'-GTTTCAGGGTATTGTCTTTGG-3', reverse 5'-GATGATAACATGCACTGTGTG-3' (WT 201 bp and TR 394 bp) and transposon-specific primer 5'-GCTTGTGGAAGGCTACTCGAA-3'. Inversion of the CAG promoter of the $Ppp1r12a^{ex1\text{-}9}$ and $Trp53bp2^{ex13\text{-}18}$ conditional alleles were detected as described by Huijbers *et al.*[39]. $Cdh1^{F}$ and $Cdh1^{\Delta}$ alleles were identified by PCR as previously described[10].

RNA preparation and sequencing Whole tumor pieces were processed as previously described[23]. Illumina TruSeq mRNA libraries were sequenced with 50 base single reads on a HiSeq2000 using V3 chemistry (Illumina Inc., San Diego). Sequence reads were trimmed using Cutadapt[61] (version 1.12), filtering any reads shorter than 20 bp after trimming. The trimmed reads were aligned to the mm10 reference genome using STAR[62] (version 2.5.2b). Gene expression counts were generated by featureCounts[63] (version 1.5.0-post3) using gene definitions from Ensembl 76.

Northern blotting Tumor tissues were homogenized in TRIzol reagent (15596-018, Ambion life technologies) using a polytron (DI 18 Disperser, IKA) according to the manufacturer's protocol. Polyadenylated RNA was purified from total RNA using Dynabeads Oligo(dT)25 (61005, ThermoFisher scientific) according to the manufacturer's protocol. 500 ng of polyadenylated RNA was separated on 1 % agarose formaldehyde gels and transferred to nitrocellulose membranes as described by Sambrook *et al.*[64]. The primers used to obtain the cDNA probes are listed in Supplementary Table S3.3.

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[22]. The mouse read counts were normalized by correcting for sequencing depth using the DESeq's[65] 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[66], 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[21].

Identification of molecular mILC subtypes The molecular mILC subtypes were identified using the NMF package[67] (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 Figure S3.14). 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[7] 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[28]. 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[61] (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 (version 2.2.8)[68]. 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[29]. 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)[30], 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:

$$b_g = \sum_i I(i) w_i \quad \text{with} \quad i \in T_g$$

where b_g is the bias of gene g, T_g is the collection of insertions in gene g, w_i 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 b_g close to 1 are biased toward sense insertions, whereas genes with b_g 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 (version 0.3.0)[69]. 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[31]. 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[7] were visualized using cBioPortal[70]. 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)[22]. After removing normal samples and duplicate patient samples, correlation scores and p values were calculated using Spearman's rank correlation statistic.

[†]http://firebrowse.org

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.

3.5.4 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).

Immunoblotting Protein lysates were made using lysis buffer (20 mmol Tris-HCl 8.0 pH, 150 mmol NaCl, 1 mmol EDTA, 1 mmol EGTA, 1 % Triton X-100, 0.5 % deoxycholate, 0.1 % SDS in milli-Q) complemented with protease and phosphatase inhibitors (Roche) and quantified using the BCA protein assay kit (Pierce). Protein lysates were loaded on a 3 % to 8 % gel Tris-acetate gradient gel (Invitrogen) and transferred overnight onto nitrocellulose membrane (Bio-Rad) in 1x transfer buffer (25 mmol Tris, 2 mol Glycine, 20 % methanol in demineralized water). Membranes were blocked in 5 % nonfat dry milk or 5 BSA in TBS-T (7.6 pH, 20 mmol Tris, 137 mmol NaCl, 0.005 % Tween-20 in demineralized water) and primary antibodies were incubated overnight in 5 % nonfat dry milk or 5 % BSA in TBS-T. Membranes were washed three times and incubated with the secondary antibodies goat anti-rabbit-HRP (1:2000, Dako P0448), rabbit anti-mouse-HRP (1:5000, Dako P0260) or donkey anti-mouse IRDye 680nm (1:5000, Li_COR 926 32222) in 5 % BSA in TBS-T. Stained membranes were washed three times in TBS-T and then developed using ECL (Pierce 32209) or captured using the Li-Cor Odyssey Infrared Imaging System and analyzed using Odyssey Application software version 3.0.16.

Immunoprecipitation and LC/MS-MS analysis HC11 cells were transduced with pBABE-puro, pBABE-*Ppp1r12a*$^{ex1\text{-}9}$ or pBABE-*Trp53bp2*$^{ex13\text{-}18}$ as previously

described[10]. Protein lysates were made using a NP40 lysis buffer (50 mmol Tris-HCl (8.0 pH), 150 mmol NaCl, 1 % NP40 and 10 % Glycerol). Protein G Dynabeads (Invitrogen) were prepared according to manufacturer's protocol and incubated for 2 hours at 4 degrees with 10 μg of rabbit anti-FLAG antibody (Sigma). The lysates (200 mg) with beads were incubated overnight and washed five times with NP40 lysis buffer. Beads were reconstituted in 30 μL XT sample buffer (Bio-Rad) and heated at 95 °C for 7 min. Bead eluates were subsequently cleared and concentrated into a small protein band by short, partial SDS-PAGE separation[71]. The gel was stained with GelCode Blue stain reagent, followed by excision of the gel bands, reduction of the proteins with DTT and alkylation with iodoacetamide. Proteins were digested with sequencing grade trypsin (3 ng μL^{-1}) overnight at 37 °C. Peptides were extracted with acetonitrile, dried down in a speed vacuum centrifuge and reconstituted in 10 % formic acid prior to mass spectrometry analysis.

Peptides were separated in 65-min gradients using the Proxeon nLC 1000 system (Thermo Scientific, Bremen) equipped with a 75 μm x 500 mm PicoFrit analytical column fitted with a PicoTip Emitter (New Objective, Woburn, MA, USA), packed in-house with ReproSil-Pur 120 C18-AQ 2.4 μm stationary phase (Dr. Maisch). The outlet of the analytical column was coupled directly to a Thermo Orbitrap Fusion (Q-OT-qIT, Thermo Scientific) using the Proxeon nanoflex source. Nanospray was achieved using a liquid junction (Thermo) operated at 1.6 kV; additional settings were as previously described[72].

Raw data were analyzed by MaxQuant[73] (version 1.5.0.30) in label-free quantitation mode, using standard settings. MS/MS data were searched against the Mus Musculus Swissprot database (16,719 entries, release 2015_10) complemented with a list of common contaminants. Trypsin/P was chosen as cleavage specificity allowing two missed cleavages. Carbamidomethylation (C) was set as a fixed modification, and oxidation (M) was used as variable modification. For label-free quantitation (LFQ) analysis, LFQ quantitation and 'match between runs' were enabled. LFQ intensities (3 independent replicate values per sample) were $\log_2$-transformed in Perseus, after which proteins were filtered for at least three valid values in one of the sample groups. Missing values were replaced by imputation based on a normal distribution using a width of 0.3 and a downshift of 1.8[74]. Volcano plots were constructed for each immunoprecipitation using a permutation-based FDR-corrected t-test. Proteins were considered interactors when $p < 0.01$ and $\log_2(\text{Abundance difference}) > 1$.

3.5.5 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 $\log_2(\text{abundance difference}) > 1$. 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 $\pm$ 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.

3.5.6 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.

3.5.7 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 10.6084/M9.FIGSHARE.4765111[¶]. Processed expression and insertion data have also been deposited in Figshare under the identifier 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.

[‡]https://github.com/jrderuiter/ilc-sb-screen
[§]https://github.com/jrderuiter/pyim
[¶]https://doi.org/10.6084/M9.FIGSHARE.4765111
[||]https://doi.org/10.6084/M9.FIGSHARE.4765111
[**]https://github.com/jrderuiter/ilc-sb-screen

3.6 Supplementary Material

3.6.1 Supplementary Figures

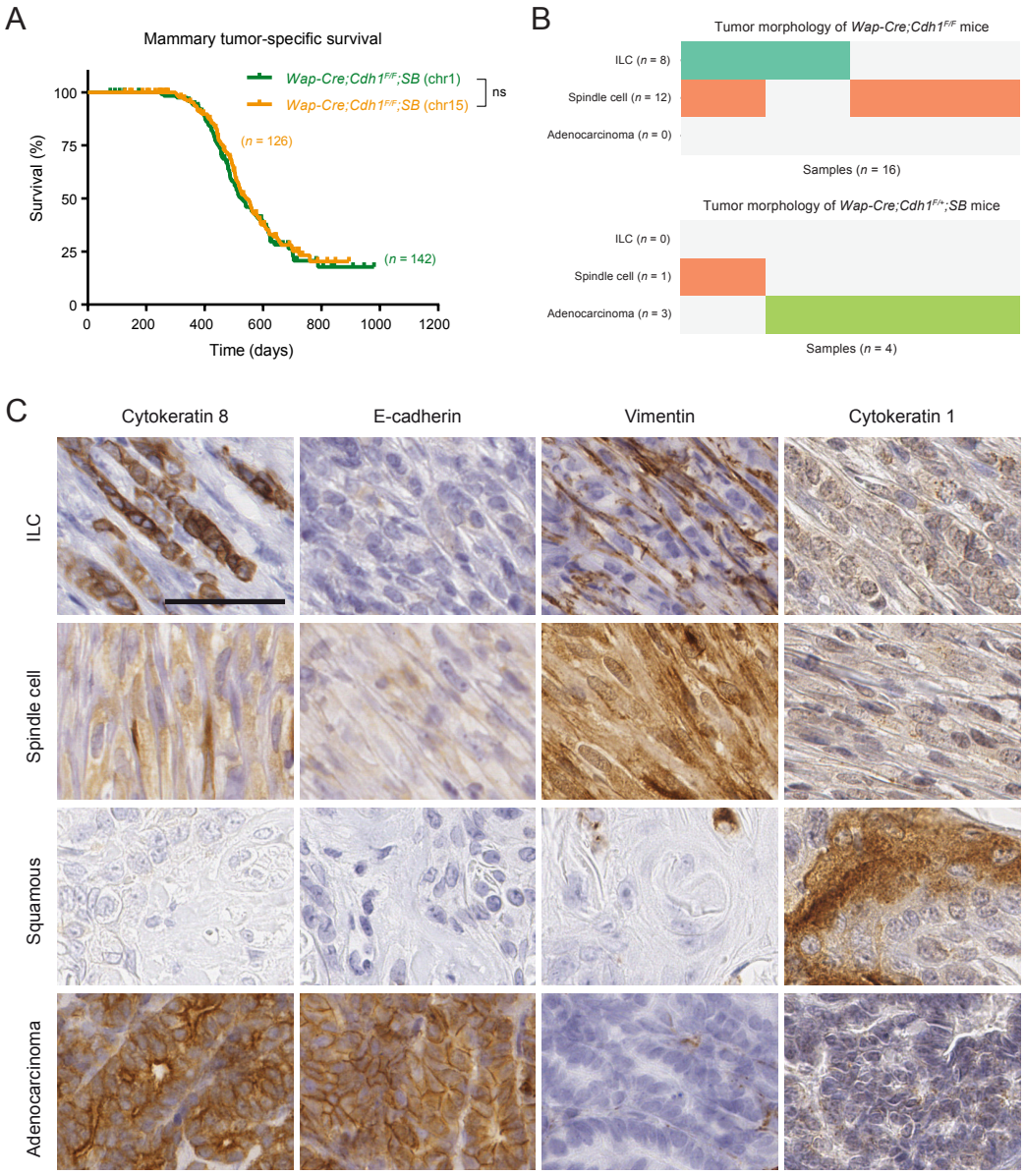


Fig. S3.1. Survival curves for the individual *T2/Onc* transgenic mouse lines, distribution and immunohistochemical stainings of the different tumor morphologies. (A) Survival curves *Wap-Cre;Cdh1$^{F/F}$;SB* mice carrying the *T2/Onc* transposon donor loci on chromosomes 1 and 15 ($n = 142$ and $n = 126$ 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.

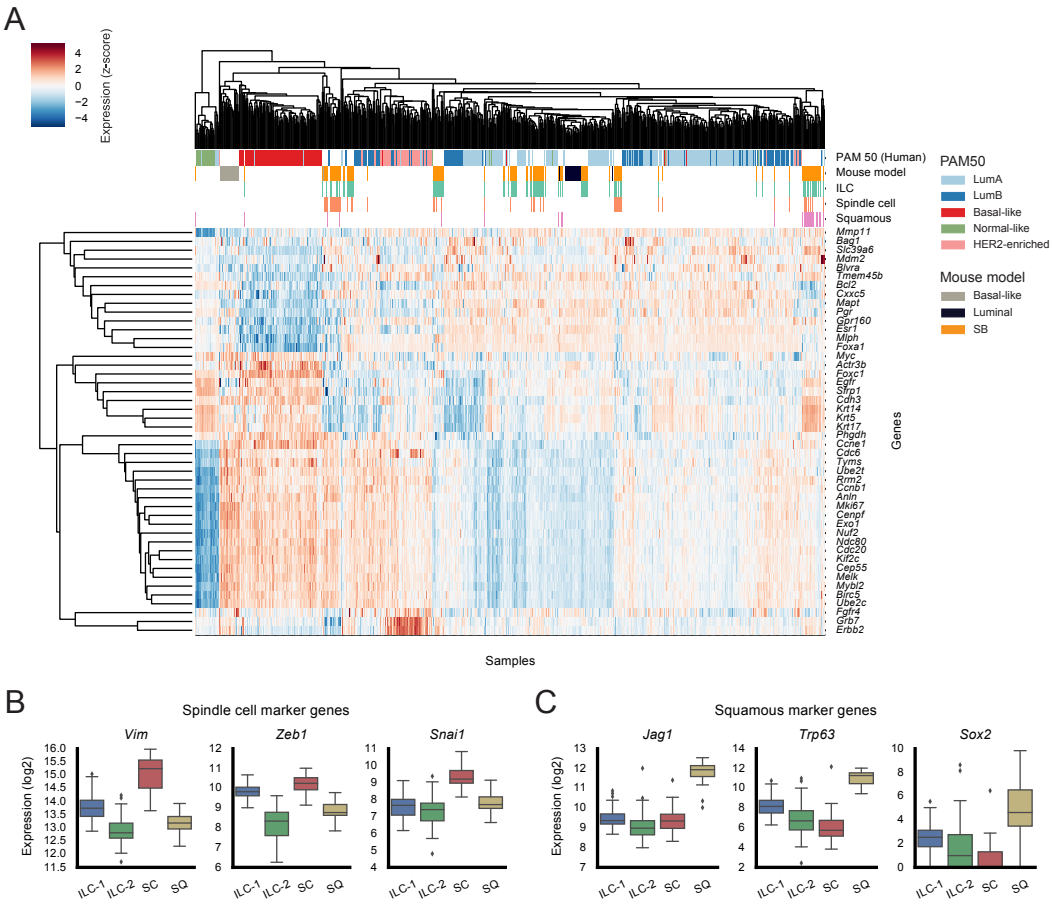


Fig. S3.2. Additional plots accompanying the PAM50 gene expression analyses. **(A)** Unsupervised clustering (Euclidean distance, average linkage) of the SB-induced tumors with reference mouse models and human breast cancer samples from TCGA, showing the expression of all 46 orthologous mouse genes from 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.

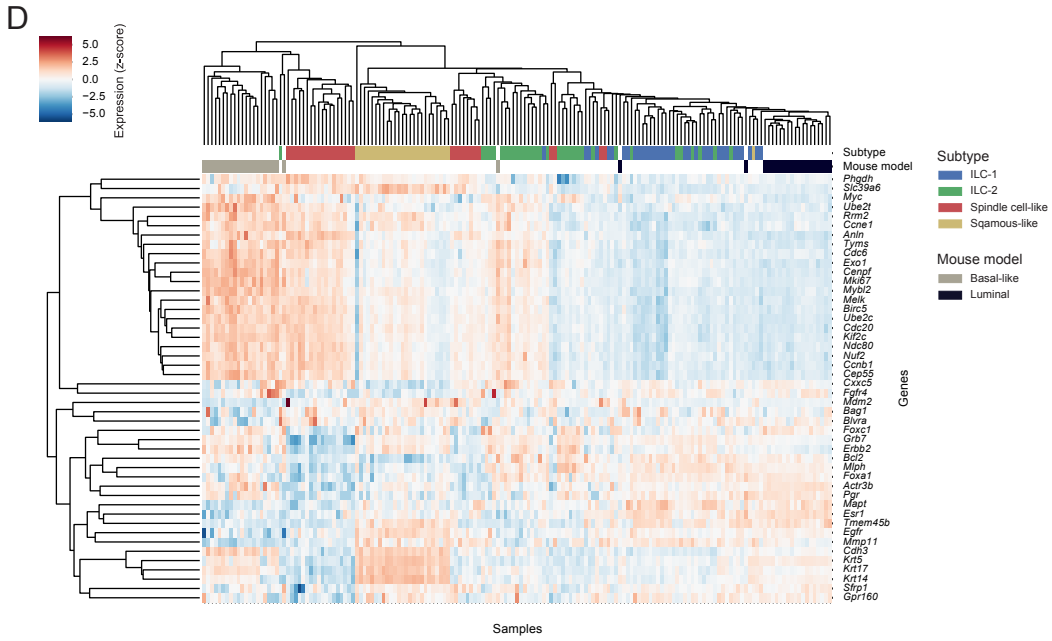


Fig. S3.2. **Continued. (D)** Unsupervised clustering (Euclidean distance, average linkage) of the SB-induced molecular subtypes with the reference luminal and basal-like mouse models.

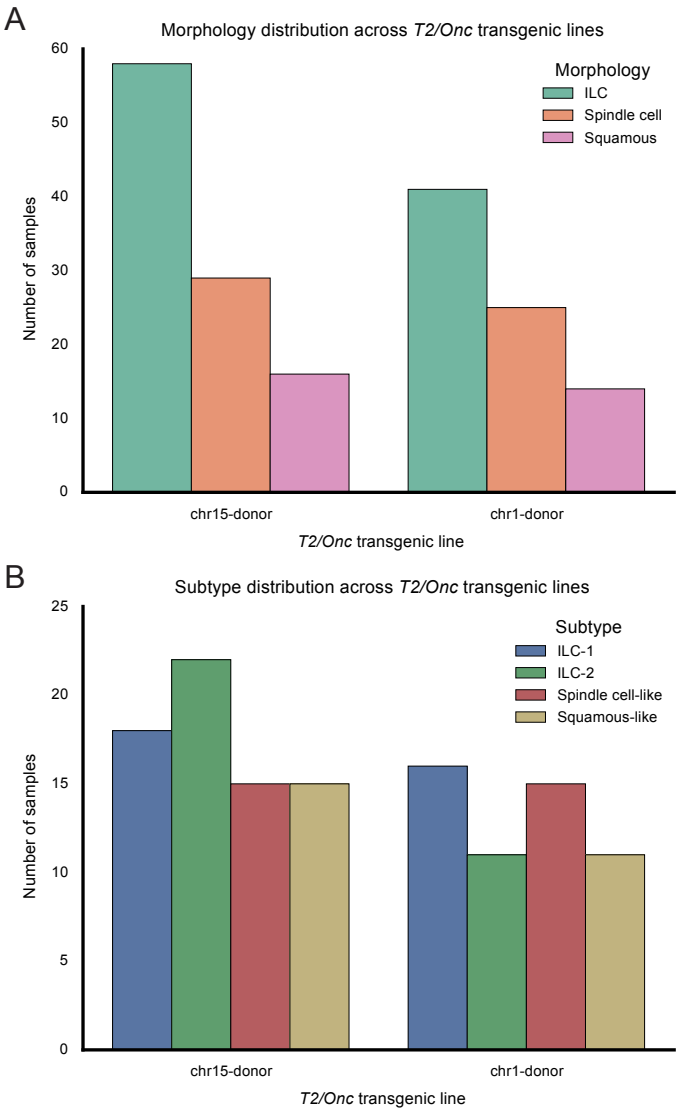


Fig. S3.3. Distribution of morphology and expression subtypes across the *T2/Onc* transgenic mouse lines. (A) Distribution of morphology across the two *T2/Onc* transgenic lines, showing that there is no significant bias between the two lines. Lack of any significant bias was confirmed using pairwise 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.

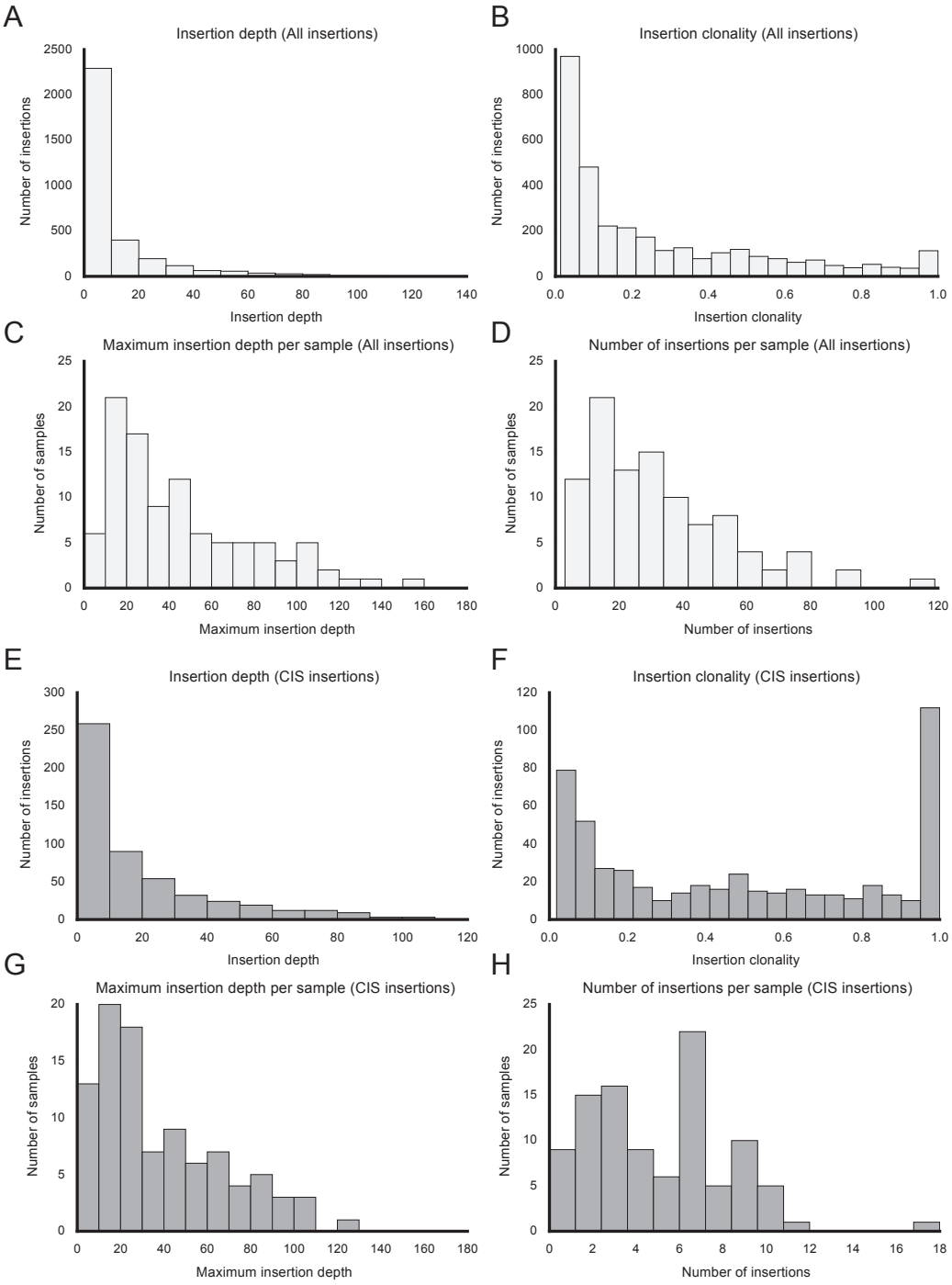


Fig. S3.4. Histograms detailing various statistics of the identified insertions, both for all insertions and for insertions within CISs. **(A)** Distribution of the insertion depths (as indicated by ShearSplink's ULP score) over all insertions. **(B)** Distribution of insertion clonality scores. **(C)** 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.

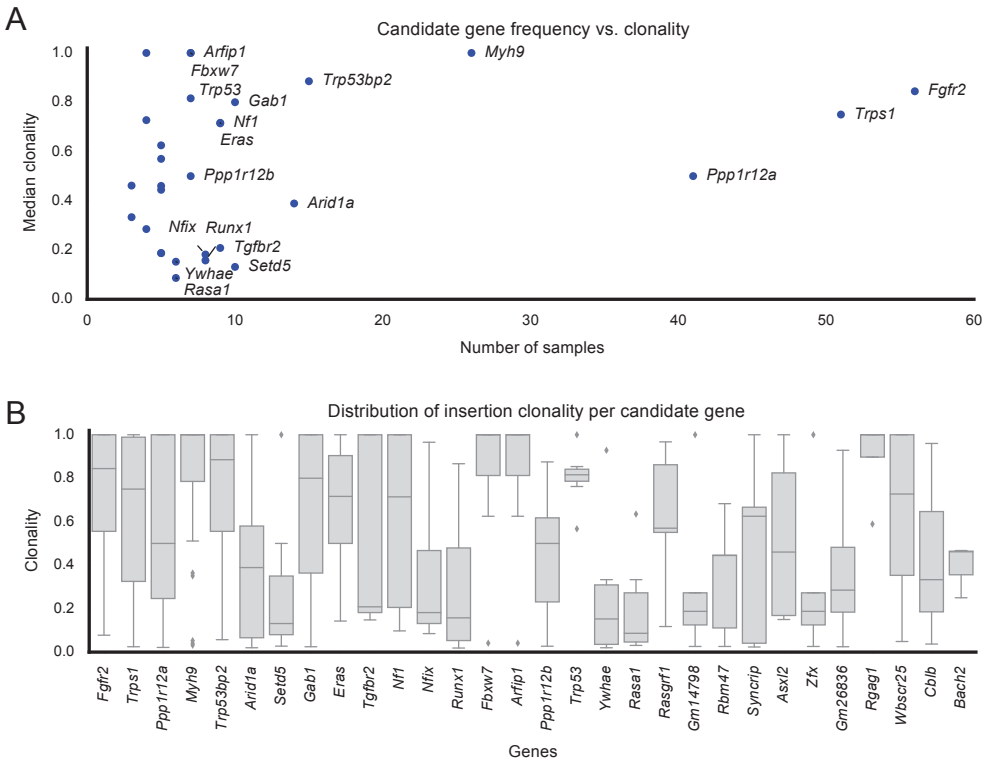


Fig. S3.5. Additional details on the clonality of the candidate genes. (A) Ranking of genes by their overall frequency and the median clonality of their insertions. For samples with multiple insertions in the same gene, we used the clonality of the strongest insertion to avoid underestimating the median 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.

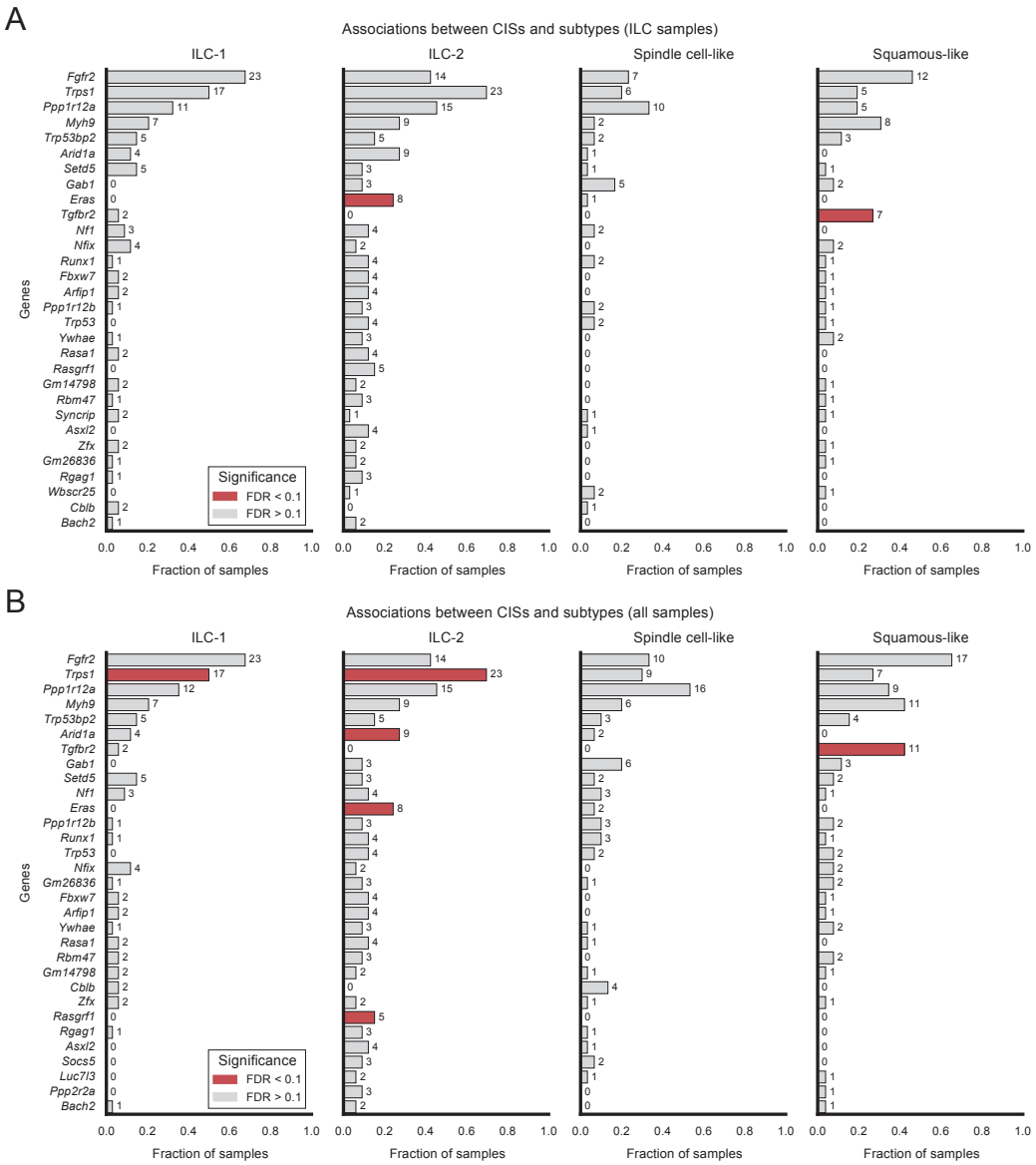


Fig. S3.6. Distribution of the candidate genes over the identified molecular subtypes. Bar plots indicating the distribution of all candidate genes over the molecular subtypes. Red bars indicate significant associations between candidate genes and the corresponding molecular subtype (FDR < 0.1, one-sided Fisher's exact test with Benjamini-Hochberg correction). 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.

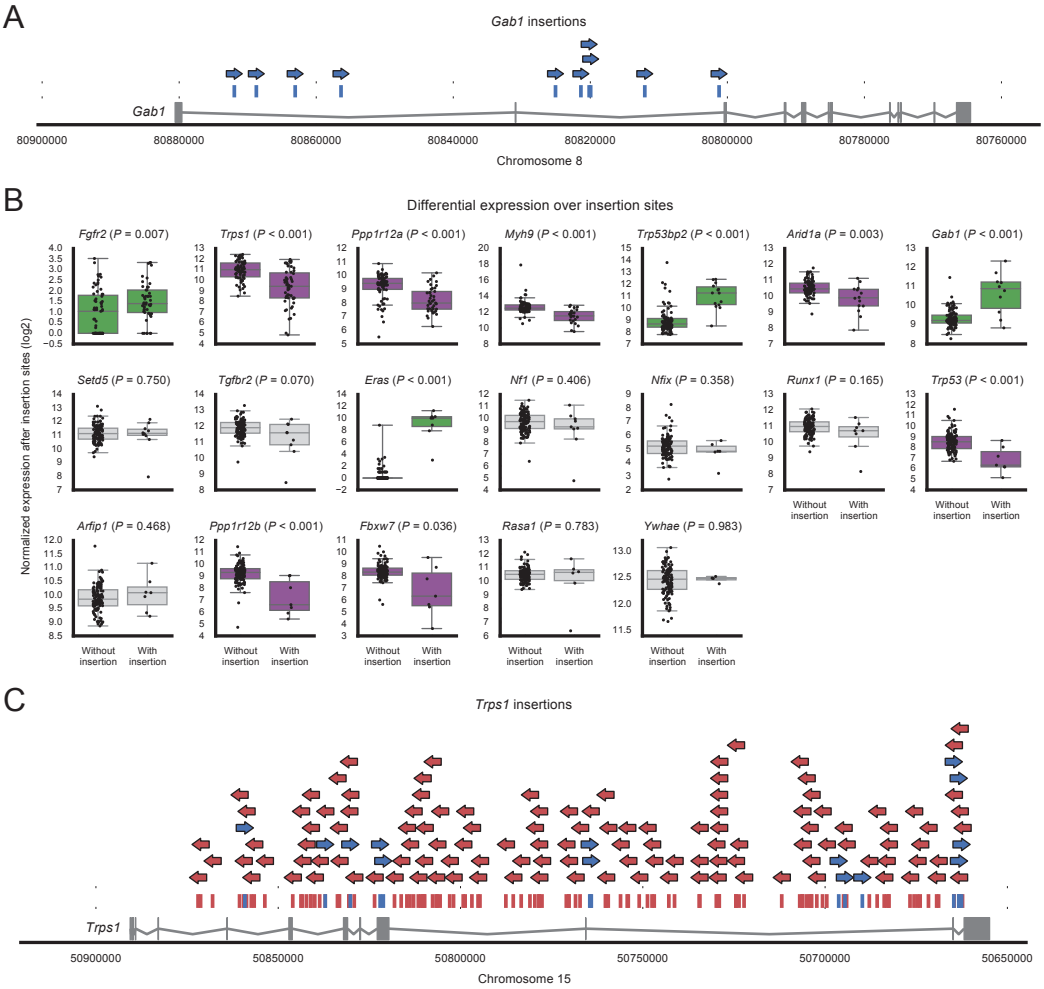


Fig. S3.7. Additional details on the insertion patterns and differential expression of the candidate genes. **(A)** Insertion pattern in *Gab1*, showing bias toward activating insertions. **(B)** Boxplots for the main candidate genes showing the difference in expression after the insertion sites in the 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.

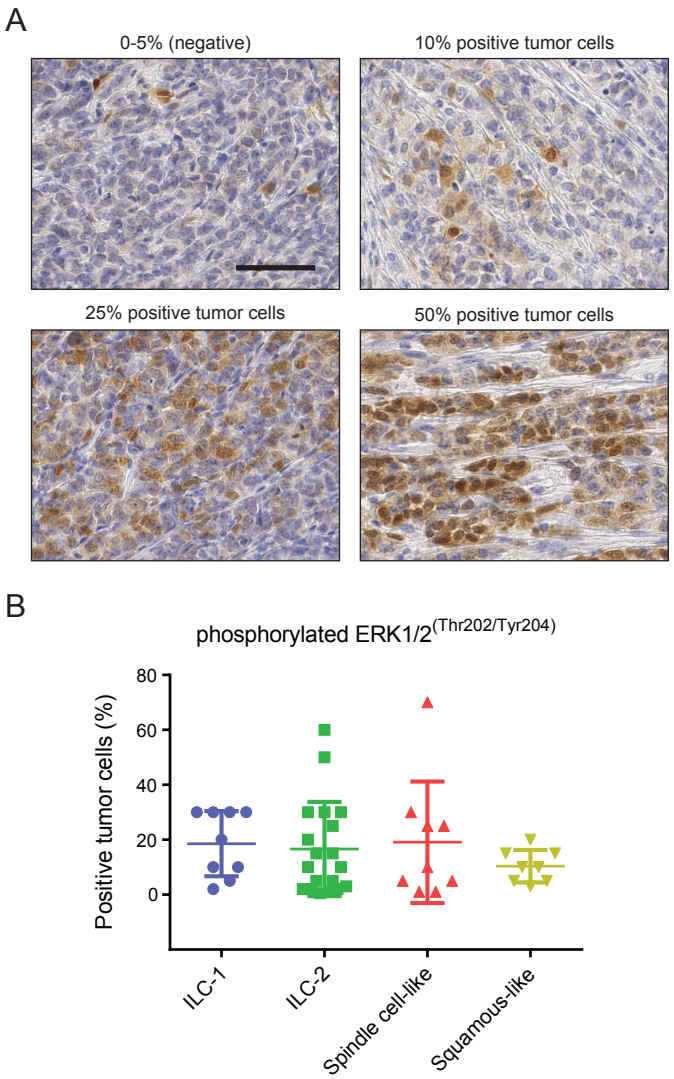


Fig. S3.8. Representative images and quantification of immunohistochemical staining for phosphorylated ERK1/2, a downstream protein of the RAS/MAPK signaling pathway. **(A)** Representative images of different percentages of phosphorylated ERK1/2$^{(Thr202/Tyr204)}$ staining in SB-induced tumors. Scale bar, 50 μm. **(B)** Quantification of phosphorylated ERK1/2$^{(Thr202/Tyr204)}$ staining in the different molecular subtypes. Mean $\pm$ s.d.

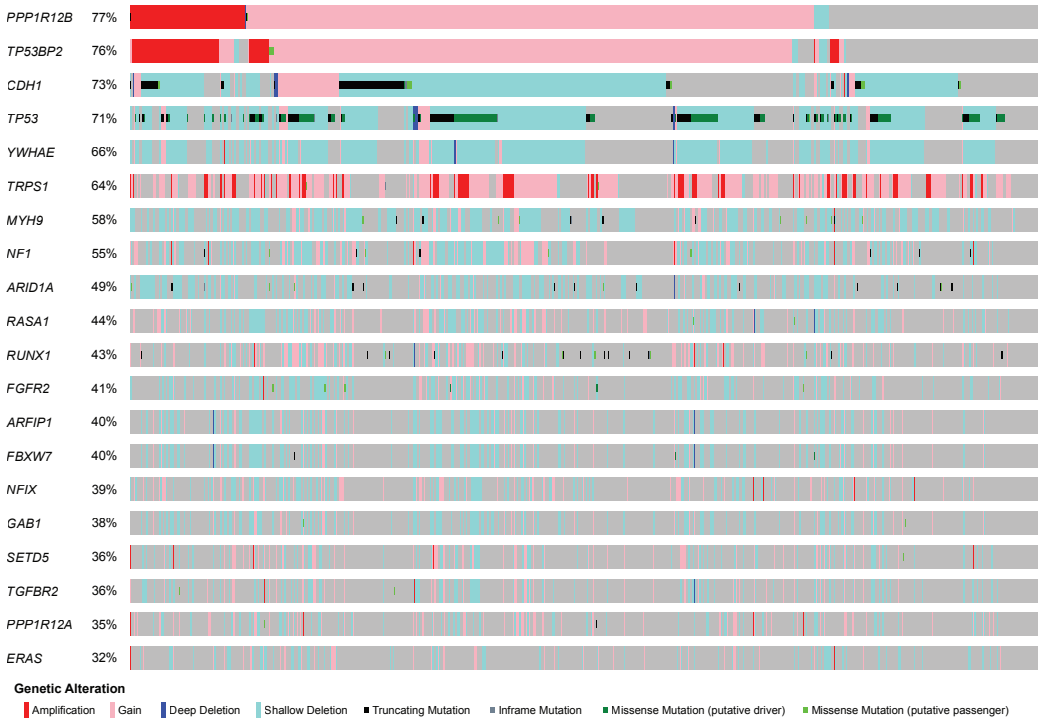


Fig. S3.9. Overview of the mutations and copy number events in the TGCA breast cancer data set (816 samples) for each of the main candidate genes and CDH1. Percentages indicate the fraction of tumors with alterations in the respective genes.

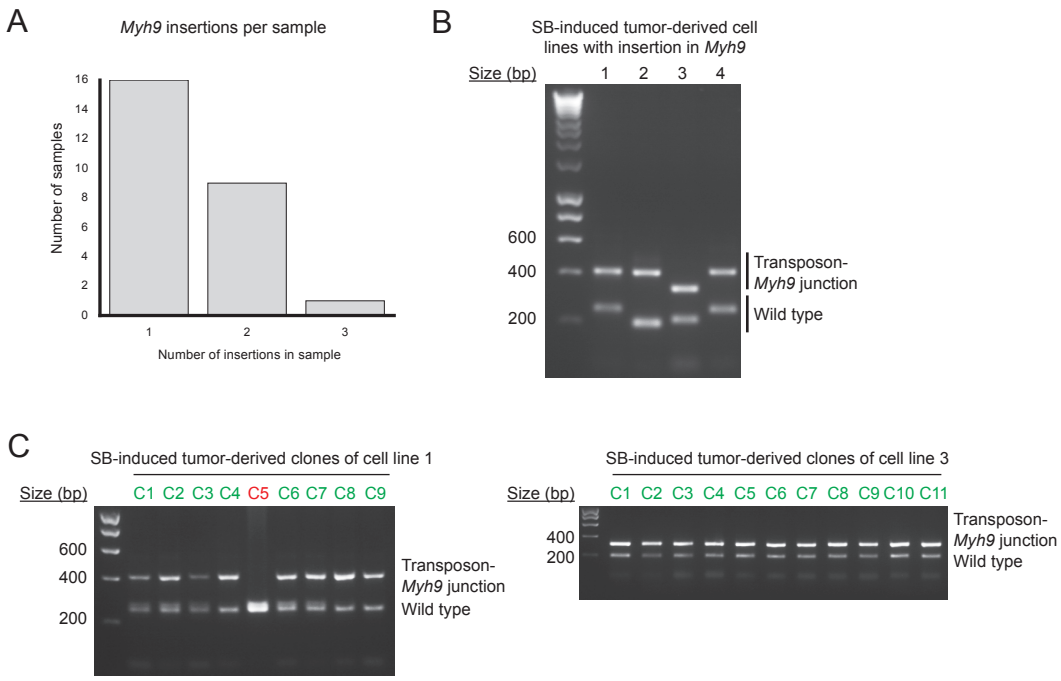


Fig. S3.10. ***Myh9* haploinsufficiency in ILC formation in SB-induced tumor-derived cell lines with or without *Myh9* insertion.** **(A)** Distribution of the number of insertions per sample in *Myh9*, showing that the majority of the tumors show single insertions in the gene. **(B)** PCR amplification 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.

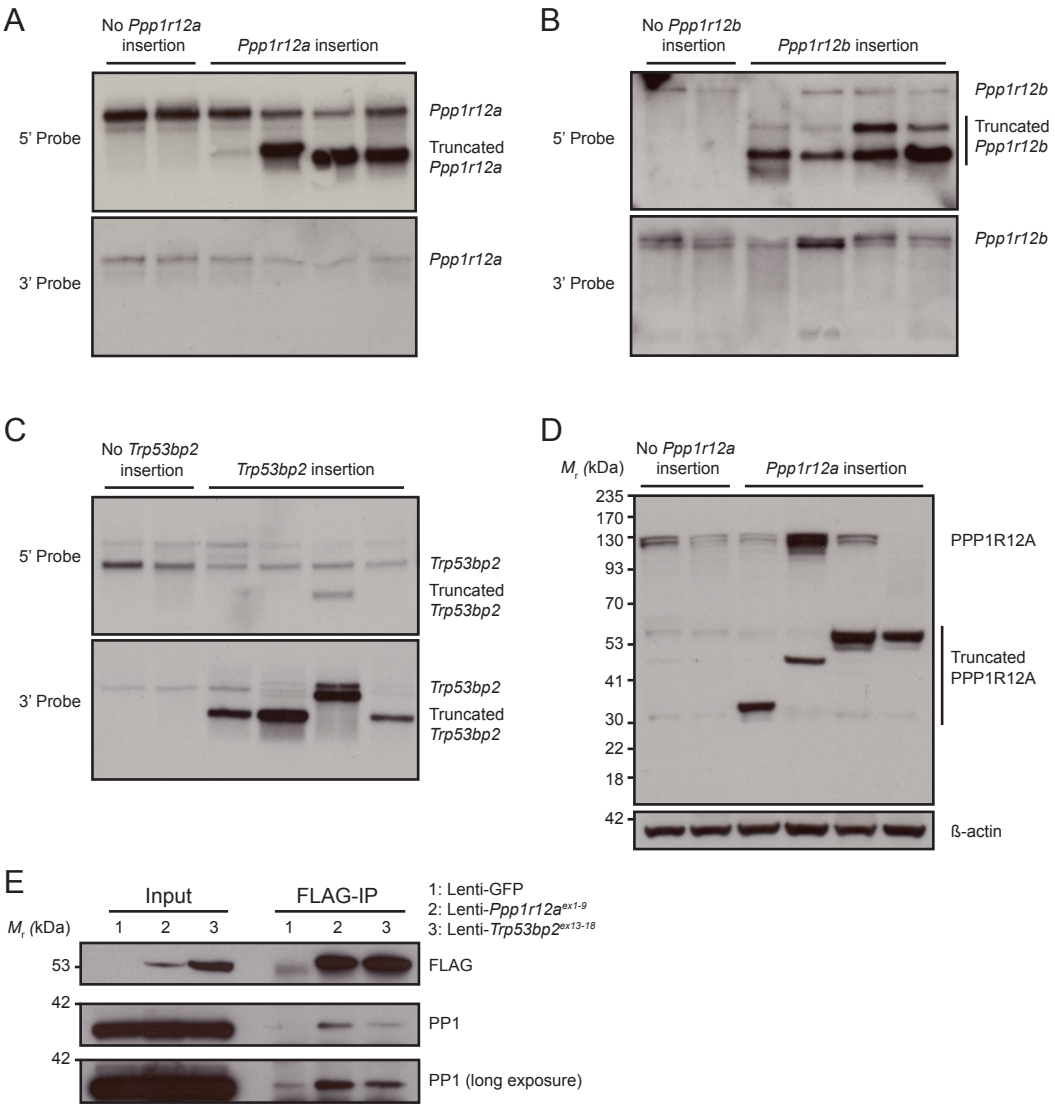


Fig. S3.11. Expression of truncated *Ppp1r12a/b* and *Trp53bp2* and PPP1R12A protein expression in SB-induced tumors. (A–C) Truncated *Ppp1r12a/b* and *Trp53bp2* in SB-induced tumors, as visualized by northern blot analysis. **(D)** Expression of PPP1R12A in SB-induced tumors with and without insertions in *Ppp1r12a*, as visualized by immunoblotting using an anti-PPP1R12A antibody. β-actin is shown as a loading control. **(E)** Co-immunoprecipitation of PP1 in HC11 cells expressing GFP, truncated PPP1R12A or TRP53BP2, as visualized by immunoblotting using anti-FLAG and anti-PP1 antibodies.

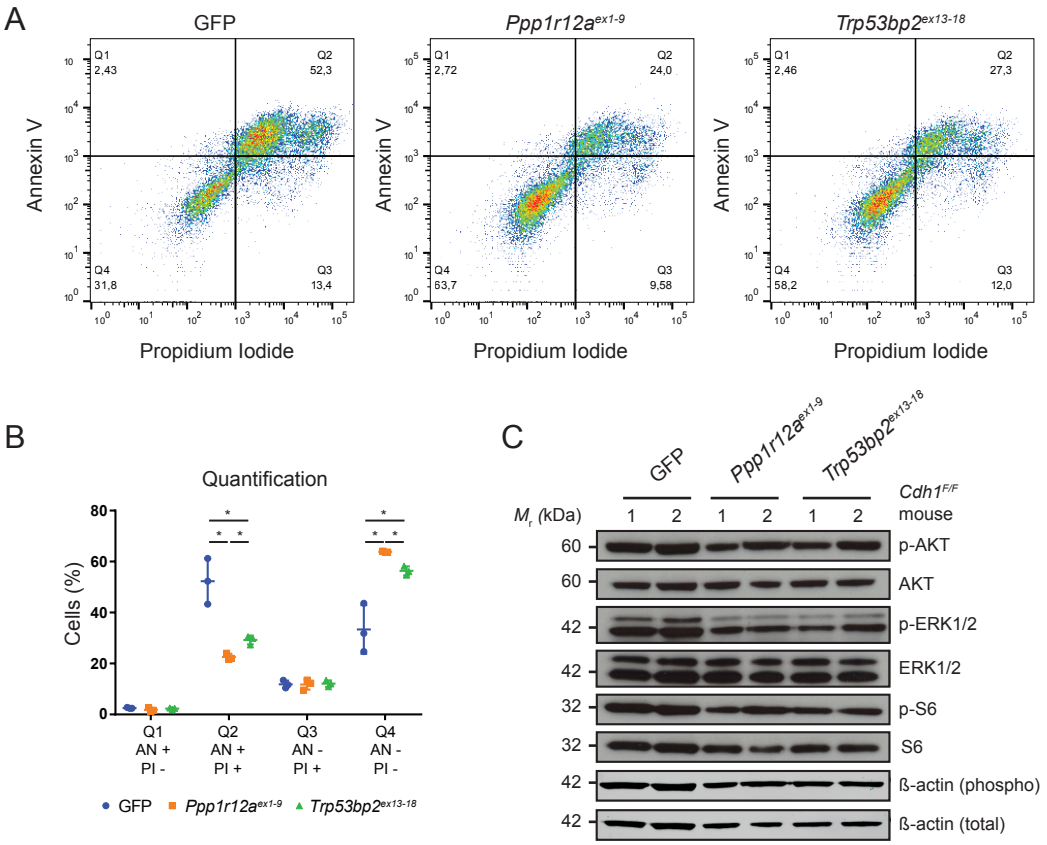


Fig. S3.12. Characterization of cell death, cell survival markers and activation of a p53 response in E-cadherin-deficient cells expressing truncated PPP1R12A or TRP53BP2 or showing reduced levels of MYH9. (A) Representative dot plots depicting the proportions of Annexin V (AN)- and/or propidium iodide (PI)-positive cells 72 h after seeding of AdCre-transduced *Cdh1*$^{F/F}$ MMECs with simultaneous transduction of Lenti-GFP, Lenti-*Ppp1r12a*$^{ex1\text{-}9}$ or Lenti-*Trp53bp2*$^{ex13\text{-}18}$. **(B)** Quantification of Annexin V– and PI-positive cells from AdCre-transduced *Cdh1*$^{F/F}$ MMECs with simultaneous transduction of Lenti-GFP, Lenti-*Ppp1r12a*$^{ex1\text{-}9}$ or Lenti-*Trp53bp2*$^{ex13\text{-}18}$. 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*$^{ex1\text{-}9}$ or Lenti-*Trp53bp2*$^{ex13\text{-}18}$, as visualized by immunoblotting. β-actin is shown as a loading control.

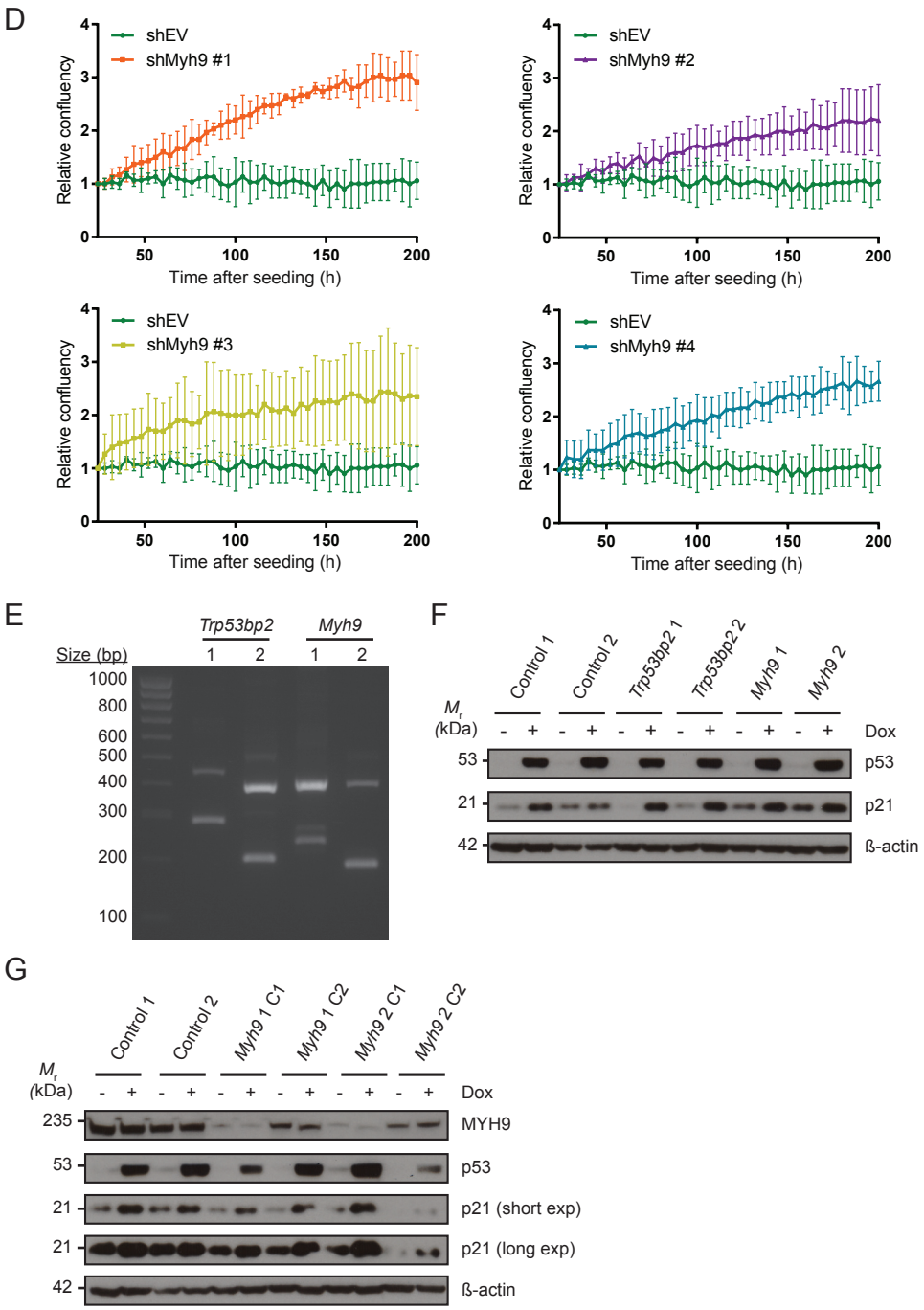


Fig. S3.12. **Continued.** **(D)** Cell survival of AdCre-transduced *Cdh1*$^{F/F}$*;mT/mG* MMECs with simultaneous shRNA- mediated knockdown of Myh9 with different shRNAs, as quantified by using real-time IncuCyte imaging for 200 h. Mean ± s.d. of three independent experiments. **(E)** PCR amplification 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 μmol). β-actin is shown as a loading control. **(G)** Expression of p53 and p21 in SB-induced tumor- derived 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 μmol). β-actin is shown as a loading control. Exp, exposure.

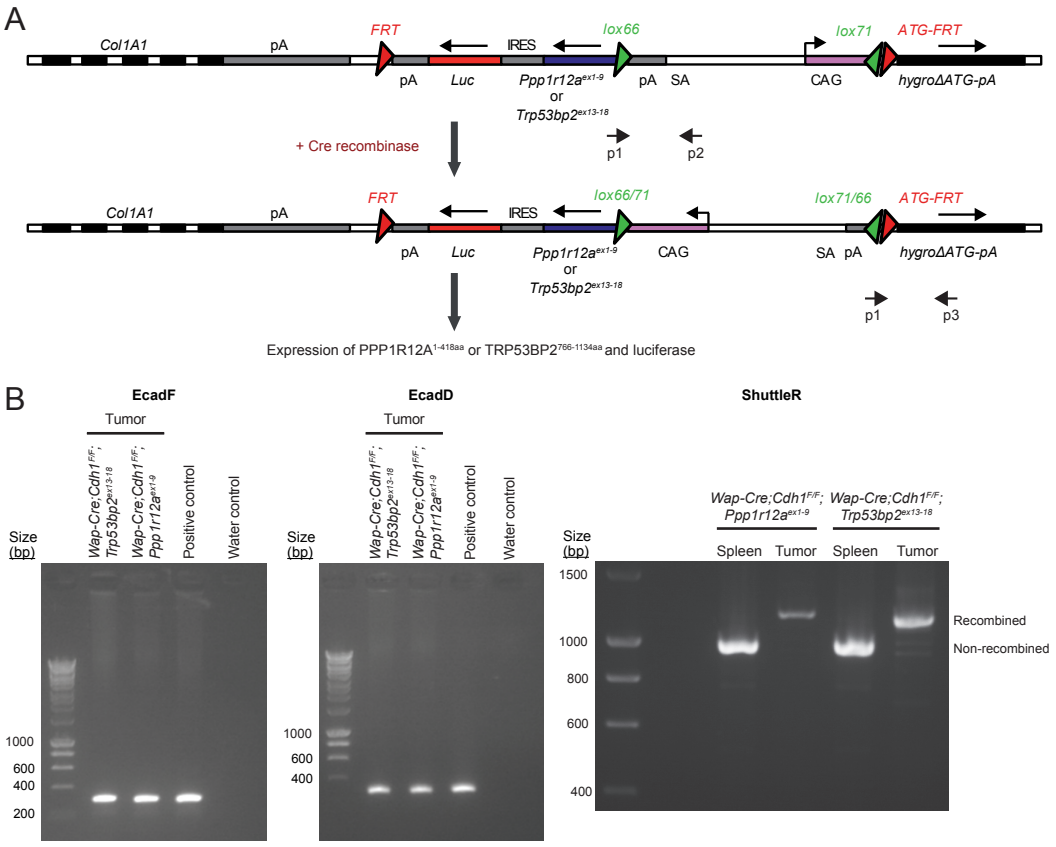


Fig. S3.13. Overview of Cre-conditional alleles and distribution of the tumor morphology in the different genetically engineered mouse models. **(A)** Depiction of Cre-conditional *invCAG-Ppp1r12a*$^{ex1-9}$*-IRES-Luc* and *invCAG-Trp53bp2*$^{ex13-18}$*-IRES-Luc* alleles in the Col1a1 locus. Cre-mediated recombination allows inversion of the CAG promoter, resulting in expression of PPP1R12A$^{1-418aa}$ or TRP53BP2$^{766-1134aa}$ accompanied by luciferase expression. **(B)** Recombination status of the Cre-conditional alleles in tumors from *Wap-Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1-9}$ and *Wap-Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13-18}$ 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).

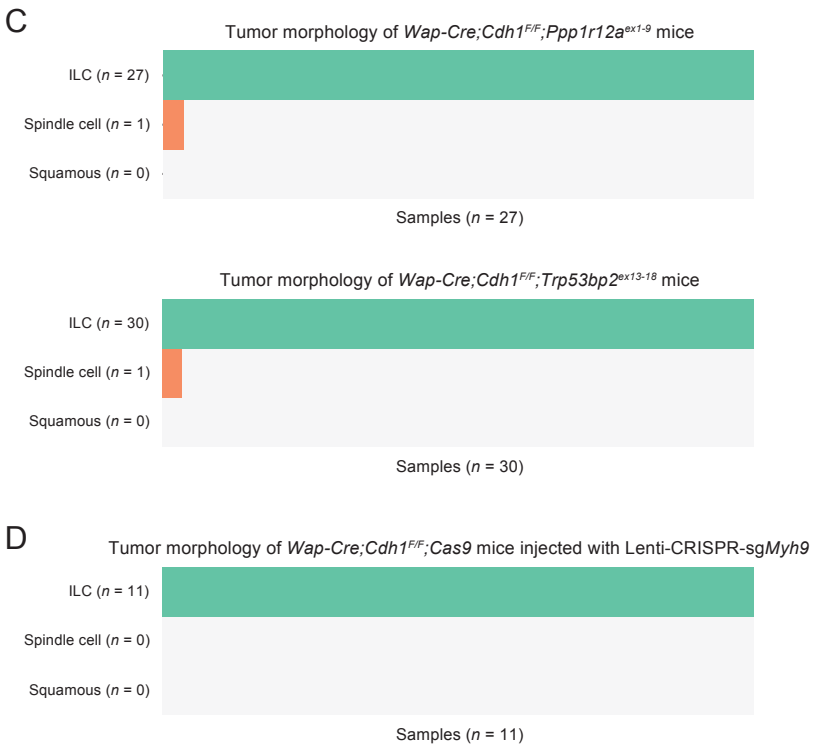


Fig. S3.13. **Continued.** **(C)** Histological classification of tumors from *Wap-Cre;Cdh1*$^{F/F}$*;Ppp1r12a*$^{ex1\text{-}9}$ ($n = 27$) and *Wap-Cre;Cdh1*$^{F/F}$*;Trp53bp2*$^{ex13\text{-}18}$ ($n = 30$) mice. **(D)** Histological classification of tumors ($n = 11$) from *Wap-Cre;Cdh1*$^{F/F}$*;Cas9* mice injected with Lenti-CRISPR-sg*Myh9*.

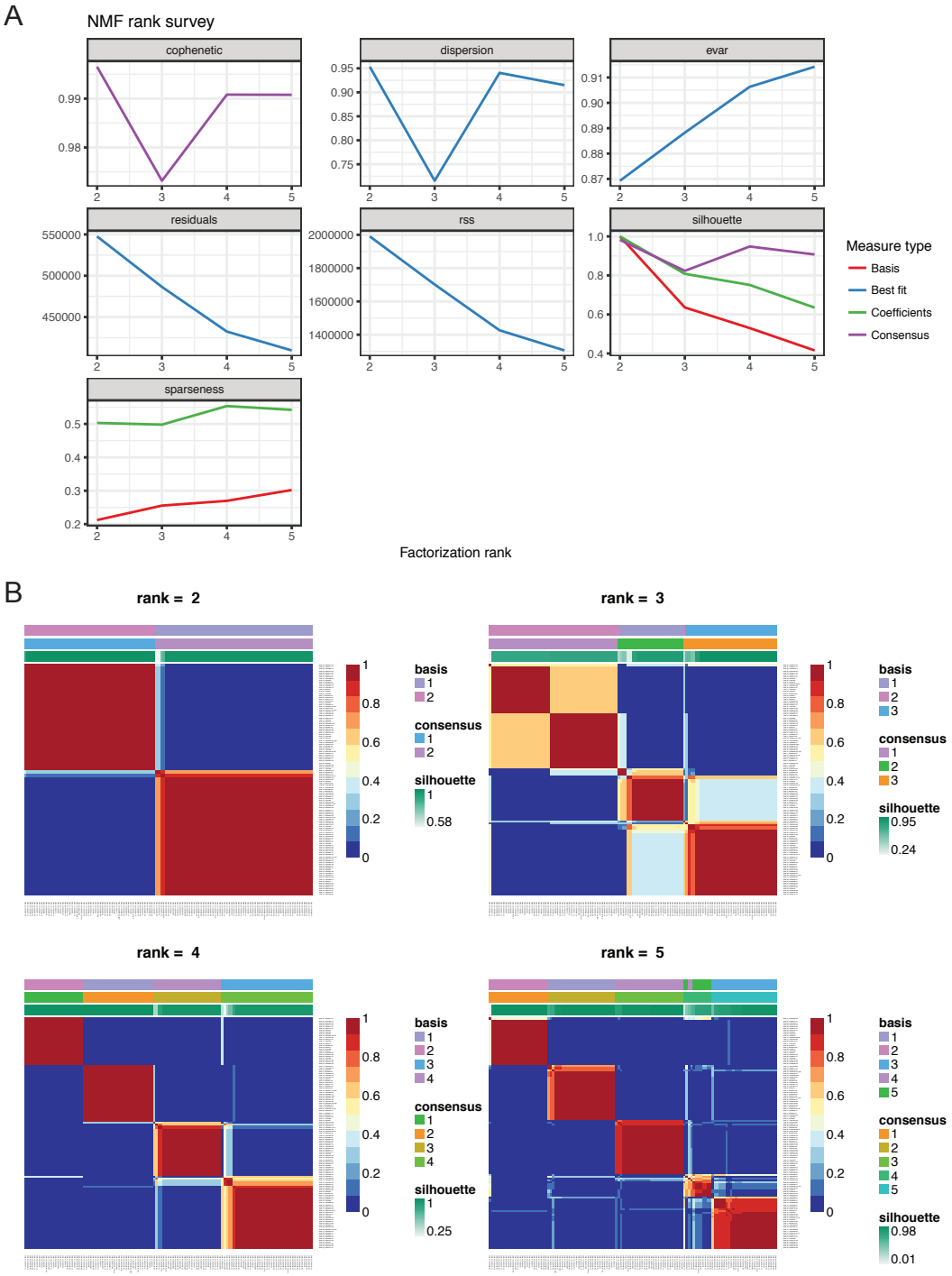


Fig. S3.14. Clustering statistics for different numbers of clusters in the NMF subtype analysis. (A) Overview of various clustering statistics for different numbers of clusters (2–5) in the NMF analysis of the SB-induced tumors. **(B)** Consensus maps for the same cluster sizes, showing the consistency of cluster assignments for the different numbers of clusters.

3.6.2 Supplementary Tables (available online)

Tab. S3.1. **Overview of sample information including sample IDs, mouse IDs, *T2/Onc* lines, histopathology types, RNA-seq IDs, metastasis sites and subtypes.**

Tab. S3.2. **Overview of the identified candidate genes and various statistics for each gene.**

Tab. S3.3. **Overview of all primers used in this study for genotyping, cloning and northern blotting.**

Tab. S3.4. **Detailed information about antibodies and antigen retrieval methods used in immunohistochemical experiments.**

3.7 References

[1] V Martinez and JG Azzopardi. “Invasive lobular carcinoma of the breast: incidence and variants”. In: *Histopathology* 3.6 (1979), pp. 467–488 (cit. on p. 44).

[2] Marilyn J Borst and John A Ingold. “Metastatic patterns of invasive lobular versus invasive ductal carcinoma of the breast”. In: *Surgery* 114.4 (1993), pp. 637–642 (cit. on p. 44).

[3] Hilda Wong, Silvia Lau, Polly Cheung, 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”. In: *BMC Cancer* 14.1 (2014), p. 826 (cit. on p. 44).

[4] Carien M Niessen and Cara J Gottardi. “Molecular components of the adherens junction”. In: *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1778.3 (2008), pp. 562–571 (cit. on p. 44).

[5] Roland Moll, Margarete Mitze, Uwe H Frixen, and Walter Birchmeier. “Differential loss of E-cadherin expression in infiltrating ductal and lobular breast carcinomas.” In: *The American Journal of Pathology* 143.6 (1993), p. 1731 (cit. on p. 44).

[6] CB Vos, AM Cleton-Jansen, Geert Berx, et al. “E-cadherin inactivation in lobular carcinoma in situ of the breast: an early event in tumorigenesis”. In: *British Journal of Cancer* 76.9 (1997), p. 1131 (cit. on p. 44).

[7] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, et al. “Comprehensive molecular portraits of invasive lobular breast cancer”. In: *Cell* 163.2 (2015), pp. 506–519 (cit. on pp. 44, 65, 66, 74, 76).

[8] Emad A Rakha, Arjun Patel, Des G Powe, et al. “Clinical and biological significance of E-cadherin protein expression in invasive lobular carcinoma of the breast”. In: *The American Journal of Surgical Pathology* 34.10 (2010), pp. 1472–1479 (cit. on pp. 44, 53).

[9] Oréda Boussadia, Stefanie Kutsch, Andreas Hierholzer, Véronique Delmas, and Rolf Kemler. “E-cadherin is a survival factor for the lactating mouse mammary gland”. In: *Mechanisms of Development* 115.1-2 (2002), pp. 53–62 (cit. on p. 44).

[10] Patrick W B Derksen, Xiaoling Liu, Francis Saridin, et al. “Somatic inactivation of E-cadherin and p53 in mice leads to metastatic lobular mammary carcinoma through induction of anoikis resistance and angiogenesis”. In: *Cancer Cell* 10.5 (2006), pp. 437–449 (cit. on pp. 44, 53, 69, 73, 78).

[11] Patrick W B Derksen, Tanya M Braumuller, Eline van der Burg, et al. “Mammary-specific inactivation of E-cadherin and p53 impairs functional gland development and leads to pleomorphic invasive lobular carcinoma in mice”. In: *Disease Models and Mechanisms* 4.3 (2011), pp. 347–358 (cit. on pp. 44, 45, 53, 68, 69).

[12] Daniel E Stange, Bernhard Radlwimmer, Falk Schubert, et al. “High-resolution genomic profiling reveals association of chromosomal aberrations on 1q and 16p with histologic and genetic subgroups of invasive breast cancer”. In: *Clinical Cancer Research* 12.2 (2006), pp. 345–352 (cit. on pp. 44, 66).

[13] P T Simpson, J S Reis-Filho, M B K Lambros, et al. “Molecular profiling pleomorphic lobular carcinomas of the breast: evidence for a common molecular genetic pathway with classic lobular carcinomas”. In: *The Journal of Pathology* 215.3 (2008), pp. 231–244 (cit. on p. 44).

[14] Fiamma Buttitta, Lara Felicioni, Fabio Barassi, et al. “PIK3CA mutation and histological type in breast carcinoma: high frequency of mutations in lobular carcinoma”. In: *The Journal of Pathology* 208.3 (2006), pp. 350–355 (cit. on p. 44).

[15] Matthias Christgen, Monika Noskowicz, Elisa Schipper, et al. “Oncogenic PIK3CA mutations in lobular breast cancer progression”. In: *Genes, Chromosomes and Cancer* 52.1 (2013), pp. 69–80 (cit. on p. 44).

[16] Cigdem Ercan, Paul J Van Diest, Bram Van Der Ende, et al. “p53 mutations in classic and pleomorphic invasive lobular carcinoma of the breast”. In: *Cellular Oncology* 35.2 (2012), pp. 111–118 (cit. on pp. 44, 53, 65).

[17] Magali Michaut, Suet-Feung Chin, Ian Majewski, et al. “Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer”. In: *Scientific Reports* 6 (2016), p. 18517 (cit. on pp. 44, 53, 65, 66).

[18] Christine Desmedt, Gabriele Zoppoli, Gunes Gundem, et al. “Genomic characterization of primary invasive lobular breast cancer”. In: *Journal of Clinical Oncology* 34.16 (2016), pp. 1872–1881 (cit. on pp. 44, 53, 65, 66).

[19] Lara S Collier, Corey M Carlson, Shruthi Ravimohan, Adam J Dupuy, and David A Largaespada. “Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse”. In: *Nature* 436.7048 (2005), pp. 272–276 (cit. on pp. 45, 46, 68, 69).

[20] H Nikki March, Alistair G Rust, Nicholas A Wright, et al. “Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis”. In: *Nature Genetics* 43.12 (2011), p. 1202 (cit. on pp. 45, 68).

[21] Joel S Parker, Michael Mullins, Maggie C U Cheang, et al. “Supervised risk predictor of breast cancer based on intrinsic subtypes”. In: *Journal of Clinical Oncology* 27.8 (2009), pp. 1160–1167 (cit. on pp. 47, 74).

[22] Cancer Genome Atlas Network et al. “Comprehensive molecular portraits of human breast tumours”. In: *Nature* 490.7418 (2012), p. 61 (cit. on pp. 47, 74, 76).

[23] Mirjam C Boelens, Micha Nethe, Sjoerd Klarenbeek, et al. “PTEN loss in E-cadherin-deficient mouse mammary epithelial cells rescues apoptosis and results in development of classical invasive lobular carcinoma”. In: *Cell Reports* 16.8 (2016), pp. 2087–2101 (cit. on pp. 47, 59, 70, 73).

[24] Xiaoling Liu, Henne Holstege, Hanneke van der Gulden, et al. “Somatic loss of BRCA1 and p53 in mice induces mammary tumors with features of human BRCA1-mutated basal-like breast cancer”. In: *Proceedings of the National Academy of Sciences* 104.29 (2007), pp. 12111–12116 (cit. on p. 47).

[25] Jason S Carroll, X Shirley Liu, Alexander S Brodsky, et al. “Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1”. In: *Cell* 122.1 (2005), pp. 33–43 (cit. on p. 49).

[26] Antoni Hurtado, Kelly A Holmes, Caryn S Ross-Innes, Dominic Schmidt, and Jason S Carroll. “FOXA1 is a key determinant of estrogen receptor function and endocrine response”. In: *Nature Genetics* 43.1 (2011), p. 27 (cit. on p. 49).

[27] A 2011 Goldhirsch, William C Wood, Alan S Coates, 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”. In: *Annals of Oncology* 22.8 (2011), pp. 1736–1747 (cit. on p. 49).

[28] Marco J Koudijs, Christiaan Klijn, Louise van der Weyden, et al. “High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors”. In: *Genome Research* 21.12 (2011), pp. 2181–2189 (cit. on pp. 50, 74).

[29] Jeroen De Ridder, Anthony Uren, Jaap Kool, Marcel Reinders, and Lodewyk Wessels. “Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens”. In: *PLoS Computational Biology* 2.12 (2006), e166 (cit. on pp. 50, 75).

[30] Johann de Jong, Jeroen de Ridder, Louise van der Weyden, et al. “Computational identification of insertional mutagenesis targets for cancer gene discovery”. In: *Nucleic Acids Research* 39.15 (2011), e105–e105 (cit. on pp. 50, 75).

[31] Sander Canisius, John WM Martens, and Lodewyk FA Wessels. “A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence”. In: *Genome Biology* 17.1 (2016), p. 261 (cit. on pp. 54, 76).

[32] Michael E Grassie, Lori D Moffat, Michael P Walsh, and Justin A MacDonald. “The myosin phosphatase targeting protein (MYPT) family: a regulated mechanism for achieving substrate specificity of the catalytic subunit of protein phosphatase type 1δ”. In: *Archives of Biochemistry and Biophysics* 510.2 (2011), pp. 147–159 (cit. on pp. 54, 58).

[33] Pingzhao Zhang, Yuanyuan Zhang, Kun Gao, et al. “ASPP1/2-PP1 complexes are required for chromosome segregation and kinetochore-microtubule attachments”. In: *Oncotarget* 6.39 (2015), p. 41550 (cit. on p. 54).

[34] Yuanyuan Zhang, Yuqi Wang, Youheng Wei, et al. “The tumor suppressor proteins ASPP1 and ASPP2 interact with C-Nap1 and regulate centrosome linker reassembly”. In: *Biochemical and Biophysical Research Communications* 458.3 (2015), pp. 494–500 (cit. on p. 54).

[35] Shahar Rotem, Chen Katz, Hadar Benyamini, et al. “The structure and interactions of the proline-rich domain of ASPP2”. In: *Journal of Biological Chemistry* 283.27 (2008), pp. 18990–18999 (cit. on p. 58).

[36] Mandar Deepak Muzumdar, Bosiljka Tasic, Kazunari Miyamichi, Ling Li, and Liqun Luo. “A global double-fluorescent Cre reporter mouse”. In: *Genesis* 45.9 (2007), pp. 593–605 (cit. on p. 59).

[37] Daniel Schramek, Ataman Sendoel, Jeremy P Segal, et al. “Direct in vivo RNAi screen unveils myosin IIa as a tumor suppressor of squamous cell carcinomas”. In: *Science* 343.6168 (2014), pp. 309–313 (cit. on p. 59).

[38] Mary Anne Conti, Anthony D Saleh, Lauren R Brinster, et al. “Conditional deletion of nonmuscle myosin II-A in mouse tongue epithelium results in squamous cell carcinoma”. In: *Scientific Reports* 5 (2015), p. 14068 (cit. on p. 59).

[39] Ivo J Huijbers, Jessica Del Bravo, Rahmen Bin Ali, et al. “Using the GEMM-ESC strategy to study gene function in mouse models”. In: *Nature protocols* 10.11 (2015), p. 1755 (cit. on pp. 61, 69, 73).

[40] Eva K Brinkman, Tao Chen, Mario Amendola, and Bas van Steensel. “Easy quantitative assessment of genome editing by sequence trace decomposition”. In: *Nucleic Acids Research* 42.22 (2014), e168–e168 (cit. on pp. 64, 72).

[41] Jorge Sergio Reis-Filho, Pete T Simpson, Nicholas C Turner, et al. “FGFR1 emerges as a potential therapeutic target for lobular breast carcinomas”. In: *Clinical Cancer Research* 12.22 (2006), pp. 6652–6662 (cit. on p. 65).

[42] Wa Xian, Leontios Pappas, Darshan Pandya, et al. “Fibroblast growth factor receptor 1–transformed mammary epithelial cells are dependent on RSK activity for growth and survival”. In: *Cancer Research* 69.6 (2009), pp. 2244–2251 (cit. on p. 65).

[43] Nicholas Turner and Richard Grose. “Fibroblast growth factor signalling: from development to cancer”. In: *Nature Reviews Cancer* 10.2 (2010), p. 116 (cit. on p. 65).

[44] Jeffrey S Ross, Kai Wang, Christine E Sheehan, et al. “Relapsed classic E-Cadherin (CDH1)–mutated invasive lobular breast cancer shows a high frequency of HER2 (ERBB2) gene mutations”. In: *Clinical Cancer Research* 19.10 (2013), pp. 2668–2676 (cit. on p. 66).

[45] Miguel Vicente-Manzanares, Xuefei Ma, Robert S Adelstein, and Alan Rick Horwitz. “Non-muscle myosin II takes centre stage in cell adhesion and migration”. In: *Nature Reviews Molecular Cell Biology* 10.11 (2009), p. 778 (cit. on p. 66).

[46] Alessandro Pecci, Ilaria Canobbio, Alessandra Balduini, et al. “Pathogenetic mechanisms of hematological abnormalities of patients with MYH9 mutations”. In: *Human Molecular Genetics* 14.21 (2005), pp. 3169–3178 (cit. on p. 66).

[47] Domenico Flagiello, Michèle Gerbault-Seureau, Xavier Sastre-Garau, et al. “Highly recurrent der (1; 16)(q10; p10) and other 16q arm alterations in lobular breast cancer”. In: *Genes, Chromosomes and Cancer* 23.4 (1998), pp. 300–306 (cit. on p. 66).

[48] Kathryn Van Hook, Zhiping Wang, Dexi Chen, et al. “ΔN-ASPP2, a novel isoform of the ASPP2 tumor suppressor, promotes cellular survival”. In: *Biochemical and Biophysical Research Communications* 482.4 (2017), pp. 1271–1277 (cit. on p. 67).

[49] Stefano Annunziato, Sjors M Kas, Micha Nethe, et al. “Modeling invasive lobular breast carcinoma by CRISPR/Cas9-mediated somatic genome editing of the mammary gland”. In: *Genes & Development* 30 (2016), pp. 1470–1480 (cit. on p. 69).

[50] Adam J Dupuy, Keiko Akagi, David A Largaespada, Neal G Copeland, and Nancy A Jenkins. “Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system”. In: *Nature* 436.7048 (2005), p. 221 (cit. on p. 69).

[51] Linda Henneman, Martine H van Miltenburg, Ewa M Michalak, et al. “Selective resistance to the PARP inhibitor olaparib in a mouse model for BRCA1-deficient metaplastic breast cancer”. In: *Proceedings of the National Academy of Sciences* 112.27 (2015), pp. 8409–8414 (cit. on pp. 69, 70).

[52] Silva Krause, Amy Brock, and Donald E Ingber. “Intraductal injection for localized drug delivery to the mouse mammary gland”. In: *Journal of Visualized Experiments* 80 (2013), e50692–e50692 (cit. on p. 69).

[53] Chris W Doornebal, Sjoerd Klarenbeek, Tanya M Braumuller, et al. “A preclinical mouse model of invasive lobular breast cancer metastasis”. In: *Cancer Research* 73.1 (2013), pp. 353–363 (cit. on pp. 69, 70).

[54] Robert D Cardiff, Miriam R Anver, Barry A Gusterson, et al. “The mammary pathology of genetically engineered mice: the consensus report and recommendations from the Annapolis meeting”. In: *Oncogene* 19.8 (2000), pp. 968–988 (cit. on p. 69).

[55] Andrew J Ewald, Audrey Brenot, Myhanh Duong, Bianca S Chan, and Zena Werb. “Collective epithelial migration and cell rearrangements drive mammary branching morphogenesis”. In: *Developmental cell* 14.4 (2008), pp. 570–581 (cit. on p. 70).

[56] RK Ball, RR Friis, CA Schoenenberger, W Doppler, and B Groner. “Prolactin regulation of beta-casein gene expression and of a cytosolic 120-kd protein in a cloned mouse mammary epithelial cell line.” In: *The EMBO journal* 7.7 (1988), p. 2089 (cit. on p. 70).

[57] Eugenio Montini, Daniela Cesana, Manfred Schmidt, et al. “The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy”. In: *The Journal of clinical investigation* 119.4 (2009), pp. 964–975 (cit. on p. 71).

[58] F Ann Ran, Patrick D Hsu, Jason Wright, et al. “Genome engineering using the CRISPR-Cas9 system”. In: *Nature Protocols* 8.11 (2013), pp. 2281–2308 (cit. on p. 71).

[59] Neville E Sanjana, Ophir Shalem, and Feng Zhang. “Improved vectors and genome-wide libraries for CRISPR screening”. In: *Nature Methods* 11.8 (2014), pp. 783–784 (cit. on p. 71).

[60] Antonia Follenzi, Laurie E Ailles, Silvia Bakovic, Massimo Geuna, and Luigi Naldini. “Gene transfer by lentiviral vectors is limited by nuclear translocation and rescued by HIV-1 pol sequences”. In: *Nature Genetics* 25.2 (2000), pp. 217–222 (cit. on p. 71).

[61] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet. journal* 17.1 (2011), pp. 10–12 (cit. on pp. 73, 75).

[62] Alexander Dobin, Carrie A Davis, Felix Schlesinger, et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21 (cit. on p. 73).

[63] Yang Liao, Gordon K Smyth, and Wei Shi. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7 (2013), pp. 923–930 (cit. on p. 73).

[64] Joseph Sambrook, Edward F Fritsch, Tom Maniatis, et al. *Molecular cloning*. Vol. 2. Cold Spring Harbor Laboratory Press, 1989 (cit. on p. 73).

[65] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (2014), pp. 1–21 (cit. on p. 74).

[66] W Evan Johnson, Cheng Li, and Ariel Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1 (2007), pp. 118–127 (cit. on p. 74).

[67] Renaud Gaujoux and Cathal Seoighe. “A flexible R package for nonnegative matrix factorization”. In: *BMC Bioinformatics* 11.1 (2010), p. 1 (cit. on p. 74).

[68] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4 (2012), pp. 357–359 (cit. on p. 75).

[69] Julian R de Ruiter, Sjors M Kas, Eva Schut, et al. “Identifying transposon insertions and their effects from RNA-sequencing data”. In: *Nucleic Acids Research* 45.12 (2017), pp. 7064–7077 (cit. on p. 76).

[70] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, et al. “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal”. In: *Science Signaling* 6.269 (2013), pl1 (cit. on p. 76).

[71] Margarita Maurer, André C Müller, Christine Wagner, et al. “Combining filter-aided sample preparation and pseudoshotgun technology to profile the proteome of a low number of early passage human melanoma cells”. In: *Journal of Proteome Research* 12.2 (2012), pp. 1040–1048 (cit. on p. 78).

[72] Robert S Jansen, Ruben Addie, Remco Merkx, et al. “N-lactoyl-amino acids are ubiquitous metabolites that originate from CNDP2-mediated reverse proteolysis of lactate and amino acids”. In: *Proceedings of the National Academy of Sciences* 112.21 (2015), pp. 6601–6606 (cit. on p. 78).

[73] Jürgen Cox, Marco Y Hein, Christian A Luber, et al. “Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ”. In: *Molecular & Cellular Proteomics* 13.9 (2014), pp. 2513–2526 (cit. on p. 78).

[74] Sally J Deeb, Rochelle C J D’Souza, Jürgen Cox, Marc Schmidt-Supprian, and Matthias Mann. “Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles”. In: *Molecular & Cellular Proteomics* 11.5 (2012), pp. 77–89 (cit. on p. 78).

4

Identifying transposon insertions and their effects from RNA-sequencing data

Julian R. de Ruiter[1,2], Sjors M. Kas[1], Eva Schut[1], David J. Adams[3], Marco J. Koudijs[1], Lodewyk F. A. Wessels[2,4], Jos Jonkers[1]

[1] Division of Molecular Pathology and Cancer Genomics Netherlands, Netherlands Cancer Institute, Plesmanlaan 121, Amsterdam 1066 CX, The Netherlands
[2] Division of Molecular Carcinogenesis and Cancer Genomics Netherlands, Netherlands Cancer Institute, Plesmanlaan 121, Amsterdam 1066 CX, The Netherlands
[3] Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK
[4] Faculty of EEMCS, Delft University of Technology, Mekelweg 4, Delft 2628 CD, The Netherlands

Published in Nucleic Acids Research, 1 June 2017
Volume 45, Issue 12, Pages 7064-7077

4.1 Abstract

Insertional mutagenesis using engineered transposons is a potent forward genetic screening technique used to identify cancer genes in mouse model systems. In the analysis of these screens, transposon insertion sites are typically identified by targeted DNA-sequencing and subsequently assigned to predicted target genes using heuristics. As such, these approaches provide no direct evidence that insertions actually affect their predicted targets or how transcripts of these genes are affected. To address this, we developed IM-Fusion, an approach that identifies insertion sites from gene-transposon fusions in standard single- and paired-end RNA-sequencing data. We demonstrate IM-Fusion on two separate transposon screens of 123 mammary tumors and 20 B-cell acute lymphoblastic leukemias, respectively. We show that IM-Fusion accurately identifies transposon insertions and their true target genes. Furthermore, by combining the identified insertion sites with expression quantification, we show that we can determine the effect of a transposon insertion on its target gene(s) and prioritize insertions that have a significant effect on expression. We expect that IM-Fusion will significantly enhance the accuracy of cancer gene discovery in forward genetic screens and provide initial insight into the biological effects of insertions on candidate cancer genes.

4.2 Introduction

Transposon-based insertional mutagenesis (TIM) is a high-throughput method for cancer gene discovery in mice[1]. In TIM, discrete DNA elements called transposons can migrate throughout the genome by a cut-and-paste mechanism, in which they are excised from their original location in the genome and randomly reintegrated elsewhere[2]. Depending on the location and orientation of their reintegration, these integrations can activate oncogenes or inactivate tumor suppressors, thereby inducing tumor development and progression[3]. By identifying genomic loci that are recurrently affected by transposon insertions in multiple independent tumors, this approach can be used to identify candidate cancer genes[1,3,4].

Transposon insertion sites are typically identified using targeted DNA-sequencing approaches, in which junction fragments containing transposon and flanking genomic sequences are selectively amplified and sequenced[5]. The genomic parts of these sequences are mapped to the reference genome to identify insertion sites and their genomic locations[6]. These insertions are then assigned to their putative target gene(s) using heuristics, typically picking genes in the direct vicinity of the insertion. Examples of such heuristics are nearest gene[6], fixed window[7] and rule-based mapping approaches[8].

A significant drawback of DNA-sequencing approaches is that they do not provide any direct evidence that an insertion actually affects a gene. In ambiguous cases with multiple genes in the vicinity of an insertion, heuristic approaches are frequently unable to identify the true target(s) of the insertion. This typically leads to an arbitrary selection of a single gene (nearest gene), potentially selecting the wrong gene or missing other targets (false negatives). Alternatively, heuristics may select many genes in the direct vicinity of the insertion (fixed window, rule-based mapping), resulting in the selection of many non-target genes (false positives).

Additionally, DNA-sequencing approaches provide limited insight into how the expression of a target gene is affected by a transposon insertion and which novel transcripts may result from the insertion. This has two main drawbacks. First, it prevents prioritizing insertions that have a strong effect on gene expression and are therefore likely of more importance than insertions without an effect on expression. This limits effective discrimination between driver and passenger insertions, resulting in long lists of candidate loci which are likely to include a substantial fraction of false positives that do not affect expression. Second, it limits our understanding of how gene expression or the expression of (novel) gene transcripts is affected by insertions. These insights may be key to ultimately understanding the biological effect of insertions and how they may contribute to tumorigenesis.

In previous work, Temiz *et al.* have demonstrated that insertions can be identified in paired-end RNA-sequencing data using their tool Fusion Finder[9]. In Fusion Finder, insertions are detected from discordant mate pair alignments, in which one mate aligns to a genomic sequence and the other to part of the transposon sequence. A drawback of this approach is that it does not use information from chimeric reads overlapping the fusion boundary between the gene and the transposon (split reads), limiting the accuracy and sensitivity of insertion detection. Additionally, the dependency on mate pair information prevents its use for analyzing datasets based on single-end RNA-sequencing.

In this work, we present an approach called IM-Fusion, which uses fusion-aware RNA-seq alignment to identify transposon insertions from splicing events between endogenous genes and the transposon. Key advantages of this approach are that it identifies exactly which gene(s) are affected by a transposon insertion and how the transposon is incorporated into the resulting gene transcript. Additionally, by using both split reads and discordant mate pairs to identify insertions, IM-Fusion is more sensitive than existing approaches and can be used to analyze single-end RNA-sequencing datasets. Finally, by combining insertions with exon-level expression data, we are able to accurately predict the consequences of integrations on gene transcripts.

4.3 Materials and Methods

4.3.1 IM-Fusion

Insertion detection First, we create an augmented reference genome by adding the transposon as an extra sequence in the reference genome. Then, for each sample, we align sequence reads to the augmented reference genome using a fusion-aware RNA-seq aligner such as STAR[10] or Tophat-Fusion[11]. By default, STAR is used for alignment, with the argument '--chimSegmentMin' to ensure that chimeric read alignments are produced. Chimeric alignments from STAR are filtered to select alignments that represent fusions between the transposon and genomic sequences. Alignments that overlap with the fusion junction (represented by split-read alignments) are grouped by the position of their breakpoints, as these reads precisely identify the location of a fusion. Each such group is considered to represent a single gene-transposon fusion. For paired-end sequencing data, alignments that do not overlap with the fusion boundary are grouped if their mate positions fall within a pre-defined distance, which depends on the insert size of the dataset. Where possible, these 'spanning' read groups are assigned as additional support for fusions identified from split-reads. For cases where no such fusion is found, approximate locations for the corresponding fusions are predicted based on the bounds provided by the spanning reads.

The identified fusions are annotated to identify which gene(s) and which transposon feature(s) are involved in each fusion. Fusions that do not involve splice acceptor (SA) or splice donor (SD) features of the transposon or fusions that represent biologically implausible situations (such as fusions between transposon features and gene exons in opposite orientations) are considered artifacts and removed from the list of fusions. Optionally, fusions supported by less than a pre-defined number of reads can be removed to avoid fusions with low support. For this filtering, we provide two distinct measures: a support score and an FFPM (fusion fragments per million) score. The support score simply indicates the number of reads/mates that supports the corresponding fusion. The FFPM score is a scaled version of the support score, which is normalized for differences in sequencing depth between samples. This score is analogous to the FFPM score used by STAR-Fusion[12]. The list of filtered fusions is used to predict approximate locations of the corresponding insertion sites, based on the breakpoints of the fusions.

Transcript assembly To identify cases in which insertions lead to the expression of non-canonical transcripts, IM-Fusion provides an optional step which uses StringTie[13] to perform a reference-guided assembly of novel transcripts using the

read alignment from STAR. The produced transcript annotation is used to assign any previously unannotated insertions to any novel transcripts that overlap with the insertion. If such a novel transcript overlaps with any known genes, the corresponding insertion is also assigned to these known genes, as the transcript likely represents an alternative transcript of these existing genes.

Commonly targeted gene selection Commonly targeted genes (CTGs) are selected by testing if genes are affected by insertions more frequently than would be expected by chance according to the Poisson distribution. The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time or space, as long as the expected number of events in a fixed window is known and events occur independently. Specifically,

$$P_g(K = k; \lambda_g) = \frac{\lambda_g^k e^{-\lambda_g}}{k!}$$

where k is the number of events and λ_g is the expected number of events in a fixed window. Here, each insertion represents an independent event and the fixed window is the genomic region of the gene of interest, optionally expanded to include a window around the gene. The expected number of insertions is calculated based on the size of the gene window, the size of the transcriptome (the union of windows for all genes) and the total number of insertions within the transcriptome windows.

In detail, we first count the number of insertions that were identified for a given gene g (by the insertion identification step) and were located within a pre-defined window (by default 20 kb) around the gene. This count is denoted as N_g. Second, we calculate the expected number of insertions in gene g (λ_g) based on its window size and the total number of insertions within the transcriptome as follows:

$$\lambda_g = \frac{W_g}{W_t} N_t$$

in which W_g corresponds to the size of the window around gene g, W_t the size of the transcriptome windows (the sum of windows for all genes in the genome, corrected for overlap between gene windows) and N_t represents the total number of insertions within the transcriptome windows. Using λ_g, we then calculate the probability of observing N_g or more insertions in gene g as:

$$p_g = P_g(K \geq N_g; \lambda_g)$$

After testing all genes of interest (by default all genes with at least one insertion in the gene), calculated p values are corrected for multiple testing using Bonferroni correction.

If the transposon employed in the screen is known to be biased toward integrating at specific nucleotide sequences, λ_g can be calculated differently to take this integration bias into account. In this case, instead of using the size of the gene windows, we use the number of occurrences of the nucleotide sequence with the gene window (S_g) and within the transcriptome windows (S_t) to calculate λ_g:

$$\lambda_g = \frac{S_g}{S_t} N_t$$

To account for a potential bias in integrations on the chromosome on which the transposon concatemer is located, insertions and genes on the donor chromosome can be excluded from the analysis. In this case, genes on the donor chromosome are also excluded when calculating the transcriptome size (W_t/S_t) and the number of insertions (N_t).

Differential expression analysis To test for differential expression, we first generate exon expression counts from the read alignments using featureCounts[14]. For this count summarization, we use a flattened version of the reference GTF file, which is similar to the flattened GTF files produced by DEXSeq[15]. This flattened GTF is required to ensure that overlapping exons from different transcripts of the same gene are only counted once by featureCounts.

Next, to test a given gene g for differential expression, we divide the exons of gene g into two groups: those before the transposon insertions in the gene (E_g^B) and those after the insertions (E_g^A). We assume that the expression counts of exons before the insertions (E_g^B) are not directly affected by the presence of an insertion and therefore reflect differences in the overall expression of the gene between samples. Based on this assumption, we normalize the counts of each sample for differences in overall expression of the gene by dividing the counts by a sample-specific normalization factor, which is calculated from the counts of the exons in E_g^B using DESeq2's median-of-ratios approach[16]. We then sum the normalized counts of exons in E_g^A per sample, to get a single (normalized) count of expression after the insertion site for each sample. Finally, to actually test for differential expression in the presence of an insertion in gene g, we use a two-tailed Mann–Whitney U test to compare the distribution of these counts between samples with an insertion in gene g and samples without an insertion in the gene.

In some cases, the above test is not possible because some samples do not have at least one exon before and after their insertion sites. This mostly occurs when insertions are located upstream of the first exon of the gene. To handle these cases, we first try to remove these problematic samples and repeat the test using the remaining samples. For cases where this does not leave us with any samples to test, we provide an additional gene-level test, which compares the expression of the overall gene between samples with/without insertions after normalizing for overall differences in sequencing depth.

By default, we do not use multiple testing correction for the differential expression test, as we primarily select CTGs using the Poisson-based test and use the differential expression test as an extra test to determine whether to keep the CTG. Additionally, not all CTGs may be subjected to the same test, as some genes may be tested using the gene-level test if the exon-level version is not applicable.

Single-sample differential expression To test for differential expression in a single sample (as opposed to the group-wise test described above), we provide an alternative approach that uses the same normalization procedure, but uses a negative binomial distribution to compare the expression of the sample of interest to samples without an insertion. In this approach, a negative binomial is fitted using the after insertion counts of samples without an insertion in the gene. The after count of the sample of interest is then compared to this distribution using a two-tailed test to determine if the gene is differentially expressed.

Implementation For convenience and reusability, we implemented the different steps of IM-Fusion in a Python package called *imfusion*, which is freely available on GitHub[*]. Jupyter notebooks containing the code and results of the various computational analyses are also available on GitHub[†].

The Python package provides commands for each main step of IM-Fusion, including the construction of the custom reference genome, identification of insertions from RNA-seq reads, selection of CTGs and analysis of differential expression. The current implementation supports the use of STAR or Tophat-Fusion to detect fusions, although support for additional fusion-aware aligners may be added in the future. For full functionality, working installations of STAR/Tophat2, StringTie and featureCounts are required; as STAR or Tophat2 (which implements Tophat-Fusion) are used to align reads and detect fusions, StringTie is used to detect novel transcripts and featureCounts is used to generate the expression counts. Optionally, STAR-Fusion[12]

[*]https://github.com/nki-ccb/imfusion
[†]https://github.com/jrderuiter/imfusion-analyses

can also be used to detect endogenous gene fusions as part of the STAR insertion detection pipeline.

4.3.2 Datasets

ILC dataset (RNA-seq) Single-end RNA-sequencing data from 123 tumors were obtained from a dataset of a *Sleeping Beauty* (SB) transposon screen in a mouse model of invasive lobular breast carcinoma (ILC)[17]. The RNA-seq data were downloaded from ENA in fastq format (accession number PRJEB14134) and analyzed using IM-Fusion (version 0.3.1) to detect SB insertion sites in each sample, as well as subsequently identify CTGs and their effects. For this analysis, we created an augmented reference genome using the mm10 version of the mouse genome and the *T2/Onc* transposon sequence[18]. STAR (version 2.5.2b) was used to perform the alignment, StringTie (version 1.3.0) was used for transcript assembly and featureCounts (version 1.5.0-post3) was used to generate expression counts. Reference genome features were downloaded from Ensembl 76.

ILC dataset (ShearSplink) DNA-sequencing data prepared using the ShearSplink protocol[19] for the same tumors as the ILC RNA-seq dataset were downloaded from Figshare[‡] and analyzed using the ShearSplink pipeline in PyIM[§] (version 0.2.0) to identify SB insertion sites. In essence, this pipeline first extracts genomic DNA from reads by removing the transposon and linker sequences. The genomic sequences are then aligned to the reference genome using Bowtie2 (version 2.2.8)[20], and the resulting alignments are grouped by sample and position to identify the location of insertion sites. Finally, identified insertions are assigned to their predicted target genes using the windows outlined in KC-RBM[8]. To reduce the number of identified target genes for each insertion, we selected a single target gene for each insertion by picking the closest gene identified by KC-RBM. In cases where this was not possible, e.g. due to overlapping genes, we retained multiple target genes.

B-ALL dataset Insertion data and paired-end RNA-seq data from 20 B-cell acute lymphoblastic leukemias (B-ALLs) were obtained from a previously published dataset of a SB screen performed in a mouse model of B-ALL[21]. The RNA-seq data were downloaded from ENA in fastq format (study ID: ERP005291, array expression ID: E-ERAD-264). The insertion data were obtained from the Supplementary Materials of the publication or through personal communication. Control samples were omitted from the performed analyses.

[‡]DOI: 10.6084/m9.figshare.4765111
[§]https://github.com/jrderuiter/pyim

4.3.3 Methods - ILC dataset

Gene-transposon fusion validation in RNA Tumor RNA was extracted as previously described[22] and 300 ng was converted to complementary DNA (cDNA) with a Moloney murine leukemia virus reverse transcriptase using random hexamer primers according to manufacturer's protocol (Tetro cDNA synthesis kit, Bioline). Gene-transposon fusions were detected by standard polymerase chain reaction (PCR) with an annealing temperature of 58 °C. The following primer sequences were used:

- SA reverse 5'-TTCCCGCGAATCCATCTTTC-3'
- En2SA reverse 5'-GTCGACTGCAGAATTCGATGA-3'
- SD forward 5'-GCCCATCAAGCTTGCTACTA-3'
- *Myh9* forward 5'-CTGTGTGGTCATCAACCCTTAT-3'
- *Trp53bp2* reverse 5'-ATCGCTCTGGTTTCGATAAGG-3'
- *Ctnnd1* forward 1 5'-GCTACATGCCTTGACAGATGA-3'
- *Ctnnd1* forward 2 5'-GAGAGGAGAAAGGCAGGAAAG-3'
- *Hprt* forward 5'-CTGGTGAAAAGGACCTCTCG-3'
- *Hprt* reverse 5'-TGAAGTACTCATTATAGTCAAGGGCA-3'

To study the effects of individual SB insertions on expression, we visualized single insertions together with the expression of each of their targets in the affected sample and tested for differential expression over the insertion site in the sample. The visualization was generated using the Python package geneviz, which is freely available on GitHub[¶]. Gene annotations for the plot were obtained from Ensembl 76. Expression profiles were generated from the RNA-seq alignment of the sample using pysam[23], by counting the number of reads overlapping each nucleotide position in the plotted range. Junction strengths were derived from the junction files (SJ.out.tab) generated by STAR during the alignment. To test for differential expression, we used the single-sample exon-level test implemented by IM-Fusion.

Effects on CTGs To identify biases in SA/SD insertions for the various CTGs, we counted the number of times each transposon feature (SD, SA, En2SA) was involved in the insertions affecting each CTG. The results were visualized to show the different distributions across CTGs. To test for differential expression, we applied IM-Fusions group-wise DE test for each CTG.

Insertion comparison To compare the overlap in insertions between IM-Fusion and ShearSplink, we matched two insertions between IM-Fusion and ShearSplink

[¶]https://github.com/jrderuiter/geneviz

under the following conditions: both insertions were identified in the same sample, had the same predicted target gene and their relative location and orientation was compatible. The latter restriction was used to ensure that a ShearSplink insertion was in the correct location to generate the fusion observed by IM-Fusion in the RNA-seq data. Insertions matched between the two approaches were marked as 'Shared', unmatched insertions were designated 'IM-Fusion only' or 'ShearSplink only' depending on the approach that identified them.

To identify features distinguishing shared insertions from insertions that were unique to either approach, we compared the set of shared insertions to the IM-Fusion- and ShearSplink-specific insertions. For both comparisons (Shared/ShearSplink and Shared/IM-Fusion), we first defined a set of features that could potentially affect insertion detection by either method. We then trained a logistic regression model on these features to predict whether an insertion was matched or unique to the corresponding approach. This model was used to determine the significance of each feature. Finally, we visualized the distributions of significant features for both the matched/unmatched insertions using kernel density estimation (KDE) plots for interpretation.

Candidate gene comparison To compare the candidate genes identified by ShearSplink and IM-Fusion, we first identified significant common insertion sites (CISs) and differentially expressed CTGs (DE CTGs) separately using the respective approaches. We then visualized the resulting gene rankings, linking genes that were identified as candidate genes by both approaches. Candidate genes were colored to distinguish whether they were (i) shared between both approaches (black), (ii) were identified to have insertions but were not selected as a CTG/CIS by the other approach (blue), (iii) were selected as a CTG/CIS but were not differential expressed (green), (iv) were not selected as a CTG/CIS and were not differentially expressed (purple) and (v) were omitted entirely by the other approach (red).

ShearSplink insertion validation in DNA Tumor DNA was isolated using a phenol–chloroform extraction. Transposon insertions were detected in 500 ng DNA by standard PCR with an annealing temperature of 58 °C. The following primer sequences were used:

- *En2SA* forward 5'-GCTTGTGGAAGGCTACTCGAA-3'
- *Nf1* 11KOU029-R5.INS_12 reverse
  5'-CTCACGTGAAGTGGGAAAGACA-3'
- *Nf1* 12SKA029-R3.INS_15 reverse
  5'-GGCGCACACCTTTAATCCTAAC-3'

- *Nf1* 12SKA033-R3.INS_10 reverse 5'-TAGCTCCCTGTGTGTTCCTTTG-3'
- *Nf1* 12SKA068-L3.INS_15 reverse 5'-AAGGGTGAAGCAGGAGGATTAC-3'
- *Nf1* 12SKA092-L2.INS_10 reverse 5'-ACGGAGAAGGAGAGAGGGAAA-3'
- *Nf1* 12SKA104-R3.INS_1 reverse 5'-CCAACATCCCTGTTGTGTGTATG-3'
- *Hprt* forward 5'-CTGGTGAAAAGGACCTCTCG-3'
- *Hprt* reverse 5'-TGAAGTACTCATTATAGTCAAGGGCA-3'

Endogenous fusion identification Endogenous gene fusions were identified by applying STAR-Fusion[12] (version 0.5.4) to the raw RNA-seq data (fastq files) using recommended settings. The resulting list of fusions were combined across samples and filtered for fusions with breakpoints at known splice junctions, as these are most likely to reflect proper gene fusions. The filtered fusions were prioritized by grouping fusions on the involved genes and ranking by the recurrence of these gene pairs across samples. The fusions involving *Fgfr2* were validated using the same approach as for the gene-transposon fusions, with the following additional primers:

- *Fgfr2* forward 5'-TGGCCAGGGATATCAACAAC-3'
- *Kif16b* reverse 5'-CTTTCCTGAGGGCTAGAGTTTG-3',
- *Myh9* reverse 5'-GATAGCGCCTTTGTCTCCTT-3',
- *Tbc1d1* reverse 5'-CCAGGCTGTGAGAAGGATTT-3'

4.3.4 Methods - B-ALL dataset

Candidate gene comparison To compare IM-Fusion with the DNA-seq results from the original publication, we applied IM-Fusion to the paired-end RNA-seq data and compared the identified DE CTGs with the published candidate genes (DE CISs). To avoid selecting CTGs with very low support in this relatively deeply sequenced dataset (as these are more likely to represent false positives), we filtered insertions with fewer than 10 supporting reads or mates from the CTG analysis.

Effect of sequencing depth The B-ALL samples were downsampled to depths of 15, 30, 50 and 70 million reads using Seqtk[||]. IM-Fusion was applied to each of these downsampled datasets to identify DE CTGs, using the same settings as were used for

[||] https://github.com/lh3/seqtk

the full dataset. The number of insertions and DE CTGs were compared between the different depths, as well as the overlap in DE CTGs between depths.

Single- versus paired-end comparison A single-end version of the dataset was simulated by supplying only the first pair as input to IM-Fusion. The results from the paired-end and single-end analyses were compared by juxtaposing DE CTGs and insertions in these genes between the two analyses.

Fusion Finder comparison We created an augmented version of the mm10 reference genome containing the *T2/Onc* transposon sequence in the same manner as described by Temiz *et al.*[9]. This reference was modified to mask the En2 and Foxf2 gene loci, which contain sequences homologous to parts of the transposon sequence. Tophat2[24] (version 2.1.0) was used to align reads to this augmented reference, after which the Fusion Finder script (version 3.1) was used to identify insertions in each sample. The results were compared with IM-Fusions DE CTGs and published candidate genes by analyzing the overlap between the identified insertions and the CTGs/CISs. To determine why certain CTGs/candidates were not identified by Fusion Finder, we visualized the distribution of the used transposon features and compared the alignments of reads supporting insertions unique to IM-Fusion between the Tophat2 and STAR alignments using pysam[23].

Endogenous fusion identification Endogenous gene fusions were identified in the same manner as for the ILC dataset.

4.4 Results

4.4.1 Identifying insertion sites from gene-transposon fusions

Transposon insertions can affect the expression of nearby genes, potentially leading to the activation of oncogenes or the inactivation of tumor suppressors. For example, consider the *T2/Onc* transposon (Figure 4.1A) that is used in this work. When integrated in the vicinity of a gene, this transposon can induce (over)expression of nearby genes by initiating transcription from its promoter sequence (MSCV) and then splicing into the gene using the SD sequence (Figure 4.1B). Alternatively, the transposon can truncate transcripts using either of its SA sites (SA/En2SA) and their corresponding polyA (pA) sites (Figure 4.1C). Depending on the gene and the location of the transposon, these truncations can inactivate the gene by resulting in

an unstable transcript or inactive protein, or activate the gene by removing inhibitory protein domains.

In both of these cases, part of the transposon sequence is incorporated into the resulting mRNA transcript(s) via splicing between the affected gene and the transposon. As such, these transcripts effectively represent fusions between the transposon sequence and the affected gene. We therefore hypothesized that it should be possible to detect transposon insertion sites from RNA-sequencing by identifying gene-transposon fusions using existing gene fusion detection tools. By further analyzing the breakpoints of each fusion, we could determine exactly which gene and which feature of the transposon are involved in the fusion, and use this information to predict the location of the corresponding insertion site.

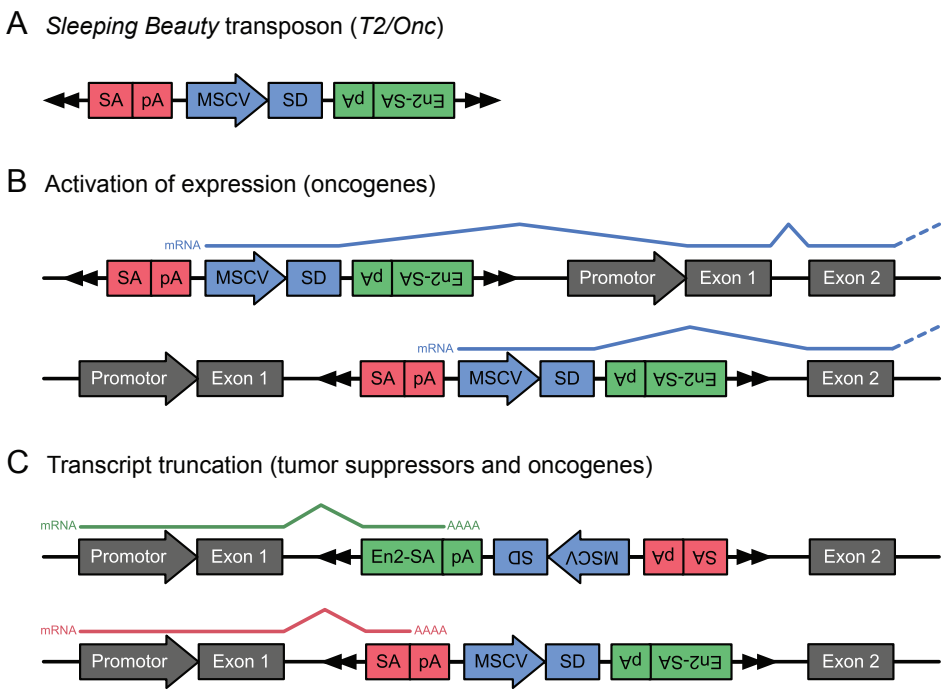


Fig. 4.1. Overview of the *T2/Onc* transposon and its effects on gene expression. **(A)** The transposon sequence contains two splice acceptor sequences (SA and En2SA) with corresponding polyA sequences (pA), and a single promoter sequence (MSCV) combined with a splice donor (SD) sequence. **(B)** Sense insertions of the transposon either within or upstream of a gene may drive overexpression of the downstream gene sequence by initiating expression from the transposons promoter and SD sequence. **(C)** Insertions within genes (in either orientation) may truncate gene transcripts by splicing to either of the SA sites (SA or En2SA). The resulting truncations may inactive tumor suppressor genes, but can also activate oncogenes by removing inhibitory domains from the resulting protein.

4.4.2 IM-Fusion

In this work, we developed a tool called IM-Fusion, that uses a three-step approach to (i) identify insertions from gene-transposon fusions in RNA-sequencing data, (ii) select genes that are more frequently affected by insertions than would be expected by chance and (iii) test if the expression of these genes is significantly changed by their insertions (Figure 4.2A). A brief description of each of the steps is provided below, more details are available in the Materials and Methods section.

Identifying insertion sites

IM-Fusion identifies transposon insertion sites from gene-transposon fusions in the RNA-seq data. To identify these fusions, IM-Fusion first creates an augmented version of the host reference genome by adding the sequence of the transposon as an extra sequence to the original reference sequence. Then, for each sample, IM-Fusion uses a fusion-aware RNA-seq aligner to align RNA-seq reads to the augmented reference and identify gene fusions. By default, STAR[10] is used for this purpose, although Tophat-Fusion[11] is also supported. The identified fusions are filtered to only select fusions between genes and the transposon sequence. These gene-transposon fusions are then analyzed to identify the involved genes and transposon features, and to infer the approximate locations of the insertions (Figure 4.2B). Optionally, the RNA-seq alignment can be used to perform a reference-guided transcript assembly, which allows IM-Fusion to detect insertions that result in the expression of novel (unannotated) transcripts.

An important advantage of IM-Fusion over DNA-sequencing based approaches is that, instead of focusing on deriving the exact location of insertion sites, it focuses on determining which genes are affected by insertions. This gene-centric approach allows us to select only those insertions that affect expressed genes and are therefore most likely to have an actual biological effect. By doing so, IM-Fusion provides an important filter that strongly enriches for biologically relevant insertions and avoids selecting many extraneous insertions that are unlikely to affect gene expression. This greatly increases the specificity of our results, providing more confidence in the identified hits.

Selecting commonly targeted genes

To identify genes that are commonly targeted by insertions, we use the Poisson distribution to test whether a given gene has more insertions than may be expected by chance (see Materials and Methods). To correct for cases in which a single insertion is detected multiple times in the same gene, either due to its involvement

A Overview of IM-Fusion

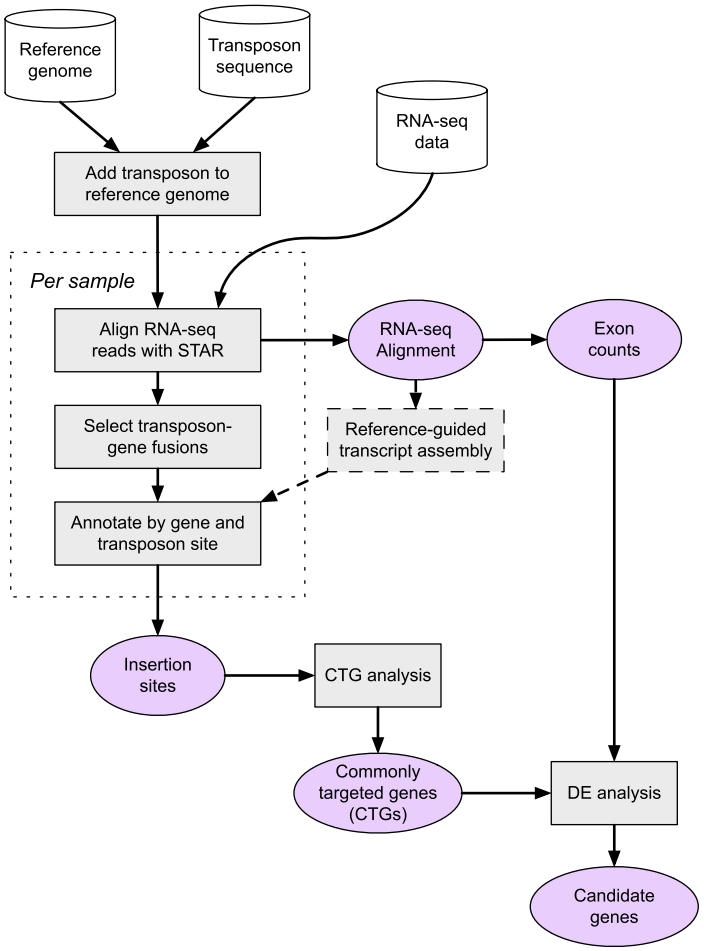


Fig. 4.2. Overview of IM-Fusion. (A) The IM-Fusion pipeline. Samples are initially processed individually to identify insertions and generate gene and exon expression counts for each sample separately. The per-sample results are then combined to identify genes that are recurrently affected across samples. For these genes, we then combine the expression and insertion data to test for differential expression over the insertion site. The results of this analysis are used to determine if insertions have a significant effect on the expression of their target genes and exactly how the insertions affect the resulting gene transcript.

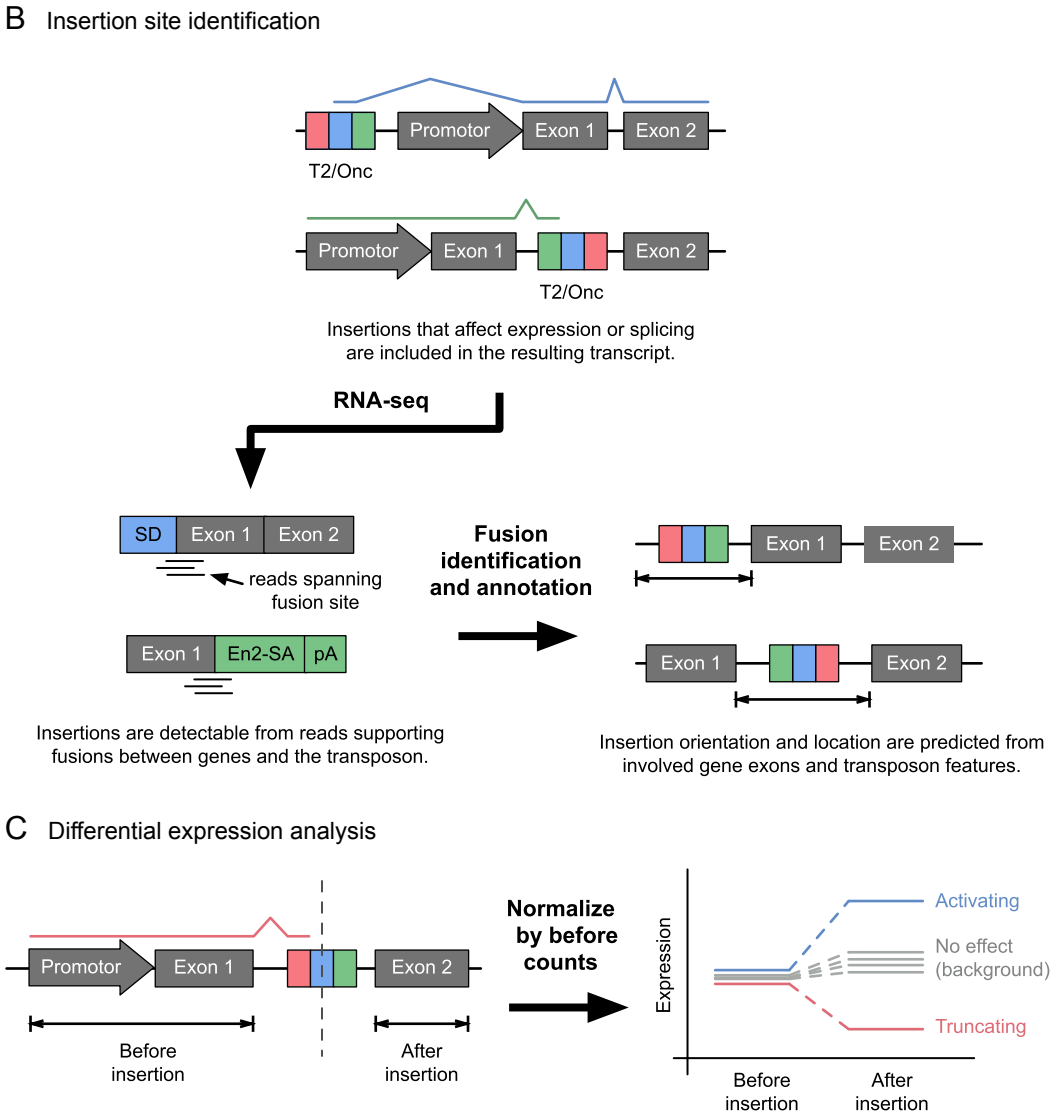


Fig. 4.2. **Continued. (B)** Transposons that affect gene expression are included in gene transcripts and are therefore detectable as fusion transcripts between genes and the transposon. These fusions are detected by reads or mate-pairs that bridge the fusion site. The breakpoints of the identified gene-transposon fusions are analyzed to identify the involved gene(s) and predict an approximate location for the corresponding insertion(s). **(C)** Insertion and expression data are combined to test if an insertion significantly affects the expression of exons downstream of the insertion site. Expression counts are calculated both before/after the insertion site for a sample with an insertion and a set of background samples without an insertion. The 'before' count is then used to normalize the sample counts, after which the normalized 'after' counts are compared to the 'before' counts to test for differential expression. Samples with a truncating insertion are expected to show a lower level of expression after the insertion relative to the background, whilst samples with an activating insertion are expected to show increased expression after the insertion.

in multiple gene isoforms or due to local hopping within the gene, insertions are by default collapsed into a single insertion per gene per sample (taking the average location of the insertions) before testing for enrichment. This ensures that selected CTGs indeed represent recurrent insertions across multiple samples, and not just multiple insertions within a single or few samples.

Testing for significant effects on expression

To establish whether the expression of a CTG is significantly altered by its insertions, we test for differential expression over the insertion sites in the gene. The main goal of this analysis is to determine if we see a significant increase in expression after the insertion site, indicating that (partial) gene transcripts are (over)expressed by the insertions, or observe a significant decrease in expression, indicating that gene transcript(s) are truncated by the insertions.

To perform the test, we first normalize for differences in overall expression of the gene across all samples based on the expression of exons before the insertion site, which we assume are not directly affected by the presence of an insertion. After this normalization, we compare the normalized expression levels after the insertion site between samples with and without an insertion in the gene to test for differential expression (Figure 4.2C). By default the test performs a group-wise comparison using the Mann–Whitney U test, in which the expression of samples with an insertion is compared to samples without insertions in the gene. Alternatively, we also provide a single-sample test based on the negative binomial distribution, which determines whether the gene is differentially expressed in a specific sample.

For cases without exons before the insertion site(s), which can occur if insertions are located upstream of the gene, an additional gene-level test is provided. This test compares the expression of the overall gene between samples with/without insertions, after normalizing for overall differences in sequencing depth.

4.4.3 Applying IM-Fusion to a mouse model of breast cancer

We tested our approach by using IM-Fusion to identify SB transposon insertions in 123 tumors from a mouse model of invasive lobular breast cancer (ILC)[17]. On average, 0.1% of the reads in each sample were chimeric reads supporting a potential fusion, of which 0.42% represented a putative gene-transposon fusion (Supplementary Table S4.1). From these fusions, IM-Fusion identified a total of 2057 transposon insertion sites across all tumors, with a median of 12 insertions per tumor (Supplementary Table S4.2). A total of 1043 genes were affected by at least one insertion,

14 of which were selected as differentially expressed (DE) CTGs (Supplementary Table S4.3). To confirm the existence of the identified insertions, a subset of insertions was validated using PCRs targeting the predicted gene-transposon fusion transcripts (Supplementary Figure S4.1).

Effects of individual insertions

To evaluate the effect of individual insertions, we visualized single insertions together with the expression of their target genes in the corresponding sample (Figure 4.3A-D). A first example is shown in Figure 4.3A, which shows an antisense insertion in the *Trps1* gene. This insertion was identified from a fusion between the transposons En2SA site and the fourth exon of the gene, indicating that the insertion truncated the gene after this exon. This hypothesis was supported by the expression profile of the gene in this sample, which showed a marked reduction in expression after the insertion site. Using the single sample DE test, we confirmed that this reduction in expression was indeed significant compared to background samples without an insertion in the gene (Figure 4.3C).

A second example (Figure 4.3B) shows a sense insertion in the *Trp53bp2* gene, which was identified from two distinct gene-transposon fusions. The first fusion, between the SA site of the transposon and exon 12 of the gene, indicated that the insertion truncated gene transcription after this exon. However, the second fusion, between the SD site and exon 13, indicated that the insertion also drove overexpression of a partial gene transcript downstream of the insertion. Taken together, this showed that the insertion simultaneously resulted in both the truncation of the original gene transcript and overexpression of a C-terminal transcript containing exons 11–18. This overexpression was clearly reflected in the expression levels of the gene, which were significantly increased after the insertion site (Figure 4.3D). Finally, from the shown splice junctions we saw that the full-length transcript of *Trp53bp2* (and/or the truncated N-terminal transcript) was still expressed in this sample, though at lower levels than the partial transcript.

General effects of insertions on CTGs

To determine how each identified CTG was affected by its insertions, we first analyzed the insertions in each CTG to identify if the gene was biased to SD or SA insertions. In this analysis, a bias to SD insertions would indicate the gene is mainly overexpressed by insertions in the gene (Figure 4.1B). Conversely, a bias toward the SA/En2SA sites would indicate the gene is mainly truncated by its insertions (Figure 4.1C). Second, we used IM-Fusion to test for differential expression across the insertion site to determine if the insertions affect the expression of the gene and whether the observed effect points to truncation or overexpression of the gene. For clarity we

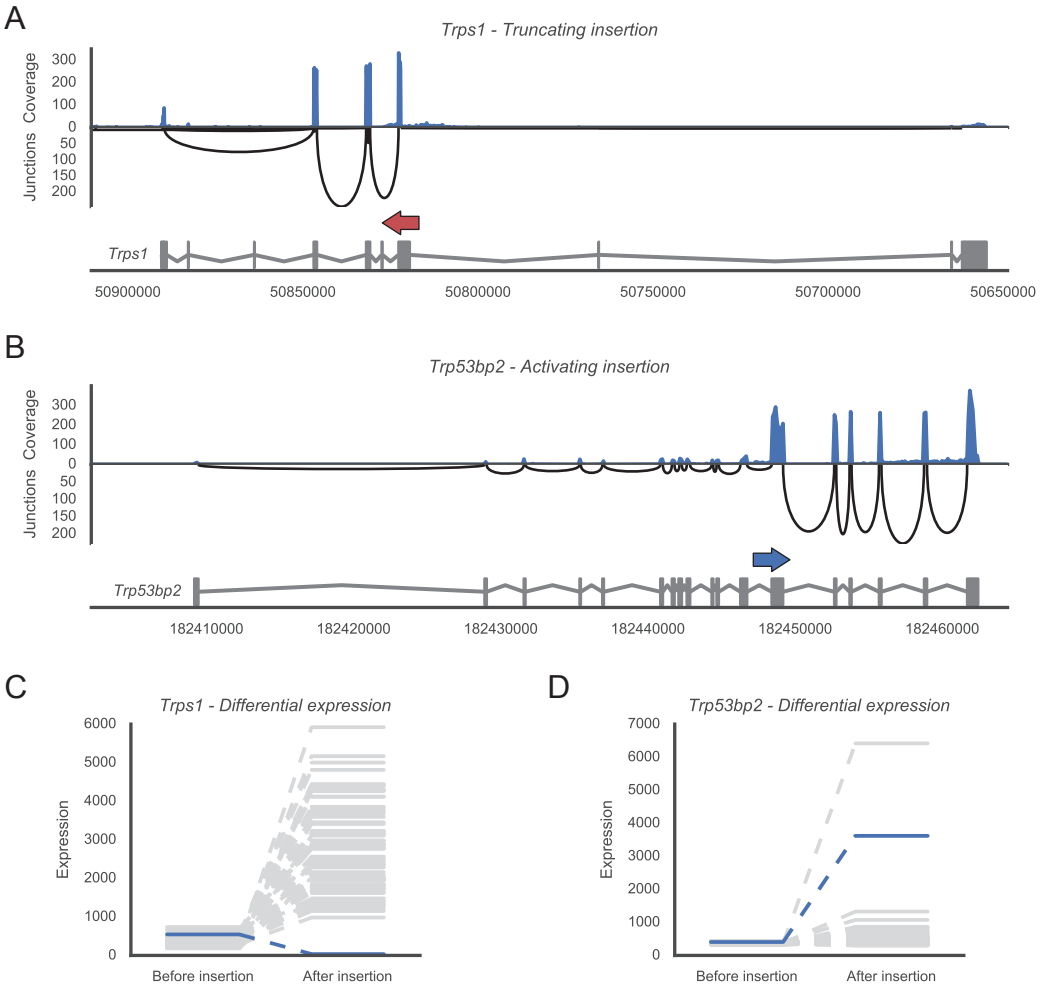


Fig. 4.3. Examples of identified insertions, CTGs and their effects on gene expression. **(A)** An example of an antisense insertion in *Trps1* that results in truncation of the gene transcript. The insertion (red arrow) is shown above the main transcript of the gene, together with expression levels of the gene. The expression of the exons is shown along the top in blue, which reflects the number of reads covering the various exons. Similarly, the black arches below indicate the strength of the splicing junctions between the different exons, with the height of the arch indicating the number of reads supporting the splice junction. Taken together, these expression profiles show a strong decrease in expression after the insertion site, supporting the hypothesized truncation. **(B)** An example of a sense insertion in *Trp53bp2* (blue arrow). This insertion results in both truncation of the gene and overexpression of a partial transcript. **(C)** Quantified expression levels before/after the insertion site for the *Trps1* insertion shown in (A). Compared to the samples without an insertion (gray), the sample with this insertion (blue) shows a significant decrease in expression after the insertion. **(D)** Quantified expression levels for the *Trp53bp2* insertion. Overexpression of the truncated transcript is clearly reflected by the increase in expression after the insertion site.

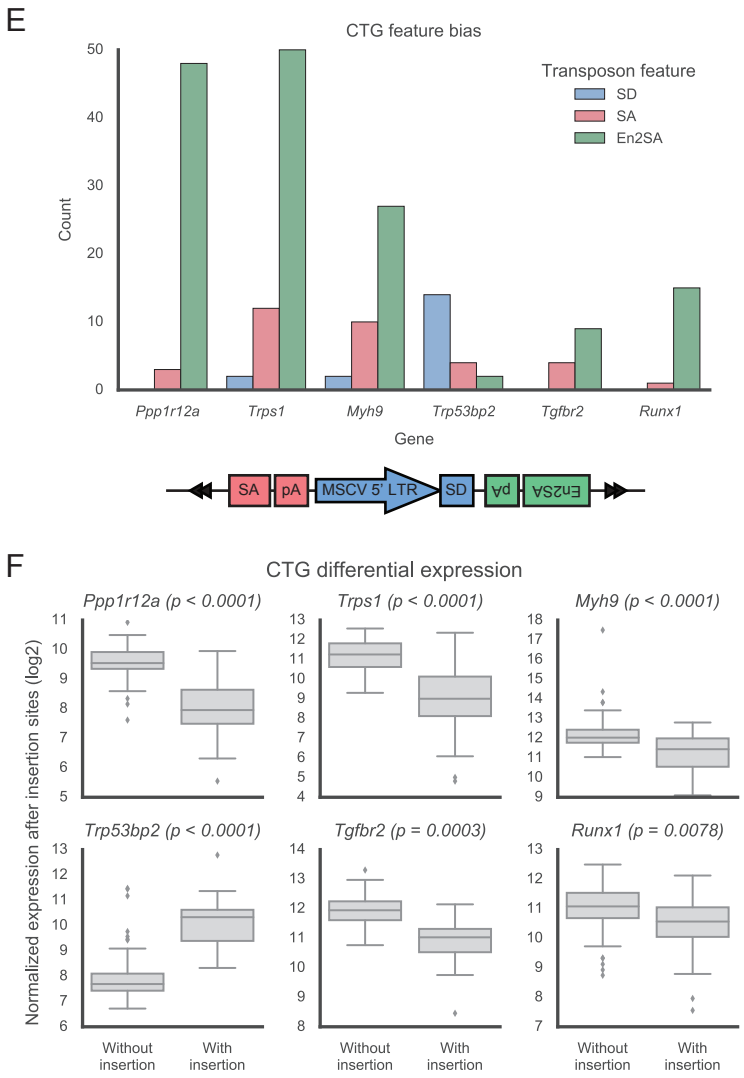


Fig. 4.3. **Continued. (E)** Frequencies of the transposon features involved in the insertions of the top six identified CTGs. A bias toward SA/En2SA favors truncation of the gene, whereas a bias toward SD favors overexpression. **(F)** Differential expression across the insertion sites for each of the CTGs. An increase in the presence of an insertion indicates overexpression, a decrease indicates truncation.

limited our analysis here to the top six candidate genes; similar results for the other candidates are available in Supplementary Figure S4.2.

This analysis showed that most top CTGs (*Ppp1r12a*, *Trps1*, *Myh9*, *Tgfbr2* and *Runx1*) were clearly biased toward SA/En2SA insertions (Figure 4.3E), indicating that transcripts of these genes were being truncated by the transposon insertions. This hypothesis was further supported by the DE tests (Figure 4.3F), which confirmed that each of these genes showed a significant decrease in expression after the insertion site, indicating that genes are indeed truncated. Conversely, for one top CTG, *Trp53bp2*, we saw a clear bias toward SD insertions, indicating that this gene is overexpressed by its insertions. This was again supported by the DE analysis,

which determined that *Trp53bp2* showed a significant increase in expression after its insertion sites.

Comparison with targeted DNA-sequencing

To assess if IM-Fusion identifies similar insertions to targeted DNA-sequencing approaches, we compared our results to those obtained by targeted DNA-sequencing of insertions using the ShearSplink protocol[19]. For this comparison, we matched insertions between the two approaches (IM-Fusion and ShearSplink) if they identified the same target gene and had compatible genomic locations and orientations. Note that, using this approach, an insertion is counted multiple times if it is assigned to multiple genes, thereby increasing the apparent total number of insertions.

Matched insertions were considered to be shared by both approaches, whereas unmatched insertions were categorized as 'ShearSplink-specific' or 'IM-Fusion-specific' depending on their source. This analysis showed that the majority of the insertions identified by IM-Fusion (578/818) were shared with ShearSplink (Figure 4.4A). However, a substantial number of insertions were unique to either IM-Fusion (240) or ShearSplink (2838), indicating a considerable disparity between the two approaches.

ShearSplink-specific insertions

To investigate why certain insertions were not identified by IM-Fusion, we compared the ShearSplink-specific insertions to the insertions identified by both approaches. The goal of this comparison was to identify features that distinguished the two sets of insertions (see Materials and Methods) and might therefore provide insight into the underlying reasons for the observed differences. Of the considered features, the following were determined to be significantly predictive: the expression level of the predicted target gene, the relative location of the insertion within its target, the distance of an insertion to its target and the support of the ShearSplink insertion.

The first two of these features point toward biases in the sequencing coverage of the RNA-seq data that affect the detection of insertions. The first feature, the expression level of the target gene, indicates that IM-Fusion had trouble identifying insertions in genes with no or low expression (Figure 4.4B). The lack of insertions in non-expressed genes was expected, as these insertions are not represented in the RNA-seq data. As these insertions are unlikely to have any biological effect, their omission is expected to increase the specificity of IM-Fusion with regard to biologically relevant insertions. The lack of insertions in genes with low expression reflects an inherent bias of RNA-seq toward highly expressed genes, which results in less sequencing coverage for genes with low expression.

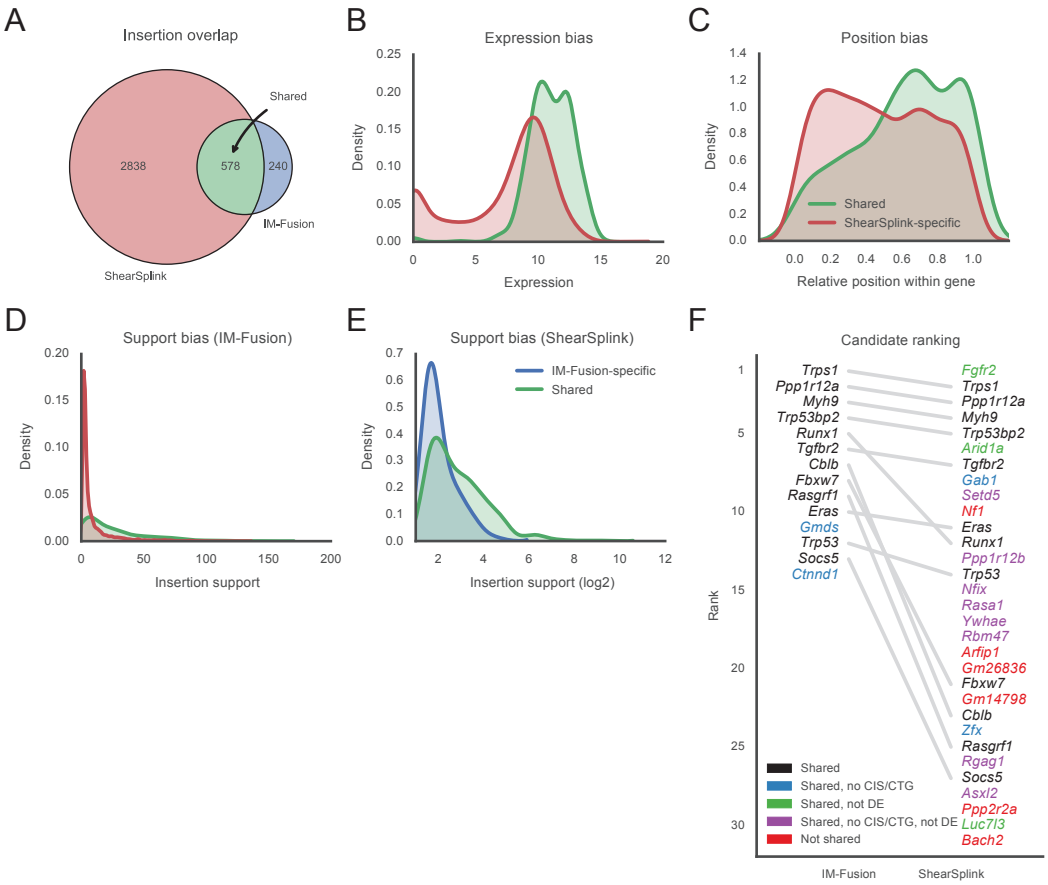


Fig. 4.4. Comparison of insertions identified by IM-Fusion and ShearSplink. **(A)** Venn diagram of the insertions identified by ShearSplink (red) and IM-Fusion (blue). Many IM-Fusion insertions are shared with ShearSplink (green), but a considerable number of insertions are unique to either approach. **(B and C)** Distribution of features reflecting biases of RNA-sequencing that affect the detection of insertions by IM-Fusion. ShearSplink-specific insertions (red) typically have low expression compared to shared insertions (green) and are therefore more difficult to detect by RNA-seq. Similarly, insertions toward the start of the gene are more frequently missed by IM-Fusion due to the 3' bias of the polyA tail selection used in the RNA-sequencing. **(D and E)** Distributions of support of DNA-seq insertions and support of RNA-seq insertions. Insertions with low DNA-seq support are more often missed by IM-Fusion, whilst insertions with low IM-Fusion support are often not detected by ShearSplink. These differences likely reflect heterogeneity of subclonal insertions present in the tumor tissue samples used for DNA-seq and RNA-seq, respectively. **(F)** Comparison of the frequency-based ranking of candidate genes identified by IM-Fusion and ShearSplink. Gray lines indicate the relative rankings of genes that were identified by both approaches. Genes missed by the other approach are marked red. Genes that were identified to have insertions but not selected as CISs/CTGs by the other approach are colored blue or purple, depending on their differential expression status. Genes that were identified as CISs/CTGs but were not differentially expressed are marked green.

Similarly, the second feature, the relative position of an insertion within the gene, showed that IM-Fusion misses more insertions at the 5' end of genes (Figure 4.4C). This is due to a well documented 3' bias of the polyA-tail selection used to enrich for mRNAs in RNA-sequencing, which results in decreasing coverage toward the

5' end of gene transcripts. Together, these two biases limit the ability of RNA-seq-based approaches such as IM-Fusion to detect insertions in lowly expressed genes, particularly at their 5' end. This effect can be mitigated by deeper sequencing and by using a different approach to enrich for mRNAs in the preparation for RNA-sequencing (such as rRNA depletion).

Another significant feature, the support of an insertion (Figure 4.4D), showed that IM-Fusion mainly missed ShearSplink insertions with a low support score. This bias may be due to one or more of the following reasons. First, our RNA-seq data may not be deep enough to detect very subclonal insertions that are only present in a very small fraction of the tumor cells. Second, the observed differences may reflect intratumoral heterogeneity, as we did not use the same tumor fragments for RNA and DNA extraction and sequencing, but instead used two separate pieces of the same tumor. For clonal insertions this is not an issue, but subclonal insertions might be present in only one of the tumor pieces, therefore leading to some of the observed differences between IM-Fusion and ShearSplink.

Finally, we found that the heuristic assignment of target genes by ShearSplink also introduced biases. Even after restricting the assignment of target genes to the closest gene, ShearSplink was unable to identify a unique target gene for some insertions. For example, insertions within the *Arfip1/Fbxw7* locus were frequently assigned by ShearSplink to both *Fbxw7* and *Arfip1*. Closer inspection of these insertions indicated that these insertions are in fact closely clustered in *Fbxw7* and are therefore unlikely to affect the *Arfip1* transcript that overlaps with *Fbxw7*. This hypothesis was supported by the IM-Fusion results, which only identified insertions in *Fbxw7*, indicating that *Arfip1* is a false positive of the heuristic assignment by ShearSplink. Similarly, the distance to target gene feature indicated that insertions further away from their target genes are rarely matched by IM-Fusion. These hits are also likely to be artifacts of the heuristic assignment of target genes by Shearsplink.

IM-Fusion-specific insertions

To determine why some insertions were only identified by IM-Fusion, we also compared the set of insertions unique to IM-Fusion to the shared insertions. This comparison identified the support score of an insertion as the most predictive feature of IM-Fusion-specific insertions (Figure 4.4E). This feature, which reflects the number of reads supporting the corresponding gene-transposon fusion, showed that ShearSplink mainly misses insertions with a low IM-Fusion support score. As these insertions are only supported by a few reads in the RNA-seq data, they are likely either false positives of IM-Fusion or subclonal insertions that are present in a small fraction of tumor cells or in specific parts of the tumor. In the latter case, the

missed insertions are again likely attributable to heterogeneity between the DNA- and RNA-seq samples, as previously explained for the ShearSplink support feature.

Comparison of identified candidate genes

To assess if IM-Fusion identified different candidate genes than ShearSplink, we compared the DE CTGs from IM-Fusion to the genes associated with CISs from the ShearSplink analysis. This comparison showed that IM-Fusion and ShearSplink identified 14 and 32 candidate genes respectively, of which 12 were shared between both approaches. From a comparison of the rankings of the candidate genes (Figure 4.4F), we saw the strongest concordance between the most frequently recurring candidate genes, with more discrepancy among the less frequent candidates.

To determine why some ShearSplink candidate genes were not identified by IM-Fusion, we examined them in more detail. Five genes (*Arfip1*, *Gm26836*, *Gm14798*, *Ppp2r2a* and *Bach2*) were not identified at all by IM-Fusion, suggesting that these are either false positives of the ShearSplink analysis, as we have already argued for *Arfip1*, or are weak/subclonal insertions that were not picked up by IM-Fusion. For *Nf1*, IM-Fusion did detect several weak insertions, which were only supported by single reads and were therefore filtered from the CTG analysis. These insertions, together with additional validation of several ShearSplink insertions (Figure S4.3 Supplementary Figure S4.3), demonstrated that *Nf1* was not a false positive of the ShearSplink analysis. However, closer inspection showed that Nf1 insertions were generally supported by few reads in the ShearSplink data, thereby explaining their omission by IM-Fusion.

Several other genes (*Setd5*, *Gab1*, *Ppp1r12b*, e.g.) were identified to have insertions by IM-Fusion, but were not detected in enough samples to be selected as a CTG. Further analysis showed that insertions in missing samples were supported by few ShearSplink reads, indicating that these insertions are missed due to their low clonality. This also explains why several of these genes (*Ppp1r12b*, *Nfix*, *Rmb47*, etc.) are not differentially expressed in the presence of an insertion, as we are less likely to pick up expression differences if the signal is weak due to subclonality.

A few candidate genes, including *Fgfr2* – the top hit from the ShearSplink analysis, were not selected as DE CTGs due to a lack of differential expression. Closer analysis showed that *Fgfr2* is affected by a mix of sense and antisense insertions. Whilst the antisense insertions merely truncate the gene, the sense insertions both truncate the gene and induce the overexpression of a partial C-terminal transcript (Supplementary Figure S4.4). Together, this results in a mix of samples with increased and decreased expression, thereby representing a more complex pattern of expression changes than the overall changes that the DE test was designed to detect. This indicates that,

although the DE test is useful for prioritizing candidate genes, frequently recurring CTGs that are not differentially expressed should be investigated in more detail to avoid filtering out more complex cases of differential expression. This can, for example, be done by grouping samples based on the orientation of their insertions (as done here) or on the involved SD/SA sites if these are expected to have different effects on expression.

Finally, besides the known candidates, IM-Fusion identified two novel candidates that were not identified by ShearSplink. Interestingly, both of these genes were identified in similar numbers of samples (two to three samples) by both ShearSplink and IM-Fusion, indicating that IM-Fusion may have more power to identify rare CTGs.

4.4.4 Application of IM-Fusion to paired-end RNA-sequencing data from B-ALL tumors

To test IM-Fusion on paired-end RNA-sequencing data, we used an additional dataset of SB-induced B-cell acute lymphoblastic leukemias (B-ALL) for which both targeted DNA-sequencing and relatively deeply sequenced paired-end RNA-sequencing (70–90 million reads) was available[21]. In the original analysis of this dataset, Van der Weyden *et al.* first identified CISs from targeted DNA-sequencing data, and then selected predicted target genes that showed significant differential expression in the presence of an insertion (DE CISs). For our comparison, we applied IM-Fusion using only the RNA-sequencing data and compared the identified insertions and CTGs to the results of the DNA-seq analysis. In light of the higher sequencing depth of the B-ALL dataset (relative to the ILC dataset), we removed insertions with fewer than 10 supporting reads in the CTG analysis to avoid selecting genes that are recurrently detected but have low support, as these are likely to represent false positives (Supplementary Figure S4.5).

CTG comparison

On average, 0.72% of the mate pairs in each sample reflected chimeric alignments, of which 0.45% supported potential gene-transposon fusions (Table S4.4). From these fusions (Supplementary Table S4.5, IM-Fusion identified six CTGs (*Jak1*, *Stat5b*, *Cblb*, *Zfp423*, *Dlx3* and *Bmi1*), of which all except *Bmi1* and *Dlx3* coincided with the six DE CISs identified by the DNA-seq analysis (Figure 4.5A-B). Two genes were only identified by the DNA-seq analysis (*Foxp1* and *Il2rb*). Closer inspection of the original DNA-seq data showed that insertions in these genes were generally supported by $<$10 reads (Supplementary Table S4.6), suggesting that these insertions are subclonal and are therefore not represented in the RNA-seq sample due to the afore-mentioned

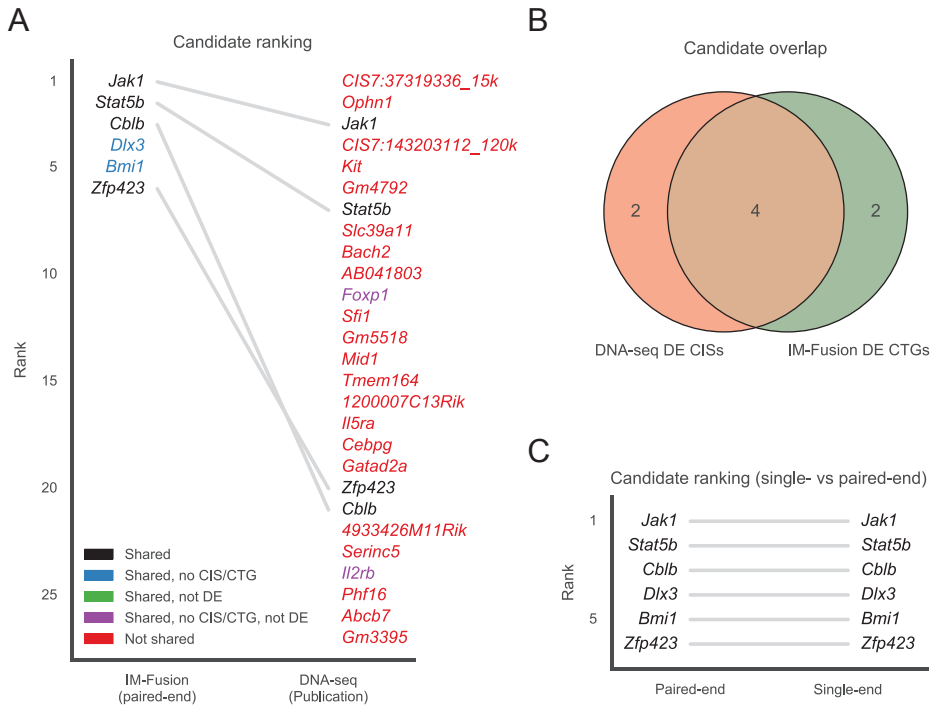


Fig. 4.5. Comparison of candidate genes identified by IM-Fusion and the original DNA-sequencing analysis in the B-ALL dataset. (A) Comparison of candidate gene rankings between IM-Fusion (left) and the original DNA-seq-based analysis (right). Colors are coded as in Figure 4.4F. **(B)** Overlap of IM-Fusions DE CTGs and the published DE CISs. **(C)** Comparison of DE CTG rankings by IM-Fusion on the single- and paired-end versions of the B-ALL dataset.

issues with sample heterogeneity. Interestingly, both of the novel CTGs (*Bmi1*, *Dlx3*) have been reported to play a role in the development of B-ALL[25,26], suggesting that these are true hits that were missed by the DNA-based analysis.

Effect of sequencing depth

To determine how sequencing depth affects the detection of insertions and CTGs, we made use of the high sequencing depth of the B-ALL dataset to repeat the analysis at reduced depth by downsampling the original dataset to 15, 30, 50 and 70 million reads per sample. These analyses showed that the number of detected insertions increases linearly with the sequencing depth (Supplementary Figure S4.6A), indicating that additional sequencing depth provides more power to detect insertions. In contrast, only one extra DE CTG (*Dlx3*) was detected at higher sequencing depths (Supplementary Figure S4.6B-C), suggesting that deep sequencing may provide limited returns when screening for candidate genes. However, less insertions were detected in some of these CTGs at the lower depths (Supplementary Table S4.7), demonstrating that a higher sequencing depth will provide more accuracy in the detection of weak insertions.

Single- versus paired-end sequencing

To study the added value of paired-end sequencing, we simulated a single-end version of the dataset by applying IM-Fusion to only one of the paired-ends. Although the analysis of the single-end dataset identified the same DE CTGs as the paired-end analysis (Figure 4.5C), the paired-end data yielded on average two times higher support scores for insertions due to the higher effective depth of paired-end sequencing, and identified a number of insertions that were not detected in the single-end sequencing data (Supplementary Table S4.8). Overall, this suggests that paired-end sequencing data is not strictly necessary for detecting insertions, but is beneficial for the detection of weak insertions.

Comparison with Fusion Finder

Finally, to compare IM-Fusion with existing approaches, we analyzed the B-ALL dataset using Fusion Finder[9], which uses Tophat2[24] to identify transposon insertions from discordant mate pairs in paired-end RNA-sequencing data. Comparison of the identified insertions showed that Fusion Finder identified recurrent insertions in Cblb and Dlx3, but was only able to identify insertions in a single sample for Jak1 and Bmi1, and was unable to detect insertions in any of the other DE CTGs identified by IM-Fusion (Supplementary Table S4.9).

More detailed analyses of the results showed that the insertions in CTGs missed by Fusion Finder are (i) biased toward SD insertions and (ii) mainly supported by chimeric reads overlapping the fusion boundary, rather than mate pairs that span the fusion (with one mate on either side of the fusion). The latter explains why the majority of these insertions were not detected by Fusion Finder, as Fusion Finder does not incorporate split read information into its insertion detection. This highlights an important advantage of using fusion-aware aligners such as STAR and Tophat-Fusion, as these aligners explicitly account for chimeric fusion reads in their alignment, resulting in increased sensitivity for the detection of these insertions.

Although Fusion Finder failed to detect insertions involving the SD site of the transposon in this dataset, it did identify SD insertions in the original study by Temiz *et al.*[9]. We expect that the differences between our result and theirs are due to differences in read lengths, as the B-ALL dataset uses 100 bp reads compared to the 50 bp read length used in their dataset. The longer read length makes it more likely that reads overlap the fusion boundary, making an approach that uses chimeric reads preferable with longer read lengths.

4.5 Discussion

We have presented IM-Fusion, a novel approach for identifying transposon insertion sites from gene-transposon fusions in RNA-sequencing data. A key advantage of this approach is that it focuses on identifying insertions that affect gene expression. As such, IM-Fusion provides a significant filter that strongly enriches for insertions that actually affect the expression of their target genes and are therefore most likely to be biologically relevant. This greatly increases the specificity of the approach, providing more confidence in detected insertions and genes and increasing our power to identify rare candidate genes. Furthermore, by combining the insertions with a differential expression analysis, IM-Fusion provides valuable insight into the effect of insertions on the affected target genes.

An important advantage of using RNA-sequencing rather than targeted DNA-sequencing for identification of transposon insertions, is that RNA-sequencing provides much more information than just the location of insertion sites. For example, IM-Fusion uses RNA-expression information to determine how a gene is affected by the presence of an insertion. The same expression data may also be used to identify more global changes in gene expression associated with tumor subtypes or with specific insertions[17], or be used to detect single nucleotide variants and somatic gene fusions that contribute to tumorigenesis. As an example of the latter, we have identified several endogenous fusions in the ILC and B-ALL datasets (Supplementary Figure S4.7 and Table S4.10), including several *Fgfr2* fusions that reflect known oncogenic fusions previously identified in human cancers[27]. Most importantly, these extra analyses can be performed on the same RNA-seq sample, thereby inherently avoiding potential discrepancies resulting from the use of different tumor material for DNA- and RNA-sequencing, an issue that we encountered in the analyses of both the ILC and the B-ALL datasets.

A potential limitation of IM-Fusion is that it requires splicing between the transposon and the affected genes to identify the corresponding insertions. As a result, it will not detect transposon insertions that affect expression via enhancer sequences, as the effects of these insertions are not mediated via splicing. In our analyses, this does not seem to be an issue, as DNA-sequencing approaches did not identify any candidate genes that were perturbed via enhancer effects. This suggests that the MSCV enhancer sequence present in the *T2/Onc* transposon is not particularly active and that the transposon therefore mainly affects expression via splicing. This notion is in agreement with previous studies reporting preferential intragenic insertion of the *T2/Onc* transposon[8], making it less likely to act as an enhancer. Enhancer effects may however play a more important role in case other transposons are employed. Similarly, IM-Fusion may be unable to detect insertions that result in transcript

instability or degradation, as these will be under-represented in the RNA-seq data. Although we do not observe evidence pointing to transcript degradation in the presence of (clonal) SB insertions (Supplementary Figure S4.8), other transposons might have different effects on transcript stability.

A strategy to identify both insertions whose effects are mediated by transcriptional enhancement and insertions that affect expression via splicing, would be to combine DNA- and RNA-sequencing methods, ideally using RNA and DNA isolated from the same sample. In such a combined approach, RNA-sequencing could be used to identify and characterize insertions that are mediated via splicing. For insertions that are uniquely identified by DNA-sequencing, the RNA-seq data could be used to analyze their effects on expression of the predicted target genes. Such a strategy would effectively unite the advantages of both approaches, by combining the unbiased identification of insertion sites by DNA-sequencing with the additional biological information provided by RNA-sequencing in a single analysis.

Although Temiz *et al.*[9] have provided a proof-of-concept showing that transposon insertions can be identified via paired-end RNA-sequencing, our analysis was performed on a much larger dataset (123 versus 20 samples), allowing us to determine biases that affect insertion detection in DNA- and RNA-sequencing data and identify potential limitations of either approach. Furthermore, IM-Fusion improves on Fusion Finder by using a fusion-aware RNA-seq aligner to identify transposon insertions, which enables the use of single-end RNA-sequencing data and increases the sensitivity and the accuracy of insertion detection by also using chimeric reads to identify gene-transposon fusions. Finally, IM-Fusion is provided as comprehensive software package that enables users to perform the entire analysis from start to finish, including the generation of augmented reference genomes, identification of CTGs and testing for differential expression.

In summary, IM-Fusion provides a convenient approach for the identification of insertion sites and their effects on target gene expression from standard single- and paired-end RNA-sequencing data. By combining the identification of insertion sites with expression data, our approach provides valuable insight into the effect of an insertion on its target gene(s) and helps prioritize insertions that are biologically relevant. We expect that this approach will significantly enhance the accuracy of cancer gene discovery in forward genetic screens and prioritization of the identified candidate cancer genes for functional validation studies.

Acknowledgments We would like to thank the animal housing, animal pathology and genomics core facilities for their help in generating the mice and the genomics

data used in this work. We would also like to thank Sander Canisius for his critical reading of the draft manuscript and for his help in testing the software.

Funding Worldwide Cancer Research grant 07-0585; Netherlands Organization for Scientific Research [NWO-NGI Zenith grant 93512009, NWO-ZonMW Vici grant 91814643, Cancer Genomics Netherlands (CGCNL), Cancer Systems Biology Center (CSBC)]; European Research Council (ERC Synergy grant 319661: CombatCancer). Funding for open access charge: ERC.

4.6 Supplementary Material

4.6.1 Supplementary Figures

A Sample overview

#	Sample	Chrom.	Position	Strand	Support	Transposon feature	Gene	Type
1	13SKA014-R3	15	77807867	1	198	En2SA	*Myh9*	Shared
2	12SKA029-R3	15	77807867	1	10	En2SA	*Myh9*	IM-Fusion-specific
3	12SKA101-L3	15	77807867	-1	10	SA	*Myh9*	Shared
4	12SKA102-R3	15	77807867	-1	2	SA	*Myh9*	IM-Fusion-specific
		1	182448400	1	7	SD	*Trp53bp2*	Shared
5	11KOU051-R3	1	182448400	1	23	SD	*Trp53bp2*	Shared
6	12SKA017	2	84615087	1	6	SA	*Ctnnd1*	IM-Fusion-specific
7	12SKA108-R3	2	84650405	1	2	SA	*Ctnnd1*	Shared
8	11KOU012-R5	2	84650405	1	9	SA	*Ctnnd1*	Shared

B Validation results

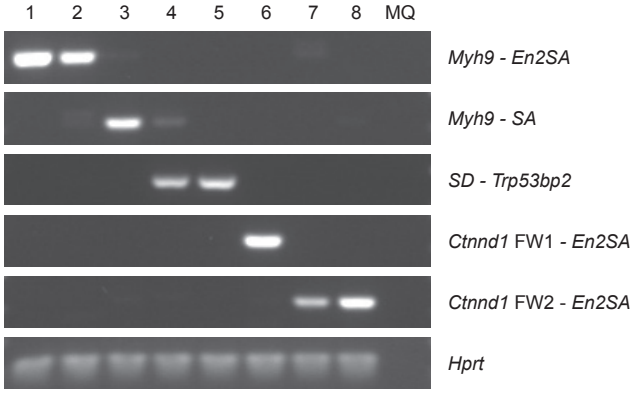


Fig. S4.1. Validation of gene-transposon fusions from the ILC dataset via targeted PCR on cDNA. A selection of predicted insertions from IM-Fusion were validated using PCR primers targeting the corresponding fusions. **(A)** Overview of the samples used in the validation. The *Type* column indicates whether the insertion was shared or only identified by IM-Fusion. Samples were chosen to (i) include a mix of SA/SD/En2SA insertions and (ii) span a range of low/high support scores. **(B)** Results of the validation, showing that each of the expected fusions is indeed detected in the cDNA of the corresponding sample. MQ (= MilliQ) is a water control, which was used a negative control.

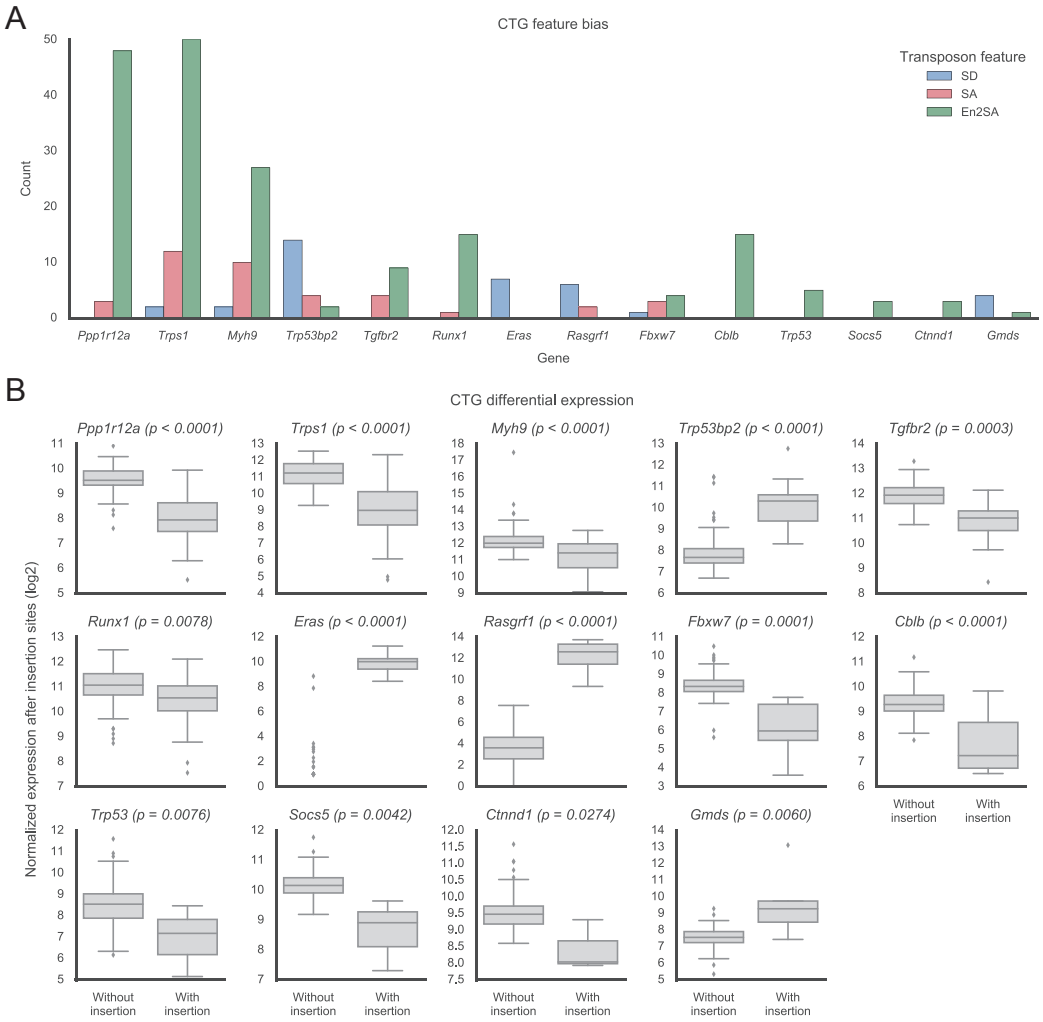


Fig. S4.2. **Feature bias and differential expression for all DE CTGs in the ILC dataset.** **(A)** Frequencies of the transposon features involved in the insertions for each CTG. A bias towards SA/En2SA favors truncation of the gene, whereas a bias towards SD favors overexpression. **(B)** Differential expression across the insertion sites for each CTG. An increase in the presence of an insertion indicates overexpression, a decrease indicates truncation.

A Sample overview

#	Insertion id	Sample	Chrom.	Position	Strand	Support	Gene	Type
1	11KOU029-R5.INS_12	1566_15_11KOU029-R5	11	79461479	1	11	*Nf1*	Shared
2	12SKA029-R3.INS_15	2049_38_12SKA029_R3	11	79359162	-1	10	*Nf1*	ShearSplink-specific
3	12SKA033-R3.INS_10	2800_1_12SKA033-R3	11	79426158	-1	6	*Nf1*	ShearSplink-specific
4	12SKA068-L3.INS_15	2800_20_12SKA068-L3	11	79439837	1	2	*Nf1*	ShearSplink-specific
5	12SKA092-L2.INS_10	2800_36_12SKA092-L2	11	79446215	-1	15	*Nf1*	ShearSplink-specific
6	12SKA104-R3.INS_1	2800_44_12SKA104-R3	11	79382459	-1	58	*Nf1*	ShearSplink-specific

B Validation results

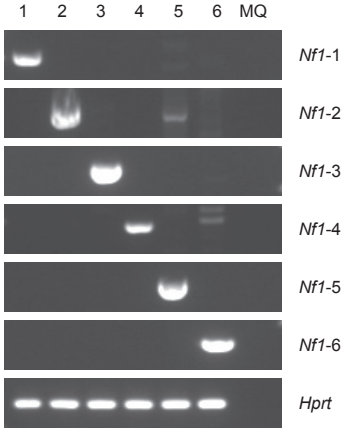


Fig. S4.3. Validation of ShearSplink insertions from the ILC dataset via targeted PCR on DNA. Several ShearSplink insertions in *Nf1* were validated using PCR primers targeting the insertion sites. **(A)** Overview of the samples and insertions used in the validation. The *Type* column indicates whether the insertion was shared or only identified by ShearSplink. **(B)** Results of the validation, showing that each of the expected insertions is indeed detected in the DNA of the corresponding sample. MQ (= MilliQ) is a water control, which was used a negative control.

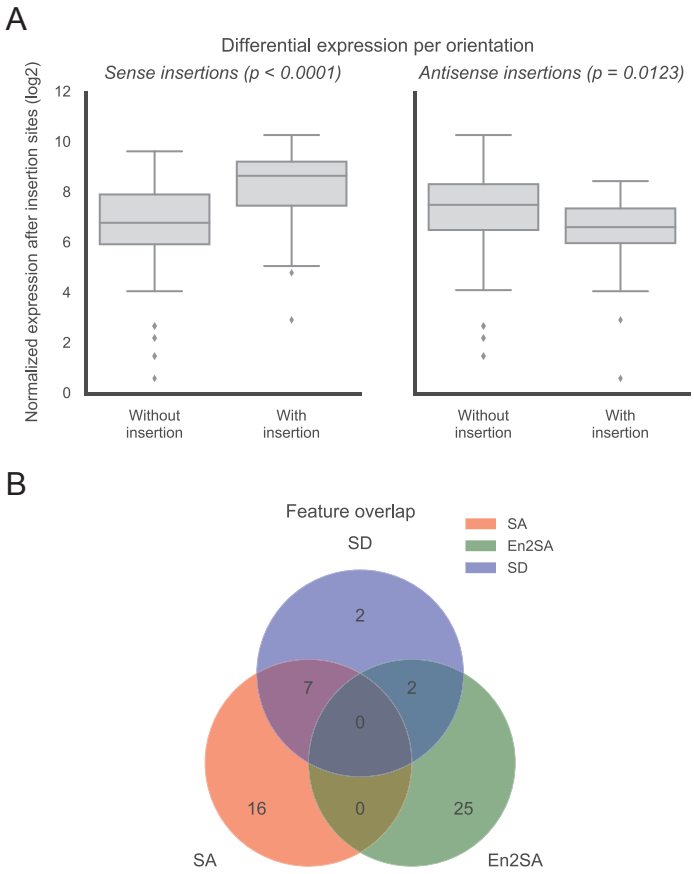


Fig. S4.4. **Detailed analysis of insertion effects in *Fgfr2* in the ILC dataset. (A)** Stratifying the differential expression test for the orientation of insertions in samples shows that samples with sense insertions (using the SA/SD sites) show significant overexpression of the end of the gene, whilst samples with antisense insertions (using the En2SA site) show a decrease in expression. **(B)** The majority of samples with a sense insertion show both truncation of the transcript via the SA site and overexpression via the SD site. This indicates that these samples effectively have a truncated gene transcript, but that the remainder of the gene is simultaneously overexpressed as a separate transcript.

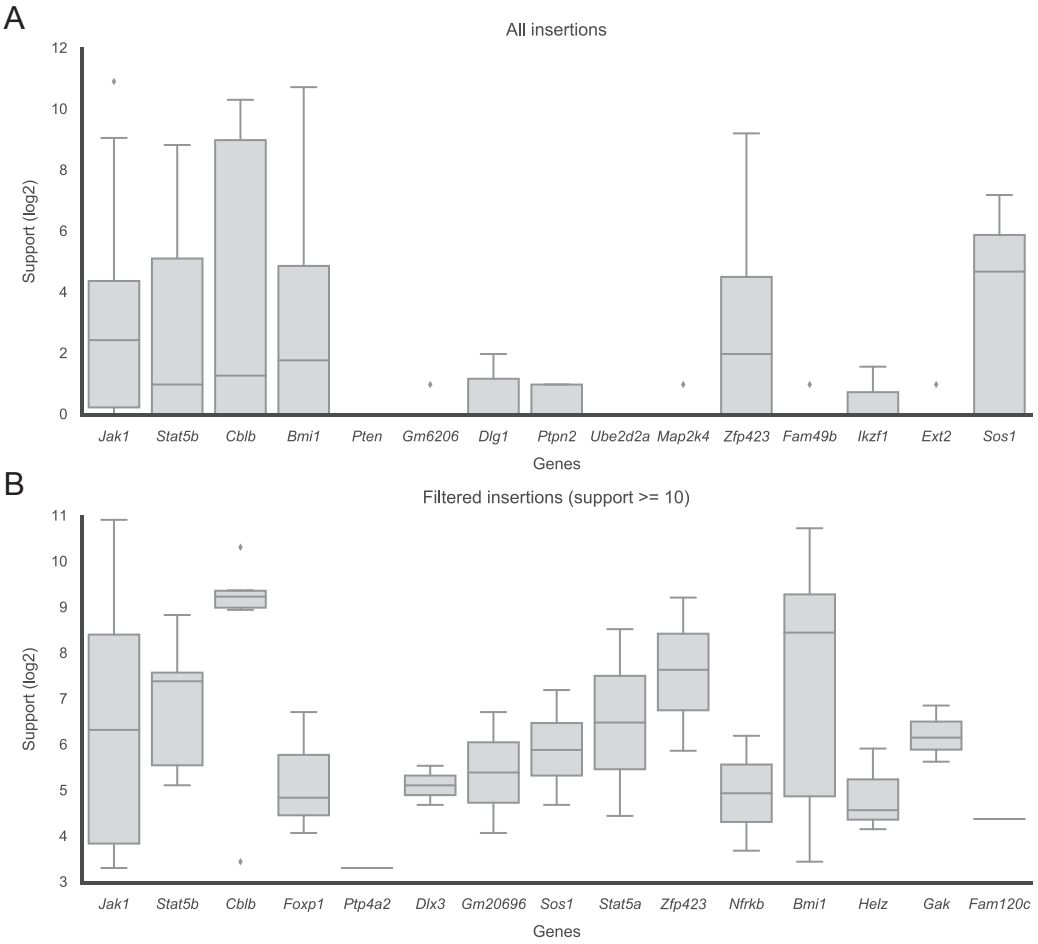


Fig. S4.5. Number of supporting mates in the B-ALL paired-end RNA-sequencing data, before and after filtering by support scores. (A) Support for top genes (ranked by insertion frequency) before filtering insertions for a minimum support of 10 reads. This shows that a number of genes (such as *Pten*, *Gm6206* and *Ube2d2a*) recur frequently, but have very low support and are therefore more likely to represent false positives or weak/subclonal insertions. **(B)** Support for the top genes after filtering for a minimum support of 10 mates. This shows that the remaining genes all have a reasonable number of supporting reads. As such, this filtering improves the confidence in any CTGs identified from this filtered set of insertions and improves detection power by limiting the number of tests (thereby reducing the multiple testing correction).

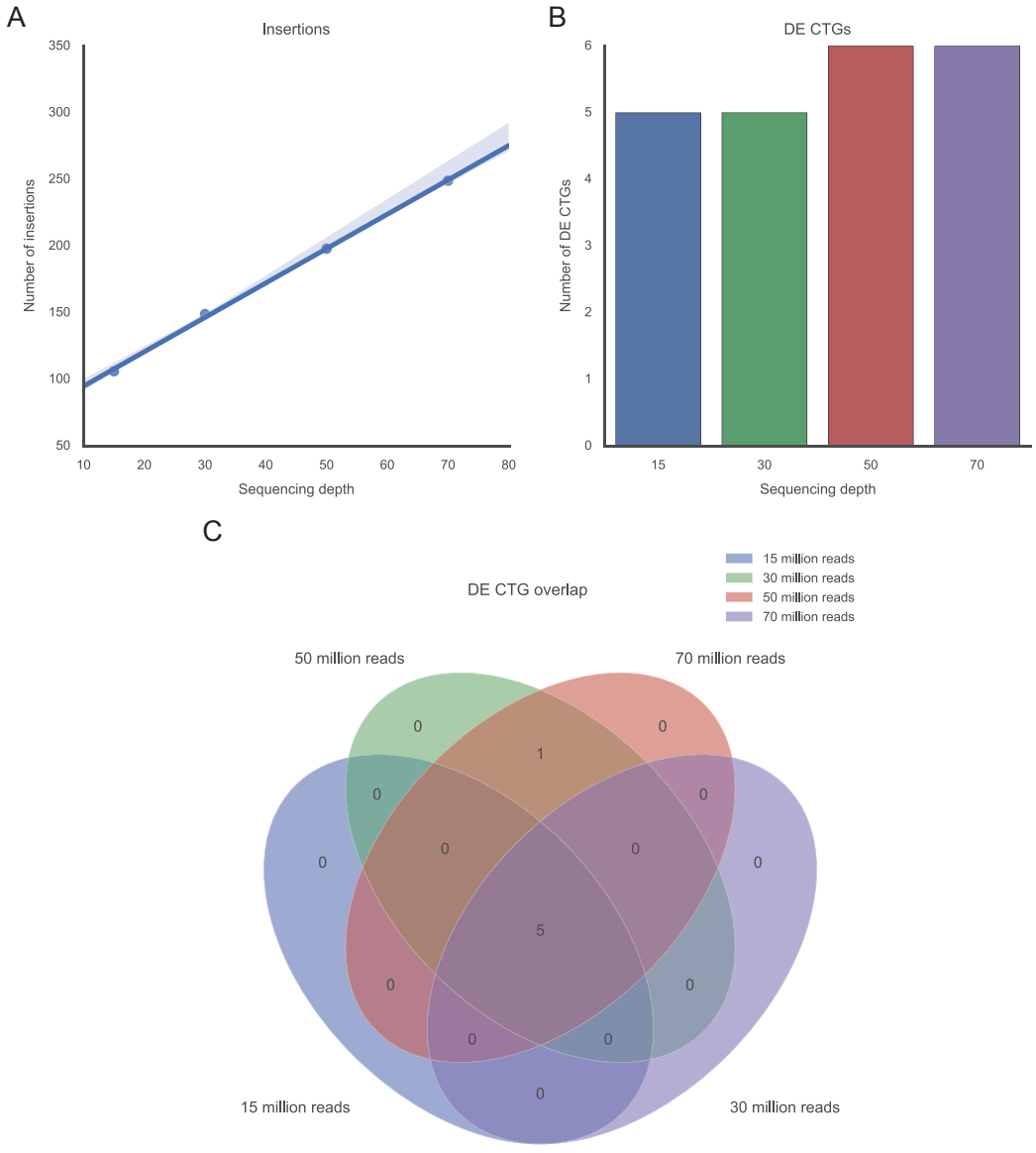


Fig. S4.6. Effects of downsampling in the B-ALL dataset. (A) The number of detected insertions as a function of sequencing depth, showing a linear relation between sequencing depth and the number of insertions. **(B)** The number of DE CTGs detected at different sequencing depths. **(C)** Overlap between DE CTGs detected across the different sequencing depths.

A Sample overview

Sample	Left gene	Right gene	Left breakpoint	Right breakpoint	Support (junction)
12SKA127-R3	*Fgfr2*	*Kif16b*	7:130167703:-	2:142834136:-	5
12SKA035-L3	*Fgfr2*	*Myh9*	7:130167703:-	15:77767663:-	3
11KOU023	*Fgfr2*	*Tbc1d1*	7:130167703:-	5:64256715:+	11

B Validation results

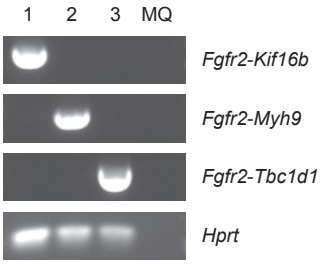


Fig. S4.7. Validation of endogenous *Fgfr2* fusions from the ILC dataset via targeted PCR on cDNA. (A) Overview of endogenous fusions from STAR-Fusion (Supplementary Table S4.10) involving *Fgfr2*, which reflect oncogenic fusions of *FGFR2* identified in human breast cancers. **(B)** Validation results for these fusions, showing that each of the expected fusions is indeed detected in the cDNA of the corresponding sample. MQ (= MilliQ) is a water control, which was used a negative control.

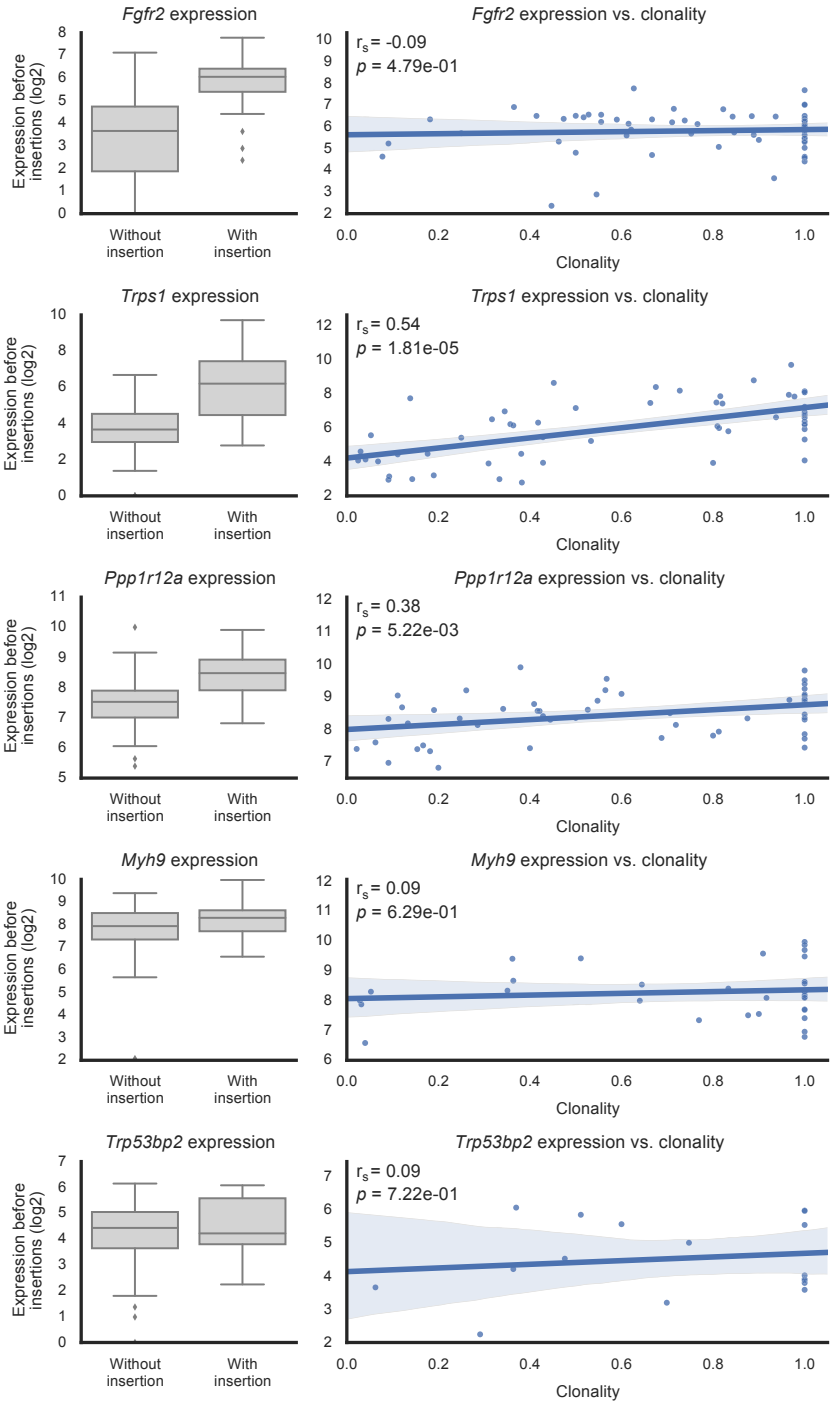


Fig. S4.8. Effects of ShearSplink insertions on transcript stability (quantified using expression before the insertion sites) in the ILC dataset. Compares expression between samples with and without an insertion (left) and across samples with varying levels of insertion clonality (right), showing that (clonal) insertions do not result in reduced expression. Expression values were quantified using the exons before the insertion sites (after normalizing for overall differences in expression between samples), as the expression of these exons should not be affected by the insertion(s). Correlations and p values were calculated using Spearman's Rank correlation (indicated as r_s and p, respectively). For brevity, results are shown for the top 10 ShearSplink candidate genes (continued on the next page).

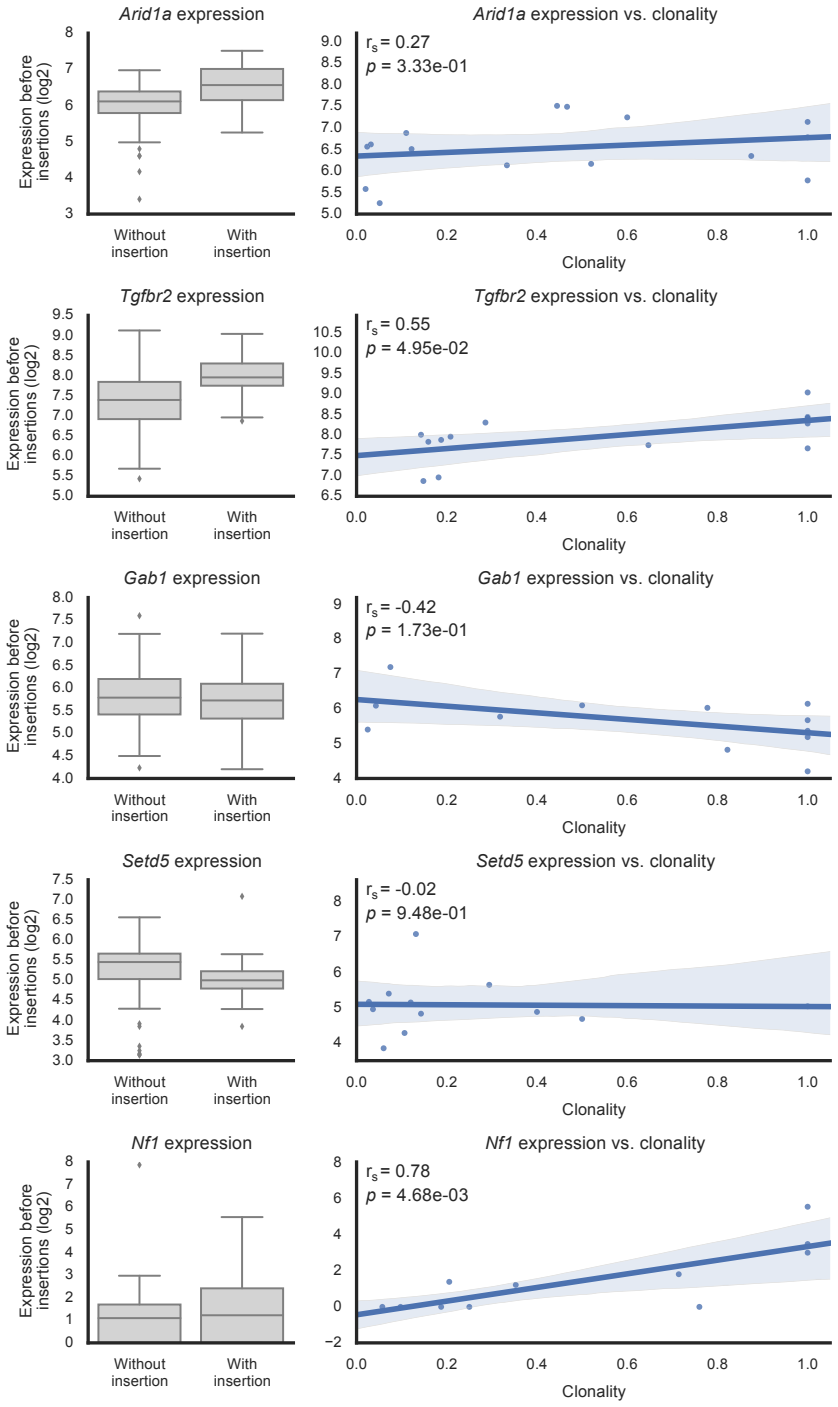


Fig. S4.8. Effects of ShearSplink insertions on transcript stability (continued).

4.6.2 Supplementary Tables (available online)

Tab. S4.1. **Quantification of fusion reads in the ILC dataset.** Quantification of the number of RNA-seq reads in each sample of the ILC dataset, together with the number of reads supporting gene-fusions and how many of these fusion reads support gene-transposon fusions. The two ratio columns indicate the fraction of fusion reads relative to the total number of reads and the fraction of fusions that support gene-transposon fusions.

Tab. S4.2. **Fusions identified by IM-Fusion in the ILC dataset.** Overview of all insertions identified by IM-Fusion in the ILC dataset. The genomic breakpoints of the corresponding fusions are described in the *Chromosome*, *Position* and *Strand* columns. The *Feature* columns describe the transposon features that are involved in the fusions and the position of the transposon breakpoints. The *Gene* columns describe the genes involved in each fusion, whilst the *Novel transcript* column indicates if a novel transcript is created by the fusion. Finally, the *Support* and *Ffpm* columns indicate the degree of support for each fusion (see Methods for more details).

Tab. S4.3. **Overview of the CTGs identified by IM-Fusion in the ILC dataset**. DE CTGs were selected with a corrected CTG p value < 0.05 and a DE p value < 0.05. The DE direction column indicates the direction of the differential expression after the insertion site.

Tab. S4.4. **Quantification of fusion reads in the B-ALL dataset.** Quantification of the number of RNA-seq reads in each sample of the B-ALL dataset, together with the number of reads supporting gene-fusions and how many of these fusion reads support gene-transposon fusions. The two ratio columns indicate the fraction of fusion reads relative to the total number of reads and the fraction of fusions that support gene-transposon fusions (same as for Supplementary Table S4.1).

Tab. S4.5. **Fusions identified by IM-Fusion in the B-ALL dataset.** Overview of all insertions identified by IM-Fusion in the B-ALL dataset. The table structure is the same as described for Supplementary Table S4.2.

Tab. S4.6. **Overview of B-ALL insertions from the original DNA-sequencing analysis for each of the published candidate genes.** Due to lack of the original annotation, insertions were selected for each gene if they occurred within 20kb of the gene. The *Support* column indicates the number of reads supporting the insertions. The *RNAseq* column states whether the insertion was also identified in the (single-end) RNA-sequencing analysis performed using IM-Fusion. Altogether, these tables show that the majority of the insertions not identified by IM-Fusion had a relatively low depth in the DNA-sequencing data. Additionally, *Foxp1* and *Ill2rb* are generally only supported by insertions with low depth.

Tab. S4.7. **Number of samples with insertions in each DE CTG of the B-ALL dataset at different sequencing depths, showing that additional insertions in *Jak1*, *Stat5b* and *Dlx3* are detected at the higher depths.**

Tab. S4.8. **Overview of single- and paired-end support scores for insertions identified by IM-Fusion for the DE CTGs in the B-ALL dataset.** For brevity, the table is limited to the strongest insertion for each gene in the corresponding sample. This table highlights four insertions that were missed by the single-end analysis, as well as differences in support scores between the single-end and paired-end dataset, although the majority of the paired-end insertions are identified from chimeric reads spanning the fusion junction.

Tab. S4.9. **Overview of insertions identified by Fusion Finder in the B-ALL dataset.** For brevity, this table is limited to the published candidate genes and DE CTGs from the IM-Fusion analysis, as these were the candidates of interest for the comparison.

Tab. S4.10. **Overview of the top endogenous fusions detected in the two RNA-seq datasets. (A)** Top 20 fusions identified in the ILC dataset. Fusions with *En2* and *Foxf2* were filtered from the results, as these genes contain sequences that are homologous with the En2SA and SD sequences of the transposon. These fusions therefore actually represent gene-transposon fusions, rather than endogenous fusions, which is also evident from the observation that all involved fusion partners coincide with candidate genes from the IM-Fusion analyses. The three predicted *Fgfr2* fusions were validated using a targeted PCR (Supplementary Figure S4.7), thus confirming the presence of these fusions. **(B)** Top 20 fusions identified in the B-ALL dataset, using the same filtering. The engineered *Etv6-Runx1* fusion was detected in most samples, supporting the validity of the results. This fusion was likely missed in the remaining four samples due to (i) differences in expression of the fusion and/or (ii) differences between the mouse/human sequences of *Runx1* (as the reference used for STAR-Fusion contains the mouse *Runx1* sequence, whilst the engineered fusion was created using the human sequence).

4.7 References

[1] Neal G Copeland and Nancy A Jenkins. “Harnessing transposons for cancer gene discovery”. In: *Nature Reviews Cancer* 10.10 (2010), pp. 696–706 (cit. on p. 106).

[2] Zoltán Ivics, Meng Amy Li, Lajos Mátés, et al. “Transposon-mediated genome manipulation in vertebrates”. In: *Nature Methods* 6.6 (2009), pp. 415–422 (cit. on p. 106).

[3] Jaap Kool and Anton Berns. “High-throughput insertional mutagenesis screens in mice to identify oncogenic networks”. In: *Nature Reviews Cancer* 9.6 (2009), pp. 389–399 (cit. on p. 106).

[4] Lara S Collier and David A Largaespada. “Hopping around the tumor genome: transposons for cancer gene discovery”. In: *Cancer Research* 65.21 (2005), pp. 9607–9610 (cit. on p. 106).

[5] David A Largaespada and Lara S Collier. “Transposon-mediated mutagenesis in somatic cells”. In: *Chromosomal Mutagenesis* (2008), pp. 95–108 (cit. on p. 106).

[6] Aaron L Sarver, Jesse Erdman, Tim Starr, David A Largaespada, and Kevin AT Silverstein. “TAPDANCE: An automated tool to identify and annotate transposon insertion CISs and associations between CISs from next generation sequence data”. In: *BMC Bioinformatics* 13.1 (2012), p. 1 (cit. on p. 106).

[7] Jeroen De Ridder, Anthony Uren, Jaap Kool, Marcel Reinders, and Lodewyk Wessels. “Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens”. In: *PLOS Computational Biology* 2.12 (2006), e166 (cit. on p. 106).

[8] Johann de Jong, Jeroen de Ridder, Louise van der Weyden, et al. “Computational identification of insertional mutagenesis targets for cancer gene discovery”. In: *Nucleic Acids Research* 39.15 (2011), e105–e105 (cit. on pp. 106, 112, 132).

[9] Nuri A Temiz, Branden S Moriarity, Natalie K Wolf, et al. “RNA sequencing of Sleeping Beauty transposon-induced tumors detects transposon-RNA fusions in forward genetic cancer screens”. In: *Genome Research* 26.1 (2016), pp. 119–129 (cit. on pp. 107, 116, 131, 133).

[10] Alexander Dobin, Carrie A Davis, Felix Schlesinger, et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21 (cit. on pp. 108, 118).

[11] Daehwan Kim and Steven L. Salzberg. “TopHat-Fusion: an algorithm for discovery of novel fusion transcripts”. In: *Genome Biology* 12.8 (2011), R72 (cit. on pp. 108, 118).

[12] Nicolas Stransky, Ethan Cerami, Stefanie Schalm, Joseph L Kim, and Christoph Lengauer. “The landscape of kinase fusions in cancer”. In: *Nature Communications* 5 (2014), p. 4846 (cit. on pp. 108, 111, 115).

[13] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, et al. “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads”. In: *Nature Biotechnology* 33.3 (2015), pp. 290–295 (cit. on p. 108).

[14] Yang Liao, Gordon K Smyth, and Wei Shi. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7 (2013), pp. 923–930 (cit. on p. 110).

[15] Simon Anders, Alejandro Reyes, and Wolfgang Huber. “Detecting differential usage of exons from RNA-seq data”. In: *Genome Research* 22.10 (2012), pp. 2008–2017 (cit. on p. 110).

[16] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (2014), pp. 1–21 (cit. on p. 110).

[17] Sjors M Kas, Julian R de Ruiter, Koen Schipper, et al. “Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular breast carcinoma”. In: *Nature Genetics* 49.8 (2017), pp. 1219–1230 (cit. on pp. 112, 121, 132).

[18] Lara S Collier, Corey M Carlson, Shruthi Ravimohan, Adam J Dupuy, and David A Largaespada. “Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse”. In: *Nature* 436.7048 (2005), pp. 272–276 (cit. on p. 112).

[19] Marco J Koudijs, Christiaan Klijn, Louise van der Weyden, et al. “High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors”. In: *Genome Research* 21.12 (2011), pp. 2181–2189 (cit. on pp. 112, 125).

[20] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4 (2012), pp. 357–359 (cit. on p. 112).

[21] Louise van der Weyden, George Giotopoulos, Alistair G Rust, et al. “Modeling the evolution of ETV6-RUNX1–induced B-cell precursor acute lymphoblastic leukemia in mice”. In: *Blood* 118.4 (2011), pp. 1041–1051 (cit. on pp. 112, 129).

[22] Mirjam C Boelens, Micha Nethe, Sjoerd Klarenbeek, et al. “PTEN loss in E-cadherin-deficient mouse mammary epithelial cells rescues apoptosis and results in development of classical invasive lobular carcinoma”. In: *Cell Reports* 16.8 (2016), pp. 2087–2101 (cit. on p. 113).

[23] Heng Li, Bob Handsaker, Alec Wysoker, et al. “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079 (cit. on pp. 113, 116).

[24] Daehwan Kim, Geo Pertea, Cole Trapnell, et al. “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”. In: *Genome Biology* 14.4 (2013), R36 (cit. on pp. 116, 131).

[25] Lynn M Heltemes-Harris, Jon D Larson, Timothy K Starr, et al. “Sleeping Beauty transposon screen identifies signaling modules that cooperate with STAT5 activation to induce B-cell acute lymphoblastic leukemia”. In: *Oncogene* 35.26 (2016), p. 3454 (cit. on p. 130).

[26] Marta Campo Dell’Orto, Barbara Banelli, Emanuela Giarin, et al. “Down-regulation of DLX3 expression in MLL-AF4 childhood lymphoblastic leukemias is mediated by promoter region hypermethylation”. In: *Oncology Reports* 18.2 (2007), pp. 417–423 (cit. on p. 130).

[27] Yi-Mi Wu, Fengyun Su, Shanker Kalyana-Sundaram, et al. “Identification of targetable FGFR gene fusions in diverse cancers”. In: *Cancer Discovery* 3.6 (2013), pp. 636–647 (cit. on p. 132).

5

Comparative oncogenomics and iterative mouse modeling identifies combinations of driver genes and drug targets in *BRCA1*-mutated breast cancer

Stefano Annunziato[1,7,*], Julian R. de Ruiter[1,2,7,*], Linda Henneman[1,3,*], Chiara S. Brambillasca[1,7,*], Catrin Lutz[1,7], François Vaillant[4,5], Frederica Ferrante[1,7], Anne Paulien Drenth[1,7], Eline van der Burg[1,7], Bjørn Siteur[6], Bas van Gerwen[6], Roebi de Bruijn[1,2,7], Martine H. van Miltenburg[1,7], Ivo J. Huijbers[3], Marieke van de Ven[6], Jane E. Visvader[4], Geoffrey J. Lindeman[4], Lodewyk F. A. Wessels[2,7,8,9], Jos Jonkers[1,7,8,9]

[1] Division of Molecular Pathology, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands
[2] Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands
[3] Transgenic Core Facility, Mouse Clinic for Cancer and Aging (MCCA), The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands
[4] ACRF Stem Cells and Cancer Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia
[5] Department of Medical Biology, University of Melbourne, Parkville, Victoria 3010, Australia
[6] Preclinical Intervention Unit, Mouse Clinic for Cancer and Aging (MCCA), The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands
[7] Cancer Genomics Netherlands, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands
[8] Oncode Institute, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands
[9] Correspondence should be addressed to J.J. (j.jonkers@nki.nl) or L.F.A.W. (l.wessels@nki.nl)
[*] these authors contributed equally to this work

Published in Nature Communications, 23 January 2019
Volume 10, Issue 1, Article number 397

5.1 Abstract

BRCA1-mutated breast cancer is primarily driven by DNA copy number alterations (CNAs) containing large numbers of candidate driver genes. Validation of these candidates requires novel approaches for high-throughput *in vivo* perturbation of gene function. Here we develop genetically engineered mouse models (GEMMs) of BRCA1-deficient breast cancer that permit rapid introduction of putative drivers by either retargeting of GEMM-derived embryonic stem cells, lentivirus-mediated somatic overexpression or *in situ* CRISPR/Cas9-mediated gene disruption. We use these approaches to validate *Myc*, *Met*, *Pten* and *Rb1* as *bona fide* drivers in BRCA1-associated mammary tumorigenesis. Iterative mouse modeling and comparative oncogenomics analysis show that MYC-overexpression strongly reshapes the CNA landscape of BRCA1-deficient mammary tumors and identifiy MCL1 as a collaborating driver in these tumors. Moreover, MCL1 inhibition potentiates the *in vivo* efficacy of PARP inhibition (PARPi), underscoring the therapeutic potential of this combination for treatment of BRCA1-mutated cancer patients with poor response to PARPi monotherapy.

5.2 Statement of significance

We demonstrate that engineering additional driver genes in progressively complex mouse models of BRCA1-deficient triple-negative breast cancer can strongly influence their evolutionary trajectory. Comparative oncogenomics analysis of recurrent copy number alterations in human and mouse tumors uncovered new culprits of tumorigenesis and MCL1 as a novel therapeutic vulnerability.

5.3 Introduction

Triple-negative breast cancer (TNBC) accounts for 10%-15% of all breast cancers and is characterized by lack of expression of the estrogen receptor (ER), the progesterone receptor (PR) and the human epidermal growth factor receptor 2 (HER2). Due to the lack of these receptors, TNBCs cannot be treated with targeted therapies that have been effective in treating other breast cancer subtypes. As a result, TNBC has a relatively poor clinical prognosis and chemotherapy remains its current standard-of-care.

At the mutational level, TNBC is primarily a DNA copy number driven disease[1], harboring a multitude of copy number alterations (CNAs) containing various driver

genes[2]. TNBCs are furthermore characterized by mutations in the TP53 tumor suppressor gene, which occur in more than 80% of cases. Moreover, approximately 50% of TNBCs show loss of BRCA1 or BRCA2, either due to germline or somatic mutations or because of promoter hypermethylation[2]. BRCA1 and BRCA2 are crucial for error-free repair of DNA double-strand breaks via homologous recombination, and loss of these genes results in high levels of chromosomal instability and a specific mutator phenotype. This results in recurrent patterns of CNAs in BRCA-deficient tumors, suggesting that these aberrations contain specific driver genes required for tumorigenesis.

Unfortunately, the high degree of genomic instability in BRCA-deficient TNBCs results in large numbers of CNAs harboring tens-to-thousands of genes, which complicates the identification of putative cancer drivers. To address this issue, several computational approaches have been developed to identify minimal regions that are recurrently gained or lost across tumors[3–6]. Other approaches have complemented these tools with comparative oncogenomic strategies, in which combined analyses of human and mouse tumors are used to identify candidate driver genes that are frequently altered in tumors from both species[7–9]. We have previously used comparative oncogenomics analyses to identify driver genes that were frequently aberrantly amplified or deleted in both mouse and human BRCA1-deficient TNBCs, including the proto-oncogene *MYC* and the tumor suppressor *RB1*[10]. However, it is currently still unclear how exactly these putative drivers of BRCA1-deficient TNBC contribute to tumorigenesis, and specifically how they may influence the mutational landscape of the resulting tumors. To address these questions, we generate additional mouse models of BRCA1-deficient TNBC harboring different candidate genes. To overcome the time-consuming nature of generating these mouse models via germline engineering, we develop somatic mouse models of BRCA1-deficient TNBC and we show that these models accurately reflect their germline counterparts. We analyze the resulting tumors to assess the contribution of candidate drivers to BRCA1-associated mammary tumorigenesis and to determine their effect on the copy number landscape. Finally, by applying comparative oncogenomics to a combined set of germline and somatic BRCA1-deficient TNBCs with MYC overexpression, we identify MCL1 as a key driver and a therapeutic target in these tumors.

5.4 Results

5.4.1 Driver landscape in human BRCA1-deficient breast cancer

To determine the mutational landscape of human *BRCA1*-mutated breast cancer, we performed a meta-analysis by combining datasets from four large-scale breast cancer sequencing studies and extracting the mutational data of all *BRCA1*-mutated tumors. This analysis identified a total of 80 breast cancers (~1.5%) with a homozygous deletion or an inactivating (putative) driver mutation in *BRCA1* (Figure 5.1A, Supplementary Table S5.1). For 18 of these cases (~23%) triple-negative (TN) status could not be determined due to missing or inconclusive immunohistochemistry data. Of the remaining 62 cases, 40 (~65%) were scored as TNBC. Association with TN status was stronger in tumors from *BRCA1* germline mutations carriers (27/30) than in tumors with *BRCA1* somatic mutations (13/32).

We next analyzed the mutational landscape of the 80 BRCA1-deficient breast cancer cases, focusing on deleterious mutations, amplifications and homozygous deletions. At the mutational level, these tumors were mainly characterized by mutations in *TP53* (52/80, ~65%) and *PIK3CA* (23/80, ~29%). At the copy number level, the most prominent events included amplifications of MYC (35/80, ~44%) and several co-amplified genes (e.g. *RAD21*, *EXT1*, *RECQL4*, *RSPO2*, *EPPK1*, *PLEC*) in the same locus (30-34%). MYC is a particularly well-known transcription factor that lies at the crossroad of several growth-promoting pathways and regulates global gene expression, resulting in increased proliferation and influencing many other cellular processes (reviewed by Meyer *et al.*[11] and Kress *et al.*[12]). The *MYC* oncogene resides in the 8q24 genomic locus, which is among the most frequently amplified regions in breast cancer[13], particularly in TNBC[14]. *MYC* expression and MYC signaling are aberrantly elevated in TNBC[15,16] and a MYC transcriptional gene signature has been correlated with basal-like breast cancer (BLBC), a subtype typical for human BRCA1-deficient breast cancer[17–19]. Altogether, this confirms that human BRCA1-deficient breast cancers are enriched for TNBCs and are mainly characterized by inactivating mutations in *TP53* and amplification of *MYC*.

5.4.2 MYC is a potent driver in BRCA1-associated tumorigenesis

To study the contribution of MYC overexpression to BRCA1-associated mammary tumorigenesis, we initially employed the *K14Cre;Brca1$^{F/F}$;Trp53$^{F/F}$* (KB1P) mouse

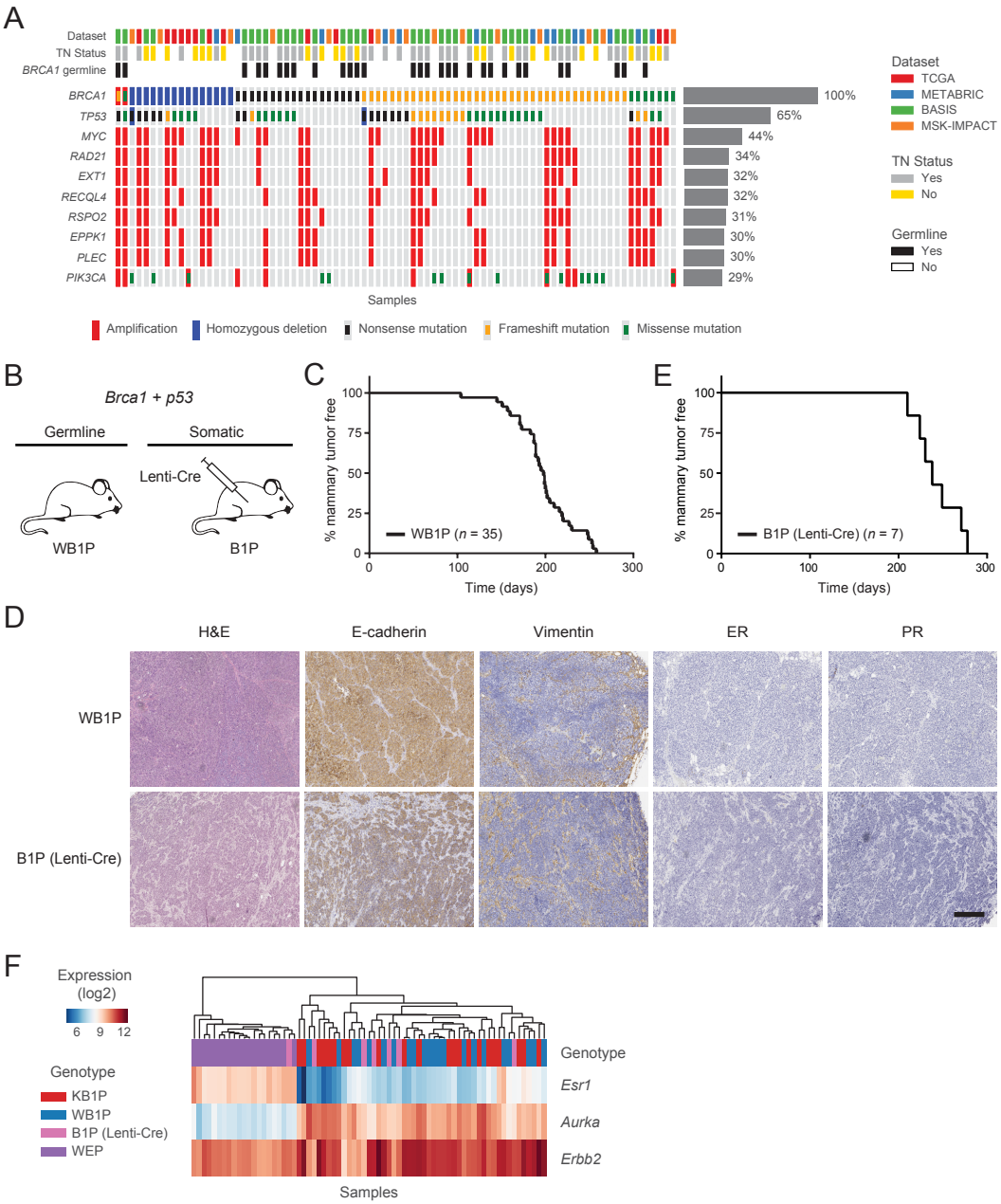


Fig. 5.1. Mutational landscape of human BRCA1-mutated TNBC and characterization of the WB1P model. **(A)** Overview of the most common deleterious mutations and copy number events in 80 *BRCA1*-mutated human breast tumor samples from four large-scale tumor-sequencing studies. **(B)** Overview of the germline and somatic mouse models for mammary gland-specific inactivation of conditional *Brca1* and *Trp53* alleles. **(C)** Kaplan-Meier curve showing mammary tumor-specific survival for *WapCre;Brca1*$^{F/F}$*;Trp53*$^{F/F}$ (WB1P) female mice. **(D)** Representative hematoxylin and eosin (HE) staining and immunohistochemical detection of E-cadherin, vimentin, ER and PR in WB1P tumors and in tumors from Lenti-Cre injected *Brca1*$^{F/F}$*;Trp53*$^{F/F}$ (B1P) mice. Bar, 400 μm. **(E)** Kaplan-Meier curve showing mammary tumor-specific survival of B1P females injected with Lenti-Cre. **(F)** Unsupervised clustering (Euclidean distance, average linkage) of the WB1P tumors with tumors derived from published mouse models of luminal (*WapCre;Cdh1*$^{F/F}$*;Pten*$^{F/F}$, WEP[20]) and basal-like (*K14Cre;Brca1*$^{F/F}$*;Trp53*$^{F/F}$, KB1P[21]) breast cancer, using a three-gene signature that distinguishes the PAM50 subtypes[22].

model[21], in which epithelium-specific loss of BRCA1 and p53 leads to the formation of mammary tumors and, to a lesser extent, other epithelial tumors including skin tumors. We used our previously established GEMM-ESC pipeline[23] to generate *K14Cre;Brca1$^{F/F}$;Trp53$^{F/F}$;Col1a1$^{invCAG\text{-}Myc\text{-}IRES\text{-}Luc/+}$* (KB1P-Myc) mice with epithelium-specific loss of BRCA1 and p53 and overexpression of MYC. Unfortunately, these mice were more prone to developing non-mammary tumors than KB1P mice and had to be sacrificed around 110 days for skin cancers and thymomas due to expression of K14Cre in these tissues.

To avoid unwanted development of non-mammary tumors, we took a two-pronged approach (Figure 5.1B). On one hand, we developed a novel GEMM (*WapCre;Brca1$^{F/F}$;Trp53$^{F/F}$*, WB1P) in which mammary-specific expression of Cre is driven by the whey acidic protein (*Wap*) gene promoter. In this WB1P model, female mice spontaneously developed mammary tumors with a median latency of 198 days ($n = 35$, Figure 5.1C), which is comparable to the latency of KB1P females (median latency of 197 days, $n = 41$). Similar to KB1P mammary tumors, WB1P tumors were either pure carcinomas (83%) or carcinosarcomas (17%). All tumors were poorly differentiated, negative for ER and PR (Figure 5.1D) and showed recombination of the *Brca1^{F}* and *Trp53^{F}* alleles. On the other hand, we employed a somatic strategy and performed intraductal injection of lentiviral vectors[24–26] expressing the Cre-recombinase (Lenti-Cre) in *Brca1$^{F/F}$;Trp53$^{F/F}$* (B1P) females. Tumors from B1P mice injected with Lenti-Cre had a median latency of 238 days after injection ($n = 7$, Figure 5.1E), and in terms of their morphology, they were indistinguishable from WB1P tumors (Figure 5.1D).

To determine if tumors from these two new mouse models reflected the basal-like subtype typical for human BRCA1-deficient breast cancer, we performed RNA-sequencing on 22 WB1P tumors and 7 tumors from B1P mice injected with Lenti-Cre, and compared their expression profile to tumors from the KB1P mouse model and a mouse model of luminal breast cancer[20] (*WapCre;Cdh1$^{F/F}$;Pten$^{F/F}$*, WEP), using a three-gene signature that distinguishes the PAM50 subtypes[22]. This analysis showed that all *Brca1$^{\Delta/\Delta}$;Trp53$^{\Delta/\Delta}$* mouse mammary tumors from the three different mouse models cluster together and are characterized by low expression of *Esr1* and high expression of the proliferation marker *Aurka* (Figure 5.1F), reflecting the expression profile of human BLBC (Supplementary Figure S5.1).

To study the effects of *Myc* amplification in WB1P mice, we applied the GEMM-ESC strategy[23] to insert the conditional *invCAG-Myc-IRES-Luc* cassette into the *Col1a1* locus of WB1P embryonic stem cells. In the resulting *WapCre;Brca1$^{F/F}$;Trp53$^{F/F}$;Col1a1$^{invCAG\text{-}Myc\text{-}IRES\text{-}Luc/+}$* (WB1P-Myc) model, mammary-specific expression of Cre induces inactivation of BRCA1 and p53 and concomitant overexpression of the MYC oncogene accompanied by luciferase expression (Fig-

ure 5.2A). WB1P-Myc female mice developed multifocal mammary tumors with a median latency of 97 days ($n = 35$, Figure 5.2B). These tumors grew exponentially (Supplementary Figure S5.2A) and animals had to be sacrificed 2-3 weeks after detection of palpable tumors. In contrast to the KB1P-Myc mice, WB1P-Myc mice developed only mammary tumors.

To test if somatic engineering could be used to overexpress MYC in the mammary gland, we performed intraductal injections of Lenti-Cre in *Brca1*$^{F/F}$*;Trp53*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Myc\text{-}IRES\text{-}Luc/+}$ (B1P-Myc, $n = 16$) females (Figure 5.2A). Moreover, we also injected lentiviral vectors encoding both Cre and *Myc* (Lenti-MycP2ACre, Supplementary Figure S5.3A) in B1P females ($n = 13$) and lentiviral vectors encoding Myc (Lenti-Myc) in WB1P mice ($n = 15$). Mice from all three groups developed mammary tumors with 100% penetrance and specifically in the injected glands (Figure 5.2C). B1P-Myc mice injected with Lenti-Cre developed tumors much faster than B1P mice injected with Lenti-Cre (126 days after injection vs 238 days after injection). B1P females injected with Lenti-MycP2ACre and WB1P females injected with Lenti-Myc developed tumors even faster (median latency of 92 and 61 days after injection, respectively), most likely due to higher *Myc* expression from the viral constructs than from the knock-in allele (Supplementary Figure S5.3B).

Histopathological analysis showed that, unlike the WB1P mouse model, WB1P-Myc females developed multifocal tumors that were all carcinomas. However, similar to WB1P tumors, WB1P-Myc tumors were poorly differentiated and ER-/PR-negative (Figure 5.2D). Furthermore, they displayed recombined *Brca1* and *Trp53* alleles and were sensitive to cisplatin and PARP inhibitors upon transplantation into nude mice (Supplementary Figure S5.2B). *WapCre;Brca1*$^{F/+}$*;Trp53*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Myc\text{-}IRES\text{-}Luc/+}$ females that were heterozygous for *Brca1*^{F} alleles ($n = 20$) developed tumors slightly but significantly slower than *WapCre;Brca1*$^{F/F}$*;Trp53*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Myc\text{-}IRES\text{-}Luc/+}$ mice with homozygous *Brca1*^{F} alleles (Supplementary Figure S5.2C). Histopathological analysis showed that mammary tumors from the somatic models were indistinguishable from the cognate tumors from the germline models (Figure 5.2D). WB1P-Myc tumors showed similar expression levels of *Esr1* and *Aurka* as the WB1P tumors, indicating that they retained their basal-like subtype (Supplementary Figure S5.2D). Besides this, WB1P-Myc tumors showed high mRNA and protein levels of MYC compared to WB1P tumors, demonstrating successful expression of the knock-in allele (Supplementary Figure S5.2E-F). Unsupervised clustering of RNA-seq data from tumors from the somatic models confirmed that they also retained their basal-like phenotypes, and PCA analysis showed that these tumors also resemble their counterparts from the germline models in terms of their global gene expression profiles (Supplementary Figure S5.3C-E). Taken together, these data

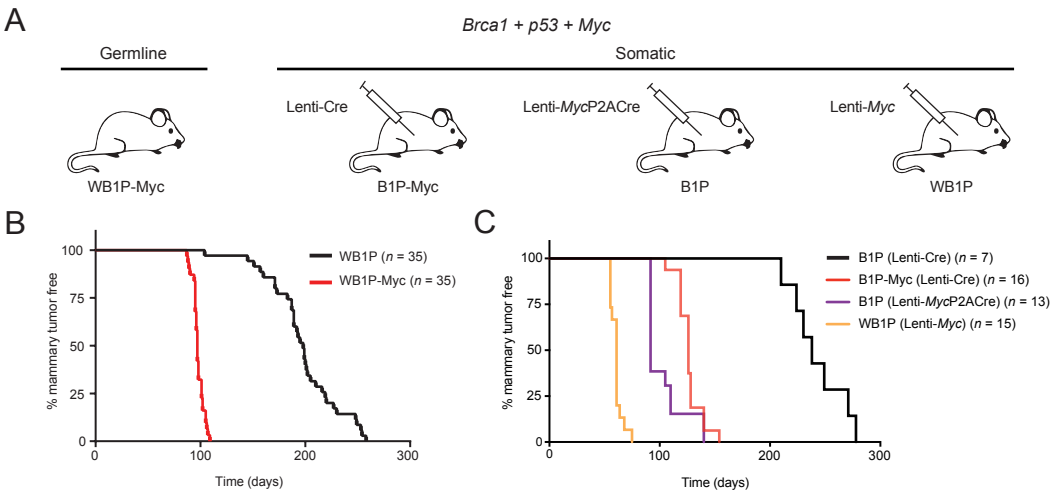


Fig. 5.2. Validation of additional drivers in WB1P mice using germline and somatic engineering. **(A)** Overview of the germline and somatic mouse models for mammary gland-specific *Myc* overexpression in mice with conditional *Brca1* and *Trp53* alleles. **(B)** Kaplan-Meier curves showing mammary tumor-specific survival for the different genotypes. *WapCre;Brca1*$^{F/F}$*;Trp53*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Myc\text{-}IRES\text{-}Luc/+}$ (WB1P-Myc) females showed a reduced mammary tumor-specific survival compared to WB1P littermates (97 days vs 198 days; $^{****}p < 0.0001$ by Mantel-Cox test). **(C)** Kaplan-Meier curves showing mammary tumor-specific survival for the different non-germline models. *Brca1*$^{F/F}$*;Trp53*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Myc\text{-}IRES\text{-}Luc/+}$ (B1P-Myc) females injected with Lenti-Cre, B1P females injected with Lenti-MycP2ACre and WB1P females injected with Lenti-Myc showed a reduced mammary tumor-specific survival compared to B1P female mice injected with Lenti-Cre (respectively 126, 92 and 61 days after injection vs 238 days after injection; $^{****}p < 0.0001$ by Mantel-Cox test).

provide functional validation in germline and somatic models of the role of MYC in BRCA1-associated mammary tumorigenesis.

5.4.3 Loss of PTEN and RB1 collaborates with MYC in tumorigenesis

After MYC amplification, the next most common alterations in our analysis of the human BRCA1-deficient TNBCs were mutations and/or amplifications of *PIK3CA* (23/80 cases), indicating that activation of PI3K signaling is an important driver in this breast cancer subtype (Figure 5.1A). Indeed, in addition to *PIK3CA* mutation/amplification, heterozygous or homozygous loss of *PTEN* (a negative regulator of PI3K signaling) was observed in 29/80 and 6/80 cases, respectively (Table S5.1). Genetic alterations of *PIK3CA/PTEN* and *MYC* co-occurred in ~29% of all tumors analyzed (23/80 cases), indicating that MYC overexpression and PI3K pathway activation collaborate in BRCA1-related breast tumorigenesis.

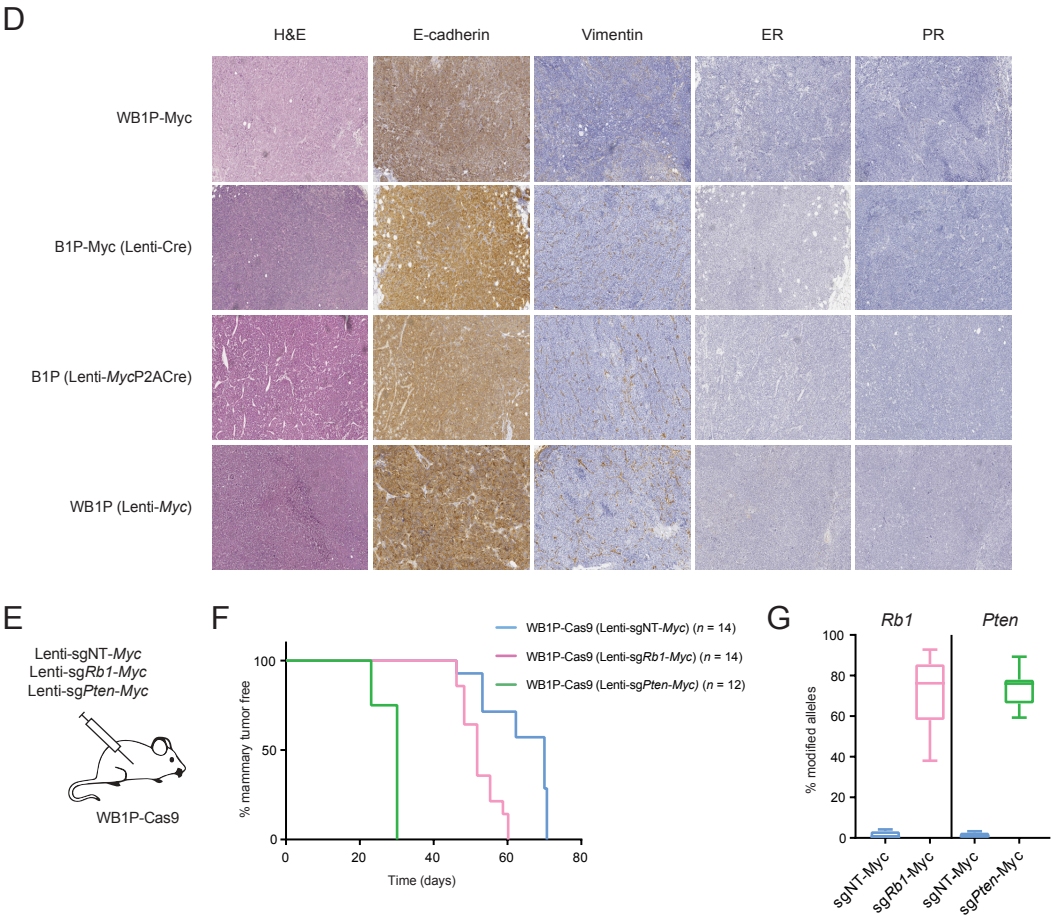


Fig. 5.2. **Continued.** **(D)** Representative hematoxylin and eosin (HE) staining and immunohistochemical detection of E-cadherin, vimentin, ER and PR in tumors from WB1P-Myc females and in tumors from Lenti-Cre injected B1P-Myc mice, Lenti-MycP2ACre injected B1P mice and Lenti-Myc injected WB1P mice. Scale bar, 400 μm. **(E)** Overview of the intraductal injections performed in *WapCre;Brca1*$^{F/F}$*;Trp53*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Cas9\text{-}IRES\text{-}Luc/+}$ (WB1P-Cas9) females with high-titer lentiviruses encoding Myc and either a non-targeting (NT) sgRNA (Lenti-sgNT-Myc), a sgRNA targeting exon 2 of *Rb1* (Lenti-sg*Rb1*-Myc) or a sgRNA targeting exon 7 of *Pten*. **(F)** Kaplan-Meier curves showing mammary tumor-specific survival for the different models. WB1P-Cas9 females injected with Lenti-sg*Pten*-Myc and Lenti-sg*Rb1*-Myc showed a reduced mammary tumor-specific survival compared to WB1P-Cas9 female mice injected with Lenti-sgNT-Myc (respectively 30 and 52 days after injection vs 70 days after injection, $^{****}p < 0.0001$ and $^{***}p < 0.001$ by Mantel-Cox test). **(G)** Boxplots depicting the fraction of modified Rb1 and Pten alleles in tumors from WB1P-Cas9 mice injected with Lenti-sgNT-Myc, Lenti-sg*Rb1*-Myc and Lenti-sg*Pten*-Myc. 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.

To assess if activation of PI3K signaling via loss of PTEN collaborates with MYC overexpression in BRCA1-deficient TNBC, we developed *WapCre;Brca1$^{F/F}$;Trp53$^{F/F}$;Col1a1$^{invCAG\text{-}Cas9\text{-}IRES\text{-}Luc/+}$* (WB1P-Cas9) mice with mammary-specific loss of BRCA1 and p53 and concomitant expression of Cas9. We then cloned and validated lentiviral vectors encoding a nontargeting sgRNA (sgNT) or a sgRNA targeting the seventh exon of *Pten* (sg*Pten*), in combination with a *Myc*-overexpression cassette. Since also RB1 loss has been implicated in BRCA1-deficient breast cancer[27] and MYC-driven TNBC[28], we also generated a similar lentiviral vector encoding MYC and a sgRNA targeting the second exon of *Rb1* (sg*Rb1*). These lentiviral vectors (Lenti-sgNT-Myc, Lenti-sg*Pten*-Myc and Lenti-sg*Rb1*-Myc) were injected intraductally into WB1P-Cas9 females (Figure 5.2E) resulting in tumor formation with high penetrance and very short latency (70, 30 and 52 days after injection, respectively; $n = 14$, 12 and 14, respectively, Figure 5.2F). Genomic DNA of mammary tumors from Lenti-sg*Pten*-Myc and Lenti-sg*Rb1*-Myc injected WB1P-Cas9 mice showed extensive modification of the target gene (Figure 5.2G; Supplementary Figure S5.4A-B), with a strong bias towards indels resulting in frameshift mutations, supporting homozygous inactivation of the tumor suppressor genes. Together, these results demonstrate that activation of PI3K signaling and RB1 loss collaborate with MYC in BRCA1-deficient TNBC.

5.4.4 MYC overexpression reshapes the copy number landscape

To identify additional collaborating driver genes in BRCA1-deficient TNBC, we decided to characterize the CNA landscape of WB1P and WB1P-Myc tumors, with the assumption that recurrent CNAs in these tumors might underscore a conserved selective pressure towards the specific gain or loss of cancer genes that collaborate with loss of BRCA1 and p53 – alone or in combination with MYC overexpression – during TNBC development. We therefore performed DNA copy number sequencing (CNV-seq) on 39 WB1P tumors and identified recurrent CNAs using RUBIC[6]. This analysis showed that WB1P tumors exhibit a high degree of genomic instability and harbor a multitude of recurrent gains and losses (Figure 5.3A; Supplementary Figure S5.5A). The most evident of these events was a focal amplification on chromosome 6 containing the *Met* oncogene. Besides *Met*, we also identified a recurrent loss on chromosome 14 (harboring *Rb1*) and several amplifications on chromosome 15 (containing *Myc*), in line with our previous studies in KB1P mice[10].

Remarkably, CNV-seq of 19 WB1P-Myc tumors showed a dramatically reshaped copy number landscape (Figure 5.3B), with significantly fewer CNAs compared to the WB1P model (Figure 5.3C; $p < 0.00001$, two-sided Mann-Whitney U test). To determine if the decreased number of CNAs observed in WB1P-Myc

tumors was not simply a result of the shortened tumor latency, we generated *WapCre;Brca1$^{F/F}$;Trp53$^{F/F}$;Col1a1$^{invCAG\text{-}Met\text{-}IRES\text{-}Luc/+}$* (WB1P-Met) mice containing the *Met* oncogene, which we found frequently amplified in the WB1P tumors. Similar to WB1P-Myc females, WB1P-Met female mice developed multifocal mammary tumors with a short latency of 89 days ($n = 11$, Supplementary Figure S5.6A). All WB1P-Met tumors were classified as poorly differentiated ER/PR-negative ductal carcinomas and showed MET overexpression and active MET signaling (Supplementary Figure S5.6B-E). These data confirm the previously reported role of MET in the onset and progression of TNBC[29]. CNV-sequencing of WB1P-Met tumors ($n = 20$) showed an intermediate number of CNAs (Supplementary Figure S5.6F), which was lower than the WB1P tumors but significantly higher than the WB1P-Myc tumors ($p < 0.001$, one-sided Mann-Whitney U test). This demonstrates that the observed differences in CNA load are not merely a function of tumor latency but also of the driver gene. Moreover, the validation of MET as a potent driver in BRCA1-associated tumorigenesis underscores the potential of iterative analysis of CNAs in progressively complex mouse models as an approach for identifying putative cancer genes that promote tumorigenesis in specific genetic contexts.

5.4.5 Comparative oncogenomics identifies *MCL1* as a driver gene

Our RUBIC analyses showed that most of the CNAs identified in WB1P tumors were no longer present in WB1P-Myc tumors, suggesting an increased evolutionary pressure to acquire only specific driver mutations (Figure 5.3B). Interestingly, a small number of losses were retained, including the *Rb1*-associated loss on chromosome 14, further supporting *Rb1* as a collaborating driver in MYC-driven BRCA1-deficient mammary tumors. Focusing on novel events, we identified a strongly recurrent amplicon on chromosome 11 encompassing the *Col1a1* locus in which we introduced the *invCAG-Myc-IRES-Luc* cassette. The recurrent amplification of this locus suggests that WB1P-Myc tumors underwent a selection for increased MYC expression via amplification of the conditional *Myc* knock-in allele. Besides this, we also identified novel recurrent amplifications on chromosome 3 and chromosome 15, which were syntenic with human 1q and 22q loci, respectively, which are commonly amplified in breast cancer patients.

To identify additional driver genes in MYC-driven BRCA1-deficient TNBC, we used a comparative oncogenomics strategy to select candidate genes that are recurrently aberrated in both WB1P-Myc tumors and human BLBCs from TCGA. In this approach (outlined in Figure 5.3D), we first identified candidate drivers in both species individually using RUBIC. For the mouse tumors, we combined CNV-seq data of tumors from both the WB1P-Myc GEMM and the somatically engineered MYC-

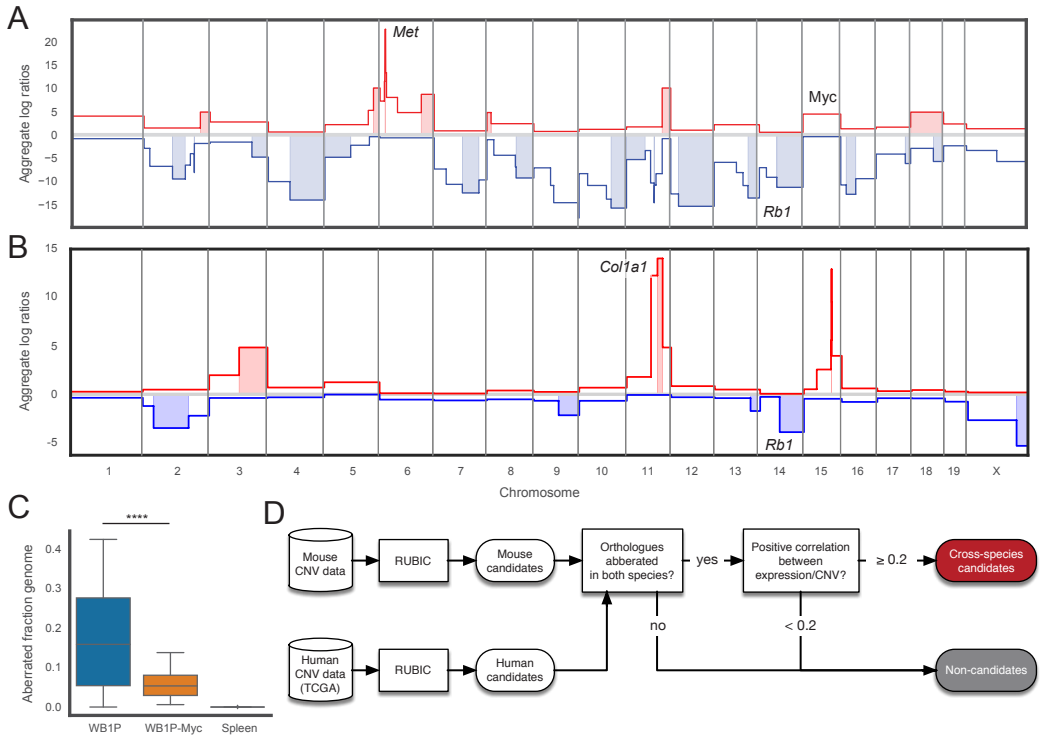


Fig. 5.3. Identification of candidate drivers in WB1P-Myc tumors using comparative oncogenomics. **(A and B)** Genome-wide RUBIC analysis of CNV profiles of WB1P tumors (A) and WB1P-Myc tumors (B). Significant amplifications and deletions are marked by light red and blue columns, respectively. **(C)** Genomic instability of WB1P and WB1P-Myc tumors. Scores for spleen samples from WB1P mice are shown as reference; $^{****}p < 0.0001$ (two-sided Mann-Whitney U test). 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. See Materials and Methods for more details. **(D)** Flowchart illustrating the comparative oncogenomics analysis pipeline used for the identification of additional cancer driver genes.

driven B1P models to increase our sample size, based on the observation that these tumors share the same distinctive CNA profile (Supplementary Figure S5.5B-C). Next, we mapped genes between species using mouse-human orthologs and took the intersection of both candidate lists. Finally, to prioritize genes that show differences in expression, we filtered the remaining candidates for genes with a positive Spearman correlation (> 0.2) between their expression and copy number status.

After applying this strategy, we focused on genes residing in the recurrent amplifications on mouse chromosomes 3 and 15, as these were the most striking new events in the WB1P-Myc model. The recurrent amplification on chromosome 11 containing the conditional *Myc* knock-in allele in the *Col1a1* locus was excluded from this analysis. While this did not identify any candidate genes in the peak on chromosome 15 (mainly due to a lack of orthologous, recurrently aberrated

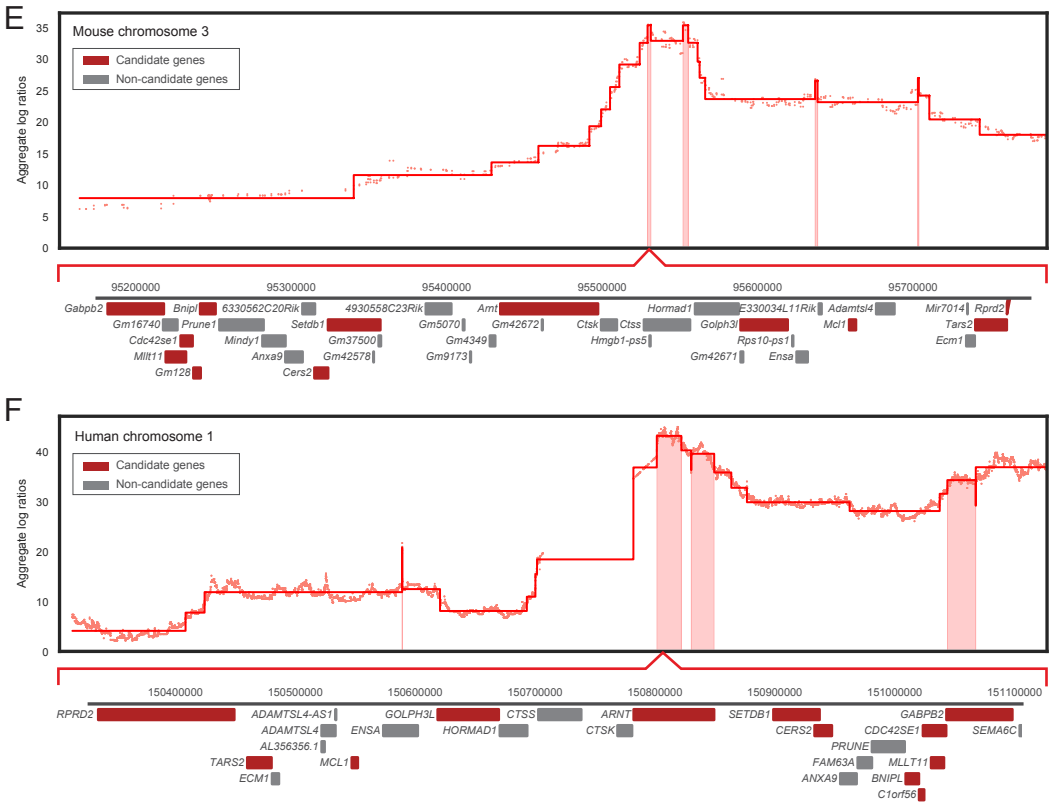


Fig. 5.3. **Continued. (E)** Chromosome 3 RUBIC analysis of the combined CNV profiles of the tumors from germline and somatic mouse models overexpressing *Myc* in the mammary gland. Significant amplifications are marked by light red columns. Genes residing in the minimal amplicon of chromosome 3 are shown. Cross-species candidate genes surviving filter criteria are colored in red. **(F)** Chromosome 1 RUBIC analysis of the CNV profiles of human TNBC. Significant amplifications are marked by light red columns. Orthologs of the genes shown in panel E are shown. Cross-species candidate genes surviving filter criteria are colored in red.

genes), it did identify a list of 12 candidate genes residing in the peaks on mouse chromosome 3 (Figure 5.3E) and human chromosome 1q (Figure 5.3F). To identify potential drivers in this list of candidates, we derived organoids from a WB1P-Myc mammary tumor using our recently established methodology[30]. We next performed a fitness screen in these WB1P-Myc organoids with a focused lentiviral shRNA library targeting candidate genes. This screen showed a marked depletion for shRNAs targeting *Mcl1* (Figure 5.4A), indicating that MCL1 expression is essential for growth of WB1P-Myc tumor cells. In line with this, WB1P-Myc tumors showed strongly elevated expression of MCL1 compared to WB1P tumors (Figure 5.4B).

To determine whether MCL1 cooperates with MYC in driving BRCA1-deficient TNBC, we generated a lentiviral vector encoding both *Cre* and *Mcl1* (Lenti-Mcl1P2ACre, Supplementary Figure S5.7A) and injected this lentivirus intraductally into B1P and B1P-Myc females ($n = 7$ and $n = 11$, respectively) to achieve simultaneous Cre-mediated recombination of the conditional alleles and overexpression of *Mcl1*

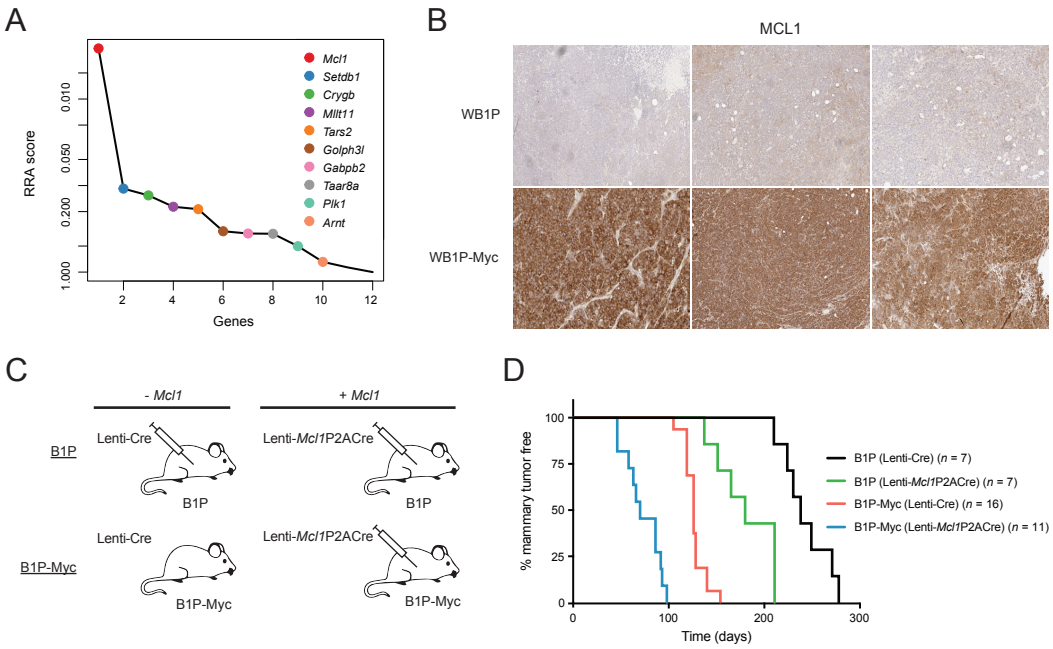


Fig. 5.4. Validation of MCL1 as a druggable driver in BRCA1-mutated TNBC. (A) MAGeCK software was used to compute RRA scores for all genes included in our focused shRNA library, showing depletion of *Mcl1* shRNAs in WB1P-Myc organoids. **(B)** Immunohistochemical detection of MCL1 in multiple independent WB1P and WB1P-Myc tumors. **(C)** Overview of the non-germline mouse models for mammary-specific *Mcl1* overexpression. Scale bar, 400 μm. **(D)** Kaplan-Meier curves showing mammary tumor-specific survival for the different models. B1P and B1P-Myc females injected with Lenti-Mcl1P2ACre showed a reduced mammary tumor-specific survival compared to B1P and B1P-Myc female mice injected with Lenti-Cre, respectively (180 days after injection vs 238 days after injection; $^{**}p < 0.01$ by Mantel-Cox test; 70 days after injection vs 126 days after injection; $^{****}p < 0.0001$ by Mantel-Cox test).

(Figure 5.4C). Co-expression of MCL1 and Cre in B1P and B1P-Myc mice resulted in a significant decrease in tumor latency compared to mice in which only Cre was delivered (180 vs 238 days and 70 vs 126 days, respectively; Figure 5.4D; Supplementary Figure S5.7B). MCL1 overexpression appeared to relieve pressure for chromosome 3 amplification in the resulting tumors (Supplementary Figure S5.7C-D). Conversely, MCL1 silencing in WB1P-Myc organoids resulted in *Myc* downregulation (Supplementary Figure S5.7E). Altogether, these analyses identify *Mcl1* as an important driver gene in the recurrent amplification on mouse chromosome 3 and demonstrate that MCL1 effectively collaborates with MYC in BRCA1-associated breast tumorigenesis.

5.4.6 MCL1-inhibition synergizes with PARP-inhibition in PDXs

Our focused shRNA library screen showed that knockdown of *Mcl1* is detrimental to WB1P-Myc organoids. To test whether WB1P-Myc tumors are also sensitive

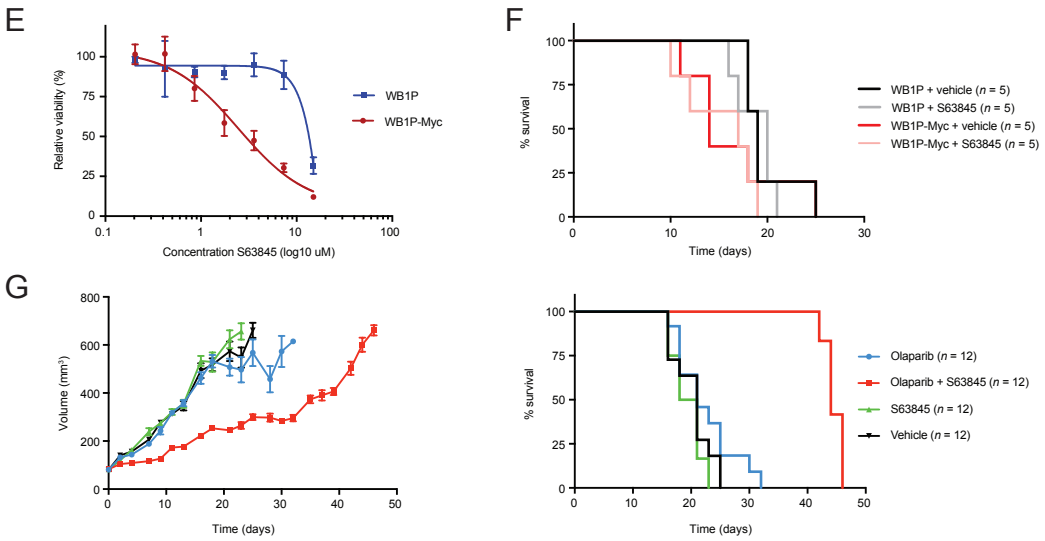


Fig. 5.4. **Continued. (E)** *In vitro* response of WB1P and WB1P-Myc organoids to MCL1 inhibitor S63845. Error bars represent standard error of the mean. **(F)** *In vivo* response of organoid-derived WB1P and WB1P-Myc tumors to S63845, as visualized by Kaplan-Meier curves. WB1P and WB1P-Myc organoid lines were transplanted in the fourth mammary fat pad of nude mice. When tumors had reached a size of 100 mm^3, mice were treated with 25 $mg\,kg^{-1}$ S63845 (i.v. once-weekly for 5 weeks) or vehicle. **(G)** Response of the BRCA1-mutated TNBC PDX-110 xenograft model to S63845 and the PARP inhibitor olaparib, as visualized by tumor volume curves (left) and Kaplan-Meier curves (right). Single-cell suspensions of PDX-110 were transplanted in the fourth mammary fat pad of *NOD-SCID-IL2R*$\gamma_c^{-/-}$ mice. When tumors had reached a size of 100 mm^3, mice were treated with 25 $mg\,kg^{-1}$ S63845 (i.v. once-weekly for 4 weeks), 50 $mg\,kg^{-1}$ olaparib (i.p. 5 days out of 7 for 4 weeks), both drugs or vehicle. Combination therapy with S63845 and olaparib prolonged survival compared to olaparib monotherapy ($^{****}p < 0.0001$ by Mantel-Cox test). Error bars represent standard error of the mean.

to pharmacological MCL1-inhibition, we tested the *in vitro* sensitivity of WB1P and WB1P-Myc organoids to the selective MCL1 inhibitor S63845[31], which was recently shown to display activity in patient-derived xenograft (PDX) models of breast cancer[32]. In contrast to other BH3 mimetics, S63845 binds to human MCL1 at sub-nanomolar concentrations whereas it does not display detectable binding to other anti-apoptotic family members[31]. Proliferation assays indicated that WB1P-Myc organoids were more sensitive to S63845 than WB1P organoids (Figure 5.4E). To examine the response of WB1P and WB1P-Myc tumors to MCL1-inhibition, we transplanted WB1P and WB1P-Myc organoids orthotopically into nude mice ($n = 10$ per line) and tested the response of the tumor outgrowths to S63845. However, in this setting we did not observe a differential sensitivity to MCL1-inhibition, as none of the tumors responded to S63845 at the tested dose (Figure 5.4F).

Based on the anti-apoptotic role of MCL1, we reasoned that MCL1-inhibition might be most effective when combined with a pro-apoptotic drug, as previously demonstrated in PDX models of HER2-amplified breast cancer and TNBC treated with

trastuzumab and docetaxel, respectively[32]. We therefore investigated whether S63845 could enhance the efficacy of the clinical PARP inhibitor (PARPi) olaparib, which is currently used for treatment of BRCA1-mutated breast cancer patients. As the olaparib-sensitivity of WB1P and WB1P-Myc tumors was too high to assess any synergistic effect of MCL1-inhibition (Supplementary Figure S5.2B; Supplementary Figure S5.8A-B), we turned to a PDX model of *BRCA1*-mutated TNBC (PDX-110), which expresses relatively high levels of MYC and MCL1 and shows limited sensitivity to PARPi[32]. To assess the effect of combined inhibition of PARP and MCL1 in this model, PDX tumor cells were orthotopically injected into *NOD-SCID-IL2R*$\gamma_c^{-/-}$ mice ($n = 48$), which were randomly allocated to vehicle-, single- or double-treatment arms once tumors reached a volume of 100 mm. Remarkably, while treatment with S63845 or olaparib alone did not elicit a clinical response, tumor growth was considerably inhibited upon treatment with both drugs and tumors relapsed only when treatment was stopped after 4 weeks (Figure 5.4G).

5.5 Discussion

In this work, we have used both germline and somatic engineering approaches to rapidly test candidate cancer drivers in the WB1P mouse model of BRCA1-deficient TNBC. Using this approach, we validated *MYC*, *MET*, *PTEN* and *RB1* as *bona fide* drivers of BRCA1-associated tumorigenesis and showed that MYC overexpression dramatically changes the mutational landscape of the resulting tumors. Finally, by applying a comparative oncogenomics strategy to uncover additional culprits of tumorigenesis, we identified *MCL1* as a druggable cancer driver that collaborates with MYC in BRCA1-deficient TNBC.

An important challenge of modeling cancer in mice, is that it requires technology that allows rapid introduction of new driver mutations to quickly create a variety of compound-mutant mouse models containing different combinations of candidate genes. Such experiments are difficult to perform using germline engineering approaches, which generally involve extensive cross-breeding of single-mutant mice to produce animals carrying the desired combination of mutations. Here, we have shown that somatic engineering using lentiviral vectors for overexpression of cDNAs and CRISPR-mediated *in situ* gene editing provides an effective alternative for rapid generation of novel mouse models of BRCA1-deficient TNBC. The limitations of cDNA-based overexpression systems – which may not fully recapitulate the desired expression levels of candidate genes – might be alleviated by implementing novel technologies for CRISPR-mediated transcriptional control (CRISPRi/CRISPRa) and base-editing of endogenous genes. We therefore expect that our somatic engineer-

ing methodology can ultimately be used to generate refined breast cancer models containing a wide variety of fine-tuned (epi)genetic permutations.

One of the key advantages of this type of iterative mouse modeling, is that it highlights the profound effect that additional driver genes can have on the mutational landscape of a baseline tumor model. This indicates that the evolutionary pressure that tumor cells experience depends strongly on the combination of driver mutations, which push cells down a specific path to acquire additional aberrations that most effectively collaborate with the pre-existing events. This notion has two important implications. First, it means that it is crucial to study driver genes in the appropriate genetic contexts as observed in human tumors, as they may have very different effects in different backgrounds. Second, predominant changes in the mutational landscape likely indicate that the pre-existing driver(s) push tumors down a relatively restricted evolutionary path, which might be exploited therapeutically.

A clear example is provided by our study, showing that MYC-overexpression pushes the evolution of BRCA1-deficient TNBC towards amplification of a druggable driver, MCL1. Although MYC is a potent inducer of cell proliferation, supraphysiologic overexpression of *MYC* also has pro-apoptotic effects and is generally not tolerated in non-transformed cells[33]. This suggests that tumor cells need to acquire additional alterations in other, collaborating cancer driver genes to counteract MYC-induced apoptosis. MCL1 belongs to the Bcl-2 family and is involved in the inhibition of apoptosis[34]. While it cannot be excluded that MYC overexpression reshapes evolution of BRCA1-deficient TNBCs via negative selection of tumor cell clones with high levels of CNAs, amplifications of *MCL1* might be particularly selected for as they may counteract the pro-apoptotic effects of MYC overexpression. *MCL1* amplifications have been identified in a variety of tumor types, including breast cancer[35]), where they correlate with poor survival[36,37]. Although the commonly amplified chromosome 1q region (where *MCL1* resides) might harbor additional driver genes, MCL1 is the main pro-survival protein upregulated in TNBC[38], and amplification of *MCL1* has been implicated in resistance to multiple therapies used in patients with TNBC, where it is often co-amplified with *MYC*[39,40]. We found amplification of *MCL1* in 15% of the *BRCA1*-mutated TNBCs we analyzed (12/80 cases), and two-thirds of these cases showed co-amplification of *MYC* (Supplementary Figure S5.8C). MCL1 and MYC were also shown to cooperate in mouse models of leukemia and non-small cell lung cancer (NSCLC), and co-expression of these two factors correlates with poor NSCLC patient survival[41,42]. This suggests that MCL1 inhibition might be particularly effective against *MYC*-overexpressing tumors by exposing them to MYC-induced apoptosis. MCL1 has only recently been recognized as an important therapeutic target, and currently several MCL1 inhibitors are being tested in pre-clinical trials, where they are showing promising activity, especially in combination therapies[32,43]. Five phase-I clinical trials are currently ongoing for testing MCL1 in-

hibitors in patients with hematopoietic malignancies (NCT02675452; NCT02992483; NCT02979366; NCT03672695; NCT03465540). Our study demonstrates that MCL1 inhibition considerably enhances response of *BRCA1*-mutated TNBC to the clinical PARPi olaparib and suggests that this combination should be prioritized for clinical evaluation, especially in *BRCA1*-mutated cancer patients with poor response to PARPi monotherapy.

Another example of identification of druggable drivers in mouse models of BRCA1-deficient TNBC was recently provided by Liu *et al.*[44], who analyzed transcriptional and CNA profiles of mammary tumors from our previously published KB1P and *K14Cre;Trp53*$^{F/F}$ (KP) models[21]. This analysis yielded a spectrum of somatic genetic alterations putatively driving tumor evolution, including gene-fusions and chromosomal amplifications and deletions (including recurrent amplification of *Met* and *Myc* and deletion of *Rb1*). Interestingly, even though KB1P and KP tumors were following diverse evolutionary trajectories, most tumors displayed enhanced activation of MAPK and/or PI3K signaling and could be treated with inhibitors specific for the aberrated drug target.

In summary, we applied novel germline and somatic technologies to functionally validate the role of candidate drivers in BRCA1-deficient TNBC *in vivo* at unprecedented speed. Our integrate approach revealed a profound effect of MYC overexpression on tumor evolution and identified MCL1 as a critical and druggable dependency in BRCA1-deficient TNBC with high expression of MYC. Combined inhibition of MCL1 and PARP might benefit a subset of BRCA1-mutated TNBC patients and warrants further investigation.

Acknowledgments We thank Sjors Kas, Eva Schut, Bastiaan Evers, Ben Morris, Renske de Korte-Grimmerink, Natalie Proost and Rebecca Theeuwsen for providing reagents, technical suggestions, and/or help with the experiments. We are grateful for excellent support from the NKI animal facility, RHPC computing facility, flow cytometry facility, animal pathology facility, transgenic facility, preclinical intervention unit, core facility molecular pathology and biobanking (CFMPB), and genomics core facility. PDX-110 was derived from a primary breast tumor provided by the Victorian Cancer Biobank (supported by the Victorian Government, Australia). This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative (e-infra160136). Financial support was provided by the Oncode Institute, the Netherlands Organization for Scientific Research (NWO: Cancer Genomics Netherlands (CGCNL), Cancer Systems Biology Center (CSBC), Netherlands Genomics Initiative (NGI) Zenith 93512009 (J.J.), VICI 91814643 (J.J.)), the European Research Council (ERC Synergy project CombatCancer), a National Roadmap grant for Large-Scale Research Facilities from NWO (J.J.) and National Health and

Medical Research Council (Australia) grants 1113133 (J.E.V. and G.J.L.), 1078730 (G.J.L.) and 1102742 (J.E.V.).

Author information These authors contributed equally: Stefano Annunziato, Julian R. de Ruiter, Linda Henneman, Chiara S. Brambillasca. These authors jointly supervised this work: Lodewyk F. A. Wessels, Jos Jonkers.

Contributions Conceived the study: S.A., J.R.d.R., L.H., C.S.B., L.F.A.W., J.J.. Designed and supervised the experiments: S.A., J.R.d.R., L.H., C.S.B., F.V., M.H.v.M., I.J.H., M.v.d.V., J.E.V., G.J.L., L.F.A.W., J.J.. Performed the experiments: S.A., J.R.d.R., L.H., C.S.B., C.L., F.V., F.F., A.D., E.v.d.B., B.S., B.v.G., R.d.B., M.H.v.M.. Analyzed and interpreted data: S.A., J.R.d.R., L.H., C.S.B., F.V., J.E.V., G.J.L., L.F.A.W., J.J.. Wrote the paper: S.A., J.R.d.R., J.J..

5.6 Materials and Methods

Lentiviral vectors The Lenti-Cre vector (pBOB-CAG-iCRE-SD, Addgene plasmid #12336) was a kind gift of Lorenzo Bombardelli. Lenti-MycP2ACre and Lenti-Mcl1-P2ACre were cloned as follows. AgeI and SalI were used to remove GFP-T2A-puro from the SIN.LV.SF-GFP-T2A-puro[45]. P2ACre was synthesized with AgeI-SalI overhangs and inserted as AgeI-SalI fragment into the SIN.LV.SF-GFP-T2A-puro backbone, resulting in SIN.LV.SF-P2ACre. *Myc* and *Mcl1* murine cDNAs were isolated with BamHI-AgeI overhangs using standard PCR from cDNA clones (Clone 8861953, Source BioScience; Clone 3491015, Dharmacon) and inserted as BamHI-AgeI fragments into the SIN.LV.SF-P2ACre vector. The Lenti-sg*Rb1*-Myc, Lenti-sg*Pten*-Myc and Lenti-sgNT-Myc vectors were cloned as follows. *Myc* cDNA was isolated with XbaI-XhoI overhangs using standard PCR from the Lenti-MycP2ACre vector, and inserted as XbaI-XhoI fragment into pGIN, a lentiviral vector for sgRNA overexpression[46]. The non-targeting sgRNA (TGATTGGGGGTCGTTCGCCA) and sgRNAs targeting mouse *Rb1* exon 2 (TCTTACCAGGATTCCATCCA) and mouse *Pten* exon 7 (CCTCAGCCATTGCCTGTGTG) were cloned as described[47]. All vectors were validated by Sanger sequencing. Concentrated stocks of VSV-G pseudotyped lentivirus were produced by transient co-transfection of four plasmids in 293T as previously described[48]. Lentiviral titers were determined using the qPCR lentivirus titration kit from Abm (LV900).

Cell culture 293T cells for lentiviral production and Cre-reporter 293T cells containing a lox-STOP-lox-GFP cassette were cultured in Iscove's medium (Invitrogen Life Technologies) containing 10 % FBS, 100 IU ml^{-1} penicillin, and 100 μg ml^{-1} streptomycin. Transductions were performed by adding diluted viral supernatant to

the cells in the presence of $8\,\mu g\,ml^{-1}$polybrene (Sigma). Cells were transduced for 24h, after which medium was refreshed. Harvesting of cells for flow cytometry or immunoblotting was performed 5 days after transduction.

Flow cytometry Cells were collected 5 days after transduction, washed in PBS, fixed in Fixation Buffer (BD Biosciences) and permeabilized with Perm Buffer III (BD Biosciences). They were then stained using the primary rabbit antibody anti-Myc (1:1000, Abcam ab32072) or anti-Mcl1 (1:1000, Cell Signaling 94296S) for 30 minutes at 4 degrees, washed in PBS and incubated for 15 minutes with an AlexaFluor647-conjugated secondary anti-rabbit antibody (1:1000, Thermofisher). Stained cells were analyzed using a Becton Dickinson LSR FORTESSA. GFP and AlexaFluor647 expression of viable cells was measured using a 488 nm and 640 nm laser for excitation, respectively. Data analysis was performed using FlowJo software version 7.6.5.

PCRs and TIDE analyses Amplification of *Rb1* exon 2 and *Pten* exon 7 was performed with specific primers spanning the target sites (FW_Rb1: TCACCATGCTAGCAGCTCTTC; RV_Rb1: AGCCAGTTCAATGGTTGTGGG; FW_Pten: TGTATTTAACCACACAGATCCTCA; RV_Pten: AACAAACTAAGGGTCGGGGC) and $1\,\mu g$ DNA template using the Q5 high-fidelity PCR kit from NEB. Amplicons were run on 1% agarose gel and gel-purified using the Isolate II PCR and Gel kit (Bioline). PCR products were Sanger-sequenced using the FW primer and CRISPR/Cas9-induced editing efficacy was predicted and quantified as described[49]. Untransduced cells were taken along as a control in each gRNA amplification.

Immunoblotting Protein lysates were made using lysis buffer (20 mmol Tris 8.0 pH , 300 mmol NaCl, 2 % NP40, 20 % glycerol, 10 mmol EDTA) complemented with protease inhibitors (Roche) and quantified using the BCA Protein Assay Kit (Pierce). Protein lysates were loaded onto a 3 % to 8 % Tris-Acetate gradient Gel (Invitrogen) and transferred overnight onto PVDF membrane (Millipore, methanol pre-wetted) in transfer buffer (38 mmol glycine, 5 mmol TRIS and 0.01 % SDS in PBS-T (0.5 % Tween-20). Membranes were blocked in 5 % ELK in TBS-T after which they were stained for four hours at room temperature using the primary antibody anti-Met (1:1000, Cell Signaling 4560S), anti-phosphoMet (1:1000, Cell Signaling 3077S), anti-Myc (1:1000, Abcam ab32072) or anti-Mcl1 (1:1000, Cell Signaling 94296S). Membranes were washed three times with 1 % milk in PBS-T and incubated for 1 hr with an HRP-conjugated secondary antibody (1:2000, DAKO). Stained membranes were washed three times in 1 % milk in PBS-T and developed using SuperSignal ECL (Pierce).

Organoid culture WB1P and WB1P-Myc mammary tumor organoids were isolated and cultured as described[30]. For cell viability assays, organoids were seeded (100,000 cells per well) in 40 μl complete mouse media/BME mixture on 24-well suspension plates and cultured for 5 days in the presence of S63845 (ApEXBiO). Cell viability was assessed using the resazurin-based Cell Titer Blue assay following manufacturer's protocol (Promega). Cell viability experiments were performed 3 times in triplicate and data were analyzed with GraphPad Prism statistical software using nonlinear regression and extra sum-of-squares F-test. For the focused shRNA library screen in WB1P-Myc organoids[30], a small library of shRNA targeting candidate genes was built from the Mission shRNA collection (mouse TRC v1.0 collection) by pooling shRNAs targeting candidate genes (*Mcl1*, *Gabpb2*, *Arnt*, *Setdb1*, *Tars*, *Golph3l*, *Lass2* and *Mllt11*) and control genes (*Plk1*, *Nlrp5*, *Crygb* and *Taar8a*). Organoids were transduced at MOI 0.3 and analyzed for shRNA representation at day 0, 7 and 14. MAGeCK software was used to compute RRA scores for all genes to identify relative shRNA depletion.

Meta-analysis of four human breast cancer datasets Curated copy number and mutation data for the METABRIC, TCGA and MSK-impact datasets were downloaded from cBioPortal*, after which the downloaded mutation data was filtered for deleterious mutations (Missense_Mutation, Nonsense_mutation, Frame_Shift_Del and Frame_Shift_Ins). Similarly, copy number data were filtered for high-level amplifications (amp) or homozygous deletions (homdel). Besides this, the MSK-impact dataset was filtered to include only breast cancers. Mutation and copy number data for the BASIS dataset were obtained from Supplementary Tables 4, 14 and 20 accompanying reference[50] and filtered using similar criteria as the other datasets. The resulting datasets were merged and, where possible, annotated with the ER, PR and HER2 status of the corresponding samples. To select for samples with deleterious missense mutations in *BRCA1*, *BRCA1* missense mutations were annotated for their expected pathogenicity using the Breast Cancer Information Core (BIC) database† and Align-GVGD‡. We only selected samples with *BRCA1* missense mutations that were considered to be pathogenic (annotated as clinically important by BIC or Align-GVGD assigned class C15, C25 and C65). The final dataset was visualized using a custom script, focusing on cancer-associated genes, as defined by cBioPortal.

RNA Sequencing Illumina TruSeq mRNA libraries were generated and sequenced with 50-65 base single reads on a HiSeq 2500 using V4 chemistry (Illumina Inc., San Diego) as previously described[51]. The resulting reads were trimmed using Cutadapt (version 1.15) to remove any remaining adapter sequences and to filter reads shorter than 20 bp after trimming to ensure good mappability. The trimmed reads were

*http://cbioportal.org, 13/10/2017
†http://research.nhgri.nih.gov/bic
‡http://agvgd.iarc.fr

aligned to the GRCm38 reference genome using STAR (version 2.5.3a). QC statistics from Fastqc (version 0.11.5) and the above-mentioned tools were collected and summarized using Multiqc (version 1.1). Gene expression counts were generated by featureCounts (version 1.5.2) using gene definitions from Ensembl GRCm38 version 89. Normalized expression values were obtained by correcting for differences in sequencing depth between samples using DESeq's median-of-ratios approach and then log-transforming the normalized counts.

Generation of CNA profiles and data analysis Sequencing was performed using the Illumina HiSeq 2500 with V4 chemistry (Illumina Inc., San Diego) as previously described[51]. The resulting reads were trimmed using Cutadapt (version 1.15) to remove any remaining adapter sequences and trim reads longer to a length of 50 bp for QDNAseq. Additionally, reads shorter than 20 bp after trimming were removed to ensure good mappability. The trimmed reads were aligned to the the GRCm38 reference genome using BWA aln (version 0.7.15). QC statistics from Fastqc (version 0.11.5) and the above-mentioned tools were collected and summarized using Multiqc (version 1.1). The resulting alignments were analyzed using QDNAseq (version 1.14.0) using the mm10 reference genome (with a 50K bin size, 50 bp read lengths and default settings for other parameters) to generate copy number logratios, segmented profiles and calls. The segmented profiles were analyzed using RUBIC[6] (version 1.0.3) to identify recurrent CNAs regions (focal threshold $= 1e+08$, min probes $= 4$ and FDR $= 0.25$). Genes with copy number values were identified using a custom script, in which missing values were imputed from surrounding bins (with window size $= 11$, requiring at least 5 non-missing values). Copy number instability was scored by calculating the fraction of bins with logratio values above/below a threshold of ± 0.5 in the segmented copy number data.

CNA analysis of BLBCs from TGCA Segmented copy number data for the TCGA breast cancer samples were downloaded from firebrowse (version 2016_01_28). These data were matched to subtype annotations from TCGA and filtered for BLBC samples. This BLBC dataset was analyzed using RUBIC (version 1.0.3) to identify recurrent CNAs (with focal threshold $= 1e + 07$, min probes $= 260000$ and FDR $= 0.25$).

Comparative oncogenomics analysis Candidate genes were initially selected by identifying mouse genes that were recurrently amplified/deleted in the RUBIC analysis of the combined germline/somatic Myc-driven mouse tumors. These candidates were subsequently annotated for their orthologous human genes using Ensembl Biomart (GRCH37) and filtered for candidates whose orthologues were also recurrently aberrated in the RUBIC analysis of the human BLBCs. To filter for correlation with expression, we calculated the Spearman correlation between copy number calls

and gene expression values of the remaining candidate genes and selected genes with an absolute correlation > 0.2, resulting in a list of cross-species candidates.

Mouse studies *Myc* murine cDNA was obtained from a cDNA clone (Clone 8861953, Source BioScience), sequence-verified and inserted as FseI-PmeI fragment into the *Frt-invCag-IRES-Luc* shuttle vector[23], resulting in *Frt-invCag-Myc-IRES-Luc*. *Frt-invCag-Met-IRES-Luc* and *Frt-invCag-Cas9-IRES-Luc* were described were described by Henneman *et al.*[51] and Annunziato *et al.*[52]. Flp-mediated knockin of the shuttle vectors in the *WapCre;Brca1$^{F/F}$;Trp53$^{F/F}$;Col1a1-frt* GEMM-ESC was performed as described[23]. Chimeric animals were crossed with WB1P or B1P mice to generate the experimental cohorts. *WapCre*, *Brca1$^{F/F}$*, *Trp53$^{F/F}$* and knockin alleles were detected using PCR as described[21,23,53]. *In vivo* bioluminescence imaging was performed as described[51] using a cooled CCD system (Xenogen Corp., CA, USA) coupled to Living Image acquisition and analysis software (Xenogen). Intraductal injections were performed as described[24,52]. Lentiviral titers ranging from 2×10^8 TU/ml to 20×10^8 TU/ml were used.

Orthotopic transplantation of WB1P and WB1P-Myc mammary tumors or organoids was performed by implanting small tumor fragments or cells into the fourth right mammary fat pad of nude mice as previously described[30]. Treatment was initiated when tumors reached a size of ~100 mm (formula for tumor volume: $0.5 * \text{length} * \text{width}^2$). Cisplatin (6 mg kg^{-1} i.v.) was administered at day 0 and 14. Olaparib (100 mg kg^{-1} i.p.) and AZD2461 (100 mg kg^{-1} per os) were administered daily for 28 consecutive days. S63845 (25 mg kg^{-1} i.v.) was administered once-weekly for 5 weeks[32]. For experiments with PDX-110[32], thawed single cell suspensions of the tumor were transplanted orthotopically into the fourth right mammary fat pad of *NOD-SCID-IL2R$\gamma_c^{-/-}$* mice. Treatment was initiated when tumors reached a size of ~100 mm. Olaparib (50 mg kg^{-1} i.p.) was administered 5 days out of 7 for 4 weeks. S63845 (25 mg kg^{-1} i.v.) was administered once-weekly for 4 weeks. Vehicle was DMSO/10% (2-hydroxypropyl-b-cyclodextrin) for olaparib and 20% (2-hydroxypropyl-b-cyclodextrin)/HCl for S63845.

Animal experiments were approved by the Animal Ethics Committees of the Netherlands Cancer Institute and the Walter and Eliza Hall Institute of Medical Research. Mice were bred and maintained in accordance with institutional, national and European guidelines for Animal Care and Use.

Histology and immunohistochemistry Tissues were formalin-fixed overnight and paraffin-embedded by routine procedures. Haematoxylin and eosin staining was performed as described[54]. Immunohistochemical stainings were processed as previously described[51,54]. For MYC and MCL1 immunohistochemistry, primary rabbit antibody

anti-Myc (1:1000, Abcam ab32072) or anti-Mcl1 (1:1000, Cell Signaling 94296S) were used. All slides were digitally processed using the Aperio ScanScope (Aperio, Vista, CA, USA) and captured using ImageScope software version 12.0.0 (Aperio).

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability The analysis pipelines for the RNA-Seq and CNV-Seq data were implemented using Snakemake (version 4.3.1) and are freely available on GitHub[§,¶]. Additional scripts are available from the corresponding author upon reasonable request.

[§]https://github.com/jrderuiter/snakemake-rnaseq
[¶]https://github.com/jrderuiter/snakemake-cnvseq

5.7 Supplementary Material

5.7.1 Supplementary Figures

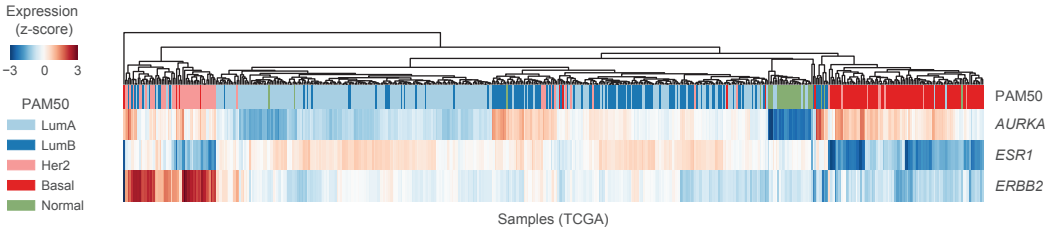


Fig. S5.1. Gene expression analysis of human breast cancer samples. Unsupervised clustering (Euclidean distance, average linkage) of the human breast cancer samples from TCGA, using a three-gene signature that distinguishes the PAM50 subtypes[22]. LumA, luminal A; LumB, luminal B.

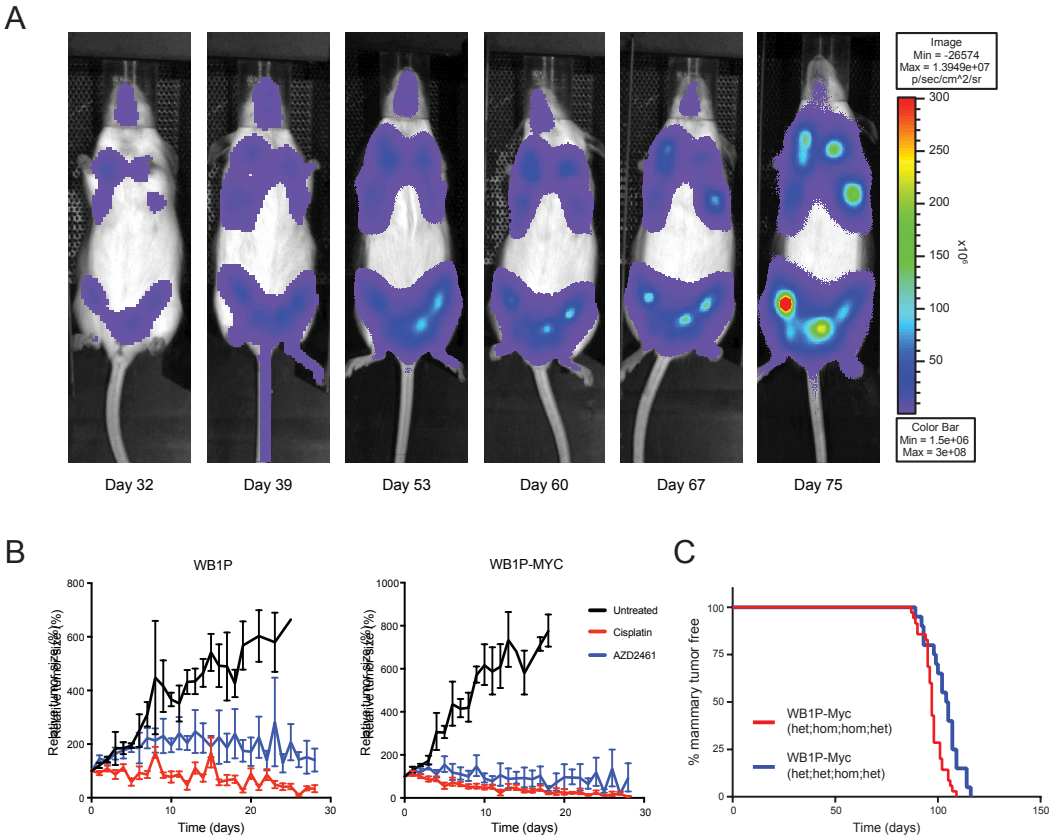


Fig. S5.2. **Characterization of the WB1P-Myc mouse model.** **(A)** Longitudinal *in vivo* bioluminescence imaging of luciferase expression in a WB1P-Myc female, showing signal build-up over time. **(B)** Response of WB1P and WB1P-Myc tumors to cisplatin and PARP inhibitors, as visualized by tumor volume curves. Small fragments of WB1P and WB1P-Myc tumors were transplanted in the fourth mammary fat pad of nude mice. When tumors had reached a size of 100 mm, mice were treated with 6 mg kg^{-1} cisplatin (administered i.v. on day 0 and day 14), 100 mg kg^{-1} AZD2461 (administered daily by per os, for 28 consecutive days) or vehicle. Error bars represent standard error of the mean. **(C)** Kaplan-Meier curves showing mammary tumor-specific survival for the different genotypes. *WapCre;Brca1*$^{F/F}$*;Trp53*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Myc\text{-}IRES\text{-}Luc/+}$ (WB1P-Myc) females showed a reduced mammary tumor-specific survival compared to *WapCre;Brca1*$^{F/+}$*;Trp53*$^{F/F}$*;Col1a1*$^{invCAG\text{-}Myc\text{-}IRES\text{-}Luc/+}$ littermates (97 days vs 105 days; $^{***}p < 0.001$ by Mantel-Cox test).

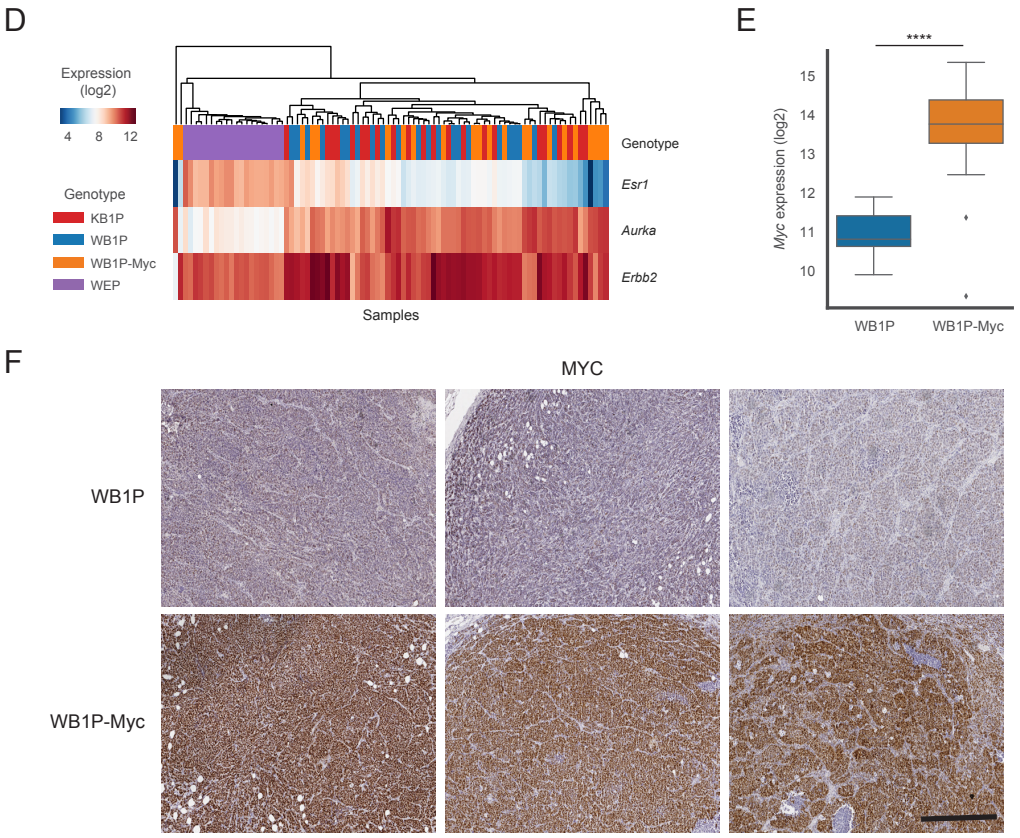


Fig. S5.2. **Continued.** **(D)** Unsupervised clustering (Euclidean distance, average linkage) of WB1P and WB1P-Myc tumors with tumors derived from published mouse models of luminal (*WapCre;Cdh1*$^{F/F}$*;Pten*$^{F/F}$, WEP) and basal-like (*K14Cre;Brca1*$^{F/F}$*;Trp53*$^{F/F}$, KB1P) breast cancer, using a three-gene signature that distinguishes the PAM50 subtypes[22]. **(E)** *Myc* expression levels in WB1P and WB1P-Myc tumors; $^{****}p < 0.0001$ (two-sided Mann-Whitney U test). 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. **(F)** Immunohistochemical detection of MYC in multiple independent WB1P and WB1P-Myc tumors. Scale bar, 400 μm.

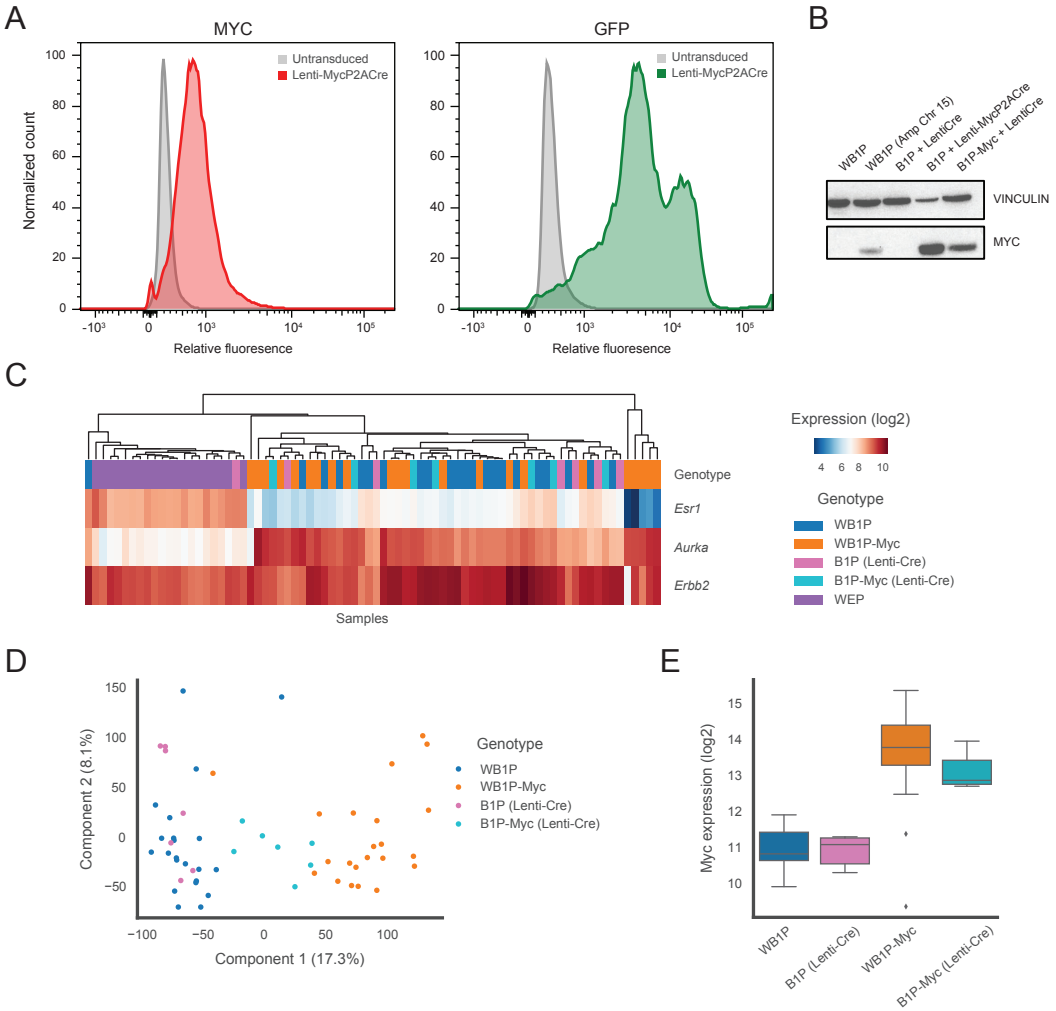


Fig. S5.3. **Non-germline models with Myc overexpression.** **(A)** *In vitro* validation of Lenti-MycP2ACre in Cre-reporter cells (containing a lox-STOP-lox-GFP cassette) 5 days after transduction. Expression of MYC as visualized by FACS using an anti-MYC antibody and FACS analysis of Cre-recombined GFP-positive cells are shown. **(B)** Expression of MYC in independent tumors as visualized by immunoblotting using anti-MYC antibody. Sample order: WB1P tumor without chromosome 15 amplification; WB1P tumor with chromosome 15 amplification; tumor from B1P mouse injected with Lenti-Cre; tumor from B1P mouse injected with Lenti-MycP2ACre; tumor from B1P-Myc mouse injected with Lenti-Cre. **(C)** Unsupervised clustering (Euclidean distance, average linkage) of the tumors from germline and somatic models using the three-gene PAM50 signature[22], showing that tumors from the somatic models retain the basal expression profiles observed in tumors from the germline models. **(D)** PCA plot comparing WB1P and WB1P-Myc tumors to tumors from the non-germline models (Lenti-Cre injected B1P and B1P-Myc mice) using global gene expression profiles. **(E)** *Myc* expression levels in tumors from germline and non-germline models. 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.

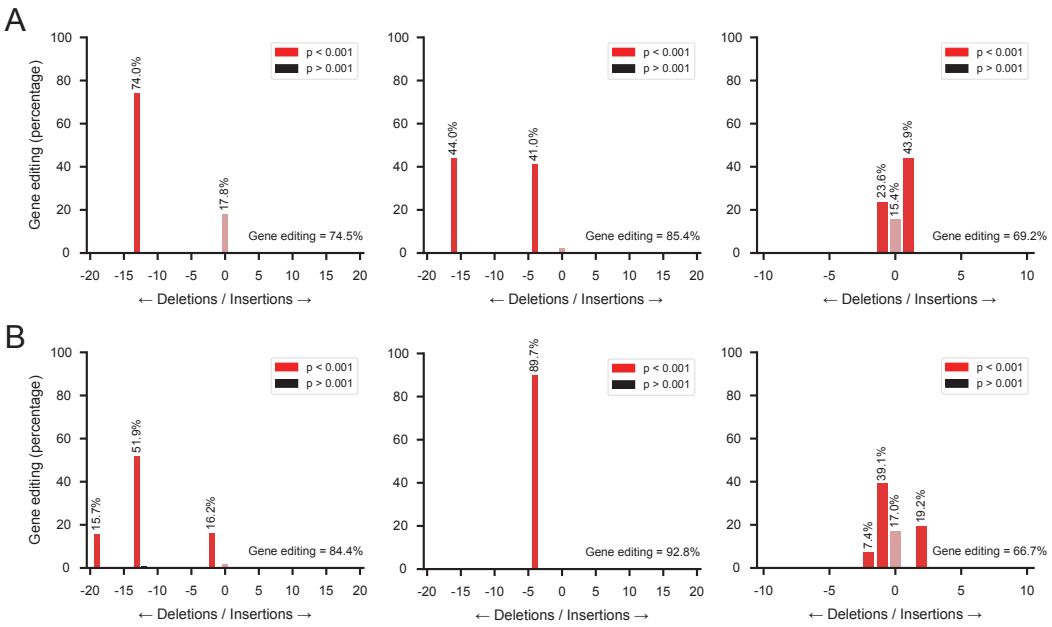


Fig. S5.4. CRISPR-mediated somatic gene disruption of *Rb1* and *Pten*. TIDE analyses showing the spectrum of insertions/deletions (indels) of the targeted *Rb1* **(A)** and *Pten* **(B)** alleles in multiple independent tumors from WB1P-Cas9 mice injected with Lenti-sg*Rb1*-Myc and Lenti-sg*Pten*-Myc, respectively.

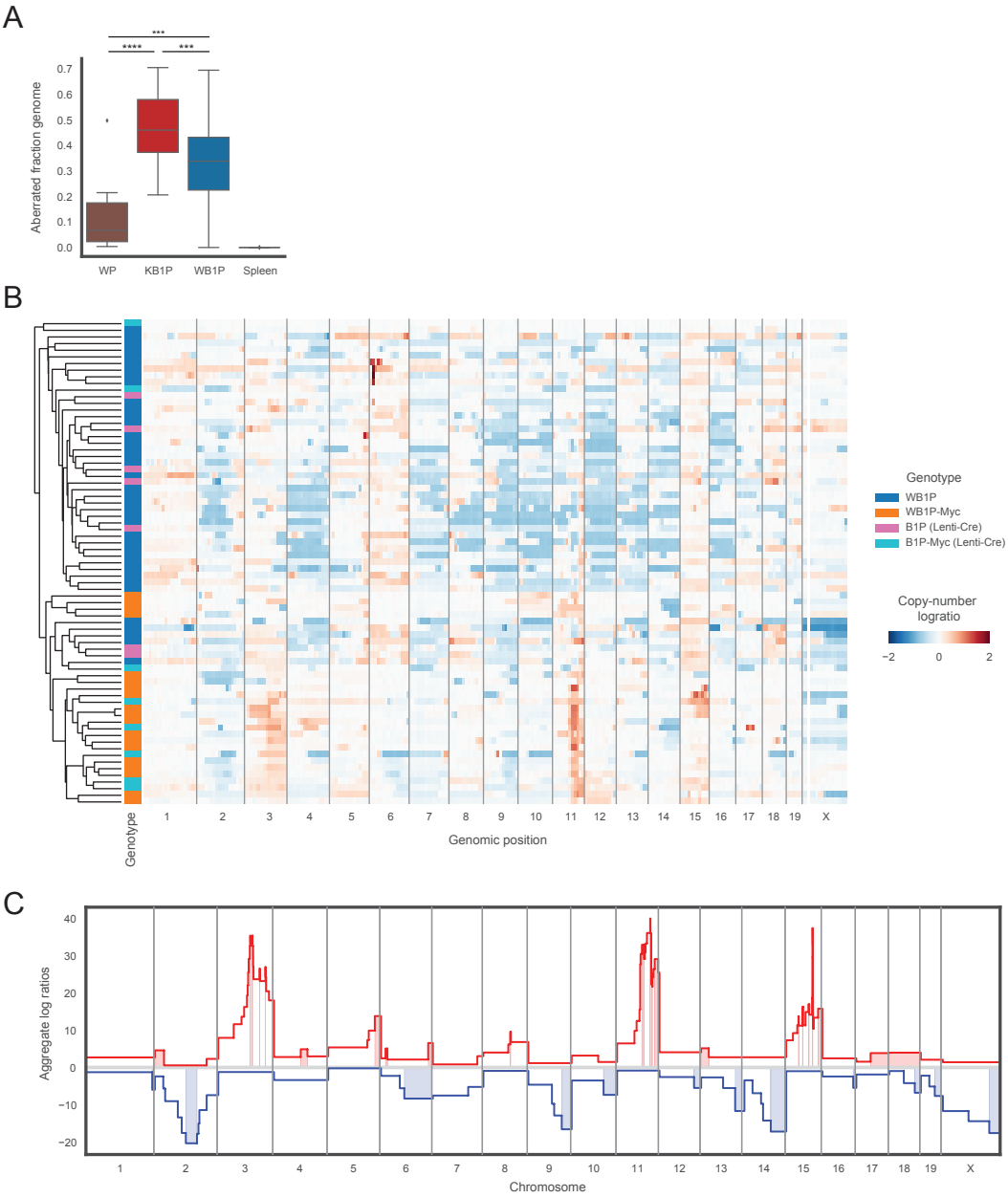


Fig. S5.5. **Genomic instability of WB1P and WB1P-Myc tumors. (A)** Genomic instability scores of WP, WB1P and KB1P tumors; $^{***}p < 0.001$ and $^{****}p < 0.0001$ (two-sided Mann-Whitney U test). Scores for spleen samples from WB1P mice are shown as reference. Boxes extend from the third (Q3) to the 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. **(B)** Unsupervised clustering (correlation distance, average linkage) of the tumors from somatic and germline models based on their copy number profiles. Tumors from the somatic models mainly cluster together with their germline counterparts, demonstrating that these tumors have similar patterns of copy number aberrations. **(C)** Overview of the recurrently aberrated regions identified by RUBIC in the combined set of MYC-driven tumors from the germline and somatic models.

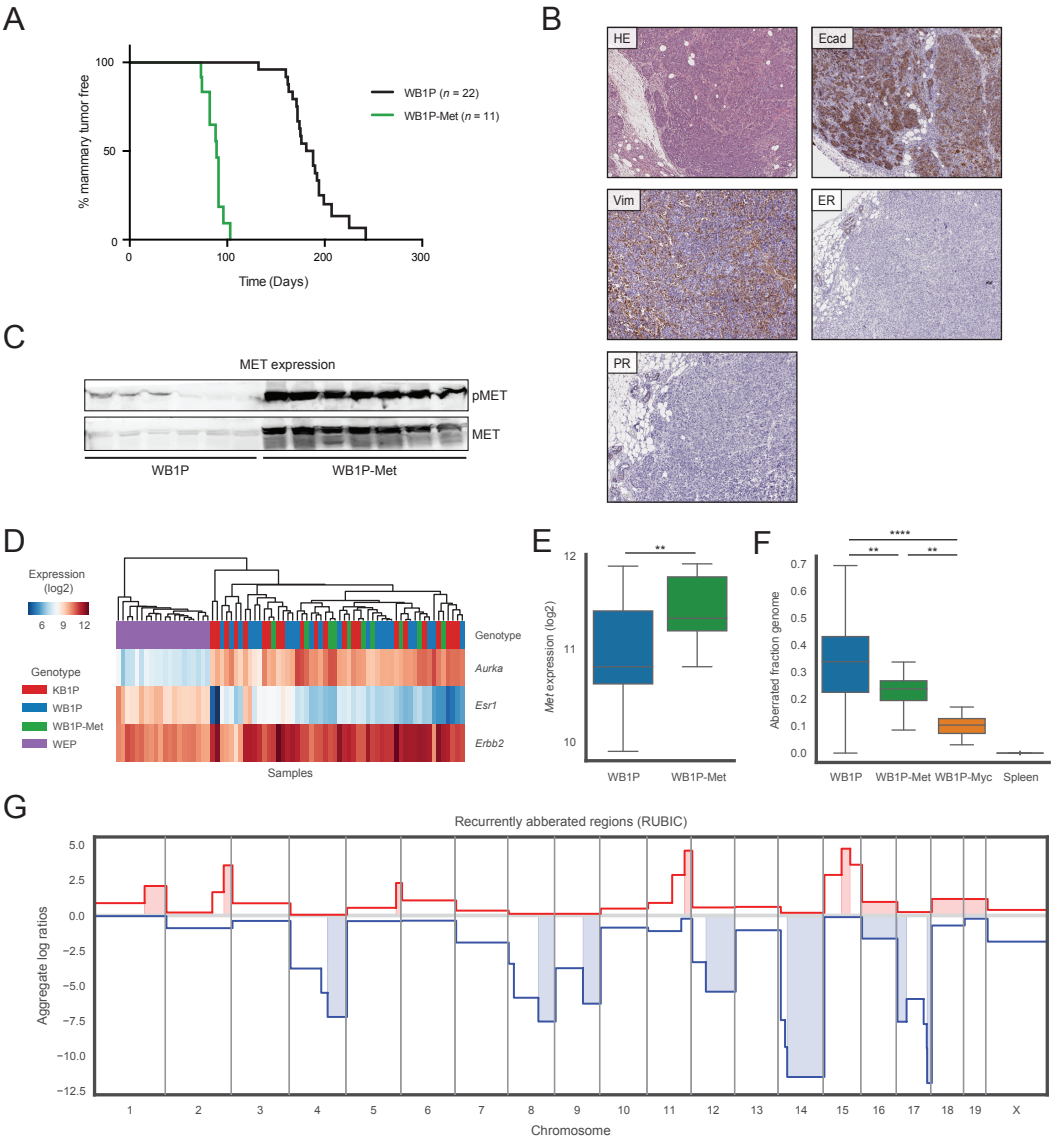


Fig. S5.6. Characterization of the WB1P-Met mouse model. **(A)** Kaplan-Meier curves showing mammary tumor-specific survival for the different genotypes. *WapCre;Brca1$^{F/F}$;Trp53$^{F/F}$;Col1a1$^{invCAG\text{-}Met\text{-}IRES\text{-}Luc/+}$* (WB1P-Met) females showed a reduced mammary tumor-specific survival compared to WB1P littermates (89 days vs 188 days; $^{****}p < 0.0001$ by Mantel-Cox test). **(B)** Representative hematoxylin and eosin (HE) staining and immunohistochemical detection of E-cadherin, vimentin, ER and PR in WB1P-Met tumors. **(C)** Expression and activity of MET in independent tumors from WB1P and WB1P-Met mice, as visualized by immunoblotting using anti-MET and anti-phosphoMET antibodies. **(D)** Unsupervised clustering (Euclidean distance, average linkage) of WB1P and WB1P-Met tumors with tumors derived from published mouse models of luminal (*WapCre;Cdh1$^{F/F}$;Pten$^{F/F}$*, WEP) and basal-like (*K14Cre;Brca1$^{F/F}$;Trp53$^{F/F}$*, KB1P) breast cancer, using the three-gene signature that distinguishes the PAM50 subtypes[22]. **(E)** *Met* expression levels in WB1P and WB1P-Met tumors; $^{**}p < 0.01$ (two-sided Mann-Whitney U test). 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. **(F)** Genomic instability of WB1P, WB1P-Met and WB1P-Myc tumors; $^{**}p < 0.01$ and $^{****}p < 0.0001$ (two-sided Mann-Whitney U test). Scores for spleen samples from WB1P mice are shown as reference. **(G)** Genome-wide RUBIC analysis of the CNV profiles of WB1P-Met tumors. Significant amplifications and deletions are marked by light red and blue columns, respectively.

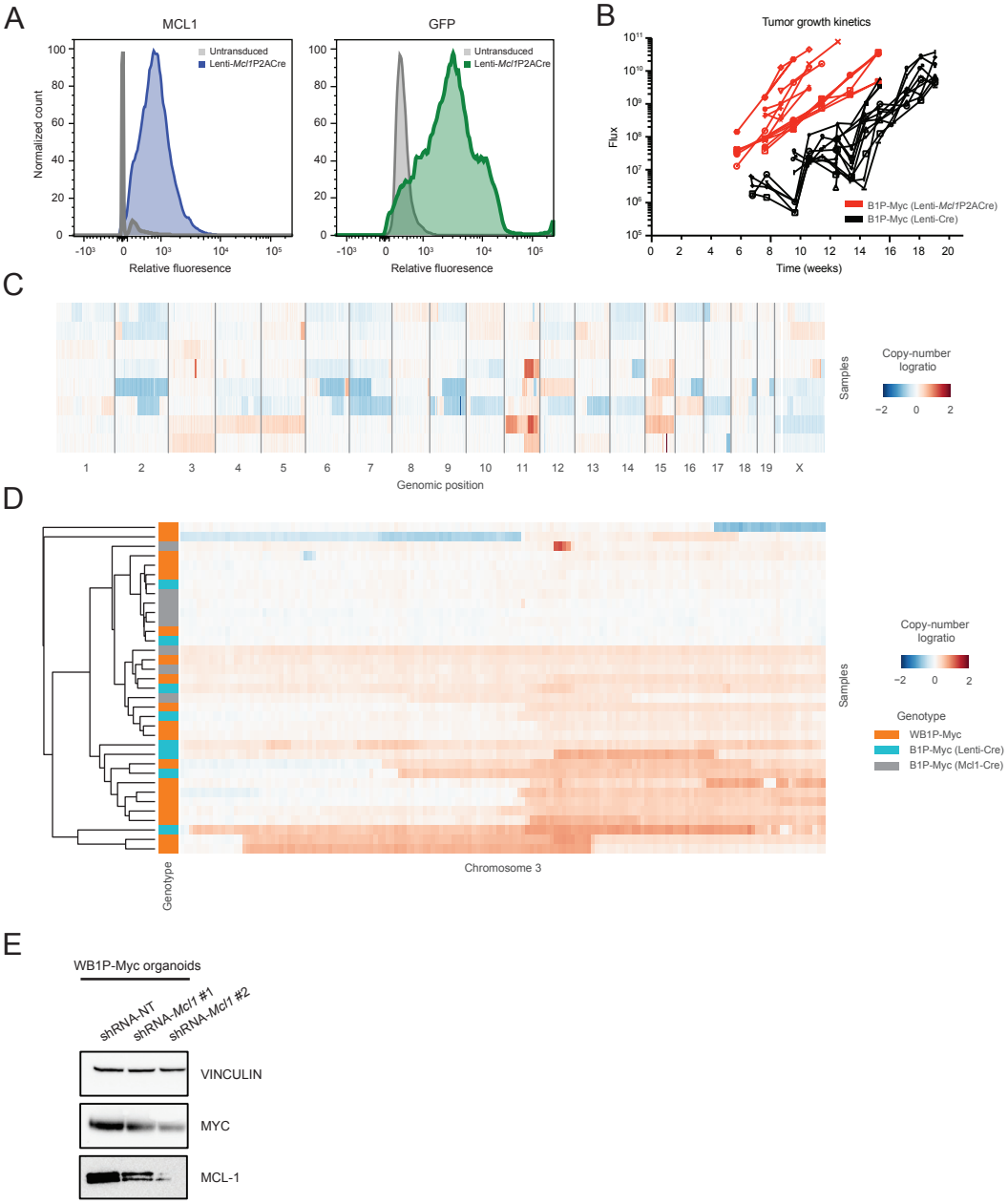


Fig. S5.7. Validation of MCL1 as a driver in BRCA1-associated TNBC. (A) *In vitro* validation of Lenti-Mcl1P2ACre in Cre-reporter cells 5 days after transduction. Expression of MCL1 as visualized by FACS using an anti-MCL1 antibody and FACS analysis of Cre-recombined GFP-positive cells are shown. **(B)** Longitudinal *in vivo* bioluminescence imaging of luciferase expression in B1P-Myc animals injected with Lenti-Cre (black lines) or Lenti-Mcl1P2ACre (red lines), showing signal build-up over time. **(C)** Heatmap showing the copy number logratios for tumors from B1P-Myc females injected with Lenti-Mcl1P2ACre, showing that recurrent gains on chromosomes 11 and 15 are retained, whilst the gain on chromosome 3 is less pronounced. **(D)** Unsupervised clustering (Euclidean distance, average linkage) of MYC-driven tumors from germline and somatic models (WB1P-Myc tumors and tumors from Lenti-Cre injected B1P-Myc mice) together with tumors from B1P-Myc females injected with Lenti-Mcl1P2ACre. The clustering shows that the focal gain surrounding *Mcl1* is lost in the majority of tumors from Lenti- Mcl1P2ACre injected B1P-Myc females. **(E)** Expression of MYC and MCL1 in WB1P-Myc organoids transduced with a non-targeting shRNA or with two independent sh*Mcl1* vectors, as visualized by immunoblotting using anti-MYC antibody 5 days after transduction.

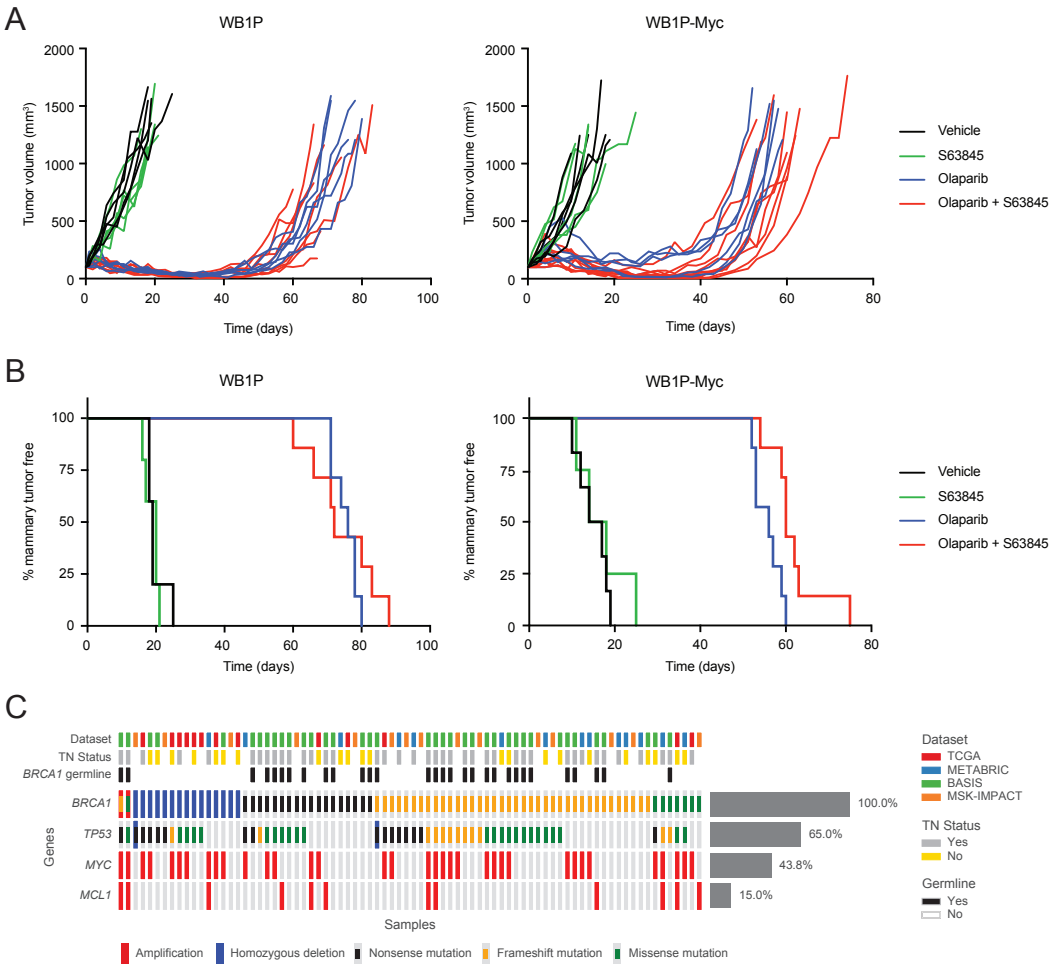


Fig. S5.8. Treatment of WB1P and WB1P-Myc tumors with MCL1 and PARP inhibitors. **(A-B)** Response of organoid-derived WB1P and WB1P-Myc tumors to MCL1 and PARP inhibitors, as visualized by tumor volume curves (A) and Kaplan Meier curves (B). WB1P and WB1P-Myc organoid lines were transplanted in the fourth mammary fat pad of nude mice. When tumors had reached a size of 100 mm, mice were treated with 25 mg kg^{-1} S63845 (administered i.v. once weekly for 5 weeks), 100 mg kg^{-1} olaparib (administered i.p. daily for 28 consecutive days), both drugs or vehicle. S63845 shows minimal additive efficacy over olaparib alone in WB1P-Myc tumors ($^*p < 0.05$ by Mantel-Cox test). **(C)** Overview of the mutations and copy number events in our panel of 80 *BRCA1*-mutated human breast tumor samples for *TP53*, *MYC* and *MCL1*.

5.7.2 Supplementary Tables (available online)

Tab. S5.1. Mutational landscape of human *BRCA1*-mutated TNBC. Deleterious mutations and copy number events in 80 *BRCA1*-mutated human breast tumor samples from four large-scale tumor-sequencing studies.

5.8 References

[1] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, et al. “Emerging landscape of oncogenic signatures across human cancers”. In: *Nature Genetics* 45.10 (2013), p. 1127 (cit. on p. 150).

[2] Cancer Genome Atlas Network et al. “Comprehensive molecular portraits of human breast tumours”. In: *Nature* 490.7418 (2012), p. 61 (cit. on p. 151).

[3] Rameen Beroukhim, Gad Getz, Leia Nghiemphu, et al. “Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma”. In: *Proceedings of the National Academy of Sciences* 104.50 (2007), pp. 20007–20012 (cit. on p. 151).

[4] Christiaan Klijn, Henne Holstege, Jeroen de Ridder, et al. “Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data”. In: *Nucleic Acids Research* 36.2 (2008), e13–e13 (cit. on p. 151).

[5] Ewald van Dyk, Marcel JT Reinders, and Lodewyk FA Wessels. “A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control”. In: *Nucleic Acids Research* 41.9 (2013), e100–e100 (cit. on p. 151).

[6] Ewald van Dyk, Marlous Hoogstraat, Jelle ten Hoeve, Marcel JT Reinders, and Lodewyk FA Wessels. “RUBIC identifies driver genes by detecting recurrent DNA copy number breaks”. In: *Nature Communications* 7 (2016), p. 12159 (cit. on pp. 151, 158, 170).

[7] Lars Zender, Mona S Spector, Wen Xue, et al. “Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach”. In: *Cell* 125.7 (2006), pp. 1253–1267 (cit. on p. 151).

[8] Minjung Kim, Joseph D Gans, Cristina Nogueira, et al. “Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene”. In: *Cell* 125.7 (2006), pp. 1269–1281 (cit. on p. 151).

[9] Jenny Mattison, Jaap Kool, Anthony G Uren, et al. “Novel candidate cancer genes identified by a large-scale cross-species comparative oncogenomics approach”. In: *Cancer Research* 70.3 (2010), pp. 883–895 (cit. on p. 151).

[10] Henne Holstege, Erik van Beers, Arno Velds, et al. “Cross-species comparison of aCGH data from mouse and human BRCA1-and BRCA2-mutated breast cancers”. In: *BMC Cancer* 10.1 (2010), p. 455 (cit. on pp. 151, 158).

[11] Natalie Meyer and Linda Z Penn. “Reflecting on 25 years with MYC”. In: *Nature Reviews Cancer* 8.12 (2008), p. 976 (cit. on p. 152).

[12] Theresia R Kress, Arianna Sabò, and Bruno Amati. “MYC: connecting selective transcriptional control to global RNA production”. In: *Nature Reviews Cancer* 15.10 (2015), p. 593 (cit. on p. 152).

[13] Ajay N Jain, Koei Chin, Anne-Lise Børresen-Dale, et al. “Quantitative analysis of chromosomal CGH in human breast tumors associates copy number abnormalities with p53 status and patient survival”. In: *Proceedings of the National Academy of Sciences* 98.14 (2001), pp. 7952–7957 (cit. on p. 152).

[14] JL Dillon, SM Mockus, Guruprasad Ananda, et al. “Somatic gene mutation analysis of triple negative breast cancers”. In: *The Breast* 29 (2016), pp. 202–207 (cit. on p. 152).

[15] Dai Horiuchi, Leonard Kusdra, Noelle E Huskey, et al. “MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition”. In: *Journal of Experimental Medicine* 209.4 (2012), pp. 679–696 (cit. on p. 152).

[16] Daniel C Koboldt, Qunyuan Zhang, David E Larson, et al. “VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing”. In: *Genome research* 22.3 (2012), pp. 568–576 (cit. on p. 152).

[17] M Chehani Alles, Margaret Gardiner-Garden, David J Nott, et al. “Meta-analysis and gene set enrichment relative to er status reveal elevated activity of MYC and E2F in the “basal” breast cancer subgroup”. In: *PloS One* 4.3 (2009), e4710 (cit. on p. 152).

[18] Sanjay Chandriani, Eirik Frengen, Victoria H Cowling, et al. “A core MYC gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response”. In: *PloS One* 4.8 (2009), e6693 (cit. on p. 152).

[19] Michael L Gatza, Joseph E Lucas, William T Barry, et al. “A pathway-based classification of human breast cancer”. In: *Proceedings of the National Academy of Sciences* 107.15 (2010), pp. 6994–6999 (cit. on p. 152).

[20] Mirjam C Boelens, Micha Nethe, Sjoerd Klarenbeek, et al. “PTEN loss in E-cadherin-deficient mouse mammary epithelial cells rescues apoptosis and results in development of classical invasive lobular carcinoma”. In: *Cell Reports* 16.8 (2016), pp. 2087–2101 (cit. on pp. 153, 154).

[21] Xiaoling Liu, Henne Holstege, Hanneke van der Gulden, et al. “Somatic loss of BRCA1 and p53 in mice induces mammary tumors with features of human BRCA1-mutated basal-like breast cancer”. In: *Proceedings of the National Academy of Sciences* 104.29 (2007), pp. 12111–12116 (cit. on pp. 153, 154, 166, 171).

[22] Benjamin Haibe-Kains, Christine Desmedt, Sherene Loi, et al. “A three-gene model to robustly identify breast cancer molecular subtypes”. In: *Journal of the National Cancer Institute* 104.4 (2012), pp. 311–325 (cit. on pp. 153, 154, 173, 175, 176, 179).

[23] Ivo J Huijbers, Rahmen Bin Ali, Colin Pritchard, et al. “Rapid target gene validation in complex cancer mouse models using re-derived embryonic stem cells”. In: *EMBO molecular medicine* 6.2 (2014), pp. 212–225 (cit. on pp. 154, 171).

[24] Silva Krause, Amy Brock, and Donald E Ingber. “Intraductal injection for localized drug delivery to the mouse mammary gland”. In: *Journal of Visualized Experiments* 80 (2013), e50692–e50692 (cit. on pp. 154, 171).

[25] Melanie R Rutkowski, Michael J Allegrezza, Nikolaos Svoronos, et al. “Initiation of metastatic breast carcinoma by targeting of the ductal epithelium with adenovirus-cre: a novel transgenic mouse model of breast cancer”. In: *Journal of Visualized Experiments: JoVE* 85 (2014), p. 51171 (cit. on p. 154).

[26] Luwei Tao, Maaike PA van Bragt, Elizabeth Laudadio, and Zhe Li. “Lineage tracing of mammary epithelial cells using cell-type-specific cre-expressing adenoviruses”. In: *Stem Cell Reports* 2.6 (2014), pp. 770–779 (cit. on p. 154).

[27] Prashant Kumar, Malini Mukherjee, Jacob PS Johnson, et al. “Cooperativity of Rb, Brca1, and p53 in malignant breast cancer evolution”. In: *PLoS Genetics* 8.11 (2012), e1003027 (cit. on p. 158).

[28] Erik S Knudsen, A Kathleen McClendon, Jorge Franco, et al. “RB loss contributes to aggressive tumor phenotypes in MYC-driven triple negative breast cancer”. In: *Cell Cycle* 14.1 (2015), pp. 109–122 (cit. on p. 158).

[29] Jennifer F Knight, Robert Lesurf, Hong Zhao, et al. “Met synergizes with p53 loss to induce mammary tumors that possess features of claudin-low breast cancer”. In: *Proceedings of the National Academy of Sciences* 110.14 (2013), E1301–E1310 (cit. on p. 159).

[30] Alexandra A Duarte, Ewa Gogola, Norman Sachs, et al. “BRCA-deficient mouse mammary tumor organoids to study cancer-drug resistance”. In: *Nature Methods* 15.2 (2018), p. 134 (cit. on pp. 161, 169, 171).

[31] András Kotschy, Zoltán Szlavik, James Murray, et al. “The MCL1 inhibitor S63845 is tolerable and effective in diverse cancer models”. In: *Nature* 538.7626 (2016), p. 477 (cit. on p. 163).

[32] Delphine Merino, James R Whittle, François Vaillant, et al. “Synergistic action of the MCL-1 inhibitor S63845 with current therapies in preclinical models of triple-negative and HER2-amplified breast cancer”. In: *Science Translational Medicine* 9.401 (2017), eaam7049 (cit. on pp. 163–165, 171).

[33] Natalie Meyer, Sam S Kim, and Linda Z Penn. “The Oscar-worthy role of Myc in apoptosis”. In: *Seminars in Cancer Biology*. Vol. 16. 4. Elsevier. 2006, pp. 275–287 (cit. on p. 165).

[34] John G Clohessy, Jianguo Zhuang, Jasper de Boer, Gabriel Gil-Gómez, and Hugh JM Brady. “Mcl-1 interacts with truncated Bid and inhibits its induction of cytochrome c release and its role in receptor-mediated apoptosis”. In: *Journal of Biological Chemistry* 281.9 (2006), pp. 5750–5759 (cit. on p. 165).

[35] Rameen Beroukhim, Craig H Mermel, Dale Porter, et al. “The landscape of somatic copy number alteration across human cancers”. In: *Nature* 463.7283 (2010), p. 899 (cit. on p. 165).

[36] Qingqing Ding, Xianghuo He, Weiya Xia, et al. “Myeloid cell leukemia-1 inversely correlates with glycogen synthase kinase-3β activity and associates with poor prognosis in human breast cancer”. In: *Cancer Research* 67.10 (2007), pp. 4564–4571 (cit. on p. 165).

[37] Kirsteen J Campbell, Sandeep Dhayade, Nicola Ferrari, et al. “MCL-1 is a prognostic indicator and drug target in breast cancer”. In: *Cell Death & Disease* 9.2 (2018), p. 19 (cit. on p. 165).

[38] CM Goodwin, OW Rossanese, ET Olejniczak, and SW Fesik. “Myeloid cell leukemia-1 is an important apoptotic survival factor in triple-negative breast cancer”. In: *Cell Death and Differentiation* 22.12 (2015), p. 2098 (cit. on p. 165).

[39] Justin M Balko, Jennifer M Giltnane, Kai Wang, et al. “Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets”. In: *Cancer Discovery* 4.2 (2014), pp. 232–245 (cit. on p. 165).

[40] Kyung-min Lee, Jennifer M Giltnane, Justin M Balko, et al. “MYC and MCL1 cooperatively promote chemotherapy-resistant breast cancer stem cells via regulation of mitochondrial oxidative phosphorylation”. In: *Cell Metabolism* 26.4 (2017), pp. 633–647 (cit. on p. 165).

[41] Zhifu Xiang, Hui Luo, Jacqueline E Payton, et al. “Mcl1 haploinsufficiency protects mice from Myc-induced acute myeloid leukemia”. In: *The Journal of Clinical Investigation* 120.6 (2010), pp. 2109–2118 (cit. on p. 165).

[42] Thaddeus D Allen, Chang Qi Zhu, Kirk D Jones, et al. “Interaction between MYC and MCL1 in the genesis and outcome of non–small-cell lung cancer”. In: *Cancer Research* 71.6 (2011), pp. 2212–2221 (cit. on p. 165).

[43] Clint Mitchell, Adly Yacoub, Hamed Hossein, et al. “Inhibition of MCL-1 in breast cancer cells promotes cell death in vitro and in vivo”. In: *Cancer Biology & Therapy* 10.9 (2010), pp. 903–917 (cit. on p. 165).

[44] Hui Liu, Charles J Murphy, Florian A Karreth, et al. “Identifying and targeting sporadic oncogenic genetic aberrations in mouse models of triple-negative breast cancer”. In: *Cancer Discovery* 8.3 (2018), pp. 354–369 (cit. on p. 166).

[45] Sjors M Kas, Julian R de Ruiter, Koen Schipper, et al. “Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular breast carcinoma”. In: *Nature Genetics* 49.8 (2017), pp. 1219–1230 (cit. on p. 167).

[46] Bastiaan Evers, Katarzyna Jastrzebski, Jeroen PM Heijmans, et al. “CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes”. In: *Nature Biotechnology* 34.6 (2016), p. 631 (cit. on p. 167).

[47] Neville E Sanjana, Ophir Shalem, and Feng Zhang. “Improved vectors and genome-wide libraries for CRISPR screening”. In: *Nature Methods* 11.8 (2014), pp. 783–784 (cit. on p. 167).

[48] Antonia Follenzi, Laurie E Ailles, Silvia Bakovic, Massimo Geuna, and Luigi Naldini. “Gene transfer by lentiviral vectors is limited by nuclear translocation and rescued by HIV-1 pol sequences”. In: *Nature Genetics* 25.2 (2000), pp. 217–222 (cit. on p. 167).

[49] Eva K Brinkman, Tao Chen, Mario Amendola, and Bas van Steensel. “Easy quantitative assessment of genome editing by sequence trace decomposition”. In: *Nucleic Acids Research* 42.22 (2014), e168–e168 (cit. on p. 168).

[50] Serena Nik-Zainal, Helen Davies, Johan Staaf, et al. “Landscape of somatic mutations in 560 breast cancer whole-genome sequences”. In: *Nature* 534.7605 (2016), p. 47 (cit. on p. 169).

[51] Linda Henneman, Martine H van Miltenburg, Ewa M Michalak, et al. “Selective resistance to the PARP inhibitor olaparib in a mouse model for BRCA1-deficient metaplastic breast cancer”. In: *Proceedings of the National Academy of Sciences* 112.27 (2015), pp. 8409–8414 (cit. on pp. 169–171).

[52] Stefano Annunziato, Sjors M Kas, Micha Nethe, et al. “Modeling invasive lobular breast carcinoma by CRISPR/Cas9-mediated somatic genome editing of the mammary gland”. In: *Genes & Development* 30 (2016), pp. 1470–1480 (cit. on p. 171).

[53] Patrick W B Derksen, Xiaoling Liu, Francis Saridin, et al. “Somatic inactivation of E-cadherin and p53 in mice leads to metastatic lobular mammary carcinoma through induction of anoikis resistance and angiogenesis”. In: *Cancer Cell* 10.5 (2006), pp. 437–449 (cit. on p. 171).

[54] Chris W Doornebal, Sjoerd Klarenbeek, Tanya M Braumuller, et al. “A preclinical mouse model of invasive lobular breast cancer metastasis”. In: *Cancer Research* 73.1 (2013), pp. 353–363 (cit. on p. 171).

6

Transcriptomics and transposon mutagenesis identify multiple *in vivo* mechanisms of resistance to the FGFR inhibitor AZD4547

Julian R. de Ruiter[1,2,10,*], Sjors M. Kas[1,10,*], Koen Schipper[1,10], Eva Schut[1,10], Lorenzo Bombardelli[3,10], Ellen Wientjens[1,10], Anne Paulien Drenth[1,10], Renske de Korte-Grimmerink[1,10], Sunny Mahakena[4], Christopher Philips[5], Paul D. Smith[6], Sjoerd Klarenbeek[7], Koen van de Wetering[4,8], Anton Berns[3,10], Lodewyk F. A. Wessels[2,9,10], Jos Jonkers[1,10]

[1] Division of Molecular Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[2] Division of Molecular Carcinogenesis, Netherlands Cancer Institute, Amsterdam, The Netherlands
[3] Division of Molecular Genetics, Netherlands Cancer Institute, Amsterdam, The Netherlands
[4] Division of Molecular Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[5] Discovery Sciences, IMED Biotech Unit, Astrazeneca, Cambridge, UK
[6] Bioscience, Oncology, IMED Biotech Unit, Astrazeneca, Cambridge, UK
[7] Experimental Animal Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[8] Present address: Department of Dermatology & Cutaneous Biology, Sidney Kimmel Medical College, Philadelphia, USA
[9] Department of EEMCS, Delft University of Technology, Delft, The Netherlands
[10] Oncode Institute, Amsterdam, The Netherlands
[*] These authors contributed equally to this work

Published in Cancer Research, 16 August 2018
Volume 78, Issue 19, Pages 5668-5679

6.1 Abstract

In human cancers, FGFR signaling is frequently hyperactivated by deregulation of FGF ligands or by activating mutations in the FGFR receptors such as gene amplifications, point mutations, and gene fusions. As such, FGFR inhibitors are considered an attractive therapeutic strategy for patients with mutations in FGFR family members. We previously identified *Fgfr2* as a key driver of invasive lobular carcinoma (ILC) in an *in vivo* insertional mutagenesis screen using the *Sleeping Beauty* transposon system. Here we explore whether these FGFR-driven ILCs are sensitive to the FGFR inhibitor AZD4547 and use transposon mutagenesis in these tumors to identify potential mechanisms of resistance to therapy. Combined with RNA sequencing-based analyses of AZD4547-resistant tumors, our *in vivo* approach identified several known and novel potential resistance mechanisms to FGFR inhibition, most of which converged on reactivation of the canonical MAPK–ERK signaling cascade. Observed resistance mechanisms included mutations in the tyrosine kinase domain of FGFR2, overexpression of MET, inactivation of RASA1, and activation of the drug-efflux transporter ABCG2. ABCG2 and RASA1 were identified only from *de novo* transposon insertions acquired during AZD4547 treatment, demonstrating that insertional mutagenesis in mice is an effective tool for identifying potential mechanisms of resistance to targeted cancer therapies.

6.2 Introduction

FGFRs are members of the receptor tyrosine kinase (RTK) family that bind to different FGF family members and are upstream of both the MAPK–ERK and PI3K–AKT signaling pathways. FGFRs dimerize upon FGF ligand binding, which results in cross-phosphorylation of the receptors cognate kinase domains and allows the binding of the adaptor protein FGFR substrate 2α (FRS2α), a key transducer of FGFR signaling[1]. Once bound, subsequent phosphorylation of FRS2α induces the recruitment of growth factor receptor-bound 2 (GRB2) and son of sevenless (SOS), resulting in activation of the MAPK–ERK signaling pathway. In contrast, activation of the PI3K–AKT signaling pathway is mediated by interactions between the FRS2α complex and GRB2-associated binding protein 1 (GAB1)[1].

In human cancers, FGFR signaling is frequently hyperactivated by deregulation of FGF ligands or by activating mutations in the receptors, which predominantly consist of gene amplifications, point mutations, and gene fusions[2,3]. As such, FGFR inhibitors are considered to be an attractive therapeutic strategy for patients with mutations in FGFR family members. Currently, no FGFR-targeted therapies are approved for the treatment of human cancer, but multiple therapeutics targeting

FGFR signaling are under investigation in several phase I/II clinical trials in different types of cancer[2,3]. These encompass several different approaches for inhibiting FGFR, including nonselective and selective FGFR small-molecule tyrosine kinase inhibitors, monoclonal antibodies against FGFRs and FGF ligand traps.

Although these initial trials have shown promising results concerning tolerability and antitumor activity of several FGFR inhibitors in a subset of patients[2,4–9], more research is required to determine the right criteria for patient selection and to tackle potential resistance mechanisms to improve FGFR-targeted therapies. Several studies have already identified resistance mechanisms to FGFR-targeting agents, including polyclonal secondary FGFR mutations (including gatekeeper mutations)[10–12], activation of alternative RTKs[13–16], and paracrine signaling of the tumor stroma[17,18]. However, because most of these mechanisms are identified in *in vitro* studies, they may not encompass the complete spectrum of resistance mechanisms to FGFR inhibitors.

In a previous study, we identified *Fgfr2* as a key driver of invasive lobular breast carcinoma (ILC) using a *Sleeping Beauty* (SB)-based transposon insertional mutagenesis screen in mice[19]. In this work, we explore how mouse ILCs (mILC) with SB transposon insertions in *Fgfr2* respond to treatment with the selective FGFR inhibitor AZD4547, and by which mechanisms they acquire resistance to FGFR inhibition. Our results show that the tumors exhibit increased FGFR signaling and initially regress upon treatment with AZD4547, but eventually develop treatment resistance. By performing a multiomics analysis focusing on the resistant tumors, we identify several known and novel mechanisms by which tumors become resistant to AZD4547 treatment. Two of these mechanisms were uniquely identified from *de novo* transposon insertions that were acquired during treatment, demonstrating that insertional mutagenesis in mice is an effective tool for identifying resistance mechanisms to targeted cancer therapies.

6.3 Materials and methods

Orthotopic transplantations and AZD4547 intervention Orthotopic transplantations of small tumor fragments were performed as previously described by Doornebal and colleagues[20]. For the WESB-*Fgfr2* tumor-derived cells, 200,000 cells were injected orthotopically into the right fat pad of 8- to 15-week-old wild-type syngeneic recipient FVB females in 20 μl Matrigel (Corning) and PBS (1:1). For the WESB-*Fgfr2*-EV and WESB-*Fgfr2*-ABCG2 cells, 200,000 cells were injected orthotopically into the right fat pad of 8-week-old NMRI-nude females (Janvier Labs) in 20 μl Matrigel (Corning) and PBS (1:1). All the drug interventions were initiated as

soon as the mammary tumors reached a size of $5 * 5$ mm (62.5 mm^3; tumor volume: length $*$ width^2 $* 0.5$). The treatments were performed daily by oral gavage for the indicated time with either the vehicle (1 %-Tween-80 in demineralized water) or AZD4547 (AstraZeneca) at a dose of 12.5 mg kg^{-1} d^{-1}. The experimental cohort was monitored and mice were sacrificed (overall survival) when the (total) mammary tumor burden reached a size of approximately 1500 mm^3 (tumor volume: length $*$ width^2 $* 0.5$) or suffered from clinical signs of distress (respiratory distress, ascites, distended abdomen, rapid weight loss, and severe anemia) caused by mammary tumor burden or metastatic disease. One hour after the last dosing, mice were sacrificed and the tumor, lungs, liver, spleen, and tumor-draining lymph nodes were collected for further analysis. The mouse technicians were blinded to the sample groups for the treatments of WESB-*Fgfr2* and WESB-*Fgfr2*-ABCG2 established tumors in mice. All animal 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.

Cell culture The isolation of primary tumor cells of the SB-induced mILCs (referred to as WESB cells) was performed as previously described by Kas and colleagues[19]. WESB cells were cultured in DMEM-F12 medium containing 10 % FBS, 100 IU ml^{-1} penicillin, 100 μg ml^{-1} streptomycin (all from Life Technologies). MEF3.8 cells were cultured in DMEM-F12 medium containing 10 % FBS, 100 IU ml^{-1} penicillin, 100 μg ml^{-1} streptomycin (all from Life Technologies). Phoenix packaging cells were cultured in Iscove's medium (Life Technologies) containing 10 % FBS, 100 IU ml^{-1} penicillin, and 100 μg ml^{-1} streptomycin. WESB-*Fgfr2* cells were transduced with LZRS-IRES-GFP or LZRS-*Bcrp1*-IRES-GFP as previously described by Allen and colleagues[21]. Single GFP+ cells were sorted and allowed to recover before they were used in the experiments. Cell line authentication was not conducted. All cell lines were kept at low passage and routinely tested for Mycoplasma contamination using the MycoAlert Mycoplasma Detection Kit (Lonza).

Additional experimental details regarding the cell viability, clonogenic, and competition assays are described in the Supplementary Material.

Vesicular Transport Assays Vesicular transport assays were performed using the rapid filtration method as previously described[22,23]. Additional experimental details are described in the Supplementary Material.

Nucleic acid isolation DNA and RNA were isolated from whole tumor pieces using the Allprep DNA/RNA Mini Kit (Qiagen) according to the manufacturer's protocol.

Additional experimental details regarding the detection of the endogenous *Fgfr2–Tbc1d1* fusion and the *Met* qPCR copy number analysis are described in the Supplementary Material.

Amplification of SB transposon insertions Transposon insertions were amplified and mapped following a previously described tagmentation-based DNA sequencing protocol[24]. Additional experimental details and the analysis of the insertions sites are described in the Supplementary Material.

Antibodies The primary antibodies to the following proteins were used: FGFR2 (1:1,000, GeneTex 10648), phospho-FGFR (1:1,000, CST 3471), FRS2 (1:1,000, ProteinTech 11503-1-AP), phospho-FRS2 (Tyr436) (1:1,000, Abcam 193363), AKT1 (1:1000, CST 2938), phospho-AKT(Ser473; 1:1,000, CST 4060), p44/42 MAP kinase (1:1,000, CST 4695), phospho-p44/42 MAPK ERK1/ERK2 (Thr202/Tyr204; 1:1,000 CST 9101), ABCG2 (1:400, Abcam 24115) and β-actin (1:50,000, Sigma A5441).

Additional details regarding immunoblotting and IHC are described in the Supplementary Material.

***In silico* modeling of the FGFR2 kinase in complex with the inhibitor AZD4547** A composite complex of AZD4547 bound to mutated FGFR2 was built using the crystal structure of the FGFR1 kinase domain in complex with AZD4547 (PDB code V405) as a template. A structural alignment of the FGFR2 kinase domain crystal structure (PDB code 2PVF) was performed and the positions of the resistance mutations were mapped onto this alignment. Molecular graphic images were prepared using the CHIMERA package[25].

Statistical analysis For the mouse studies, no statistical tests were performed to determine the appropriate sample size. Survival probabilities were estimated using the Kaplan–Meier method and compared using the Mantel–Cox test. The effect of AZD4547 treatment on tumor growth of WESB-*Fgfr2*-EV and WESB-*Fgfr2*-ABCG2 established tumors was tested using mixed linear models (details in Supplementary Material). To test for differential expression of *Abcg2*, *Rasa1*, and *Pcdh15* over the insertion sites across all SB-induced tumors, we used the group-wise differential expression test implemented in IM-Fusion[26]. The investigators were not blinded to the sample groups for all experiments. Graphs and error bars represent mean ± SD. Python 3.5, R 3.3.1 and GraphPad Prism 7.03 were used for the statistical analyses. p values < 0.05 were considered significant.

Data availability Raw tagmentation and RNA-sequencing data are available in ENA under accession number PRJEB25507.

6.4 Results

6.4.1 Activation of Fgfr2 induces mILC formation

In a previous study, we performed an SB insertional mutagenesis screen in mice with mammary-specific inactivation of *Cdh1* (encoding E-cadherin) to identify genes and pathways driving the development of ILC[19]. Analysis of common insertion sites (CIS) in the SB-induced mILCs showed that a majority (56 of 99) of these tumors had SB insertions in Fgfr2, providing strong evidence that *Fgfr2* is a driver of ILC.

To confirm active FGFR signaling in mILCs with SB insertions in *Fgfr2*, we established cell lines from two SB-induced tumors, one with an insertion upstream of *Fgfr2* (WESB-*Fgfr2*) and one without an insertion in or near *Fgfr2* (WESB). We next compared the expression of FGFR2 and downstream signaling proteins between the two tumor cell lines. Although immunoblot analysis with an anti-FGFR2 antibody revealed no expression of native FGFR2 in either of the cell lines, we observed a doublet of approximately 240 kDa that was only expressed in the WESB-*Fgfr2* cell line (Figure 6.1A). This size coincided with the predicted protein size of a gene fusion between *Fgfr2* and *Tbc1d1* (Supplementary Figure S6.1A-B), which we previously identified in RNAseq data from this SB-induced mILC[26]. Similar *FGFR2* gene fusions were previously identified in several other studies, which demonstrated that these fusions result in increased FGFR signaling[27]. In line with this, comparison of signaling proteins downstream of FGFR2 showed increased expression of phosphorylated FRS2α in WESB-*Fgfr2* tumor-derived cells compared with WESB cells.

These results demonstrate that WESB-*Fgfr2* cells show increased expression of an *Fgfr2–Tbc1d1* fusion gene, which is driven by an SB insertion upstream of *Fgfr2*. The increased expression of this fusion gene results in activation of FGFR signaling, suggesting that FGFR inhibition could be an interesting therapeutic strategy in these tumors.

6.4.2 Mouse ILCs with SB insertions in *Fgfr2* are dependent on FGFR signaling

To determine if WESB-*Fgfr2* cells were indeed sensitive to FGFR inhibition, we treated these cells with the selective FGFR inhibitor AZD4547, which is currently being evaluated in several early-phase clinical trials[2]. After treatment with 100 nmol l^{-1} AZD4547, WESB-*Fgfr2* cells showed a decrease in expression of phosphorylated FGFR, FRS2α, and ERK1/2 (Figure 6.1B), confirming inhibition of the FGFR signaling pathway. Compared with WESB cells, the WESB-*Fgfr2* cells also showed reduced

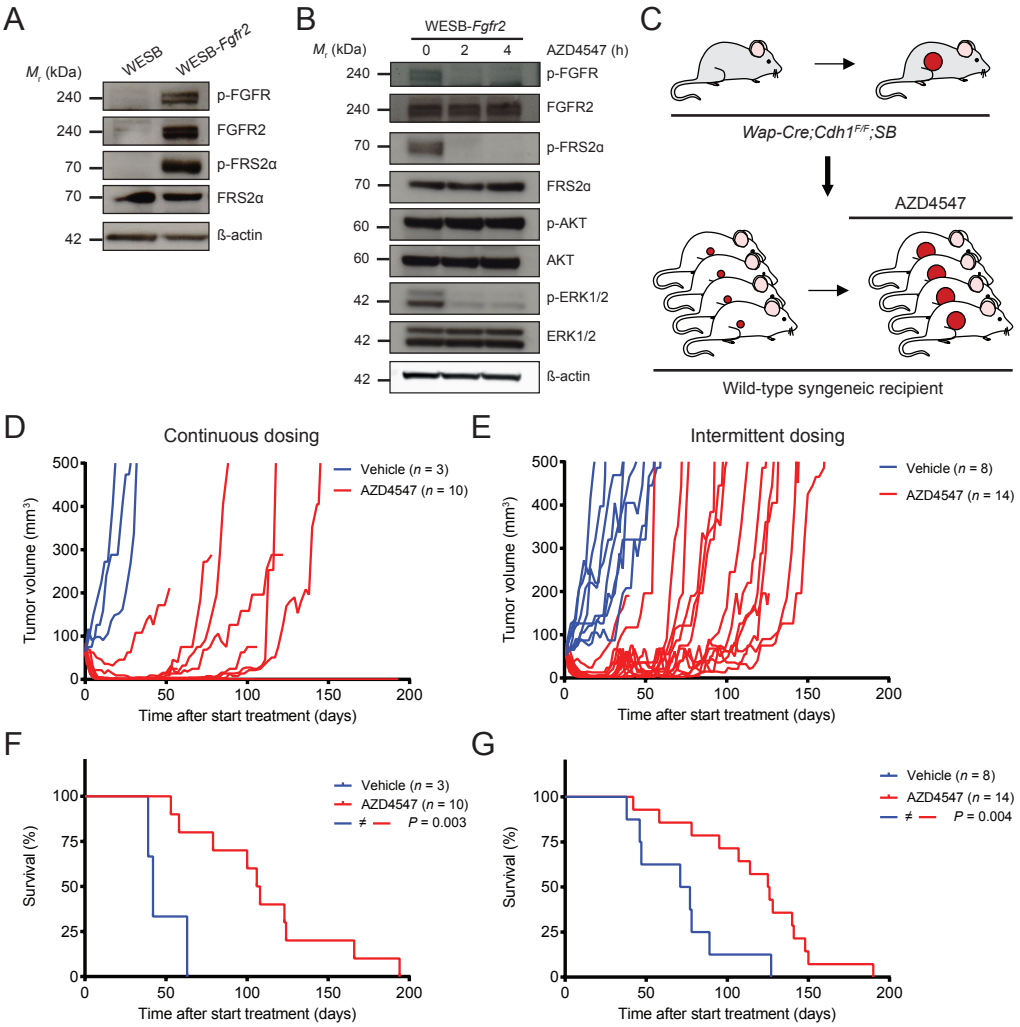


Fig. 6.1. Intervention study with AZD4547 in mILCs with active FGFR signaling. **(A)** Representative immunoblot ($n = 3$) for the expression of FGFR2 and its downstream signaling proteins in WESB and WESB-*Fgfr2* cells. β-Actin was used as a loading control. **(B)** The effect of FGFR inhibition on FGFR signaling in WESB-*Fgfr2* cells after short-term treatment with the FGFR inhibitor AZD4547 (100 nmol l^{-1}), as visualized by immunoblotting with antibodies detecting total and phosphorylated FGFR, FRS2α, AKT, and ERK1/2. β-Actin was used as a loading control. **(C)** Schematic overview depicting the orthotopic transplantation of SB-induced mILC fragments into wild-type syngeneic recipient mice and the subsequent drug intervention with AZD4547. **(D and E)** Tumor growth kinetics of orthotopically transplanted WESB-*Fgfr2* tumors under the continuous (D) and intermittent (E) treatment schedules with vehicle (blue) or AZD4547 (red). **(F and G)** Kaplan–Meier curves showing the overall survival of tumor-bearing mice under continuous (F) and intermittent (G) treatment with vehicle (blue) or AZD4547 (red). p values were calculated using a Mantel–Cox test.

viability upon exposure to increasing concentrations of AZD4547 (Supplementary Figure S6.1C), indicating that WESB-*Fgfr2* cells are dependent on FGFR signaling for their survival *in vitro*.

To determine the efficacy of AZD4547 *in vivo*, we orthotopically transplanted WESB-*Fgfr2* tumor fragments into multiple wild-type syngeneic recipient FVB animals and treated these animals with vehicle or AZD4547 ($12.5\,\mathrm{mg\,kg^{-1}\,d^{-1}}$) daily via oral gavage (Figure 6.1C). To reduce the potential toxicity of prolonged treatment and to test the effect of a 'drug holiday' on tumor growth, the animals were treated using either a continuous or an intermittent dosing schedule (Supplementary Figure S6.1D). In both dosing schedules, tumors treated with AZD4547 showed decreased expression of downstream FGFR signaling and increased expression of cleaved caspase-3 (Supplementary Figure S6.1E–G), which resulted in tumor regression within 10 to 20 days after start of the treatment (Figure 6.1D-E). Furthermore, the majority of the AZD4547-treated tumors (9 of 10 continuous-treated and 13 of 14 intermittent-treated tumors) did not show any regrowth within the first treatment cycle of 24 days.

In the majority of mice, continuous treatment with AZD4547 resulted in tumor control for at least 40 days after start of the treatment, resulting in an increased overall survival compared with the vehicle-treated animals (Figure 6.1F). Notably, 2 of 10 sacrificed animals did not show any remaining tumor cells. In contrast, all intermittently treated mice showed tumor regrowth after the first treatment cycle of 24 days. However, these tumors remained sensitive to multiple additional cycles of AZD4547 treatment (Supplementary Figure S6.2A), resulting in an increased overall survival compared with the vehicle-treated animals (Figure 6.1G). Similar results were obtained with continuous AZD4547 treatment of mice after the orthotopic injection of WESB-*Fgfr2* tumor-derived cells (Supplementary Figure S6.2B-C).

Although the overall survival of AZD4547-treated mice was increased compared with vehicle-treated animals, there was no significant difference in survival between the continuous dosing (107 days) and the intermittent dosing (126 days) groups. However, in the continuous dosing group an increased number of animals succumbed due to clinical signs of distress (respiratory distress, ascites, distended abdomen, rapid weight loss, and severe anemia), suggesting that the intermittent treatment schedule is less toxic for the animals (Supplementary Figure S6.2D). In spite of the potent anticancer activity of AZD4547, both treatment schedules failed to deliver long-term tumor control, most likely due to the emergence of acquired therapy resistance. This reflects the major problem observed in patients with cancer treated with targeted anticancer therapies.

6.4.3 Transcriptome analysis identifies known and novel secondary FGFR2 mutations and increased MET expression in AZD4547-resistant tumors

To explore potential resistance mechanisms to FGFR inhibition, we performed RNA-sequencing of AZD4547-sensitive and -resistant tumors obtained from vehicle-treated and AZD4547-treated animals (Supplementary Table S6.1), respectively, and compared their mutational spectra to identify mutations that were acquired during AZD4547 treatment. In this approach, we initially focused on known resistance mechanisms to FGFR-targeting therapeutics[10–16], which include upregulation of alternative RTKs and secondary FGFR mutations.

To assess if upregulation of other RTKs could explain the resistance of these tumors, we used RNA-sequencing data to determine changes in gene expression for *Kit*, *Met*, and all FGFR-, EGFR-, and IGF-related RTKs. For this purpose, we used DIDS, an algorithm that is specifically designed to identify differentially expressed genes in heterogeneous populations[28]. Although this analysis did not identify any RTKs that were significantly differentially expressed across multiple samples, it did identify two AZD4547-resistant tumors with increased expression of *Met* [also known as hepatocyte growth factor receptor (HGFR)] and one AZD4547-resistant tumor with increased expression of insulin-like growth factor 1 receptor (*Igf1r*), compared with vehicle-treated tumors (Figure 6.2A; Supplementary Figure S6.3A). Subsequent analysis of IGF1R and MET protein expression by IHC did not support IGF1R as a potential resistance mechanism, because no correlation was observed between expression levels of Igf1r mRNA and IGF1R protein (Figure 6.2B). In contrast, both AZD4547-resistant tumors with high *Met* mRNA expression showed amplification of *Met* (Supplementary Figure S6.3B) and overexpression of MET protein (Figure 6.2C), whereas vehicle-, AZD4547-treated, and other AZD4547-resistant tumors were negative for MET expression (Supplementary Figure S6.3C). These *in vivo* results are in line with previous *in vitro* studies showing that upregulation of MET attenuates the efficacy of FGFR inhibition in tumor cells[14,15], indicating that upregulation of MET may also counteract the therapeutic efficacy of AZD4547 *in vivo*. In contrast to previous *in vitro* studies[13,16], we did not observe an obvious increase in mRNA expression of the *Egfr* or other EGFR-family members in any of the AZD4547-resistant tumors.

Next, to determine if any mutations in RTKs or members of the MAPK–ERK pathway could explain the resistance of these tumors, we used the RNA-sequencing data to identify mutations in the above-mentioned RTKs and genes involved in the MAPK–ERK signaling pathway (Supplementary Table S6.2). Using this approach, we identified 12 missense mutations in *Fgfr2* affecting 11 different amino acids, of

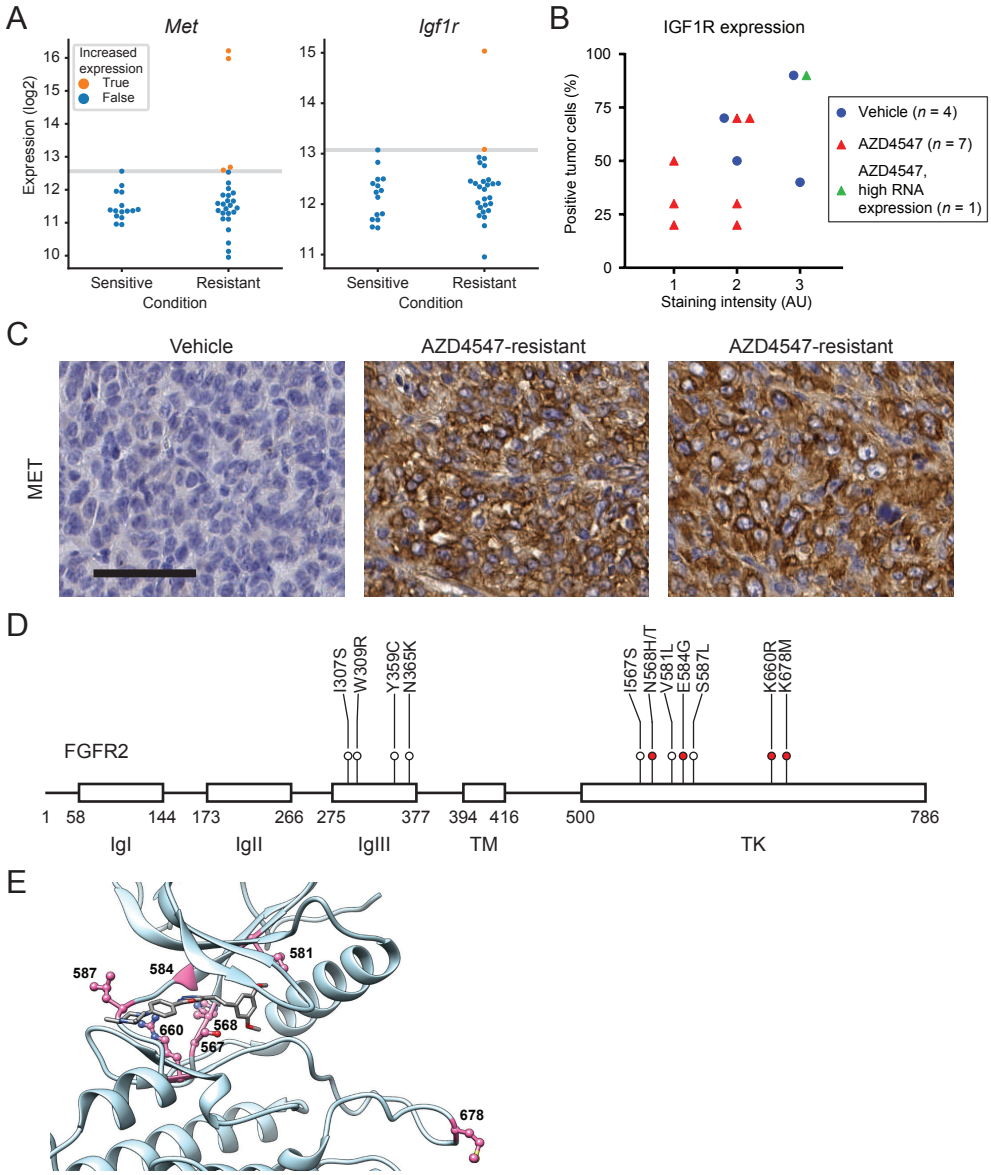


Fig. 6.2. **Transcriptome analysis of AZD4547-resistant tumors.** **(A)** Differential expression analysis of *Met* and *Igf1r* in AZD4547-sensitive ($n = 15$) and -resistant ($n = 27$) tumors using DIDS, showing outliers for *Met* ($n = 2$) and *Igf1r* ($n = 1$). **(B)** Quantification of IGF1R expression in AZD4547-sensitive ($n = 4$) and -resistant ($n = 8$) tumors. AU, arbitrary unit. **(C)** Representative immunohistochemical stainings of MET in AZD4547-sensitive (vehicle) and -resistant tumors. Scale bar, 50 μm. **(D)** Schematic overview showing the locations of mutations identified for *Fgfr2* in the AZD4547-resistant tumors. Red mutations have been previously reported in patients with FGFR2 fusion–positive cholangiocarcinoma, whose tumors acquired resistance to the selective FGFR inhibitor NVP-BGJ398. Numbers indicate amino acid residue positions (mouse). Ig, immunoglobulin-like domain; TM, transmembrane domain; TK, tyrosine kinase domain. **(E)** *In silico* modeling of AZD4547 (middle) in the ATP-binding pocket of FGFR2. The FGFR2 kinase domain is depicted as a sky-blue ribbon with residues of interest colored by atom type: carbon, pink; nitrogen, blue; oxygen, red. AZD4547 is colored by atom type: carbon, gray; nitrogen, blue; oxygen, red.

which four were located in the third immunoglobulin (Ig)-like domain (IgIII) and seven were located in the tyrosine kinase domain (Figure 6.2D).

To predict the effects of the mutations in the FGFR2 kinase domain on AZD4547 binding, we mapped the seven missense mutations onto the FGFR2 protein structure and observed that residues I567, N568, V581, E584, S587 reside in the ATP-binding pocket of FGFR2 (Figure 6.2E). As a consequence, these mutated residues directly perturb the binding site of AZD4547. In addition, the E584G mutation is located in the kinase hinge and introduces flexibility at the key recognition motif for AZD4547. The K660R mutation is adjacent to the binding site, which suggests that the binding of AZD4547 is indirectly perturbed. Finally, the K678M mutation is located in the kinase activation loop, suggesting that this mutation alters the dynamics of the activation loop and favors the active conformation.

Interestingly, four out of the seven missense mutations in the tyrosine kinase domain reflected recurrent point mutations that were previously reported in patients with FGFR2 fusion–positive cholangiocarcinoma, whose tumors acquired resistance to the selective FGFR inhibitor NVP-BGJ398[10]. In this previous work, structural characterization combined with functional *in vitro* studies showed that these FGFR2 kinase mutations either induce a steric clash with NVP-BGJ398 in the ATP-binding pocket or destabilize the inactive conformation of the kinase. Together, these data indicate that the seven mutations in the ATP-binding pocket of FGFR2 disrupt the binding of selective FGFR tyrosine kinase inhibitors and therefore hamper their therapeutic efficacy.

6.4.4 AZD4547-resistant SB-induced tumors show *de novo* insertions in candidate resistance genes

Because of the presence of a constitutively active SB transposase, the SB-induced mILCs could be capable of developing resistance by acquiring *de novo* transposon insertions in or near resistance genes during AZD4547 treatment. To determine if SB-mediated mutagenesis might indeed be driving resistance in some of these tumors, we performed an insertion analysis of 27 AZD4547-resistant SB-induced tumors and compared the identified SB insertions to those found in the donor tumor and 15 vehicle-treated tumors. Globally, this analysis showed that the majority of the clonal insertions in the donor tumor (e.g., *Fgfr2*, *Ppp1r12a*, *Slc16a9*, and *Trps1*) were maintained after orthotopic transplantation and long-term treatment of the tumor-bearing mice (Supplementary Figure S6.4A; Supplementary Tables S6.1 and S6.3). Interestingly, additional SB insertions were observed in both vehicle- and AZD4547-treated tumors, indicating that mobilization of transposons still occurs after transplantation of SB-induced tumors.

To specifically identify *de novo* insertions that might be driving resistance to AZD4547, we filtered for genes that contained SB insertions in the AZD4547-resistant tumors but not in the donor or vehicle-treated tumors (Supplementary Figure S6.4B). This analysis revealed three candidate resistance genes (*Abcg2*, *Rasa1*, and *Pcdh15*) with insertions in at least three AZD4547-resistant tumors (Figure 6.3A). Of these three genes, *Abcg2* contained several independent insertions that were mainly in the sense orientation and located upstream of the transcription start site (Figure 6.3B), indicating that these insertions result in increased *Abcg2* expression. In support of this, these insertions coincided with increased mRNA and protein expression of ABCG2 (Figure 6.3C; Supplementary Figure S6.5A–C). Variable ABCG2 expression highlights intratumor heterogeneity in mechanisms of AZD4547 resistance. In contrast to *Abcg2*, *Rasa1*, and *Pcdh15* contained either a mix of sense/antisense insertions or purely antisense insertions, suggesting that these genes are inactivated (Figure 6.3D–G). Further analysis revealed decreased expression of exons downstream of the insertion sites in *Rasa1*, supporting inactivation of *Rasa1* via truncation of the gene, whereas expression of *Pcdh15* was not markedly affected.

To investigate whether insertions from the donor tumor might contribute to intrinsic treatment resistance, we compared the relative support scores of insertions between untreated tumors (vehicle-treated tumors and the donor tumor) and AZD4547-resistant tumors to determine if insertions in specific genes were enriched after AZD4547 treatment. This analysis identified six genes (*Arid1a*, *Myh9*, *Fbxw7*, *Matr3*, *Slc16a9*, and *Map4k4*) with increased support scores in AZD4547-resistant tumors, indicating that subclones with insertions in these genes are selected for during treatment (Supplementary Figure S6.5D). These genes might therefore be involved in intrinsic resistance to AZD4547. Interestingly, the top three genes (*Arid1a*, *Myh9*, and *Fbxw7*) were previously identified as candidate driver genes in ILC formation[19].

Collectively, these results show that persistent mobilization of transposons in SB-induced mILCs allows them to acquire new insertions during treatment and that this approach can be used to identify novel resistance mechanisms. Our analysis implicates upregulation of *Abcg2* and inactivation of *Rasa1* as additional resistance mechanisms to AZD4547, which were not previously identified with our mutational analyses. This demonstrates that combining insertional mutagenesis with drug treatments poses an effective strategy for identifying resistance mechanisms to targeted therapies in mice.

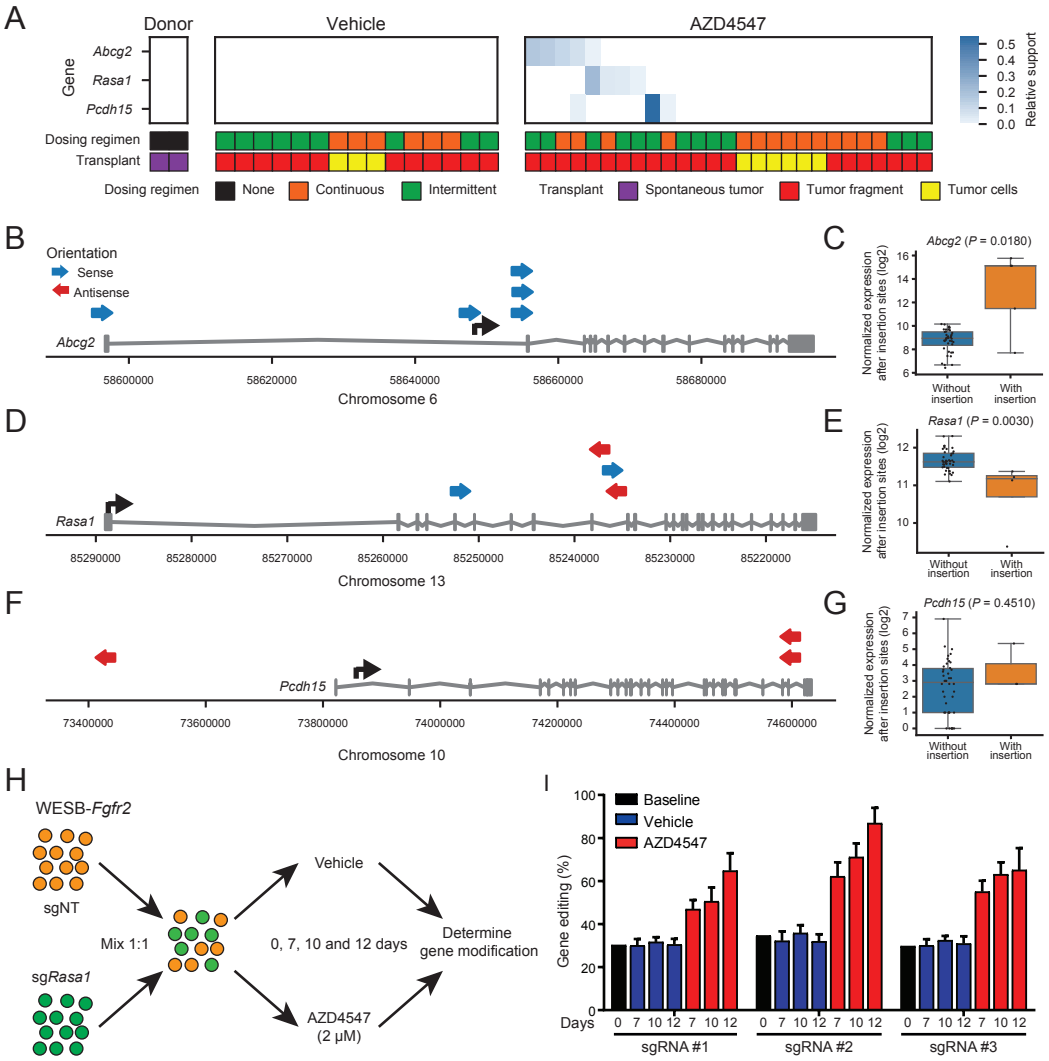


Fig. 6.3. Overview of insertions and corresponding gene expression of candidate resistance genes. (A) Overview of insertions in candidate resistance genes that were mutated in at least three AZD4547-resistant tumors ($n = 27$) and were not mutated in any of the donor (two technical replicates) or vehicle-treated ($n = 15$) tumors. Relative clonality of insertions is indicated by relative support scores (blue), which were calculated by counting the number of mate pairs supporting an insertion and normalizing these support scores to the highest score of the corresponding sample. **(B–G)** Left, visualization of SB insertions (arrows) in *Abcg2* (B), *Rasa1* (D), and *Pcdh15* (F). Right, normalized gene expression values after the insertion sites of *Abcg2* (C), *Rasa1* (E), and *Pcdh15* (G) in all *SB*-induced tumors with and without insertions in the respective genes. p values were calculated using a Mann–Whitney U test, as implemented in IM-Fusion. **(H)** Schematic overview of the *in vitro* competition assay performed with WESB-*Fgfr2* cells transfected with modified pX330 vectors containing sgRNAs targeting *Rasa1* (sg*Rasa1*) or a nontargeting sgRNA (sgNT). I, Competition assay of WESB-*Fgfr2*-sgNT and WESB-*Fgfr2*-sg*Rasa1* cells mixed in a 1:1 ratio at day 0. After 7, 10, and 12 days of vehicle or AZD4547-treatment ($2\,\mu\text{mol}\,\text{l}^{-1}$), the allele distributions of the polyclonal populations were quantified by the percentage of frameshift mutations in *Rasa1* using the TIDE algorithm[29]. The percentages of gene modifications are mean ± SD of at least four replicates.

6.4.5 Loss of RASA1 reduces sensitivity of WESB-*Fgfr2* tumor cells to AZD4547

To test whether inactivation of *Rasa1* induces resistance to AZD4547 treatment, we transfected WESB-*Fgfr2* tumor-derived cells with modified pX330 vectors containing single guide RNAs (sgRNA) targeting three different genomic regions of *Rasa1* or a nontargeting sgRNA (sgNT). All the *Rasa1* targeting sgRNAs induced efficient modification of the *Rasa1* target sites in the transfected cell populations (Supplementary Figure S6.5E–G), as determined by tracking of insertions or deletions [indels] by decomposition (TIDE) analysis[29]. To test for drug sensitivity, we performed an *in vitro* competition assay with a mixture of WESB-*Fgfr2*-sgNT and WESB-*Fgfr2*-sg*Rasa1* cells (1:1 ratio) in the presence or absence of $2\,\mu\text{mol}\,\text{l}^{-1}$ AZD4547 and subsequently quantified the allele distribution of the polyclonal population using the frequency of frameshift mutations in *Rasa1* (Figure 6.3H). After prolonged AZD4547 treatment, the polyclonal population was enriched for *Rasa1* frameshift mutations for all three *Rasa1*-targeted regions (Figure 6.3I), indicating that Rasa1-depleted cells were less sensitive to AZD4547 treatment compared with control cells. In contrast, the allele distributions were not affected when cells were cultured without AZD4547, demonstrating that the observed effect was not due to a difference in proliferation between WESB-*Fgfr2*-sgNT and WESB-*Fgfr2*-sg*Rasa1* cells. Altogether, these data show that inactivation of *Rasa1* reduces the sensitivity of WESB-*Fgfr2* cells to AZD4547 treatment.

Abcg2 is an ATP-binding cassette (ABC) efflux transporter, suggesting that overexpression of this gene may induce resistance through increased extrusion of AZD4547 from the tumor cells. To determine if this is indeed the case, we first sought to confirm that AZD4547 is a substrate for the ABCG2 transporter. To this end, we performed a vesicular transport assay (Figure 6.4A), in which we measured the uptake of tritium-labeled methotrexate ([^{3}H]-MTX) in inside-out Sf9-membrane vesicles expressing ABCG2 (Sf9-ABCG2), both in the presence and absence of increasing concentrations of AZD4547, and compared the results to the uptake of [^{3}H]-MTX in control Sf9-membrane vesicles (Sf9-control). This showed that ATP-dependent uptake of [^{3}H]-MTX by ABCG2 was inhibited by AZD4547 (Figure 6.4B), indicating that AZD4547 is indeed a substrate of ABCG2.

6.4.6 Overexpression of ABCG2 reduces sensitivity to AZD4547

To further explore whether increased expression of Abcg2 reduces the sensitivity of cells to AZD4547, we used mouse embryonic fibroblasts (MEF) derived from

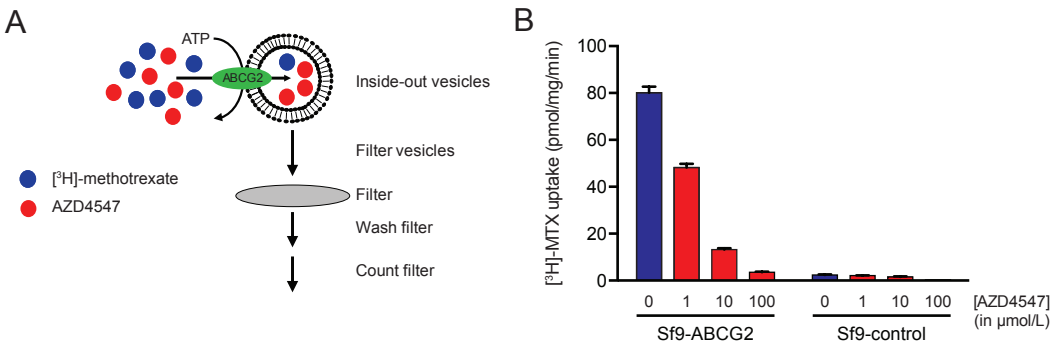


Fig. 6.4. AZD4547 is a substrate for ABCG2. (A) Schematic overview of the vesicular transport assay. Control or ABCG2-containing inside-out membrane vesicles were incubated with [^{3}H]-MTX (blue) and increasing concentrations of AZD4547 (red) in the presence or absence of ATP. After 10 minutes incubation, the membrane vesicles were captured using rapid filtration and the filters were washed to eliminate the excess of [^{3}H]-MTX that was not transported into the vesicles. The retained radioactivity inside the membrane vesicles was measured using liquid scintillation counting. **(B)** Inhibition of ABCG2-mediated [^{3}H]-MTX transport by increasing concentrations of AZD4547 (red). Values are corrected for transport in the absence of ATP. Data are mean ± SD of three independent experiments, which were each performed in triplicate.

Abcb1a$^{-/-}$*;Abcb1b*$^{-/-}$*;Abcc1*$^{-/-}$ mice (hereafter referred to as MEF3.8), which have very low background expression of endogenous ABCG2[21]. Furthermore, these MEFs lack both P-glycoprotein (P-gp, encoded by *Abcb1a* and *Abcb1b*), and the multidrug resistance-associated protein 1 (MRP1, encoded by *Abcc1*), allowing us to exclude confounding influences of these other drug efflux transporters. To test for drug sensitivity, MEF3.8 cells were transduced with an empty retroviral expression vector (MEF3.8-EV) or a vector containing *Abcg2* (MEF3.8-ABCG2) and subsequently exposed to increasing concentrations of AZD4547 in long-term clonogenic assays (Figure 6.5A). Compared with MEF3.8 cells, MEF3.8-ABCG2 cells were able to survive higher concentrations of AZD4547 (Figure 6.5B-C), indicating that increased ABCG2 expression reduces the efficacy of AZD4547.

To confirm that increased expression of ABCG2 also reduces the sensitivity of treatment-naive WESB-*Fgfr2* cells to AZD4547, we transduced these cells with an empty retroviral expression vector (WESB-*Fgfr2*-EV) or a vector containing *Abcg2* (WESB-*Fgfr2*-ABCG2) and treated the transduced cells with AZD4547. Short-term treatment of WESB-*Fgfr2*-EV cells with AZD4547 resulted in decreased phosphorylation of FGFR, FRS2α, and ERK1/2, whereas the phosphorylation levels of these proteins were less affected in AZD4547-treated WESB-*Fgfr2*-ABCG2 cells (Figure 6.5D).

To test the effect of ABCG2 overexpression on the responsiveness of established tumors to AZD4547, we injected WESB-*Fgfr2*-EV and WESB-*Fgfr2*-ABCG2 cells into the mammary glands of immunocompromised NMRI-nude mice and these animals

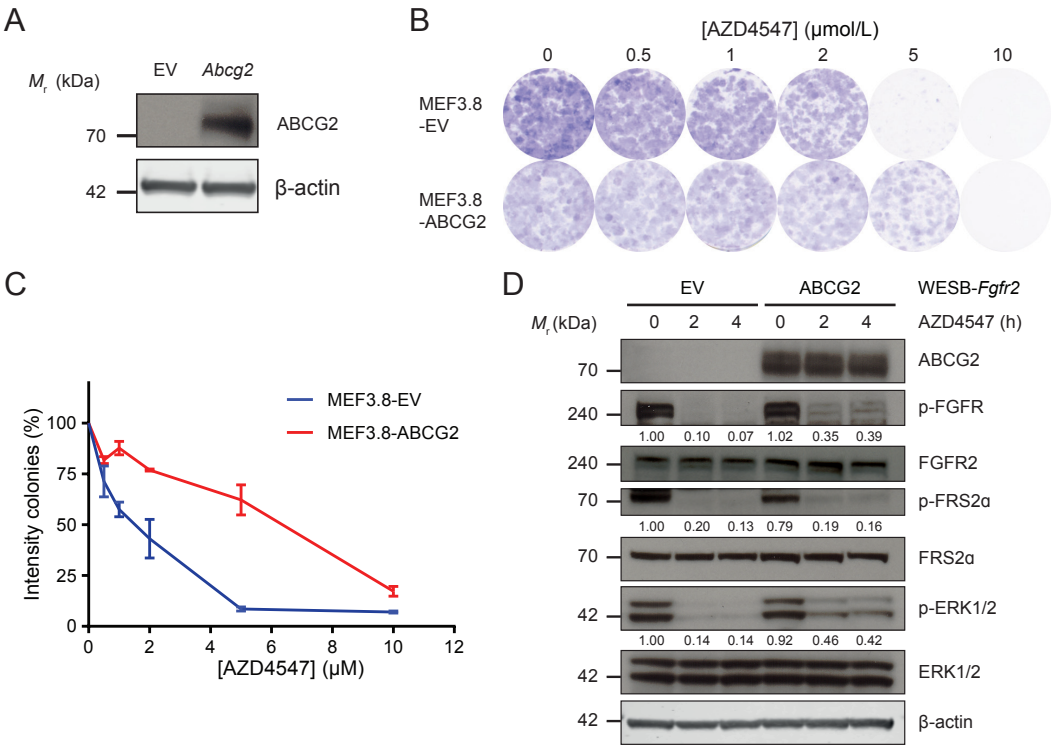


Fig. 6.5. ABCG2 reduces sensitivity to AZD4547 *in vitro* and *in vivo*. **(A)** Representative immunoblot ($n = 3$) for the expression of ABCG2 in MEF3.8-EV and MEF3.8-ABCG2 cells. β-Actin was used as a loading control. **(B and C)** Representative images (B) and quantification (C) of the clonogenic assays (11 days after seeding the cells) of MEF3.8 cells treated with increasing concentrations of AZD4547. Data are mean ± SD of three independent experiments. **(D)** Representative immunoblot ($n = 3$) showing the effect of short-term AZD4547 treatment (100 nmol l^{-1}) in WESB-*Fgfr2* cells with and without ABCG2 expression. β-actin was used as a loading control.

were treated with either vehicle or AZD4547 (12.5 mg kg^{-1} d^{-1}) daily for 30 days when the tumors reached the size of 62.5 mm. Interestingly, the NMRI-nude mice did not show tumor regression upon treatment with AZD4547, in contrast to the previously used FVB syngeneic animals (Supplementary Figure S6.2B), suggesting that an intact immune system might enhance the therapeutic efficacy of AZD4547. Nonetheless, mice with WESB-*Fgfr2*-EV tumors did show stable disease, whereas WESB-*Fgfr2*-ABCG2 tumors progressed during treatment (Figure 6.5E–G). These results show that increased ABCG2 expression also reduces the sensitivity of FGFR2-activated tumors to AZD4547 *in vivo*, confirming that upregulation of this drug efflux transporter can drive resistance to AZD4547.

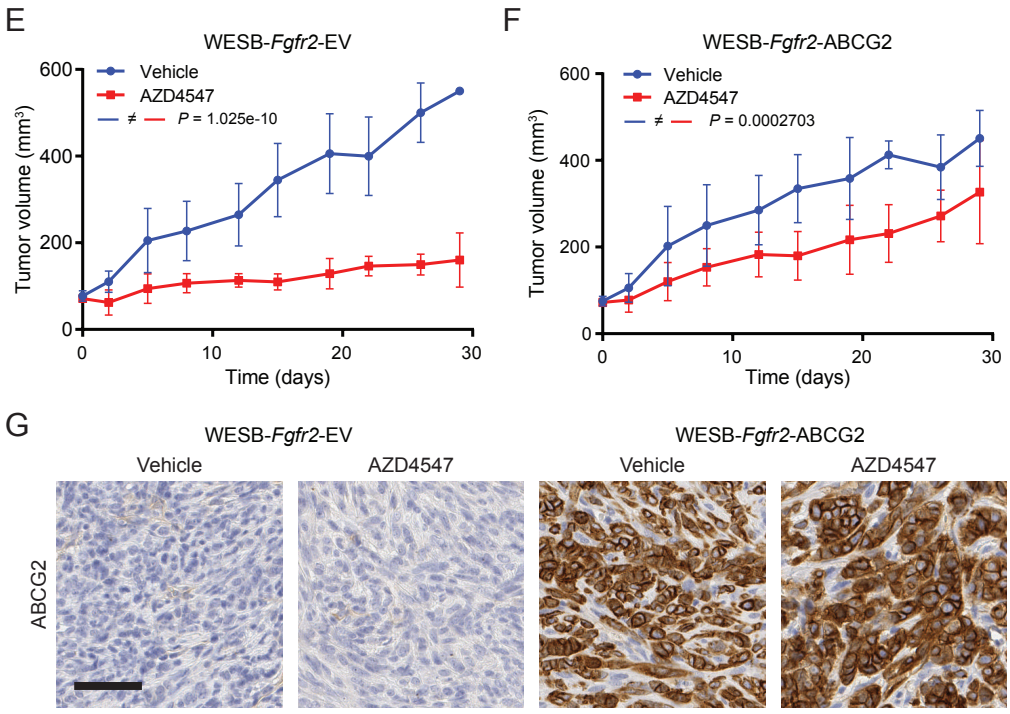


Fig. 6.5. Continued. (E and F) Tumor growth kinetics of WESB-*Fgfr2*-EV (E) and WESB-*Fgfr2*-ABCG2 (F) cells in NMRI-nude female mice under vehicle (blue) or AZD4547 (red) treatment (for 30 consecutive days). Datapoints are mean ± SD and p values were calculated using mixed linear models (details are described in the Supplementary Material). Vehicle-treated WESB-*Fgfr2*-EV versus WESB-*Fgfr2*-ABCG2 tumors, $p = 0.5511$, not significant; AZD4547-treated WESB-*Fgfr2*-EV versus WESB-*Fgfr2*-ABCG2 tumors, $p = 3.983e-05$. **(G)** IHC detection of ABCG2 in sections of WESB-*Fgfr2*-EV and WESB-*Fgfr2*-ABCG2 tumors. Scale bar, 50 μm.

6.5 Discussion

In this work, we performed an SB-based insertional mutagenesis screen in a mouse model of ILC to identify genes that are involved in the development of resistance to FGFR-targeting therapies. As a starting point for this screen, we used SB-induced mILCs, in which we previously identified *Fgfr2* as the most frequently mutated candidate gene[19]. By orthotopically transplanting an SB-induced mILC with activated FGFR signaling into multiple recipient mice, we showed that treatment with the FGFR inhibitor AZD4547 initially results in tumor regression and provides long-term tumor control, but eventually results in acquired treatment resistance. Our mutational analysis of the AZD4547-resistant tumors identified several potential resistance mechanisms, including secondary mutations in FGFR2, inactivation of RASA1, a negative regulator of RAS signaling, and overexpression of MET and the drug-efflux transporter ABCG2. Together, these mechanisms explain acquired resistance to AZD4547 in (21 of 27) tumors (Figure 6.6A). Resistance mechanisms in the remaining six tumors remain to be identified.

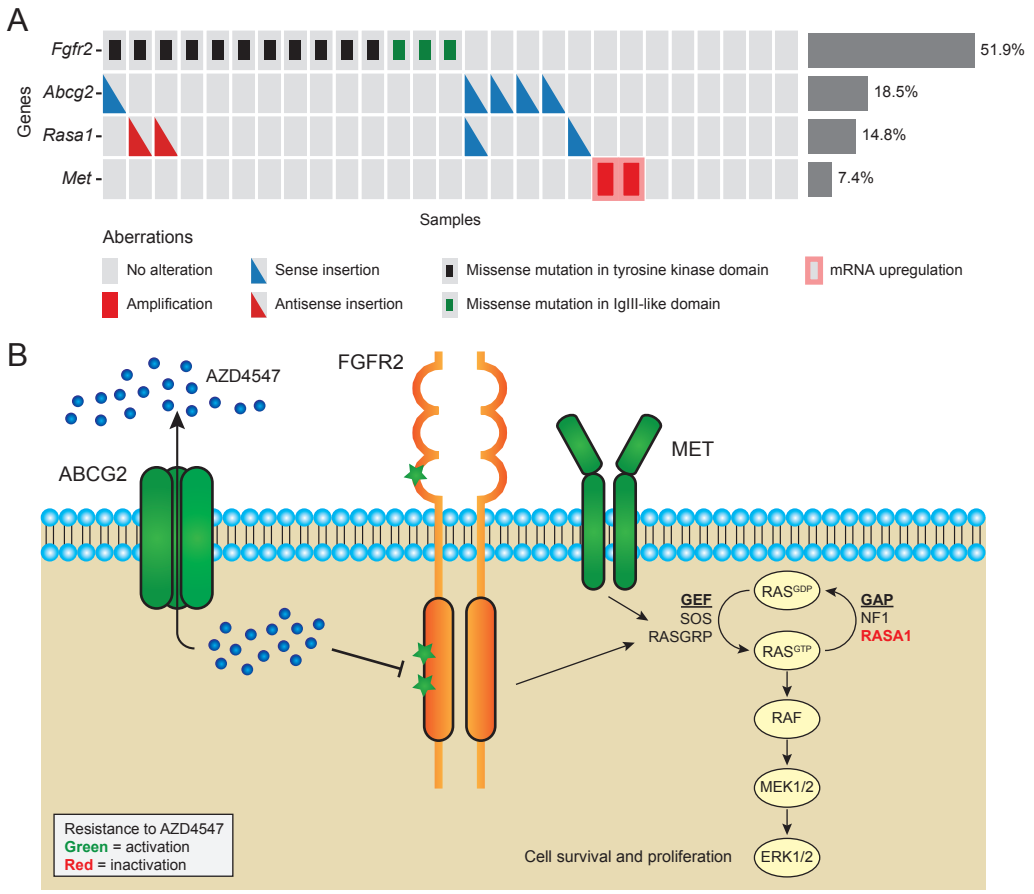


Fig. 6.6. Schematic overview of the resistance mechanisms in AZD4547-resistant tumors identified in this study. (A) Overview of the genetic alterations and gene expression changes in the respective genes identified in the AZD4547-resistant tumors. Red bars, amplification of the respective gene. Blue and red triangles indicate insertions in sense or antisense orientation relative to the gene, respectively. Bars with a red outline indicate the upregulation of the respective genes. Green and black squares represent missense mutations in the IgIII-like domain or the tyrosine kinase domain of FGFR2, respectively. **(B)** Schematic representation of the identified resistance mechanisms in the FGFR signaling pathway. Activating and inactivating events resulting in resistance to AZD4547 are depicted in green and red, respectively.

In line with previous studies in a mouse model of melanoma[30] and *Arf*$^{-/-}$ mice[31], we show that transposon mutagenesis in mice cannot only be used to identify candidate cancer genes, but is also an effective strategy to identify genes involved in *in vivo* drug resistance. In our mutational analysis of the AZD4547-resistant tumors, we exploited the constitutive activity of the SB-mediated insertional mutagenesis system in SB-induced mILCs to identify potential resistance mechanisms in an unbiased, genome-wide fashion. This allowed us to identify two resistance mechanisms (activation of ABCG2 and inactivation of RASA1), which might not have been identified without SB mutagenesis. However, resistance mechanisms that involve specific amino acid substitutions may not be uncovered by transposon mutagenesis, but only arise from spontaneous mutations. A comprehensive characterization of the various

mechanisms of resistance to targeted anticancer therapeutics may therefore require a multipronged approach, combining transposon mutagenesis with other sequencing modalities to identify spontaneous mutations and/or transcriptional changes that may be driving resistance. Given enough sequencing depth, RNA-sequencing based approaches for identifying transposon insertions may be able to provide the most comprehensive analysis from a single dataset[26,32], by allowing detection of transposon insertions, mutations, gene-fusions, and transcriptional changes in RNA-sequencing data. However, targeted DNA-sequencing approaches (as we have used here to detect SB transposon insertions) are likely to yield more detailed detection of insertions and/or mutations with a low frequency, by effectively providing deeper sequencing at a lower cost.

The diverse spectrum of identified resistance mechanisms illustrates the major challenge that (intra-)tumor heterogeneity poses for the prevention of therapy resistance, as we observe multiple resistance mechanisms arising from and within a single (donor) tumor. All of the identified mechanisms center on reactivation of the canonical MAPK–ERK signaling pathway, suggesting that this is a dominant mechanism for overcoming vulnerability to FGFR inhibition (Figure 6.6B). Reactivation of MAPK–ERK signaling has also been identified as a predominant resistance mechanism to EGFR inhibitors[33]. In our analysis of AZD4547-resistant tumors, we observed recurrent alterations in several components of the MAP–ERK pathway, including secondary mutations in FGFR2, overexpression of the MET receptor and inactivation of RASA1. MET overexpression can induce resistance by driving reactivation of signaling pathways downstream of FGFR2, as has previously been shown in the context of FGFR and other RTK inhibitors[2,14,15,34]. Also loss of RASA1, which is a negative regulator of RAS, may cause resistance to FGFR inhibition via reactivation of the MAPK–ERK pathway[35].

Our analysis of the secondary mutations in FGFR2 showed that the majority of these mutations occurred in the tyrosine kinase domain, suggesting that they mainly provide resistance by preventing the inhibitor from binding to the ATP-binding pocket and thereby reactivating the FGFR signaling pathway. This finding agrees with previous studies with other FGFR inhibitors, which identified polyclonal secondary FGFR mutations (including gatekeeper mutations) as a main resistance mechanism to FGFR-targeting treatments[10–12]. Our observations are further supported by studies with other RTK inhibitors, which also describe secondary mutations in the receptor as one of the main resistance mechanisms to tyrosine kinase inhibitors[34].

Our validation of the drug efflux transporter *Abcg2* showed that increased ABCG2 expression can induce resistance by reducing the concentration of AZD4547 within tumor cells, which results in decreased inhibition of FGFR and reactivation of the FGFR signaling pathway. In patients, overexpression of the drug efflux pump MDR1

(encoded by *ABCB1*) has been observed in chemotherapy-resistant ovarian cancer[36]. Our results suggest that drug efflux transporters such as ABCG2 cannot only drive therapy resistance in hematologic malignancies[37], but may also have similar effects on therapy efficacy in solid tumors.

Recent approaches have aimed to overcome resistance to FGFR-targeting therapies either by combining multiple existing RTK inhibitors[14,15], or by designing irreversibly binding inhibitors such as FIIN-2, FIIN-3, and PRN1371[38,39], which cannot be disrupted as easily by secondary mutations in the receptor. However, our results suggest that combining FGFR and MEK/ERK inhibitors might be a more effective strategy, as this prevents reactivation of MAPK–ERK signaling. In addition, to avoid resistance resulting from drug efflux transporters, novel inhibitors should be specifically designed to be poor substrates for common transporters. Alternatively, CRISPR/Cas9 genetic screens could be used to identify synthetic lethal interactions with FGFR inhibitors in the context of FGFR inhibitor-resistant tumors to design rational and more effective combination therapies to overcome drug resistance[40].

In summary, SB insertional mutagenesis in mice is an effective tool to identify mechanisms of drug resistance. A comprehensive analysis of AZD4547-resistant mILCs, in which SB-based mutagenesis is combined with targeted DNA- and RNA-sequencing, allowed us to explain the mechanism of resistance in 78% of the resistant tumors, of which all converged to the reactivation of the canonical MAPK–ERK signaling cascade. Altogether, our findings suggest that FGFR-targeting drugs might be improved by designing FGFR inhibitors that are poor substrates of drug efflux transporters and irreversibly bind to the ATP-binding pocket of the receptor to prevent secondary mutations in the tyrosine kinase domain. In addition, combining these novel FGFR inhibitors with MEK/ERK inhibitors might be an even more effective strategy for preventing resistance to FGFR-targeted therapies.

Acknowledgments We are grateful to Els Wagenaar, Sander Canisius, Chris Doornebal, Eline van der Burg, and Ute Boon for providing technical suggestions and/or assistance with the experiments. We thank Alfred Schinkel (The Netherlands Cancer Institute, Amsterdam) for the MEF3.8 and MEF3.8-ABCG2 cells. We thank Elaine Kilgour, Lindsay Bridgett, and Paul Smith (Oncology IMED, IMED Biotech Unit, AstraZeneca) for providing the AZD4547. We thank the NKI animal facility, RHPC computing 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), Netherlands Genomics Initiative (NGI) grant Zenith 93512009 (J. Jonkers), infrastructural

grant from the National Roadmap for Large-Scale Research Facilities], the EU Seventh Framework Program (Infrafrontier-I3 project 312325 to J. Jonkers), and the European Research Council (ERC Synergy project CombatCancer to J. Jonkers). This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative (e-infra160136 to J. Jonkers). This work is part of the Oncode Institute, which is partly financed by the Dutch Cancer Society (KWF; to J. Jonkers).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Disclosure of potential conflicts of interest C. Phillips and P.D. Smith have ownership interest (including patents) in AstraZeneca PLC. L.F.A. Wessels reports receiving a commercial research grant from Genmab. No potential conflicts of interest were disclosed by the other authors.

Author contributions

Conception and design: S.M. Kas, J.R. de Ruiter, J. Jonkers.
Development of methodology: S.M. Kas, J.R. de Ruiter, K. Schipper, L. Bombardelli, A. Berns, J. Jonkers.
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): S.M. Kas, K. Schipper, A.P. Drenth, R. de Korte-Grimmerink, S. Mahakena, K. van de Wetering.
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): S.M. Kas, J.R. de Ruiter, K. Schipper, C. Phillips, P.D. Smith, S. Klarenbeek, K. van de Wetering, L.F.A. Wessels, J. Jonkers.
Writing, review, and/or revision of the manuscript: S.M. Kas, J.R. de Ruiter, C. Phillips, P.D. Smith, K. van de Wetering, L.F.A. Wessels, J. Jonkers.
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): E. Schut, L. Bombardelli, E. Wientjens, A.P. Drenth, R. de Korte-Grimmerink, A. Berns.
Study supervision: J. Jonkers.

6.6 Supplementary Material

6.6.1 Supplementary experimental details

Cell viability assay WESB cells were seeded in triplicate with 500 cells per well in 96-well plates. After 24 hours, the medium was refreshed with either DMSO (as a control) or AZD4547 with the indicated concentrations. Three days later, cell viability was assayed in an Envision plate reader (Perkin Elmer) using resazurin (cell titer blue; Promega).

Clonogenic assay MEFs were trypsinized and 2000 cells were seeded in 6-well plates. After 24 hours, the medium was refreshed with either DMSO (as a control) or AZD4547 with the indicated concentrations. Eleven days later, the cells were fixed with 4 % formalin in PBS and stained with 0.1 % crystal violet in demineralized water. Quantification was performed by dissolving the crystal violet with 10 % acetic acid in demineralized water and determining the absorbance at 590 nm. The experiment was performed three times.

Competition assay WESB-*Fgfr2* cells were transfected with pX330.pgkpur constructs containing three independent sgRNAs targeting *Rasa1* (s*gRasa1*) or a non-targeting as control (sgNT) using Lipofectamine 2000 (ThermoFisher Scientific 11668027) according to manufacturer's protocol. Transfected cells were selected using puromycin ($4\,\mu g\,ml^{-1}$) for 48 hours. The pX330.pgkpur construct is a modified version of the pX330 backbone[41], which contains a puromycin resistance ORF under the hPGK promoter[42]. The pX330-U6-Chimeric_BB-CBh-hSpCas9 construct was a gift from Dr Feng Zhang (Addgene plasmid #42230). WESB-*Fgfr2* cells containing sgNT or s*gRasa1* were mixed and seeded at a one-to-one ratio on 6-well plates in medium supplemented with 5 % FBS and either DMSO or 2 μmol AZD4547. The medium was refreshed every 4 days and DNA was isolated at days 0, 7, 10 and 12 using the Gentra Puregene genomic DNA isolation kit (Qiagen). PCR amplifications of *Rasa1* exon 2, 6 and 8 were performed with specific primers spanning the target sites and 100 ng to 200 ng DNA template, using the Q5 High-Fidelity PCR kit (NEB M0492). Amplification PCR reactions were diluted 20 times with Milli-Q and subsequently Sanger sequenced using the FW primers. CRISPR/Cas9-induced editing efficacy was quantified using the TIDE algorithm[29]. Cells with only sgNT were used as a negative control in all genomic DNA amplifications and only TIDE outputs with $R^2 > 0.9$ were considered.

sgRNA sequence *Rasa1*-1: 5'-TTATAAGAGAGAGTGATCGG-3'
sgRNA sequence *Rasa1*-2: 5'-CGAGAAGAAGATCCACACGA -3'

sgRNA sequence *Rasa1*-3: 5'-ATCTCCAGGAGTATTATCTG-3'
Rasa1 sgRNA1 PCR FW 5'-TTGTGTTCTCACAGACCTGAAT-3' (557 bp)
Rasa1 sgRNA1 PCR RV 5'-TCAATCTGTGATCTCCAAGCC-3' (557 bp)
Rasa1 sgRNA2 PCR FW 5'-TGTAGGCAAGAGAGCCAAATTA-3' (697 bp)
Rasa1 sgRNA2 PCR RV 5'-GTTCAAGGCCAGTCTGATCTAC-3' (697 bp)
Rasa1 sgRNA3 PCR FW 5'-GAGTTCTTTCAGAGAGCGAAGG-3' (406 bp)
Rasa1 sgRNA3 PCR RV 5'-GAGTTCTTTCAGAGAGCGAAGG-3' (406 bp)

Preparation of membrane vesicles and vesicular transport assays Membrane vesicles from Sf9 cells were obtained after infection with a control or a human ABCG2-containing baculovirus at a multiplicity of infection of 1[22]. After incubation at 27 °C for 3 days, cells were harvested by centrifugation at $500xg$ for 5 min. Cells were then resuspended in ice-cold hypotonic buffer (0.5 mmol sodium phosphate and 0.1 mmol EDTA, 7.4 pH) supplemented with a protease inhibitor cocktail (Roche, Basel, Switzerland) and incubated at 4 °C for 90 min under constant agitation. Thereafter, the cell lysate was homogenized using a tight-fitting Dounce homogenizer. Next, cell debris and nuclei were removed by slow-speed centrifugation at $500xg$ (4 °C for 10 min). The supernatant was collected and centrifuged at 4 °C at $100,000xg$ for 40 min. The membrane pellet was resuspended in TS buffer (50 mmol Tris-HCl and 250 mmol sucrose, 7.4 pH) and passed through a 27-gauge needle 25 times. The vesicles were dispensed in aliquots, snap-frozen in liquid nitrogen, and stored at −80 °C until use. Vesicular transport assays were performed using the rapid filtration method as previously described[22,23]. Briefly, ABCG2 or control Sf9 membrane vesicles containing 20 μg of protein were incubated with 1 mmol [^{3}H]-MTX in 50 μl of TS buffer in the presence of 4 mmol ATP or AMP, 10 mmol $MgCl_2$, 10 mmol creatine phosphate, and 100 g of creatine kinase/ml. After 10 minutes, 40 μl of the reaction mixture was diluted in 200 μl of ice-cold TS buffer and immediately filtrated using a MultiScreen$_{HTS}$ vacuum manifold in combination with Multiscreen$_{HTS}$ FB 96-well filter plates (Millipore, Bedford, MA). Membranes were washed four times with 200 μl of ice-cold phosphate-buffered saline and the radioactivity retained on the membranes was counted by liquid scintillation counting.

Immunohistochemistry Tissues were formalin-fixed and paraffin-embedded (FFPE) by routine procedures. Immunohistochemical stainings of MET, ABCG2 (BCRP) and cleaved Caspase-3 were processed as previously described[20,43]. The following primary antibodies were used for the respective proteins: MET (1:100, R&D Systems AF527), BCRP (1:400, Abcam 24115) and cleaved Caspase-3 (1:400, CST 9661). Citrate buffer was used as antigen retrieval for MET and BCRP. TRIS/EDTA 9.0 pH was used for cleaved Caspase-3. Immunohistochemical staining of IGF1R was performed on a Discovery Ultra autostainer (Ventana Medical Systems). Briefly, paraffin sections were cut at 3 μm, heated at 75 °C for 28 minutes and deparaffinised

in the instrument with EZ prep solution (Ventana Medical Systems). Heat-induced antigen retrieval was carried out using Cell Conditioning 1 (CC1, Ventana Medical Systems) for 64 minutes at 95 °C. IGF1R was detected using clone G11 (Ready-to-use, 16 minutes at 37 °C, Ventana Medical Systems), bound antibody was detected using the OMap anti-Rb HRP (Ventana Medical Systems) for 12 minutes after which the ChromoMap DAB Kit (Ventana Medical Systems) was applied. Slides were counterstained with Hematoxylin and Bluing Reagent (Ventana Medical Systems). All slides were digitally processed using the Aperio ScanScope (Aperio, Vista, CA, USA) and captured using ImageScope software version 12.3.2.8013 (Aperio). Cleaved Caspase-3 and ABCG2 immunohistochemical stainings were reviewed and scored by a veterinary pathologist (Sjoerd Klarenbeek) in a blinded manner. 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. The number of cleaved Caspase-3 positive cells were counted in four independent fields (0.34 mm^2) per tumor and the average number of positive cells per mm^2 was calculated. Necrotic areas in these tumors were excluded from the analysis.

Immunoblotting Protein lysates were made using lysis buffer (25 mmol Tris-HCl 7.6 pH, 150 mmol NaCl, 1 % NP-40, 1 % sodium deoxycholate, 0.1 % SDS in Milli-Q) complemented with protease and phosphatase inhibitors (Roche) and quantified using the BCA protein assay kit (Pierce). Equal amounts of proteins were separated on a 4 % to 12 % Bis-Tris gradient gel (Invitrogen) and transferred overnight onto nitrocellulose membrane (Bio-Rad) in 1x transfer buffer (25 mmol Tris, 2 mol Glycine, 20 % methanol in demineralized water). Membranes were blocked in 5 % w/v bovine serum albumin (BSA) in PBS-T (7.5 pH, 0.005 % Tween-20 in demineralized water) and incubated overnight with the primary antibodies in 5 % w/v BSA in PBS-T. Membranes were washed three times and incubated with the secondary antibodies goat anti-rabbit-HRP (1:2000, Dako P0448), rabbit anti-mouse-HRP (1:5000, Dako P0260), rabbit anti-rat-HRP (1:2000, Invitrogen 61-9520) or donkey anti-mouse IRDye 680nm (1:5000, Li_COR 926-32222) in 5 % w/v BSA in PBS-T. Stained membranes were washed three times in PBS-T and then developed using ECL (Pierce 32209), ECL 2 Substrate (Pierce 80196) or captured using the Li-Cor Odyssey Infrared Imaging System and analyzed using Odyssey Application software version 3.0.16. The intensities of the bands were quantified using ImageJ software version 2.0.0-rc-65/1.52b.

RNA sequencing and analysis Illumina TruSeq mRNA libraries were generated and sequenced with 50-65 base single reads on a HiSeq 2500 using V4 chemistry (Illumina Inc., San Diego) as previously described by Boelens *et al.*[44]. The resulting reads were trimmed using Cutadapt[45] (version 1.13) to remove any remaining adapter sequences, filtering reads shorter than 20 bp after trimming to ensure good

mappability. The trimmed reads were aligned to the GRCm38 reference genome using STAR[46] (version 2.5.3a). QC statistics from Fastqc[47] (version 0.11.5) and the above-mentioned tools were collected and summarized using Multiqc[48] (version 1.1). Gene expression counts were generated by featureCounts[49] (version 1.5.2) using gene definitions from Ensembl GRCm38 version 89. Normalized expression values were obtained by correcting for differences in sequencing depth between samples using DESeq's median-of-ratios approach and then log-transforming the normalized counts[50]. Differentially expressed genes were identified using DIDS[28] (version 0.10.1), using a threshold of $p < 0.05$ for statistical significance. Variants in RTKs and genes involved in downstream FGFR signaling were called using Vardict[51] (version 2017.04.18) and annotated using Ensembl VEP[52] (version 90.7). The entire analysis pipeline (including the alignment, expression estimation and variant calling) was implemented using Snakemake[53] and is freely available on GitHub*.

Amplification of SB transposon insertions Transposon insertions were amplified following a previously described tagmentation-based DNA sequencing protocol[24]. Briefly, recombinant Tn5 transposase was prepared as previously described by Picelli *et al.*[54], and diluted in glycerol buffer to a final concentration of 3.7 μmol. The Tn5-adapter complex was prepared by incubating 30 minutes at 37 °C equimolar amounts of Tn5 and separately annealed adapters pairs Tn5ME-A+ 3'dT5P-oligo and Tn5ME-B+3'dT5P-oligo as previously described by David L. Stern[24]. Each tagmentation reaction was prepared by combining 2 μl of genomic DNA (100 ng in total), 4 μl of 5x TAPS-PEG buffer[54], 1 μl of Tn5-adapters complex and 13 μl water and incubated for 10 minutes at 55 °C. Tn5 was stripped off from DNA by adding 4 μl of 0.2 % SDS and incubating the reaction 5 min at 55 °C. The enrichment PCR was performed by combining 3 μl of tagmented DNA, 1 μl of enrichment primer at 1 μmol, 6 μl water and 10 μl Phusion Flash 2x mix (cat. #F548L, Thermo Scientific, Waltham, MA USA). PCR1 was performed by combining 5 μl of enrichment PCR reaction, 8 μl water, 1 μl of P5-indexed primer and 1 μl of transposon-specific primer SB-PCR1 and 10 μl Phusion Flash 2x mix. PCR2 was performed by combining 2 μl of PCR1 reaction, 8 μl water, 1 μl of P7-indexed primer + 1 μl FC2 primer and 10 μl Phusion Flash 2x mix. Equal amounts of PCR2 products were pooled and run on an agarose gel. Fragments above 600 basepairs were excised from the gel, purified on Qiagen columns and eluted in water. The pool of tagmented DNA was sequenced with 150 base paired-end reads on a MiSeq 300 using the micro kit v2 reagents (Illumina Inc., San Diego). The following primer sequences and PCR cycler settings were used:

Primer sequences

SB-enrich: GCTTGTGGAAGGCTACTCGAAATGTTTGACCC

*https://github.com/jrderuiter/snakemake-rnaseq

SB_pcr1: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGTATGTAAACTTCCGACTTCAAC
FC2: AATGATACGGCGACCACCGA
Tn5ME-A-adaptor: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
3'dT5P-oligo: CTGTCTCTTATACACATCTGAC (must be 5' phosphorylated and 3' OH blocked by an inverted thymindine)
Tn5ME-B-adaptor: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
A-idx-i5-1: AATGATACGGCGACCACCGAGATCTACACTAATGTGGTCGTCGGCAGCGTC
A-idx-i5-2: AATGATACGGCGACCACCGAGATCTACACGCACTCAGTCGTCGGCAGCGTC
A-idx-i5-3: AATGATACGGCGACCACCGAGATCTACACAACAGCGGTCGTCGGCAGCGTC
A-idx-i5-4: AATGATACGGCGACCACCGAGATCTACACCCATATGATCGTCGGCAGCGTC
A-idx-i5-5: AATGATACGGCGACCACCGAGATCTACACTGGAAAGCTCGTCGGCAGCGTC
A-idx-i5-6: AATGATACGGCGACCACCGAGATCTACACAGCAACGCTCGTCGGCAGCGTC
A-idx-i5-7: AATGATACGGCGACCACCGAGATCTACACCCCTTGCATCGTCGGCAGCGTC
A-idx-i5-8: AATGATACGGCGACCACCGAGATCTACACCCCTCTTGTCGTCGGCAGCGTC
A-idx-i5-9: AATGATACGGCGACCACCGAGATCTACACTTCGAGCCTCGTCGGCAGCGTC
A-idx-i5-10: AATGATACGGCGACCACCGAGATCTACACAGTAGTTATCGTCGGCAGCGTC
A-idx-i5-11: AATGATACGGCGACCACCGAGATCTACACAGAAAGTGTCGTCGGCAGCGTC
A-idx-i5-12: AATGATACGGCGACCACCGAGATCTACACTGCCGGTATCGTCGGCAGCGTC
A-idx-i5-13: AATGATACGGCGACCACCGAGATCTACACGCAAACTGTCGTCGGCAGCGTC
A-idx-i5-14: AATGATACGGCGACCACCGAGATCTACACGGTTGAGATCGTCGGCAGCGTC
A-idx-i5-15: AATGATACGGCGACCACCGAGATCTACACATAGATGTTCGTCGGCAGCGTC
A-idx-i5-16: AATGATACGGCGACCACCGAGATCTACACCAAACATTTCGTCGGCAGCGTC
A-idx-i5-17: AATGATACGGCGACCACCGAGATCTACACCTGAGCGTTCGTCGGCAGCGTC
A-idx-i5-18: AATGATACGGCGACCACCGAGATCTACACTAGTCTCTTCGTCGGCAGCGTC
B-idx-i7-1: CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGG
B-idx-i7-2: CAAGCAGAAGACGGCATACGAGATATCCGCATGTCTCGTGGGCTCGG
B-idx-i7-3: CAAGCAGAAGACGGCATACGAGATATTGAAGTGTCTCGTGGGCTCGG
B-idx-i7-4: CAAGCAGAAGACGGCATACGAGATCTCTGCGTGTCTCGTGGGCTCGG
B-idx-i7-5: CAAGCAGAAGACGGCATACGAGATGATACGCAGTCTCGTGGGCTCGG
B-idx-i7-6: CAAGCAGAAGACGGCATACGAGATTACGTTCGGTCTCGTGGGCTCGG
B-idx-i7-7: CAAGCAGAAGACGGCATACGAGATTGAATCCTGTCTCGTGGGCTCGG
B-idx-i7-8: CAAGCAGAAGACGGCATACGAGATGGCTATAAGTCTCGTGGGCTCGG
B-idx-i7-9: CAAGCAGAAGACGGCATACGAGATCACAACCTGTCTCGTGGGCTCGG

PCR cycler settings

Enrichment PCR (2-step): 98 °C 30 s; 45 cycles of (98 °C 8 s, 72 °C 35 s)
PCR1: 98 °C 30 s; 15-18 cycles of (98 °C 8 s, 63 °C 5 s, 72 °C 30 s)
PCR2: 98 °C 30 s; 15-18 cycles of (98 °C 8 s, 63 °C 5 s, 72 °C 30 s)

Insertion site analysis Insertion sites were identified using the taqmap pipeline in PyIM[†] (version 0.3.0). Briefly, this analysis pipeline first trimmed the TaqMap paired-end reads using Cutadapt[45] (version 1.12) to remove any matepairs not containing the transposon and to remove any Nextera transposase sequences. The remaining mate pairs were aligned to the mm10 reference genome using Bowtie2[55] (version 2.3.0). After the alignment, redundant sequences mapping to the same genomic location and belonging to the same tumor were collapsed into a single insertion. To avoid issues with slight variations in the alignment, insertions from the same sample that occurred within 10 bp of each other were collapsed into a single insertion. Insertions were assigned to genes using the rule-based mapping approach[56] with the 'SB' preset and gene definitions from Ensembl GRCm38 89. Support scores were calculated as the number of unique mate pairs supporting a given insertion. Relative support scores (used as a proxy for clonality) were calculated by normalizing support scores to the highest support score of the corresponding sample.

Genes associated with *de novo* resistance were identified by selecting genes that did not have any insertions in untreated tumors (vehicle-treated tumors and the donor tumor) and had insertions in at least two AZD4547-resistant samples. The *de novo* candidate genes were then ranked by their frequency of occurrence. Genes associated with intrinsic resistance were selected by performing a Welch's t-test between the clonality scores of insertions in the AZD4547-resistant tumors and the vehicle treated tumors, as well as determining the difference between the means of both groups (to ensure a minimum effect size). Candidate genes were selected by filtering for genes with a difference in means > 0.1 and a t-test p value < 0.25, after which the candidates were ranked by their mean differences.

Validation of the endogenous *Fgfr2-Tbc1d1* fusion The *Fgfr2-Tbc1d1* fusion was detected in WESB-*Fgfr2* cells as previously described[26]. WESB cells were used as negative control. The following primer sequences were used:

Fgfr2 FW: 5'-TGGCCAGGGATATCAACAAC-3'
Tbc1d1 RV: 5'-CCAGGCTGTGAGAAGGATTT-3'

***Met* qPCR copy number analysis** DNA was isolated from AZD4547-resistant tumors and wild-type FVB spleen as a control. The qPCR was performed on a Quantstudio 6 flex Real-Time PCR system (Applied Biosystems) using low ROX SYBR green (Bioline) with *Met* and β-catenin (*Ctnnb1*) specific primers.

Met FW: 5'-TCTCGGAGCCACAAACTACA-3'
Met RV: 5'-GCAGTCCCGACAAGGTAAAC-3'

[†]https://github.com/jrderuiter/pyim

Ctnnb1 FW: 5'-TCAGGGCAGGTGAAACTGTA-3'
Ctnnb1 RV: 5'-GACTCCCAGCACACTGAACTTA-3'

The relative copy number levels of *Met* and *Ctnnb1* were quantified using a five-point standard curve. The *Met* relative copy number was normalized to the *Ctnnb1* relative copy number for each sample and subsequently normalized to the normalized relative abundance of wild-type FVB spleen.

Statistical analysis The effect of AZD4547 treatment on tumor growth of WESB-*Fgfr2*-EV and WESB-*Fgfr2*-ABCG2 established tumors was tested using mixed linear models. Prior to this analysis, the tumor size measurements were scaled so that each tumor's size at the first time point was equal to 1. Inspection of the individual tumor growth curves suggested an approximately linear increase in tumor size over time. The exact growth rate, i.e. the slope of the growth curve, showed some inter-tumor variability. Therefore, we modeled growth rate using a fixed effect population-level slope β, a random effect tumor-level slope b_i to account for inter-tumor variability, and a fixed effect term γ for the interaction of time and treatment to model the effect of treatment on tumor growth. Additional inter-tumor variability is allowed by random intercepts a_i, which complement the fixed effect intercept α. This leads to the following model formulation for the size of tumor i as a function of time and treatment:

$$\text{Tumor size}_i = \alpha + a_i + (\beta + b_i) * \text{time} + \gamma * \text{time} * \text{treatment}_i$$

The significance of the treatment effect was established using an ANOVA comparing the models with and without the interaction term γ.

6.6.2 Supplementary Figures

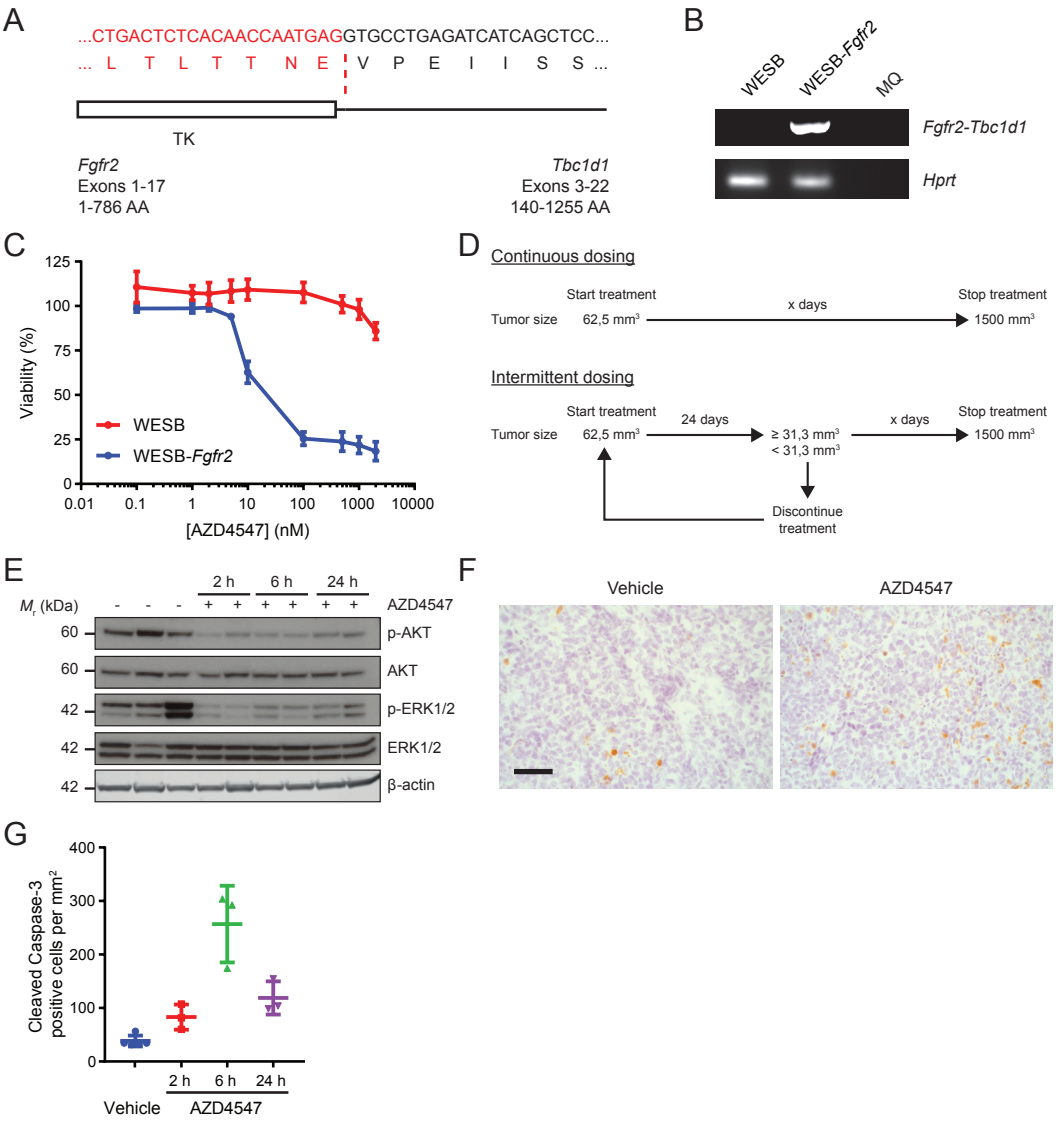


Fig. S6.1. Validation of the Fgfr2-Tbc1d1 gene fusion and the effects of AZD4547 treatment on viability and downstream FGFR signaling in WESB cells. **(A)** Schematic overview of the *Fgfr2-Tbc1d1* gene fusion showing the genomic and amino acid sequences of *Fgfr2* (red) and *Tbc1d1* (black). **(B)** Validation of the *Fgfr2-Tbc1d1* fusion using targeted PCR on cDNA of WESB-*Fgfr2* tumor-derived cells. WESB cells and MQ (= Milli-Q) were used as negative controls. **(C)** The viability of WESB and WESB-*Fgfr2* cells after 72 hours of AZD4547 treatment. Data are mean ± s.d. of three independent experiments in which the cells were seeded in triplicate. **(D)** Schematic representation of the treatment schedules with AZD4547 in tumor-bearing mice. **(E)** Immunoblot showing the effect of a single vehicle/AZD4547 treatment on total and phosphorylated AKT and ERK1/2 in established WESB-*Fgfr2* tumors. β-actin was used as a loading control. **(F and G)** Representative images (F) and quantification (G) of immunohistochemical detection of cleaved Caspase-3 in tumors after a single vehicle ($n = 5$) or AZD4547 treatment ($n = 3$ for each indicated time point after administration of AZD4547). Scale bar, 50 μm.

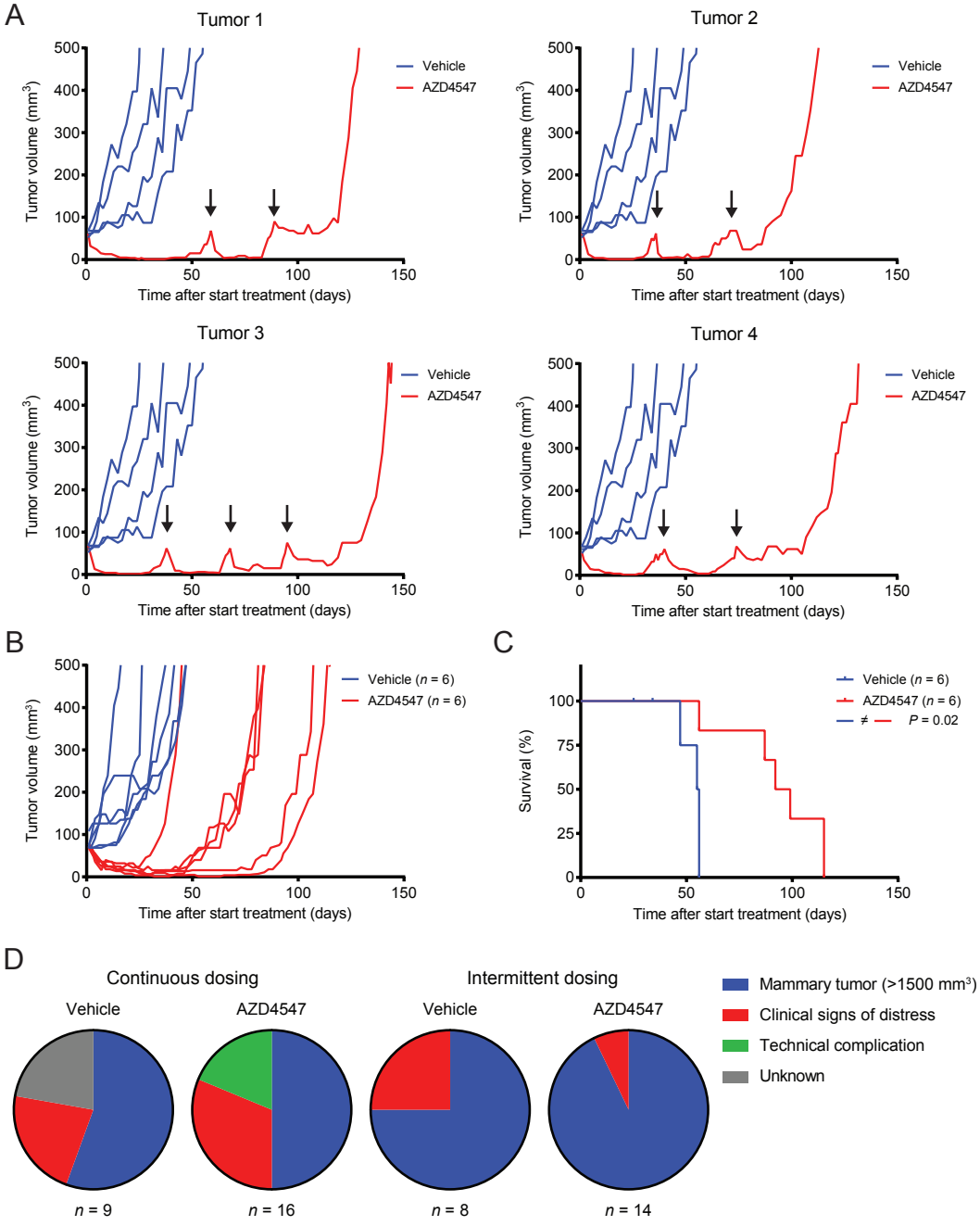


Fig. S6.2. **Tumor growth kinetics of established tumors during treatment and causes of death observed under the different treatment schedules.** **(A)** Tumor growth kinetics of vehicle (blue) and four different AZD4547-treated (red) tumors. The black arrows indicate the start of a treatment cycle of 24 days. **(B)** Tumor growth kinetics of WESB-*Fgfr2* cells injected into the mammary gland of wild-type syngeneic recipient female mice under continuous treatment with vehicle (blue) or AZD4547 (red). **(C)** Kaplan-Meier curve showing the overall survival of tumor-bearing mice upon continuous dosing with vehicle (blue) or AZD4547 (red). p value was calculated using a Mantel–Cox test. **(D)** Comparison of the causes of death observed in the different treatment schedules.

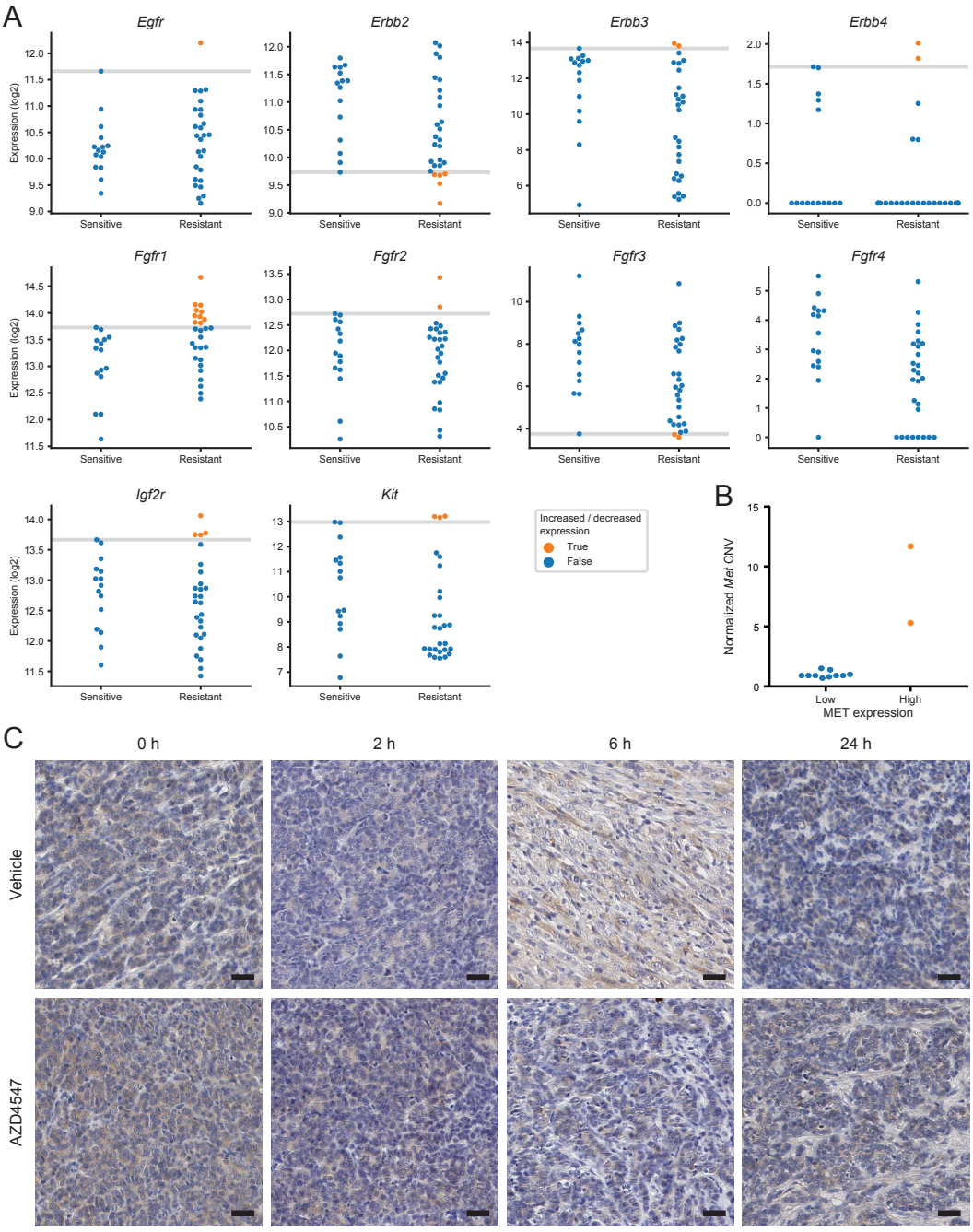


Fig. S6.3. Differential gene expression analysis of multiple RTK family members and the status of MET in vehicle-treated, AZD4547-sensitive and resistant tumors. (A) Differential gene expression analysis of all FGFR-, EGFR-, IGF-related family members and *Kit* in AZD4547-sensitive ($n = 15$) and resistant ($n = 27$) tumors using DIDS. **(B)** Quantitative PCR (qPCR) of *Met* copy number variation (CNV) in AZD4547-resistant tumors with high ($n = 2$) or low ($n = 10$) MET expression. *Met* CNV is normalized to *Ctnnb1* (control). **(C)** Representative images of MET immunohistochemical stainings of WESB-*Fgfr2* tumors after a single vehicle- or AZD4547-treatment. Scale bar, 25 μm.

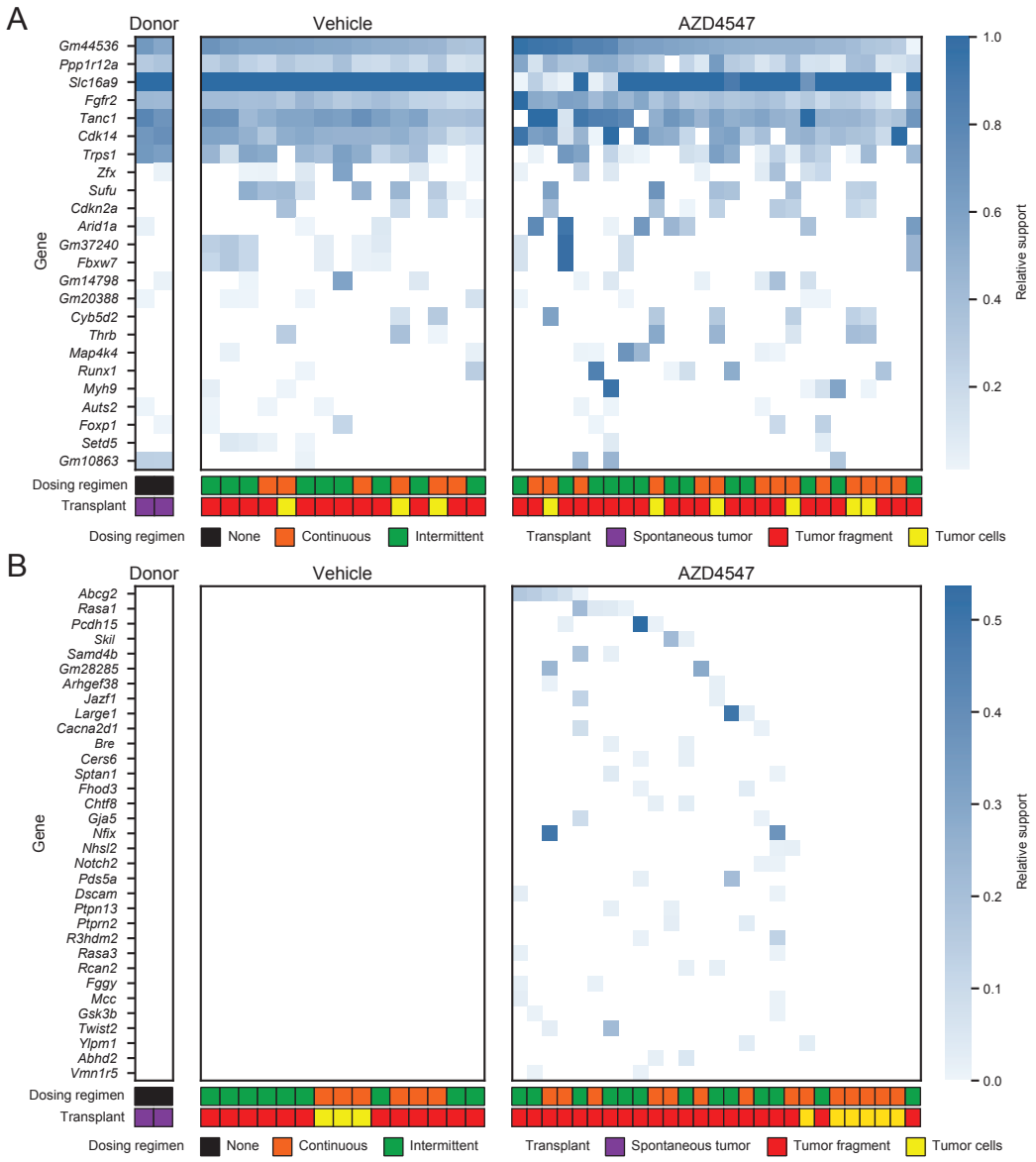


Fig. S6.4. Overview of insertions in genes across donor, vehicle-treated and AZD4547-resistant tumors. (A) Overview of insertions in genes across donor (two technical replicates), vehicle-treated ($n = 15$) and AZD4547-resistant ($n = 27$) tumors. Relative support scores of insertions (depicted in blue) were calculated by counting the number of mate pairs supporting the insertion and normalizing these support scores to the highest score of the corresponding sample. **(B)** Overview of insertions in genes that were mutated in two or more AZD4547-resistant ($n = 27$) tumors but were not mutated in any of the donor or vehicle tumors. The relative support of the insertions within each sample is depicted in blue.

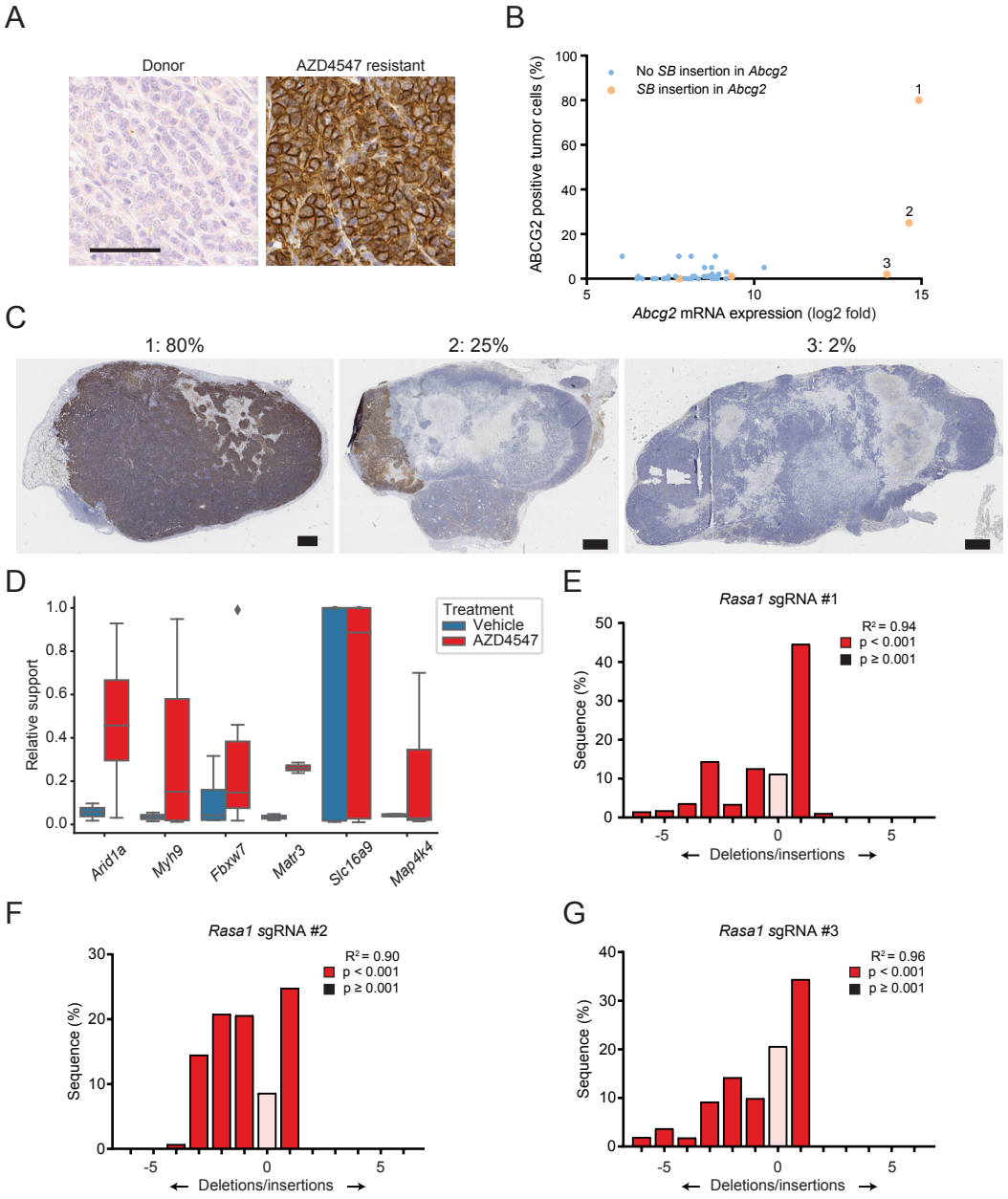


Fig. S6.5. Immunohistochemical stainings of ABCG2 in correlation to mRNA expression in AZD4547-resistant tumors, boxplots of genes that show intrinsic resistance to AZD4547 and efficacy of *Rasa1* gene editing in WESB-*Fgfr2* cells. **(A)** Representative immunohistochemical stainings of ABCG2 in the WESB-*Fgfr2* donor tumor and an AZD4547-resistant tumor. Scale bar, 50 μm. **(B)** Correlation plot of mRNA and protein expression of ABCG2 in AZD4547-resistant tumors ($n = 27$) with and without insertions in *Abcg2*. **(C)** Immunohistochemical stainings of ABCG2 in the three AZD4547-resistant tumors with the highest *Abcg2* mRNA expression as depicted in (B). Scale bar, 1 mm. **(D)** Boxplot showing the relative support scores for genes that show enriched support in AZD4547-resistant compared to vehicle-treated tumors (differential support score > 0.1 and $p < 0.25$, t-test). **(E-G)** Representative spectrum of insertions and deletions (indels) for each sgRNA targeting *Rasa1* in WESB-*Fgfr2* cells as quantified using the TIDE algorithm.

6.6.3 Supplementary Tables (available online)

Tab. S6.1. **Overview of the sequenced tumor samples, including sample IDs, treatment information and the type of transplanted material.**

Tab. S6.2. **Mutations (SNVs) identified for RTKs and genes involved in the MAPK-ERK signaling pathway in the donor, vehicle- and AZD4547-treated tumors.**

Tab. S6.3. **Overview of insertions identified in the donor, vehicle- and AZD4547-treated tumors.**

6.7 References

[1] David M Ornitz and Nobuyuki Itoh. “The fibroblast growth factor signaling pathway”. In: *Wiley Interdisciplinary Reviews: Developmental Biology* 4.3 (2015), pp. 215–266 (cit. on p. 190).

[2] Irina S Babina and Nicholas C Turner. “Advances and challenges in targeting FGFR signalling in cancer”. In: *Nature Reviews Cancer* 17.5 (2017), pp. 318–332 (cit. on pp. 190, 191, 194, 207).

[3] Rut Porta, Roberto Borea, Andreia Coelho, et al. “FGFR a promising druggable target in cancer: Molecular biology and new drugs”. In: *Critical reviews in oncology/hematology* 113 (2017), pp. 256–267 (cit. on pp. 190, 191).

[4] Young Kwang Chae, Keerthi Ranganath, Peter S Hammerman, et al. “Inhibition of the fibroblast growth factor receptor (FGFR) pathway: the current landscape and barriers to clinical application”. In: *Oncotarget* 8.9 (2017), p. 16052 (cit. on p. 191).

[5] Fabrice André, Thomas Bachelot, Mario Campone, et al. “Targeting FGFR with dovitinib (TKI258): preclinical and clinical data in breast cancer”. In: *Clinical Cancer Research* 19.13 (2013), pp. 3693–3702 (cit. on p. 191).

[6] Lucia Nogova, Lecia V Sequist, Jose Manuel Perez Garcia, et al. “Evaluation of BGJ398, a fibroblast growth factor receptor 1-3 kinase inhibitor, in patients with advanced solid tumors harboring genetic alterations in fibroblast growth factor receptors: results of a global phase I, dose-escalation and dose-expansion study”. In: *Journal of Clinical Oncology* 35.2 (2017), pp. 157–165 (cit. on p. 191).

[7] Josep Tabernero, Rastislav Bahleda, Rodrigo Dienstmann, et al. “Phase I dose-escalation study of JNJ-42756493, an oral pan–Fibroblast growth factor receptor inhibitor, in patients with advanced solid tumors”. In: *Journal of Clinical Oncology* 33.30 (2015), pp. 3401–3408 (cit. on p. 191).

[8] Milind M. Javle, Rachna T. Shroff, Andrew Zhu, et al. “A phase 2 study of BGJ398 in patients (pts) with advanced or metastatic FGFR-altered cholangiocarcinoma (CCA) who failed or are intolerant to platinum-based chemotherapy”. In: *Journal of Clinical Oncology* 34.4_suppl (2016), pp. 335–335 (cit. on p. 191).

[9] Paul K Paik, Ronglai Shen, Michael F Berger, et al. “A phase Ib open-label multicenter study of AZD4547 in patients with advanced squamous cell lung cancers”. In: *Clinical Cancer Research* (2017) (cit. on p. 191).

[10] Lipika Goyal, Supriya K Saha, Leah Y Liu, et al. “Polyclonal Secondary FGFR2 Mutations Drive Acquired Resistance to FGFR Inhibition in Patients with FGFR2 Fusion–Positive Cholangiocarcinoma”. In: *Cancer Discovery* 7.3 (2017), pp. 252–263 (cit. on pp. 191, 197, 199, 207).

[11] V Chell, K Balmanno, AS Little, et al. “Tumour cell responses to new fibroblast growth factor receptor tyrosine kinase inhibitors and identification of a gatekeeper mutation in FGFR3 as a mechanism of acquired resistance”. In: *Oncogene* 32.25 (2013), pp. 3059–3070 (cit. on pp. 191, 197, 207).

[12] Sara A Byron, Huaibin Chen, Andreas Wortmann, et al. “The N550K/H mutations in FGFR2 confer differential resistance to PD173074, dovitinib, and ponatinib ATP-competitive inhibitors”. In: *Neoplasia* 15.8 (2013), 975IN28–988IN30 (cit. on pp. 191, 197, 207).

[13] Maria Teresa Herrera-Abreu, Alex Pearson, James Campbell, et al. “Parallel RNA interference screens identify EGFR activation as an escape mechanism in FGFR3-mutant cancer”. In: *Cancer Discovery* 3.9 (2013), pp. 1058–1071 (cit. on pp. 191, 197).

[14] SM Kim, H Kim, MR Yun, et al. “Activation of the Met kinase confers acquired drug resistance in FGFR-targeted lung cancer therapy”. In: *Oncogenesis* 5.7 (2016), e241 (cit. on pp. 191, 197, 207, 208).

[15] Jinjia Chang, Shanshan Wang, Zhe Zhang, et al. “Multiple receptor tyrosine kinase activation attenuates therapeutic efficacy of the fibroblast growth factor receptor 2 inhibitor AZD4547 in FGFR2 amplified gastric cancer”. In: *Oncotarget* 6.4 (2015), p. 2009 (cit. on pp. 191, 197, 207, 208).

[16] Jun Wang, Oliver Mikse, Rachel G Liao, et al. “Ligand-associated ERBB2/3 activation confers acquired resistance to FGFR inhibition in FGFR3-dependent cancer cells”. In: *Oncogene* 34.17 (2015), pp. 2167–2177 (cit. on pp. 191, 197).

[17] Timothy R Wilson, Jane Fridlyand, Yibing Yan, et al. “Widespread potential for growth-factor-driven resistance to anticancer kinase inhibitors”. In: *Nature* 487.7408 (2012), p. 505 (cit. on p. 191).

[18] Fred Harbinski, Vanessa J Craig, Sneha Sanghavi, et al. “Rescue screens with secreted proteins reveal compensatory potential of receptor tyrosine kinases in driving cancer growth”. In: *Cancer Discovery* 2.10 (2012), pp. 948–959 (cit. on p. 191).

[19] Sjors M Kas, Julian R de Ruiter, Koen Schipper, et al. “Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular breast carcinoma”. In: *Nature Genetics* 49.8 (2017), pp. 1219–1230 (cit. on pp. 191, 192, 194, 200, 205).

[20] Chris W Doornebal, Sjoerd Klarenbeek, Tanya M Braumuller, et al. “A preclinical mouse model of invasive lobular breast cancer metastasis”. In: *Cancer Research* 73.1 (2013), pp. 353–363 (cit. on pp. 191, 211).

[21] John D Allen, Sonja C Jackson, and Alfred H Schinkel. “A mutation hot spot in the Bcrp1 (Abcg2) multidrug transporter in mouse cell lines selected for Doxorubicin resistance”. In: *Cancer Research* 62.8 (2002), pp. 2294–2299 (cit. on pp. 192, 203).

[22] Koen van de Wetering and Sunny Sapthu. “ABCG2 functions as a general phytoestrogen sulfate transporter in vivo”. In: *The FASEB Journal* 26.10 (2012), pp. 4014–4024 (cit. on pp. 192, 211).

[23] Koen van de Wetering, Alexander Burkon, Wouter Feddema, et al. “Intestinal breast cancer resistance protein (BCRP)/Bcrp1 and multidrug resistance protein 3 (MRP3)/Mrp3 are involved in the pharmacokinetics of resveratrol”. In: *Molecular Pharmacology* 75.4 (2009), pp. 876–885 (cit. on pp. 192, 211).

[24] David L Stern. “Tagmentation-based mapping (TagMap) of mobile DNA genomic insertion sites”. In: *BioRxiv* (2017), p. 037762 (cit. on pp. 193, 213).

[25] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, et al. “UCSF Chimera—a visualization system for exploratory research and analysis”. In: *Journal of Computational Chemistry* 25.13 (2004), pp. 1605–1612 (cit. on p. 193).

[26] Julian R de Ruiter, Sjors M Kas, Eva Schut, et al. “Identifying transposon insertions and their effects from RNA-sequencing data”. In: *Nucleic Acids Research* 45.12 (2017), pp. 7064–7077 (cit. on pp. 193, 194, 207, 215).

[27] Yi-Mi Wu, Fengyun Su, Shanker Kalyana-Sundaram, et al. “Identification of targetable FGFR gene fusions in diverse cancers”. In: *Cancer Discovery* 3.6 (2013), pp. 636–647 (cit. on p. 194).

[28] Jorma J de Ronde, Guillem Rigaill, Sven Rottenberg, Sjoerd Rodenhuis, and Lodewyk FA Wessels. “Identifying subgroup markers in heterogeneous populations”. In: *Nucleic Acids Research* 41.21 (2013), e200–e200 (cit. on pp. 197, 213).

[29] Eva K Brinkman, Tao Chen, Mario Amendola, and Bas van Steensel. “Easy quantitative assessment of genome editing by sequence trace decomposition”. In: *Nucleic Acids Research* 42.22 (2014), e168–e168 (cit. on pp. 201, 202, 210).

[30] Daniele Perna, Florian A Karreth, Alistair G Rust, et al. “BRAF inhibitor resistance mediated by the AKT pathway in an oncogenic BRAF mouse melanoma model”. In: *Proceedings of the National Academy of Sciences* 112.6 (2015), E536–E545 (cit. on p. 206).

[31] Emilie A Chapeau, Agnieszka Gembarska, Eric Y Durand, et al. “Resistance mechanisms to TP53-MDM2 inhibition identified by in vivo piggyBac transposon mutagenesis screen in an Arf-/- mouse model”. In: *Proceedings of the National Academy of Sciences* 114.12 (2017), pp. 3151–3156 (cit. on p. 206).

[32] Nuri A Temiz, Branden S Moriarity, Natalie K Wolf, et al. “RNA sequencing of Sleeping Beauty transposon-induced tumors detects transposon-RNA fusions in forward genetic cancer screens”. In: *Genome Research* 26.1 (2016), pp. 119–129 (cit. on p. 207).

[33] Julia Rotow and Trever G Bivona. “Understanding and targeting resistance mechanisms in NSCLC”. In: *Nature Reviews Cancer* 17.11 (2017), p. 637 (cit. on p. 207).

[34] Matthew J Niederst and Jeffrey A Engelman. “Bypass mechanisms of resistance to receptor tyrosine kinase inhibition in lung cancer”. In: *Science Signaling* 6.294 (2013), re6–re6 (cit. on p. 207).

[35] Takuo Hayashi, Patrice Desmeules, Roger S Smith, et al. “RASA1 and NF1 are preferentially co-mutated and define a distinct genetic subset of smoking-associated non-small cell lung carcinomas sensitive to MEK inhibition”. In: *Clinical Cancer Research* (2017), p. 2343 (cit. on p. 207).

[36] Ann-Marie Patch, Elizabeth L Christie, Dariush Etemadmoghadam, et al. “Whole-genome characterization of chemoresistant ovarian cancer”. In: *Nature* 521.7553 (2015), p. 489 (cit. on p. 208).

[37] Karthika Natarajan, Yi Xie, Maria R Baer, and Douglas D Ross. “Role of breast cancer resistance protein (BCRP/ABCG2) in cancer drug resistance”. In: *Biochemical Pharmacology* 83.8 (2012), pp. 1084–1103 (cit. on p. 208).

[38] Li Tan, Jun Wang, Junko Tanizaki, et al. “Development of covalent inhibitors that can overcome resistance to first-generation FGFR kinase inhibitors”. In: *Proceedings of the National Academy of Sciences* 111.45 (2014), E4869–E4877 (cit. on p. 208).

[39] Eleni Venetsanakos, Ken A Brameld, Vernon T Phan, et al. “The irreversible covalent fibroblast growth factor receptor inhibitor PRN1371 exhibits sustained inhibition of FGFR after drug clearance”. In: *Molecular Cancer Therapeutics* 16.12 (2017), pp. 2668–2676 (cit. on p. 208).

[40] Diede Brunen and René Bernards. “Drug therapy: Exploiting synthetic lethality to improve cancer therapy”. In: *Nature Reviews Clinical Oncology* 14.6 (2017), p. 331 (cit. on p. 208).

[41] Le Cong, F Ann Ran, David Cox, et al. “Multiplex genome engineering using CRISPR/Cas systems”. In: *Science* (2013), p. 1231143 (cit. on p. 210).

[42] Tim Harmsen, Sjoerd Klaasen, Henri van de Vrugt, and Hein te Riele. “DNA mismatch repair and oligonucleotide end-protection promote base-pair substitution distal from a CRISPR/Cas9-induced DNA break”. In: *Nucleic acids research* 46.6 (2018), pp. 2945–2955 (cit. on p. 210).

[43] Linda Henneman, Martine H van Miltenburg, Ewa M Michalak, et al. “Selective resistance to the PARP inhibitor olaparib in a mouse model for BRCA1-deficient metaplastic breast cancer”. In: *Proceedings of the National Academy of Sciences* 112.27 (2015), pp. 8409–8414 (cit. on p. 211).

[44] Mirjam C Boelens, Micha Nethe, Sjoerd Klarenbeek, et al. “PTEN loss in E-cadherin-deficient mouse mammary epithelial cells rescues apoptosis and results in development of classical invasive lobular carcinoma”. In: *Cell Reports* 16.8 (2016), pp. 2087–2101 (cit. on p. 212).

[45] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet. journal* 17.1 (2011), pp. 10–12 (cit. on pp. 212, 215).

[46] Alexander Dobin, Carrie A Davis, Felix Schlesinger, et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21 (cit. on p. 213).

[47] S. Andrews. *FastQC: a quality control tool for high throughput sequence data*. 2017. URL: `http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc` (cit. on p. 213).

[48] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (2016), pp. 3047–3048 (cit. on p. 213).

[49] Yang Liao, Gordon K Smyth, and Wei Shi. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7 (2014), pp. 923–930 (cit. on p. 213).

[50] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (2014), pp. 1–21 (cit. on p. 213).

[51] Zhongwu Lai, Aleksandra Markovets, Miika Ahdesmaki, et al. “VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research”. In: *Nucleic Acids Research* 44.11 (2016), e108–e108 (cit. on p. 213).

[52] William McLaren, Laurent Gil, Sarah E Hunt, et al. “The ensembl variant effect predictor”. In: *Genome Biology* 17.1 (2016), p. 122 (cit. on p. 213).

[53] Johannes Köster and Sven Rahmann. “Snakemake—a scalable bioinformatics workflow engine”. In: *Bioinformatics* 28.19 (2012), pp. 2520–2522 (cit. on p. 213).

[54] Simone Picelli, Åsa K Björklund, Björn Reinius, et al. “Tn5 transposase and tagmentation procedures for massively scaled sequencing projects”. In: *Genome Research* 24.12 (2014), pp. 2033–2040 (cit. on p. 213).

[55] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4 (2012), pp. 357–359 (cit. on p. 215).

[56] Johann de Jong, Jeroen de Ridder, Louise van der Weyden, et al. “Computational identification of insertional mutagenesis targets for cancer gene discovery”. In: *Nucleic Acids Research* 39.15 (2011), e105–e105 (cit. on p. 215).

7

Selective loss of PARG restores PARylation and counteracts PARP inhibitor-mediated synthetic lethality

Ewa Gogola[1,14], Alexandra A. Duarte[1,14,*], Julian R. de Ruiter[1,2,14,*], Wouter W. Wiegant[6], Jonas A. Schmid[7], Roebi de Bruijn[1,2,14], Dominic I. James[8], Sergi Guerrero Llobet[9], Daniel J. Vis[2,14], Stefano Annunziato[1,14], Bram van den Broek[3], Marco Barazas[1,14], Ariena Kersbergen[5], Marieke van de Ven[4], Madalena Tarsounas[10], Donald J. Ogilvie[8], Marcel van Vugt[9], Lodewyk F.A. Wessels[2,14], Jirina Bartkova[11,12], Irina Gromova[11], Miguel Andújar-Sánchez[13], Jiri Bartek[11,12], Massimo Lopes[7], Haico van Attikum[6], Piet Borst[5], Jos Jonkers[1,14,17], Sven Rottenberg[1,15,16,17]

[1] Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[2] Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[3] Division of Cell Biology and BioImaging Facility, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[4] Mouse Clinic for Cancer and Aging (MCCA), Preclinical Intervention Unit, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[5] Division of Molecular Oncology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[6] Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands
[7] Institute of Molecular Cancer Research, University of Zurich, Zurich, Switzerland
[8] Drug Discovery Unit, Cancer Research UK, University of Manchester, Manchester, UK
[9] Department of Medical Oncology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands
[10] CRUK/MRC Oxford Institute for Radiation Oncology, University of Oxford, Oxford, UK
[11] Danish Cancer Society Research Center, Copenhagen, Denmark
[12] Karolinska Institute, Department of Medical Biochemistry and Biophysics, Division of Genome Biology, Science for Life Laboratory, Stockholm, Sweden
[13] Pathology Department, Complejo Hospt. Univ. Insular Materno Infantil, Las Palmas, Spain
[14] Cancer Genomics Netherlands, Oncode Institute, Amsterdam, the Netherlands
[15] Institute of Animal Pathology, Vetsuisse Faculty, University of Bern, Bern, Switzerland
[16] Lead contact
[17] Correspondence: j.jonkers@nki.nl (J.J.), sven.rottenberg@vetsuisse.unibe.ch (S.R.)
[*] These authors contributed equally to this work

Published in Cancer Cell, 11 June 2018
Volume 33, Issue 6, Pages 1078-1093

7.1 Summary

Inhibitors of poly(ADP-ribose) (PAR) polymerase (PARPi) have recently entered the clinic for the treatment of homologous recombination (HR)-deficient cancers. Despite the success of this approach, drug resistance is a clinical hurdle, and we poorly understand how cancer cells escape the deadly effects of PARPi without restoring the HR pathway. By combining genetic screens with multi-omics analysis of matched PARPi-sensitive and -resistant *Brca2*-mutated mouse mammary tumors, we identified loss of PAR glycohydrolase (PARG) as a major resistance mechanism. We also found the presence of PARG-negative clones in a subset of human serous ovarian and triple-negative breast cancers. PARG depletion restores PAR formation and partially rescues PARP1 signaling. Importantly, PARG inactivation exposes vulnerabilities that can be exploited therapeutically.

7.2 Significance

To explore defects in the DNA damage response in cancer therapy, exciting opportunities have been achieved using the 'synthetic lethal' approach. A successful example is the development of PARP inhibitors to kill cancer cells that are defective in HR; e.g., due to lack of function of BRCA1 or BRCA2. Thus, there is a real opportunity to cure patients with HR-deficient cancers if we overcome the hurdle of drug resistance. At present, it is largely unknown how tumor cells escape PARP inhibition without restoring BRCA2-mediated HR. Here, we show that loss of PARG governs PARPi resistance in HR-deficient tumors by restoring PARP1 signaling. Importantly, inactivation of PARG results in vulnerabilities that can be exploited to combat resistance.

7.3 Introduction

Defects in the DNA damage response (DDR) result in genomic instability and are implicated in many types of cancer[1]. DDR alterations are responsible for the accumulation of mutations that result in tumorigenesis, and they can be specifically exploited for targeted cancer therapy. A prime example of such a tailored approach is the application of poly (ADP-ribose) polymerase (PARP) inhibitors (PARPi) in the treatment of tumors defective in homology-directed DNA repair due to BRCA1 or BRCA2 inactivation[2,3]. PARP1, a founding member of the PARP family, is a nuclear protein functioning in various cellular processes, including chromatin remodeling and DNA damage repair[4]. Upon DNA damage, PARP1 is rapidly recruited to DNA nicks where it induces the synthesis of protein-conjugated polymers of ADP-ribose

(poly(ADP-ribose) [PAR]). PARP1 itself is a prime target of PARylation and the resulting PAR chains serve as a platform for the recruitment of downstream repair factors. PARylation is a transient and reversible modification, as it is counteracted by the activity of PAR glycohydrolase (PARG), which degrades PAR[5]. Inhibition of PARP1 leads to the accumulation of unresolved single-strand breaks (SSBs)[6]. Moreover, several PARPi trap PARP1 onto chromatin[7,8], resulting in the collapse of replication forks (RF) that hit trapped PARP1. This leads to DNA double-strand breaks (DSBs) and cells depend on BRCA1/2-mediated repair to resolve these DSBs in an error-free way. Hence, PARP1 inhibition causes synthetic lethality in tumors with defects in homologous recombination (HR)[2,3]. Indeed, this lethality was also observed in mouse models for *BRCA1/2*-mutated breast cancer[9,10] as well as in patients with *BRCA1/2* mutations who developed breast or ovarian cancer[11–13]. On the basis of these positive clinical results, three different PARPi were recently approved as a monotherapy for the treatment of *BRCA1/2*-mutated ovarian cancers[14].

Drug resistance often follows the introduction of therapeutics in the clinic, and unfortunately PARPi are no exception[11,12]. Using cell lines and mouse models, several mechanisms of PARPi resistance have been identified, including upregulation of the P-glycoprotein (P-gp; also known as ABCB1) drug efflux transporter[9,10] and restoration of HR activity (reviewed by Annunziato *et al.*[15]). While the clinical significance of P-gp-driven resistance remains controversial, HR restoration has been observed in human tumors that re-established BRCA1/2 function[16,17]. Nevertheless, secondary *BRCA1/2* mutations explain only some of the cases of PARPi resistance[18]. The requirement of BRCA1 for HR activity can be bypassed by the loss of the 53BP1-RIF1-REV7 pathway, as shown by various studies[15]. In contrast, there is no evidence that HR can be rescued in the absence of BRCA2, suggesting that BRCA2-deficient tumors employ distinct, HR-independent pathways to overcome PARPi toxicity.

Little is known thus far about HR-independent resistance to PARPi. Loss of the drug target PARP1 has been described as a mechanism of resistance in HR-proficient cells[19], but this cannot explain resistance in the PARPi target group, since PARP1 loss causes synthetic lethality of *BRCA1/2*-mutated cells[2,3]. In this study, we set out to determine how cells with an irreversible and complete defect in the HR pathway develop PARPi resistance.

7.4 Results

7.4.1 Functional genetic screens identify loss of PARG as a PARPi resistance factor

To identify HR-independent mechanisms of PARPi resistance, we carried out functional genetic screens in two types of *in vitro* cultures that we derived from *Brca2*$^{-/-}$*;p53*$^{-/-}$ mouse mammary tumors from *K14cre;p53*$^{F/F}$*;Brca2*$^{F/F}$ (KB2P) mice: two-dimensional (2D) tumor cell lines (KB2P1.21, KB2P3.4) and three-dimensional (3D) cancer organoids (ORG-KB2P26S.1)[9,20,21]. In these cells we introduced a library of 1,976 short hairpin RNA (shRNA) constructs targeting 391 DDR-related genes (on average five shRNAs/gene)[22]. The cells were then selected for 3 weeks with the PARPi olaparib or AZD2461 (Figure 7.1A) at a concentration lethal to the parental cells (data not shown). Sequencing of PARPi-surviving populations revealed a reproducible enrichment of multiple hairpins targeting PARG. The strong effect of PARG depletion is reflected by the overall top score of *Parg* among all positively selected genes, as determined by the MAGeCK (Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout) algorithm[23] (Figure 7.1B-C; Table S7.1). We applied the same screening approach to a cell line isolated from BRCA-proficient mouse mammary tumors from *K14cre;p53*$^{F/F}$ (KP) mice[9] and also identified *Parg* among the top outliers. In fact, *Parg* was the only common hit in both BRCA-deficient and -proficient screens (Figure 7.1D-E). In contrast, shRNAs targeting PARP1 were only enriched in the BRCA-proficient KP3.33 cells (Figure 7.1C-E), providing functional evidence that PARP1 loss confers PARPi resistance in BRCA-proficient cells, presumably by preventing PARP1 trapping, but not in BRCA2-deficient cells that depend on PARP1 for survival.

7.4.2 PARG is frequently lost in PARPi-Resistant KB2P mouse mammary tumors

Although high-throughput genetic screens are powerful tools for the identification of gene candidates, *in vitro* conditions do not fully recapitulate the complexity of drug response observed in real tumors. We therefore generated a panel of KB2P mouse mammary tumors that had acquired PARPi resistance *in vivo*. For this purpose, 21 individual spontaneous KB2P carcinomas were orthotopically transplanted into multiple syngeneic mice to allow differential treatment of the original donor tumor. Upon outgrowth, the tumors were either treated with vehicle control or with the PARPi AZD2461 (Figure 7.2A). As expected, KB2P tumors were initially highly sensitive to PARPi treatment but eventually developed drug resistance (Figure 7.2B-

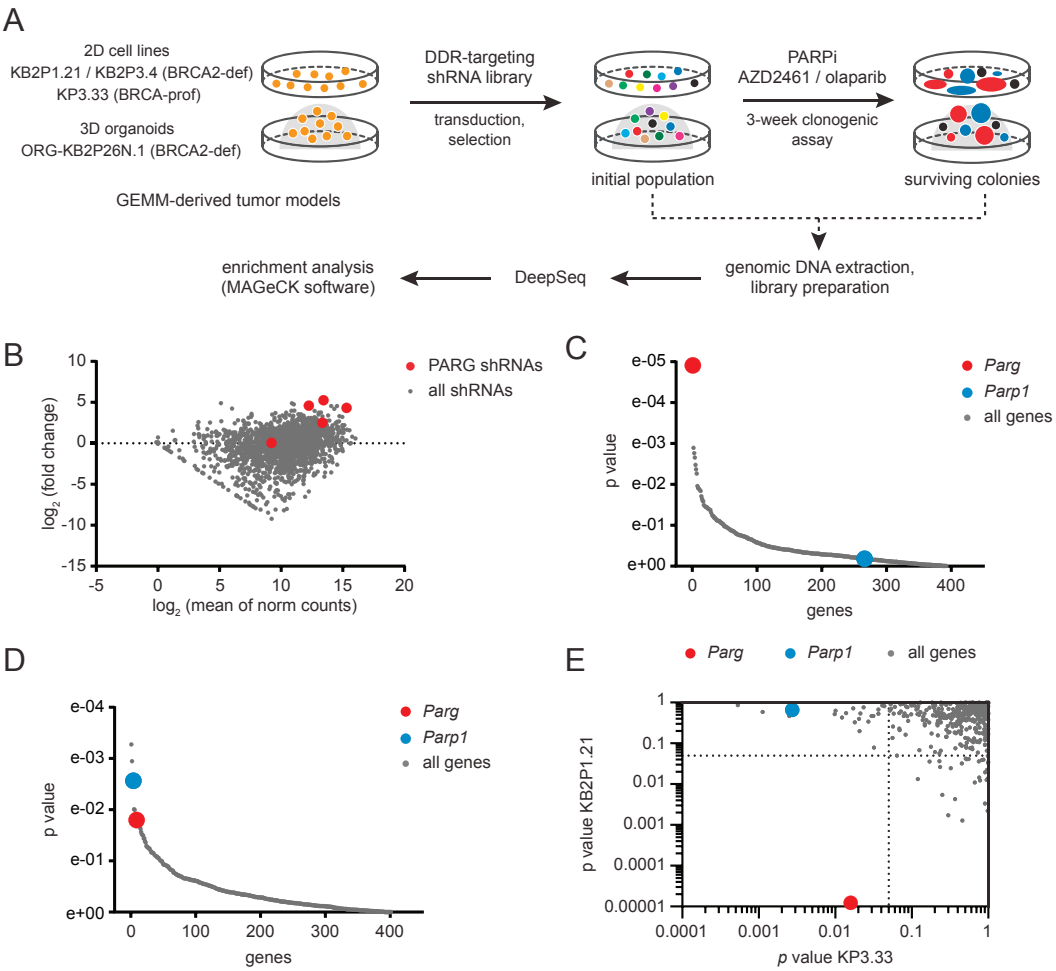


Fig. 7.1. Functional shRNA-based screens in BRCA2-deficient and -proficient cells identify PARG as PARPi resistance factor. (A) Outline of the functional shRNA screen. **(B)** Log ratio (fold change) versus abundance (mean of norm counts) plot representing the screening outcome in KB2P1.21 cells treated with AZD2461. **(C-D)** Distribution of the one-sided p value (gene enrichment) for all 391 genes targeted by the shRNA-based library in KB2P1.21 cells (C) and KP3.33 (D) cells upon PARPi treatment. **(E)** Comparison of the screening outcome between indicated cell lines; dotted grid line indicates p value = 0.05. All p values were generated per gene with MAGeCK software; each screen was performed and analyzed in triplicate. See also Table S7.1.

C). The observed resistance cannot be explained by BRCA2 restoration, which is prevented by the irreversible intragenic deletion in *Brca2*, nor by upregulation of P-gp (Figure S7.1A), because of the low affinity of AZD4261 to this transporter[24,25].

Our extensive *in vivo* studies yielded a unique collection of matched PARPi-naive ($n = 21$) and PARPi-resistant tumors ($n = 34$; for some of the donors more than one resistant tumor was generated). We have recently shown that the resistance phenotype is stable upon transplantation into allografts[21,26]. We now used this collection of tumors to identify genetic factors contributing to PARPi resistance. For this purpose, we generated transcriptome (RNA sequencing [RNA-seq]) and

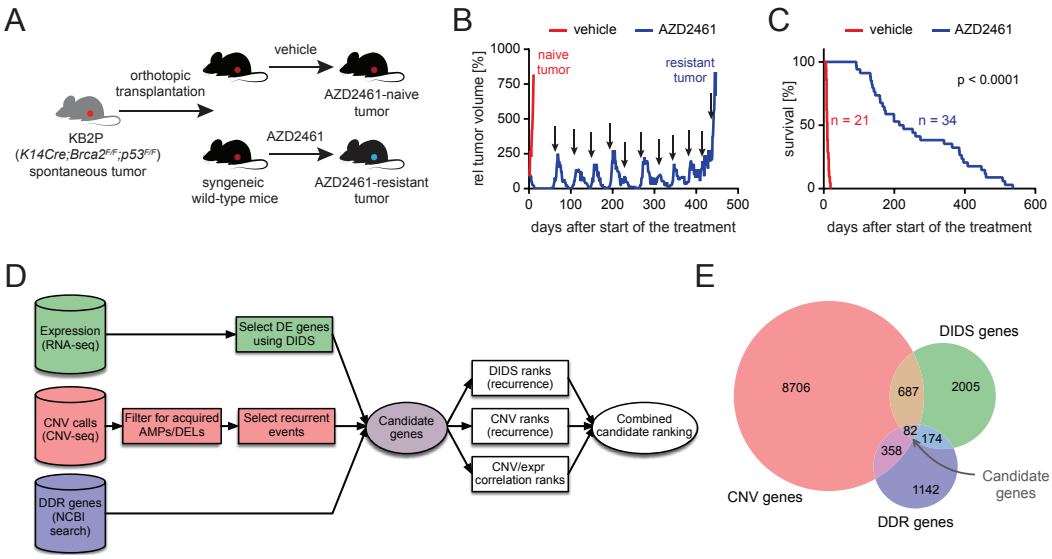


Fig. 7.2. **PARG is frequently lost in KB2P Tumors that acquired PARPi resistance *in vivo*. (A)** Generation of matched PARPi-naive and -resistant KB2P tumors. **(B)** Treatment response of individual KB2P tumor treated with either vehicle or AZD2461, orally for 28 consecutive days. Treatment was resumed when the tumors reached a size of 100% (initial volume at the start of the treatment) and the treatment cycles were repeated until acquired resistance (black arrows mark the beginning of repeated cycles). Graph shows relative tumor volume as a function of time. **(C)** Kaplan-Meier curve showing overall survival of mice bearing KB2P tumors treated with either vehicle or AZD2461. Log rank (Mantel-Cox) p value is indicated. **(D)** Flowchart illustrating the steps of multi-omics approach used for the discovery of resistance factors in a panel of KB2P tumors. **(E)** Venn diagram showing overlap of potential gene candidates identified within indicated datasets.

DNA copy number variation (CNV) sequencing (CNV-seq) data for all tumors and carried out an integrative analysis of naive versus resistant samples (Figure 7.2D-E). First, we identified differentially expressed genes using the DIDS (detection of imbalanced differential signal) algorithm (cutoff $p < 0.05$), which is specifically designed for the detection of subgroup markers in heterogeneous populations[27]. In parallel, we selected acquired copy number events, present only in resistant, but not in naive, samples. Since KB2P tumors exhibit high levels of genomic instability and accumulate many genetic alterations, we decided to focus on DDR-related genes, as their contribution to the PARPi response is most plausible. We generated a list of approximately 1,800 genes that have been implicated in DDR processes (Table S7.2) and combined it with the significant hits from the DIDS and CNV analyses. The 82 genes that survived these selection criteria were ranked based on their recurrence or correlation between expression and CNV data. To integrate these rankings, we used three different aggregation statistics: mean aggregation, Stuart aggregation, and robust rank aggregation (RRA)[28]. Consistent with our *in vitro* screens, all three algorithms placed *Parg* at the top of the list of gene candidates (Table S7.3). *Parg* also ranked among the top outliers in a non-curated,

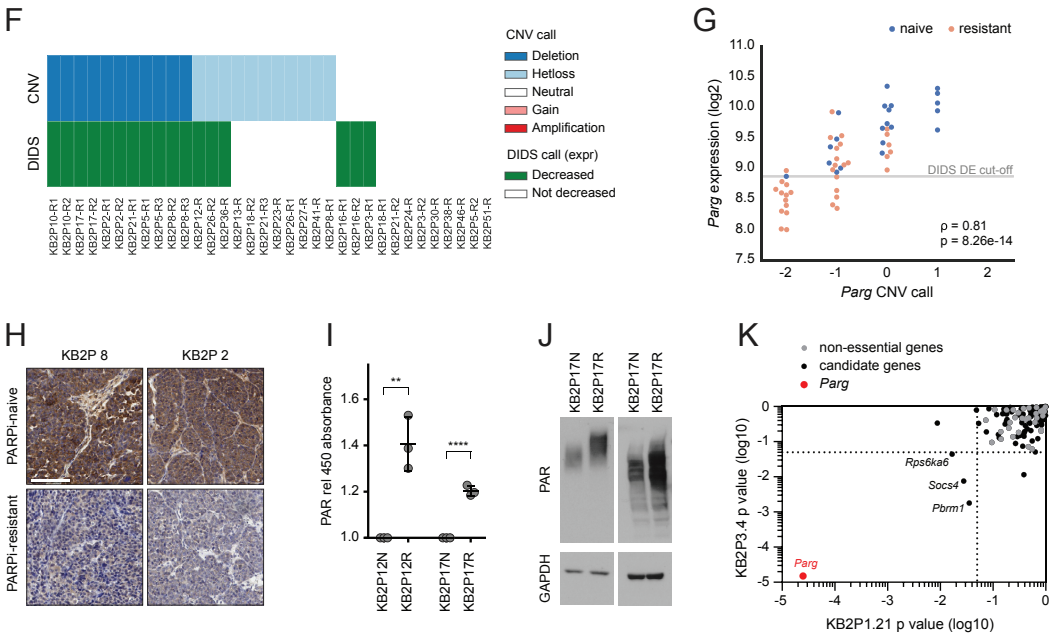


Fig. 7.2. **Continued.** **(F)** Overview of genomic alterations in *Parg* acquired by a panel of 34 KB2P PARPi-resistant tumors (KB2Px-Ry: x, original donor identification [ID] number; y, ID of individual resistant tumors derived from the same donor tumor). **(G)** Correlation between *Parg* expression and copy number estimation for the whole panel of KB2P tumors. Spearman correlation coefficient (ρ), p value, and expression threshold generated by DIDS algorithm (gray line) are indicated. **(H)** Representative images of PARG IHC staining in KB2P tumors; scale bar, 100 μm. **(I)** ELISA-based PARG activity assay in tumor organoids (N, naive; R, resistant); $^{**}p < 0.01$, $^{****}p < 0.0001$ (two-tailed t-test); experiment repeated three times, data presented as mean $\pm$ SD of replicates. **(J)** Western blot analysis of PAR in tumor-derived organoids; data representative of two independent experiments. **(K)** Enrichment analysis of the secondary genetic screen in KB2P cells: comparison of the p values for all genes targeted, shown for the cell lines indicated; one-sided p value was determined by the MAGeCK algorithm; the screen was performed in triplicate. See also Figure S7.1 and Tables S7.1 to S7.4.

genome-wide comparison (Table S7.3). In our panel of 34 PARPi-resistant tumors, we observed decreased expression of *Parg* in 17 tumors, and acquired copy number loss of the *Parg* locus in 22 tumors (11 deletions, 11 heterozygous loss events), with a substantial overlap between both datasets (Figure 7.2F-G and S1B). The difference in PARG levels between PARPi-naive and -resistant tumors was also confirmed by immunohistochemistry (IHC) (Figure 7.2H). Blinded semi-quantitative analysis of the PARG staining revealed a significant difference between resistant versus naive samples ($p < 0.015$, Mann-Whitney U test). This was further validated using an ELISA assay in which we monitored the loss of biotinylated PAR from immobilized histones and thereby directly measured the relative activity of endogenous PARG in 3D cancer organoids derived from PARG-deficient PARPi-resistant tumors and PARPi-sensitive controls (Figure 7.2I). As expected, PARPi-resistant organoids showed reduced ability to degrade synthetic PAR (Figure 7.2I) and overall exhibited elevated levels of endogenous PAR (Figure 7.2J).

To investigate the contribution of the other candidate genes to PARPi resistance, we performed a secondary genetic loss-of-function screen using an shRNA library targeting the identified candidates and 32 non-essential genes as internal controls (Table S7.4). *Parg* was again identified as a top outlier by the enrichment analysis in both KB2P cell lines (Figure 7.2K and Table S7.1), suggesting that loss of PARG is one of the strongest mechanisms involved in PARPi resistance in our model.

To test whether PARPi-induced loss of PARG is specific to KB2P tumors, we also studied genetic alterations in *Parg* in our previously described collection of PARPi-naive and -resistant BRCA1-deficient mammary tumors from *K14cre;Brca1F/F;Trp53F/F* (KB1P) mice[24]. Also in this cohort, combined RNA-seq and CNV-seq analysis identified several PARPi-resistant tumors with significantly lower expression and acquired copy number loss of *Parg* (Figure S7.1C). Taken together, our *in vivo* data confirm and extend the results from the *in vitro* screens and suggest that PARG depletion alleviates PARPi toxicity.

7.4.3 PARG downregulation causes PARPi resistance *in vitro*

To validate the role of PARG depletion in PARPi resistance, we introduced two individual shRNAs against PARG (PARG sh1, PARG sh4) in KB2P1.21 and KB2P3.4 cells, resulting in substantial decrease of *Parg* mRNA levels (Figure 7.3A and Figure S7.2A) and reduced PARG activity (Figure 7.3B and Figure S7.2B) Consistently, genetic depletion of PARG in KB2P cells led to the accumulation of PAR under genotoxic stress induced by the alkylating agent methyl methanesulfonate (MMS) (Figure S7.2C-D) but did not affect *Parp1* expression levels (Figure S7.2E).

The shRNA-mediated loss of PARG resulted in increased resistance to the PARPi olaparib and AZD2461 in long-term clonogenic survival assays. This effect was observed in cell lines derived from both KB2P and KB1P tumor models, in which PARPi-induced loss of PARG was observed *in vivo* (Figure 7.3C and Figure S7.3A-C). To exclude off-target effects of the shRNAs, we also targeted the *Parg* locus in KB2P cells using CRISPR-Cas9 technology (Figure 7.3D and Figure S7.3D-F). In contrast to the control cells, *Parg*-targeted cells formed many resistant colonies after 14 days of PARPi selection. This effect was specific to Parg inactivation, as shown by the TIDE (Tracking of Indels by Decomposition) analysis[29]. In the initial tumor cell population, roughly half of the alleles carried frameshift mutations, and vehicle (DMSO) treatment did not significantly affect allele composition. In contrast, PARPi selection resulted in a substantial increase in frameshift disruptions ($> 90\%$), showing that the surviving populations are predominantly PARG deficient (Figure 7.3D and Figure S7.3D-F, Table S7.5).

We obtained further evidence for the role of PARG in PARPi resistance with a recently developed small-molecule PARG inhibitor (PARGi) PDDX-004 (PDD00017272), which is very active against mouse PARG[30]. In line with this, PDDX-004 caused a dose-dependent accumulation of nuclear PAR upon MMS-induced DNA damage in our cell lines (Figure S7.3G-H). Consistent with our genetic inhibition experiments, the clonogenic assays in KB2P cells also showed an increased PARPi survival upon chemical inhibition of PARG (Figure 7.3E and Figure S7.3I). Moreover, the viability of cells exposed to the combination of PARPi and PARGi correlated with the degree of PARG inhibition, while PDDX-004 alone did not affect cell growth nor PARPi response of PARG-depleted cells (Figure S7.3J). In conclusion, both genetic depletion or inactivation and chemical inhibition of PARG lead to PARPi resistance in KB2P cells, confirming an important functional role of PARG in mediating PARPi toxicity.

7.4.4 PARG-Depleted KB2P cells remain HR deficient and fail to protect stalled RFs

The sequence of events that leads to PARPi-induced death of BRCA-deficient cells includes the inhibition of PAR synthesis, RF collapse, and the formation of DSBs. In collaboration with the Nussenzweig laboratory, we have recently shown that RF protection can explain resistance in some of the PARPi-resistant KB2P mouse mammary tumors[26]. Given its role in PAR catabolism, however, we did not expect that the tumors in which we find PARG downregulation would correct the BRCA2 defect by protecting stalled RFs or by BRCA2-independent restoration of HR. To verify this, we measured the capacity of PARG-depleted KB2P cells to protect stalled RF using DNA fiber assays. In both control and PARG knockdown KB2P cells, the induction of replication stress resulted in the degradation of nascent tracts (Figure S7.4A), suggesting that PARG loss cannot bypass the requirement of BRCA2 for RF stabilization. Next, we assessed the capability of KB2P cells to form RAD51 ionizing radiation (IR)-induced foci (IRIF), a hallmark of HR activity. As expected, we did not detect any RAD51 IRIF in KB2P cells, regardless of *Parg* expression levels (Figure S7.4B-C). Moreover, the same phenotype was observed in PARPi-resistant KB2P tumors, in which PARG loss was confirmed at the genomic level (Figure S7.4D). These results demonstrate that loss of PARG causes PARPi resistance independently of BRCA2 and that resistance cannot be explained by HR restoration or RF protection.

7.4.5 PARG downregulation rescues PARylation upon PARPi treatment

To assess how PARG depletion causes PARPi resistance, we studied its effect on PARylation. Upon PARPi treatment, inhibition of PARP enzymes serves as the major

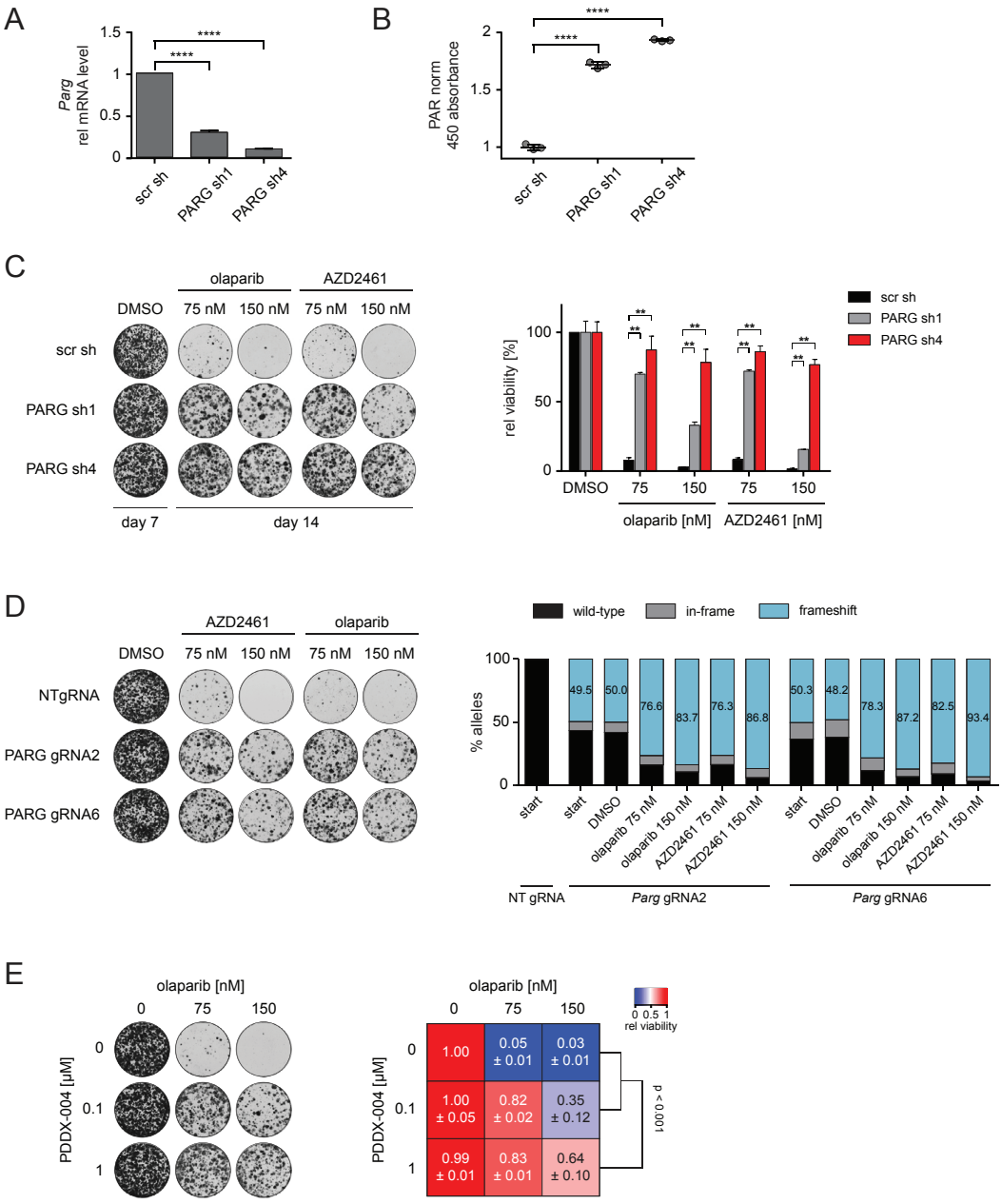


Fig. 7.3. Downregulation of PARG causes PARPi resistance in BRCA2-deficient cells *in vitro*. **(A)** RT-qPCR analysis of *Parg* expression levels in KB2P1.21 cell lines expressing indicated shRNAs; data represent mean ± SD of triplicate; $^{****}p < 0.001$ (two-tailed t-test). **(B)** PAR ELISA assay in KB2P1.21 cells; data shown as mean ± SD of triplicate, $^{****}p < 0.001$ (two-tailed t-test). **(C)** Representative images (left) and quantification (right) of long-term clonogenic assay with KB2P1.21 cells, treated with PARPi or untreated (DMSO). Data represent mean ± SD of three independent repeats; $^{**}p < 0.01$ (t-test). **(D)** Representative images (left) and TIDE quantification (right) of long-term clonogenic assay with KB2P1.21 cells expressing indicated guide RNAs. **(E)** Representative images (left) and quantification (right) of long-term clonogenic assay with KB2P1.21 cells treated as indicated. The experiment was repeated three times; data plotted as mean ± SD; $p < 0.001$ (ANOVA). See also Figures S7.2 and S7.3 and Table S7.5.

barrier to PAR formation, but this is reinforced by the PARG-mediated degradation of PAR, which acts in the same direction as PARPi. We therefore hypothesized that PARPi alone does not fully inhibit PARP and loss of PARG would allow sufficient PAR formation in the presence of PARPi. We tested this hypothesis by measuring endogenous PAR levels in KB2P cells treated with the PARPi olaparib and the PARGi PDDX-004 (Figure 7.4A). To discriminate between stabilization of pre-existing and *de novo* synthesized PAR, we first incubated cells with the inhibitors for 1 hr and then exposed them to MMS to stimulate PARP activity. As predicted, olaparib treatment resulted in a strong reduction of PAR, already at nanomolar doses. Inhibition of PARG overcame this reduction and MMS treatment led to a further increase of the PAR signal. These data indicate that olaparib concentrations sufficient to kill cells do not completely inhibit PARP and that this residual activity is sufficient for PAR formation if PARG activity is suppressed. We conclude that endogenous PARG activity is required for efficient inhibition of PAR signaling by PARPi.

We next investigated the effect of different PARP family members on the PARPi response using BRCA-proficient KP3.33 cells, in which CRISPR/Cas9-mediated disruption of *Parp1*, *Parp2*, or *Parp3* was well tolerated (Figure S7.5A-C). While the PARPi sensitivity of *Parp2*$^{-/-}$ and *Parp3*$^{-/-}$ KP3.33 cells was significantly reduced by PARG inhibition, *Parp1* depletion resulted in partial resistance to olaparib (consistent with Murai *et al.*[7]), which was not further increased by PARGi treatment (Figure S7.5D). This result is in line with a previous report suggesting that up to 90% of cellular PAR results from PARP1 activity[31] and shows that PARG-related PARPi resistance is mainly mediated by PARP1 signaling.

7.4.6 PARG inhibition reduces PARP1-DNA complexes induced by PARPi treatment

It has been shown that PARP1 association to and dissociation from chromatin is regulated by its PARylation[5] and persistent PARP1-DNA complexes, induced by clinical PARPi, are toxic to cells[7]. We therefore measured the levels of chromatin-bound PARP1 in KB2P cells using a previously described trapping assay[7]. Immunoblot analysis showed olaparib-dependent accumulation of PARP1 in chromatin fractions, which was reduced in cells expressing PARG-targeting shRNAs (Figure 7.4B). Since stable PARG depletion could result in a substantial proportion of free PARP1 in a PARylated state, and therefore lower its affinity to chromatin, we repeated the PARP1 trapping assay in cells exposed to short-time inhibition of PARG (Figure 7.4C). Although single treatment with PDDX-004 led to decreased levels of chromatin-associated PARP1, simultaneous inhibition of both PARP1 and PARG resulted in PARP1 trapping comparable with olaparib alone. We further corroborated this finding by measuring PARP1 association kinetics at multi-photon laser-induced DNA damage

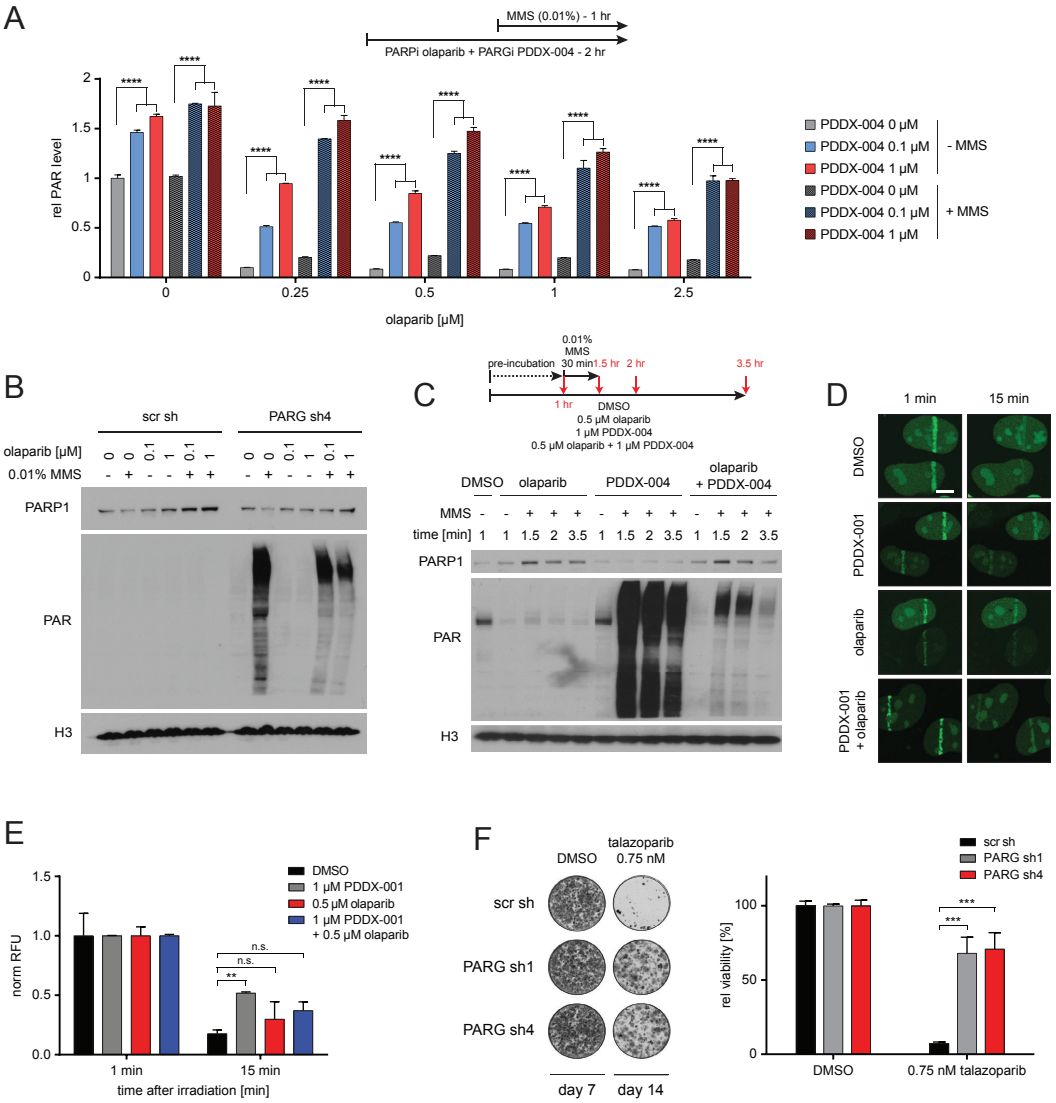


Fig. 7.4. PARG deficiency partially rescues PARylation and reduces the accumulation of PARP1-DNA complexes. (A) ELISA PAR assay in KB2P3.4 cells treated as indicated; data shown as mean ± SD of triplicate (t-test). $^{****}p < 0.0001$. **(B-C)** Immunoblot analysis of PARP1 in chromatin-bound fractions upon genetic (B) and chemical (C) inhibition of PARG in KB2P cells, treated as indicated; data representative for two independent experiments. **(D-E)** Representative images (left) and quantification (right) of analysis of PARP1 recruitment kinetics to multiphoton tracks in U2OS PARP1-GFP cells, following the indicated treatments. $^{**}p < 0.01$; n.s., not significant; two-tailed t-test, data represented as mean ± SEM. Scale bar, 10 μm. **(F)** Representative images of stained wells (right) and quantification (left) of clonogenic assay in KB2P cells expressing the indicated shRNAs and treated with talazoparib; data presented as mean ± SD of two experiments; $^{***}p < 0.001$ (two-tailed t-test). See also Figures S7.4 and S7.5.

sites in U2OS cells (Figure 7.4D-E). Cells were exposed to 0.5 μmol olaparib and/or the PARGi PDDX-001/PDD00017273 (1 μmol)[30], which alone efficiently inhibited downstream signaling of both proteins (Figure S7.5E-G). We utilized U2OS GFP-PARP1 cells and quantified the intensities of laser tracks; first, 1 min post irradiation, when under native conditions PARP1 accumulation reached a maximum, and then

15 min after the induction of DNA damage, when most of the chromatin had evicted PARP1 (Figure S7.5H). Both olaparib treatment alone and the combination with PDDX-001 resulted in a slight increase of chromatin-associated PARP1 15 min post irradiation (Figure 7.4D-E). Of note, even more PARP1 protein remained associated with damaged sites in cells exposed to PARGi only. The results, however, do not show any evidence that PARG depletion results in more rapid release of chromatin-bound PARP1. Taken together, our data demonstrate that PARG depletion does not enhance PARP1 dissociation from chromatin and therefore do not diminish PARP1 trapping *per se*. Instead, long-term suppression of PARG prevents excessive PARP1 binding and thus reduces PARPi-dependent accumulation of toxic PARP1-DNA complexes. The relevance of this finding is further supported by the fact that PARG depletion also results in resistance to talazoparib, a highly potent PARP1-DNA trapping agent in clinical use (Figure 7.4F).

7.4.7 PARG depletion alleviates PARPi-induced DNA damage

Following different forms of genotoxic stress, PARP1 activity has recently been shown to limit the rate of RF progression[32,33] by modulating fork reversal and preventing premature restart of reversed RF[34,35]. Deregulated RF remodeling by PARP inhibition was suggested to contribute to the synthetic lethality of PARPi with HR defects, as it increases the fraction of SSBs being processed into DSBs[36]. As shown in Figure 7.5A, we confirmed that PARPi treatment increases the DNA fiber track length upon induction of DNA damage with MMS or camptothecin (CPT). When PARG was also inhibited in these cells, the track length was significantly decreased, suggesting that PARG depletion prevents unrestrained RF progression in PARPi-treated cells (Figure 7.5A and Figure 7.5I). Concomitantly, PARG inhibition reduced the formation of DSBs in these cells, as measured by the neutral comet assay (Figure 7.5B and Figure 7.5J).

Given the role of PARP1 in DNA repair, we next investigated the PARP1-mediated recruitment of the scaffold protein XRCC1, a PAR interactor and a key player in the BER pathway[37]. To study the effects of PARPi and/or PARGi on XRCC1 recruitment, we applied the laser micro-irradiation assay to U2OS cells expressing a XRCC1-GFP fusion protein. We found that under drug-free conditions XRCC1-GFP was rapidly recruited to sites of laser-induced DNA damage (Figure 7.5C-D). Although a large proportion of the protein dissociated from chromatin within 60 min after irradiation, a substantial amount of XRCC1-GFP remained at the sites of DNA damage. In line with previous reports[38], treatment of cells with the PARPi olaparib abrogated XRCC1-GFP localization to laser-inflicted damage (Figure 7.5C-D). Inhibition of PARG mitigated the inhibitory effect of olaparib, however, and partially rescued XRCC1-GFP

recruitment. Importantly, the quantitative analysis of laser track intensities showed that the restored accumulation, although delayed in time, resulted in a similar retention of XRCC1-GFP at 1 hr post irradiation as in the control cells (Figure 7.5C-D). This effect of PARGi is specific to PARylation-induced recruitment of DNA repair factors, as we did not observe any differences in chromatin association of MDC1, which localizes to damaged sites in a PARP1-independent manner[39] (Figure 7.5E). As a readout for PARP1 function in the repair of SSBs, we employed the previously described Fast Micro-method SSB assay[40]. Consistent with our previous results, *Parg*-depleted cells exhibited increased capacity to repair SSBs in comparison with control cells (Figure 7.5F). This was further confirmed in cells exposed to olaparib for 24 hr. Immunofluorescence (IF) analysis of γH2AX foci revealed that *Parg*-depleted cells accumulated less olaparib-induced DNA damage (Figure 7.5G).

Based on our data, we conclude that PARG suppression not only reduces PARP1-DNA complexes but also rescues controlled RF progression and promotes the recruitment of DNA repair enzymes to sites of damage in cells exposed to PARPi. Altogether, this leads to a reduction of PARPi-induced DNA damage and improved PARPi survival (Figure 7.5H).

7.4.8 PARG deficiency overcomes PARPi toxicity in human cancer cells

The anticancer efficacy of PARPi has been validated in various clinical studies and several PARPi were recently approved for the treatment of patients with *BRCA1/2*-mutated tumors. We therefore determined whether PARG depletion confers PARPi resistance in human cancer cells by introducing two individual shRNAs targeting PARG in *BRCA1*-mutated SUM149PT (carrying a protein-truncating 2288delT mutation) and BRCA2-deficient DLD-1 cells. Both shRNAs efficiently suppressed PARG expression and conferred resistance to olaparib (Figure 7.6A-B and Figure S7.6A). Similarly, chemical inhibition of PARG led to increased survival of both cell lines in the presence of PARPi (Figure S7.6B-C).

Given that PARG loss causes PARPi resistance independently of BRCA1/2, we extended our analysis to a recently published pharmacogenomics dataset of 1,001 human cancer cell lines[41]. In particular, we assessed the correlation between half maximal inhibitory concentration (IC_{50}) values of PARPi and gene expression levels of 1,800 DDR-related factors (Table S7.2). Gene expression data and drug responses to four different PARPi (olaparib, veliparib, rucaparib, and talazoparib) were available for 935 cell lines from this panel. Statistical analysis revealed a significant negative association between PARG expression and IC_{50} values of all four PARPi (Figure 7.6C and Figure S7.6D); i.e., higher *PARG* RNA levels were related to in-

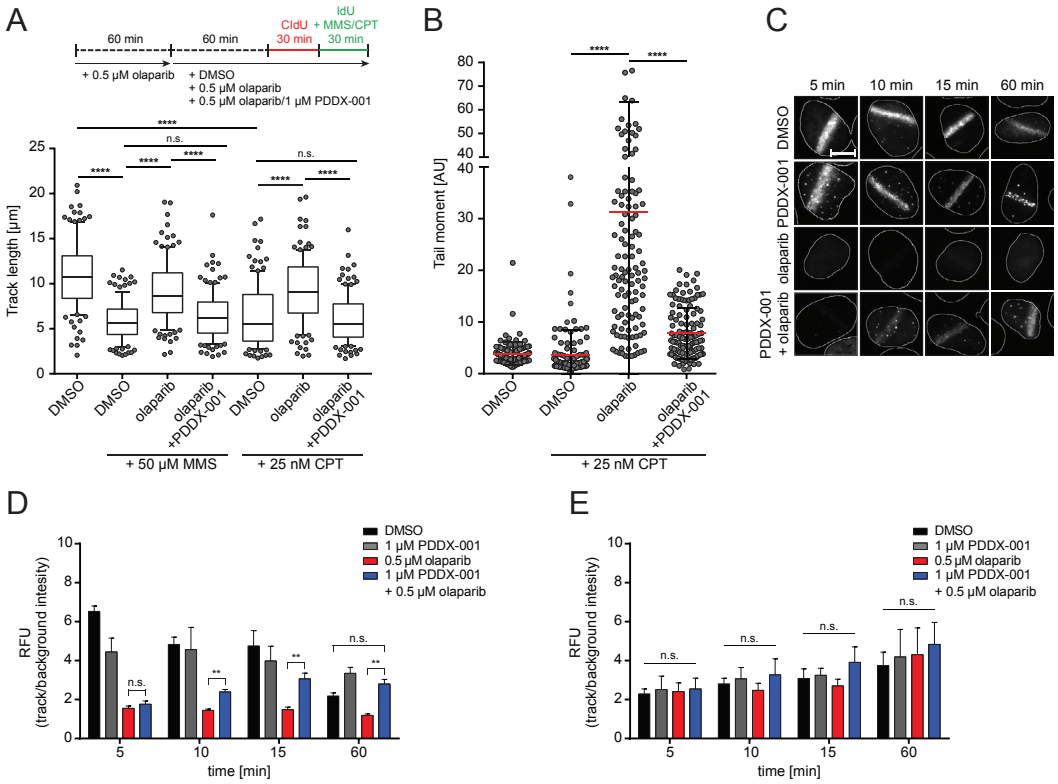


Fig. 7.5. PARG inhibition alleviates PARPi-induced DNA damage. (A) RF progression assay in U2OS cells exposed to indicated treatments; the experiment was repeated twice; box extends from 25th to 75th percentile, with a middle line representing the median and whiskers drawn down to the 10th percentile and up to the 90th; Mann-Whitney U test; $^{****}p < 0.001$; n.s., not significant. **(B)** Neutral comet assay in U2OS cells treated as in (A); Mann-Whitney U test, $^{****}p < 0.001$, data shown as mean $\pm$ SD of a replicate; the experiment was repeated twice. **(C-D)** Representative images (C) and quantification (D) of time course analysis of GFP-XRCC1 recruitment in U2OS cells treated as indicated; $^{**}p < 0.01$; n.s., not significant; two-tailed unpaired t-test; data represent mean $\pm$ SEM of three independent experiments. Scale bar, 10 μm. **(E)** Quantification of MDC1 tracks following immunostaining; statistical analysis as in (D).

creased sensitivity to these drugs. A similar negative association was also observed for *PARP1* gene expression, in agreement with the concept that more PARP1 leads to more trapping of PARP1 onto DNA in the presence of PARPi.

7.4.9 PARG depletion occurs in triple-negative breast and serous ovarian cancer

To further assess the clinical relevance of PARG depletion, we measured the heterogeneity of PARG expression in large sections of 56 treatment-naive triple-negative human breast cancer (TNBC) biopsies from high-risk women eligible for PARPi treatment[42,43]. IHC analysis (Figure 7.6D-E, and Figure S7.6E) revealed that, although PARG protein was expressed in a vast majority of the biopsies, PARG-negative

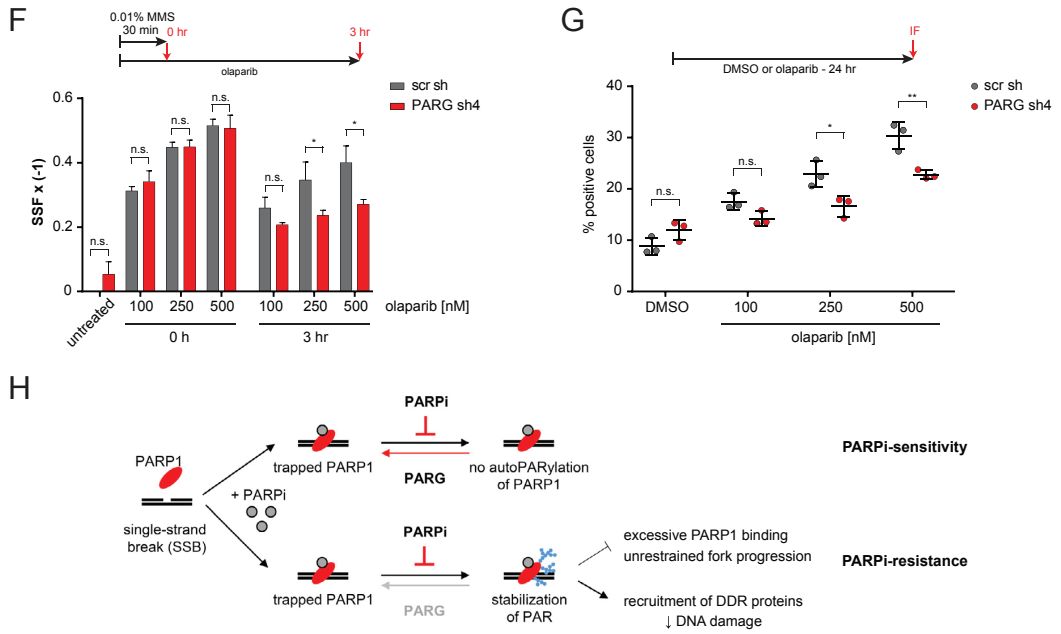


Fig. 7.5. **Continued. (F)** SSB assay in KB2P cells, treated as indicated; SSF, strand scission factor; data representative for two independent experiments, shown as mean ± SD of a replicate; two-tailed unpaired t-test; $^{*}p < 0.05$; n.s., not significant. **(G)** IF analysis of gH2AX foci in KB2P cells, treated as indicated; statistical analysis as in (F); $^{**}p < 0.01$; $^{*}p < 0.05$. **(H)** Proposed model: in the presence of PARPi, PARP1 maintains residual activity but is counteracted by PARG. Loss of PARG leads to stabilization of PAR chains and, consequently, limits accumulation of toxic PARP-DNA complexes, prevents unrestrained RF progression, and promotes the recruitment of the downstream factors. Together, this results in reduced PARPi-induced DNA damage and improves PARPi survival. See also Figure S7.5.

areas were found in a sizable proportion of samples. Specifically, 29 (52%) and 14 (25%) out of 56 cancers showed lack of PARG in areas corresponding to >10% and >20% of the tumor, respectively. Moreover, PARG-negative cells were positive for PAR, and in some of the samples PAR levels were substantially increased (Figure 7.6E). Of note, the variable degree of PAR could also reflect the degree of the endogenous DNA damage among the cases, as PARP enzymes are activated by DNA damage and these patients did not receive any genotoxic therapy. A similar PARG expression spectrum was also found in a cohort of serous ovarian carcinomas[44], a cancer type that has been recently approved for PARPi treatment (Figure 7.6D and Figure S7.6F). Taken together, our data show that PARG-depleted cells pre-exist in a substantial proportion of clinically relevant tumors and could potentially be selected by PARPi treatment.

7.4.10 PARG suppression results in acquired vulnerabilities

Molecular alterations that render cells resistant to targeted therapies may also cause synthetic dependencies, which in turn could be exploited therapeutically to prevent

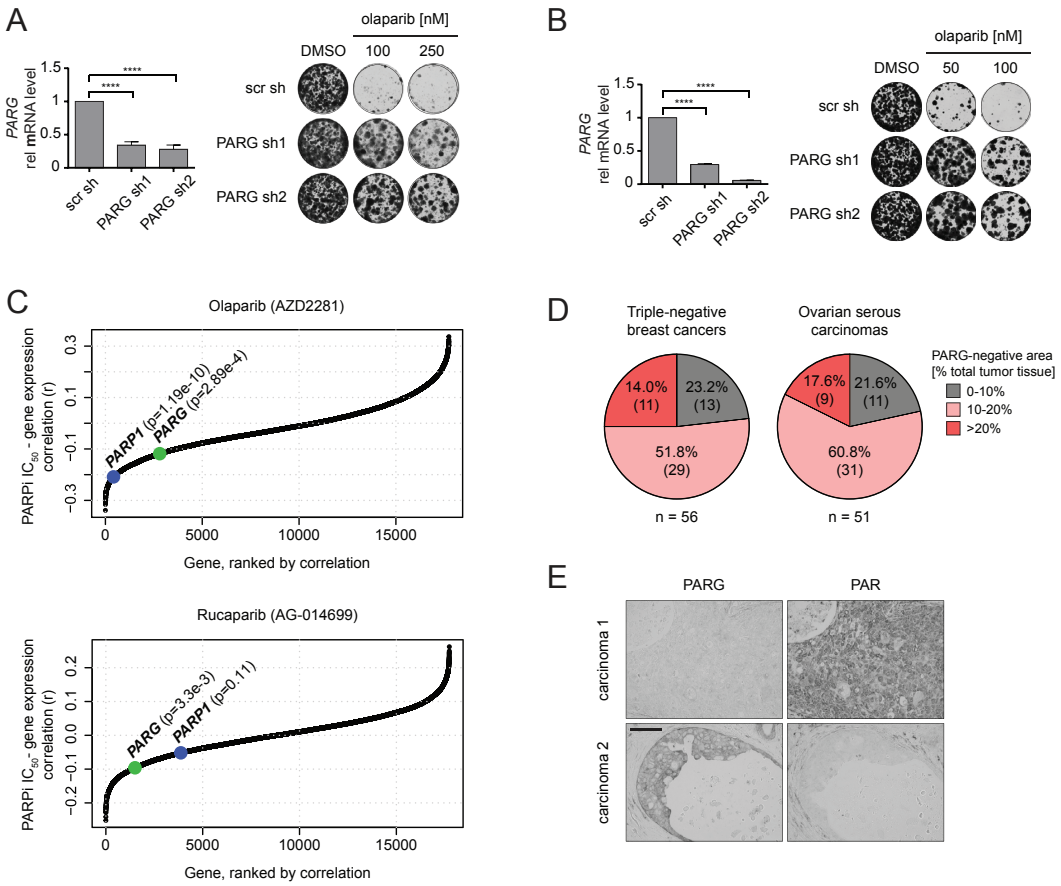


Fig. 7.6. PARG depletion in human cancer cells leads to PARPi resistance but results in acquired vulnerabilities. (A-B) Long-term clonogenic assay with SUM149PT (A) and DLD-1 BRCA2-deficient (B) cells expressing indicated shRNAs and treated with olaparib; PARG expression levels were determined by qRT-PCR (left; mean $\pm$ SD of triplicate, $^{****}p < 0.001$, t-test) and representative images of stained wells (right) are shown. **(C)** Correlation between IC_{50} of olaparib or rucaparib and expression of DDR genes; PARP1 and PARG are highlighted; p values were determined using the relation between estimated coefficient and Student's t distribution. **(D)** Summary of IHC analysis of PARG expression in TNBCs and ovarian serous carcinomas. **(E)** Representative images of PARG and PAR IHC of TNBC biopsies. Scale bar, 100 μm.

cancer progression. Perturbed PAR signaling due to downregulation of PARG has been shown to increase the sensitivity to IR[45]. Susceptibility to IR is also one of the characteristics of cells with dysfunctional BRCA1/2 proteins[46]. We therefore set out to determine whether PARG suppression could potentiate IR toxicity in *BRCA1/2*-mutated cells. For this purpose, SUM149PT cells with shRNA-mediated *PARG* knockdown were exposed to a range of IR doses and grown for another 7 days. Viability measures showed increased IR sensitivity of *PARG*-depleted cells in comparison with control populations (Figure 7.6F). Furthermore, dose-dependent sensitization was also achieved by chemical inhibition of PARG in SUM149PT, BRCA2-deficient DLD-1, and KB2P cells (Figure 7.6G-H and Figure S7.6G). Notably, we also observed synergistic effects between PARG inhibition and treatment with temozolo-

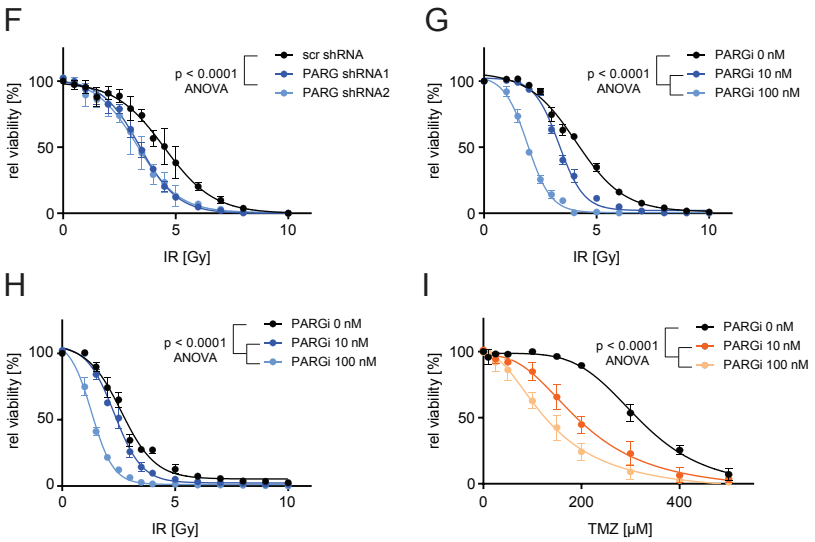


Fig. 7.6. **Continued. (F-G)** Response to irradiation of SUM149PT cells, expressing indicated shRNAs (F) or treated with PARGi (G). **(H)** Response to irradiation of DLD-1 BRCA2-deficient cells treated as indicated. **(I)** Response to temozolomide of SUM149PT cells treated as indicated. In (F)-(I) data are presented as mean ± SD of triplicates. See also Figure S7.6.

mide (Figure 7.6I and Figure S7.6H), an alkylating agent that has been previously shown to potentiate PARPi toxicity[8].

Together, these results illustrate that PARG suppression, although detrimental to PARPi efficacy, provides therapeutic vulnerabilities that could be used to target resistant tumors.

7.5 Discussion

In this study, we show that loss of PARG is a frequent mechanism of PARPi resistance in *Brca2*-mutated tumors. Our data provide an HR-independent mechanism for tumor cells to adapt their DDR in order to escape the lethal effects of PARPi. PARG is the main enzyme responsible for degrading nuclear PAR and thereby counteracting the action of PARP enzymes. Hence, PARG works in the same direction as PARPi and prevents PAR accumulation. Our finding that PARG depletion causes PARPi resistance in BRCA2-deficient tumors highlights an important aspect of PARPi therapy: the endogenous PARG activity in tumor cells is crucial for therapy success. As PARPi do not fully block PARP activity, loss of PARG activity is sufficient to restore PAR formation and rescue downstream signaling of PARP1.

Within the PARP family of ADP-ribosyltransferases, three family members, PARP1, PARP2, and PARP3, have been linked to DNA repair[47]. PARP1 is the most abundant

of these and has been shown to play critical roles in the DDR[48]. Upon DNA damage, RFs are rapidly and globally reversed and are maintained in the reversed state by transient PARP-mediated inhibitory ADP ribosylation of RECQ1 helicase, the enzymatic activity specifically required for restart of reversed RFs[34,35]. In this way, PARP1 represents a molecular switch to control transient fork reversal and RF restart following genotoxic stress[35]. Neelsen and Lopes[36] therefore suggested that the synthetic lethality of PARPi with HR defects results not only from an increasing load of SSBs but also from a greater fraction of these lesions being processed into DSBs. Whereas untreated cells gain extra time to repair DNA damage through RF reversal, PARPi-treated cells are unable to reverse forks efficiently, resulting in increased DNA breakage and the requirement for HR-mediated DSB repair. In line with this hypothesis, we found that PARG depletion restores controlled RF progression in the presence of PARPi and reduces subsequent DNA damage. We also found a mechanism at the level of DNA repair that contributes to PARPi resistance induced by PARG loss: PAR stabilization rescues the recruitment of the downstream scaffolding protein XRCC1, which is known to bring together a variety of components required for efficient SSB repair[49].

In BRCA-proficient tumors, the toxic effect of PARPi can also be counteracted by the loss of the drug target PARP1. Consistent with the data of Pettitt *et al.*[19], we found a significant enrichment for *Parp1*-targeting shRNAs in our drug resistance screen in BRCA-proficient mammary tumor cells. In accordance with the concept of synthetic lethality, however, this hit did not show up when we screened BRCA2-deficient mammary tumor cells. Previous screens in BRCA1-deficient tumor cells also did not yield *Parp1* as a hit[22].

Most PARPi do not only block the catalytic activity of PARP1 but also induce toxic PARP1-DNA complexes. Our study shows that PARG inhibition reduces the amount of trapped PARP1 by preventing its excessive binding. This result underscores the delicate balance between enzymatic PARP1 activity and its toxicity when trapped on DNA.

Since PAR synthesis and degradation go hand in hand in orchestrating the DNA damage response, the use of PARGi has been put forward for the treatment of cancers with DDR defects[50], and the possibility of a synthetic lethal interaction between PARG and BRCA proteins has received considerable interest. However, several studies that addressed this question have produced contradictory results[51–53], which suggest that sensitivity to PARG depletion may depend on the cell line and the degree of PARG suppression. Importantly, Koh *et al.*[54] showed that PARG depletion, although embryonically lethal, can be tolerated in embryonic stem cells cultured in the presence of PARPi. In our cell lines, both genetic depletion and chemical inhibition of PARG were well tolerated and did not affect cellular viability. Moreover,

homozygous loss of *Parg* was acquired *in vivo* in a substantial fraction of KB2P tumors.

PARG-negative clones were also found in a sizable proportion of human high-risk TNBCs or serous ovarian cancers, potential target groups for PARPi treatment due to the increased presence of HR-deficient cancers. Taken together, these data suggest that PARG-negative clones can be specifically selected by PARPi treatment and modulate therapy response.

While the clinical application of PARPi has initially focused on *BRCA1/2*-mutated tumors, the therapeutic scope of these drugs is now being extended to other molecular defects (reviewed by Lord and Ashworth[55]). Since PARG acts directly at PAR structures and independently of the HR pathway, stabilization of PARylation via PARG suppression might represent a generic mechanism of PARPi resistance, relevant for a broad spectrum of cancers. Although this is bad news for the clinical use of PARPi, loss of PARG can also be exploited as a potential Achilles' heel for cancer treatment, as it confers sensitivity to IR[45]. Our data indeed show that PARG suppression potentiates the toxicity of radiation therapy in BRCA-deficient cells. Additionally, we show that PARG inhibition synergizes with temozolomide, a chemotherapeutic agent that is now being evaluated in the clinic in combination with PARPi[56].

Our research has yielded a collection of matched PARPi-naive and -resistant *Brca2*$^{-/-}$ mouse mammary tumors, which can be further utilized in a search for additional resistance mechanisms. Although PARG loss was observed in the majority of the PARPi-resistant KB2P tumors, it cannot explain resistance in all cases. Three other candidates, *Rps6ka6*, *Socs4*, and *Pbrm1*, were validated as additional significant hits in a secondary screen. Since all three of these genes are connected to chromatin, it will be interesting to understand the underlying mechanism of how they affect PARPi response in a follow-up study.

In collaboration with the Durocher and Lord laboratories, we have recently identified that PARPi resistance can also be caused by loss of the Shieldin (SHLD) complex, which acts downstream of 53BP1 in blocking DNA end resection (Noordermeer *et al.*, manuscript submitted). Importantly, loss of the SHLD complex is not a generic PARPi resistance mechanism, since it causes PARPi resistance specifically in BRCA1- but not in BRCA2-deficient cells. In contrast, loss of PARG explains PARPi resistance in both *BRCA1*- and *BRCA2*-mutated tumors, as it operates independently of the HR pathway.

Taken together, our findings suggest that PARG is an important mediator of PARPi response. The presence of PARG-negative cells in treatment-naive tumors from the clinically relevant groups of high-risk women suggests that PARG loss should be

assessed as a potential cause of clinical PARPi resistance. In this case, measurement of PARG activity should further improve clinical decision making for patients with tumors that lack homology-directed DNA repair.

Author contributions Conceptualization, E.G., P.B., J.J., and S.R.; Methodology, E.G., A.A.D., W.W.W., D.I.J., and B.v.d.B.; Investigation, E.G., A.A.D., J.R.d.R., W.W.W., J.A.S., S.G.L., J. Bartkova, I.G., M.A.S., R.d.B., S.A., M.B., and A.K.; Software and Formal Analysis, E.G., J.R.d.R., R.d.B., and D.J.V.; Resources, D.I.J, D.J.O., and M.T.; Writing - Original Draft, Review & Editing - E.G., P.B., J.J., and S.R.; Supervision, M.v.d.V., M.v.V., L.F.A.W., J. Bartek., M.L., H.v.A., P.B., J.J., and S.R.; Funding Acquisition, D.J.O., J. Bartek., H.v.A., J.J., and S.R.

Acknowledgments We thank the members of the Preclinical Intervention Unit of the Mouse Clinic for Cancer and Ageing (MCCA) at the NKI, Wendy Sol, and Guotai Xu for their technical support with the animal studies, and we thank the NKI core facilities: Animal Pathology facility, Digital Microscopy facility, Genomics Core facility, and Animal facility for their excellent service. We are also grateful to Peter Bouwman (NKI) and Thanos Halazonetis (Geneve University) for help in generating the list of DDR-associated genes, to SurfSara for allowing us to use their HPC Cloud service for the computational analyses, and to Chief Pathologist Vera Timmermans Wielenga (Department of Pathology, Copenhagen University Hospital, Denmark) and the technician Maria Grønvig Nielsen (Danish Cancer Society Research Center) for their contribution to IHC analysis of TNBC biopsies. Financial support came from the Dutch Cancer Society (KWF 2011-5220 and 2014-6532), the Netherlands Organisation for Scientific Research (VICI 91814643 and a National Roadmap grant for Large-Scale Research Facilities), the Netherlands Genomics Initiative (93512009), Cancer Research UK (C480/A1141 and C5759/A17098), the Danish Cancer Society, Swedish Research Council, Cancerfonden and Danish National Research Foundation (project CARD), the Swiss National Science Foundation (310030_156869), the Swiss Cancer League (KLS-4282-08-2017), and the European Research Council (CoG-617485, CoG-681572, SyG-319661).

Conflict of interest The authors declare no potential conflicts of interest.

7.6 STAR Methods

7.6.1 Contact for reagent and resource sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Sven Rottenberg (sven.rottenberg@vetsuisse.unibe.ch).

7.6.2 Experimental model and subject details

Mice All animal experiments were approved by the Animal Ethics Committee of The Netherlands Cancer Institute (Amsterdam, the Netherlands) and performed in accordance with the Dutch Act on Animal Experimentation (November 2014). *Brca2*$^{-/-}$*;p53*$^{-/-}$ mammary tumors were generated in *K14cre;Brca2*$^{F/F}$*;p53*$^{F/F}$ (KB2P) female mice, described previously[20]. Tumor implantation experiments were performed in syngeneic, wild-type F1 (first filial generation) FVB:129/Ola females, at the age of 6 weeks. Parental FVB (FVB/NRj) and 129/Ola animals were purchased from Janvier Labs and Harlan Olac, respectively, and crossed at the NKI Animal Facility. Animals were assigned randomly to the treatment groups and the treatments were supported by animal technicians who were blinded regarding the hypothesis of the treatment outcome. Collection of *Brca1*$^{-/-}$*;p53*$^{-/-}$ (KB1P) PARPi-naive and -resistant mouse mammary tumors was described before[24].

Human samples of triple-negative breast and serous ovarian cancer Retrospective Triple Negative Breast Cancer (TNBCs) biopsies from 56 clinical high-risk patients (high-risk definition according to the Danish Breast Cooperative Group; www.dbcg.dk accessed 22.10.2009) that underwent mastectomy between 2003 and 2015 were selected and classified as being triple negative according to the criteria set in the ASCO/CAP guidelines (ER $< 1\%$, PR $< 1\%$, HER2 0, 1+ or 2+ but FISH/ CISH negative). The patients presented a unifocal tumor of an estimated size of more than 20 mm. None of the patients had previous surgery to the breast and did not receive preoperative treatment[42,43]. This study was conducted in compliance with the Helsinki II Declaration and written informed consent was obtained from all participants and approved by the Copenhagen and Frederiksberg regional division of the Danish National Committee on Biomedical Research Ethics (KF 01-069/03).

Paraffin-embedded material from the cohort of ovarian tumors was collected at the Department of Pathology, University Hospital, Las Palmas, Gran Canaria, Spain, from surgical operations performed in the period 1995-2005. For the purpose of the present study, only samples from serous ovarian carcinoma (the type approved for

treatment by PARP inhibitors) were used from a larger cohort that was reported previously[44], and included also other histological types of ovarian tumors. The use of long-term stored tissue samples in this study was in accordance with the Spanish codes of conduct (Ley de Investigación Biomédica) and was approved by the review board of the participating institution. Patients were informed that samples may be used for research purposes under the premise of anonymity.

Cell lines All 2D cell lines used in this study were described previously: KB2P1.21, KB2P3.4, KP3.33[9], KB1P-G3[24], U2OS (RRID:CVCL_0042), SUM149PT (RRID:CVCL_3422), DLD-1 BRCA2(-/-) (Horizon Discovery, #HD 105-007; RRID:CVCL_HD57), HEK293FT (RRID:CVCL_6911).

For these lines, cell growth media were supplemented with 10 % (v/v) fetal calf serum (FCS, Sigma) and 50 units ml^{-1} penicillin-streptomycin (Gibco). KB2P1.21, KB2P3.4, KB1P-G3 and KP3.33 cells were grown in Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 (DMEM/F12; Gibco) containing 5 $\mu g\,ml^{-1}$ Insulin (Sigma, #I0516), 5 ng ml^{-1} cholera toxin (Sigma, #C8052) and 5 ng ml^{-1} murine epidermal growth-factor (EGF, Sigma, #E4127). U2OS cells were cultured in DMEM (Gibco) media supplemented with GlutaMAX (Gibco, #35050-061). SUM149PT cells were grown in RPMI1640 (Gibco) media, similarly to DLD-1 BRCA2(-/-) cells for which growth media was additionally enriched with 2 mmol L-glutamine and 25 mmol sodium bicarbonate. HEK293FT cells were cultured in Iscove's Modified Dulbecco's Media (IMDM, Gibco) supplemented with 2 mmol glutamine.

Tissue culture was carried out under standard conditions (37 °C, 5 % CO_2), except for KB2P1.21, KB2P3.4 and KB1P-G3 cells which were cultured under low oxygen conditions (3 % O2). All cell lines used in this study are of female origin, except for DLD-1 BRCA2(-/-) cells (male). Testing for mycoplasma contamination was performed on a regular basis.

Tumor-derived organoids KB2P26S.1, KB2P17 and KB2P12 tumor organoids were derived from a mammary KB2P PARPi-naive tumor (female donor), genotyped and cultured as described before[21]. Briefly, cultures were embedded in Culturex Reduced Growth Factor Basement Membrane Extract Type 2 (BME, Trevigen; 40 μl BME:growth media 1:1 drop in a single well of 24-well plate) and grown in Advanced DMEM/F12 (AdDMEM/F12, Gibco) supplemented with 1 mol HEPES (Sigma), GlutaMAX (Gibco) 50 units ml^{-1} penicillin-streptomycin (Gibco), B27 (Gibco), 125 mmol N-acetyl-L-cysteine (Sigma), 50 ng ml^{-1} murine epidermal growth factor (Sigma), 10 % (v/v) Rspo1-conditioned medium (kindly provided by Calvin Kuo, Stanford University) and 10 % (v/v) Noggin-conditioned medium[57]. Organoids were cultured

under standard conditions (37 °C, 5 % CO2) and regularly tested for mycoplasma contamination.

7.6.3 Method details

Generation of PARPi-naive and -resistant KB2P tumors

Spontaneous mammary tumors that arouse in KB2P (*K14cre;Brca2$^{F/F}$;p53$^{F/F}$*) mice were harvested, genotyped, sampled and cryopreserved (DMEM/F12, 10 % (v/v) FCS, antibiotics-free), as described before[20,58]. To obtain matched PARPi-naive and -resistant tumor panel, 21 individual spontaneous tumors were engrafted as tumor fragments in the fourth right mammary fat pad of wild-type FVB:129/Ola female mice (F1). Each tumor donor was transplanted into multiple animals (at least 2, 1 for control and 1 for PARPi treatment) and starting from 2 weeks after transplantation, tumor size was monitored at least three times a week. Tumor volume was determined by caliper measurements (length and width in mm) and calculated by using the following formula: $0.5 * \text{length} * \text{width}^2$. All treatments were initiated when tumors reached approx. 200 mm (100 % relative tumor volume). For PARPi treatment, AZD2461 powder was reconstituted in deionized water solution of 0.5 % (w/v) hydroxypropyl methylcellulose (HPMC) to a final concentration of 10 mg ml^{-1}. AZD2461 solution (100 mg kg^{-1}) or vehicle control (0.5 % HPMC) were administered to animals orally for 28 consecutive days. Upon tumor relapse to 100 % relative tumor volume, the treatment was repeated and continued for another 28 days, until acquired resistance. Animals were sacrificed when the tumors reached a volume of 1.500 mm. Tumor sampling included cryopreserved tumor pieces, fresh frozen tissue and formalin-fixed material (4 % (w/v) formaldehyde in PBS).

Generation of deep sequencing data and analysis

RNA preparation, sequencing and DIDS analysis Fresh-frozen tumor tissues were subjected to high-speed shaking in 2 ml microcentrifuge tubes containing 1 ml of TRIsure reagent (Bioline) and stainless steel beads (TissueLyser LT, Qiagen; 10 min, 50 Hz, room temperature). Homogenized lysates were further processed for RNA isolation following TRIsure manufacturer's protocol. Quality and quantity of the total RNA was assessed by the 2100 Bioanalyzer using a Nano chip (Agilent, Santa Clara, CA). Total RNA samples having RIN > 8 were subjected to library generation.

Strand-specific libraries were generated using the TruSeq Stranded mRNA sample preparation kit (Illumina Inc., San Diego, RS122-2101/2) according to the manufacturer's instructions (Illumina, Part #15031047 Rev. E). Briefly, polyadenylated RNA from intact total RNA was purified using oligo-dT beads. Following purification, the

RNA was fragmented, random primed and reverse transcribed using SuperScript II Reverse Transcriptase (Invitrogen, part #18064-014) with the addition of Actinomycin D. Second strand synthesis was performed using Polymerase I and RNaseH with replacement of dTTP for dUTP. The generated cDNA fragments were 30 end adenylated and ligated to Illumina Paired-end sequencing adapters and subsequently amplified by 12 cycles of PCR. The libraries were analyzed on a 2100 Bioanalyzer using a 7500 chip (Agilent, Santa Clara, CA), diluted and pooled equimolar into a 10 nmol sequencing stock solution.

Illumina TruSeq mRNA libraries were sequenced with 50 base single reads on a HiSeq2000 using V3 chemistry (Illumina Inc., San Diego). The resulting reads were trimmed using Cutadapt[59] (version 1.12) to remove any remaining adapter sequences, filtering reads shorter than 30 bp after trimming to ensure good mappability. The trimmed reads were aligned to the GRCm38 reference genome using STAR[60] (version 2.5.2b). QC statistics from Fastqc[61] (version 0.11.5) and the abovementioned tools were collected and summarized using Multiqc[62]. Gene expression counts were generated by featureCounts[63] (version 1.5.0-post3) using gene definitions from Ensembl GRCm38 version 76. Normalized expression values were obtained by correcting for differences in sequencing depth between samples using DESeq's median-of-ratios approach[64] and then log-transforming the normalized counts. Differentially expressed genes were identified using DIDS[27], using a threshold of $p < 0.05$ for statistical significance. Given that generated p value was a very conservative estimate of the true p value and was only used as a heuristic filter, we did not apply multiple testing correction (in accordance with de Ronde *et al.*[27]). Selected genes were subsequently ranked by the number of (resistant) samples that were considered to show differential expression according to the DIDS criteria.

Genomic DNA extraction and CNV sequencing Genomic DNA was isolated from fresh-frozen tumor material using standard phenol:chloroform extraction. CNV-Seq was performed using double stranded DNA (dsDNA), quantified with the QubitO®dsDNA HS Assay Kit (Invitrogen, #Q32851). To obtain fragment sizes of 160-180 bp, 2μg of dsDNA were fragmented by Covaris shearing and purified using 1.8X Agencourt AMPure XP PCR Purfication beads according to manufacturer's protocol (Beckman Coulter, #A63881). Next, sheared DNA was quantified and qualified on a BioAnalyzer system with the DNA7500 assay kit (Agilent Technologies, #5067-1506). Library preparation for Illumina sequencing was carried out with 1μg of DNA and KAPA HTP Library Preparation Kit (KAPA Biosystems, #KK8234). To obtain a sufficient yield for sequencing, 4-6 PCR cycles were performed during the library enrichment step. Prepared libraries were cleaned up using 1X AMPure XP beads and analyzed on a BioAnalyzer system using the DNA7500 chips to determine the molarity. Finally, up to 11 uniquely indexed samples were

pooled (equimolar pooling) in a final concentration of 10 nmol and sequenced on an Illlumina HiSeq2500 machine in one lane of a single read 65 bp run, according to manufacturer's instructions. The resulting reads were trimmed using Cutadapt[59] (version 1.12) to ensure a uniform length of 50 bp reads between samples and to remove any remaining adapter sequences. After trimming, reads shorter than 30 bp were removed to ensure good mappability. The trimmed reads were aligned to the GRCm38 reference genome using BWA aln[65] (version 0.7.15). The resulting alignmentswere sorted and marked for duplicates using Picard tools (version 2.5.0). QC statistics from Fastqc[61], Samtools[66] (version 1.2) and the above-mentioned tools were collected and summarized using Multiqc[62]. Copy number calls were generated using the QDNAseq[67] and QDNAseq.mm10 packages from Bioconductor (versions 1.8.0 and 1.4.0, respectively). To select for acquired copy number events, the CNV calls from resistant samples were filtered to remove any calls that were already present in the matched sensitive sample. After this filtering, genes were ranked by their recurrence (number of resistant samples with a loss/gain), dropping any genes that were only aberrant in a single sample.

DDR-related gene list The DDR gene set was generated based on the gene list from Thanos Halazonetis (Geneve University) and the NCBI search (terms: 'DNA repair', 'DNA damage response', 'DNA replication', 'telomere-associated genes'). See Table S7.2 for the complete gene list.

Final ranking of gene candidates To obtain a final ranking for the 82 candidate genes, we first created two new DIDS and CNV rankings, containing only these candidate genes. Next, to prioritize candidates with a strong correlation between copy number and expression, we created a third ranking by calculating the spearman correlation between the copy number values and expression of each gene across all samples, and sorting the genes by decreasing correlation. These three ranks were aggregated using the R package RobustRankAggreg[68] (version 1.1) to obtain the final rankings using three different aggregation methods (Stuart, RRA and Mean). To handle ties within the DIDS/CNV ranks (due to the discrete nature of these ranks), genes with the same values in these ranks were assigned the same rank value before the aggregation.

Immunohistochemistry analysis

PARG IHC analysis in KB2P tumor panel Immunohistochemical staining was performed on formalin-fixed paraffin embedded (FFPE) material of the KB2P tumor panel. First, antigen retrieval was performed by cooking the samples in citrate buffer 6.0 pH (ScyTek Laboratories) for 15 min in pressure cooker (110 °C).

Next, endogenous peroxidase activity was blocked by incubation a with methanol solution of 3 % (v/v) H_2O_2 for 20 min. 10 % (v/v) milk solution in PBS was used as a blocking buffer (30 min, room temperature) and PBS containing 1 % (w/v) bovine serum albumin and 1.25 % (v/v) normal goat serum was used as antibody diluent. Incubation with primary PARG antibody (Thermo Fisher Scientific; #PA5-14158; diluted 1:100, final concentration 20 $\mu g\,ml^{-1}$) was carried out overnight at 4 °C and followed by 30 min incubation with secondary Goat-a-Rabbit-Bio antibody (DAKO, #E0432; 1:1000) at room temperature. For detection, samples were exposed to PBS solution containing DAB substrate (Sigma, #D5905) and 0.025 % (v/v) H_2O_2 (Sigma, #A31642) for 20 min and hematoxylin counterstaining. Semi-quantitative (scoring: 1-low signal, 2-high signal) PARG expression analysis was carried out by a pathologist who was blinded regarding the identity of the samples.

PARG/PAR analysis in TNBC and serous ovarian carcinoma cohorts Five-mm tissue sections were cut from the formalin-fixed, paraffin-embedded tissue blocks and mounted on Super Frost Plus slides (Menzel-Gläser, Braunschweig, Germany), baked at 60 °C for 60 min, deparaffinized, and rehydrated through graded alcohol rinses. Heat induced antigen retrieval was performed by immersing the slides in citrate pH 6.0 buffer and heating them in a 750 W microwave oven for 15 min followed by immunohistochemistry staining with the primary antibodies as follows: anti-PARG antibody from Thermo Scientific (PA5 14158; diluted 1:2000) and anti-PAR antibody from GeneTex (10H, GTX75054, diluted 1:2500. The sections were incubated with the primary antibodies overnight in a cold-room, followed by subsequent processing by the indirect streptavidin-biotin-peroxidase method using the Vectastain Elite kit (Vector Laboratories, Burlingame, CA, USA) and nickel-sulphate-based chromogen enhancement detection as previously described[69], without nuclear counterstaining. For negative controls, sections were incubated with non-immune sera. The results were evaluated by two experienced researchers, including a senior oncopathologist, and the data expressed as percentage of positive tumor cells within each lesion.

Constructs, lentiviral transductions and genome editing

Constructs A collection of 1,976 lentiviral hairpins targeting 391 DDR-related mouse genes (pLKO.1; DDR library) were derived from the Sigma Mission library (TRCMm1.0) as described before[22]. Custom-made shRNA library (pLKO.1) targeting 82 candidate genes (identified by the multi-omics analysis of KB2P tumors) and 32 non-essential genes (552 shRNAs in total, on average 5 shRNAs/gene) was obtained from the Sigma Mission collection (TRCMm1.0) (see also Table S7.4). Non-essential genes were used as negative controls for the enrichment analysis, and were selected based on the work of Hart and colleagues[70] and RNA-Seq data from KB2P tumors (non-expressed genes).

Individual hairpin constructs used in the validation studies were selected from the TRC library: mouse PARG - sh1: TRCN0000126559, sh4: TRCN0000126562; human PARG - sh1: TRCN0000051303, sh2: TRCN0000051305) (see also Table S7.6). For CRISPR/Cas9-based genome editing two different systems were used: (1) for targeting *Parg* locus (Figure 7.3D and Figure S7.3D-F) the Cas9 expressing pGSC_Cas9_Neo vector (kind gift from Bastiaan Evers, NKI) was used and individual gRNAs (see Table S7.6) were cloned into the iKRUNC-Puro system described previously[71]; (2) for targeting *Parp1*, *Parp2*, *Parp3* loci (Figure S7.5A-D) lentiCRISPRv2 vector was used and individual gRNAs (see Table S7.6) were cloned as described previously[72]. For laser micro-irradiation experiments, pEGFP-C1-XRCC1 (gift from Simon Bekker-Jensen) and pEGFP-c3-PARP1 (gift from Valerie Schreiber) vectors were used. pEGFP-N1-CHD2 vector has been described before[73]. All constructs were verified by Sanger sequencing.

Lentiviral transductions Lentiviral stocks, pseudotyped with the VSV-G envelope, were generated by transient transfection of HEK293FT cells, as described before[74]. Lentiviral titers were determined using the qPCR Lentivirus Titration Kit (Applied Biological Materials), following the manufacturer's instructions. For all experiments the amount of lentiviral supernatant used was calculated to achieve the MOI (multiplicity of infection) of 50, except for the transduction of the lentiviral library (genetic screens) for which an MOI of 1.5 was used. To ensure efficient transduction, cells were incubated with lentiviral supernatants overnight in the presence of polybrene ($8\,\mu g\,ml^{-1}$). Antibiotic selection was initiated 24 hr post-transduction and was carried out for 5 consecutive days. Tumor-derived organoids were transduced according to a previously established protocol[21].

Genome editing For CRISPR/Cas9-mediated genome editing with the iKRUNC system, KB2P1.21 or KB2P3.4 cells were first transduced with the lentiviral pGS-Cas9 (Neo) construct (MOI 50) and grown under G418 selection ($500\,\mu g\,ml^{-1}$) for 5 days. Next, neomycin-selected cells were incubated with lentiviral supernatants of iKRUNC-Puro vectors (gRNA-encoding constructs, MOI 50) and exposed to $3\,\mu g\,ml^{-1}$ puromycin for 5 days. To induce gRNA expression, puromycin-surviving cells were treated for another 5 days with $3\,\mu g\,ml^{-1}$ doxycycline (Sigma). For CRISPR/Cas9-mediated genome editing with lentiCRISPRv2 system, KP3.33 cells were transduced with the lentiviral supernatant (MOI 50) and grown under Puromycin ($3\,\mu g\,ml^{-1}$) selection for 5 days. To assess the modification rate, genomic DNA was extracted (Puregene Core Kit A, Qiagen) and $100\,\mu g$ was used as an input for the PCR amplification of the targeted sequence. PCR reaction was performed with Thermo Scientific Phusion High-Fidelity PCR Master Mix (Thermo Scientific), according to manufacturer's instructions (3-step protocol: annealing - 60 °C for 5 s, extension time 15 s) and using primers listed in Table S7.6. Resulting PCR products were

purified using the MinElute PCR Purification Kit (Qiagen) and eluted in 50 μl of water. Finally, 2 μl of purified DNA served as a template for the BigDye Terminator v3.1 reaction (Thermo Fisher). BigDye PCR reactions were performed with the same forward primers as in the preceding PCR reactions (no reverse primer used) and according to the BigDye manufacturer's protocol. Allele composition was determined with the TIDE analysis[29] by comparing sequences from modified and parental (non-transduced control) cells.

Functional genetic screens

The DDR shRNA library and the shRNA library targeting candidate genes (secondary screen) were stably introduced into 2D lines (KP3.33, KB2P1.21, KB2P3.4) and organoids (ORG-KB2P26S.1) by lentiviral transduction (MOI 1.5). After antibiotic selection (puromycin, 3 μl ml^{-1}, 5 days) cells were seeded for the clonogenic assay with PARPi or pelleted for the genomic DNA isolation (day 0; control samples for the enrichment analysis). The total number of cells used in a single screen was calculated as following: library complexity $*$ coverage (1000x). Cells were seeded at low confluency to avoid contact inhibition between single clones (2D cells - 30,000 cells per 10 cm dish; organoids - 50,000 cells/well, 24-well format) and in the presence of PARPi (KB2P1.21/KB2P3.4: 200 nmol olaparib, 300 nmol AZD2461; KP3.33: olaparib 10 μmol; ORG-KB2P26S.1: 25 nmol AZD2461). Cells were selected with PARPi for 3 weeks, and media was refreshed at least twice per week. PARPi-surviving clones were pooled and genomic DNA was extracted (QIAmp DNA Mini Kit, Qiagen). shRNA sequences were retrieved by a two-step PCR amplification, as described before[22]. To maintain screening coverage, the amount of genomic DNA used as an input for the first PCR reaction was taken into account (6 μg of genomic DNA per 106 genomes, 1 μg per PCR reaction). Resulting PCR products were purified using MinElute PCR Purification Kit (Qiagen) and submitted for Illumina sequencing. Sequence alignment and enrichment analysis (day 0 vs PARPi-treated population) was carried out using MaGECK software[23].

Long-term clonogenic assays

Long-term clonogenic assays were performed in 6-well (KB2P1.21, KB2P3.4, KB1P-G3, KP3.33) or 12-well plates (SUM149PT, DLD-1 BRCA2(-/-)). Cells were seeded at low density to avoid contact inhibition between the clones (KB2P1.21 - 3,000 cells/well, KB2P3.4 and KP3.33 - 2,000 cells/well, KB1P-G3 - 5,000 cells/well, SUM149PT and DLD-1 BRCA2(-/-) - 3,000 cells/well) and cultured for 2 weeks, except for control (DMSO-treated) KB2P1.21, KB2P3.4, KB1P-G3 and KP3.33 cells, which were stopped after 7 days. Media was refreshed at least twice a week. For the quantification, cells were incubated with Cell-Titer Blue (Promega) reagent and later fixed with 2 % formaldehyde and stained with 0.1 % crystal violet. Clonogenic assays

with cells with CRISPR/Cas9-based genome editing were quantified with TIDE, similarly to a previously described method[75]. Drug treatments: cells were grown in the continuous presence of temozolomide, PARPi (olaparib, talazoparib or AZD2461), and/or PARGi (PDDX-004), at the indicated concentrations. All compounds were reconstituted in DMSO (PARP/PARG inhibitors: 10 mmol, temozolomide: 5 mg ml^{-1}). For the ioninizing irradiation studies, cells received single irradiation doses 24 hr after seeding. IR treatments were carried out using Gammacell 40 Extractor (Best Theratronics Ltd.).

RT-qPCR

To determine gene expression levels, total RNA was extracted from cultured cells using ISOLATE II RNA Mini Kit (Bioline) and used as a template to generate cDNA with Tetro cDNA Synthesis Kit (Bioline; oligo $(dT)_{18}$ mix). Quantitative RT-PCR was performed using SensiMix SYBR Low-ROX Kit (Bioline; annealing temperature - 60 °C) in a Lightcycler 480 384-well plate (Roche), and analyzed using Lightcycler 480 Software v1.5 (Roche). Mouse *Hrtp* and human *HPRT* were used as housekeeping genes (control). The primer sequences used in this study are listed in Table S7.6.

PAR immunoblotting

Tumor-derived organoids or 2D cells were washed with PBS, trypsinized and then lysed for 30 min in RIPA buffer supplemented with protease inhibitors (cOmplete Mini EDTA-free, Roche) and 1 μmol PARG inhibitor ADP-HPD (Merck). The protein concentration was determined using Pierce BCA Protein Assay Kit (Thermo Scientific). SDS-Page was carried out with the Invitrogen NuPAGE SDS-PAGE Gel System (Thermo Fisher; gel: 4 % to 12 % Bis-Tris, buffer: MOPS, input: 50 μg protein), according to the manufacturer's protocol. Next, proteins were electrophoretically transferred to a nitrocellulose membrane (Biorad) and then the membrane was blocked in 5 % (w/v) milk (PAR) solution in Tris-buffered saline Triton X-100 buffer (TBS-T; 100 mmol Tris, 7.4 pH, 500 mmol NaCl, 0.1 % (v/v) Triton X-100). Membranes were incubated overnight with primary antibodies in blocking buffer, at 4 °C. Horseradish peroxidase (HRP)-conjugated secondary antibody incubation was performed for 1 hr at room temperature in blocking buffer and signals were visualized by ECL (Pierce ECL Western Blotting Substrate, Thermo Scientific). Primary antibodies used in this study: mouse monoclonal anti-PAR (H10) (Millipore), 1:1000; rabbit polyclonal anti-PARP1 (#9542, Cell Signaling), 1:1000; rabbit polyclonal anti-PARP3 (#ALX-210-541- R100, Enzo Life Sciences), 1:500; rabbit polyclonal anti-Histone H3 (#ab1791, Abcam), 1:5000; mouse monoclonal anti-GADPH (6C5) antibody (Santa Cruz), 1:5000 dilution. Secondary antibody used in this study: rabbit polyclonal anti-mouse immunoglobulins/HRP (Dako), diluted 1:5000.

PARP1 trapping assay

The PARP1 trapping assay was adapted from Murai *et al.*[7]. In brief, 24 hr prior the assay, KB2P cells were seeded on 10 cm dishes to achieve ~90 % confluency. Drug treatments: (1) KB2P cells with genetic depletion of PARG (Figure 7.4B) were treated with olaparib (0, 0.1 or 1 μmol) and 0.01 % MMS for 2 hr; (2) KB2P cells with chemical inhibition of PARG (time-course experiment, Figure 7.4C) were first pre-incubated with 0.5 μmol olaparib and/or1 μmol PDDX-004 for 1 hr, and then exposed to the same treatments but in a presence of 0.01 % MMS for 30 min; following incubation with MMS cells were further incubated with olaparib and/or PDDX-004 for 0.5 or 2 hr, as shown on Figure 7.4C. After indicated treatments cells were trypsinized and subsequently lysed to isolate chromatin-bound fractions. Fractionation was performed with Subcellular Protein Fractionation Kit from Thermo Scientific (#78840, Rockford, IL, USA), following the manufacturer's instructions and in the presence of 1 mM of PARGi ADP-HPD (#118415, Calbiochem) in the lysis buffers. Immunoblotting was carried out as described in previous section (Immunoblotting). Experiments were repeated three times.

PARG activity assay

Enzymatic activity of endogenous PARG was measured using HT Colorimetric PARG Assay Kit (#4683-096-K, Trevigen), following the manufacturer's instructions. Experiments were repeated three times.

PAR ELISA assay

To measure endogenous PAR levels cells were seeded on 6-well plate 24 hr prior to PARPi, PARGi or combined PARPi/PARGi treatment to achieve ~90 % confluency at the day of the assay. Cells were treated with different doses of inhibitors for 2 hr, and additionally some of the samples were exposed to 0.01 % MMS for the last 60 min. Cell lysis, protein isolation and PAR ELISA were carried out using HT PARP In Vivo Pharmacodynamic Assay II kit (#4520-096-K, Trevigen), according to the manufacturer's protocol. Protein concentration was measured with Pierce BCA Protein Assay kit (Thermo Scientific).

PAR immunofluorescence analysis

PAR levels were measured using the adapted immunofluorescent PAR cell assay described before[76]. Briefly, cells were seeded on Corning 96-well special optics plates (#CLS3720, Sigma) 24 hr prior the assay to achieve ~90 % confluency. Next, cells were treated with a range of PARGi doses for 2 hr, and for the last 60 min cells were additionally exposed to 0.01 % MMS. After incubation with drugs, plates

were fixed with ice-cold 95 % (v/v) methanol/PBS (100 ml/well) for 15 min at −20 °C. Plates were then washed twice with PBS and cells were permeabilized by adding 100 μl per well of 0.1 % (v/v) Triton X-100 in PBS and incubating for 20 min at room temperature. Incubation with the primary mouse monoclonal anti-PAR (H10) antibody (Millipore), diluted 1:4000 in PBS solution containing 5 % (v/v) FBS and 0.05 % (v/v) Tween-20, was carried out overnight at 4 °C. After three washes with PBS, cells were incubated for 1 hr (room temperature) with polyclonal AlexaFluor488 goat anti-mouse immunoglobulins (1:1000) and Hoechst (1:5000; Thermo Scientific) diluted in 5 % (v/v) FBS/0.1 % (v/v) Triton X-100 in PBS. PAR immunofluorescent signal was detected with a Leica TCS SP8 confocal system (Leica Microsystems), using a HC PL APO 40x/1.10 W objective. Total nuclear intensities were measured per nuclei with ImageJ software. For each well, four different areas (200 cells on average) were imaged and analyzed. Each experiment was repeated three times.

RAD51/53BP1 IRIF analysis

Cultured cells Cells were seeded on Millicell EZ slides (#PEZGS0816, Millipore) 24 hr prior the assay to achieve ∼90 % confluency. Cells were then irradiated using the Gammacell 40 Extractor (Best Theratronics Ltd.) at the dose of 10 Gy and allowed to recover for 4 hr. Next, cells were fixed with 4 % (w/v) solution of formaldehyde in PBS for 30 min and permeabilized in 0.2 % (v/v) Triton X-100 in PBS^{++} (PBS solution containing 1 mmol $CaCl_2$ and 0.5 mmol $MgCl_2$) for 20 min. To minimize the background, cells were further incubated for 30 min in staining buffer (1 % (w/v) BSA, 0.15 % (w/v) glycine and 0.1 % (v/v) Triton X-100 in PBS^{++}). Staining buffer was also used as a solvent for antibodies - primary antibodies: rabbit anti-RAD51 (gift from Roland Kanaar, Erasmus MC, Rotterdam), diluted 1:5000, rabbit polyclonal anti-53BP1 (Abcam), diluted 1:1000, secondary antibody - goat polyclonal anti-rabbit, Alexa FluorO®658-conjugated, diluted 1:1000. Incubation with primary and secondary antibodies was done for 2 hr and 1 hr, respectively. All incubations were performed at room temperature. Samples were mounted with VECTASHIELD Hard Set Mounting Media with DAPI (#H-1500; Vector Laboratories). Images were captured with Leica SP5 (Leica Microsystems) confocal system and analyzed using an in-house developed macro to automatically and objectively evaluate the DNA damage-induced foci[22]. Fraction of positive cells was determined for each sample using following criteria: RAD51 > 5 foci/nucleus, 53BP1 > 10 foci/nucleus. Experiment was performed in triplicate (on average 100 cells/replicate). As a positive control for RAD51 staining, BRCA-proficient KP3.33 cells were used.

***In situ* analysis of GEMM tumors** Matched PARPi-naive and -resistant KB2P tumors, and KP (*K14cre;p53$^{F/F}$*) tumor (positive control) were re-transplanted into wild-type

syngeneic mice. Upon tumor outgrowth to 500 mm tumors were locally irradiated (dose: 15 Gy) using CT-guided high precision cone beam micro-irradiator (X-RAD 225Cx) or left untreated (control). Two hr post-irradiation tumors were isolated and fixed in 4 % (w/v) solution of formaldehyde in PBS. FFPE material was then used for immunofluorescent staining. First, samples were deparaffinized and antigen retrieval was done by cooking samples in DAKO Target Retrieval Solution pH 9 (#S236784, DAKO) for 20 min in microwave at ~600 W. Next, samples were permeabilized in 0.2 % (v/v) Triton X-100 in PBS for 20 min and further incubated with DNase (1000 u ml^{-1}; #04536282001, Roche) in humidified chamber for 1 hr at 37 °C. Incubation with antibodies, mounting, imaging and analysis were carried out as described for cultured cells. At least five different areas were imaged and analyzed for each sample. All incubations were performed at room temperature, unless otherwise stated.

Immunofluorescent staining of Olaparib-Induced γH2AX foci Cells were seeded on Corning 96-well special optics plates (#CLS3720, Sigma) 24 hr prior the olaparib treatment to achieve ~60 %. Next, cells were incubated with olaparib for the next 24 hr and subsequently fixed and stained following the same protocol as described for RAD51/53BP1 IRIF assay (cultured cells). Primary antibody used in this assay: Phospho-Histone H2A.X (Ser139) rabbit monoclonal antibody, Cell Signaling, #2577 (1:200 in staining buffer). Immunofluorescent signal was detected with a Leica TCS SP8 confocal system (Leica Microsystems), using a HC PL APO 40x/1.10 W objective. Foci were quantified using in-house developed macro to automatically and objectively evaluate the DNA damage-induced foci[22]. For each condition, four different areas (200 cells on average) were imaged and analyzed. Experiment was performed in triplicate and repeated twice.

DNA fiber assay

Replication fork progression assay Fork progression was measured using an adapted method described previously in Zellweger *et al.*[35]. Briefly, asynchronously growing U2OS cells were first incubated with 0.5 μmol PARPi (olaparib or AZD2461) for 1 hr and further treated with 0.5 μmol PARPi alone, or in combination with 1 μmol PDDX-001, for another 60 min. Next, cells were labeled with 30 μmol CIdU (Sigma-Aldrich) for 30 min, washed twice with PBS, and labeled with 250 μmol IdU in the presence of genotoxic agents - 50 μmol MMS or 25 nmol CPT for another 30 min. Following pulse labelling, cells were quickly trypsinized and resuspended in PBS at $2.5 * 10^5$ cells per ml. Next, labeled cells were diluted 1:8 with unlabeled cells, and 2.5 ml of cells were mixed with 7.5 ml of lysis buffer (200 mmol Tris-HCl, 7.5 pH, 50 mmol EDTA, and 0.5 % (w/v) SDS) on a glass slide. After 9 min, the slides were tilted at 15-45, and the resulting DNA spreads were air dried, fixed in 3:1

methanol/acetic acid, and refrigerated overnight. The DNA fibers were denatured with 2.5 mol HCl for 1 hr, washed with PBS, and blocked with 2 % (w/v) BSA in PBST (PBS and Tween 20) for 40 min. The newly replicated CldU and IdU tracks were labeled (for 2.5 hr in the dark, at RT) with anti-BrdU antibodies recognizing CldU (rat; Abcam) and IdU mouse (BD), respectively. After washing for 533 min in 0.2 % (v/v) PBS-T, the following secondary antibodies were used (1 hr incubation, in the dark, at RT): anti-mouse Alexa Fluor 488 (Molecular Probes) and anti-rat Cy3 (Jackson ImmunoResearch Laboratories, Inc.). After washing for 533 min in PBS-T (0.2 % (v/v)), the slides were air dried completely, mounted with 20 μl per slide Antifade gold (Invitrogen), and sealed to a coverslip by transparent nail polish. Microscopy was performed with a fluorescence microscope (IX81; Olympus; objective lenses: LC Plan Fluor 603, 1.42 NA oil) and acquired with a charge-coupled device camera (Orca AG; Hamamatsu Photonics). The images were processed with CellR software (version 2.6; Olympus). On average 120 fibers were quantified per condition; experiment was repeated twice.

Replication fork protection assay KB2P cells were seeded 24 hr prior the assay to achieve ~75 % confluency. On the next day, cells were pulse-labeled with CldU (25 μmol) for 20 min, washed with warm medium and pulse-labeled with IdU (250 μmol) for another 20 min. Next, cells were incubated for 3 hr with 0.9 μmol MMS to induce replication stress. After MMS incubation, cells were trypsinized, washed and lysed on microscopy slides in lysis buffer (0.5 % (w/v) sodium dodecyl sulfate (SDS), 200 mmol Tris 7.4 pH, 50 mmol EDTA). DNA fibers were spread by tilting the slide, air-dried and fixed in methanol/acetic acid (3:1) for 10 min. Fixed DNA spreads were treated with 2.5 mol HCl for 75 min. Immunofluorescent staining was performed as described in the previous section (RF progression assay). Images were acquired on a Leica DM-6000RXA fluorescence microscope, with Leica Application Suite software. CldU and IdU track lengths were measured using ImageJ software. At least 100 IdU and CldU ratios per condition were analyzed.

Neutral comet assay Asynchronously growing U2OS cells were first incubated with 0.5 μmol olaparib for 1 hr and further treated with 0.5 μmol olaparib alone, or in combination with 1 μmol PDDX-001, for another 60 min. Next, DNA damage was induced by 30 min incubation with 25 nmol CPT in the presence of PARPi and/or PARGi. Cells were then collected by trypsinization and resuspended in PBS at the concentration of 106 cells per ml. 20 μl of cell suspension was then loaded onto 600 μl of 0.8 % (w/v) Low Melting Point (LMP) agarose (Lonza) in PBS, previously equilibrated to 37 °C (for 60 min). Next, 60 μl of cells resuspended in LMP was spread on the comet slide (CometAssay®Kit, Trevigen). Slides were incubated at 4 °C for 15 min to allow solidification of the cells-LMP mixture and further incubated in a lysis buffer (CometAssayO®Lysis Solution, Trevigen) overnight at 4 °C. On the

next day, slides were first incubated in electrophoresis buffer (300 mmol sodium acetate, 100 mmol Tris, 8.3 pH) for 1 hr and then electrophoresis was performed using the comet chamber at 21 Volt for 30 min, at 300 mA and 4 °C. Next, slides were rinsed twice in water, placed in 70 % ethanol at 4 °C for 20 min and incubated at 37 °C until dry. Slides were than stained with SYBR®Gold (Thermo Fisher Scientific) diluted at 1:30,000 in Tris-EDTA (10 mmol Tris-HCl 7.5 pH, 1 mmol EDTA) for 30 min in dark. Images were recorded on a Leica DM6 B upright digital research microscope at 10x magnification. The images were analyzed using the Open Comet plugin (http://www.cometbio.org/) for Fiji. At least 95 cells were analyzed per sample and plotted using GraphPad PRISM 7.0c. The experiment was repeated twice.

Fast micromethod DNA single-strand break assay The assay was performed as described before[40]. Briefly, KB2P1.21 cells were seeded on Corning 96-well special optics plates (#CLS3720, Sigma) 24 hr prior the assay (10,000 cells/well). On the next day, cells were first treated with olaparib and 0.01% MMS for 30 min to induce base damage. After that incubation plates were either processed (time point 0) or further incubated with olaparib for 3 hr (time point 3 hr). Following incubation with drugs, media was removed and cells were incubated for 1 hr in dark with 50 ml of 1 % (v/v) solution of PicoGreen (Thermofisher) in PBS. This incubation was performed at 4 °C to avoid additional DNA repair. After 1 hr 250 μl of NaOH solution, adjusted to a final pH of 12.45 with EDTA, was added to cells and the measurement of the fluorescence signal (485 nm/538 nm) was immediately started using Tecan Infinite 200 PRO (Tecan) plate reader. Measurements were taken every 60 s for 25 min. The data were analyzed as described before[40]. BCA protein measurements were taken to assure equal amount of cells per condition. The experiment was performed in triplicate and repeated twice.

Laser Micro-Irradiation Assays

PARP1-GFP: Multiphoton Laser Mico-Irradiation U2OS cells were transiently transfected (using Lipofectamine 2000, Invitrogen; according to the manufacturer's protocol) with pEGFP-c3-PARP1 (kind gift from Valerie Schreiber) vector. Cells grown on 18 mm coverslips were placed in a Chamlide CMB magnetic chamber and the growth medium was replaced by CO_2-independent Leibovitz's L15 medium supplemented with 10 % (v/v) FCS and penicillin-streptomycin. Cells were treated with DMSO (control), 0.5 μmol olaparib, 1 μmol PDDX-001 or olaparib/PDDX-001 combination for 1 hr prior to micro-irradiation. Laser micro-irradiation was carried out on a Leica SP5 confocal microscope equipped with an environmental chamber set to 37 °C. DSB-containing tracks (1.5 μm width) were generated with a Mira modelocked titanium-sapphire (Ti:Sapphire) laser (l = 800 nm, pulse length =

200 fs, repetition rate = 76 MHz, output power = 80 mW) using a UV-transmitting 633 1.4 NA oil immersion objective (HCX PL APO; Leica). Confocal images were recorded before and after laser irradiation at 5 s or 10 s time intervals over a period of 2-3 min. Images recorded after multi-photon micro-irradiation of living cells were analyzed using ImageJ software. The average pixel intensity of laser tracks was measured within the locally irradiated area function (I_{damage}), in the nucleoplasm outside the locally irradiated area ($I_{\text{nucleoplasm}}$) and in a region not containing cells in the same field of view ($I_{\text{background}}$). The relative level of accumulation expressed relative to the protein level in the nucleoplasm was calculated as follows: $(I_{\text{damage}} - I_{\text{background}}) - (I_{\text{nucleoplasm}} - I_{\text{background}})$.

XRCC1-GFP: UV-A Laser Micro-Irradiation The laser micro-irradiation assay was performed as described before[73]. Briefly, U2OS cells were transiently transfected (using Lipofectamine 2000, Invitrogen; according to the manufacturer's protocol) with pEGFP-C1-XRCC1 (gift from Simon Bekker-Jensen). Cells were grown on 18 mm coverslips and sensitized with 10 μmol 50-bromo-2-deoxyuridine (BrdU) for 24 hr. On the next day, the growth medium was replaced with CO_2-independent Leibovitz's L15 medium supplemented with 10 % FCS and penicillin-streptomycin and cells were kept at 37 °C. Cells were treated with DMSO (control), 0.5 umol olaparib, 1 μmol PDDX-001 or olaparib/PDDX-001 combination for 1 hr prior to micro-irradiation. For micro-irradiation, the cells were placed in a Chamlide TC-A live-cell imaging chamber that was mounted on the stage of a Leica DM IRBE widefield microscope stand (Leica) integrated with a pulsed nitrogen laser (Micropoint Ablation Laser System; Andor). The pulsed nitrogen laser (16 Hz, 364 nm) was directly coupled to the epifluorescence path of the microscope and focused through a Leica 403 HCX PLAN APO 1.25-0.75 oil-immersion objective. The laser output power was set to 60 to generate strictly localized subnuclear DNA damage, resulting in XRCC1, but not in XRCC4 accumulation[73]. Following micro-irradiation, cells were incubated for the indicated time points at 37 °C in Leibovitz's L15 and subsequently fixed with 4 % formaldehyde before immunostaining. Typically, an average of five cells was micro-irradiated (2 iterations per pixel) within 1 min using Andor IQ software (Andor). For each condition, 20-25 cells were micro-irradiated, and experiments were repeated three times (60-75 cells in total). Following DNA damage, cells were further incubated for 5, 10, 15 or 60 min. Next, cells were fixed with 2 % formaldehyde in PBS for 20 min. Cells were post-extracted with 0.5 % Triton-X100 (Sigma) in PBS, and blocked with 20 mmol glycine and 0.5 % BSA in PBS for 15 min and used for antibody incubations. Samples were incubated with primary antibodies overnight at 4 °C and with secondary antibodies supplemented with 0.1 mg ml^{-1} DAPI for 1 hr at room temperature. Samples were subsequently mounted in Polymount (Polysciences). Antibodies used in this study are: primary rabbit polyclonal anti-MDC1 (#ab11171-50, Abcam), diluted 1:1000 and secondary

goat anti-rabbit AlexaFluor-555-conjugated, diluted 1:1500 (#A-21429, Thermo Fisher). The MDC1 (AF-555), XRCC1-GFP (GFP) and CHD2-GFP (GFP) signal was detected using Zeiss AxioImager D2 widefield fluorescence microscope equipped with 403, 633 and 1003 PLAN APO (1.4 NA) oil-immersion objectives (Zeiss) and an HXP 120 metal-halide lamp used for excitation. Images were recorded using the ZEN 2012 software. Tracks were evaluated in ImageJ, by calculating the ratio of track to background (nuclear) intensities for each channel.

PARPi toxicity analysis in a panel of 1,001 cancer cell lines

The pharmacogenomics dataset of 1,001 human cancer cell lines, published recently[41], was used to assess the corelation between PARG or PARP1 expression and IC_{50} values of PARPi. For this purpose, expression data for DDR-related genes (see Table S7.2) and drug response data to olaparib, veliparib, rucaparib and talazoparib was extracted, resulting in a smaller dataset for 935 out of 1,001 cell lines. For statistical analysis Pearson correlation was estimated for each gene and the correlation p values were computed using the relation between the estimated coefficient and the student-t distribution. Negative correlation means that a lower IC_{50} associates with higher expression (= more sensitive).

7.6.4 Quantification and statistical analysis

Statistical parameters including sample size, precision measures and statistical significance are reported in the figures, corresponding figure legends and Method Details sections.

Genetic screens See Figure 7.1, Figure 7.2K and Table S7.1. Genetic screens were performed in triplicate and statistical analysis was carried out using the MAGeCK software.

***In vivo* studies** See Figure 7.2.

Survival analysis The effect of PARPi treatment in mice carrying KB2P tumors was measured by survival analysis of control ($n = 21$) vs AZD2461-treated group ($n = 34$). Data are presented as Kaplan-Meier curves and the p value was computed using Log-Rank (Mantel Cox) statistics (Figure 7.2C).

Statistical analysis of deep-sequencing data See Figure 7.2D, Figure 7.2G, and Figure S7.1. Detailed description of the bioinformatics analysis is reported in the Method Details section.

Immunohistochemistry analysis of PARG expression Semi-quantitative comparison of the PARG staining of resistant ($n = 34$) versus naive ($n = 21$) KB2P tumors was carried out by a trained pathologist who was blinded regarding the sample identity. Group comparison was performed using Mann-Whitney U test. Representative images are shown (Figure 7.2H).

qRT-PCR analysis Gene expression measurements were performed in triplicate, normalized to expression of house keeping genes, and presented as mean $\pm$ SD of replicates. Statistical significance was estimated with the two-tailed unpaired t-test. See Figure 7.3A, Figure 7.6A-B, Figure S7.2A, Figure S7.2E, and Figure S7.3B.

PARG activity assays and PAR ELISA assay Experiments were performed in triplicate, repeated three times and presented as mean $\pm$ SD of repeats. Statistical significance was estimated with the two-tailed unpaired t-test. See Figure 7.2I, Figure 7.3B, and Figure S7.2B for the PARG activity assays and Figure 7.4A for the PAR ELISA assay.

PAR immunofluorescence assay Intensities of PAR signal were measured per nuclei, for four different imaging areas per condition, each containing on average 200 cells; the experiment was repeated three times. Results were presented as a mean for each area (12 areas in total) $\pm$ SD. Statistical analysis was performed using the two-tailed unpaired t-test. See Figure S7.2D.

Long-term clonogenic assays See Figures 7.3C, 7.3E, 7.4F, 7.6F-I, S7.3A, S7.3C, S7.3I, S7.3J, S7.5D, S7.6A, S7.6C, S7.6G, and S7.6H. All experiments were repeated three times, unless otherwise stated, and data are presented as mean $\pm$ SD of replicates. For statistical analysis the two-tailed unpaired t-test (Figures 7.3C, 7.4F, S7.3A, S7.3C, S7.3H, S7.5D, and S7.6A) or ANOVA (Figures 7.3E, S7.6G-J, S7.3G, S7.6B, S7.5F, and S7.6G) test was used.

Analysis of DNA damage-induced foci See Figures S7.5G, S7.4B, and S7.4D. For the analysis of γH2A foci (Figure 7.5G), foci were counted per nuclei, in four different imaging areas per condition, each containing on average 200 cells. Results were presented as a mean $\pm$ SD of the three replicates. Statistical analysis was performed using the two-tailed unpaired t-test. For the analysis of RAD51/53BP1 irradiation-induced foci (Figures S7.4B and S7.4D), foci were counted per nuclei and data are presented as mean $\pm$ SD of replicates (triplicate for cultured cells and five replicates for GEMM samples) and group comparison was performed using the two-tailed unpaired t-test. On average 100 cells were analyzed per replicate.

Fast micromethod SSB assay See Figure 7.5F. Assay was performed in triplicate and repeated twice, data are shown as mean ± SD of the two repeats. Statistical significance was assessed using two-tailed unpaired t-test.

DNA fiber assays See Figures 7.5A, S7.5A, and S7.5I. For the fork protection assay (Figure S7.4A), at least 100 fibers were measured per condition and statistical significance was assessed with the Mann-Whitney U test. For fork progression analysis (Figures 7.4C and S7.5I), on average 120 fibers were quantified and group comparison was performed with the two-tailed unpaired t-test. Experiments were repeated twice.

Neutral comet assay On average 100 comets (cells) were analyzed per condition and each data point represents tail moment (Figure 7.5B) or olive moment (Figure S7.5J) of a single comet, together with a mean ± SD of all cells per condition. Statistical significance was assessed using Mann-Whitney U test. Experiments were repeated twice.

Micro-irradiation assays See Figures 7.4E, 7.5D-E, and S7.5F-G. In each experiment on average 15-20 cells were micro-irradiated and analyzed per condition. Experiments were repeated three times and data are presented as mean ± SEM of independent experiments. Statistical analysis was performed using two-tailed unpaired t-test.

Immunoblotting See Figures 7.2J, 47.4B-C, and S7.5A-B. Immunoblotting experiments were repeated three times. Representative images are shown.

PARPi Toxicity Analysis in a Panel of 1,001 Cancer Cell Lines See Figures 7.6C and S7.6D. Correlation between gene expression and IC_{50} values for PARPi was assessed using Pearson correlation, which was estimated for each gene and the correlation p values were computed using the relation between the estimated coefficient and the student-t distribution.

7.6.5 Data and software availability

The deep sequencing data generated in this study is available in the European Nucleotide Archive (ENA) under accession number PRJEB20535. Raw data from this manuscript: https://doi.org/10.17632/r9cm2kh8mt.1.

7.7 Supplementary Material

7.7.1 Supplementary Figures

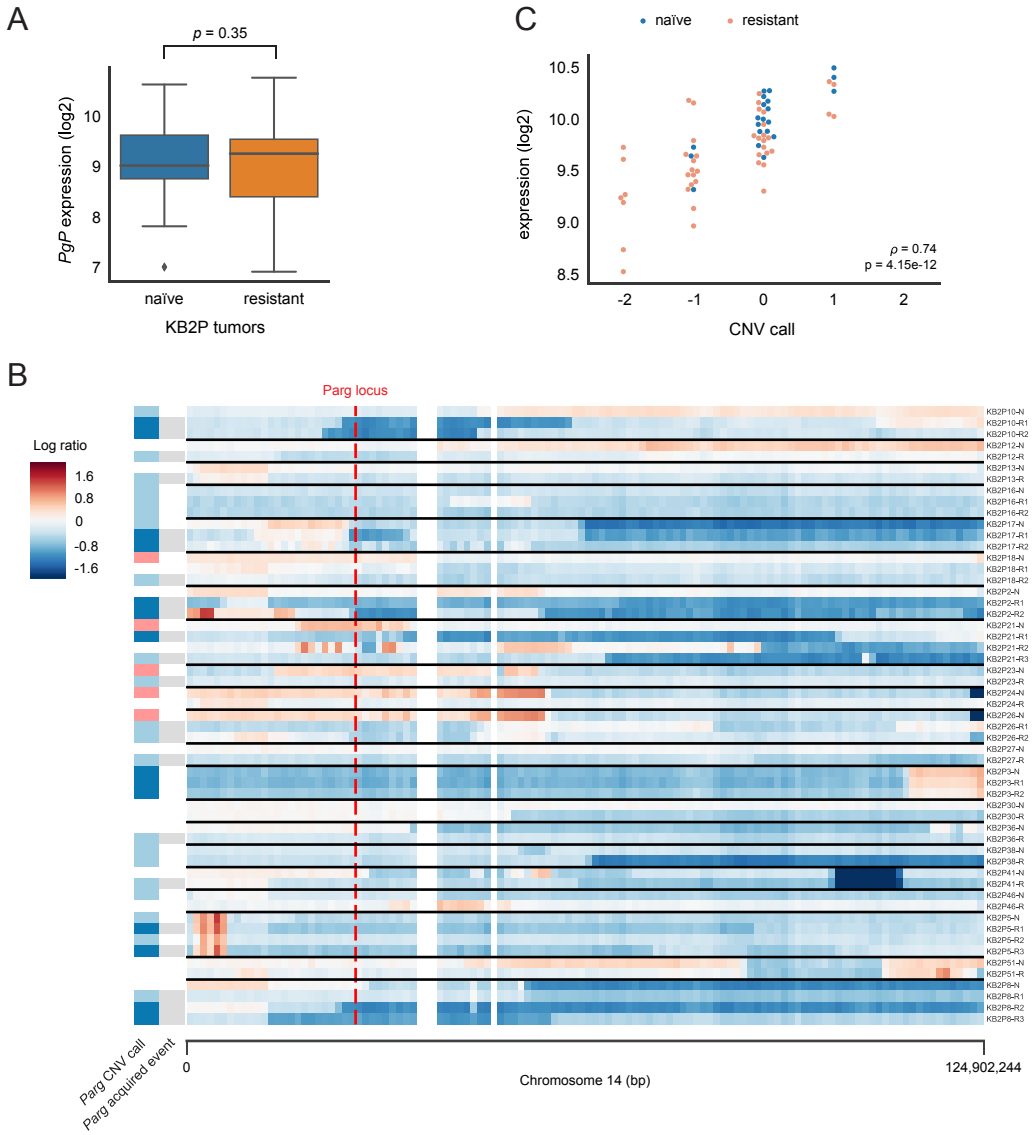


Fig. S7.1. **Multi-omics analysis of PARPi-resistant mouse mammary tumors. Related to Figure 7.2. (A)** Expression (RNA-seq) of Pgp drug efflux transporter in KB2P PARPi-naive and resistant tumors; p value: two-tailed t-test. The box extends from the lower to upper quartile values of the data, with a line at the median; whiskers (whis) are a function of the inner quartile range, they extend to the most extreme data point within (whis $*$ $(75\% - 25\%)$) data range. **(B)** Copy number $\log_2$ ratios of chromosome 14 shown for all KB2P tumors; the *Parg* locus is indicated by the red dashed line. KB2Px-R/Ny: x - original donor ID number, y - ID of individual resistant tumors derived from the same donor tumor, R - resistant, N - naive. **(C)** Correlation between *Parg* expression and copy number estimation for a panel of KB1P tumors. Spearman correlation coefficient (ρ) and p value (DIDS algorithm) are indicated.

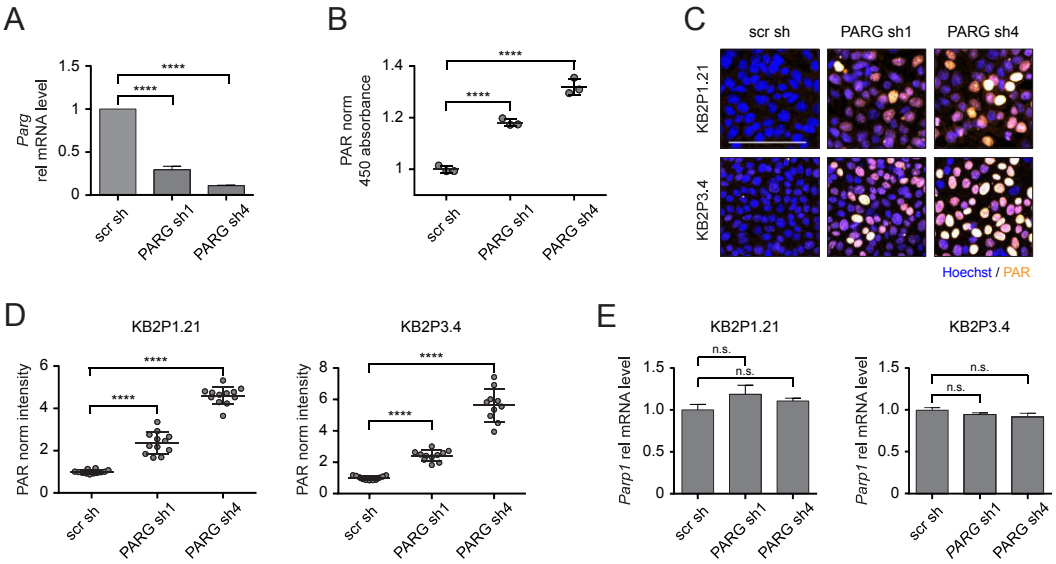


Fig. S7.2. Suppression of PARG in KB2P cells. Related to Figure 7.3. **(A)** RT-qPCR analysis of Parg expression in KB2P3.4 cells expressing indicated shRNAs; $^{****}p < 0.001$ (two-tailed t-test), data shown as mean $\pm$ SD of triplicate. **(B)** ELISA-based PARG activity assay in the cell lines indicated; data represent mean $\pm$ SD of triplicate, $^{****}p < 0.001$ (two-tailed t-test). **(C-D)**Immunofluorescence-based PAR detection in the indicated cell lines upon MMS (0.01 %) treatment. Representative microscopic images (C) and quantification of the assay (D) are shown; $^{****}p < 0.001$ (two-tailed t-test), data are presented as mean $\pm$ SD of single areas imaged (experiment repeated three times, four areas/experiment, average 300 cells/area); scale bar, 100 μm. **(E)** RT-qPCR analysis of *Parp1* gene expression in KB2P PARG kd cells; n.s. - not significant (two tailed t-test), data are shown as mean $\pm$ SD of triplicate experiments.

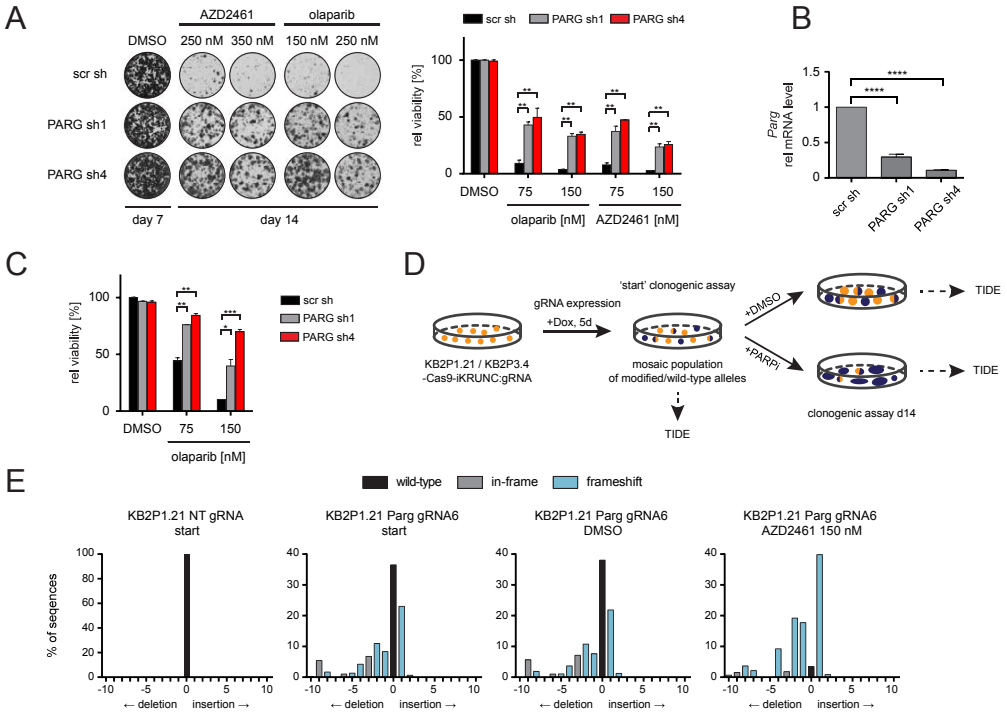


Fig. S7.3. Downregulation of PARG drives PARPi resistance in KB2P cell lines. Related to Figure 7.3. **(A)** Long-term clonogenic assay using KB2P3.4-scr sh (scramble shRNA, control), -PARG sh1 and -PARG sh4 cells and the PARPi AZD2461 or olaparib. Representative images of stained wells (left) and quantification of the assay (right; mean ± SD of three repeats) are shown; $^{**}p < 0.01$ (two-tailed t-test). **(B)** qRT-PCR analysis of *Parg* expression in KB1P-G3 cells expressing the hairpins indicated; $^{****}p < 0.001$ (two-tailed t-test), data shown as mean ± SD of triplicate. **(C)** Clonogenic assay in KB1P-G3 cells expressing the hairpins indicated and treated with olaparib; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (two-tailed t-test); data represents mean ± SD of two repeats. **(D)** Outline of the long-term clonogenic assays in KB2P cells expressing control (NT) or gRNAs targeting *Parg*. **(E)** Example of the TIDE analysis: spectrum of alleles identified by the algorithm in the samples indicated.

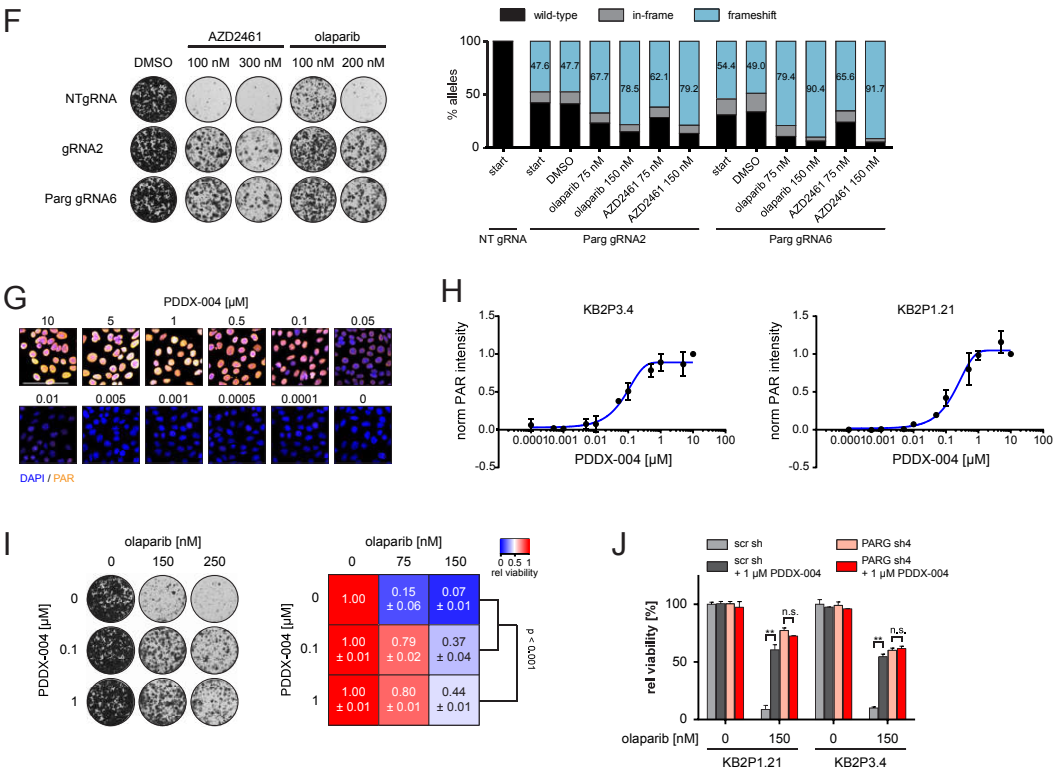


Fig. S7.3. **Continued. (F)** Long-term clonogenic assay using KB2P3.4 cells expressing the indicated gRNAs. Representative images (left) and TIDE quantification of allele composition (right) are shown. **(G-H)** Chemical inhibition of PARG using PDDX-004 in the cell lines indicated. Representative images of PAR immunofluorescent staining (G) and quantification (H) are shown; data represent mean ± SD of three repeats; scale bar, 100 μm. **(I)** Long-term clonogenic assay using KB2P3.4 cells treated with the PARPi olaparib and PARGi PDDX-004 alone or in combination. Representative images of stained cells (left) and quantification of the assay (right) are shown. Data on the graph represent experiments repeated three times (mean ± SD); p value (ANOVA) is indicated. **(J)** Clonogenic assay in KB2P cells expressing the shRNAs indicated, and treated with olaparib; $^{**}p < 0.01$, n.s. - not significant (two-tailed t-test); data shown as mean ± SD of triplicate.

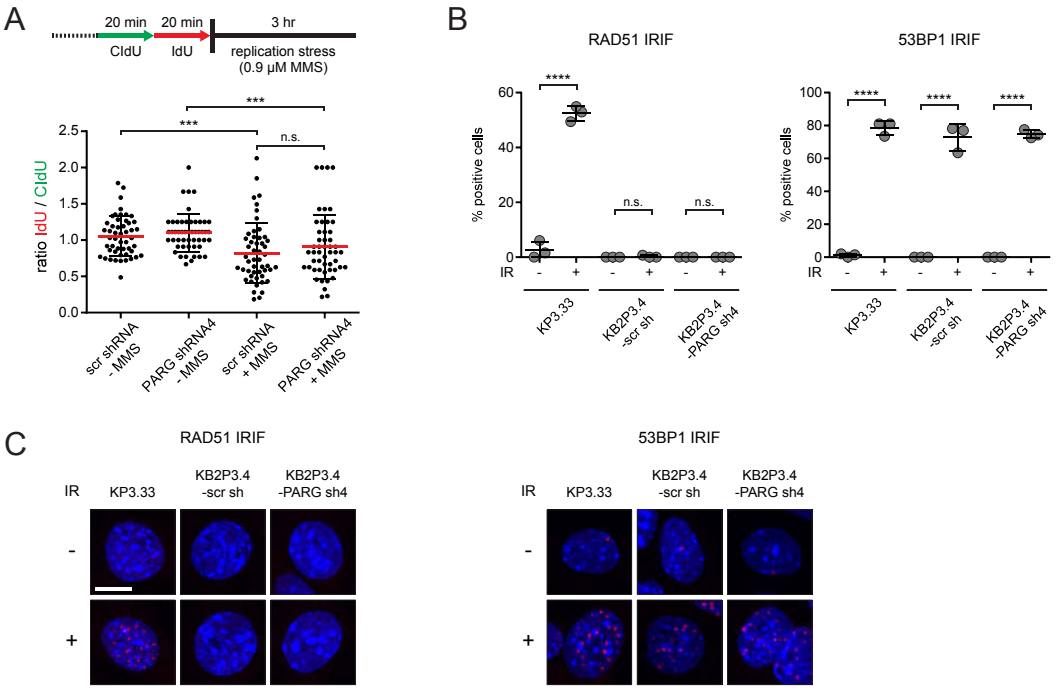


Fig. S7.4. **PARG loss causes PARPi resistance independently of HR and RF protection. Related to Figure 7.4. (A)** DNA fiber assay using KB2P3.4 cells treated with MMS (0.9 μmol); $^{***}p < 0.001$, n.s. - not significant (Mann-Whitney U test). Error bars represent SD and red lines represent mean values. **(B-C)** RAD51/53BP1 IRIF formation assay in the given cell lines; KP3.33 ($p53^{-/-}$) cells were used as HR-proficient control in this assay; 53BP1 served as DNA damage marker. Quantification (B) and representative images (C) are shown. A cell was considered positive when: RAD51 $\geq$ 5 foci/nuclei, 53BP1 $\geq$ 10 foci/nuclei. Data represent mean $\pm$ SD; $^{****}p < 0.0001$, n.s. - not significant (two-tailed t-test); scale bar, 10 μm.

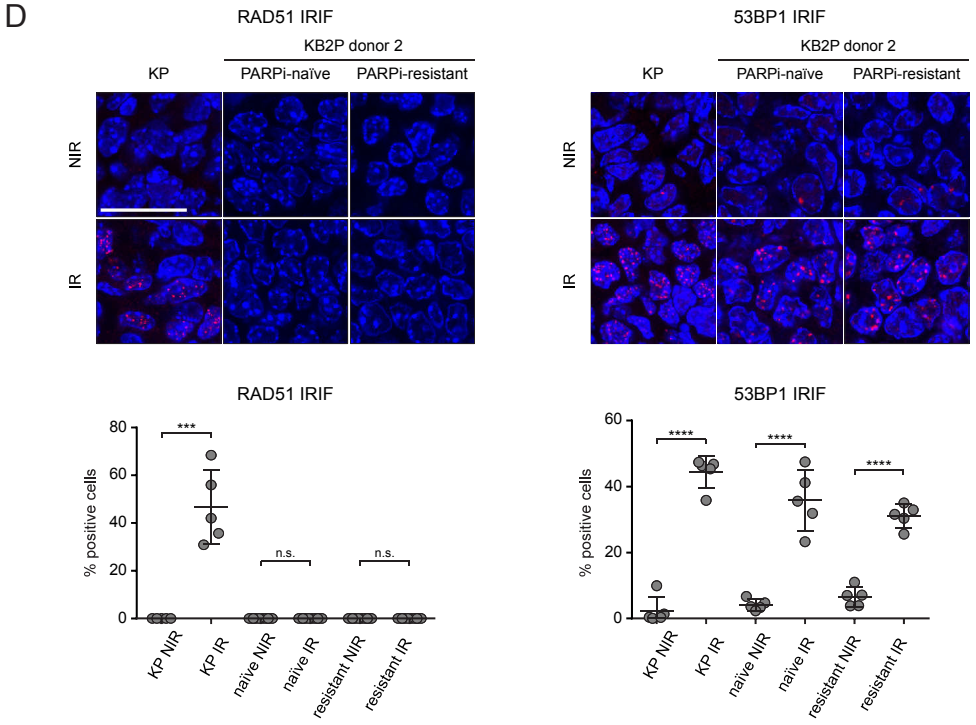


Fig. S7.4. **Continued. (D)** Example of RAD51 foci formation analysis in a KB2P2 tumor pair, representative for all PARPi-resistant tumors for which alterations in the *Parg* gene were identified by DeepSeq analysis. Matched naive samples, as well as KP (p53$^{-/-}$) tumor (positive control) were taken along for the analysis. Microscopy images (upper panels) and foci quantification (lower panels) are shown. For quantification, a fraction of positive cells ($\geq$ 5 foci/nuclei for RAD51 and $\geq$ 10 foci/nuclei for 53BP1) was calculated for five different areas of each tumor (shown as single data point on the graph, error bars represent SD). $^{***}p < 0.0001$, $^{****}p < 0.0001$, n.s. - not significant (two-tailed t-test). 53BP1 was used as a marker of DNA damage. Scale bar, 25 μm.

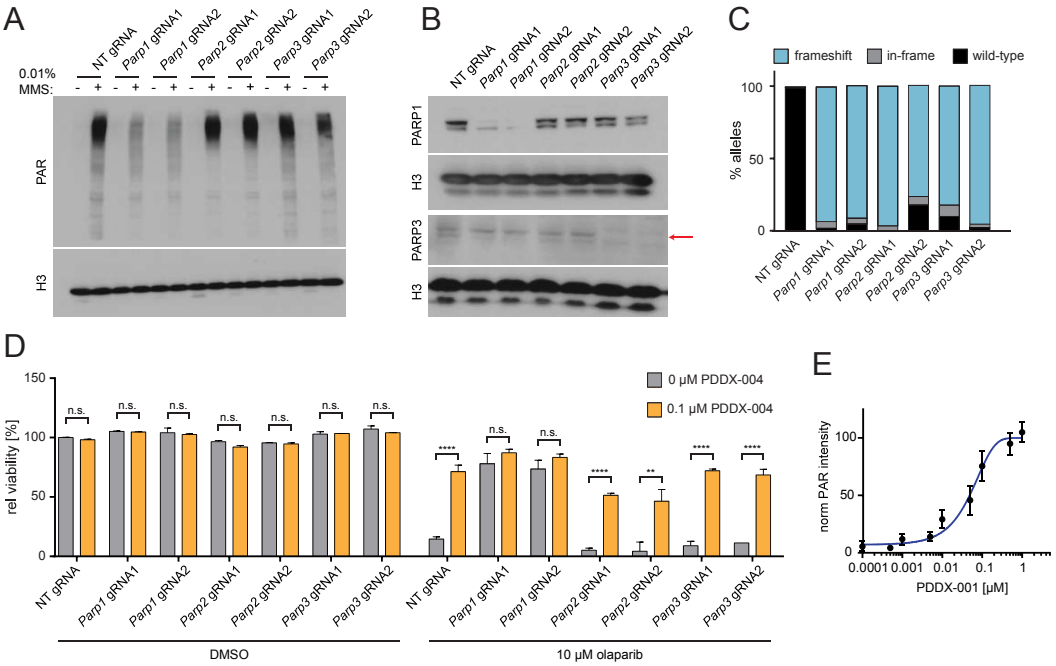


Fig. S7.5. **PARG depletion partially restores downstream signaling of PARP1. Related to Figures 7.4 and 7.5. (A-B)** Immunoblot analysis of PAR (A), PARP1 and PARP3 (B) cellular levels in KP3.33 expressing the gRNAs indicated. Red arrow (B) indicates band specific for PARP3. Histone 3 (H3) was used as a loading control. Data representative of experiments repeated twice. **(C)** TIDE analysis of allele modification rates in KP3.33 cells expressing the gRNAs indicated; data representative for experiment repeated twice. **(D)** Clonogenic assay with KP3.33 cells expressing the gRNAs indicated, treated with olaparib; $^{**}p < 0.01$, $^{****}p < 0.0001$, n.s. - not significant; data shown as mean ± SD of two repeats. **(E)** PARG inhibition in U2OS cells by PDDX-001; data presented as mean ± SD of two repeats.

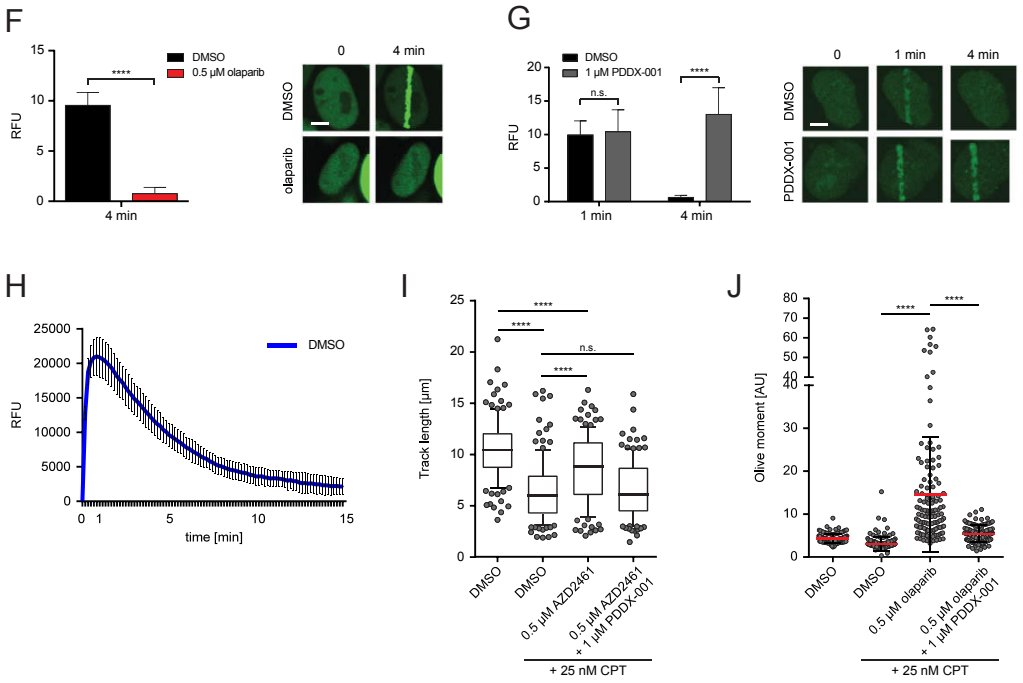


Fig. S7.5. **Continued. (F)** GFP-XRCC1 recruitment analysis in DMSO (control) or olaparib-treated GFP-XRCC1 U2OS cells; representative images (right) and quantification (left) are shown. XRCC1 recruitment was used as a readout for PARP1 inhibition. **(G)** GFP-CHD2 binding to the site of DNA damage in GFP-CHD2 U2OS cells treated with DMSO (control) or PDDX-001; representative images (right) and quantification (left panel) are shown. CHD2 association to chromatin was used as a positive control for PARG inhibition. **(H)** Time-course recruitment analysis of GFP-PARP1 to laser tracks in control cells (DMSO). For experiments shown in (F-H) data are represented as mean ± SEM of two independent repeats; $^{****}p < 0.0001$ (t-test); in all images scale bar, 10 μm. **(I)** Fork progression assay in U2OS cells exposed to the treatments indicated. Box extends from 25th to 75th percentile, with a middle line representing the median and whiskers drawn down to the 10th percentile and up to the 90th. **(J)** Neutral comet assay in U2OS cells treated as indicated. Data shown as mean ± SD of a replicate. Experiments shown in (I-J) were repeated twice; $^{****}p < 0.001$, n.s. - not significant, Mann-Whitney U test.

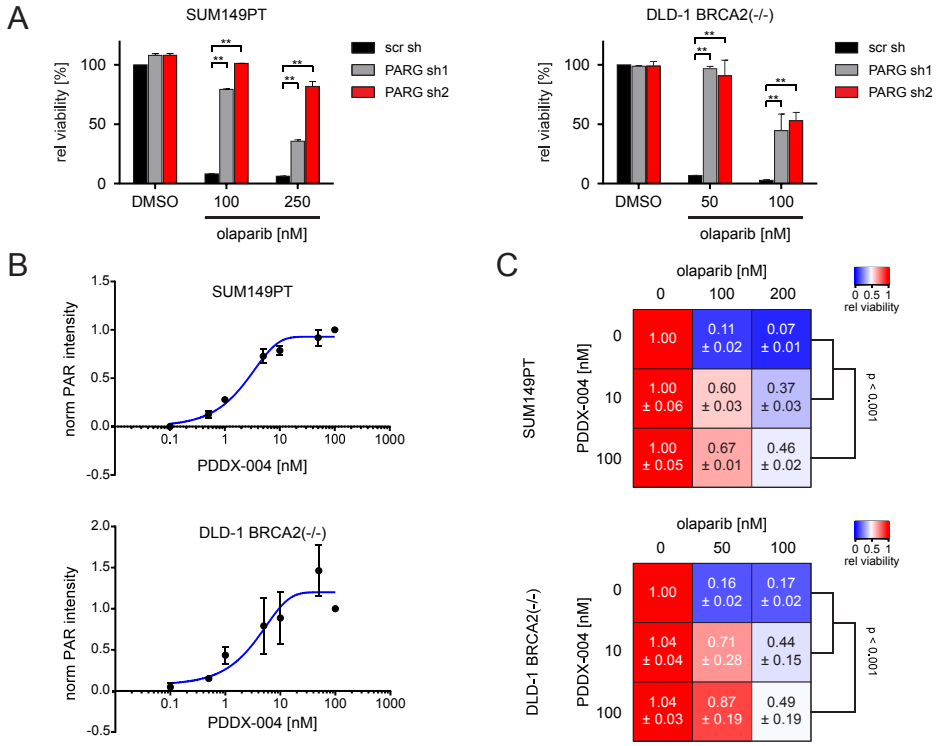


Fig. S7.6. PARPi response of PARG-depleted human cancer cells. Related to Figure 7.6. **(A)** Quantification of long-term clonogenic assays shown in Figure 7.5A. Graph represents mean ± SD values of triplicates; $^{**}p < 0.01$ (two-tailed t-test). **(B)** Chemical inhibition of PARG with PDDX-004 in indicated cells. Data represent mean ± SD of three repeats. **(C)** Quantification of a long-term clonogenic assay in SUM149PT (upper panel) and DLD-1 BRCA2(-/-) (lower panel) cells treated with the PARPi olaparib and the PARGi PDDX-004 alone or in combination. Data presented as mean ± SD values of three repeats.

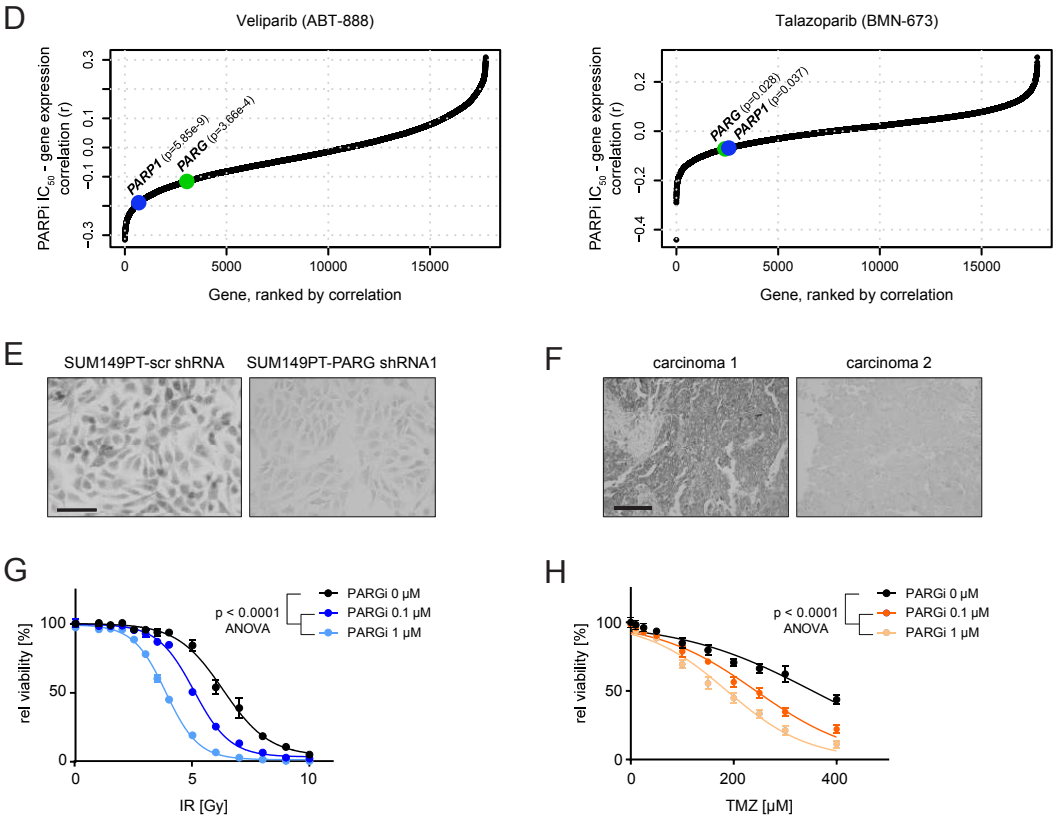


Fig. S7.6. **Continued. (D)** Correlation analysis between IC_{50} values of PARPi (veliparib and talazoparib) and expression of DDR genes in a panel of 935 human cancer cell lines; *PARP1* and *PARG* are highlighted; Pearson correlation was computed and p values were determined using the relation between estimated coefficient and the student-t distribution. **(E)** Validation of PARG antibody used in IHC analyses; representative images of SUM149PT cells expressing indicated shRNAs are shown; scale bar, 100 μm. **(F)** Example of PARG IHC staining in two ovarian serous carcinomas; scale bar, 100 μm. **(G-H)** Response of KB2P1.21 cells to ionizing radiation (IR) and temozolomide (TMZ), treated with PARGi PDDX-004. Data is presented as mean $\pm$ SD.

7.7.2 Supplementary Tables (available online)

Tab. S7.1. Genetic screens identify *Parg* as a PARPi-resistance factor. Related to Figures 7.1 and 7.2.

Tab. S7.2. DDR-related genes. Related to Figures 7.2 and 7.6.

Tab. S7.3. Overview of candidate genes identified in multi-omics analysis of matched KB2P PARPi-naive and -resistant tumors. Related to Figure 7.2.

Tab. S7.4. Custom-made shRNA library targeting candidate genes (identified by multi-omics analysis of KB2P tumors) and additional 32 non-essential genes (controls). Related to Figure 7.2.

Tab. S7.5. TIDE analysis of clonogenic assays using KB2P cells expressing gRNAs targeting *Parg*. Related to Figure 7.3.

Tab. S7.6. Oligonucleotides used in this study. Related to STAR methods.

7.8 References

[1] Felix Dietlein, Lisa Thelen, and H Christian Reinhardt. “Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches”. In: *Trends in Genetics* 30.8 (2014), pp. 326–339 (cit. on p. 230).

[2] Helen E Bryant, Niklas Schultz, Huw D Thomas, et al. “Specific killing of BRCA2-deficient tumours with inhibitors of poly (ADP-ribose) polymerase”. In: *Nature* 434.7035 (2005), pp. 913–917 (cit. on pp. 230, 231).

[3] Hannah Farmer, Nuala McCabe, Christopher J Lord, et al. “Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy”. In: *Nature* 434.7035 (2005), pp. 917–921 (cit. on pp. 230, 231).

[4] Bryan A Gibson and W Lee Kraus. “New insights into the molecular and cellular functions of poly (ADP-ribose) and PARPs”. In: *Nature Reviews Molecular Cell Biology* 13.7 (2012), pp. 411–424 (cit. on p. 230).

[5] John M Pascal and Tom Ellenberger. “The rise and fall of poly (ADP-ribose): an enzymatic perspective”. In: *DNA Repair* 32 (2015), pp. 10–16 (cit. on pp. 231, 239).

[6] Guotai Xu, Jos Jonkers, and Sven Rottenberg. “PARP Inhibitor Resistance—What Is Beyond BRCA1 or BRCA2 Restoration?” In: *PARP Inhibitors for Cancer Therapy*. Springer, 2015, pp. 453–471 (cit. on p. 231).

[7] Junko Murai, N Huang Shar-yin, Benu Brata Das, et al. “Trapping of PARP1 and PARP2 by clinical PARP inhibitors”. In: *Cancer Research* 72.21 (2012), pp. 5588–5599 (cit. on pp. 231, 239, 259).

[8] Junko Murai, Shar-Yin N Huang, Amèlie Renaud, et al. “Stereospecific PARP trapping by BMN 673 and comparison with olaparib and rucaparib”. In: *Molecular Cancer Therapeutics* 13.2 (2014), pp. 433–443 (cit. on pp. 231, 246).

[9] Bastiaan Evers, Rinske Drost, Eva Schut, et al. “Selective inhibition of BRCA2-deficient mammary tumor cell growth by AZD2281 and cisplatin”. In: *Clinical Cancer Research* 14.12 (2008), pp. 3916–3925 (cit. on pp. 231, 232, 251).

[10] Sven Rottenberg, Janneke E Jaspers, Ariena Kersbergen, et al. “High sensitivity of BRCA1-deficient mammary tumors to the PARP inhibitor AZD2281 alone and in combination with platinum drugs”. In: *Proceedings of the National Academy of Sciences* 105.44 (2008), pp. 17079–17084 (cit. on p. 231).

[11] M William Audeh, James Carmichael, Richard T Penson, et al. “Oral poly (ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial”. In: *The Lancet* 376.9737 (2010), pp. 245–251 (cit. on p. 231).

[12] Peter C Fong, David S Boss, Timothy A Yap, et al. “Inhibition of poly (ADP-ribose) polymerase in tumors from BRCA mutation carriers”. In: *New England Journal of Medicine* 361.2 (2009), pp. 123–134 (cit. on p. 231).

[13] Andrew Tutt, Mark Robson, Judy E Garber, et al. “Oral poly (ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial”. In: *The Lancet* 376.9737 (2010), pp. 235–244 (cit. on p. 231).

[14] Akihiro Ohmoto and Shinichi Yachida. “Current status of poly (ADP-ribose) polymerase inhibitors and future directions”. In: *OncoTargets and Therapy* 10 (2017), p. 5195 (cit. on p. 231).

[15] Stefano Annunziato, Marco Barazas, Sven Rottenberg, and Jos Jonkers. “Genetic dissection of cancer development, therapy response, and resistance in mouse models of breast cancer”. In: *Cold Spring Harbor Symposia on Quantitative Biology*. Vol. 81. Cold Spring Harbor Laboratory Press. 2016, pp. 141–150 (cit. on p. 231).

[16] Stacey L Edwards, Rachel Brough, Christopher J Lord, et al. “Resistance to therapy caused by intragenic deletion in BRCA2”. In: *Nature* 451.7182 (2008), pp. 1111–1115 (cit. on p. 231).

[17] Elizabeth M Swisher, Wataru Sakai, Beth Y Karlan, et al. “Secondary BRCA1 mutations in BRCA1-mutated ovarian carcinomas with platinum resistance”. In: *Cancer Research* 68.8 (2008), pp. 2581–2586 (cit. on p. 231).

[18] Joo Ern Ang, Charlie Gourley, C Bethan Powell, et al. “Efficacy of chemotherapy in BRCA1/2 mutation carrier ovarian cancer in the setting of PARP inhibitor resistance: a multi-institutional study”. In: *Clinical Cancer Research* 19.19 (2013), pp. 5485–5493 (cit. on p. 231).

[19] Stephen J Pettitt, Farah L Rehman, Ilirjana Bajrami, et al. “A genetic screen using the PiggyBac transposon in haploid cells identifies Parp1 as a mediator of olaparib toxicity”. In: *PloS One* 8.4 (2013), e61520 (cit. on pp. 231, 247).

[20] Jos Jonkers, Ralph Meuwissen, Hanneke van der Gulden, et al. “Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer”. In: *Nature Genetics* 29.4 (2001), pp. 418–425 (cit. on pp. 232, 250, 252).

[21] Alexandra A Duarte, Ewa Gogola, Norman Sachs, et al. “BRCA-deficient mouse mammary tumor organoids to study cancer-drug resistance”. In: *Nature Methods* 15.2 (2018), p. 134 (cit. on pp. 232, 233, 251, 256).

[22] Guotai Xu, J Ross Chapman, Inger Brandsma, et al. “REV7 counteracts DNA double-strand break resection and affects PARP inhibition”. In: *Nature* 521.7553 (2015), pp. 541–544 (cit. on pp. 232, 247, 255, 257, 260, 261).

[23] Wei Li, Han Xu, Tengfei Xiao, et al. “MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens”. In: *Genome Biology* 15.12 (2014), p. 554 (cit. on pp. 232, 257).

[24] Janneke E Jaspers, Ariena Kersbergen, Ute Boon, et al. “Loss of 53BP1 causes PARP inhibitor resistance in Brca1-mutated mouse mammary tumors”. In: *Cancer Discovery* 3.1 (2013), pp. 68–81 (cit. on pp. 233, 236, 250, 251).

[25] Lenka Oplustil O’Connor, Stuart L Rulten, Aaron N Cranston, et al. “The PARP inhibitor AZD2461 provides insights into the role of PARP3 inhibition for both synthetic lethality and tolerability with chemotherapy in preclinical models”. In: *Cancer Research* 76.20 (2016), pp. 6084–6094 (cit. on p. 233).

[26] Arnab Ray Chaudhuri, Elsa Callen, Xia Ding, et al. “Replication fork stability confers chemoresistance in BRCA-deficient cells”. In: *Nature* 535.7612 (2016), pp. 382–387 (cit. on pp. 233, 237).

[27] Jorma J de Ronde, Guillem Rigaill, Sven Rottenberg, Sjoerd Rodenhuis, and Lodewyk FA Wessels. “Identifying subgroup markers in heterogeneous populations”. In: *Nucleic Acids Research* 41.21 (2013), e200–e200 (cit. on pp. 234, 253).

[28] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. “Robust rank aggregation for gene list integration and meta-analysis”. In: *Bioinformatics* 28.4 (2012), pp. 573–580 (cit. on p. 234).

[29] Eva K Brinkman, Tao Chen, Mario Amendola, and Bas van Steensel. “Easy quantitative assessment of genome editing by sequence trace decomposition”. In: *Nucleic Acids Research* 42.22 (2014), e168–e168 (cit. on pp. 236, 257).

[30] Dominic I James, Kate M Smith, Allan M Jordan, et al. “First-in-class chemical probes against poly (ADP-ribose) glycohydrolase (PARG) inhibit DNA repair with differential pharmacology to Olaparib”. In: *ACS Chemical Biology* 11.11 (2016), pp. 3179–3190 (cit. on pp. 237, 240).

[31] Mi Young Kim, Tong Zhang, and W Lee Kraus. “Poly (ADP-ribosyl) ation by PARP-1:PAR-laying’NAD+ into a nuclear signal”. In: *Genes & Development* 19.17 (2005), pp. 1951–1967 (cit. on p. 239).

[32] Kazuto Sugimura, Shin-ichiro Takebayashi, Hiroshi Taguchi, Shunichi Takeda, and Katsuzumi Okumura. “PARP-1 ensures regulation of replication fork progression by homologous recombination on damaged DNA”. In: *Journal Cell of Biology* 183.7 (2008), pp. 1203–1212 (cit. on p. 241).

[33] Arnab Ray Chaudhuri, Yoshitami Hashimoto, Raquel Herrador, et al. “Topoisomerase I poisoning results in PARP-mediated replication fork reversal”. In: *Nature Structural & Molecular Biology* 19.4 (2012), pp. 417–423 (cit. on p. 241).

[34] Matteo Berti, Arnab Ray Chaudhuri, Saravanabhavan Thangavel, et al. “Human RECQ1 promotes restart of replication forks reversed by DNA topoisomerase I inhibition”. In: *Nature Structural & Molecular Biology* 20.3 (2013), pp. 347–354 (cit. on pp. 241, 247).

[35] Ralph Zellweger, Damian Dalcher, Karun Mutreja, et al. “Rad51-mediated replication fork reversal is a global response to genotoxic treatments in human cells”. In: *Journal of Cell Biology* 208.5 (2015), pp. 563–579 (cit. on pp. 241, 247, 261).

[36] Kai J Neelsen and Massimo Lopes. “Replication fork reversal in eukaryotes: from dead end to dynamic response”. In: *Nature Reviews Molecular Cell Biology* 16.4 (2015), pp. 207–220 (cit. on pp. 241, 247).

[37] Zhanna K Nazarkina, Svetlana N Khodyreva, Stéphanie Marsin, Olga I Lavrik, and J Pablo Radicella. “XRCC1 interactions with base excision repair DNA intermediates”. In: *DNA Repair* 6.2 (2007), pp. 254–264 (cit. on p. 241).

[38] In-Kwon Kim, Roderick A Stegeman, Chris A Brosey, and Tom Ellenberger. “A quantitative assay reveals ligand specificity of the DNA scaffold repair protein XRCC1 and efficient disassembly of complexes of XRCC1 and the poly (ADP-ribose) polymerase 1 by poly (ADP-ribose) glycohydrolase”. In: *Journal of Biological Chemistry* 290.6 (2015), pp. 3775–3783 (cit. on p. 241).

[39] Qingsheng Yan, Rong Xu, Liya Zhu, et al. “BAL1 and its partner E3 ligase, BBAP, link Poly (ADP-ribose) activation, ubiquitylation, and double-strand DNA repair independent of ATM, MDC1, and RNF8”. In: *Molecular and Cellular Biology* 33.4 (2013), pp. 845–857 (cit. on p. 242).

[40] Heinz C Schröder, Renato Batel, Heiko Schwertner, Oleksandra Boreiko, and Werner EG Müller. “Fast micromethod DNA single-strand-break assay”. In: *DNA Repair Protocols*. Springer, 2006, pp. 287–305 (cit. on pp. 242, 263).

[41] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, et al. “A landscape of pharmacogenomic interactions in cancer”. In: *Cell* 166.3 (2016), pp. 740–754 (cit. on pp. 242, 265).

[42] Teresa Cabezón, Irina Gromova, Pavel Gromov, et al. “Proteomic profiling of triple-negative breast carcinomas in combination with a three-tier orthogonal technology approach identifies Mage-A4 as potential therapeutic target in estrogen receptor negative breast cancer”. In: *Molecular & Cellular Proteomics* 12.2 (2013), pp. 381–394 (cit. on pp. 243, 250).

[43] Irina Gromova, Pavel Gromov, Naoko Honma, et al. “High level PHGDH expression in breast is predominantly associated with keratin 5-positive cell lineage independently of malignancy”. In: *Molecular Oncology* 9.8 (2015), pp. 1636–1654 (cit. on pp. 243, 250).

[44] Pavel Moudry, Kenji Watanabe, Kamila M Wolanin, et al. “TOPBP1 regulates RAD51 phosphorylation and chromatin loading and determines PARP inhibitor sensitivity”. In: *Journal of Cell Biology* 212.3 (2016), pp. 281–288 (cit. on pp. 244, 251).

[45] Jean-Christophe Amé, Elise Fouquerel, Laurent R Gauthier, et al. “Radiation-induced mitotic catastrophe in PARG-deficient cells”. In: *Journal of Cell Science* 122.12 (2009), pp. 1990–2002 (cit. on pp. 245, 248).

[46] Charlene Kan and Junran Zhang. “BRCA1 Mutation: A Predictive Marker for Radiation Therapy?” In: *International Journal of Radiation Oncology * Biology * Physics* 93.2 (2015), pp. 281–293 (cit. on p. 245).

[47] Fabricio G Sousa, Renata Matuo, Daniele G Soares, et al. “PARPs and the DNA damage response”. In: *Carcinogenesis* 33.8 (2012), pp. 1433–1440 (cit. on p. 246).

[48] Silvana De Lorenzo, Anand Patel, Rachel Hurley, and Scott HMD Kaufmann. “The elephant and the blind men: making sense of PARP inhibitors in homologous recombination deficient tumor cells”. In: *Frontiers in Oncology* 3 (2013), p. 228 (cit. on p. 247).

[49] Julie K Horton, Mary Watson, Donna F Stefanick, et al. “XRCC1 and DNA polymerase β in cellular protection against cytotoxic DNA single-strand breaks”. In: *Cell Research* 18.1 (2008), pp. 48–63 (cit. on p. 247).

[50] Nilufer Jasmine Selimah Fauzee, Juan Pan, and Ya-lan Wang. “PARP and PARG inhibitors—new therapeutic targets in cancer treatment”. In: *Pathology & Oncology Research* 16.4 (2010), pp. 469–478 (cit. on p. 247).

[51] Catherine Fathers, Ross M Drayton, Svetlana Solovieva, and Helen E Bryant. “Inhibition of poly (ADP-ribose) glycohydrolase (PARG) specifically kills BRCA2-deficient tumor cells”. In: *Cell Cycle* 11.5 (2012), pp. 990–997 (cit. on p. 247).

[52] Aurélia Noll, Giuditta Illuzzi, Jean-Christophe Amé, Françoise Dantzer, and Valérie Schreiber. “PARG deficiency is neither synthetic lethal with BRCA1 nor PTEN deficiency”. In: *Cancer Cell International* 16.1 (2016), p. 53 (cit. on p. 247).

[53] Polly Gravells, Emma Grant, Kate M Smith, Dominic I James, and Helen E Bryant. “Specific killing of DNA damage-response deficient cells with inhibitors of poly (ADP-ribose) glycohydrolase”. In: *DNA Repair* 52 (2017), pp. 81–91 (cit. on p. 247).

[54] David W Koh, Ann M Lawler, Marc F Poitras, et al. “Failure to degrade poly (ADP-ribose) causes increased sensitivity to cytotoxicity and early embryonic lethality”. In: *Proceedings of the National Academy of Sciences* 101.51 (2004), pp. 17699–17704 (cit. on p. 247).

[55] Christopher J Lord and Alan Ashworth. “PARP inhibitors: Synthetic lethality in the clinic”. In: *Science* 355.6330 (2017), pp. 1152–1158 (cit. on p. 248).

[56] Amy Drean, Christopher J Lord, and Alan Ashworth. “PARP inhibitor combination therapy”. In: *Critical Reviews in Oncology/Hematology* 108 (2016), pp. 73–85 (cit. on p. 248).

[57] Toshiro Sato, Robert G Vries, Hugo J Snippert, et al. “Single Lgr5 stem cells build crypt villus structures in vitro without a mesenchymal niche”. In: *Nature* 459.7244 (2009), pp. 262–265 (cit. on p. 251).

[58] Sven Rottenberg, Anders OH Nygren, Marina Pajic, et al. “Selective induction of chemotherapy resistance of mammary tumors in a conditional mouse model for hereditary breast cancer”. In: *Proceedings of the National Academy of Sciences* 104.29 (2007), pp. 12117–12122 (cit. on p. 252).

[59] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet. journal* 17.1 (2011), pp. 10–12 (cit. on pp. 253, 254).

[60] Alexander Dobin, Carrie A Davis, Felix Schlesinger, et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21 (cit. on p. 253).

[61] S. Andrews. *FastQC: a quality control tool for high throughput sequence data*. 2017. URL: `http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc` (cit. on pp. 253, 254).

[62] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (2016), pp. 3047–3048 (cit. on pp. 253, 254).

[63] Yang Liao, Gordon K Smyth, and Wei Shi. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7 (2014), pp. 923–930 (cit. on p. 253).

[64] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (2014), pp. 1–21 (cit. on p. 253).

[65] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. In: *Bioinformatics* 25.14 (2009), pp. 1754–1760 (cit. on p. 254).

[66] Heng Li, Bob Handsaker, Alec Wysoker, et al. “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079 (cit. on p. 254).

[67] Ilari Scheinin, Daoud Sie, Henrik Bengtsson, et al. “DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly”. In: *Genome Research* 24.12 (2014), pp. 2022–2032 (cit. on p. 254).

[68] Raivo Kolde and Sven Laur. *RobustRankAggreg: Methods for robust rank aggregation*. R package version 1.1. 2013 (cit. on p. 254).

[69] Jirina Bartkova, Zuzana Horejsi, Karen Koed, et al. “DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis”. In: *Nature* 434.7035 (2005), pp. 864–870 (cit. on p. 255).

[70] Traver Hart, Kevin R Brown, Fabrice Sircoulomb, Robert Rottapel, and Jason Moffat. “Measuring error rates in genomic perturbation screens: gold standards for human functional genomics”. In: *Molecular Systems Biology* 10.7 (2014), p. 733 (cit. on p. 255).

[71] Anirudh Prahallad, Guus JJE Heynen, Giovanni Germano, et al. “PTPN11 is a central node in intrinsic and acquired resistance to targeted cancer drugs”. In: *Cell Reports* 12.12 (2015), pp. 1978–1985 (cit. on p. 256).

[72] Neville E Sanjana, Ophir Shalem, and Feng Zhang. “Improved vectors and genome-wide libraries for CRISPR screening”. In: *Nature Methods* 11.8 (2014), pp. 783–784 (cit. on p. 256).

[73] Martijn S Luijsterburg, Inge de Krijger, Wouter W Wiegant, et al. “PARP1 links CHD2-mediated chromatin expansion and H3. 3 deposition to DNA repair by non-homologous end-joining”. In: *Molecular Cell* 61.4 (2016), pp. 547–562 (cit. on pp. 256, 264).

[74] Antonia Follenzi, Laurie E Ailles, Silvia Bakovic, Massimo Geuna, and Luigi Naldini. “Gene transfer by lentiviral vectors is limited by nuclear translocation and rescued by HIV-1 pol sequences”. In: *Nature Genetics* 25.2 (2000), pp. 217–222 (cit. on p. 256).

[75] Marco Barazas, Stefano Annunziato, Stephen J Pettitt, et al. “The CST complex mediates end protection at double-strand breaks and promotes PARP inhibitor sensitivity in BRCA1-deficient cells”. In: *Cell Reports* 23.7 (2018), pp. 2107–2118 (cit. on p. 258).

[76] Dominic I James, Stephen Durant, Kay Eckersley, et al. “An assay to measure poly (ADP ribose) glycohydrolase (PARG) activity in cells”. In: *F1000 Research* 5 (2016) (cit. on p. 259).

8

General discussion

8.1 Introduction

In this work, we have made extensive use of several mouse models for studying tumor development and therapy resistance in two different types of breast cancer: invasive lobular carcinoma (ILC) and BRCA1-deficient triple-negative breast cancer (TNBC). This allowed us to identify several known and novel drivers of ILC (Chapters 3 and 4) and BRCA1-deficient TNBC (Chapter 5), which were also shown to be frequently aberrated in human tumors. Besides this, we also demonstrated how mouse models can be used to identify resistance mechanisms in both ILC (Chapter 6) and BRCA-deficient TNBC (Chapter 7), by treating mice with transplanted tumor fragments with targeted treatments and comparing therapy-sensitive and -resistant tumors.

In the following sections, we focus on some of the concepts that underlie the work presented in this thesis. Specifically, we will first elaborate on our approaches used to identify cancer genes and therapy resistance mechanisms and reflect on several advantages and shortcomings of the presented approaches. Next, we touch on several recent advances in sequencing technologies and challenges in increasing the accessibility of sequencing data for future studies. Finally, we reflect on the current role(s) of mouse models in cancer research and how this is impacted by increased human sequencing efforts and the developments of organoid technologies.

8.2 Strengths and weaknesses of insertional mutagenesis strategies

One of the major goals/challenges in cancer research involves identifying which genes are driving tumor development, based on the idea that these genes can be targeted specifically in new therapies. In this work, we have shown how transposon-based insertional mutagenesis (IM) can be used to screen for candidate cancer genes, allowing us to identify driver genes in a mouse model of invasive lobular carcinoma (ILC) (Chapter 3). Moreover, we also have demonstrated how transposon insertions can be identified from RNA-sequencing data, and how this approach provides an important biological filter that focuses on transposon insertions that are actually expressed in the tumors, and therefore most likely to be relevant for tumor formation (4)

One of the main advantages of IM strategies, compared to other mutagenesis approaches, is that they provide a relatively unbiased genome-wide approach to forward genetic screening, whilst allowing efficient retrieval of transposon insertions via targeted sequencing. Moreover, depending on the used transposon, IM-based

screens can be used to simultaneously identify both candidate oncogenes and tumor suppressors in a single screen. As such, IM-based screening approaches have been successfully used to identify cancer driver genes in a wide variety of cancer types[1–7]. In these settings, IM has been particularly effective in identifying cancer drivers that are not readily apparent in human sequencing data, due to complex amplicons or more subtle events such as alternative splicing. A clear example is given by our transposon mutagenesis screen, which identified several ILC driver genes that were located in an amplicon on human chromosome 1q. Due to the size of this amplicon, these genes would have been difficult to identify without our screening results.

An important limitation of IM strategies is that different transposon systems have specific biases[8], which can lead to different results depending on the used transposon. Part of these biases are inherent to the used transposon systems (e.g., *piggyBac*[9,10], *Sleeping Beauty*[11,12]), which have been shown to have integration biases towards specific genetic sequences and/or certain gene features (e.g., integration within or upstream of genes)[13]. Other biases stem from the sequences used to construct the transposon, which may affect the mutagenicity of the transposon in certain types of cells. For example, the *T2/Onc2* transposon used in our screen contains an MSCV promoter, which is likely to drive oncogene expression in hematopoietic cells at a higher rate than in other cell types. For this reason, other transposons such as the *T2/Onc3*[14] transposon have been developed for targeting different cell types by including different promoter sequences.

These biases may explain to some extent why our screen identified a relatively concise set of candidate genes, compared to other recent IM screens[1,4,7]. For instance, SB transposons generally have a lower integration rate than other, more aggressive transposon systems such as *piggyBac*[15]. This might explain why we identify fewer integration sites than studies with other transposons[7], but it does not explain differences with SB-based studies[1,4]. Similarly, the used transposon (*T2Onc/2*) may not be particularly active in mammary gland epithelium, due to limited activity of the MSCV promoter in this tissue. Alternatively, biological constraints might be limiting the number of potential ILC driver genes, as is suggested by the limited number of driver mutations associated with ILC[16]. Finally, sequencing issues (e.g. limited sequencing depth) might have reduced the sensitivity of our insertion detection. However, orthogonal analyses using different sequencing approaches (IM-Fusion[17], Tagmap[18]) identified similar patterns of insertions, indicating that this is unlikely to be the case.

Another limitation of IM strategies is that targeted sequencing approaches aimed at identifying transposon insertions are unlikely to identify other types of mutations generated by competing, spontaneous mutational processes occurring in tumors. As such, naive IM screening approaches may miss additional candidate genes that are,

for example, aberrated by spontaneous point-mutations or copy number changes. The risk of additional mutations depends highly on the used mouse model, with more genomically unstable tumors being more likely to acquire mutations via alternative mechanisms. This issue can be mitigated by either applying IM in genomically stable models, or by combining targeted sequencing strategies with other sequencing modalities such as exome-sequencing[19]. Alternatively, RNA-sequencing-based approaches[17,20] such as IM-Fusion can be used for simultaneous identification of transposon insertions and mutations, as a single RNA-sequencing dataset can be used for insertion detection, variant calling, gene-fusion detection and differential expression analyses.

Altogether, IM strategies are effective approaches for identifying candidate cancer driver genes and potential mechanisms of therapy resistance. Moreover, when combined with somatic engineering approaches, IM provides an especially powerful method for identifying candidate cancer genes in specific genetic backgrounds and rapidly validating these candidates in the same *in vivo* context. Besides this, deeper sequencing and single-cell sequencing approaches[21] may provide more insight into the heterogeneity of IM-induced tumors and allow researchers to study clonal evolution within these tumors. This type of approach – optionally combined with multiplexed *in vivo* CRISPR approaches that permit studying interactions between multiple cancer driver genes[22,23] – may help determine how heterogeneous patient tumors are likely to evolve over time and in response to specific treatments, enabling the development of better treatment strategies.

8.3 Trade-offs in uncovering therapy resistance mechanisms

Acquired therapy resistance is currently a major challenge in the clinic that limits the effectiveness of many chemotherapy drugs and targeted therapies. In this work, we aimed to identify resistance mechanisms to two different targeted therapies, FGFR inhibition (FGFRi) in ILC (Chapter 6) and PARP inhibition (PARPi) in BRCA1-deficient TNBC (Chapter 7). To do so, we transplanted donor tumors into multiple syngeneic mice, which were subjected to treatment. By subsequently contrasting therapy-resistant and -sensitive (i.e. vehicle-treated) tumors, this approach allowed us to identify potential resistance mechanisms to both targeted treatments.

In Chapter 6, we focused on identifying FGFRi resistance mechanisms using insertional mutagenesis, based on the premise that SB-induced tumors can acquire new transposon insertions during treatment and in doing so become resistant to the applied therapy[24,25]. One advantage of using IM for this purpose, is that potential

resistance mechanisms are relatively easy to identify by using targeted sequencing to reveal the locations of new insertions in resistant tumors. However, a limitation of the approach is that the range of identifiable resistance mechanisms may be limited by integration biases of the used transposon and by the types of mutations that can be generated by transposon insertions (e.g. gene overexpression and gene truncation, but not point mutations). Moreover, competing mutagenic processes may result in additional mutations that may contribute to resistance but cannot be detected by targeted insertion sequencing, as evidenced by the FGFR2 mutations that we identified in RNA-sequencing data from several FGFRi-resistant ILC tumors.

In Chapter 7, we used genomic (copy number variation (CNV)-sequencing) and transcriptomic (RNA-sequencing) approaches to identify potential PARPi resistance mechanisms in BRCA2-deficient mammary tumors from KB2P mice. One of the main challenges of this type of approach is that genomically unstable tumors (such as KB2P tumors) are likely to contain many additional aberrations after treatment, of which only a very small fraction is likely to be causally involved in treatment resistance. As such, identifying potential resistance mechanisms requires additional prioritization of mutations, either using computational algorithms or by integrating data from additional sequencing approaches, other experiments or external databases. In the KB2P model, we addressed this issue by requiring genes to be frequently amplified/deleted and to show a strong effect on expression, providing additional evidence that these mutations have a functional effect. Additional evidence from a separate shRNA screen was used to further validate our results, by providing orthogonal evidence that candidate genes were indeed involved in resistance.

An alternative approach for identifying acquired resistance mechanisms involves correction of copy number profiles of resistant tumors for events that were present before treatment, and subsequent application of algorithms such as RUBIC[26] or GISTIC[27] to identify recurrent copy number events that are acquired after treatment. This would potentially allow us to identify DNA copy number alterations (CNAs) that arose in response to treatment. The main advantage of this approach is that it may restrict the number of potential candidate genes by focusing on the minimal regions of recurrent CNAs. However, this approach requires a snapshot of the pre-treatment copy number state for each tumor to avoid calling events that were already present in the pre-treatment tumor samples. In our case, these samples were not available. Moreover, it may be challenging to obtain representative pre-treatment samples for each donor, due to differences in pre-treatment tumor fragments stemming from intra-tumor heterogeneity. In addition, this approach assumes that the same resistance mechanism occurs in each tumor, and it may therefore not have enough statistical power in a population displaying heterogeneous resistance mechanisms.

In essence, the main difference between our two resistance screens lies in their focus. In the IM-based approach, we first used RNA-sequencing data to identify if any known mechanisms could explain the observed resistance, allowing us to perform a relatively focused analysis of this genome-wide dataset. We then used insertion-site sequencing to identify novel resistance mechanisms in an unbiased fashion. Due to the selective nature of the used transposon, this analysis identified a relatively short list of novel mechanisms, which might not have been identified through genome-wide analyses. In contrast, our analysis of the KB2P tumors was performed in a genome-wide fashion and relied on additional data sources (such as the curated list of DNA damage response (DDR) genes and the secondary shRNA screen) to prioritize candidate genes. In this respect, IM-based approaches have the advantage of being cleaner than a genome-wide analysis, at the risk of missing novel resistance mechanisms that are not captured by the transposon system. On the other hand, the sequencing approach has the advantage of potentially being more comprehensive but runs the risk of missing mechanisms in the analysis stage, due to the added complexity of the analysis, biases introduced by prior knowledge, and/or lack of statistical power resulting from small sample size and heterogeneous resistance mechanisms. As such, the choice of which approach to use likely depends on various aspects of the used mouse model (such as its degree of genomic instability) and pre-existing knowledge of potential resistance mechanisms.

Treatment resistance will most likely remain a substantial challenge for many anti-cancer therapies in the foreseeable future. One particularly challenging aspect of addressing therapy resistance is that it is essentially a race against evolution, in which tumors are likely to develop new resistance mechanisms in response to any applied therapies. As such, effective therapeutic strategies will need to be designed to minimize or prevent the emergence of resistance, for example by targeting convergent signaling pathways[28] or changing treatment strategies (e.g. by employing drug holidays[29] or combination therapies[30]). However, testing different (combination) treatments in different genetic backgrounds is likely to be prohibitively expensive and time consuming, due to the sheer amount of possible combinations involved. As such, we expect that computational models trained on existing experimental data will become crucial in modeling how tumor cells and cell populations respond to specific (combination) treatments, allowing us to predict which (combinations of) anti-cancer drugs are most likely to be most effective in treating a given tumor and preventing resistance.

8.4 Advances in tumor sequencing technologies

For the studies described in this thesis, we have made substantial use of second-generation sequencing technologies, which have greatly improved our ability to perform in-depth genomic and transcriptomic characterization of cell populations. However, one limitation of these technologies, is that they essentially reduce (cancer) genomes into a large collection of very short sequence reads, which must be pieced back together using a predefined reference genome to extract anything meaningful from the data. The fact that these short sequencing reads do not allow *de novo* assembly of cancer genomes inherently limits the power of our analyses. Moreover, this fragmented snapshot is not extracted from individual cells, but from bulk populations of thousands to millions of tumor cells, meaning that any derivation is going to be an average blur of a population of cells, rather than a detailed picture of all individual cells that captures all of their differences.

In this respect, one of the most exciting recent developments is the emergence of single-cell sequencing technologies, which can be used to characterize the genomic or transcriptomic state of individual cells[31,32]. This makes single-cell approaches especially suitable for studying intra-tumor heterogeneity, which is considered to be one of the main drivers of tumor progression, metastasis and escape from therapy[33]. As such, single-cell RNA-sequencing approaches have frequently been used to identify distinct populations of cells within tumors[34–36] and to explore how these populations may be involved in tumor progression and therapy resistance[37–39]. Similarly, single-cell CNV-sequencing and exome-sequencing approaches have been used to track cancer evolution by deriving how distinct tumor subclones evolved during tumor progression[40,41]. Limiting aspects of current single-cell technologies are low sequencing depths and high costs due to the large numbers of cells that need to be sequenced for a comprehensive analysis. However, with dropping sequencing costs and the development of more cost-effective approaches, we expect that single-cell sequencing will become an increasingly important pillar in studying tumor heterogeneity and cancer evolution.

Besides this, third-generation sequencing approaches (e.g., Nanopore, PacBio) are poised to further revolutionize sequencing analyses by providing much longer reads (10,000-100,000 bp) than current (second-generation) technologies, which are generally limited to reads of hundreds of base pairs. Increased read lengths provide substantial benefits for many sequencing analyses, as longer reads are much easier to stitch back together than short sequences, as demonstrated by *de novo* assemblies based on long-read sequencing[42]. Similarly, identification of complex structural variations in patient samples have been shown to benefit greatly from longer read lengths, allowing Nanopore-based approaches to achieve substantial accuracy in

identifying variants, even with relatively low sequencing depths[43]. Furthermore, in RNA-sequencing based approaches, long-read sequencing has been used to identify novel gene transcripts[44] and quantify isoform expression[45] more accurately than current sequencing approaches.

Together, we expect that advances in single-cell and third-generation sequencing will substantially increase our ability to detect more complex mutational events, such as structural variants and alternative splicing, whilst simultaneously providing detailed insight into the heterogeneous tumor environment. However, properly leveraging these technologies will also require the development of new computational approaches for analyzing the produced data, as is evidenced by the widespread development of software packages for the analysis of single-cell and long-read sequencing data.

8.5 Increasing the reproducibility and accessibility of sequencing data

To optimally exploit the full compendium of sequencing data that is currently being generated, it will be important to generate and share sequencing datasets in a way that they can be easily queried, integrated and visualized by fellow researchers. One of the main challenges in this respect is the heterogeneity of many analyses, which is fueled by lack of consensus on best practices for analyzing different types of sequencing data. This has led to the development of many different variants of sequencing analysis pipelines, which generally differ in their choice of software for different steps of the analysis. As a result, these pipelines can produce considerably different results, even when applied to the same datasets[46–48], thus complicating the integration of datasets from different research groups.

To establish a set of best practices, several efforts have been made to address this issue by benchmarking different approaches on gold-standard datasets[46–50]. Ideally, the results of these benchmarks will be used to develop templates for many standard analyses in community projects (e.g. bcbio-nextgen*, Snakemake workflows†) and to facilitate the sharing and democratization of sequencing pipelines. An important part of these projects is the development of open-source workflow management systems (e.g. Snakemake[51], Nextflow[52]), which allow users to define abstract workflows that can be shared and reused to analyze new datasets. Together with tools for managing software installations in virtual environments (e.g., Bioconda[53], Docker‡,

*http://bcbio-nextgen.readthedocs.io/en/latest
†https://bitbucket.org/johanneskoester/snakemake-workflows
‡https://www.docker.com

Singularity[54]), these developments make it easier to develop workflows that can be easily shared with, and applied by others, thus increasing the reproducibility and uniformity of sequencing analyses.

Additionally, with the large amounts of data that are being generated, it is becoming increasingly important to facilitate the deposition of these data into large centralized repositories, where they can be easily be integrated and analyzed by other researchers. Unfortunately, processed datasets are rarely shared in this fashion, as existing repositories (e.g. GEO[55], ENA[56]) are generally restricted to specific data types and therefore provide limited capabilities for integrating and visualizing datasets from different sequencing modalities. Portals such as cBioportal[57] are a great step forward in this respect, as they provide basic functionality for querying and visualizing different datasets without requiring specific computational expertise. In future work, it will be important to expand these existing initiatives to include many more public datasets, ideally upon publication of the corresponding manuscripts. Besides this, we expect that integration with similar portals containing data from non-human organisms (such as mouse models) will play an important role in disseminating insights from these model systems and improve cross-pollination with human sequencing efforts.

8.6 A future for mouse model systems?

Model systems have long been important cornerstones of cancer research, as they provide renewable and manipulatable biological systems that can be used to formulate and test hypotheses. Compared to other systems, mouse models have been particularly useful as they provide a controlled and experimentally tractable *in vivo* setting that remains relatively close to human biology. As a result, many different types of mouse models of human cancer have been developed over the years, of which we provided an overview in Chapter 1. These developments have given rise to a vast collection of mouse models with different genetic backgrounds, mimicking important aspects of many human cancer types[58,59].

An important advantage of model systems such as mouse models, compared to human sequencing projects, is that they can provide unambiguous, causative evidence of whether a given mutation is involved in tumorigenesis or therapy resistance. For example, by introducing a mutation into a mouse model that does by itself not develop tumors, we can unequivocally establish whether the added mutation results in increased tumor formation[60]. Human sequencing projects are much more limited in this respect, as without further validation these projects can generally only provide correlative evidence[61]. As such, model systems play an important role in

the follow-up of driver genes identified in these projects, by allowing us to validate their role(s) in tumor formation and determine how they may affect other tumor characteristics such as therapy response.

The utility of mouse models in cancer research will likely be enhanced by the development of lentiviral and CRISPR-based somatic engineering approaches[62,63], which allow the rapid introduction of mutations into existing baseline mouse models (Chapter 5). As such, these approaches can be used to speed-up the creation of new mouse models containing specific mutations, greatly expanding the capacity to further validate and characterize candidate driver genes. Moreover, by combining vectors containing/targeting different driver genes, somatic engineering can also be used to generate complex mouse models with different combinations of mutations, enabling more detailed dissection of interactions between different driver genes and how these influence tumor development and therapy response.

However, mouse models are not alone and face competition from human-based models. For instance, human tumor-derived organoid systems have recently gained much popularity in cancer research by enabling efficient derivation of three-dimensional (3D) cell culture models that maintain the genomic complexity of tumors more faithfully than conventional cell line models[64,65]. For this reason, organoids have been touted as an effective approach for generating *in vitro* models of patient tumors, which can be sequenced and screened with anti-cancer drugs to identify mutations and correlate these with therapy response[66]. Besides this, organoids can also be manipulated using lentiviral and CRISPR-based somatic technologies, allowing validation of driver genes or mutations that are predicted to influence therapy response. Combined with falling sequencing costs and increasing patient sequencing efforts, these efforts in organoid models are poised to provide a wealth of data correlating human tumor genotypes with drug response. Given enough data, this may substantially improve our ability to build computational models predicting a patient's therapy response, without requiring detailed modelling in (mouse) model systems.

Organoid-based approaches are however limited by the fact that they are grown in a rather artificial *in vitro* environment, and as such do not faithfully reflect the complexity of the tumor microenvironment and/or interactions with an active immune system. As a result, (humanized) mouse models are much better suited for studying the effects of immune-based therapies, which have gained much popularity due to their unparalleled success in treating specific cancer types[67–70]. Besides this, mouse models of *de novo* tumorigenesis provide opportunities for studying all stages of tumor initiation, progression and metastasis, which is challenging to do with models based on end-stage patient tumors. Finally, detailed *in vivo* characterization is likely to remain important to gain insight into the role of specific drivers in different stages of tumorigenesis, which can be used to design and develop novel

therapies targeting driver-specific vulnerabilities. As such, we expect that mouse model systems will continue to play an important role in cancer research. However, to maximize their impact, studies should focus on exploiting the unique properties of mouse models to complement patient- and organoid-based approaches, rather than mimicking studies that are better performed in more human-like (*in vitro*) models.

8.7 References

[1] Roberto Rangel, Song-Choon Lee, Kenneth Hon-Kim Ban, et al. “Transposon mutagenesis identifies genes that cooperate with mutant Pten in breast cancer progression”. In: *Proceedings of the National Academy of Sciences* 113.48 (2016), E7749–E7758 (cit. on p. 287).

[2] Roberto Rangel, Liliana Guzman-Rojas, Takahiro Kodama, et al. “Identification of new tumor suppressor genes in triple-negative breast cancer”. In: *Cancer Research* 77.15 (2017), pp. 4089–4101 (cit. on p. 287).

[3] Sjors M Kas, Julian R de Ruiter, Koen Schipper, et al. “Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular breast carcinoma”. In: *Nature Genetics* 49.8 (2017), pp. 1219–1230 (cit. on p. 287).

[4] Michael B Mann, Michael A Black, Devin J Jones, et al. “Transposon mutagenesis identifies genetic drivers of Braf V600E melanoma”. In: *Nature Genetics* 47.5 (2015), p. 486 (cit. on p. 287).

[5] Takahiro Kodama, Justin Y Newberg, Michiko Kodama, et al. “Transposon mutagenesis identifies genes and cellular processes driving epithelial-mesenchymal transition in hepatocellular carcinoma”. In: *Proceedings of the National Academy of Sciences* 113.24 (2016), E3384–E3393 (cit. on p. 287).

[6] Haruna Takeda, Alistair G Rust, Jerrold M Ward, et al. “Sleeping Beauty transposon mutagenesis identifies genes that cooperate with mutant Smad4 in gastric cancer development”. In: *Proceedings of the National Academy of Sciences* 113.14 (2016), E2057–E2065 (cit. on p. 287).

[7] Roland Rad, Lena Rad, Wei Wang, et al. “A conditional piggyBac transposition system for genetic screening in mice identifies oncogenic networks in pancreatic cancer”. In: *Nature Genetics* 47.1 (2015), p. 47 (cit. on p. 287).

[8] Johann de Jong, Waseem Akhtar, Jitendra Badhai, et al. “Chromatin landscapes of retroviral and transposon integration profiles”. In: *PLoS Genetics* 10.4 (2014), e1004250 (cit. on p. 287).

[9] Sheng Ding, Xiaohui Wu, Gang Li, et al. “Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice”. In: *Cell* 122.3 (2005), pp. 473–483 (cit. on p. 287).

[10] Roland Rad, Lena Rad, Wei Wang, et al. “PiggyBac transposon mutagenesis: a tool for cancer gene discovery in mice”. In: *Science* 330.6007 (2010), pp. 1104–1107 (cit. on p. 287).

[11] Adam J Dupuy, Keiko Akagi, David A Largaespada, Neal G Copeland, and Nancy A Jenkins. “Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system”. In: *Nature* 436.7048 (2005), p. 221 (cit. on p. 287).

[12] Lara S Collier, Corey M Carlson, Shruthi Ravimohan, Adam J Dupuy, and David A Largaespada. “Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse”. In: *Nature* 436.7048 (2005), pp. 272–276 (cit. on p. 287).

[13] Johann de Jong, Jeroen de Ridder, Louise van der Weyden, et al. “Computational identification of insertional mutagenesis targets for cancer gene discovery”. In: *Nucleic Acids Research* 39.15 (2011), e105–e105 (cit. on p. 287).

[14] Adam J Dupuy, Laura M Rogers, Jinsil Kim, et al. “A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice”. In: *Cancer Research* 69.20 (2009), pp. 8150–8156 (cit. on p. 287).

[15] Sareina Chiung-Yuan Wu, Yaa-Jyuhn James Meir, Craig J Coates, et al. “piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells”. In: *Proceedings of the National Academy of Sciences* 103.41 (2006), pp. 15008–15013 (cit. on p. 287).

[16] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, et al. “Comprehensive molecular portraits of invasive lobular breast cancer”. In: *Cell* 163.2 (2015), pp. 506–519 (cit. on p. 287).

[17] Julian R de Ruiter, Sjors M Kas, Eva Schut, et al. “Identifying transposon insertions and their effects from RNA-sequencing data”. In: *Nucleic Acids Research* 45.12 (2017), pp. 7064–7077 (cit. on pp. 287, 288).

[18] David L Stern. “Tagmentation-based mapping (TagMap) of mobile DNA genomic insertion sites”. In: *BioRxiv* (2017), p. 037762 (cit. on p. 287).

[19] Louise van der Weyden, George Giotopoulos, Kim Wong, et al. “Somatic drivers of B-ALL in a model of ETV6-RUNX1; Pax5+/- leukemia”. In: *BMC Cancer* 15.1 (2015), p. 585 (cit. on p. 288).

[20] Nuri A Temiz, Branden S Moriarity, Natalie K Wolf, et al. “RNA sequencing of Sleeping Beauty transposon-induced tumors detects transposon-RNA fusions in forward genetic cancer screens”. In: *Genome Research* 26.1 (2016), pp. 119–129 (cit. on p. 288).

[21] Karen M Mann, Justin Y Newberg, Michael A Black, et al. “Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq”. In: *Nature Biotechnology* 34.9 (2016), pp. 962–972 (cit. on p. 288).

[22] Roman Maresch, Sebastian Mueller, Christian Veltkamp, et al. “Multiplexed pancreatic genome engineering and cancer induction by transfection-based CRISPR/Cas9 delivery in mice”. In: *Nature Communications* 7 (2016), p. 10770 (cit. on p. 288).

[23] Zoë N Rogers, Christopher D McFarland, Ian P Winters, et al. “A quantitative and multiplexed approach to uncover the fitness landscape of tumor suppression in vivo”. In: *Nature Methods* 14.7 (2017), p. 737 (cit. on p. 288).

[24] Daniele Perna, Florian A Karreth, Alistair G Rust, et al. “BRAF inhibitor resistance mediated by the AKT pathway in an oncogenic BRAF mouse melanoma model”. In: *Proceedings of the National Academy of Sciences* 112.6 (2015), E536–E545 (cit. on p. 288).

[25] Emilie A Chapeau, Agnieszka Gembarska, Eric Y Durand, et al. “Resistance mechanisms to TP53-MDM2 inhibition identified by in vivo piggyBac transposon mutagenesis screen in an Arf-/- mouse model”. In: *Proceedings of the National Academy of Sciences* 114.12 (2017), pp. 3151–3156 (cit. on p. 288).

[26] Ewald van Dyk, Marlous Hoogstraat, Jelle ten Hoeve, Marcel JT Reinders, and Lodewyk FA Wessels. "RUBIC identifies driver genes by detecting recurrent DNA copy number breaks". In: *Nature Communications* 7 (2016), p. 12159 (cit. on p. 289).

[27] Craig H Mermel, Steven E Schumacher, Barbara Hill, et al. "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy number alteration in human cancers". In: *Genome Biology* 12.4 (2011), R41 (cit. on p. 289).

[28] Kirstine Jacobsen, Jordi Bertran-Alamillo, Miguel Angel Molina, et al. "Convergent Akt activation drives acquired EGFR inhibitor resistance in lung cancer". In: *Nature Communications* 8.1 (2017), p. 410 (cit. on p. 290).

[29] Chong Sun, Liqin Wang, Sidong Huang, et al. "Reversible and adaptive resistance to BRAF (V600E) inhibition in melanoma". In: *Nature* 508.7494 (2014), p. 118 (cit. on p. 290).

[30] Adam C Palmer and Peter K Sorger. "Combination cancer therapy can confer benefit via patient-to-patient variability without drug additivity or synergy". In: *Cell* 171.7 (2017), pp. 1678–1691 (cit. on p. 290).

[31] Charles Gawad, Winston Koh, and Stephen R Quake. "Single-cell genome sequencing: current state of the science". In: *Nature Reviews Genetics* 17.3 (2016), p. 175 (cit. on p. 291).

[32] Serena Liu and Cole Trapnell. "Single-cell transcriptome sequencing: recent advances and remaining challenges". In: *F1000 Research* 5 (2016) (cit. on p. 291).

[33] Nicholas McGranahan and Charles Swanton. "Clonal heterogeneity and tumor evolution: past, present, and the future". In: *Cell* 168.4 (2017), pp. 613–628 (cit. on p. 291).

[34] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, et al. "Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer". In: *Nature Communications* 8 (2017), p. 15081 (cit. on p. 291).

[35] Chunhong Zheng, Liangtao Zheng, Jae-Kwang Yoo, et al. "Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing". In: *Cell* 169.7 (2017), pp. 1342–1356 (cit. on p. 291).

[36] Boris J Winterhoff, Makayla Maile, Amit Kumar Mitra, et al. "Single cell sequencing reveals heterogeneity within ovarian cancer epithelium and cancer associated stromal cells". In: *Gynecologic Oncology* 144.3 (2017), pp. 598–606 (cit. on p. 291).

[37] Alice Giustacchini, Supat Thongjuea, Nikolaos Barkas, et al. "Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia". In: *Nature Medicine* 23.6 (2017), p. 692 (cit. on p. 291).

[38] Nicholas E Navin. "Tumor evolution in response to chemotherapy: phenotype versus genotype". In: *Cell Reports* 6.3 (2014), pp. 417–419 (cit. on p. 291).

[39] Charissa Kim, Ruli Gao, Emi Sei, et al. "Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing". In: *Cell* 173.4 (2018), pp. 879–893 (cit. on p. 291).

[40] Nicholas Navin, Jude Kendall, Jennifer Troge, et al. "Tumour evolution inferred by single-cell sequencing". In: *Nature* 472.7341 (2011), p. 90 (cit. on p. 291).

[41] Yong Wang, Jill Waters, Marco L Leung, et al. “Clonal evolution in breast cancer revealed by single nucleus genome sequencing”. In: *Nature* 512.7513 (2014), p. 155 (cit. on p. 291).

[42] Jeong-Sun Seo, Arang Rhie, Junsoo Kim, et al. “De novo assembly and phasing of a Korean human genome”. In: *Nature* 538.7624 (2016), p. 243 (cit. on p. 291).

[43] Mircea Cretu Stancu, Markus J Van Roosmalen, Ivo Renkens, et al. “Mapping and phasing of structural variation in patient genomes using nanopore sequencing”. In: *Nature Communications* 8.1 (2017), p. 1326 (cit. on p. 292).

[44] Lucy C de Jong, Simone Cree, Vanessa Lattimore, et al. “Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events”. In: *Breast Cancer Research* 19.1 (2017), p. 127 (cit. on p. 292).

[45] Mohan T Bolisetty, Gopinath Rajadinakaran, and Brenton R Graveley. “Determining exon connectivity in complex mRNAs by nanopore sequencing”. In: *Genome Biology* 16.1 (2015), p. 204 (cit. on p. 292).

[46] Sohyun Hwang, Eiru Kim, Insuk Lee, and Edward M Marcotte. “Systematic comparison of variant calling pipelines using gold standard personal exome variants”. In: *Scientific Reports* 5 (2015), p. 17875 (cit. on p. 292).

[47] Adam Cornish and Chittibabu Guda. “A comparison of variant calling pipelines using genome in a bottle as a reference”. In: *BioMed Research International* 2015 (2015), p. 456479 (cit. on p. 292).

[48] Tyler S Alioto, Ivo Buchhalter, Sophia Derdak, et al. “A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing”. In: *Nature Communications* 6 (2015), p. 10001 (cit. on p. 292).

[49] Xiangtao Liu, Shizhong Han, Zuoheng Wang, Joel Gelernter, and Bao-Zhu Yang. “Variant callers for next-generation sequencing data: a comparison study”. In: *PloS One* 8.9 (2013), e75619 (cit. on p. 292).

[50] Justin M Zook, David Catoe, Jennifer McDaniel, et al. “Extensive sequencing of seven human genomes to characterize benchmark reference materials”. In: *Scientific Data* 3 (2016), p. 160025 (cit. on p. 292).

[51] Johannes Köster and Sven Rahmann. “Snakemake—a scalable bioinformatics workflow engine”. In: *Bioinformatics* 28.19 (2012), pp. 2520–2522 (cit. on p. 292).

[52] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, et al. “Nextflow enables reproducible computational workflows”. In: *Nature Biotechnology* 35.4 (2017), p. 316 (cit. on p. 292).

[53] Björn Grüning, Ryan Dale, Andreas Sjödin, et al. “Bioconda: sustainable and comprehensive software distribution for the life sciences”. In: *Nature Methods* 15.7 (2018), pp. 475–476 (cit. on p. 292).

[54] Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. “Singularity: Scientific containers for mobility of compute”. In: *PloS One* 12.5 (2017), e0177459 (cit. on p. 293).

[55] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, et al. “NCBI GEO: archive for functional genomics data sets—update”. In: *Nucleic Acids Research* 41.D1 (2012), pp. D991–D995 (cit. on p. 293).

[56] Rasko Leinonen, Ruth Akhtar, Ewan Birney, et al. “The European nucleotide archive”. In: *Nucleic Acids Research* 39.suppl_1 (2010), pp. D28–D31 (cit. on p. 293).

[57] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, et al. “The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data”. In: *Cancer Discovery* 2.5 (2012), pp. 401–404 (cit. on p. 293).

[58] Jessica C Walrath, Jessica J Hawes, Terry van Dyke, and Karlyne M Reilly. “Genetically engineered mouse models in cancer research”. In: *Advances in Cancer Research*. Vol. 106. Elsevier, 2010, pp. 113–164 (cit. on p. 293).

[59] Kristopher K Frese and David A Tuveson. “Maximizing mouse cancer models”. In: *Nature Reviews Cancer* 7.9 (2007), p. 654 (cit. on p. 293).

[60] Douglas Hanahan, Erwin F Wagner, and Richard D Palmiter. “The origins of oncomice: a history of the first transgenic mice genetically engineered to develop cancer”. In: *Genes & Development* 21.18 (2007), pp. 2258–2270 (cit. on p. 293).

[61] William G Kaelin Jr. “Common pitfalls in preclinical cancer target validation”. In: *Nature Reviews Cancer* 17.7 (2017), p. 425 (cit. on p. 293).

[62] Wen Bu, Li Xin, Michael Toneff, Lei Li, and Yi Li. “Lentivirus vectors for stably introducing genes into mammary epithelial cells in vivo”. In: *Journal of Mammary Gland Biology and Neoplasia* 14.4 (2009), p. 401 (cit. on p. 294).

[63] Stefano Annunziato, Sjors M Kas, Micha Nethe, et al. “Modeling invasive lobular breast carcinoma by CRISPR/Cas9-mediated somatic genome editing of the mammary gland”. In: *Genes & Development* 30 (2016), pp. 1470–1480 (cit. on p. 294).

[64] Jarno Drost, Wouter R Karthaus, Dong Gao, et al. “Organoid culture systems for prostate epithelial and cancer tissue”. In: *Nature Protocols* 11.2 (2016), p. 347 (cit. on p. 294).

[65] Norman Sachs and Hans Clevers. “Organoid cultures for the analysis of cancer phenotypes”. In: *Current opinion in genetics & development* 24 (2014), pp. 68–73 (cit. on p. 294).

[66] Marc van de Wetering, Hayley E Francies, Joshua M Francis, et al. “Prospective derivation of a living organoid biobank of colorectal cancer patients”. In: *Cell* 161.4 (2015), pp. 933–945 (cit. on p. 294).

[67] F Stephen Hodi, Steven J O’day, David F McDermott, et al. “Improved survival with ipilimumab in patients with metastatic melanoma”. In: *New England Journal of Medicine* 363.8 (2010), pp. 711–723 (cit. on p. 294).

[68] Julie R Brahmer, Scott S Tykodi, Laura QM Chow, et al. “Safety and activity of anti–PD-L1 antibody in patients with advanced cancer”. In: *New England Journal of Medicine* 366.26 (2012), pp. 2455–2465 (cit. on p. 294).

[69] Roy S Herbst, Jean-Charles Soria, Marcin Kowanetz, et al. “Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients”. In: *Nature* 515.7528 (2014), p. 563 (cit. on p. 294).

[70] Thomas Powles, Joseph Paul Eder, Gregg D Fine, et al. “MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer”. In: *Nature* 515.7528 (2014), p. 558 (cit. on p. 294).

Appendices

A

A.1 English summary

Cancer is a disease in which normal cells are deregulated by disruption of their cellular processes, resulting in increased proliferation, survival and invasion of surrounding tissues. This disruption is generally attributed to (epi)genetic alterations in so-called driver genes, which provide cells with a selective growth advantage and drive their malignant transformation. To improve cancer treatments, personalized medicine approaches have aimed to identify exactly which mutations are driving tumor development in a given patient and specifically target these mutations using precision therapies. However, one of the main challenges of this approach is identifying which mutations are true drivers, as tumors typically contain many additional passenger mutations that do not actually contribute to tumor development. Besides this, many patients often relapse after prolonged treatment due to the emergence of acquired resistance, limiting the clinical effectiveness of targeted treatments.

To address these issues, it is crucial to (i) determine exactly which mutations are driving the development of certain cancer types and (ii) identify potential resistance mechanisms to targeted treatments. However, identifying driver genes in human sequencing studies has been challenging, due to the large amounts of potential drivers that are typically identified in these studies. Similarly, the identification of resistance mechanisms is generally limited by lacking availability of pre- and post-treatment tumor samples from patients, complicating the discovery of all potential resistance mechanisms for a given treatment. Model systems such as mouse models provide several complementary approaches for studying tumor development, by enabling the development of renewable models resembling specific subtypes of human cancer, which can then be studied in more detail to identify candidate cancer genes and therapy resistance mechanisms. An overview of several of these approaches is provided in **Chapter 1**.

In this thesis, we mainly focus on using genetically engineered mouse models to identify candidate cancer genes and therapy resistance mechanisms in two different types of breast cancer: invasive lobular carcinoma (ILC) and triple-negative breast cancer (TNBC) (**Chapter 2**). ILC is a histological subtype of breast cancer that accounts for 10-15% of all breast cancer cases and is characterized by mutations in E-cadherin (encoded by the *Cdh1* gene). Although ILCs generally show increased expression of ERα, long-term outcomes of ILC are typically worse than other ER-positive breast cancers, indicating that biological differences are influencing therapy response. TNBCs account for another 10-17% of breast cancers and are characterized by low expression of HER2 and the steroid hormone receptors ERα and PR. As such, TNBCs cannot be treated with specialized hormone therapies, leaving chemotherapy

as standard-of-care. Combined with the aggressive nature of these tumors, this results in a relatively poor prognosis for patients with TNBC.

A common theme in both ILC and TNBC, is that they respond less favorably to existing therapies than other breast cancer subtypes. As such, patients would benefit from further research identifying which genes are driving the development of these breast cancers, allowing the development of novel therapies targeting vulnerabilities stemming from the identified drivers.

In **Chapter 3**, we aimed to identify potential driver genes of ILC by using *Sleeping Beauty* (SB) transposon-based insertional mutagenesis (IM) to screen for candidate cancer drivers in a mouse model with mammary-gland specific inactivation of *Cdh1*. This showed that mice with combined loss of *Cdh1* and activation of IM develop tumors resembling human ILC, demonstrating the relevance of our model. Analysis of the insertions in these tumors identified several known and novel drivers of ILC, providing leads for the development of future therapies. Most notably, we identified several recurrent and mutually exclusive SB transposon insertions in *Myh9*, *Ppp1r12a*, *Ppp1r12b* and *Trp53bp2*, whose products have been implicated in the regulation of the actin cytoskeleton. *MYH9*, *PPP1R12B* and *TP53BP2* were also found to be frequently aberrated in human tumors, indicating that this is a novel oncogenic pathway in (human) ILC that might be exploited in future therapies.

One of the main challenges in the analysis of IM screens is that they can identify many potential cancer genes, of which only a fraction is likely involved in tumorigenesis. Additionally, the DNA-based approaches used to identify transposon insertions and associated candidate genes do not provide much insight into how genes are affected by the insertions. In **Chapter 4**, we addressed these issues by developing an approach called IM-Fusion, which identifies transposon insertions from gene-transposon fusions in RNA-sequencing data. Moreover, by combining insertion detection with differential expression analysis, this approach is capable of providing detailed insight into the effects of insertions on gene expression. To demonstrate the utility of our approach, we applied IM-Fusion to two existing RNA-sequencing datasets with matched DNA-based data. This demonstrated that IM-Fusion detects transposon insertions and their true target genes more accurately than DNA-based analyses.

In **Chapter 5**, we aimed to explore how the mutational landscape of BRCA1-deficient TNBC is affected by the presence of several established driver genes (*Myc*, *Met* and *Rb1*), in order to identify additional genes that collaborate with these established drivers and may therefore be targetable in novel therapies. Our analyses showed that BRCA1-deficient TNBCs with MYC overexpression exhibit a dramatically different copy number landscape than TNBCs from other models, indicating that high MYC

levels strongly affect the evolution of these tumors. By comparing recurrently aberrated regions between mouse and human TNBCs, we identified *Mcl1* as the main driver gene in a locus that is specifically amplified in MYC-driven TNBC, suggesting that MCL1 collaborates with MYC in TNBC development. Supporting this, additional experiments in a BRCA1-deficient patient-derived xenograft (PDX) model showed that MCL1 inhibition increases the effectiveness of the poly(ADP-ribose) polymerase (PARP) inhibitor olaparib (which targets the BRCA-deficiency of these tumors), underscoring the therapeutic potential of this combination treatment for BRCA-deficient TNBC.

In the remaining chapters, we focused on identifying potential resistance mechanisms to targeted treatments in ILC and TNBC.

First, in **Chapter 6**, we sought to explore the effects of FGFR inhibition in ILC, as *Fgfr2* was frequently activated by transposon insertions in our insertional mutagenesis screen (Chapter 3). To this end, we transplanted fragments of an ILC with activated FGFR signaling into multiple recipient mice and treated these mice with the FGFR inhibitor AZD4547. This showed that tumors initially regress upon treatment with the inhibitor, but quickly become resistant to treatment. To identify potential resistance mechanisms, we exploited the ongoing insertional mutagenesis in these tumors to identify transposon insertions that were acquired during treatment and may explain the observed resistance. Combined with a transcriptomic analysis of these tumors, this approach identified several known and novel resistance mechanisms. Notably, two novel resistance mechanisms were only identified by insertional mutagenesis, demonstrating that IM is an effective tool for identifying resistance mechanisms to targeted treatments in mice.

Next, in **Chapter 7**, we aimed to identify potential resistance mechanisms in BRCA2-deficient breast cancer to PARP inhibition (PARPi), which specifically targets the HR-deficiency of BRCA-deficient tumors. To do so, we performed *in vitro* PARPi resistance screens in BRCA2-deficient mammary tumor cells. Besides this, we also performed an *in vivo* analysis by transplanting pieces of BRCA2-deficient mammary tumors into multiple recipient mice, treating these mice with PARPi until resistance and subsequently contrasting the sensitive and resistant tumors. Strikingly, both these analyses identified loss of poly(ADP-ribose) glycohydrolase (PARG) as a major resistance mechanism. Further characterization showed that depletion of PARG induces resistance by restoring PAR formation, rescuing controlled DNA replication fork progression and promoting the recruitment of downstream DNA repair factors. The potential relevance of PARG in clinical PARPi resistance is underscored by the presence of PARG-negative clones in a subset of human TNBC and serous ovarian cancers.

Finally, in **Chapter 8**, we reflected on the methods and results presented in this thesis and how they may be applied or extended in future work. Besides this, we also consider several technological advances and important challenges that remain to be addressed in the field.

A.2 Nederlandse samenvatting

Kanker is een ziekte waarbij normale cellen worden gedereguleerd door verstoring van hun cellulaire processen, wat resulteert in verhoogde proliferatie, overleving en invasie van de omringende weefsels. Kanker wordt over het algemeen veroorzaakt door mutaties in zogenaamde 'driver-genen', die cellen een selectief groeivoordeel geven en hun kwaadaardige transformatie aansturen. Om kankerbehandelingen te verbeteren, proberen onderzoekers en artsen precies te identificeren welke mutaties in een bepaalde tumor aanwezig zijn om vervolgens gerichte antikankermedicijnen toe te passen die op deze mutaties aangrijpen. Eén van de grootste problemen van deze zogeheten 'personalized medicine' strategie is dat het lastig is om te bepalen welke mutaties de echte 'drivers' zijn omdat de meeste tumoren grote aantallen 'passenger' mutaties bevatten die niet bijdragen aan de ontwikkeling van de tumor. Daarnaast worden tumoren vaak resistent na langdurige behandelingen, waardoor de klinische effectiviteit van veel gerichte antikankermedicijnen beperkt is.

Om deze tekortkomingen te verhelpen, is het cruciaal om (i) precies te bepalen welke mutaties een rol spelen in de ontwikkeling van bepaalde kankertypes en (ii) potentiële resistentiemechanismen voor bepaalde therapieën te identificeren. Het identificeren van driver-genen in humane studies is echter nog steeds een uitdaging vanwege de grote hoeveelheden potentiële driver-genen die worden geïdentificeerd in deze studies. Daarnaast wordt de identificatie van alle mogelijke resistentiemechanismen bemoeilijkt door lage beschikbaarheid van biopten van tumoren voor en na behandeling. Modelsystemen zoals muizen bieden verschillende complementaire benaderingen voor het bestuderen van tumorontwikkeling omdat ze het mogelijk maken om hernieuwbare modellen te genereren die veel aspecten van verschillende menselijke kankersoorten bevatten. Deze modellen kunnen vervolgens worden gebruikt om potentiële driver-genen te testen en verschillende mechanismen van therapie resistentie te identificeren. Een overzicht van alle verschillende modelsystemen en hun toepassingen wordt gegeven in **Hoofdstuk 1**.

In dit proefschrift hebben we ons voornamelijk gericht op het gebruik van genetisch gemodificeerde muizen om driver-genen en therapieresistentie-mechanismen te identificeren in twee verschillende soorten borstkanker: invasief lobulair carcinoom (ILC) en triple-negatieve borstkanker (TNBC) (**Hoofdstuk 2**). ILC is een histologisch subtype van borstkanker dat verantwoordelijk is voor 10-15% van alle gevallen van borstkanker. ILC wordt gekenmerkt door mutaties in E-cadherine. Hoewel ILCs over het algemeen verhoogde expressie van ERα vertonen, zijn de overlevingskansen voor ILC-patiënten doorgaans slechter dan voor patiënten met andere ER-positieve borsttumoren, wat aangeeft dat biologische verschillen de therapiegevoeligheid beïnvloeden. TNBCs zijn verantwoordelijk voor 10-17% van alle gevallen van borstkanker

en worden gekenmerkt door lage expressie van HER2 en de hormoonreceptoren ERα en PR. TNBCs kunnen daarom niet worden behandeld met anti-hormonale therapieën of geneesmiddelen tegen HER2, waardoor chemotherapie nog steeds de standaardbehandeling vormt. Gecombineerd met de agressieve aard van deze tumoren resulteert dit in een relatief slechte prognose voor patiënten met TNBC.

Een overeenkomst van ILCs en TNBCs is dat ze minder gunstig reageren op bestaande therapieën dan andere soorten borstkanker. Het is daarom belangrijk om te onderzoeken welke driver-genen de ontwikkeling van deze borstkankersoorten stimuleren, waardoor het mogelijk wordt om gerichte geneesmiddelen te ontwikkelen die specifiek aangrijpen op kwetsbaarheden die deze driver-genen met zich meebrengen.

In **Hoofdstuk 3** hebben wij het *Sleeping Beauty* (SB) transposon systeem gebruikt om een insertiemutagenese (IM) screen uit te voeren in een muismodel met borstklier-specifiek verlies van E-cadherine, teneinde mogelijke driver-genen van ILC te identificeren. Gezamenlijk verlies van E-cadherine en activering van het SB transposon systeem leidde in muizen tot ontwikkeling van borsttumoren die leken op menselijke ILCs, wat de relevantie van ons model ondersteunde. Analyse van de transposon-inserties in deze tumoren leidde tot de identificatie van verschillende bekende en nieuwe driver-genen, die mogelijk gebruikt kunnen worden om nieuwe therapieën te ontwikkelen. Een opvallende ontdekking is dat vier van deze driver-genen (*Myh9*, *Ppp1r12a*, *Ppp1r12b* en *Trp53bp2*) vaak voorkwamen maar nooit samen gemuteerd waren, wat aangeeft dat ze waarschijnlijk een rol spelen in hetzelfde biologische proces. Verder onderzoek wees uit dat alle vier de genen betrokken zijn bij de regulatie van het actine cytoskelet in de tumorcellen. Drie van deze genen (*MYH9*, *PPP1R12B* en *TP53BP2*) zijn ook vaak gemuteerd in borsttumoren bij de mens, wat aangeeft dat dit nieuwe oncogene proces mogelijk een aangrijpingspunt vormt voor de ontwikkeling van toekomstige therapieën.

Een van de belangrijkste uitdagingen bij het analyseren van insertiemutagenese screens is dat de transposon-inserties veel potentiële kankergenen kunnen identificeren waarvan waarschijnlijk slechts een klein deel daadwerkelijk betrokken is bij tumorontwikkeling. Bovendien geven de DNA-gebaseerde benaderingen die vaak worden gebruikt om kandidaat-genen te identificeren, weinig inzicht in hoe de expressie van genen wordt beïnvloed door de transposon-inserties, waardoor biologisch inzicht in het effect van inserties wordt beperkt. In **Hoofdstuk 4** hebben wij daarom een nieuwe methode ontwikkeld, genaamd IM-Fusion, die deze problemen omzeilt door transposon-inserties te identificeren via detectie van fusies tussen de transposons en de betrokken genen in RNA-sequencing data. Deze methode verbetert de detectie van inserties door precies vast te stellen welke genen betrokken zijn bij een fusie en daarnaast te testen op het effect van de insertie op genexpressie. Om onze methode te toetsen, hebben we IM-Fusion toegepast op twee bestaande

RNA-sequencing datasets van tumoren waarvoor ook DNA-sequencing datasets beschikbaar waren. Onze vergelijking van IM-Fusion met de DNA-gebaseerde insertie analyses liet zien dat IM-Fusion accurater is bij het detecteren van transposon-inserties en de betrokken genen.

In **Hoofdstuk 5** hebben we vervolgens onderzocht hoe BRCA1-gemuteerde TNBC wordt beïnvloed door de aanwezigheid van verschillende bekende driver-genen (*Myc*, *Met* en *Rb1*). We hebben gekeken of genetisch gemodificeerde muizen met deze driver-genen gebruikt kunnen worden om genen te identificeren die collaboreren met een specifiek driver-gen in de ontwikkeling van BRCA1-gemuteerde TNBC. Ons onderzoek liet zien dat tumoren met MYC-overexpressie een compleet ander mutatieprofiel vertoonden dan andere tumoren, wat aangaf dat de evolutie van deze tumoren sterk werd beïnvloed door de hoge MYC-activiteit. Door gemuteerde regio's in de muizentumoren te vergelijken met de gemuteerde regio's in TNBCs van patiënten, ontdekten we MCL1 als een belangrijke driver die specifiek geamplificeerd is in MYC-gedreven TNBC, wat suggereert dat MCL1 samenwerkt met MYC in de ontwikkeling van BRCA1-gemuteerde TNBC. Deze hypothese wordt verder ondersteund door experimenten in een 'patient-derived tumor xenograft' muismodel voor BRCA1-gemuteerde TNBC, die aantoonden dat MCL1 remming de effectiviteit van de poly(ADP-ribose) polymerase (PARP) remmer olaparib (een remmer die specifiek aangrijpt op de BRCA1 mutatie in deze tumoren) verhoogt. Dit geeft aan dat de combinatie van PARP-remmers met MCL1-remmers mogelijk een kansrijke therapie is voor patiënten met BRCA1-gemuteerde TNBC die niet goed reageren op behandeling met alleen PARP-remmers.

In de resterende hoofdstukken hebben we ons gericht op het identificeren van mogelijke resistentiemechanismen voor gerichte behandelingen in ILC en TNBC.

Naar aanleiding van onze insertiemutagenese screen in Hoofdstuk 3, onderzochten we in **Hoofdstuk 6** de effectiviteit van FGFR-remming in ILC, aangezien *Fgfr2* vaak werd geactiveerd door transposon-inserties in de tumoren. Om het effect van FGFR-remming in deze tumoren te onderzoeken, transplanteerden we tumorfragmenten van een ILC met geactiveerde FGFR-signalering in meerdere muizen, die vervolgens werden behandeld met de FGFR-remmer AZD4547. Dit liet zien dat tumoren aanvankelijk goed reageerden op de behandeling, maar snel resistent werden tegen de remmer. Om mogelijke resistentiemechanismen te identificeren, hebben we de actieve insertiemutagenese in deze tumoren gebruikt om nieuwe transposon-inserties te identificeren die tijdens de behandeling optraden en mogelijk de waargenomen resistentie kunnen verklaren. Combinatie van deze insertie-analyse met transcriptomische analyse van dezelfde tumoren leidde tot de identificatie van verschillende bekende en onbekende resistentiemechanismen. Een belangrijk punt is dat twee resistentiemechanismen alleen werden geïdentificeerd via de insertie-analyse, wat

aantoont dat insertiemutagenese een effectief hulpmiddel is voor het identificeren van resistentiemechanismen tegen gerichte behandelingen in muizen.

In **Hoofdstuk 7** hebben we onderzocht hoe BRCA2-gemuteerde TNBCs resistentie ontwikkelen tegen PARP-remmers (PARPi), een nieuwe klasse antikankermedicijnen die specifiek aangrijpen op de HR-deficiëntie van BRCA-gemuteerde tumoren. Hiertoe voerden we een tweetal *in vitro* PARPi-resistentie screens uit in BRCA2-deficiënte borsttumorcellen. Daarnaast hebben we ook een *in vivo* analyse uitgevoerd door tumorfragmenten van BRCA2-deficiënte borsttumoren te transplanteren in meerdere muizen en deze muizen te behandelen met PARP-remmers, waarna we vervolgens de mutaties in PARPi-gevoelige en -resistente tumoren vergeleken. Opvallend genoeg, identificeerden beide analyses het verlies van poly(ADP-ribose) glycohydrolase (PARG) als het belangrijkste PARPi-resistentiemechanisme in BRCA2-deficiënte borsttumoren. Verdere experimenten toonden aan dat PARG-depletie PAR-vorming herstelt en daarnaast de progressie van de DNA-replicatievork herstelt en de rekrutering van DNA-reparatiefactoren bevordert. De klinische relevantie van PARG-verlies als een resistentiemechanisme wordt verder ondersteund door de aanwezigheid van PARG-negatieve cellen in een subset van humane TNBCs en eierstokkankers.

Ten slotte reflecteerden we in **Hoofdstuk 8** op de methoden en resultaten die in dit proefschrift worden gepresenteerd en hoe deze kunnen worden toegepast of uitgebreid in toekomstig werk. Daarnaast beschreven we de implicaties van verschillende technologische ontwikkelingen en belangrijke uitdagingen die nog resteren voor de toekomst.

A.3 Curriculum vitae

Julian de Ruiter was born on December 14th, 1988 in Farnborough, Great Britain. He spent his first school years at Green Street Green Primary School, before moving back to the Netherlands with his parents in 1999. In 2001, he started secondary school at the Gymnasium Felisenum in Velsen, from which he graduated in 2007. In the same year, he started the Computer Science Bachelor program at the Delft University of Technology (TUD), from which he graduated with distinction in 2009. During the third and final year of this program, he and several others developed an application for visualizing and statistically analyzing aCGH copy number profiles as part of their final bachelor project, which won the UFD Imtech grant, an award that is given to the best five bachelor projects developed by students at the TUD in a given year.

This bachelor project piqued Julian's interest in the biological sciences, which is why he enrolled in the Bioinformatics track of the TUD Computer Science master program in 2009. During the first year of this master, he also enrolled in an honors track program, allowing him to follow additional computer science courses outside of the standard curriculum. Next to his studies, Julian was chairman of the 21st board of the student sport club DSZ WAVE, which supports and promotes the sports of water polo, swimming and triathlon for students in Delft. In the second year of his master, Julian spent four months in the group of Ilya Shmulevich at the Institute for Systems Biology in Seattle, where he explored the use of multi-scale signal analysis for studying patterns in chromatin modification at different scales along the genome. After this internship, he performed his master thesis work on developing approaches for interpreting Random Forest classifiers and using these to gain insight into complex biological problems. With the completion of his thesis work, Julian graduated from the master program with distinction in 2012.

After obtaining his master's degree, Julian started as a PhD student in the group of Jos Jonkers at the Netherlands Cancer Institute in November 2012, under the joint supervision of Jos Jonkers and Lodewyk Wessels. During his PhD project, he used insertional mutagenesis strategies and *de novo* sequencing approaches to identify genes involved in breast cancer development and therapy resistance. The results of this work are described extensively in this thesis. In February 2018, Julian concluded his PhD work and accepted a position as a Machine Learning Engineer at GoDataDriven, where he aims to further expand his skills in developing data intensive applications and applying these to solve complex machine learning problems.

A.4 List of publications

Kuilman, T., Velds, A., Kemper, K., Ranzani, M., Bombardelli, L., Hoogstraat, M., ... de Ruiter, J.R., ... Krijgsman, O. (2015). **CopywriteR: DNA copy number detection from off-target sequence data.** *Genome Biology*, 16(1).

Ray Chaudhuri, A., Callen, E., Ding, X., Gogola, E., Duarte, A.A., Lee, J.E., ... de Ruiter, J.R., ... Nussenzweig, A. (2016). **Replication fork stability confers chemoresistance in BRCA-deficient cells.** *Nature*, 535(7612), 382–387.

Ter Brugge, P., Kristel, P., Van Der Burg, E., Boon, U., De Maaker, M., Lips, E., ... de Ruiter, J.R., ... Jonkers, J. (2016). **Mechanisms of therapy resistance in patient-derived xenograft models of BRCA1-deficient breast cancer.** *Journal of the National Cancer Institute, 108(11)*.

Boelens, M.C., Nethe, M., Klarenbeek, S., de Ruiter, J.R., Schut, E., Bonzanni, N., ... Jonkers, J. (2016). **PTEN Loss in E-Cadherin-Deficient Mouse Mammary Epithelial Cells Rescues Apoptosis and Results in Development of Classical Invasive Lobular Carcinoma.** *Cell Reports*, 16(8), 2087–2101.

de Ruiter, J.R., Knijnenburg, T., de Ridder, J. (2017). **Mining the forest: uncovering biological mechanisms by interpreting Random Forests.** *Bioarxiv.*

Kas, S.M.*, de Ruiter, J.R.*, Schipper, K.*, Annunziato, S., Schut, E., Klarenbeek, S., ... Jonkers, J. (2017). **Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular breast carcinoma.** *Nature Genetics, 49(8), 1219–1230.*
**Contributed equally to this work.*

de Ruiter, J.R., Kas, S.M., Schut, E., Adams, D.J., Koudijs, M.J., Wessels, L.F.A., Jonkers, J. (2017). **Identifying transposon insertions and their effects from RNA-sequencing data.** *Nucleic Acids Research*, 3(12), 636–647.

Aben, N., de Ruiter, J.R., Bosdriesz, E., Kim, Y., Bounova, G., Vis, D., ... Michaut, M. (2018). **Identifying biomarkers of anti-cancer drug synergy using multi-task learning.** *BioRxiv*, 243568.

Duarte, A.A., Gogola, E., Sachs, N., Barazas, M., Annunziato, S., de Ruiter, J.R., ... Rottenberg, S. (2018). **BRCA-deficient mouse mammary tumor organoids to study cancer-drug resistance.** *Nature Methods*, 15(2), 134–140.

Kas, S. M.*, de Ruiter, J. R.*, Schipper, K., Schut, E., Bombardelli, L., Wientjens, E., ... & Smith, P. D. (2018). **Transcriptomics and transposon mutagenesis identify multiple mechanisms of resistance to the FGFR inhibitor AZD4547**. *Cancer research*, 78(19), 5668-5679.
*Contributed equally to this work.

Gogola, E., Duarte, A. A.*, de Ruiter, J. R.*, Wiegant, W. W., Schmid, J. A., de Bruijn, R., ... & van den Broek, B. (2018). **Selective Loss of PARG Restores PARylation and Counteracts PARP Inhibitor-Mediated Synthetic Lethality.** *Cancer cell*, 33(6), 1078-1093.
*Contributed equally to this work

de Ruiter, J. R., Wessels, L. F. A., & Jonkers, J. (2018). **Mouse models in the era of large human tumour sequencing studies.** *Open biology*, 8(8), 180080.

Kluin, R. J., Kemper, K., Kuilman, T., de Ruiter, J. R., Iyer, V., Forment, J. V., ... & Klarenbeek, S. (2018). **XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data.** *BMC bioinformatics*, 19(1), 366.

Annunziato, S.*, de Ruiter, J.R.*, Henneman, L.*, Brambillasca, C.S.*, Lutz, C., Vaillant, F., Ferrante, F., Drenth, A.P., van der Burg, E., Siteur, B., van Gerwen, B., de Bruijn, R., van Miltenburg, M.H., Huijbers, I.J., van de Ven, M., Visvader, J.E., Lindeman, G.J., Wessels, L.F.A., Jonkers, J. (2019) **Comparative oncogenomics and iterative mouse modeling identifies combinations of driver genes and drug targets in BRCA1-mutated breast cancer.** *Nature Communications*, 10, 397.
*Contributed equally to this work.

Kas, S.M., de Ruiter, J.R., Schut, E., Bleijerveld, O.B., Klarenbeek, S., Wientjens, E., van Miltenburg, M.H., Drenth, A.P., van der Burg, E., Boon, U., Wessels, L.F.A., Jonkers, J. ***In vivo* oncogenicity of FGFR2 is modulated by multiple C-terminal domains.** *Manuscript in preparation.*

A.5 Acknowledgements

Over the past years, I have been fortunate to work in a stimulating environment with many wonderful, enthusiastic and dedicated people — without whom this thesis would not have been possible.

Jos, I have yet to meet someone with more passion for science than you. Thank you for your boundless enthusiasm, which made it easy to believe in our work. I enjoyed the many inspiring discussions we've had together — both inside and outside the lab — and I hope that we'll continue to cross paths!

Lodewyk, thank you for the amazing opportunity to work in not one, but two groups with experts in both machine learning and biology. With your keen but critical eye, you and Jos were a perfect duo for pushing our work to a higher level. I hope you keep doing great things together!

Sjors, our complementary skills made us an excellent team and we can definitely be proud of the papers that we wrote together. I think you're a great scientist and am curious to see what you'll achieve in the future. We'll definitely come to visit you in Manchester!

Koen, thank you for all your contributions to the SB paper, which were crucial in pushing the story to the next level. Outside work, I've had a lot of fun skiing/snowboarding together in the Alps and I hope that we can keep doing these trips in the future!

Stefano and Chiara, you were great office mates and I very much enjoyed working together on the oncogenomics project. Thank you for also stimulating our social development by organizing many Italian dinners, go-karting sessions and the memorable ski-trip to Cervinia!

Sander, Nanne and Daniel, you frequently provided me with a listening ear for debating topics of a more computational nature whenever I wanted to escape the biology. Thank you for these discussions and the many beers we enjoyed together after and outside work.

Even though they are with too many to list here, the other members of **the Jonkers/Wessels labs** and **the CombatCancer team** also provided me with crucial support during my PhD. As such, I would like to thank **everyone** for the fantastic time I have had at the NKI and for their valuable contributions.

Jeroen and Theo, even though we never quite got round to publishing my master thesis work, I always enjoyed our meetings together. Your critical advice has been of great value to me over the years and I hope we continue to keep in touch.

I would also like to thank the members of my **Doctorate Committee** for their feedback on my thesis, as well as **Piet**, **Sven**, **Jelle** and my **OIO Committee** (**Roderick**, **Bas** and **Wilbert**) for their critical feedback and advice during our meetings and my presentations.

Daarnaast wil ik graag mijn **familie en vrienden** bedanken voor hun eindeloze interesse in mijn onderzoek en de nodige afleidingen eromheen. Ik ben ook bijzonder dankbaar voor mijn **(schoon)ouders**, die altijd voor mij klaar stonden en mij de ruimte gaven om mijn dromen na te jagen. En als laatste natuurlijk **Anne Paulien**, mijn kersverse vrouw! Bedankt voor jouw onvoorwaardelijke steun en oneindige geduld door de jaren heen. Ik ben benieuwd wat de toekomst ons zal brengen en kijk ernaar uit om dat samen te ontdekken.