Supplemental Experimental Procedures

Pathology review of 817 breast cancer cases

For scoring histologic type, pathologists in the expert pathology committee (EPC) applied the same criteria used in clinical practice to diagnose histologic type (options included IDC, ILC, and Mixed IDC/ILC, along with other rarer histologic types). We then created a final consensus diagnosis (DX) for the lobular project incorporating the path report (PR) and the EPC majority diagnosis (EPC), according to these rules:

- If (EPC=IDC AND PR=IDC) OR (EPC=IDC AND PR=MIXED)
    - Then DX=IDC
- If (EPC=ILC AND PR=ILC) OR (EPC=ILC AND PR=MIXED) OR (EPC=MIXED AND PR=ILC)
    - Then DX=ILC
- If (EPC=ILC AND PR=IDC) OR (EPC=MIXED AND PR=MIXED) OR (EPC=IDC AND PR=ILC) OR (EPC=MIXED AND PR=IDC)
    - Then DX=MIXED
- If (EPC=OTHER OR PR=OTHER)
    - Then DX=OTHER

Somatic Mutation Analysis

WUSTL Read Realignment
Imported data were realigned to GRCh37-lite with bwa v0.5.9. Defaults are used in both bwa aln and bwa sampe (or bwa samse if appropriate) with the exception that for bwa aln four threads are utilized (-t 4) and bwa's built in quality-based read trimming (-q 5). ReadGroup entries were added to resulting SAM files using gmt sam add-read-group-tag. This SAM file was converted to a BAM file using Samtools v0.1.16, name sorted (samtools sort -n), mate pairings assigned (samtools fixmate), resorted by position (samtools sort), and indexed using gmt sam index-bam.

WUSTL Read Duplication Marking and Merging
Duplicate reads from the same sequencing library were merged using Picard v1.46 MergeSamFiles and duplicates are then marked per library using Picard MarkDuplicates v1.46. Lastly, each per-library BAM with duplicates marked is merged together to generate a single BAM file for the sample. For MergeSamFiles we run with SORT_ORDER=coordinate and MERGE_SEQUENCE_DICTIONARIES=true. For both tools, ASSUME_SORTED=true and VALIDATION_STRINGENCY=SILENT are specified. All other parameters are set to defaults. Samtools flagstat is run on each BAM file generated (per-lane, per-library, and final merged).

WUSTL Somatic Mutation Calling
We detected somatic point mutations using Samtools v0.1.16 (samtools pileup –cv -A -B), SomaticSniper v1.0.2 (bam-somaticsniper -F vcf -q 1 -Q 15), Strelka v1.0.10

(with default parameters except for setting isSkipDepthFilters = 0), and VarScan v2.2.6 (--min-coverage 3 --min-var-freq 0.08 --p-value 0.10 --somatic-p-value 0.05 --strand-filter 1). We detected somatic indels using the GATK 1.0.5336 (-T IndelGenotyperV2 --somatic --window_size 300 -et NO_ET), retaining only those which were called as Somatic, Pindel v0.2.2 (-w 10; with a config file generated to pass both tumor and normal BAM files set to an insert size of 400), Strelka v1.0.10 (with default parameters except for setting isSkipDepthFilters = 0), and VarScan v2.2.6 (--min-coverage 3 --min-var-freq 0.08 --p-value 0.10 --somatic-p-value 0.05 --strand-filter 1).

WUSTL Annotation, Readcounts, and Filtering

Somatic mutations annotated using GENCODE release 14 downloaded from Ensembl 69. Variants were filtered if they occurred exclusively in Intronic, Intergenic, 3'UTR, or 5'UTR, or gene flanking regions. Supporting readcounts were obtained from the tumor and normal BAM using bam-readcount (https://github.com/genome/bam-readcount). Variants were filtered if the normal aliquot had less than 8x coverage of the reference allele or more than 1 variant supporting read in the normal BAM. A minimum threshold of two supporting reads and a minimum variant allele fraction (VAF) of 10% were required in the tumor BAM. Recurrent artifacts and common germline dbSNPs identified with a GMAF>0 in dbSNP137 were also filtered.

Mutation Calling

The breast cancer mutation list (MAF file) from the latest available TCGA DCC Archive (genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.1.1.0) was downloaded and checked. Some missing somatic variants were recovered from the intermediate variant lists generated by the variant calling bioinformatics pipeline at the TCGA Genome Sequencing Center (GSC). These variant were previously filtered out by the pipeline, because of a dbSNP-based false-positive filter. AKT1 E17K and PTEN R130Q are among several submissions to dbSNP that are incorrectly tagged as germline sites. After recovering the missed calls, calls were removed from two FFPE tumors (TCGA-A7-A26E-01B, TCGA-AC-A3OD-01B) with excessively more calls than their fresh frozen counterparts (also in the cohort). Also removed calls from a sample (TCGA-A8-A08C) that the GSC determined to be a tumor/normal sample swap, based on observing loss-of-heterozygosity events in the matched normal (TCGA-A8-A08C). Removed calls with fewer than 8 total reads in either tumor or normal.

Additional point mutations were called by running UNCeqR (Wilkerson et al., 2014) on Exome-seq and RNA-seq data, and additional indels were called using bwa-mem (Li and Durbin, 2009) for alignment, Abra (Mose et al., 2014) for local reassembly, and Strelka (Saunders et al., 2012) for calling somatic indels. Of these 127946 calls, 8755 were removed at germline sites with a global minor allele frequency (GMAF) >0.05%, based on 1000 genomes Phase 1 data. Further removed calls with fewer than 8 total reads in either tumor or normal, calls with >1% variant allele fraction (VAF) in normal, calls with tumor DNA+RNA variant supporting reads <2, and calls with tumor DNA+RNA VAF <10%.

Column names were standardized; the mutation lists concatenated, de-duplicated, and sorted by sample ID and genomic loci. Adjacent SNPs with matched sample IDs were merged together as DNPs. Heterozygosity status in columns 12 and 13 of the MAF was standardized across all calls using a simple 80% VAF cutoff. The vcf2maf tool (DOI:10.5281/zenodo.14107) was used with the Gencode v19 transcript database, and Ensembl's VEP v75 annotator, to standardize the selection of isoforms to which variant effects are mapped. Gene names were updated to the latest HUGO aliases based on genenames.org, and Entrez IDs were retrieved and backfilled using NCBI's Entrez tools.

RNA-seq analysis

RNA sequencing was performed at University of North Carolina at Chapel Hill on the Illumina HiSeq and data were processed using methods previously described (Hoadley et al., 2014). Briefly, resulting sequencing reads were aligned to the human hg19 genome assembly using MapSlice (Wang et al., 2010). Gene expression was quantified for the transcript models corresponding to the TCGA GAF 2.13 using RSEM4 and normalized within samples to a fixed upper quartile. Gene expression data is available at the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/). Upper quartile normalized RSEM data were log2 transformed. Genes with a value of zero following log2 transformation were set to the missing value and genes with missing values in greater than 20% of samples were excluded from analyses. PAM50 classification, including calculation of the Proliferation signature, was performed as previously described (Parker et al., 2009).

Significance Analysis of Microarray (SAM) analysis was used to identify differentially expressed genes by comparing each subgroup to all other samples; an FDR of 0 was considered significant. To investigate pathway activity, the 11-gene PAM50 Proliferation signature (Parker et al., 2009) as well as Macrophage-associated signatures including those that measure CD68, Macrophage Colony Stimulating Factor (MacCSF), Macrophage Th1 (MacTh1), and T-cell Receptor Signaling (TCR,) signatures (Iglesia et al., 2014). A t-test was used to statistically assess differences between samples in a given subgroup and all other ILC tumors.

To determine breast cancer intrinsic subtypes based on the PAM50 signature, first, the TCGA mRNA-seq data were subsampled to match the ER distribution of the training set used for the PAM50. Second, the entire TCGA 817 data set was adjusted to the median gene expression calculated for the PAM50 genes determined from the ER balanced subset; intrinsic subtyping was then done as previously described (Cancer Genome Atlas, 2012).

miRNA-seq analysis

We generated microRNA sequence (miRNA-seq) data for 817 tumor samples using methods described previously portraits (Cancer Genome Atlas, 2012). We aligned

reads to the GRCh37/hg19 reference human genome, and annotated miRNA read count abundance with miRBase v16. While we used only exact-match read alignments for this, the BAM files that are available from cgHUB (cghub.ucsc.edu) include all sequence reads. We used miRBase v20 to assign 5p and 3p mature strand names to MIMAT accession IDs.

We identified groups of samples that had similar abundance profiles using unsupervised non-negative matrix factorization (NMF, v0.5.06) consensus clustering with default settings (Gaujoux and Seoighe, 2010). The input was a reads-per-million (RPM) data matrix for the ~300 (25%) most-variant 5p or 3p mature strands, which we parsed from the level 3 isomiR data files that are available from the TCGA data portal. After running a rank survey with 30 iterations per solution, we chose a preferred clustering solution from the cophenetic and average silhouette width score profiles, and then used 500-iterations for the main clustering run. We calculated a profile of silhouette widths from the NMF consensus membership matrix, and considered samples with relatively low widths within a cluster as atypical cluster members.

For the heatmap displayed, we included all miRs used in the NMF and ordered the samples by then NMF cluster solution. We transformed each row of the matrix by $\log_{10}(\text{RPM} + 1)$, then used the pheatmap v0.7.7 R package to scale and then cluster only the rows with a Euclidean distance measure.

To identify miRs that were differentially abundant (DA) between sample groups (eg. ILC vs IDC, mRNA class1 vs other, miRNA cluster1 vs other, etc), we used unpaired two-class SAMseq analyses with a read-count input matrix and an FDR threshold of 0.05 by samr 2.0 (Tusher et al., 2001) in R 2.15.0 (Table S1). For the figures, we filtered the results by removing miRs with median expression less than 50 RPKM in at least one of the two groups, and miRs for which the Wilcoxon adjusted p-value was greater than 0.05. The RPM filtering acknowledged potential sponge effects from competitive endogeneous RNAs (ceRNAs) that can make weakly abundant miRs less influential. Given this, we support assessing fold change at the same time as absolute miR abundance by adding, to each fold change barplot, a boxwhisker plot that shows the distribution of miR abundance in the two sample groups.

SNP-based copy number analysis

DNA from each tumor or germline sample was hybridized to Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute as previously described (McCarroll et al., 2008). Briefly, from raw .CEL files, Birdseed was used to infer a preliminary copy-number at each probe locus (Korn et al., 2008). For each tumor, genome-wide copy number estimates were refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor (Cancer Genome Atlas Research, 2011) (and Tabak B. and Beroukhim R. Manuscript in preparation). This linear combination of normal samples tends to

match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. Individual copy-number estimates then underwent segmentation using Circular Binary Segmentation (Olshen et al., 2004). As part of this process of copy-number assessment and segmentation, regions corresponding to germline copy-number alterations were removed by applying filters generated from either the TCGA germline samples from the ovarian cancer analysis or from samples from this collection. Segmented copy number profiles for tumor and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude to the set of inferred copy-number changes underlying each segmented copy number profile (Mermel et al., 2011). Significant focal copy number alterations were identified from segmented data using GISTIC 2.0 (Mermel et al., 2011). Allelic copy number, whole genome doubling and purity and ploidy estimates were calculated using the ABSOLUTE algorithm (Carter et al., 2012).

Array-based DNA methylation assay

Illumina Infinium DNA methylation HumanMethylation 27 (HM27) and HumanMethylation 450 (HM450) platforms were used to obtain DNA methylation profiles of 1,000 breast tumor tissue samples and 125 adjacent non-malignant prostate tissue samples. In order to monitor technical variations, each batch of samples was assayed with control cell line technical replicates. The HM27 array contains 27,578 probes, which target CpG sites near the transcription start site of 14,475 consensus coding sequencing (CCDS) in the NCBI Database. The HM450 array contains 485,777 probes, which include 482,421 CpG sites, 3,091 CpH sites, and 65 SNPs in human genome. It covers 96% of CpG islands and 99% of Refseq genes with multiple probes per gene located in promoter, 5'UTR, first exon, gene body, and 3'UTR. The detailed information of HM27 and HM450 is available from Illumina (www.illumina.com).

Sample and data processing
In order to profile DNA methylation, 1 ug of genomic DNA from each sample was bisulfite converted using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA). The completeness of bisulfite conversion and the amount of bisulfite-converted DNA was assayed by conducting MethyLight-based quality control (QC) reactions (Campan et al., 2009). All the samples that passed QC tests were whole-genome amplified and enzymatically fragmented to hybridize in the arrays. All arrays were scanned using the Illumina iScan technology and IDAT files were produced. IDAT files were processed with the R/Bioconductor package *methylumi*. DNA methylation data of TCGA BRCA samples were generated using the *EGC.tools* R package (https://github.com/uscepigenomecenter/EGC.tools).

TCGA Data Packages

There are 3 data levels for DNA methylation data. The description of each data level and file is available on the TCGA Data Portal website (http://tcga-data.nci.nih.gov/tcga).

The following data archives were used for the analyses described in this manuscript.
Jhu-usc.edu_BRCA.HumanMethylation27.Level_3.1.1.0
Jhu-usc.edu_BRCA.HumanMethylation27.Level_3.2.1.0
Jhu-usc.edu_BRCA.HumanMethylation27.Level_3.3.1.0
Jhu-usc.edu_BRCA.HumanMethylation27.Level_3.4.1.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.1.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.2.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.3.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.4.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.5.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.6.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.7.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.8.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.9.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.10.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.11.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.12.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.13.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.14.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.15.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.16.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.17.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.18.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.19.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.20.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.21.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.22.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.23.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.24.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.25.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.26.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.27.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.28.8.0
Jhu-usc.edu_BRCA.HumanMethylation450.Level_3.29.8.0

Merging HM27 and HM450 data
In order to merge DNA methylation data of HM27and HM450, we first fitted a LOESS regression model between two platforms using cell line control technical replicates. M values ($\log_2$ (Methylated intensity/Unmethylated intensity)) of 25,978 probes from HM450, found common in the HM27, were normalized against HM27. Out of 25,978 probes, 20,297 probes were selected for the analyses since some of probes 1) have a detection P value greater than 0.05, 2) have a SNP within 10 bp of the

interrogated CpG site, 3) are located in a repeat element (*BSgenome.Hsapiens.UCSC.hg19* R package), 4) are not uniquely aligned to the human genome (UCSC hg19, Feb 2009), 5) span known regions of small insertions and deletions (indels) in the human genome (UCSC hg19, Feb 2009), 5) show high technical variances after the platform correction across technical replicates. For downstream analyses, M values were transformed to β values (0 indicates unmethylation and 1 indicates methylation).

Recurrent Genomic Alterations in Breast Cancer and Breast Cancer Subtypes

We search for statistically significant recurrence of copy number alterations and somatic mutations across all 817 breast cancer samples using the GISTIC 2.0 (Beroukhim et al., 2010; Mermel et al., 2011) and MutSigCV (Lawrence et al., 2013) algorithms, respectively. MutSigCV takes into account gene-specific differences in background mutation rate by using genomic covariates. Genes with q-values less than 0.1 were considered significant. We performed MutSigCV and GISTIC analyses independently on all three ILC expression clusters, on ILC, ILC Luminal A, IDC, IDC Luminal A, IDC Luminal B, IDC Her2+, and IDC Basal-like subtypes and on the complete data set. We then combined the resulting recurrently mutated genes and recurrent regions of copy number gain and losses to define a consolidated set of recurrent genomic alterations in breast cancer, which accounts for the intrinsic heterogeneity of the disease. We used these selected set of events to derive binary alteration calls for each sample (1 = altered, 0 = wild-type) as previously described (Cancer Genome Atlas, 2012). Binary alteration calls were used to define the alteration frequency of each event within each breast cancer subtype. For each comparison between subtypes, only alterations occurring in at least 6 samples (corresponding to ~1% of the combined IDC and ILC dataset – n=617 - and 2% of the combined IDC Luminal A and ILC Luminal A dataset – n=307) were used and statistical significant differences were evaluated by Fisher's exact test.

DNA methylation of *CDH1* gene

For promoter region, DNA methylation profiles of probes, located in 1,500bp windows of *CDH1* transcription start site, were studied using level3 HM27 and HM450 data (Suppl Fig 2d-h). Supplemental figure 2f was generated by using 553 tumors, 69 normals, and 2 leukocyte samples, which were arrayed on HM450. Thirteen available HM450 probes in *CDH1* promoter region were sorted based on their genomic coordinates, and tumors grouped by histology were ordered by increasing *CDH1* gene expression level. The heatmap shown in supplemental figure 2g was plotted using the merged HM27 and HM450 data of all 817 breast tumors in freeze list, 90 normals, and 2 leukocyte samples. Leukocyte fraction was estimated based on the methods we described previously (Carter et al., 2012). All 817 tumor samples were ordered by increasing leukocyte fraction estimate. Six probes found in merged HM27 and HM450 data were ordered by genomic location (Suppl Fig 2e). In order to assess the gene expression level associated with DNA methylation change, level 3 RNA-seq RSEM data were obtained from the TCGA Data Portal website

(http://tcga-data.nci.nih.gov/tcga). Level 3 RNA-seq RSEM data were log2 transformed (log_2 (RSEM+1)) to generate scatterplots (x-axis: DNA methylation level, y-axis: gene expression level) (Suppl Fig 2e).

We also used whole genome bisulfite sequencing to characterize DNA methylation levels at 157 CpGs located in *CDH1* promoter region (1,500 bp upstream to 1,500bp downstream of *CDH1* transcription start site). Among these, 7 CpGs intersected with HM450 probes and 4 CpGs intersected with HM27 probes. DNA methylation levels at these CpGs were highly correlated (Suppl Fig 2h).

In order to investigate DNA methylation levels including enhancer regions of *CDH1* gene (Rhie et al., 2014)z a total of 34 HM450 probes spanning genomic loci 50kb upstream and 50kb downstream of *CDH1* transcription start site were visualized in the Suppl Fig 2i. In this heatmap, probes were ordered based on their genomic coordinates, and tumors were grouped by histology then unsupervised clustering was performed.

FOXA1 DNA-amino acid and amino acid-amino acid interactions

Experimentally validated DNA interactions between the FOXA1 protein and residues in the Fork-head domain have been derived from (Gajiwala and Burley, 2000), whereas predicted DNA interactions have been computed using the PDA algorithm (Kim and Guo, 2009) through the web service WebPDA (http://bioinfozen.uncc.edu/webpda/).

We evaluate amino acid proximity in the 3D space for residues in the FOXA1 fork-head domain, as the minimal distance among all atomic distances between each residue pair. Atomic coordinates for residues in the fork-head domain have been derived from the 3D crystal structure of FOXA3 fork-head domain (PDBid: 1VTN).
Graphical representations of the fork-head domain 3D structures have been generated using PyMOL (OSX version MacPyMOL). Structural elements described in this manuscript can be isolated from the whole structure using the following PyMOL script:

```
sele DNA, resi 1-33
sele forkhead_domain, resi 117-218
sele w2_loop, resi 196-218
sele msh_mutations, resi 125+175+196+199+202-205+208+209+212+215
sele other_mutations, resi 143+163+182
sele DNA_contact_residues, resi 162+165+169+172-174+191+193+209-211
```

DNA methylation at FOXA1 binding sites

FOXA1 ChIP-seq data sets from breast cancer cells were obtained from previous studies (Ross-Innes et al., 2012; Wang et al., 2012). HM450 probes, within 100bp of FOXA1 binding sites, were selected to investigate DNA methylation levels at FOXA1

binding sites (n=85,242). The heatmap in Figure 3f was generated using median DNA methylation level of the 3,976 most variable probes at FOXA1 binding sites and of the 2,000 most variable probes at non-FOXA1 binding sites (n=400,335) with median DNA methylation levels in normal samples within the same range of the 3,976 probe set (0.5 < beta < 0.7). Tumor samples in Figure 3f only includes samples profiled with the HM450 platform (n = 659) and were ordered by decreasing FOXA1 mRNA expression (from left to right). The same sorting criterion was applied to normal samples.

Differential expression analysis between FOXA1 mutant and wild-type cases

All differential expression analyses have been performed using the `limma` R package with `voom` correction (Law et al., 2014) to enable the analysis of RNA-seq data. FOXA1 targets defined by the presence of FOXA1 binding motif in the promoter were derived from the Molecular Signature DataBase (MSigDB) (Liberzon et al., 2011), gene set ID: *V$HNF3ALPHA_Q6*. FOXA1 targets were also defined by genomic loci corresponding to the most variable methylation probes matching FOXA1 binding sites (identified as previously described) (Suppl. Table 3). Comparisons have been separately for all FOXA1 mutations and for FOXA1 mutations within the mutation structural domain (MSH) we identified. Genes obtaining an FDR adjusted p-value < 0.1 were considered as significantly differentially expressed (Suppl. Tables 4 and 5). Gene Set Enrichment Analysis (GSEA) was performed on the gene sets containing FOXA1 targets using the `romer` function included in the `limma` package.

RPPA analysis

Data were generated, processed and normalized as previously (Hoadley et al., 2014). Replication Based Normalized (RBN) Reverse Phase Protein Array (RPPA) data containing expression levels for 187 protein and phosphorylated proteins for 633 samples within the larger dataset (n=817) were utilized to identify differentially expressed proteins (Suppl. Table 6). To identify proteins and phosphoproteins that are differentially expressed between lobular and ductal tumors, we restricted our analyses to the Luminal A Lobular (n=65) and Luminal A ductal (n=158) samples to account for differences in the distribution of molecular subtype between the histological subtypes. A t-test was used to identify proteins and phosphorylated proteins that were expressed at significantly different levels ($p<0.05$) between each subset of patients. To identify significantly expressed proteins and phosphoproteins between each ILC subtype, samples in each subset of tumor were compared to all other samples by t-test; proteins expressed at significantly different levels are shown in Figure 5b.

To assess pathway activity using RPPA data, tumor samples were scored using a series of protein expression signatures, as previously described (Akbani et al., 2014), and a t-test used to assess differences in pathway activity between a given subgroup and all other samples (Suppl. Tables 1 and 6). To assess the relationship

between mRNA-defined molecular subtype and RPPA subtype, samples were assigned to RPPA-defined subtype, as previously described (Cancer Genome Atlas, 2012), and a Fisher's exact test used to assess the relationships.

PARADIGM integrated pathway analysis of copy number and expression data

Integration of copy number, mRNA expression and pathway interaction data was performed on 817 BRCA samples using the PARADIGM software (Vaske et al., 2010). Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes using pathway interactions, copy number and expression data from each patient sample.

Pathways were obtained in BioPax Level 3 format, from the NCIPID and BioCarta databases (http://pid.nci.nih.gov) and the Reactome database (http://reactome.org). Gene identifiers were unified by UniProt ID then converted to Human Genome Nomenclature Committee's HUGO symbols using mappings provided by HGNC (http://www.genenames.org/). Altogether, 1,524 pathways were obtained. Interactions from all of these sources were then combined into a merged Superimposed Pathway (SuperPathway). Genes, complexes, and abstract processes (e.g. "cell cycle" and "apoptosis") were retained and henceforth referred to collectively as *pathway features*. The resulting pathway structure contained a total of 19504 features, representing 7369 proteins, 9354 complexes, 2092 families, 82 RNAs, 15 miRNAs and 592 abstract processes.

Thresholded gene level copy number data from GISTIC was obtained from Firehose. Log2 transformed, median-centered mRNA data was rank transformed based on the global ranking across all samples and all genes and discretized (+1 for values with ranks in the highest tertile, -1 for values with ranks in the lowest tertile, and 0 otherwise) prior to PARADIGM analysis. From these data, the PARADIGM algorithm infers an integrated pathway level (IPL) for each gene that reflects a gene's activity in a tumor sample relative to the median activity across all tumors. PARADIGM IPLs of the 19504 features within the SuperPathway is available within the Lobular Breast Cancer data snapshot.

PARADIGM inferred pathway biomarkers differentiating Luminal A invasive ductal and Luminal A invasive lobular carcinomas

We considered in this analysis 201 Luminal A (LumA) invasive ductal carcinomas (IDC) and 106 LumA invasive lobular carcinomas (ILC). An initial minimum activity filter (at least 1 sample with absolute activity > 0.05) was applied, resulting in 16267 features (6490 proteins, 7446 complexes, 1937 families, 13 RNAs, 15 miRNAs and 366 abstract processes). PARADIGM IPLs differentially activated between LumA IDC and LumA ILC were identified using the t-test and Wilcoxon Rank Sum test with BH FDR correction. Only features deemed significant (FDR corrected $p<0.05$) by both tests and showing an absolute difference in-group means > 0.05 were selected. Differentially activated IPLs were then filtered by connectivity

within the SuperPathway, such that only interconnected features through regulatory interactions (i.e. activation, inhibition) were retained. This regulatory sub-network of differentially activated IPLs was further pruned to include only features linked through regulatory nodes with >5 outgoing edges and was visualized using Cytoscape. A zoomed-in view of the first-degree neighbors of the PARADIGM feature 'Active AKT family' within this pruned regulatory subnetwork of differentially activated IPLs was created from Cytoscape.

PARADIGM inferred pathway biomarkers of ILC subtypes

All 127 ILCs were considered in this analysis. A minimum activity filter (at least 1 sample with absolute activity > 0.05) is applied, resulting in 16222 features. IPLs differentially activated between ILC Class 1 (n=50) and the other subtypes were identified using the t-test and Wilcoxon Rank Sum test with Benjamini-Hochberg (BH) FDR correction. Only features deemed significant (FDR corrected $p<0.05$) by both tests and with absolute difference in-group means > 0.05 were selected. Differentially activated IPLs were then filtered by connectivity within the SuperPathway structure, such that only interconnected features through regulatory interactions (i.e. activation, inhibition) were retained. From this regulatory sub-network of differentially activated features, nodes with ≥ 5 outgoing edges were selected. Similar analyses were performed to identify regulatory nodes with differential IPLs in ILC Class 2 (n=50) and ILC Class 3 (n=27). The IPLs of the resulting regulatory hubs were scaled to median 0 and standard deviation 1 and visualized in a heatmap generated using the heatmap.plus package.

Mutually exclusive alterations in Invasive Lobular Carcinoma

The MEMo algorithm (Ciriello et al., 2012) was used to identify recurrent and mutually exclusive alterations in 127 ILC cases. In total we identify 31 modules with Step-down adjusted p-value < 0.05 (Suppl. Table 7). Many of these modules are sub-modules of each other and most of them include alterations converging or downstream of the PI3K/Akt pathway. Besides *PTEN* homozygous deletion and mutations, which are enriched in ILC, mutually exclusive alterations activating Akt signaling identified by MEMo include *AKT1*-E17K activating mutations, *KRAS* activating mutations (G12C/S), *NF1* loss of function mutations, and DNA amplification and overexpression of *GAB2,* all acting upstream of the PI3K complex (Gu and Neel, 2003; Shaw and Cantley, 2006). Amplification and overexpression of mir21, a *PTEN* targeting micro-RNA (Lou et al., 2010; Meng et al., 2007), was also observed. Additional alterations in the module are events acting downstream of Akt, such as amplification and overexpression of *IKBKB*, a negative regulator of the TSC-complex inhibiting mTOR (Cully et al., 2006), *RPS6KB1* encoding for the p70S6K protein and of the oncogene MYC, and loss-of-function mutations and deletions targeting *MAP2K4* and *MAP3K1*.

Mutually exclusive alterations upstream of the pathway were singled out in Figure 4d and separately tested using MEMo statistical framework that preserves both

number of alterations per gene and number of alterations per sample. In Figure 4e, the average RPPA Z-score for phospho-Akt at T308 and S473 was compared in samples with at least one alteration upstream of Akt and in wild type samples for these events.

ILC mRNA subtypes

To identify molecular subtypes of lobular breast tumors, we utilized mRNAseq expression data from the 106 Luminal A samples that comprise 83% of the lobular tumors in our cohort in order to limit the confounding influence of molecular subtype. Using this subset of tumors, we first filtered the mRNA expression data to those genes that were present in more than 80% of all samples. These data were further filtered to the 1,000 most differentially expressed genes based on standard deviation (std dev >1.735) and the data were then imputed to replace missing values Consensus Cluster Plus Analysis (Wilkerson and Hayes, 2010) was then used to assess the optimal number of subgroups between 2 and 10 subgroups. Consensus CDF and delta were used to determine k=3 as the optimal number of tumor subgroups (Suppl. Fig 5a-c). Principal component analysis (PCA) demonstrated variability between each group of tumors but also suggested that some common features would be identified (Suppl. Fig 5d). To build a quantitative classifier such that future samples could be assessed, we further restricted our training data to those samples that have a positive silhouette width for each subgroup (n=89) and ClaNC (Classification to Nearest Centroid) (Dabney, 2005) was used to identify a 60-gene classifier (Suppl. Table 8) which showed the lowest level of cross-validation (CV) error and that largely recapitulates these subgroups with 92% concordance between the two strategies (Suppl. Fig. 5e-g). Using this classifier we assigned all 127 ILC samples in our dataset (Suppl. Table 8). To classify the 148 lobular samples in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset, we merged the mRNA expression data from the TCGA (n=817) and METABRIC (n=1992) datasets at the gene level. Once the data were merged, the list of genes was restricted to the 57 (out of 60) genes in the ILC classifier that are present in the METABRIC data. To remove variation between the datasets, the mean of each gene was set to 0 and the standard deviation set to 1 across the combined dataset. Samples were then assigned to each subgroup using CLaNC (Suppl. Fig. 5j-k and Suppl. Table 8).

To assess differences in disease-specific survival and overall survival between subgroups, we analyzed ILC samples from the METABRIC cohort (Curtis et al., 2012). Clinical data were acquired August, 2013. The 148 tumors defined as ILC by Curtis et al were classified into each of the three ILC subgroups as detailed about. Differences in overall survival and disease-specific survival were determined for each pair of subgroups; these results are reported in Figure 5d and 5e, respectively. To investigate the effect of proliferation on prognosis, the 11-gene PAM50 proliferation signature was calculated for each of the 148 ILC tumors, the dataset was divided into 'high' and 'low' proliferation using the median and overall and disease-specific survival were determined (Suppl. Fig 5l-o). Survival analyses are

not reported for 127 sample TCGA cohort due to the immaturity of the clinical data. Tumor purity was assessed by ABSOLUTE (Carter et al., 2012) (Fig. 5a) and differences between groups assessed by a t-test (Suppl. Fig. 5p).

Molecular classification of mixed ductal/lobular breast tumors

ISOpure

To study the individual contribution of IDC and ILC origin in mixed ductal/lobular invasive carcinoma of the breast we implement the ISOpure step 1 algorithm using Matlab using standard parameter (Quon et al., 2013). To deconvolute tumor sample heterogeneity this method build a statistical model representing the tumor component explained by multiple reference samples based on RNA-seq expression data. For each mixed IDC/ILC sample, we calculated the fraction of tumor explained by IDC and ILC component as well as the fraction of sample that cannot be explained by the reference samples. We used randomly selected 50 IDC and 50 ILC cases as reference populations. As controls, we used as queries all IDC (n=440) and ILC (n=77) cases not included as reference and 153 GBM samples from the TCGA dataset (Brennan et al., 2013). To illustrate the ILC and IDC like component in mixed ductal/lobular invasive carcinoma, we report the ratio of the two components.

Query-OncoSign

To assess genetic similarity between mixed tumors and ILC and IDC based on a set of selected recurrent mutations and copy number alterations (Suppl. Table 2), we used a modified version of the OncoSign algorithm (Ciriello et al., 2013). Briefly, OncoSign builds a bipartite network where nodes are either samples or alterations, and each alteration is connected to the set of samples where it was observed. Given this network representation, OncoSign partitions samples into classes while maximizing the bipartite network modularity associated to each candidate partition. The partition with the maximal bipartite modularity is returned as solution [ref]. Here, we started from an already existing partition where ILC and IDC samples were pre-classified in the corresponding histological subgroups and IDC samples were further subdivided by PAM50 subtypes (normal-like cases were excluded from the analysis). We refer to this set of classes as *reference classes* and we defined in total 5 reference classes: ILC, IDC Luminal A, IDC Luminal B, IDC Her2+, IDC Basal-like. Mixed cases were each assigned to a separate set, each containing only one sample. We refer to these singleton sets as *query elements*. Each query element was iteratively assigned to one of the existing reference classes by maximizing the overall bipartite network modularity. It should be clear that this approach does not define a classifier. The reference classes are indeed defined independently of the features (CNA and mutations) and therefore such features are not necessarily discriminant of the pre-defined classes. To account for potential biases induced by the order followed to assign the query elements and to test whether the set of features we used are discriminant of the reference classes, we ran this approach over 100 boot-strapped iterations where at each iteration 5% of samples from the reference classes were added to the list of query elements. At the end each mixed sample receives an assignment score for each reference class defined as the fraction

of iterations it has been assigned to each class. Alteration frequencies were scaled to prevent most frequent alterations from dominating the assignments: each alteration had therefore an associated weight $w = (1-f)^k$, where f if the alteration frequency, and k a scaling parameter. In this study we chose k = 3 as the integer k that maximizes the fraction of correct re-assignment of ILC and IDC samples to the original group (62% for all 5 references classes, 70% when IDC samples are counted as one class).

ElasticNet

Elastic net modeling was used to assess the genetic relationships of tumors with a mixed ILC-IDC histology as compared to those tumors classified as purely ILC or IDC taking into account copy number alterations, somatic mutations, pathway signaling as determined by gene expression modules, and mRNA expression data. In total, we considered 961 features including 409 gene expression modules (Fan et al., 2011; Gatza et al., 2014) and 123 genes that were found to be mutated in the dataset at a frequency greater than 2.3%; 428 copy number alterations, including each chromosomal arm (n=44) and 384 additional focal regions that have been previously reported to be highly significant (Beroukhim et al., 2010; Weigman et al., 2012) as well as *CDH1* mRNA expression levels. To perform our analysis, we first excluded samples histologically classified as 'Other' as well as IDC and ILC samples characterized as basal-like. The remaining samples were divided into training (66.6%, n=339) and testing (33.4%, n=170) cohorts stratified by IDC, ILC and PAM50 subtypes. IDC samples were coded as 1 while ILC samples were coded as 0. To be certain that the training and testing datasets were balanced in terms of IDC, ILC and PAM50 subtype composition, the R package "sampling": Survey Sampling (http://cran.r-project.org/web/packages/sampling/index.html) was used. We next utilized the R package "glmnet": Lasso and Elastic-Net Regularized Generalized Linear Models (http://cran.r-project.org/web/packages/glmnet/index.html) to build a model capable of predicting IDC and ILC subtype using only the training subset of the data. Using the training data, we performed a 10 fold cross validation (CV) (family="binomial", type.measure="auc") to identify each parameter of the elastic net (alpha and lambda) model. By calculating the AUC (Area Under ROC Curve) of the validation dataset, we selected as the optimal parameters those that generated the highest AUC. Using the training data and the optimal parameters as determined by 10-fold CV, we built a final, optimized model. This model was then applied to both the training and the testing data, and the score calculated, as a continuous variable, each sample. The optimized model was then used to generate an ROC curve for the training data. Finally, in order to compute optimal thresholds such that samples with a mixed IDC-ILC histology could be classified as ILC-like or IDC-like, we used the R Package "OptimalCutpoints": Computing optimal cut-points in diagnostic tests (http://cran.r-project.org/web/packages/OptimalCutpoints/index.html); this analysis was performed on the training data alone. For the testing data, a sample with a model score below the threshold was predicted as ILC-like whereas samples with a model score greater than the cut-point were predicted as IDC-like.

References

Akbani, R., Ng, P.K., Werner, H.M., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.Y., Yoshihara, K., Li, J.*, et al.* (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nature communications *5*, 3887.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M.*, et al.* (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899-905.

Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H.*, et al.* (2013). The somatic genomic landscape of glioblastoma. Cell *155*, 462-477.

Campan, M., Weisenberger, D.J., Trinh, B., and Laird, P.W. (2009). MethyLight. Methods in molecular biology *507*, 325-337.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

Cancer Genome Atlas Research, N. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A.*, et al.* (2012). Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology *30*, 413-421.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. Genome research *22*, 398-406.

Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nature genetics *45*, 1127-1133.

Cully, M., You, H., Levine, A.J., and Mak, T.W. (2006). Beyond PTEN mutations: the PI3K pathway as an integrator of multiple inputs during tumorigenesis. Nature reviews Cancer *6*, 184-192.

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y.*, et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346-352.

Dabney, A.R. (2005). Classification of microarrays to nearest centroids. Bioinformatics *21*, 4148-4154.

Fan, C., Prat, A., Parker, J.S., Liu, Y., Carey, L.A., Troester, M.A., and Perou, C.M. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. BMC medical genomics *4*, 3.

Gajiwala, K.S., and Burley, S.K. (2000). Winged helix proteins. Current opinion in structural biology *10*, 110-116.

Gatza, M.L., Silva, G.O., Parker, J.S., Fan, C., and Perou, C.M. (2014). An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. Nature genetics *46*, 1051-1059.

Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC bioinformatics *11*, 367.

Gu, H., and Neel, B.G. (2003). The "Gab" in signal transduction. Trends in cell biology *13*, 122-130.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D., Niu, B., McLellan, M.D., Uzunangelov, V.*, et al.* (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929-944.
Iglesia, M.D., Vincent, B.G., Parker, J.S., Hoadley, K.A., Carey, L.A., Perou, C.M., and Serody, J.S. (2014). Prognostic B-cell signatures using mRNA-seq in patients with subtype-specific breast and ovarian cancer. Clinical cancer research : an official journal of the American Association for Cancer Research *20*, 3818-3829.
Kim, R., and Guo, J.T. (2009). PDA: an automatic and comprehensive analysis program for protein-DNA complex structures. BMC genomics *10 Suppl 1*, S13.
Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., Darvishi, K.*, et al.* (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature genetics *40*, 1253-1260.
Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology *15*, R29.
Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A.*, et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.
Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.
Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739-1740.
Lou, Y., Yang, X., Wang, F., Cui, Z., and Huang, Y. (2010). MicroRNA-21 promotes the cell proliferation, invasion and migration abilities in ovarian epithelial carcinomas through inhibiting the expression of PTEN protein. International journal of molecular medicine *26*, 819-827.
McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A.*, et al.* (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature genetics *40*, 1166-1174.
Meng, F., Henson, R., Wehbe-Janek, H., Ghoshal, K., Jacob, S.T., and Patel, T. (2007). MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer. Gastroenterology *133*, 647-658.
Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome biology *12*, R41.
Mose, L.E., Wilkerson, M.D., Hayes, D.N., Perou, C.M., and Parker, J.S. (2014). ABRA: improved coding indel detection via assembly-based realignment. Bioinformatics *30*, 2813-2815.
Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics *5*, 557-572.

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z.*, et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *27*, 1160-1167.
Quon, G., Haider, S., Deshwar, A.G., Cui, A., Boutros, P.C., and Morris, Q. (2013). Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome medicine *5*, 29.
Rhie, S.K., Hazelett, D.J., Coetzee, S.G., Yan, C., Noushmehr, H., and Coetzee, G.A. (2014). Nucleosome positioning and histone modifications define relationships between regulatory elements and nearby gene expression in breast epithelial cells. BMC genomics *15*, 331.
Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R.*, et al.* (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389-393.
Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811-1817.
Shaw, R.J., and Cantley, L.C. (2006). Ras, PI(3)K and mTOR signalling controls tumour cell growth. Nature *441*, 424-430.
Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America *98*, 5116-5121.
Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*, i237-245.
Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y.*, et al.* (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome research *22*, 1798-1812.
Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M.*, et al.* (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research *38*, e178.
Weigman, V.J., Chao, H.H., Shabalin, A.A., He, X., Parker, J.S., Nordgard, S.H., Grushko, T., Huo, D., Nwachukwu, C., Nobel, A.*, et al.* (2012). Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. Breast cancer research and treatment *133*, 865-880.
Wilkerson, M.D., Cabanski, C.R., Sun, W., Hoadley, K.A., Walter, V., Mose, L.E., Troester, M.A., Hammerman, P.S., Parker, J.S., Perou, C.M.*, et al.* (2014). Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. Nucleic acids research *42*, e107.
Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572-1573.