**Supplementary Methods**

**Statistical analysis and software**

All custom code, statistical analysis, and visualizations were performed in Python or R. We used *Nextflow* to manage some of the computational pipelines (1). All code used for this publication is available online at: <https://github.com/murphycj/HuiEtAl2017>.

**RNA-seq analysis**

The quality of the raw FASTQ files were checked with FastQC, then mapped to mouse reference GRCm38 using STAR (v2.4.0f1, 2-pass mode). Duplicate reads were marked (but not removed), reads were realigned around indels, and then base quality scores were recalibrated using the Genome Analysis Toolkit (v3.6, default parameters) (2). Gene expression (FPKM) was estimated using Cufflinks (v2.2.1, default parameters).

**Breast cancer subtype classification**

To classify the mouse tumors into human breast cancer subtypes using the AIMs classifier we first converted mouse gene symbols to human Entrez gene IDs using BioMart and the vertebrate homology list provided by Mouse Genome Informatics (3-5). Figure S1C was produced by clustering (spearman correlation distance, average linkage) the mouse tumors with breast cancer RNA-seq expression profiles from TCGA on the 137 shared genes between mouse and human that were part of the AIMs signature. RNA-seq expression profiles for TCGA breast cancer (BRCA) samples were downloaded from Broad GDAC Firehose ([http://gdac.broadinstitute.org/)](http://gdac.broadinstitute.org/%29) data version 2016\_01\_28. Breast sample subtype information was determined using the PAM50 classification and ER, PR, and Her2 status (for TNBC patients) from the original TCGA BRCA publication (6).

To classify the mouse tumors into human breast cancer subtypes using the PAM50 classifier we first converted mouse gene symbols to human gene symbols using the vertebrate homology list provided by Mouse Genome Informatics. The original PAM50 algorithm supplied centroids used for classification, but we re-computed the centroids using RNA-seq expression profiles from TCGA BRCA data and PAM50 assignments in the original TCGA BRCA publication as ground truth for cross validation (6,7). After computing the within-sample rank normalization on the RPKM measurements, we achieved 82% classification accuracy after 10-fold cross-validation on the human BRCA TCGA data (log2(RPKM) and median-centered normalization methods performed equally as well). We then classified the mouse tumors after rank normalizing the mouse FPKM values.

ER, PR, and HER2 status for each mouse tumor was determined by building a logistic regression model based on ER, PR, and HER2 status and RNA-seq expression in TCGA BRCA samples {Rantalainen, 2016 #260}. Prior to building the logistic regression model, the mouse and TCGA BRCA RNA-seq data were quantile normalized together using the *normalizeQuantiles* function from the *limma* R package.

**Gene fusion detection**

Gene fusions were called using FusionCatcher (v0.99.5a) with mouse reference genome GRCm38 (8). Fusions found in any of the three normal mammary controls, were marked as read-throughs, or occurred with the same breakpoint and gene partners in three or more independent, primary tumors were marked as false positives and removed.

**Whole-exome sequence analysis**

The quality of the raw FASTQ files were checked with FastQC, then mapped to mouse reference GRCm38 using BWA (v0.7.12, default parameters) (9). Duplicate reads were removed, reads were realigned around indels, and then base quality scores were recalibrated using the Genome Analysis Toolkit (v3.6, default parameters) (2).

The SureSelect and NimbleGen mouse exome capture kits used for sequencing were first lifted over using the UCSC liftOver tool from mm9 to mm10 coordinates. For the purpose of filtering mutations, we took the union of regions from the two exome capture kits using bedtools (v2.26.0) (10).

**Mutation detection**

We called mutations using Varscan2 on each tumor-normal pair (11). If no normal tail sample was available for a tumor, we combined the three normal liver samples as a normal control for somatic mutation calling. For calling mutations from RNA-seq we applied the additional pre-processing step of removing PCR duplicate reads. Varscan2 default parameters were used except: minimum read mapping quality of 15, minimum base quality of 15, at least four supporting reads, minimum sequencing depth of 12, minimum variant frequency of 5%, and p-value ≤ 0.01. Since not all sequenced tumors had matched normal controls, we applied a stringent filtering criterion to reduce potential false positive germline mutations. We removed mutations that had 4 or more variant supporting reads in any of the control samples (with minimum base and mapping qualities of 10) or for which there were fewer than 10 normal control samples with coverage less than 12. We then removed any mutations that were outside the exon capture regions, were found in any of the immunoglobulin, histocompatibility, or killer cell lectin like receptor genes (see the online code repository for the list of genes), were found in mouse dbSNP (Mouse Genomes Project Release Version 5) or were in known RNA-editing sites (12). The database of known RNA-editing sites was downloaded (ftp://ftp-mouse.sanger.ac.uk/REL-1202-RNAEditing/RNA-editing.vcf.gz) then converted from mm9 to mm10 coordinates using CrossMap (13). We noticed hundreds of mutations that were specific to and shared among many of the RNA-seq samples and reasoned they were likely false positives. We filtered those mutations using the Fisher’s exact test where the contingency table has the number of samples that have three or more variant supporting reads or less than three for the rows, and whether the sample is from RNA-seq or WEX for the columns. We removed mutations with p-value ≤ 0.01. Finally, we annotated the mutations using SnpEff (14).

**Mutation signature analysis**

We computed the mutational signatures using the *SomaticSignatures* (v2.10.0) R package (15). To compare the mouse mutation signatures to the 22 validated signatures identified in human cancer we first downloaded the matrix of signature values from Alexandrov *et al.* (16). For each mouse sample we computed the relative frequency of each mutation type, then proceeded with two normalization steps before comparing to human. We first normalized each mouse sample by sequencing depth using samtools mpileup (-q 15, -Q 15) by counting the number of times the middle base in each 3-mer was sequenced to a depth of at least 12 (this is the minimum sequencing depth required for calling a mutation) (17). We then normalized the frequency of each mutation type by the relative frequency of each 3-mer in the mouse genome (but restricted to the exon capture regions) and the human genome (hg19). We then combined samples by Brca1 genotype by averaging the mutation type frequencies and renormalizing so the frequencies sum to one. Finally, we computed the cosine similarity between each Brca1 genotype mutational signature and each human cancer mutation signatures.

**Identification of tumor drivers and drug vulnerabilities**

We compared the somatic alterations found in our mouse tumors to the OncoKB database to identify oncogenic and actionable alterations, and the Cancer Hotspots database to identify mouse mutations which overlap with identified hotspots in human cancer (18,19). We also identified which somatic alterations overlap genes in the PI3K and MAPK pathways (as defined by KEGG) to discern which tumors have alterations that potentially activate those pathways.

To identify overlapping point mutations, we first globally aligned the peptide sequences of canonical isoforms of mouse-human homologs using Needleman-Wunsch aligner with the BLOSUM62 matrix (20). We then converted the amino acid coordinate of the mouse mutation to human coordinates. To identify overlapping CNAs, we defined deleted mouse genes as those with copy number 0 and amplified genes as those with copy number greater than 3. Finally, we used a z-score threshold of +/- 1.5 to identify significant over or under-expressed genes.

**Copy number analysis**

Before calling CNAs with CNVkit (v0.8.1) we filtered out reads with mapping quality less than 30 and read pairs that map to different chromosomes using samtools (21). We used separate strategies to call CNAs in tumors sequenced with the SureSelect and NimbleGen kits.

CNAs of tumors sequenced with the SureSelect kit were called using a panel of three livers as controls. However, the built-in GC-content bias correction within CNVkit was not sufficient. We observed increased variance in log2 coverage ratios (tumor over liver) in capture regions with lower GC content (data not shown); therefore, we developed an algorithm to correct the variance. Let $L\_{ik}$ be the log2 coverage ratio in capture region k and sample i. Let $GC\_{j}$ denote the set of all capture regions $k$ where GC content is j (in intervals of 1%). Then we compute $m\_{ij}$ as the median absolute deviation (MAD) of all $L\_{ik}$, where $k\in GC\_{j}$. We then computed the value M, which is the median of all the $m\_{ij}$ values. We then calculated, $Lc\_{ik}$, the corrected log2 coverage ratio with the following equation:

$$Lc\_{ik}=\frac{M}{m\_{ij}}L\_{ik}, where k\in GC\_{j}$$

We demonstrated that this correction significantly improves segmentation results as determined by comparison to an independent algorithm, CopywriteR (v2.4), which can call CNAs without the need for a control sample (22). CopywriteR was not used as the primary CNA analysis because its results are noisier than CNVkit results, but provides an important benchmark by which we demonstrated the validity of our CNA calling strategy. Using the same set of filtered reads as used for CNVkit we ran CopywriteR with window size set to 40kbp and the SD parameter set to 1 for the segmentation step. We compared CNVkit and CopywriteR results by computing the weighted Euclidean distance between final segmentation results, where the weight is the length in base pairs of each segment. But before we computed the Euclidean distance we standardized the segmentation values by subtracting the mean and dividing by the standard deviation computed from the capture region log2 ratios. The variance correction significant decreased the weighted Euclidean distance between CNVkit and CopywriteR results (p-value < 0.01, paired student t-test).

Tumors sequenced using the NimbleGen kit had no suitable reference due to poor quality tail DNA and subsequent low quality sequencing results; and the livers could not be used as a control either due to being sequenced with a different capture kit. Instead, all tumors sequenced with the NimbleGen kit were used as a combined reference using CNVkit. After CNVkit corrected the combined reference for GC and mappability biases, we applied an additional correction that mitigated any bias introduced by any recurrent regions of amplification or deletion in the tumors using local regression (LOWESS). We expect there to be even coverage across the genome, which we confirmed by looking at the median-centered log2 coverage in the liver samples plotted along the genome. First, we sorted the probes by genomic position, and for each chromosome separately, applied our LOWESS correction using the *loess* R function with span=0.1 and where data was weighted by their normal density from the *dnorm* R function where *mean* was set to median and *sd* set to the standard deviation across all log2 coverage values. Weighting the data was necessary to reduce the impact of outlier and low-coverage regions.

We used a previously published approach to identify thresholds for copy number changes in each sample (23). Briefly, we sorted the log2 ratios of each capture region and calculated the median and standard deviation from the 50% central exon capture regions. We defined copy number segments with log2 ratio -2.5 or -7 standard deviations below the median as copy number losses and deep deletions, respectively. Copy number segments with log2 ratio +2 or +6 standard deviations above the median were copy number gains and amplifications, respectively. Finally, for the purpose of visualization and comparison across samples, we rescaled the log2 copy number ratios for each sample by dividing the negative log2 ratios by the deep deletion threshold and positive copy number ratios by the amplification ratio.

**Somatic mutation rate comparison**

We compared the mutation, copy number, and gene fusion rates between mouse and human TCGA cancers. The mutation and copy number data were downloaded for all cancer types available on Broad GDAC Firehose ([http://gdac.broadinstitute.org/)](http://gdac.broadinstitute.org/%29) data version 2016\_01\_28. The gene fusion data was obtained from Yoshihara et al[23]. Two measures were used for comparing copy number changes: number of copy number breakpoints and genomic instability index, as previously defined [24].

**Antibodies:**

The following antibodies were obtained from Cell Signaling Technology (Beverly, MA) and were used at 1:1000 dilution: pAKT (#3787), pERK (#4370), pS6 (# 2317), pFGFR(#3476), pMET (#3077), AKT (#4691), ERK (#4695), S6 (#2317), FGFR2 (#23328), MET (#3127), PI3K p110alpha (#4255), HA tag (#2367), HER2 (#2165) and PTEN (#9188). Other antibodies used in this paper are as follows: DLG1 (Santa Cruz Biotechnologies, sc-26530), BRAF (Santa Cruz Biotechnologies, SC-166), RAF1 (Santa Cruz Biotechnologies, sc-227), DNM3 (Sigma, SAB2107111), pFRS2 (R&D Systems, #AF5126), ERalpha (Santa Cruz SC-542), PR (Abcam, Ab63605) and p110-beta (Abcam, Ab28356).

**Mass spectrometry**: Gel bands were excised from the silver stained gel, and cut into 1-mm3 cubes. Gel cubes were destained with a 30 mM potassium ferricyanide/100 mM sodium thiosulfate solution, washed with milli-Q water and dehydrated in a SpeedVac vacuum centrifuge. Proteins were reduced with 20 mM DTT for 1 hour at 56°C and alkylated with 60 mM iodoacetamide for 45 minutes at room temperature and in-gel digested with 12.5 ng/uL trypsin (modified sequencing grade, Promega) in 100 mM ammonium bicarbonate, pH 8 for 45 min on ice, then left to incubate overnight in 100 mM ammonium bicarbonate, pH 8 at room temperature. Peptides were extracted from the gel pieces with 50% acetonitrile/5 % formic acid and 100% acetonitrile. Organic solvent was evaporated from extracted peptides in a SpeedVac vacuum centrifuge. Peptides were resuspended in 0.1% acetic acid before loading onto a precolumn [100 μm ID × 10 cm packed with 10 μm C18 beads (YMC gel, ODS-A, 12 nm, S-10 μm, AA12S11)], which was rinsed with 0.2 M acetic acid for 10 min before LC-MS analysis. The washed precolumn was connected in series with an in-house packed analytical capillary column [50 μm ID × 12 cm packed with 5 μm C18 beads (YMC gel, ODS-AQ, 12 nm, S-5 μm, AQ12S05)] with an integrated electrospray tip (∼1 μm orifice). Peptides were eluted using a 60-minute gradient from 9 to 70% acetonitrile in 0.2 M acetic acid at a flow rate of 0.2 ml/min, with a flow split of ~10,000:1, yielding a final electrospray flow rate of ~20 nL/min. Phosphopeptides were analyzed using a Thermo Q Exactive Hybrid Quadrupole-Orbitrap Plus mass spectrometer. Standard mass spectrometric parameters were as follows: spray voltage, 2 kV; no sheath or auxiliary gas flow, heated capillary temperature, 250°C; S-lens radio frequency level of 50%. The Q Exactive was operated in data-dependent acquisition mode. Full-scan MS spectra [mass/charge ratio (m/z), 350 to 2000; resolution, 70,000 at m/z 200] were detected in the Orbitrap analyzer after accumulation of ions at 3e6 target value based on predictive AGC from the previous scan. For every full scan, the 15 most intense ions were isolated (isolation width of 0.4 m/z) and fragmented (collision energy: 29%) by higher-energy collisional dissociation with a maximum injection time of 350 ms and 35,000 resolution. Dynamic exclusion was set to 30 s. Raw mass spectral data files were loaded into Proteome Discoverer version 1.4.1.14 (DBversion: 79) (Thermo) and searched against the mouse SwissProt database (sequence entries: 16,981) using Mascot version 2.4 (Matrix Science). Results were filtered for commonly occurring protein contaminants.

**Primary Mammary Epithelial Cell Isolation and Immortalization**: Primary mammary epithelial cells from Brca1-mutation carriers or healthy donors were collected according to a Institutional Review Board (IRB)- approved protocol. Briefly, fresh tissue was minced, suspended in medium containing Collagenase/Hyaluronidase (StemCell Technologies) and agitated for 3 hours. Following trypsinization for 5 minutes, cells were plated in MCF10 medium on 0.1% gelatin-coated tissue culture dishes. Primary cells were infected with retrovirus expressing SV40 large antigen and human telomerase reverse transcriptase (hTERT) at passage 3, and infected cells were selected with 2ug/ml puromycin.

**Retrovirus, Lentivirus Production and Cell Infection:** Briefly, 293 cells (ATCC) were transfected with 2ug target vector, 1.33μg of vector containing retrovirus or lentivirus Gag/Pol and 0.73μg of VSVG vector using Lipofectamine 2000 (ThermoFisher). Virus containing supernatant were collected, filtered through 4.5μm filter (Whatman), and were used to infect target cells in the presence of PolyBrene (Sigma).

**References**

1. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. Nature Research; 2017;35:316–9.

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

3. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011;2011:bar030–0.

4. Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. Nucleic Acids Research. Oxford University Press; 2017;45:D723–9.

5. Paquet ER, Hallett MT. Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype. JNCI Journal of the National Cancer Institute. 2014;107:dju357–7.

6. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.

7. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology. 2009;27:1160–7.

8. Nicorici D, atalan MS, Edgren H, Kangaspeska S, gi AM, Kallioniemi O, et al. **FusionCatcher**-a tool for finding somatic fusion genes in paired-end RNA-sequencing data. biorxiv [Internet]. 2014. Available from: http://www.biorxiv.org/content/biorxiv/early/2014/11/19/011650.full.pdf

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

10. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

11. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. Cold Spring Harbor Lab; 2012;22:568–76.

12. Danecek P, Nellåker C, McIntyre RE, Buendia-Buendia JE, Bumpstead S, Ponting CP, et al. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. Genome Biol. BioMed Central; 2012;13:26.

13. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. Oxford University Press; 2014;30:1006–7.

14. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2014;6:80–92.

15. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics. Oxford University Press; 2015;31:3673–5.

16. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. Nature Research; 2013;500:415–21.

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

18. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol. 2016;34:155–63.

19. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology. American Society of Clinical Oncology; 2017;:1–16.

20. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7.

21. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016;12:e1004873.

22. Kuilman T, Velds A, Kemper K, Ranzani M, Bombardelli L, Hoogstraat M, et al. CopywriteR: DNA copy number detection from off-target sequence data. Genome Biol. BioMed Central; 2015;16:49.

23. Geyer FC, Berman SH, Marchiò C, Burke KA, Guerini-Rocco E, Piscuoglio S, et al. Genetic analysis of microglandular adenosis and acinic cell carcinomas of the breast provides evidence for the existence of a low-grade triple-negative breast neoplasia family. Mod Pathol. Nature Publishing Group; 2017;30:69–84.

23 Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene. 2015;34:4845–54.

24. Burrell RA, McClelland SE, Endesfelder D, Groth P, Weller M-C, Shaikh N, et al. Replication stress links structural and numerical cancer chromosomal instability. Nature. 2013;494:492–6.