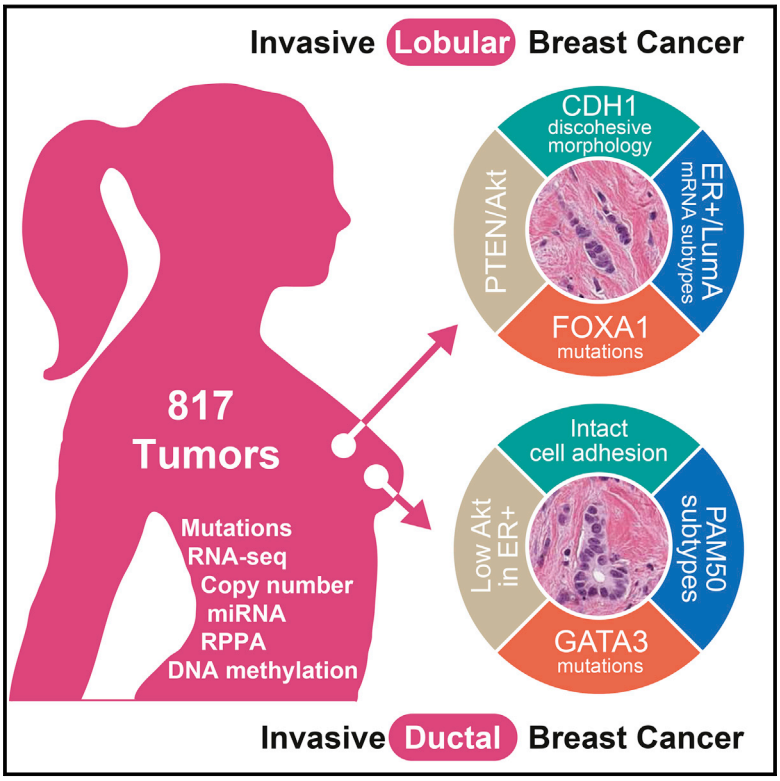
Resource

Cell

Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer

Graphical Abstract



Highlights

- Invasive lobular carcinoma (ILC) is a clinically and molecularly distinct disease
- ILCs show CDH1 and PTEN loss, AKT activation, and mutations in *TBX3* and *FOXA1*
- Proliferation and immune-related gene expression signatures define 3 ILC subtypes
- Genetic features classify mixed tumors into lobular-like and ductal-like subgroups

Authors

Giovanni Ciriello, Michael L. Gatza, Andrew H. Beck, ..., Tari A. King, TCGA Research Network, Charles M. Perou

Correspondence

cperou@med.unc.edu

In Brief

A comprehensive analysis of 817 breast tumor samples determines invasive lobular carcinoma as a molecularly distinct disease with characteristic genetic features, providing key information for patient stratification that may allow a more informed clinical follow-up.

Ciriello et al., 2015, Cell *163*, 506–519
October 8, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.09.033

CellPress

Cell

Resource

Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer

Giovanni Ciriello,[1,2,23] Michael L. Gatza,[3,4,23] Andrew H. Beck,[5] Matthew D. Wilkerson,[6] Suhn K. Rhie,[7] Alessandro Pastore,[2] Hailei Zhang,[8] Michael McLellan,[9] Christina Yau,[10] Cyriac Kandoth,[11] Reanne Bowlby,[12] Hui Shen,[13] Sikander Hayat,[2] Robert Fieldhouse,[2] Susan C. Lester,[5] Gary M.K. Tse,[14] Rachel E. Factor,[15] Laura C. Collins,[5] Kimberly H. Allison,[16] Yunn-Yi Chen,[18] Kristin Jensen,[16,17] Nicole B. Johnson,[5] Steffi Oesterreich,[19] Gordon B. Mills,[20] Andrew D. Cherniack,[8] Gordon Robertson,[12] Christopher Benz,[10] Chris Sander,[2] Peter W. Laird,[13] Katherine A. Hoadley,[3] Tari A. King,[21] TCGA Research Network,[22] and Charles M. Perou[3,*]

[1]Department of Medical Genetics, University of Lausanne (UNIL), 1011 Lausanne, Switzerland
[2]Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA
[3]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA
[4]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08903, USA
[5]Department of Pathology, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, MA, 02215, USA
[6]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA
[7]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, 90033, USA
[8]The Eli and Edythe L. Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA
[9]The Genome Institute, Washington University School of Medicine, MO, 63108, USA
[10]Buck Institute For Research on Aging, Novato, CA, 94945, USA
[11]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA
[12]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, V5Z4S6, Canada
[13]Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI, 49503, USA
[14]Department of Anatomical and Cellular Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong
[15]Department of Pathology, School of Medicine, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA
[16]Department of Pathology, School of Medicine, Stanford University Medical Center, Stanford University, Stanford, CA, USA
[17]VA Palo Alto Healthcare System, Palo Alto, 94304, CA, USA
[18]Department of Pathology and Laboratory Medicine, University of California, San Francisco, CA, 94143, USA
[19]Department of Pharmacology and Chemical Biology, Women's Cancer Research Center, University of Pittsburgh Cancer Institute, Pittsburgh, PA, 15232, USA
[20]MD Anderson Cancer Center, The University of Texas, Houston, TX, 77230, USA
[21]Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA
[22]http://cancergenome.nih.gov/
[23]Co-first author
*Correspondence: cperou@med.unc.edu
http://dx.doi.org/10.1016/j.cell.2015.09.033

SUMMARY

Invasive lobular carcinoma (ILC) is the second most prevalent histologic subtype of invasive breast cancer. Here, we comprehensively profiled 817 breast tumors, including 127 ILC, 490 ductal (IDC), and 88 mixed IDC/ILC. Besides E-cadherin loss, the best known ILC genetic hallmark, we identified mutations targeting PTEN, TBX3, and FOXA1 as ILC enriched features. PTEN loss associated with increased AKT phosphorylation, which was highest in ILC among all breast cancer subtypes. Spatially clustered FOXA1 mutations correlated with increased FOXA1 expression and activity. Conversely, GATA3 mutations and high expression characterized luminal A IDC, suggesting differential modulation of ER activity in ILC and IDC. Proliferation and immune-related signatures determined three ILC transcriptional subtypes associated with survival differences. Mixed IDC/ILC cases were molecularly classified as ILC-like and IDC-like revealing no true hybrid features. This multidimensional molecular atlas sheds new light on the genetic bases of ILC and provides potential clinical options.

INTRODUCTION

Invasive lobular carcinoma (ILC) is the second most frequently diagnosed histologic subtype of invasive breast cancer, constituting ~10%–15% of all cases. The classical form (Foote and Stewart, 1946) is characterized by small discohesive neoplastic cells invading the stroma in a single-file pattern. The discohesive phenotype is due to dysregulation of cell-cell adhesion, primarily driven by lack of E-cadherin (CDH1) protein expression observed in ~90% of ILCs (McCart Reed et al., 2015; Morrogh et al., 2012). This feature is the ILC hallmark, and immunohistochemistry (IHC) scoring for CDH1 expression is often used to discriminate between lesions with borderline ductal versus lobular histological features. ILC variants have also been described, yet all display loss of E-cadherin expression (Dabbs et al., 2013).

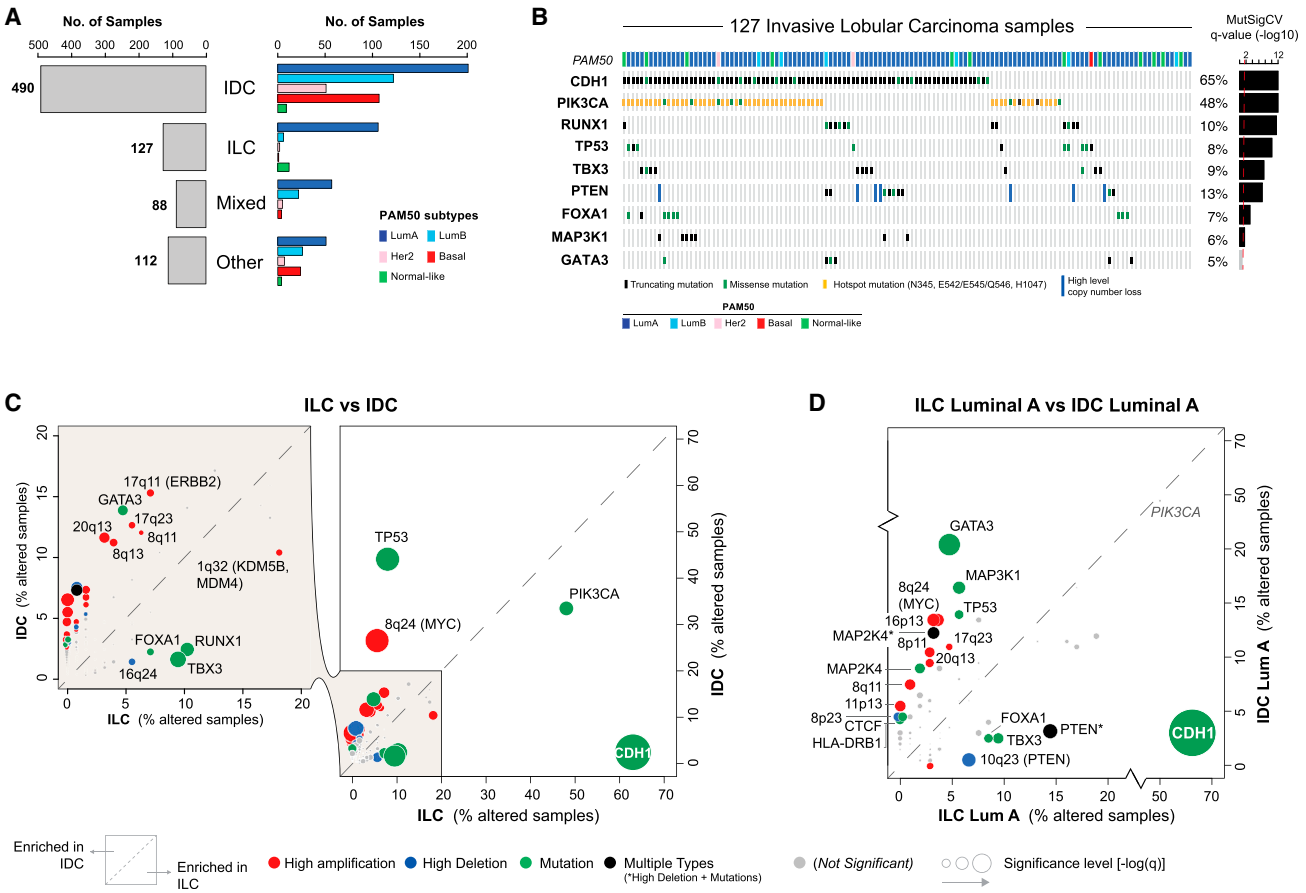


Figure 1. Molecular Determinants of Invasive Lobular Breast Cancer

(A) Histopathological breast cancer subtypes: invasive ductal (IDC), invasive lobular (ILC), mixed ductal/lobular (Mixed), and other-type (Other) carcinoma. PAM50 intrinsic subtypes are not equally distributed across breast cancer subtypes.

(B) Recurrently mutated genes (MutSigCV2) in ILC.

(C) Comparison of the alteration frequency for 153 recurrent genomic alterations in ILC versus IDC.

(D) Comparison of the alteration frequency for 153 recurrent genomic alterations in ILC LumA versus IDC LumA.

Classic ILCs are typically of low histologic grade and low to intermediate mitotic index. They express estrogen and progesterone receptors (ER and PR) and rarely show HER2 protein overexpression or amplification. These features are generally associated with a good prognosis, yet some studies suggest that long-term outcomes of ILC are inferior to stage-matched invasive ductal carcinoma (IDC) (Pestalozzi et al., 2008). Importantly, ILC infiltrative growth pattern complicates both physical exam and mammographic findings and its patterns of metastatic spread often differ from those of IDC (Arpino et al., 2004).

To date, genomic studies of ILC have provided limited insight into the biologic underpinnings of this disease, mostly focusing on mRNA expression and DNA copy-number analysis (McCart Reed et al., 2015). The first TCGA breast cancer study (Cancer Genome Atlas, 2012) reported on 466 breast tumors assayed on six different technology platforms. ILC was represented by only 36 samples, and no lobular-specific features were noted besides mutations and decreased mRNA and protein expression of CDH1. Here, we analyzed nearly twice as many breast tumors from TCGA (n = 817), including 127 ILC. This study identified multiple genomic alterations that discriminate between ILC and IDC demonstrating at the molecular level that ILC is a distinct breast cancer subtype and providing new insight into ILC tumor biology and therapeutic options.

RESULTS

Genetic Determinants of Invasive Lobular Cancer

A total of 817 breast tumor samples were profiled with five different platforms as previously described (Cancer Genome Atlas Research Network, 2014) and 633 cases were also profiled by reverse-phase protein array (RPPA). A pathology committee reviewed and classified all tumors into 490 IDC, 127 ILC, 88 cases with mixed IDC and ILC features, and 112 with other histologies (Table S1). As expected, lobular tumors were predominantly classified as luminal A (LumA) (Figure 1A) and being typically ER+ tumors characterized by low levels of proliferation markers (Table S1). ER status was clinically determined by

immunohistochemistry on 120 of 127 ILC cases, with 94% (n = 113) scoring positively.

Within 127 ILC, we identified 8,173 total coding mutations, integrating information from both DNA and RNA sequencing (Wilkerson et al., 2014). Recurrently mutated genes in ILC were identified by MutSigCV2 (Lawrence et al., 2013) and included many genes previously implicated in breast cancer (Figure 1B, Table 1) (Cancer Genome Atlas, 2012). Similarly, recurrent copy-number alterations in ILC estimated by GISTIC (Mermel et al., 2011) recapitulated known breast cancer gains and losses, in particular those observed in ER+/luminal tumors (Figure S1A). However, the frequency of these alterations (both mutations and copy-number changes) often differed significantly between IDC and ILC.

To investigate these differences, we identified recurrent alterations across all 817 samples and separately in ILC and IDC PAM50 subtypes (luminal A, n = 201, luminal B, n = 122, HER2-enriched, n = 51, and basal-like, n = 107). In total, we identified 178 events, including 68 mutated genes, 47 regions of gain, and 63 regions of loss (Table 1 and Table S2). Several of these had different incidence in ILC and in IDC (Figure 1C, Table S3). ILC cases were significantly enriched for *CDH1* mutations (63% in ILC versus 2% in IDC, q = 3.94E-53), most of them truncating, and mutations affecting *TBX3* (9% versus 2%, q = 0.003), *RUNX1* (10% versus 3%, q = 0.008), *PIK3CA* (48% versus 33%, q = 0.02), and *FOXA1* (7% versus 2%, q = 0.08). By contrast, alterations typically observed in ER−/basal-like tumors were less frequent in ILC, including *TP53* mutations (8% in ILC versus 44% in IDC, q = 1.9E-14) and focal amplification of *MYC* (6% versus 27%, q = 7.42E-7) and *CCNE1* (0% versus 7%, q = 0.01). These results partly reflect genetic differences between ER+/luminal and ER−/basal-like breast cancer, given that ILC tumors were predominantly LumA. Nonetheless, unexpected differences did emerge including a lower incidence of *GATA3* mutations in ILC compared to IDC (5% in ILC versus 13% in IDC, q = 0.03) (Figure 1C).

To better identify ILC discriminatory features, we limited our analyses to LumA samples, representing 41% of IDC (n = 201) and 83% of ILC (n = 106) (Figure 1D). This analysis confirmed a high incidence of *CDH1* (q = 1.4E-30), *TBX3* (q = 0.05), and *FOXA1* (q = 0.065) mutations in ILC, while the frequency of *RUNX1* and *PIK3CA* mutations was no longer significantly different. *GATA3* mutations (5% ILC versus 20% IDC, q = 0.003) were the second most discriminant event after *CDH1* mutations, mostly affecting IDC tumors. Interestingly, both *FOXA1* and *GATA3* are key regulators of ER activity (Liu et al., 2014), suggesting IDC and ILC may preferentially rely on different mechanisms to mediate the ER transcriptional program. Finally, homozygous losses of the *PTEN* locus (10q23) were more frequent in ILC (q = 0.035) as were *PTEN* mutations (8% versus 3%). Collectively, *PTEN* inactivating alterations were identified in 14% of LumA ILC versus 3% of LumA IDC (p = 9E-4), making this the third most discriminant feature between LumA IDC and LumA ILC (Figure 1D).

E-Cadherin Loss in Invasive Lobular Carcinoma

Loss of the epithelial specific cell-cell adhesion molecule E-cadherin (CDH1) is the key hallmark of ILC (Dabbs et al., 2013; Moll et al., 1993). *CDH1* loss is believed to confer the highly discohesive morphology characteristic of this tumor subtype and is often associated with tumor invasion and metastasis in other tumor types, including diffuse gastric cancer (Brinck et al., 2004; Cancer Genome Atlas Research Network, 2014; Richards et al., 1999). Loss-of-function mutations targeting *CDH1* are present in 50%–60% of ILC and are believed to be an early event often observed in matching lobular carcinoma in situ (LCIS) (McCart Reed et al., 2015). *CDH1* mutations typically occur in combination with chromosome 16q loss, where *CDH1* is located, thus inducing complete loss of the protein.

We identified 108 mutations in the coding sequence of *CDH1* in 107/817 patients (63%); 80 of these occurred in ILC cases. These mutations were rather uniformly distributed along the coding sequence, and 83% of them were predicted to be truncating (Figure 2A). *CDH1* mutations almost invariably co-occurred with heterozygous loss of 16q (affecting 89% of ILC cases) and were associated with downregulation of both CDH1 transcript and protein levels (Figures 2B and S2A). By combining somatic mutations, copy-number losses, and mRNA and low protein expression (the latter when available), we identified E-cadherin alterations in 120/127 (95%) cases with DNA and RNA data, and in all 79 cases with DNA, RNA, and protein data (Figures S2A–SC).

Previous studies reported sporadic cases of multiple cancer types with high DNA methylation levels at the *CDH1* promoter, suggesting epigenetic silencing as an alternative mechanism for downregulation of CDH1 (Graff et al., 1997; Richards et al., 1999; Sarrió et al., 2003; Zou et al., 2009). We analyzed the DNA methylation levels in breast tumors at CpG sites spanning from upstream of the *CDH1* promoter, across the promoter CpG island, and extending into the first intron (Figure S2D). Despite four of these probes matching DNA positions previously reported as methylated in ILC (Graff et al., 1997; Sarrió et al., 2003; Zou et al., 2009), we did not detect significant DNA hyper-methylation at these probes (Figure S2E), nor in any of the other *CDH1* associated probes analyzed (Figures 2B and S2F and S2G). Moderately increased methylation was observed in a few cases near exon 2; however, DNA methylation at this site correlated with lower tumor purity and increased leukocyte infiltration and indeed it mimicked methylation levels at this CpG site in normal leukocytes (Figures S2F and S2G). Infinium DNA methylation results were validated by whole-genome bisulfite sequencing in five samples (Figure S2H). Altogether, these data confirm that *CDH1* expression was substantially lower in ILC than in IDC and that this expression difference did not appear associated with DNA methylation at the *CDH1* promoter. Our results on 817 invasive breast tumors thus confirmed E-cadherin loss as ILC defining molecular feature but do not support the reported occurrence of *CDH1* epigenetic silencing in invasive breast cancer. The discrepancy with prior literature may be attributable in part to the reliance on highly sensitive, but non-quantitative, methylation-specific PCR assays in past studies (Herman et al., 1996) and will require further investigation.

FOXA1 Mutations in Breast Cancer

FOXA1 is a key ER transcriptional modulator (Carroll et al., 2005; Hurtado et al., 2011) coordinating ER DNA binding within a large

Table 1. Recurrently Mutated Genes in Breast Cancer

	ILC (n = 127)		ILC Luminal A (106)		IDC (490)		IDC Luminal A (201)		IDC Luminal B (122)		IDC Her2-enriched (51)		IDC Basal-like (107)		ALL Breast Cancer (817)	
Gene	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*
PIK3CA	61	1.02E-12	54	9.18E-13	164	6.09E-13	93	6.79E-13	43	6.76E-13	19	9.14E-13	7	6.22E-02	282	2.54E-13
RUNX1	13	1.02E-12	9	9.18E-13	13	n.s.	9	1.32E-05	3	n.s.	0	n.s.	1	n.s.	32	2.54E-13
CDH1	80	3.40E-12	68	6.12E-12	10	7.63E-03	7	5.33E-02	2	n.s.	0	n.s.	0	n.s.	107	2.54E-13
TP53	10	8.26E-11	6	2.22E-04	215	6.09E-13	28	6.79E-13	52	6.76E-13	37	9.14E-13	92	1.83E-12	280	2.54E-13
TBX3	12	2.54E-08	10	4.01E-06	8	n.s.	5	n.s.	2	n.s.	0	n.s.	1	n.s.	26	1.11E-08
PTEN	9	8.43E-08	8	8.86E-09	27	5.61E-11	6	5.63E-03	11	9.64E-12	4	3.21E-02	6	4.65E-03	42	2.54E-13
FOXA1	9	5.52E-04	9	6.53E-04	11	n.s.	5	n.s.	3	n.s.	2	n.s.	1	n.s.	30	4.52E-13
MAP3K1	7	2.95E-02	6	7.54E-02	40	4.06E-12	33	1.25E-11	2	n.s.	1	n.s.	4	n.s.	69	2.54E-13
GATA3	6	n.s.	5	n.s.	66	6.09E-13	40	6.79E-13	22	6.76E-13	3	n.s.	0	n.s.	96	2.54E-13
AKT1	3	n.s.	3	n.s.	15	5.61E-11	11	1.25E-11	3	5.91E-02	1	n.s.	0	n.s.	20	2.54E-13
NBL1	3	n.s.	2	n.s.	10	1.08E-10	8	2.04E-12	0	n.s.	1	n.s.	1	n.s.	16	5.24E-11
KMT2C	9	n.s.	8	n.s.	37	1.49E-08	17	2.94E-02	12	n.s.	3	n.s.	5	n.s.	64	4.89E-06
DCTD	0	n.s.	0	n.s.	6	1.02E-05	3	1.54E-02	1	n.s.	1	n.s.	0	n.s.	6	7.61E-04
RB1	0	n.s.	0	n.s.	16	1.46E-04	4	n.s.	7	n.s.	1	n.s.	4	n.s.	18	n.s.
SF3B1	4	n.s.	4	n.s.	12	3.20E-04	6	1.12E-03	4	n.s.	1	n.s.	1	n.s.	16	3.68E-04
CBFB	2	n.s.	2	n.s.	15	1.51E-03	13	5.68E-06	1	n.s.	1	n.s.	0	n.s.	24	8.14E-13
ARHGAP35	1	n.s.	1	n.s.	13	1.62E-03	5	n.s.	5	1.24E-02	0	n.s.	3	n.s.	18	7.74E-03
OR9A2	0	n.s.	0	n.s.	5	1.77E-03	2	n.s.	1	n.s.	1	n.s.	1	n.s.	5	6.47E-03
NCOA3	6	n.s.	4	n.s.	24	1.77E-03	7	n.s.	7	n.s.	4	n.s.	6	n.s.	40	3.25E-07
RBMX	2	n.s.	2	n.s.	10	2.83E-03	3	n.s.	2	n.s.	1	n.s.	4	6.27E-02	12	4.01E-08
MAP2K4	2	n.s.	2	n.s.	24	2.83E-03	18	5.29E-12	5	n.s.	1	n.s.	0	n.s.	30	1.37E-05
TROVE2	0	n.s.	0	n.s.	6	4.51E-03	1	n.s.	2	n.s.	1	n.s.	2	n.s.	8	2.77E-03
NADK	0	n.s.	0	n.s.	4	4.51E-03	0	n.s.	4	5.85E-05	0	n.s.	0	n.s.	6	3.61E-03
CASP8	1	n.s.	1	n.s.	9	6.00E-03	2	n.s.	3	n.s.	0	n.s.	3	n.s.	11	1.81E-03
CTSS	0	n.s.	0	n.s.	5	6.00E-03	1	n.s.	2	n.s.	0	n.s.	2	n.s.	5	8.91E-02
ACTL6B	2	n.s.	1	n.s.	5	7.63E-03	2	n.s.	1	n.s.	0	n.s.	2	n.s.	10	7.33E-05
LGALS1	0	n.s.	0	n.s.	4	9.78E-03	2	n.s.	2	n.s.	0	n.s.	0	n.s.	5	6.34E-03
KRAS	2	n.s.	1	n.s.	4	1.54E-02	3	7.32E-03	0	n.s.	0	n.s.	1	n.s.	7	2.49E-04
KCNN3	2	n.s.	2	n.s.	8	1.81E-02	1	n.s.	2	n.s.	2	n.s.	3	2.45E-02	16	4.53E-02
FBXW7	2	n.s.	2	n.s.	6	2.19E-02	0	n.s.	0	n.s.	0	n.s.	6	8.28E-04	11	n.s.
LRIG2	0	n.s.	0	n.s.	4	3.08E-02	2	n.s.	0	n.s.	1	n.s.	1	n.s.	6	n.s.
PIK3R1	0	n.s.	0	n.s.	9	3.08E-02	2	n.s.	3	n.s.	2	n.s.	2	n.s.	13	1.56E-03
PARP4	3	n.s.	3	n.s.	7	3.08E-02	3	n.s.	4	n.s.	0	n.s.	0	n.s.	12	n.s.
ZNF28	3	n.s.	3	n.s.	7	3.25E-02	1	n.s.	5	n.s.	0	n.s.	1	n.s.	11	1.72E-02
HLA-DRB1	0	n.s.	0	n.s.	13	3.52E-02	9	1.49E-02	2	n.s.	0	n.s.	2	n.s.	16	n.s.
ERBB2	5	n.s.	4	n.s.	7	6.42E-02	3	n.s.	1	n.s.	2	n.s.	1	n.s.	18	3.36E-06
ZMYM3	0	n.s.	0	n.s.	9	8.83E-02	3	n.s.	1	n.s.	1	n.s.	4	n.s.	11	n.s.
RAB42	1	n.s.	1	n.s.	2	n.s.	0	n.s.	0	n.s.	0	n.s.	2	6.27E-02	4	1.82E-03
CTCF	0	n.s.	0	n.s.	12	n.s.	9	7.05E-08	1	n.s.	1	n.s.	1	n.s.	18	1.93E-03
ATAD2	0	n.s.	0	n.s.	9	n.s.	2	n.s.	4	7.32E-02	2	n.s.	1	n.s.	12	n.s.
CDKN1B	3	n.s.	2	n.s.	5	n.s.	4	9.59E-02	1	n.s.	0	n.s.	0	n.s.	11	1.14E-03
GRIA2	0	n.s.	0	n.s.	6	n.s.	5	5.33E-02	0	n.s.	0	n.s.	1	n.s.	6	n.s.
NCOR1	8	n.s.	8	n.s.	23	n.s.	12	4.81E-03	7	n.s.	1	n.s.	3	n.s.	39	3.61E-03
HRNR	4	n.s.	4	n.s.	13	n.s.	3	n.s.	3	n.s.	3	n.s.	4	n.s.	23	7.65E-02
GPRIN2	1	n.s.	1	n.s.	6	n.s.	3	n.s.	1	n.s.	0	n.s.	2	n.s.	11	1.16E-05
PAX2	1	n.s.	0	n.s.	2	n.s.	2	n.s.	0	n.s.	0	n.s.	0	n.s.	4	4.80E-02

(*Continued on next page*)

Table 1. *Continued*

	ILC (n = 127)		ILC Luminal A (106)		IDC (490)		IDC Luminal A (201)		IDC Luminal B (122)		IDC Her2-enriched (51)		IDC Basal-like (107)		ALL Breast Cancer (817)	
Gene	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*	*n*	*q*
ACTG1	1	n.s.	1	n.s.	4	n.s.	2	n.s.	0	n.s.	0	n.s.	2	n.s.	8	9.39E-02
AQP12A	0	n.s.	0	n.s.	3	n.s.	1	n.s.	1	n.s.	0	n.s.	1	n.s.	5	2.69E-02
PIK3C3	2	n.s.	2	n.s.	5	n.s.	2	n.s.	0	n.s.	1	n.s.	2	n.s.	11	3.23E-02
MYB	1	n.s.	1	n.s.	7	n.s.	3	n.s.	2	n.s.	0	n.s.	2	n.s.	12	8.91E-02
IRS4	1	n.s.	1	n.s.	6	n.s.	3	n.s.	0	n.s.	0	n.s.	3	n.s.	8	9.38E-02
TBL1XR1	3	n.s.	2	n.s.	3	n.s.	1	n.s.	1	n.s.	1	n.s.	0	n.s.	12	4.71E-04
RPGR	4	n.s.	3	n.s.	11	n.s.	3	n.s.	3	n.s.	2	n.s.	3	n.s.	19	1.26E-03
CCNI	1	n.s.	1	n.s.	2	n.s.	0	n.s.	2	n.s.	0	n.s.	0	n.s.	3	6.93E-02
ARID1A	7	n.s.	5	n.s.	16	n.s.	7	n.s.	4	n.s.	3	n.s.	2	n.s.	33	7.91E-09
CD3EAP	1	n.s.	0	n.s.	2	n.s.	0	n.s.	0	n.s.	0	n.s.	2	n.s.	5	1.29E-02
ADAMTS6	1	n.s.	1	n.s.	3	n.s.	1	n.s.	0	n.s.	0	n.s.	2	n.s.	8	1.81E-03
OR2D2	0	n.s.	0	n.s.	4	n.s.	0	n.s.	3	n.s.	0	n.s.	1	n.s.	5	5.67E-02
TMEM199	0	n.s.	0	n.s.	3	n.s.	0	n.s.	2	n.s.	1	n.s.	0	n.s.	4	3.36E-02
MST1	0	n.s.	0	n.s.	5	n.s.	2	n.s.	2	n.s.	0	n.s.	1	n.s.	7	9.46E-02
RHBG	0	n.s.	0	n.s.	3	n.s.	0	n.s.	0	n.s.	1	n.s.	2	n.s.	4	7.91E-02
ZFP36L1	1	n.s.	1	n.s.	5	n.s.	2	n.s.	2	n.s.	0	n.s.	1	n.s.	8	3.37E-02
TCP11	2	n.s.	0	n.s.	3	n.s.	2	n.s.	0	n.s.	0	n.s.	1	n.s.	6	4.80E-02
CASZ1	4	n.s.	4	n.s.	3	n.s.	0	n.s.	0	n.s.	1	n.s.	2	n.s.	11	2.03E-02
GAL3ST1	1	n.s.	1	n.s.	2	n.s.	0	n.s.	1	n.s.	0	n.s.	1	n.s.	4	7.74E-03
FRMPD2	1	n.s.	1	n.s.	7	n.s.	2	n.s.	4	n.s.	0	n.s.	1	n.s.	9	8.91E-02
GPS2	1	n.s.	1	n.s.	4	n.s.	3	n.s.	0	n.s.	1	n.s.	0	n.s.	8	8.91E-02
ZNF362	0	n.s.	0	n.s.	3	n.s.	3	n.s.	0	n.s.	0	n.s.	0	n.s.	3	8.91E-02

n: number of mutations ***q***: MutSigCV2 q value.

protein complex by modifying chromatin accessibility and mediating long-range DNA interactions (Liu et al., 2014). High FOXA1 expression has been previously reported in breast and prostate cancer (Habashy et al., 2008; Sahu et al., 2011) and somatic mutations in the *FOXA1* gene have been reported in these tumor types in about 3%–4% of the cases (Barbieri et al., 2012; Cancer Genome Atlas, 2012; Robinson et al., 2015).

Here, we observed a total of 33 *FOXA1* mutations in 30/817 (3.7%) tumors (Figure 3A), and the large sample set allowed us to identify regional hotspots in the *FOXA1* mutation distribution. Mutations clustered in the fork-head DNA binding (FK) and C terminus transactivation domains (Figure 3A). A similar mutational pattern was observed by combining multiple prostate cancer sequencing studies (Baca et al., 2013; Barbieri et al., 2012; Grasso et al., 2012; Robinson et al., 2015) (Figure S3A), and confirmed by the TCGA prostate cancer project (Robinson et al., 2015). Thus, regional *FOXA1* mutation hotspots are selectively altered in a tissue-independent fashion.

Eleven *FOXA1* mutations were observed in 9/127 (7%) ILC cases. All *FOXA1* mutations in ILC were in the FK domain, whereas mutations in IDC (n = 11) were observed both in FK (n = 6) and other structural elements (n = 5), without a specific preference. The FK domain includes three α helices (H1, H2, H3), three β strands (S1, S2, S3) and two loops, typically referred to as "wings" (W1, W2) (Figures 3B and S3B). *FOXA1* mutations in FK clustered prevalently in the W2 loop. Notable exceptions were recurrently mutated residues I176 (n = 4) and D226 (n = 3). These residues are far from W2 in sequence space and located in different secondary structure elements; however, they are close (within 5 to 10 Å) to residues in W2 in the 3D space (Figures 3C, 3D and S3C). In total, 22 out of 25 FK-mutations in our dataset fall into a restricted 3D space or "mutation structural hotspot" (MSH) (Figure 3D) indicating a selective pressure for targeting protein interactions and functions mediated by this region. Notably, 8/127 ILC cases have *FOXA1* mutations within this MSH compared to 4/490 IDC cases (p = 6E-4), further supporting *FOXA1* selected mutations as an ILC feature.

FOXA1 DNA binding occurs mostly through helix H3, that recognizes the binding motif and is stabilized by interactions mediated by its "wings" (Cirillo and Zaret, 2007; Gajiwala and Burley, 2000; Kohler and Cirillo, 2010). Only a few residues predicted or experimentally shown to interact with the DNA were mutated (Figures 3B–3E) suggesting that these events are unlikely to affect FOXA1 DNA binding. FOXA1 is a pioneer factor that binds condensed chromatin and triggers DNA demethylation of its binding sites, making them accessible to transcription factors such as ER (Cirillo et al., 2002; Sérandour et al., 2011). FOXA1 activity can therefore be estimated by the methylation status of these sites where occupied FOXA1 DNA binding sites tend to be demethylated. We analyzed DNA methylation levels of the

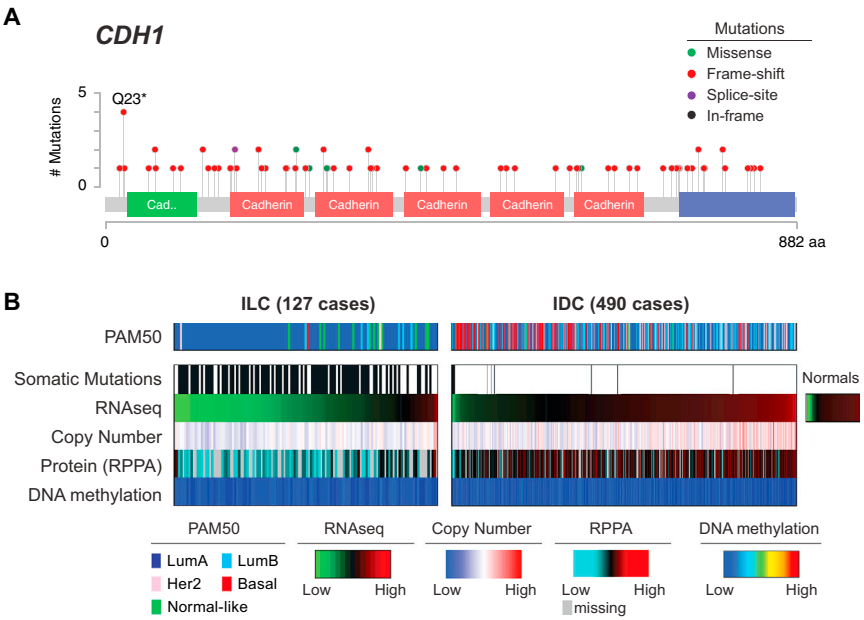


Figure 2. E-Cadherin Loss in ILC

(A) Mutations targeting the CDH1 gene target residues across the whole-sequence and are mostly predicted to be truncating (red).

(B) Comparison of E-cadherin status between ILC and IDC reveals frequent hemizygous copy-number losses at the CDH1 locus and downregulation of both mRNA [$\log_2$(RSEM)] and protein levels. See also Figures S2A–S2C. Average DNA methylation level of 6 probes at the CDH1 promoter shows no change in DNA methylation in both ILC and IDC samples. See also Figures S2D–S2I.

3,976 most variable methylation probes mapping to FOXA1 binding sites (Table S3) and methylated in normal samples (Ross-Innes et al., 2012; Wang et al., 2012). DNA methylation of these sites was substantially lower when *FOXA1* and *ESR1* were highly expressed, while it remained high in *FOXA1*-negative cases and adjacent normal tissue (Figure 3F). Inverse correlation with *FOXA1* expression (Pearson's coefficient $\rho = -0.54$) was specifically observed for DNA methylation at FOXA1 binding sites. Indeed, no correlation was found with methylation at 2,000 most variable probes with the same methylation level in normal samples as FOXA1 binding sites ($\rho = 0.07$) (Figure 3F). These data support the hypothesis that *FOXA1* mRNA expression correlates with its activity. *FOXA1* mutations were positively associated with its mRNA expression ($p = 0.002$) and maintained a similar anti-correlation with DNA methylation at FOXA1 binding sites (Figure 3F). Finally, by examining mRNA expression of *FOXA1* targets, defined as genes with a FOXA1 binding motif in the promoter or matching the genomic loci covered by the 3,976 methylation probes we analyzed (Table S3), no significant differences were identified and only a few genes showed moderate expression changes (Table S4). These data collectively indicate that *FOXA1* mutations do not abolish protein function and, in fact, they may activate alternative mechanisms to affect ER transcriptional programs.

Differential expression analyses between *FOXA1* mutant and wild-type cases within distinct subsets of samples found consistent upregulation of neuroendocrine secretory proteins *SCG1* (*CHGB*) and *SCG2*, chemokine-like factor *CMTM8*, neuroendocrine tumor associated transcription factor *NKX2-2* and Kallikrein serine proteases *KLK12*, *KLK13*, and *KLK14* (Figure S3D and Table S5). While the relatively low number of *FOXA1* mutations and breast cancer heterogeneity prevented the identification of strong transcriptional signals associated with *FOXA1* mutations, several upregulated targets in *FOXA1* mutant cases with part of them consistently found as significant suggests these lesions might drive novel binding events.

Interestingly, while ILC cases were enriched for *FOXA1* mutations, and in particular for those targeting the FK domain, ILC showed significantly fewer *GATA3* mutations, another key ER modulator. Mutations in *GATA3* were more frequent in LumA IDC (Figure 1D) and mutually exclusive with *FOXA1* events. Moreover, LumA ILC tumors show lower GATA3 mRNA ($p = 0.007$) and protein ($p = 2E\text{-}4$) levels than LumA IDC (Figures S3E and S3F). Taken together, the differential expression patterns and enrichment for hotspot mutations of GATA3 in IDC and of *FOXA1* in ILC, suggest a preferential requirement for distinct ER modulators in ILC and IDC.

Akt Signaling Is Strongly Activated in ILC

PTEN inactivation emerged as a discriminant feature between luminal A ILC and luminal A IDC. *PTEN* genetic alterations across all ILC cases included homozygous deletions (6%) and somatic mutations (7%), and were largely mutually exclusive with *PIK3CA* mutations (48%) (Figure S4A).

Unbiased differential protein expression analysis (Table S6) based on RPPA data revealed significant lower PTEN protein expression ($p = 4E\text{-}4$) in LumA ILC compared to LumA IDC (Figure 4A). Consistent with PTEN function as a negative regulator of Akt activity (Cantley and Neel, 1999; Song et al., 2012), ILC tumors also showed significantly increased Akt phosphorylation at both S473 ($p = 0.004$) and T308 ($p = 7E\text{-}5$) (Figure 4A). Upstream of the Akt pathway, we found significant upregulation of total EGFR ($p = 1E\text{-}4$) and phospho-EGFR at Y1068 ($p = 0.005$) and Y1173 ($p = 0.007$), as well as phospho-STAT3 at Y705 ($p = 7E\text{-}4$), supporting upregulation of signaling axes converging on Akt activation (Wu et al., 2013). We also identified increased phospho-p27 at T157 ($p = 0.002$), an Akt substrate, and phospho-p70S6 kinase at T389 ($p = 1E\text{-}4$), a direct mTOR target. Notably, ILC phospho-Akt levels were comparable with those typically observed in the more aggressive HER2+ and ER−/basal-like breast tumors (Figure S4B), which have uniformly high levels of PI3K/Akt signaling (Cancer Genome Atlas, 2012). Consistent with these results, we found significant upregulation of a PI3K/Akt pathway-specific protein and phospho-protein expression signature (Akbani et al., 2014) in LumA ILC compared

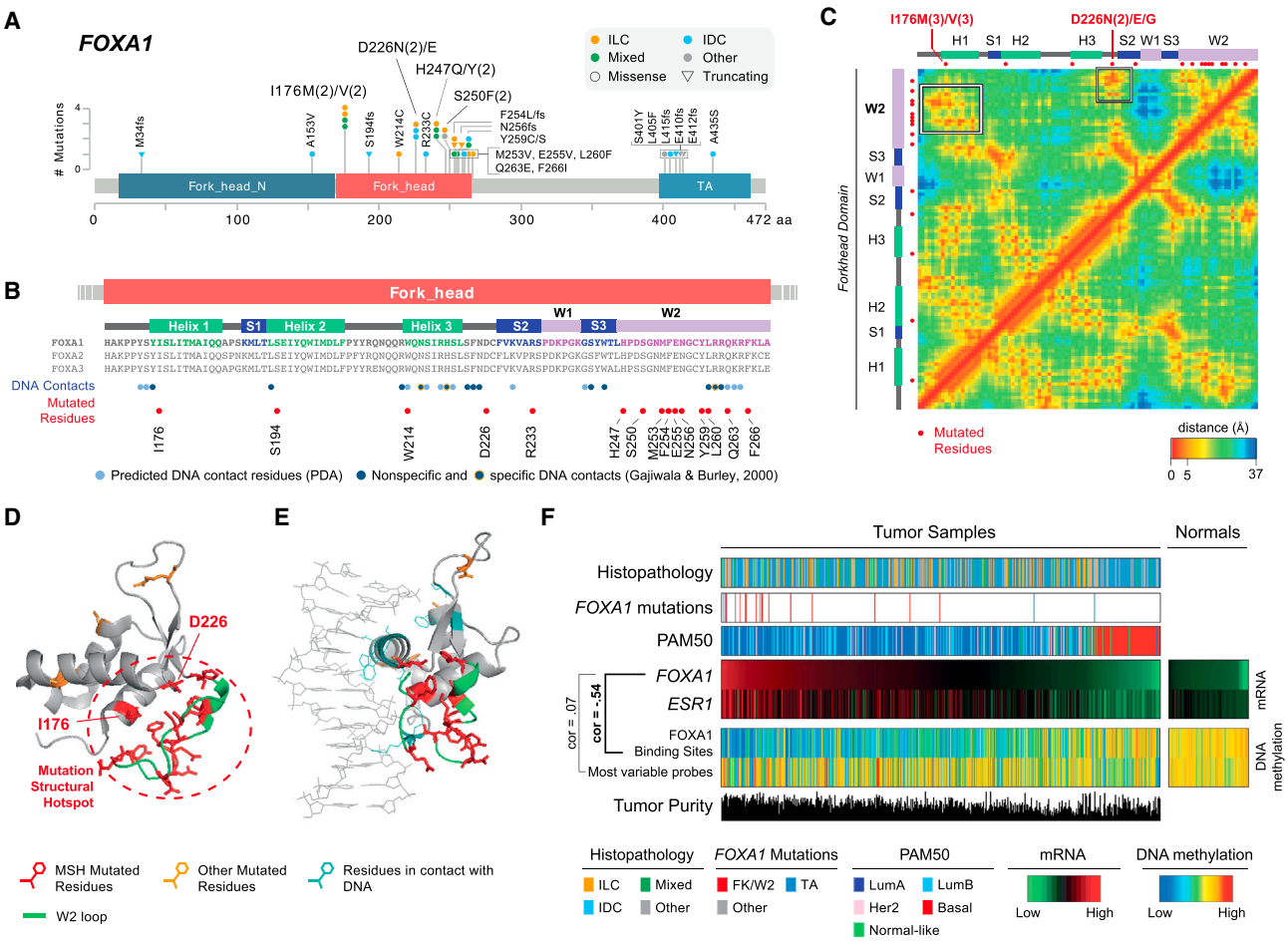


Figure 3. Recurrent FOXA1 Mutations Cluster in the 3D Space and Correlate with High FOXA1 Activity

(A) Recurrent FOXA1 mutations in 817 breast tumors cluster in the Fork-head DNA binding (FK) domain and in the C terminus trans-activation (TA) domain.

(B) Secondary structure elements of the FK domain are not equally mutated. FOXA1 mutations cluster in the W2 loop and rarely target residues interacting with the DNA.

(C) Residue-residue minimum distances for all residues in the FK domain using the 3D structure of FOXA3 FK domain (PDB ID: 1VTN). Frequently mutated residues I176 and D226 are close in the 3D space (but not in the sequence) to the residues in the W2 loop. See also Figure S3C.

(D) 3D structure of the FK domain. Mutations in the W2 loop, in I176, and in D226 form a mutational structural hotspot (MSH).

(E) 3D structure of FK domain bound to the DNA molecule shows mutated residues (red) are not those in contact with the DNA (light blue).

(F) Across all breast cancer subtypes (histopathology and PAM50), FOXA1 mutations are associated with FOXA1 high mRNA expression. FOXA1 mRNA expression is highly correlated with ER mRNA expression [$\log_2$(RSEM)] and anti-correlated with DNA methylation at FOXA1 binding sites consistent with FOXA1 activity. DNA methylation of randomly selected probes was used as control.

to LumA IDC (Figure 4B, Tables S1 and S6). Based on this signature, we found nearly equivalent levels of PI3K/Akt signaling in LumA ILC and basal-like and HER2+ IDC (Figure S4C). Finally, PARADIGM analyses (Vaske et al., 2010) showed increased activation of Akt signaling in LumA ILC relative to LumA IDC (Figure 4D).

PTEN protein loss and increased Akt phosphorylation were observed in association with *PTEN* genetic alterations, as well as in multiple ILC *PTEN* wild-type cases indicating that additional mechanisms contribute to the activation of the pathway. While *PIK3CA* mutations were frequent in LumA ILC tumors, these mutations were not associated with increased levels of phospho-Akt or pathway activity in our dataset. Using MEMo (Ciriello et al., 2012), we highlighted multiple genetic alterations converging on Akt/mTOR signaling in 45% of the samples (Figure S4D, Table S7). Among these, alterations acting upstream of Akt were identified in 40% of ILC cases and were associated with increased Akt phosphorylation and PI3K/Akt score (Figure 4E), providing an apparent molecular explanation for Akt activation in these samples. Interestingly, these events included *ERBB2* amplification and mutations (Figure 4E), both of which have been identified in relapsed ILC (Ross et al., 2013).

ILC mRNA Subtypes

Using mRNA-seq expression data from LumA ILC samples (n = 106), we identified three ILC subtypes termed *reactive-like*,

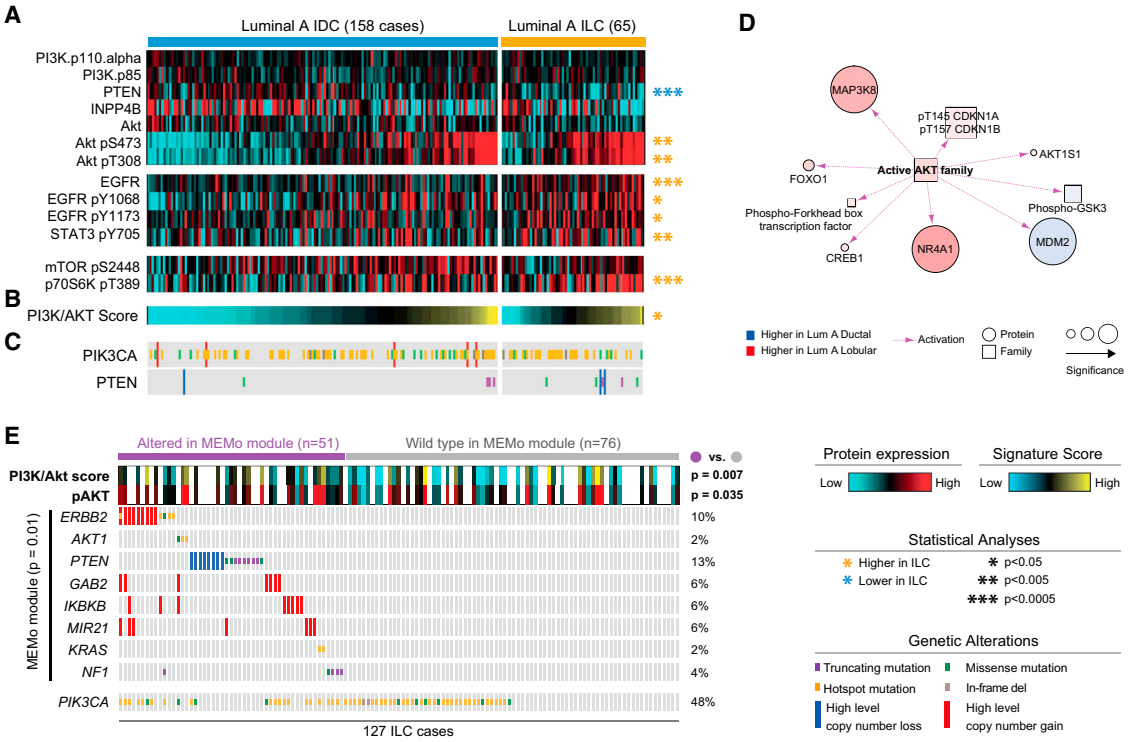


Figure 4. Akt Signaling Is Highest in ILC Tumors

(A) Differential protein and phospho-protein analysis between ILC LumA and IDC LumA reveals significant lower levels of PTEN, and higher levels of Akt, phospho-Akt, EGFR, phospho-EGFR, phopsho-STAT3, and phospho-p70S6K in ILC LumA.

(B) A PI3K/Akt protein expression signature is significantly upregulated in ILC tumors. See also Figures S4B–S4C.

(C) Mutation and copy-number alterations in *PIK3CA* and *PTEN*

(D) PARADIGM identifies increased Akt activity in LumA ILC tumors.

(E) MEMo identified multiple mutually exclusive alterations in ILC converging on Akt signaling and associated with increased phospho-Akt and PI3K/Akt protein signature in these tumors. Hotspot are defined as follow: *PIK3CA* E542, E545, Q546, and H1047; *ERBB2* L755, I767, V777; *AKT1* E17; *KRAS* G12.

immune-related, and *proliferative* (Figures S5A–S5I, Table S8). We then used a 3-class ILC subtype classifier (60 genes, Table S13) to score all ILC samples in the TCGA (n = 127) (Figure 5A) and METABRIC (Curtis et al., 2012) datasets (Table S12). Our analyses identified many significant genomic features that distinguished each ILC subtype at the mRNA and protein/phospho-protein level; but no distinguishing somatic mutations or DNA copy-number alterations.

Significant analysis of microarray (SAM) analysis (Tusher et al., 2001) identified 1,277 genes differentially expressed between ILC subtypes (q = 0) (Figure 5A, Table S8). Of these, 1,005 were highly expressed in *reactive-like* tumors, which had lower tumor purity as determined by ABSOLUTE (Carter et al., 2012) (Figures 5A and S5P), and included genes consistent with epithelial and stromal-associated signaling including keratin, kallikrein, and claudin genes as well as the oncogenes *EGFR*, *MET*, *PDGFRA*, and *KIT* (Table S8). The remaining 272 genes were highly expressed in *immune-related* tumors and include modulators of immunogenic signaling such as interleukins (IL), chemokine receptors and ligands, major histocompatibility complex, and tumor necrosis factors, as well as *IDO1* and *IFNG* (Figure 5A and Table S8). Interestingly, immune activity in this subset of tumors appears to be predominantly associated with macrophage-associated signaling as increased levels of CD68 ($p < 0.05$), macrophage-associated colony stimulating factor (MacCSF), macrophage-associated TH1 (MacTH1), and T cell receptor (TCR) gene expression signatures (Iglesia et al., 2014) were observed in both the TCGA (Figure 5A) and METABRIC (Figure S5J–S5K) datasets. Finally, *proliferative* tumors were defined by low expression of each of these 1,277 genes (Figure 5A). Intriguingly, in each dataset (Figures 5 and S5K) *proliferative* tumors had higher levels of proliferation relative to *reactive-like* tumors (TCGA: p = 3.3E-09; METABRIC: p = 0.018) and slightly higher or equivalent levels compared to *immune-related* ones (TGCA: p = 0.29; METABRIC: p = 0.008). Regardless of ILC subtype, ILC tumor proliferation was generally lower than all IDC subtypes (Figures S5L–S5M).

With respect to previously reported RPPA-based subtypes (reactive or non-reactive), *reactive-like* ILC largely, but not entirely, comprised tumors classified as reactive (Figure 5B; $p < 1E-4$), a subgroup characterized by strong microenvironment and/or cancer fibroblast signaling (Cancer Genome Atlas, 2012). Examining protein and phospho-protein expression differences between ILC subtypes identified many significant features (Figure 5B and Table S6). Reactive-like tumors had higher levels of c-Kit (p = 4E-4), consistent with mRNA expression, total

(p = 0.004) and phosphorylated PKC alpha (S657, p = 0.002); beta catenin (p = 0.012) and E-cadherin (p = 0.011), although both beta-catenin and E-cadherin levels are significantly lower than in all IDC subtypes. Decreased levels were observed instead for p70S6 kinase (p = 0.017), Raptor (p = 0.027) and eIF4G (p = 0.024).

Immune-related tumors had higher levels of immune modulator STAT5 alpha (p = 0.019), PI3K/Akt targets phospho-PRAS40 (T246, p = 0.016) and mTOR (S2448, p = 0.019), and total (p = 0.004) and phospho-MEK1 (S217-S221, p = 0.022). Consistent with the mRNA proliferation signature, tumors in the *proliferative* subtype have increased expression of cell-cycle proteins cyclin E1 (p = 0.036), FoxM1 (p = 0.019), PCNA (p = 0.019), and propsho-Chk1 (S345, p = 0.038) as well as DNA repair components Rad50 (p = 0.007), Rad51 (p = 0.007), XRCC1 (p = 0.028), and BRCA2 (p = 0.038). Decreased expression was observed for total (p = 0.014) and phospho-MAPK (T202-Y204, p = 0.038), and phosphorylated MEK1 (S217-S221, p = 0.019), PKC alpha (S657, p = 0.006), PKC beta (S660, p = 0.037), and Src (Y527, p = 0.026). Protein pathway signatures (Akbani et al., 2014) recapitulated these findings with *proliferative* tumors having increased levels of the cell cycle (p = 0.005) and DNA damage response (p = 0.014) signatures and a lower RAS-MAPK signature (p = 0.031) score (Tables S1 and S6).

Using an integrative genomics approach, PARADIGM predicted increased activation of the TP53, TP63, TP73 TCF/beta-catenin PKC, and JUN/FOS pathways in *reactive-like* tumors; increased activation of immune-modulators IL12 and IL23, IL12R and IL23R, JAK2 and TYK2 in *immune-related* ILCs (Baay et al., 2011; Duvallet et al., 2011; Strobl et al., 2011), and decreased activation of each of these pathways along with lower levels of MAPK3, RB1, and ERK1 (Figure 5C) in *proliferative* ILC tumors.

Lastly, we determined that *reactive-like* ILC patients had a significantly better disease-specific (DSS) (p = 0.038, HR: 0.47) and overall survival (OS) (p = 0.023, HR: 0.50) compared to *proliferative* ILC patients in the METABRIC dataset, which has a median follow-up of 7.2 years (compared to the TCGA median follow-up of less than 2 years), (Figure 5D). Consistent with these results, patients with more proliferative lobular tumors (i.e., greater than the median PAM50 proliferation signature score) had worse DSS (p = 0.025, HR: 2.0) and a tendency toward worse OS (p = 0.058, HR: 0.63) compared to patients with a lower proliferation score (Figures S5N–S5O). No significant differences in DSS or OS were identified between the *immune-related* subgroup and either the *proliferative* or *reactive-like* subgroup. These results are consistent with previous studies reporting that the reactive stromal phenotype is associated with a good prognosis in breast cancer while proliferation is one of the strongest indicators of worse outcome in luminal/ER+ breast cancers (Ciriello et al., 2013b).

Tumors with Mixed ILC and IDC Histology

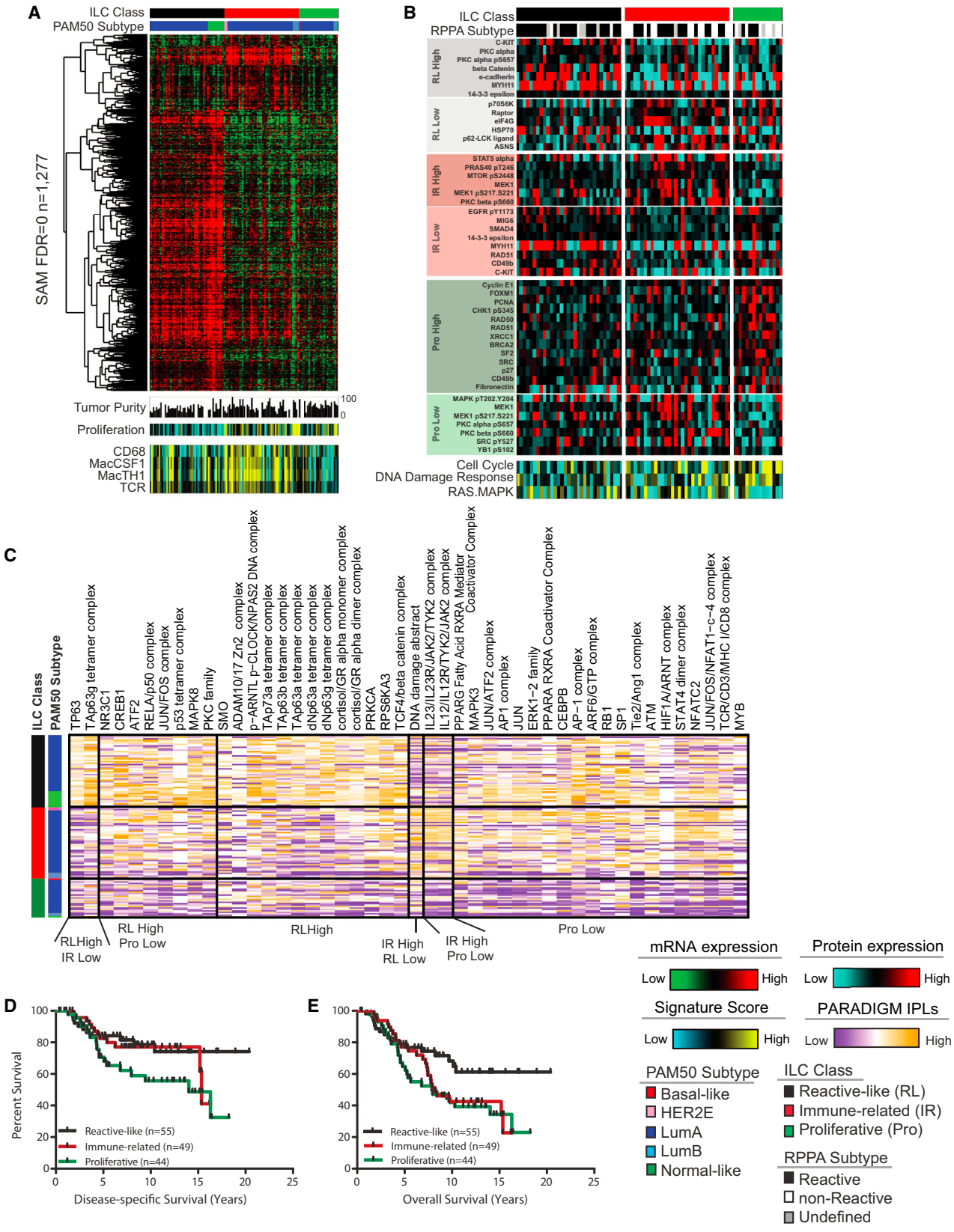
Histologically, ~3%–6% of breast tumors present both a ductal and a lobular component (Figure 6A). Pathologists currently classify these tumors as *mixed ductal/lobular breast carcinoma* or *invasive ductal cancers with lobular features* (Arps et al., 2013). There are, however, no defined criteria or uniform terminology for the classification of *mixed* tumors and as a consequence discordant clinical and molecular features have been reported (Bharat et al., 2009). Molecular profiling has the power to provide quantitative endpoints to compare the genetics of mixed tumors with those of pure ILC and IDC. In our dataset, 88/817 tumors (11%) were classified by as mixed ductal/lobular breast carcinomas. We characterized these *mixed* tumors using multiple computational approaches integrating different data-types thus to determine whether they molecularly resembled IDC (IDC-like), ILC (ILC-like), or neither.

We first analyzed the transcriptional landscape of mixed tumors using the ISOpure algorithm (Quon et al., 2013), which deconvolves the transcriptional signal of each queried tumor to estimate how much of it can be explained by one or more reference populations and how much is unique. Interestingly, mRNA expression profiles of all *mixed* cases could almost completely be explained by either IDC or ILC reference populations, suggesting that these tumors separate into IDC-like and ILC-like cases and do not represent a molecularly distinct subtype (Figure S6A). Based on this analysis, 32/88 *mixed* cases received an ILC-score greater than the IDC-score, and were therefore classified as ILC-like (Figure 6B).

We next evaluated the resemblance of *mixed* tumors to IDC and ILC based on the previously determined selected set of copy-number alterations (CNAs) and mutations (Table S2). *Mixed* tumors were enriched for IDC recurrent CNAs and mutations when compared to ILC, and vice versa (Figure S6B), indicating ILC and IDC genetic alterations were both present in these tumors, either simultaneously or in separate IDC-like and ILC-like subgroups. We then compared each *mixed* tumor to ILC and IDC based on their genomic features, by adapting the OncoSign algorithm (Ciriello et al., 2013a). This approach identified 19 ILC-like mixed samples characterized by ILC genetic features (Figures 6B and S6D). All *CDH1*-mutated *mixed* cases were classified as ILC-like, indicating *CDH1* status as a dominant feature in this analysis. A few *CDH1* wild-type mixed cases were also classified as ILC-like and characterized by ILC-enriched events such as mutations in *RUNX1* (3/4 mutated cases), *TBX3* (2/4), and *FOXA1* (2/6). ILC-enriched alterations did not co-occur with IDC-enriched ones, further indicating that mixed tumors can be categorized into ILC-like or IDC-like subgroups and do not constitute a molecularly distinct subtype.

Finally, we combined 428 CNA, including focal and arm-level alterations, 409 gene expression modules (Fan et al., 2011; Gatza et al., 2014) and somatic mutations for 128 genes mutated in more than 3% of the cases into a single ElasticNet classifier (Zou and Hastie, 2005). This integrated ElasticNet predictor identified 27/88 mixed tumors as ILC-like. These were enriched for the LumA subtype, *CDH1* mutations and loss of E-cadherin mRNA expression (Figures 6B and S6E).

Overall, these approaches were highly concordant (Figures 6B and S6F) with 24/88 cases (18/57 LumA cases) being called ILC-like by at least two approaches, and 64 being called IDC-like (Table S1). ILC-like and IDC-like mixed tumors when compared to pure ILC and IDC, respectively, do not show significant enrichment for specific genomic alterations, being molecularly similar to either one or the other subtype. Our analyses demonstrate that mixed histology tumors overwhelmingly tend



(legend on next page)

to resemble either ILC or IDC as opposed to representing a third distinct group. Moreover, IDC and ILC discriminant molecular features, in particular *CDH1* status, could be used to stratify *mixed* tumors into ILC-like and IDC-like tumor subgroups.

DISCUSSION

In this study we provide the most comprehensive molecular portrait to date of ILC. E-cadherin loss was confirmed the ILC hallmark lesion, and we could identify CDH1 loss at the DNA, mRNA, and protein level in almost all ILC cases. Moreover, 12/27 *CDH1* mutations in non-ILC cases occurred in *mixed* tumors strongly resembling ILC at the molecular level. Surprisingly, we did not identify DNA hyper-methylation of the CDH1 promoter in any breast tumor, suggesting that E-cadherin loss is not epigenetically driven. In addition, ILC and IDC differed in the *FOXA1* and *GATA3* mutational spectra, *PTEN* loss, and Akt activation. The lower incidence of *GATA3* mutations in ILC and lower GATA3 mRNA and protein expression suggest that in LumA ILC tumors there is a preferential occupancy of ER in FOXA1 bounded sites (Theodorou et al., 2013). Differential ER activity is also observed at the protein level where both total ER ($p = 0.005$) and phospho-ER ($p = 2E-05$) levels are reduced in LumA ILC versus LumA IDC. These findings in the context of recent data suggesting an improved response to the aromatase inhibitor letrozole as compared to tamoxifen in ILC (Metzger et al., 2012; Sikora et al., 2014) warrants further investigation.

The chromatin remodeling factor EP300, also involved in ER modulation, is able to directly acetylate FOXA1, and EP300 driven acetylation prevents FOXA1 DNA binding, but does not affect the protein when already bound (Kohler and Cirillo, 2010). Intriguingly, five acetylation sites have been identified in the wings of the fork-head domain; three of them in W2 (K264, K267, and K270), where most of our newly observed FOXA1 mutations cluster. These observations lead to the hypothesis that FOXA1 mutations could alter EP300 dependent acetylation of FOXA1 without affecting EP300 modulation of ER. While a rigorous evaluation of the role of EP300 in breast cancer and how FOXA1 mutations interfere with it goes beyond the scope of this study, FOXA1 mutations, its correlation with FOXA1 expression and lack of DNA methylation at its binding sites, and exclusivity with *GATA3* mutations support these as events activating FOXA1 function and, thus, ER transcriptional program.

The PI3K/Akt pathway is among the most altered in cancer providing tumor cells with enhanced growth and survival capabilities. Integrating protein and phospho-protein data with gene expression and pathway activity signatures, we consistently identified increased Akt signaling in ILC versus IDC. Notably, E-cadherin loss has been associated with Akt activation and EGFR overexpression (Lau et al., 2011; Liu et al., 2013). Lack of E-cadherin expression, which characterizes almost all ILC tumors, may thus provide a favorable cellular context for Akt activation. Recently, PI3K and Akt inhibitors entered clinical trials for several cancer types including breast cancer. Here we showed that ILC has on average the highest levels of Akt activation, measured by phospho-Akt and PI3K/Akt signaling among all breast cancer subtypes (comparable to IDC basal-like), making selective inhibition of this pathway in ILC a particularly attractive strategy.

Unbiased characterization of the ILC transcriptome showed a high degree of internal variability giving rise to three main subgroups: *reactive-like, immune-related*, and *proliferative*. While additional validation studies will clearly be required, we do observe increased expression of many druggable pathways/targets including increased levels of phospho-mTOR and phospho-MEK1 expression in the *immune-related* subgroup as well as increased SMO and ERK pathway activity in the *reactive-like* subgroup. These results, coupled with difference in clinical outcome, suggest that these subgroups will be important for future studies focused on both the clinical and biological aspects of ILC.

Finally, we showed that *mixed* ILC/IDC tumors could be separated into two major groups based on their molecular resemblance to either ILC (ILC-like) or IDC (IDC-like). The ability to classify cancers with *mixed* phenotypes based on the underlying biology has implications for clinical practice as well as furthering our understanding of the etiology of such lesions. Indeed, ILC carcinomas often metastasize to body sites not colonized by IDCs (e.g., gastrointestinal [GI] tract and peritoneal surfaces). ILC are also typically of low histologic grade and with low to intermediate mitotic index, thus limiting their response to primary chemotherapy (Cristofanilli et al., 2005) and their ability to be detected on PET scans. As such, clinicians must be aware of nonspecific symptomatology and favor diagnostic approaches such as anatomical scanning (CT scan) for ILCs. Finally, the identification of ILC enriched molecular features may ultimately lead to the design of ILC-targeted therapies. A more refined classification of mixed cancers as IDC-like or ILC-like will improve our understanding, detection, and follow-up of the disease, and enable a more informed and targeted treatment selection.

Figure 5. ILC Molecular Subtypes

(A) Three molecular subtype of lobular breast cancer were identified based on differential gene expression and show unique patterns highly expressed genes ($n = 1277$, SAM FDR = 0, upper panel), minor difference in tumor purity measured by ABSOLUTE, and differences in gene expression signatures measuring proliferation, CD68, Macrophage-associated CSF1, Macrophage–associated TH1, and T Cell Receptor Signaling (lower panel). Proliferation is highest in the *proliferative* (Pro) and *immune-related* (IR) subgroups; macrophage associated signaling is highest in *immune-related* tumors.

(B) Differences in protein expression profiles as determined by RPPA analysis. The *reactive-like* (RL) subgroup shows a significant association ($p < 1E-4$, Fisher's Exact test) with the RPPA-defined reactive subgroup of breast cancer. Differences in subgroup-specific patterns of protein expression ($p < 0.05$, t test) for individual proteins (upper panel) as well as for protein expression signatures (lower panel) were identified. The proliferative subgroup shows higher expression of the cell-cycle and DNA damage response pathways and lower levels of Ras-MAPK signaling ($p < 0.05$).

(C) Subgroup-associated signaling features identified by PARADIGM.

(D and E) *Reactive-like* ($n = 55$) tumors have significantly better (D) disease specific ($p = 0.038$, HR: 0.47, log-rank) and (E) overall survival ($p = 0.022$, HR = 0.50) compared to *proliferative* ($n = 44$) tumors in the METABRIC cohort.

See also Figure S5.

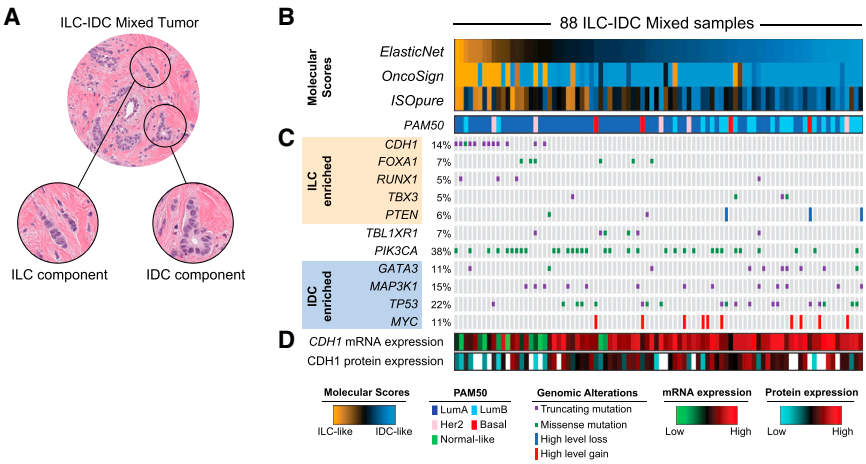


Figure 6. Molecular Classification of Mixed Ductal/Lobular Carcinoma

(A) Mixed ductal/lobular tumors present at the same time both a lobular and ductal component.

(B) We used three algorithmic approaches (ElasticNet, OncoSign, and ISOpure) to evaluate the resemblance of mixed tumors to either ILC (ILC-like) or IDC (IDC-like) based on molecular features. ILC-IDC scores are shown for all three approached at the top. See also Figures S6B–S6D.

(C) Genetic alterations enriched in ILC tumors are frequently found in ILC-like mixed cases (in particular CDH1 mutations), whereas those enriched in IDC are more frequent in IDC-like mixed cases.

(D) ILC-like mixed-cases are characterized by both low E-cadherin mRNA and protein level. See also Figure S6E.

This multi-platform study identified numerous molecular features discriminating between breast ILC and IDC, demonstrating different pathways underlying their pathogenesis, defining new ILC subtypes with different clinical outcomes, and pointing to previously unrecognized therapeutic possibilities. Importantly, we provided here a curated and integrated dataset for 817 breast tumors, including the largest collection to date of comprehensively profiled ILC. To facilitate the exploration of this dataset, we created a public web-service (http://cbio.mskcc.org/cancergenomics/tcga/brca_tcga) organizing all analyses and data used in this manuscript. We believe this resource will serve as a reference for many to further advance our understanding of human breast cancers.

EXPERIMENTAL PROCEDURES

Tumor and matched normal specimen were collected as previously described (Cancer Genome Atlas, 2012). In total 817 primary tumor samples were assayed by whole-exome DNA sequencing, RNA sequencing, miRNA sequencing, SNP arrays, and DNA methylation arrays. A subset of 633 samples was assayed by reverse phase protein array (RPPA). Histological subtypes have been determined based on consensus by a pathology committee. Intrinsic breast cancer subtyping was performed on all 817 cases, using the PAM50 classifier (Parker et al., 2009). Data generation and processing were performed as previously described (Cancer Genome Atlas Research Network, 2014).

Enrichment analyses for selected events were performed using Fisher's exact tests and a binary representation of copy-number alterations and mutations (1 is altered, 0 is wild-type).

DNA methylation of the CDH1 promoter was assessed at probes within a window 1,500 bp upstream and downstream CDH1 transcription start site using both HM27 and HM450 data. Whole-genome bisulfite sequencing was performed to characterize DNA methylation levels at 157 CpGs.

Distances between of FOXA1 mutations have been determined from the tertiary structure of FOXA3 fork-head domain (PDB ID: 1VTN). Predicted DNA interactions were derived by WebPDA (http://bioinfozen.uncc.edu/webpda). Differential expression analyses on RNA-seq data were performed using the limmavoom package (Law et al., 2014).

Replication Based Normalized (RBN) RPPA data containing expression levels for 187 protein and phosphorylated proteins for 633 samples were used for protein differential expression analysis. Differential pathway activity was assessed by t test.

ILC subtypes were determined using Consensus Cluster Plus Analysis (Wilkerson and Hayes, 2010) based on the 1,000 most differentially expressed genes and a classifier was built using ClaNC (Dabney, 2006).

Detailed description of each analysis presented in this study can be found within the Supplemental Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and nine tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.033.

AUTHOR CONTRIBUTIONS

G.C., M.L.G. and C.M.P coordinated overall study design and analyses. A.H.B., S.C.L, G.M.K.T., R.E.F., L.C.C., K.H.A., Y-Y.C., K.J., and N.B.J. coordinated pathology analyses. M.D.W., M.McL., and C.K. coordinated DNA sequencing analyses. M.L.G., K.A.H. and C.M.P coordinated RNA-seq analyses. S.K.R., H.S., and P.W.L. coordinated DNA methylation analyses. H.Z. and A.D.C. coordinated copy-number analyses. R.B. and G.R. coordinated miRNA-seq analyses. G.B.M. coordinated RPPA analyses. A.P., C.Y., S.H., R.F. were involved in bioinformatics analyses. C.B. and C.S. supervised bioinformatics analyses. S.O. and T.A.K. coordinated clinical contributions. G.C. developed web-resource for breast cancer data. G.C., M.L.G and C.M.P. wrote the manuscript, which all authors reviewed.

CONSORTIA

Rehan Akbani, J. Todd Auman, Miruna Balasundaram, Saianand Balu, Thomas Barr, Andrew Beck, Christopher Benz, Stephen Benz, Mario Berrios, Rameen Beroukhim, Tom Bodenheimer, Lori Boice, Moiz S. Bootwalla, Jay Bowen, Reanne Bowlby, Denise Brooks, Andrew D. Cherniack, Lynda Chin, Juok Cho, Sudha Chudamani, Giovanni Ciriello, Tanja Davidsen, John A. Demchok, Jennifer B. Dennison, Li Ding, Ina Felau, Martin L. Ferguson, Scott Frazer, Stacey B. Gabriel, JianJiong Gao, Julie M. Gastier-Foster, Michael L. Gatza, Nils Gehlenborg, Mark Gerken, Gad Getz, William J. Gibson, D. Neil Hayes, David I. Heiman, Katherine A. Hoadley, Andrea Holbrook, Robert A. Holt, Alan P. Hoyle, Hai Hu, Mei Huang, Carolyn M. Hutter, E. Shelley Hwang, Stuart R. Jefferys, Steven J.M. Jones, Zhenlin Ju, Jaegil Kim, Phillip H. Lai, Peter W. Laird, Michael S. Lawrence, Kristen M. Leraas, Tara M. Lichtenberg, Pei Lin, Shiyun Ling, Jia Liu, Wenbin Liu, Laxmi Lolla, Yiling Lu, Yussanne Ma, Dennis T. Maglinte, Elaine Mardis, Jeffrey Marks, Marco A. Marra, Cynthia McAllister, Michael McLellan, Shaowu Meng, Matthew Meyerson, Gordon B. Mills, Richard A. Moore, Lisle E. Mose, Andrew J. Mungall, Bradley A. Murray, Rashi Naresh, Michael S. Noble, Steffi Oesterreich, Olufunmilayo Olopade, Joel S. Parker, Charles M. Perou, Todd Pihl, Gordon Saksena, Steven E. Schumacher, Kenna R. Mills Shaw, Nilsa C. Ramirez, W. Kimryn Rathmell, Suhn K. Rhie, Jeffrey Roach, A. Gordon Robertson, Gordon Saksena, Chris Sander, Jacqueline E. Schein, Nikolaus Schultz, Hui Shen, Margi Sheth, Yan Shi,

Juliann Shih, Carl Simon Shelley, Craig Shriver, Janae V. Simons, Heidi J. Sofia, Matthew G. Soloway, Carrie Sougnez, Charlie Sun, Roy Tarnuzzer, Daniel G. Tiezzi, David J. Van Den Berg, Doug Voet, Yunhu Wan, Zhining Wang, John N. Weinstein, Daniel J. Weisenberger, Matthew D. Wilkerson, Richard Wilson, Lisa Wise, Maciej Wiznerowicz, Junyuan Wu, Ye Wu, Liming Yang, Christina Yau, Travis I. Zack, Jean C. Zenklusen, Hailei Zhang, Jiashan Zhang, Erik Zmuda.

ACKNOWLEDGMENTS

We wish to thank the many members of the TCGA Network, including the Tissue Source Sites, and patients, whom contributed samples to this study. This study was supported by funds from the TCGA Project (U24-CA143848), the NCI Breast SPORE program grant P50-CA58223-09A1, and the Breast Cancer Research Foundation. G.C. is supported by the Gabriella Giorgi-Cavaglieri Foundation, M.L.G. is supported by the National Cancer Institute of the US NIH award number K99-CA166228, A.H.B. is supported by National Library of Medicine of the NIH under Award Number K22LM011931, and A.P. is supported by Mildred-Scheel Postdoctoral Research Fellowship of the Deutsche Krebshilfe e.V. (No. 111354). C.M.P is an equity stock holder, and Board of Director Member, of BioClassifier. C.M.P is also listed an inventor on patent applications on the Breast PAM50 assay. A.D.C. receives research funding from Bayer AG.

Received: June 10, 2015
Revised: August 4, 2015
Accepted: September 10, 2015
Published: October 8, 2015

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. Nat. Commun. *5*, 3887.

Arpino, G., Bardou, V.J., Clark, G.M., and Elledge, R.M. (2004). Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome. Breast Cancer Res. *6*, R149–R156.

Arps, D.P., Healy, P., Zhao, L., Kleer, C.G., and Pang, J.C. (2013). Invasive ductal carcinoma with lobular features: a comparison study to invasive ductal and invasive lobular carcinomas of the breast. Breast Cancer Res. Treat. *138*, 719–726.

Baay, M., Brouwer, A., Pauwels, P., Peeters, M., and Lardon, F. (2011). Tumor cells and tumor-associated macrophages: secreted proteins as potential targets for therapy. Clin. Dev. Immunol. *2011*, 565187.

Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666–677.

Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. *44*, 685–689.

Bharat, A., Gao, F., and Margenthaler, J.A. (2009). Tumor characteristics and patient outcomes are similar between invasive lobular and mixed invasive ductal/lobular breast cancers but differ from pure invasive ductal breast cancers. Am. J. Surg. *198*, 516–519.

Brinck, U., Jacobs, S., Neuss, M., Tory, K., Rath, W., Kulle, B., and Füzesi, L. (2004). Diffuse growth pattern affects E-cadherin expression in invasive breast cancer. Anticancer Res. *24*, 2237–2242.

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

Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202–209.

Cantley, L.C., and Neel, B.G. (1999). New insights into tumor suppression: PTEN suppresses tumor formation by restraining the phosphoinositide 3-kinase/AKT pathway. Proc. Natl. Acad. Sci. USA *96*, 4240–4245.

Carroll, J.S., Liu, X.S., Brodsky, A.S., Li, W., Meyer, C.A., Szary, A.J., Eeckhoute, J., Shao, W., Hestermann, E.V., Geistlinger, T.R., et al. (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell *122*, 33–43.

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. Nat. Biotechnol. *30*, 413–421.

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

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

Ciriello, G., Sinha, R., Hoadley, K.A., Jacobsen, A.S., Reva, B., Perou, C.M., Sander, C., and Schultz, N. (2013b). The molecular diversity of Luminal A breast tumors. Breast Cancer Res. Treat. *141*, 409–420.

Cirillo, L.A., and Zaret, K.S. (2007). Specific interactions of the wing domains of FOXA1 transcription factor with DNA. J. Mol. Biol. *366*, 720–724.

Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol. Cell *9*, 279–289.

Cristofanilli, M., Gonzalez-Angulo, A., Sneige, N., Kau, S.W., Broglio, K., Theriault, R.L., Valero, V., Buzdar, A.U., Kuerer, H., Buchholz, T.A., and Hortobagyi, G.N. (2005). Invasive lobular carcinoma classic type: response to primary chemotherapy and survival outcomes. J. Clin. Oncol. *23*, 41–48.

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.; METABRIC Group (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.

Dabbs, D.J., Schnitt, S.J., Geyer, F.C., Weigelt, B., Baehner, F.L., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Lakhani, S.R., et al. (2013). Lobular neoplasia of the breast revisited with emphasis on the role of E-cadherin immunohistochemistry. Am. J. Surg. Pathol. *37*, e1–e11.

Dabney, A.R. (2006). ClaNC: point-and-click software for classifying microarrays to nearest centroids. Bioinformatics *22*, 122–123.

Duvallet, E., Semerano, L., Assier, E., Falgarone, G., and Boissier, M.C. (2011). Interleukin-23: a key cytokine in inflammatory diseases. Ann. Med. *43*, 503–511.

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 Med. Genomics *4*, 3.

Foote, F.W., Jr., and Stewart, F.W. (1946). A histologic classification of carcinoma of the breast. Surgery *19*, 74–99.

Gajiwala, K.S., and Burley, S.K. (2000). Winged helix proteins. Curr. Opin. Struct. Biol. *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. Nat. Genet. *46*, 1051–1059.

Graff, J.R., Herman, J.G., Myöhänen, S., Baylin, S.B., and Vertino, P.M. (1997). Mapping patterns of CpG island methylation in normal and neoplastic cells implicates both upstream and downstream regions in de novo methylation. J. Biol. Chem. *272*, 22322–22329.

Grasso, C.S., Wu, Y.M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing, X., Lonigro, R.J., Brenner, J.C., et al. (2012). The mutational landscape of lethal castration-resistant prostate cancer. Nature *487*, 239–243.

Habashy, H.O., Powe, D.G., Rakha, E.A., Ball, G., Paish, C., Gee, J., Nicholson, R.I., and Ellis, I.O. (2008). Forkhead-box A1 (FOXA1) expression in breast cancer and its prognostic significance. Eur. J. Cancer *44*, 1541–1551.

Herman, J.G., Graff, J.R., Myöhänen, S., Nelkin, B.D., and Baylin, S.B. (1996). Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc. Natl. Acad. Sci. USA *93*, 9821–9826.

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

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. Clin. Cancer Res. *20*, 3818–3829.

Kohler, S., and Cirillo, L.A. (2010). Stable chromatin binding prevents FoxA acetylation, preserving FoxA chromatin remodeling. J. Biol. Chem. *285*, 464–472.

Lau, M.T., Klausen, C., and Leung, P.C. (2011). E-cadherin inhibits tumor cell growth by suppressing PI3K/Akt signaling via β-catenin-Egr1-mediated PTEN expression. Oncogene *30*, 2753–2766.

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 Biol. *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.

Liu, X., Su, L., and Liu, X. (2013). Loss of CDH1 up-regulates epidermal growth factor receptor via phosphorylation of YBX1 in non-small cell lung cancer cells. FEBS Lett. *587*, 3995–4000.

Liu, Z., Merkurjev, D., Yang, F., Li, W., Oh, S., Friedman, M.J., Song, X., Zhang, F., Ma, Q., Ohgi, K.A., et al. (2014). Enhancer activation requires trans-recruitment of a mega transcription factor complex. Cell *159*, 358–373.

McCart Reed, A.E., Kutasovic, J.R., Lakhani, S.R., and Simpson, P.T. (2015). Invasive lobular carcinoma of the breast: morphology, biomarkers and 'omics. Breast Cancer Res. *17*, 12.

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 Biol. *12*, R41.

Metzger, O., Giobbie-Hurder, A., Mallon, E., Viale, G., Winer, E., Thurlimann, B., Gelber, R.D., Colleoni, M., Ejlertsen, B., Bonnefoi, H., et al. (2012). Abstract S1-1: Relative effectiveness of letrozole compared with tamoxifen for patients with lobular carcinoma in the BIG 1-98 trial. In Thirty-Fifth CTRC-AACR San Antonio Breast Cancer Symposium (San Antonio, TX).

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

Morrogh, M., Andrade, V.P., Giri, D., Sakr, R.A., Paik, W., Qin, L.X., Arroyo, C.D., Brogi, E., Morrow, M., and King, T.A. (2012). Cadherin-catenin complex dissociation in lobular neoplasia of the breast. Breast Cancer Res. Treat. *132*, 641–652.

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. J. Clin. Oncol. *27*, 1160–1167.

Pestalozzi, B.C., Zahrieh, D., Mallon, E., Gusterson, B.A., Price, K.N., Gelber, R.D., Holmberg, S.B., Lindtner, J., Snyder, R., Thürlimann, B., et al.; International Breast Cancer Study Group (2008). Distinct clinical and prognostic features of infiltrating lobular carcinoma of the breast: combined results of 15 International Breast Cancer Study Group clinical trials. J. Clin. Oncol. *26*, 3006–3014.

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 Med. *5*, 29.

Richards, F.M., McKee, S.A., Rajpar, M.H., Cole, T.R., Evans, D.G., Jankowski, J.A., McKeown, C., Sanders, D.S., and Maher, E.R. (1999). Germline E-cadherin gene (CDH1) mutations predispose to familial gastric cancer and colorectal cancer. Hum. Mol. Genet. *8*, 607–610.

Robinson, D., Van Allen, E.M., Wu, Y.M., Schultz, N., Lonigro, R.J., Mosquera, J.M., Montgomery, B., Taplin, M.E., Pritchard, C.C., Attard, G., et al. (2015). Integrative clinical genomics of advanced prostate cancer. Cell *161*, 1215–1228.

Ross, J.S., Wang, K., Sheehan, C.E., Boguniewicz, A.B., Otto, G., Downing, S.R., Sun, J., He, J., Curran, J.A., Ali, S., et al. (2013). Relapsed classic E-cadherin (CDH1)-mutated invasive lobular breast cancer shows a high frequency of HER2 (ERBB2) gene mutations. Clin. Cancer Res. *19*, 2668–2676.

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.

Sahu, B., Laakso, M., Ovaska, K., Mirtti, T., Lundin, J., Rannikko, A., Sankila, A., Turunen, J.P., Lundin, M., Konsti, J., et al. (2011). Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. EMBO J. *30*, 3962–3976.

Sarrió, D., Moreno-Bueno, G., Hardisson, D., Sánchez-Estévez, C., Guo, M., Herman, J.G., Gamallo, C., Esteller, M., and Palacios, J. (2003). Epigenetic and genetic alterations of APC and CDH1 genes in lobular breast cancer: relationships with abnormal E-cadherin and catenin expression and microsatellite instability. Int. J. Cancer *106*, 208–215.

Sérandour, A.A., Avner, S., Percevault, F., Demay, F., Bizot, M., Lucchetti-Miganeh, C., Barloy-Hubler, F., Brown, M., Lupien, M., Métivier, R., et al. (2011). Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. Genome Res. *21*, 555–565.

Sikora, M.J., Cooper, K.L., Bahreini, A., Luthra, S., Wang, G., Chandran, U.R., Davidson, N.E., Dabbs, D.J., Welm, A.L., and Oesterreich, S. (2014). Invasive lobular carcinoma cell lines are characterized by unique estrogen-mediated gene expression patterns and altered tamoxifen response. Cancer Res. *74*, 1463–1474.

Song, M.S., Salmena, L., and Pandolfi, P.P. (2012). The functions and regulation of the PTEN tumour suppressor. Nat. Rev. Mol. Cell Biol. *13*, 283–296.

Strobl, B., Stoiber, D., Sexl, V., and Mueller, M. (2011). Tyrosine kinase 2 (TYK2) in cytokine signalling and host immunity. Front. Biosci. (Landmark Ed.) *16*, 3214–3232.

Theodorou, V., Stark, R., Menon, S., and Carroll, J.S. (2013). GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. Genome Res. *23*, 12–22.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA *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–i245.

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 Res. *22*, 1798–1812.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

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., and Hayes, D.N. (2014). Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. Nucleic Acids Res. *42*, e107.

Wu, K., Chang, Q., Lu, Y., Qiu, P., Chen, B., Thakur, C., Sun, J., Li, L., Kowluru, A., and Chen, F. (2013). Gefitinib resistance resulted from STAT3-mediated Akt activation in lung cancer cells. Oncotarget *4*, 2430–2438.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc., B *67*, 301–320.

Zou, D., Yoon, H.S., Perez, D., Weeks, R.J., Guilford, P., and Humar, B. (2009). Epigenetic silencing in non-neoplastic epithelia identifies E-cadherin (CDH1) as a target for chemoprevention of lobular neoplasia. J. Pathol. *218*, 265–272.

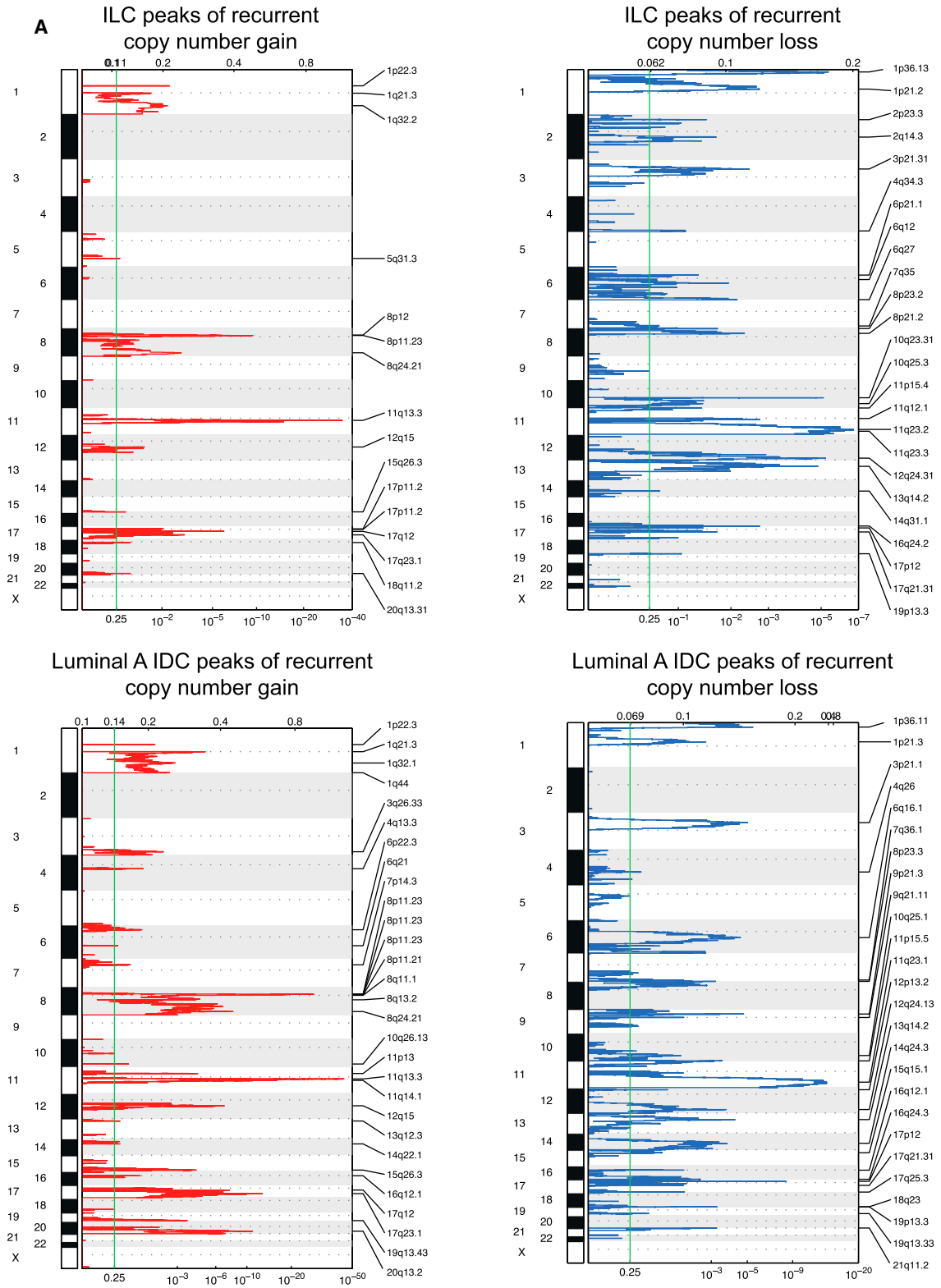
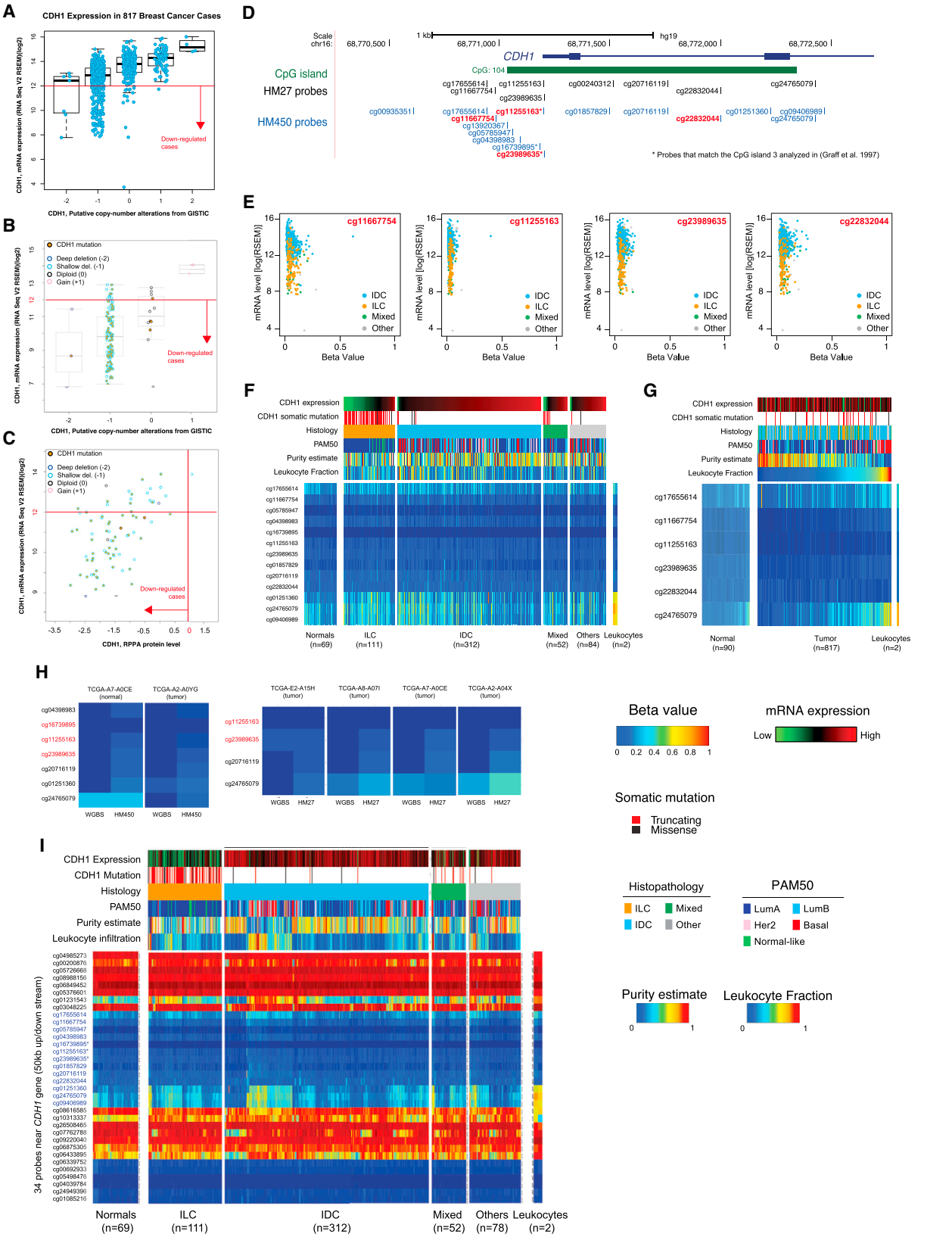


Figure S1. Significant Peaks of Copy-Number Gain and Loss Identified by GISTIC in ILC, IDC, luminal A ILC, and luminal A IDC, Related to Figure 1



(legend on next page)

Figure S2. Related to Figure 2

(A and B) CDH1 mRNA expression in (A) all 817 breast cancer and (B) in 127 ILC cases measured by RNA-seq (log2 RSEM) with respect to copy-number status. Based on the observed distribution of CDH1 mRNA levels, cases with log2(RSEM) < 12 were called downregulated for CDH1 mRNA expression.

(C) CDH1 mRNA expression in 127 ILC cases measured by RNA-seq (log2 RSEM) compared to protein level measure by RPPA. Cases with RPPA z-score < 0 were called downregulated for CDH1 protein expression.

(D) DNA methylation probes matching CDH1 CpG island.

(E) DNA methylation levels at representative probed sites are consistently low and independent of CDH1 mRNA expression and tumor histology.

(F) DNA methylation status of all probes matching the CDH1 CpG island from the HM450 Infinium platform. Samples are sorted first by histological subtype and then by CDH1 mRNA expression. DNA methylation status is also reported for two normal leukocytes.

(G) DNA methylation status of probes matching the CDH1 CpG island and present in both the HM450 and HM27 Infinium platforms. Samples are sorted by estimated leukocyte fraction.

(H) DNA methylation levels measured by whole-genome bisulfate sequencing in 5 tumor and 1 normal samples.

(I) DNA methylation levels of 34 probes covering genomic loci from 50kb upstream to 50kb down-stream of CDH1. Samples are clustered within each histological subtypes. Probes are sorted by genomic locus with probes in blue matching the CDH1 CpG island and starred probes corresponding to the genomic loci previously reported as methylated in Graff et al. (1997).

Boxplots have been generated such that the box is delimited by the lower and upper quartile, the thick line indicates the sample median, and whiskers reach whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

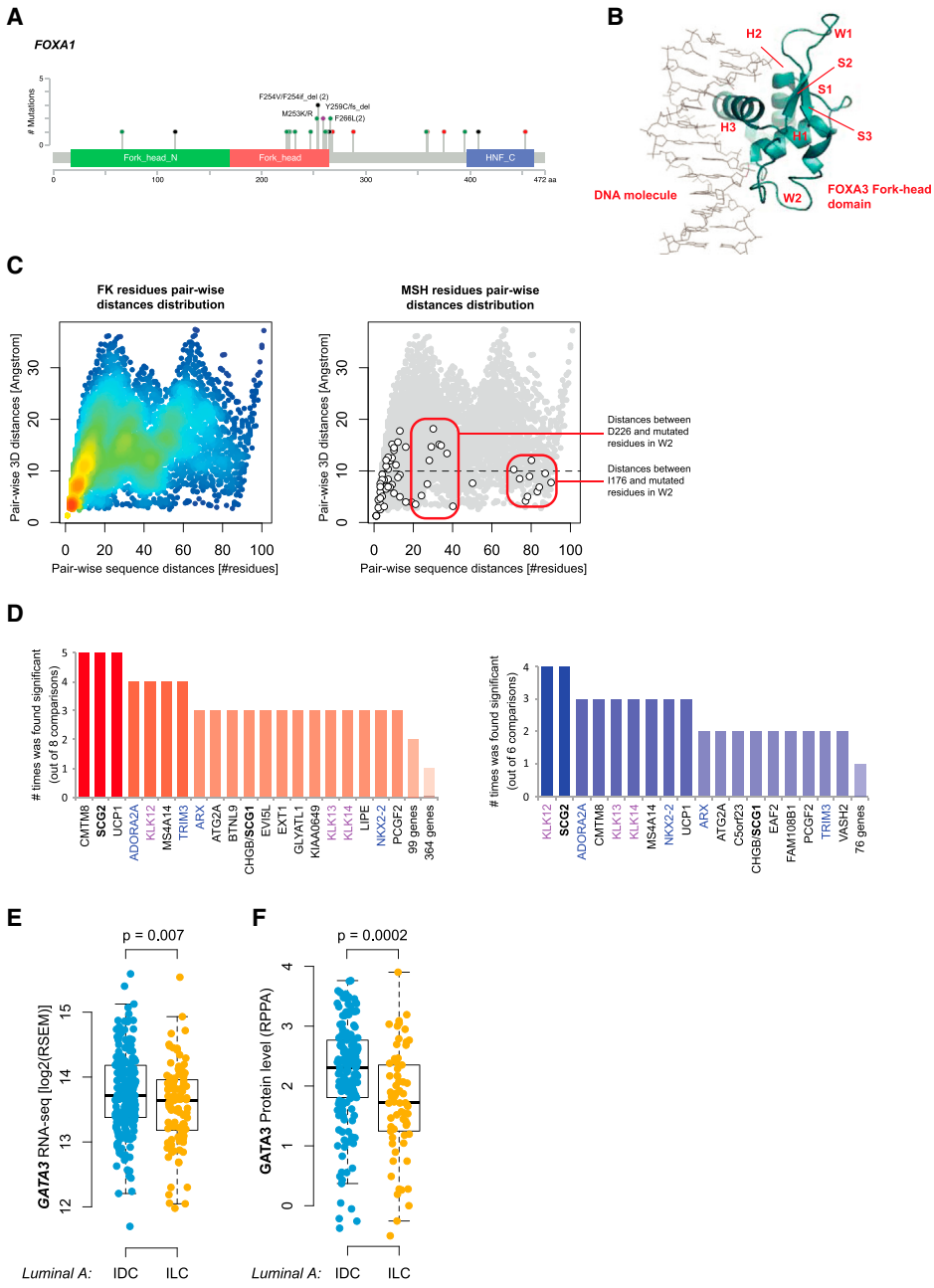


Figure S3. Related to Figure 3

(A) FOXA1 mutations from 3 prostate cancer studies show the same regional hotspots observed in breast cancer.

(B) DNA-bounded fork-head domain three-dimensional structure for FOXA3 (PDB ID: 1VTN). Secondary structural elements include 3 helices (H1, H2, H3), 3 beta-strands, (S1, S2, S3) and two loops or wings (W1, W2).

(C) For each pair of residues in the fork-head domain, pair-wise distances in the sequence space are compared to pair-wise distances in the 3D space. Uneven density of points are color-coded in the scatterplot with blue indicating the lowest density and red the highest. Pair-wise distances between mutated residues in the Mutation Structural Hotspot (MSH) are highlighted on the right plot.

(D) Differentially expressed genes across at least 3 out of all 8 comparisons (red) and across at least 2 out of 6 comparisons including ILC and luminal A samples (blue).

(E) GATA3 mRNA expression is significantly lower in ILC luminal A versus IDC lumina A.

(F) GATA3 protein level is significantly lower in ILC luminal A versus IDC lumina A.

Boxplots have been generated such that the box is delimited by the lower and upper quartile, the thick line indicates the sample median, and whiskers reach whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

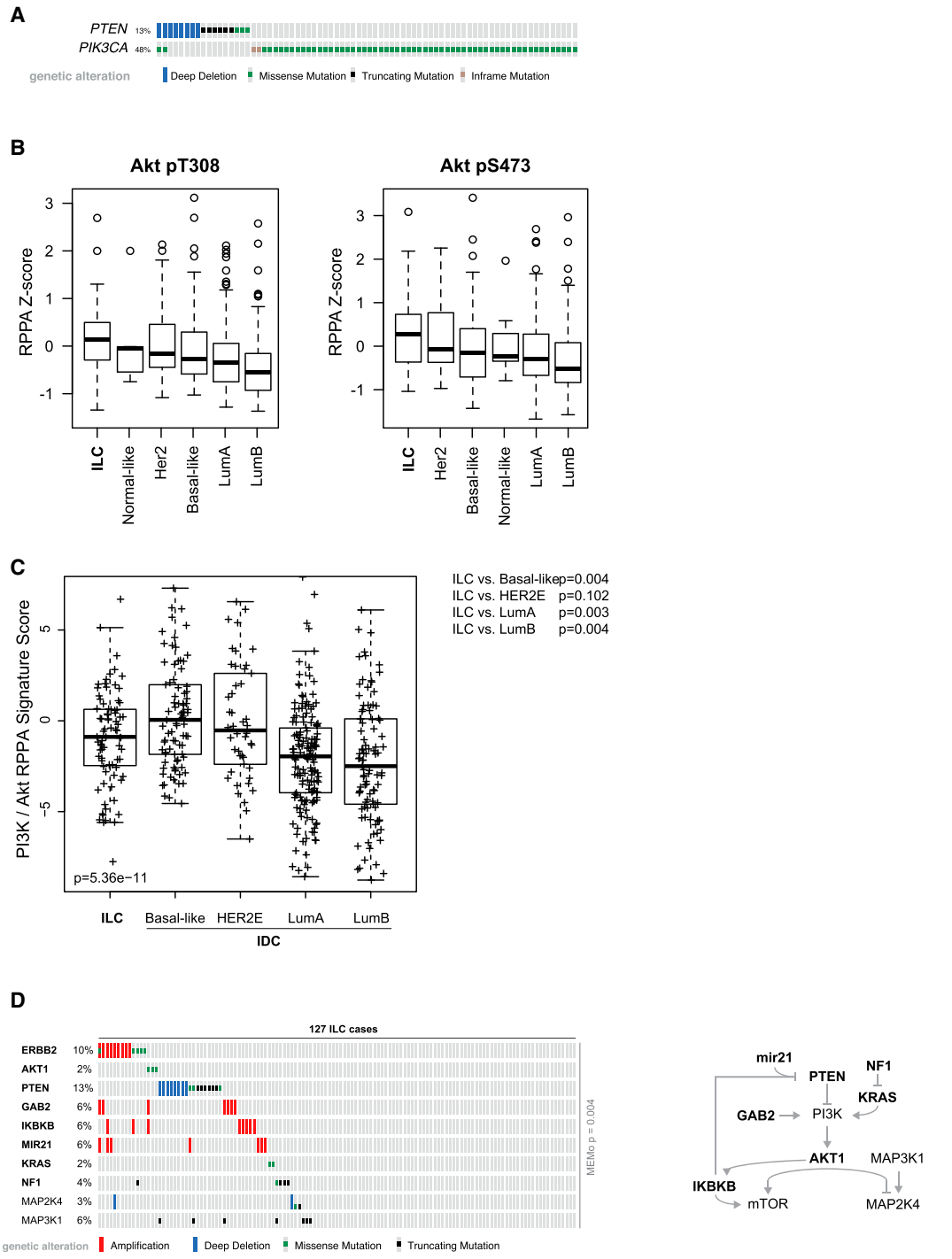


Figure S4. Related to Figure 4

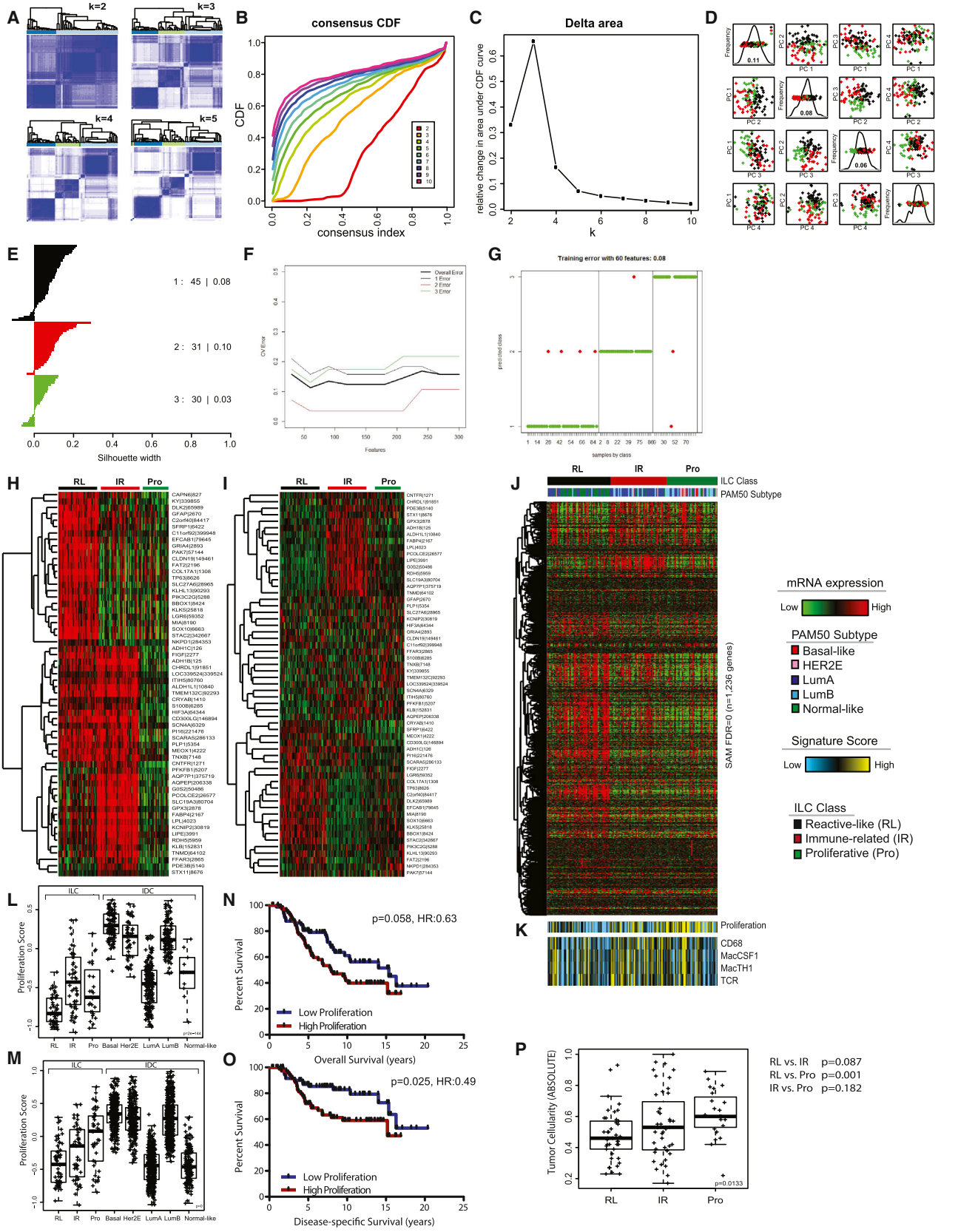
(A) PTEN deep deletions and mutations are almost perfectly mutually exclusive with PIK3CA mutations.

(B) Phospho-AKT levels have been measured by RPPA at both T308 and S473. On average, phosphorylation at these sites is at its highest in ILC tumors.

(C) PI3K/Akt RPPA protein expression signature defines differences in down-stream Akt signaling between ILC and PAM50 IDC subtype.

(D) MEMo identifies significant mutual exclusivity between alterations affecting the PI3K/AKT pathway as well as upstream and downstream targets in 45% of ILC cases. PIK3CA mutations, even if not part of the MEMo module, are shown at the bottom for completeness.

Boxplots have been generated such that the box is delimited by the lower and upper quartile, the thick line indicates the sample median, and whiskers reach whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.



(legend on next page)

Figure S5. Related to Figure 5

(A) Consensus clustering was used to determine ILC mRNA-based subtypes with K = 3 representing the chose solution. Solutions with greater Ks preserve the main 3-clusters only separating a few samples into very small-size clusters.

(B and C) Cumulative Distribution Functions (CDF) for (B) the consensus index and (C) the relative change in the area under the CDF curve support the 3 cluster solution.

(D) Principal component analysis shows PC1, PC2, PC3, and PC4 discriminant power between the 3 ILC subtypes.

(E) Positive silhouette width of each the original three subtypes was used to select 'core' members for development of quantitative classifier using CLaNC

(F) Cross Validation (CV) error for CLaNC-derived 3-class classifier using 10 to 300 genes shows 60 genes has the lowest CVerror rate.

(G) Comparison between 3 classes generated using CLaNC (y axis) and consensus clustering (x axis) shows Training Error of 8%.

(H) Hierarchical clustering of the 60 gene classifier in the TCGA (n = 127) lobular dataset identified subsets of genes upregulated in each class.

(I) Hierarchical clustering of the METABRIC dataset (n = 148) lobular tumors shows consistent patterns of expression of the 60 genes in the classifier.

(J) Hierarchical clustering of the 1,236 of 1,277 subtype-enriched genes (SAM FDR = 0) present in the METABRIC dataset show comparable patterns of expression; original classes 1, 2, and 3 have been renamed Reactive-like (RL), immune-related (IR), and proliferative (PRO) subtypes.

(K) Gene expression signatures measuring proliferation and macrophage-associated signaling: CD68, Macrophage Colony Stimulating Factor, Macrophage TH1, and TCR (T cell receptor signaling) show consistent patterns of activity as compared the TCGA subtypes.

(L) Box and whisker plot showing the level of proliferation (as measured by the PAM50 proliferation gene expression signature) in the lobular subtypes and ductal PAM50 subtypes; lobular tumors show consistently lower levels of proliferation.

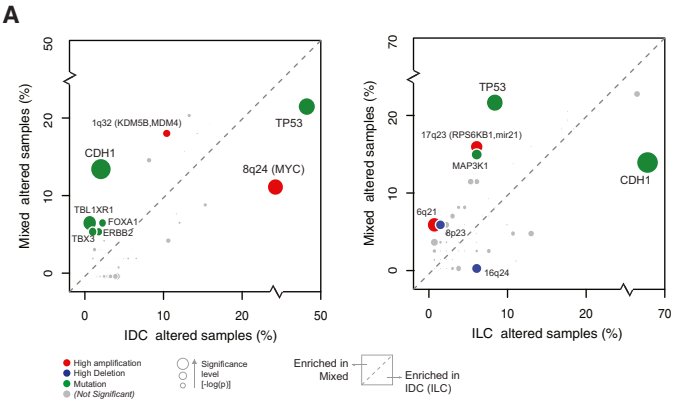
(M) Box and whisker plot showing the distribution of proliferation levels in the METBRIC dataset; consistent with (l) differences are present between lobular subtypes but ILC tumors generally have lower levels of proliferation compared to IDC tumors.

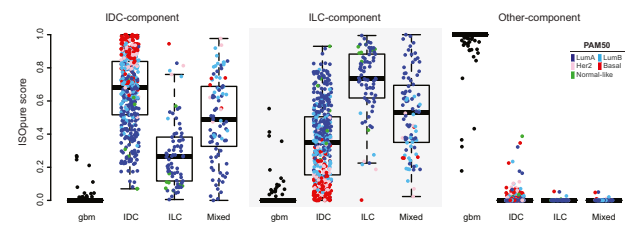
(N) Differences in overall survival in the METBRIC ILC samples (n = 148) based on PAM50 proliferation score (median) shows a trend toward better OS for patients with less proliferative tumors.

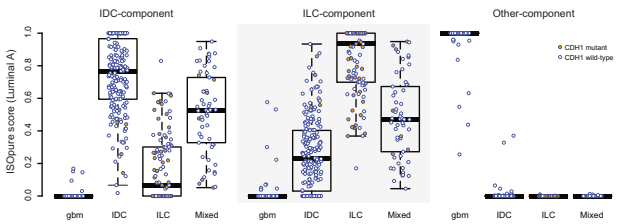
(O) Kaplan-Meier plot of disease specific survival (DSS) shows a significantly better prognosis for patients with less proliferative tumors.

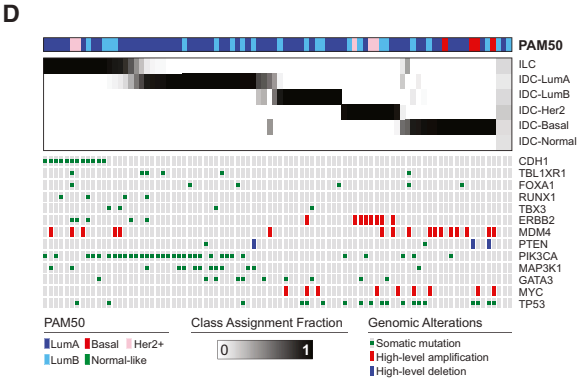
(P) ABSOLUTE was used to calculate tumor cellularity (t test) between ILC tumor subtypes.

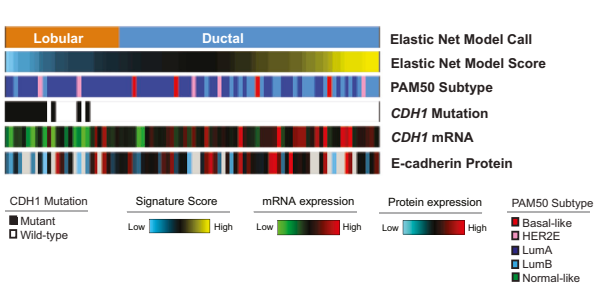
Boxplots have been generated such that the box is delimited by the lower and upper quartile, the thick line indicates the sample median, and whiskers reach whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

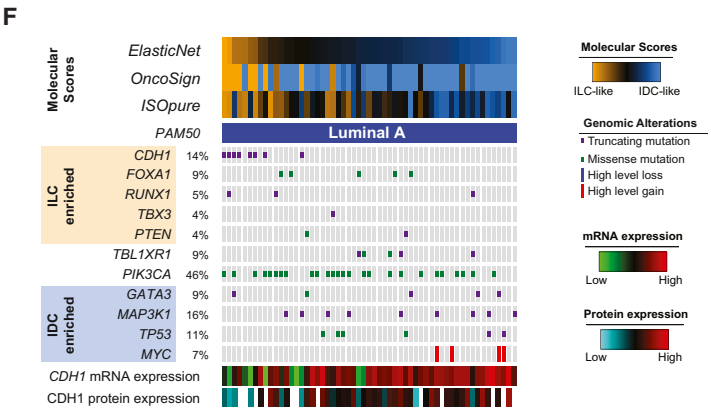












(legend on next page)

Figure S6. Related to Figure 6

(A) The alteration frequency of recurrent events in breast cancer is compared between IDC and Mixed tumors, and between ILC and Mixed tumors.
(B) ISOpure scores quantify how much of the transcriptional landscape of a tumor relates to IDC (IDC-score), ILC (ILC-score), or none of them (Other-score). ISOpure score have been computed for all 88 Mixed tumors as well as for 440 IDC cases, 79 ILC cases, and 153 Glioblastoma (GBM) as controls. PAM50 subtypes reveal luminal A tumors tend to have higher ILC-score independently of their histology.
(C) ISOpure scores have been recomputed using luminal A samples only.
(D) OncoSign assignments of Mixed tumors to either ILC or IDC broken down by PAM50 subtypes (top). Genomic alterations enriched in ILC are more frequently observed in ILC-like mixed cases, vice versa IDC-enriched alterations are more frequent in IDC-like mixed cases (bottom).
(E) ElasticNet scores and classification in ILC-like and IDC-like mixed tumors reflect Pam50 subtypes and E-cadherin status.
(F) Combined molecular classification for luminal A mixed tumors.
Boxplots have been generated such that the box is delimited by the lower and upper quartile, the thick line indicates the sample median, and whiskers reach whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

Cell

Supplemental Information

Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer

Giovanni Ciriello, Michael L. Gatza, Andrew H. Beck, Matthew D. Wilkerson, Suhn K. Rhie, Alessandro Pastore, Hailei Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, Reanne Bowlby, Hui Shen, Sikander Hayat, Robert Fieldhouse, Susan C. Lester, Gary M.K. Tse, Rachel E. Factor, Laura C. Collins, Kimberly H. Allison, Yunn-Yi Chen, Kristen Jensen, Nicole B. Johnson, Steffi Oesterreich, Gordon B. Mills, Andrew D. Cherniack, Gordon Robertson, Christopher Benz, Chris Sander, Peter W. Laird, Katherine A. Hoadley, Tari A. King, TCGA Research Network, and Charles M. Perou

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.