Multi-omic Dissection of Oncogenically Active Epiproteomes Identifies Drivers of Proliferative and Invasive Breast Tumors

John A. Wrobel, Ling Xie, Li Wang, ..., Jian Jin, Michael L. Gatza, Xian Chen
xianc@email.unc.edu

HIGHLIGHTS
ChaC dissects tumor heterogeneity for identifying oncogenic-active proteins
An oncogenic-active G9a-interactome represents the invasive tumor in a tissue
iC-MAP identifies multi-omics aberrations that drive invasive tumors
Patient-specific iC-MAP of select interactor genes are of prognostic value
Multi-omic Dissection of Oncogenically Active Epiproteomes Identifies Drivers of Proliferative and Invasive Breast Tumors

John A. Wrobel,1,2,7 Ling Xie,1,2,7 Li Wang,1,2,7 Cui Liu,1 Naim Rashid,2,3 Kristalyn K. Gallagher,4 Yan Xiong,1,2 Kyle D. Konze,5 Jian Jin,5 Michael L. Gatza,6 and Xian Chen1,2,8,*

SUMMARY

Proliferative and invasive breast tumors evolve heterogeneously in individual patients, posing significant challenges in identifying new druggable targets for precision, effective therapy. Here we present a functional multi-omics method, interaction-Correlated Multi-omic Aberration Patterning (iC-MAP), which dissects intra-tumor heterogeneity and identifies in situ the oncogenic consequences of multi-omics aberrations that drive proliferative and invasive tumors. First, we perform chromatin activity-based chemoproteomics (ChaC) experiments on breast cancer (BC) patient tissues to identify genetic/transcriptomic alterations that manifest as oncogenically active proteins. ChaC employs a biotinylated small molecule probe that specifically binds to the oncogenically active histone methyltransferase G9a, enabling sorting/enrichment of a G9a-interacting protein complex that represents the predominant BC subtype in a tissue. Second, using patient transcriptomic/genomic data, we retrospectively identified some G9a interactor-encoding genes that showed individualized iC-MAP. Our iC-MAP findings represent both new diagnostic/prognostic markers to identify patient subsets with incurable metastatic disease and targets to create individualized therapeutic strategies.

INTRODUCTION

Proliferative and invasive tumors evolve heterogeneously in a tissue microenvironment (Koren and Bentires-Alj, 2015). Such heterogeneity contributes to patient variability in rates of tumor growth, proliferation, metastasis, and susceptibility to anti-cancer therapies (Yates et al., 2015); critically, tumor heterogeneity can mislead diagnosis and treatment. Next-generation sequencing has identified cancer-related genetic alterations (Garnett et al., 2012; Green and Guyer, 2011) that, however, showed low congruence with disease prognosis or diagnosis (Torga and Pienta, 2017). At the transcriptomic level, a 50-gene expression pattern has been created to classify distinct subtypes of breast cancer (BC; PAM50-subtypes), including basal-like/triple-negative (BLBC/TNBC), luminal-A and -B, Her2+, and normal-like BC (Parker et al., 2009). However, these gene-expression signatures are insufficient to discriminate, within single PAM50-subtypes, BC patient subpopulations having different clinical outcomes, particularly subpopulations with highly proliferative and invasive tumors and poor prognosis.

Yates et al. used multi-region genome sequencing to reveal genetic aberrations that define subclonal heterogeneity across different isolates from a single tumor (Yates et al., 2015). To dissect this intra-tumor heterogeneity, substantial efforts, such as xenotransplantation amplification of tumor fractions or genome-wide screening of huge numbers of samples are necessarily employed to identify particular aberrations driving proliferative or invasive subclones (Ng et al., 2016). Gatza et al. introduced an integrated genomics method that identifies drivers of proliferative tumors in luminal BC subtype (Gatza et al., 2014). This approach identifies copy-number aberrations of tumor-essential genes whose subtypic expression patterns correlate with oncogenic pathway activity. A genome-wide RNAi screen further validated these driver genes that are essential for cell viability in an oncogenic pathway-dependent manner. However, these RNAi data that indicated gene-specific effects on BC cell viability were obtained from immortalized cell lines; hence, the data have limited clinicopathological accuracy.

Proteins are the actual disease executors; thus, mass spectrometry (MS)-based quantitative proteomics is widely used to analyze genome-wide, differentially expressed proteins in tumor versus healthy cells. However, the limited sensitivity of MS instruments often yields bulk measurements averaged over different cell
types in a tissue, obscuring protein changes that precisely define a tumor phenotype. Some researchers employ laser-capture microdissection (LCM) to enrich tumor cells, followed by MS to define the tumor-specific proteome (Liu et al., 2014b). However, the LCM-MS technique is not amenable to routine clinic sampling because processing of small, heterogeneous biospecimens is expensive, low-throughput, and plagued by sample-to-sample variability (Liu et al., 2014b).

Evidently, chromatin state aberrations that stem from complex interactions between gene susceptibility and environmental perturbation drive the heterogeneous evolution of proliferative and invasive tumors (Baker, 2011; Easwaran et al., 2014; Helin and Dhanak, 2013; Wee et al., 2014). During tumorigenesis, the chromatin states associated with particular genes are dynamically regulated in a cell-type-specific manner via interactions with various transcription factors, histone modifiers, chromatin regulators, and chromatin-remodeling enzymes that, together, constitute epigenetic regulatory proteomes (epiproteomes) (Johnson and Dent, 2013). To identify the constituents of these tumor-phenotypic epiproteomes in situ, we applied our chromatin activity-based chemoproteomics (ChaC) method (Liu et al., 2014a), a breakthrough functional proteomic technology. ChaC employs UNC0965 (Konze et al., 2014), a biotin-tagged small molecule (Vedadi et al., 2011) that specifically binds the enzymatically active form of the histone lysine methyltransferases G9a and G9a-like protein (GLP). The enzymatic activity of G9a is directly correlated with its oncogenic function in tumor cells (Vedadi et al., 2011; Wozniak et al., 2007), wherein G9a interacts with specific chromatin proteins to drive tumorigenesis (Tu et al., 2018) and metastasis (Dong et al., 2012; Si et al., 2015). Until now, and primarily by clinically incompatible co-transfection and co-immunoprecipitation (IP) approaches, only a few G9a interactors have been characterized, one-gene/protein-at-a-time, for their roles in cancer (Maier et al., 2015; Si et al., 2015). Furthermore, conventional antibody-based immunoprecipitation (IP)-MS approaches (Huttlin et al., 2015; Malovannaya et al., 2011), which capture protein complexes based on epitope abundance, cannot discriminate IP complexes from tumor and non-malignant cells in a tissue. Our G9a ChaC probe UNC0965 is superior to any antibody-dependent, abundance-based IP approaches because it enables a specific, one-step sorting of the protein complexes associated with oncogenically active G9a/GLP separate from less enzymatically active G9a in other cell types in a tissue, especially non-malignant cells. ChaC-MS analysis revealed that the compositions of UNC0965-captured epiproteomes represent the proliferative or invasive potential of a patient tissue. For the first time, with higher sensitivity, specificity, and reproducibility, and directly from clinical specimens, ChaC identified multiple G9a-interacting proteins that function in concert as candidate tumor-phenotypic determinants or drivers.

Despite this advance, current search engines for MS data-dependent protein identification use only one individual's genomic sequence as Ref-seq, so that information about individual patient's (individualized) protein aberrations is missing for all ChaC MS-identified G9a/GLP interactors. Conversely, genome-wide studies of two large BC patient cohorts, TCGA (Ciriello et al., 2015) and METABRIC (Molecular Taxonomy of BC International Consortium) (Curtis et al., 2012), have correlated individualized transcriptomic/genomic aberrations with clinical/pathological data from >2,600 patients with BC. To identify which G9a/GLP interactors possess the characteristics of the individual patients with proliferative or invasive tumors, we retrospectively analyzed (Wang et al., 2018) both TCGA and METABRIC data to identify individualized aberrations of genes that encode luminal- or BLBC-specific G9a/GLP interactors. We then identified particular G9a/GLP interactor genes that showed individualized interaction-Correlated Multi-omic Aberration Patterning, or iC-MAP, in which the G9a/GLP affinities of these UNC0965-captured proteins correlate with genomic or/and transcriptomic aberrations of their encoding genes. Kaplan-Meier (KM) survival analyses further distinguished, from >18,000 genes, patient mRNA overexpression patterns or copy-number aberrations of those interactor genes that are of prognostic value. As a result, particular interactor genes that were amplified uniquely in highly proliferative luminal patients or patients with invasive BLBC/TNBC were identified as determinants or drivers of proliferation or invasion based on their G9a/GLP affinity-correlated alterations in mRNA expression, DNA copy numbers, and/or proliferation score in patient subsets with poor prognosis. This ChaC-based iC-MAP paradigm is generally applicable to dissect real-time heterogeneity of any tumor types and identify in situ tumor-phenotypic drivers (proteins) that represent individualized diagnostic/prognostic markers and new druggable targets.

RESULTS

iC-MAP Identifies Drivers of Proliferative and Invasive Breast Tumor

As shown in Figure 1A, ChaC with UNC0965 sorts in situ and identifies candidate drivers in their oncogenically active states, i.e., within G9a/GLP-interacting epiproteomes, directly from patient tissues without
We compared multiple BC cell lines and primary tissues of similar PAM50 subtypes to determine the specificity of UNC0965 (Konze et al., 2014) in dissecting the intra-tumor heterogeneity, i.e., sorting an oncogenically active G9a/GLP-interacting epiproteome from a heterogeneous tissue. Following label-free quantitation (LFQ)-based MS characterization of the UNC0965-captured epiproteomes, we applied the Perseus method (Tyanova et al., 2016) that has built-in multiple testing to determine statistically significant LFQ ratios (FDR < 0.05); PAM50-subtypic G9a/GLP interactors are those that show quantitatively correlated LFQ ratios in protein affinities to G9a/GLP regardless of the prevalence of a phenotype in a heterogeneous tissue or in a homogeneous cell line. Next, we used TCGA (Ciriello et al., 2015) and METABRIC patient databases to retrospectively...

Figure 1. Dissection of the Tumor-phenotypic Heterogeneity by iC-MAP
(A) Schematic iC-MAP design. “CNA” refers to copy-number amplifications.
(B) UNC0965-precipitation-PCR shows that, in different BC subtypes, G9a/GLP is differentially enriched or activated in the chromatin associated with tumor suppressor genes DSC3 (left) and maspin (right). UNC0965 pulls down the enzymatically active G9a/GLP with associated DNA. qPCR indicates the differentially active G9a in different BC subtypes and the G9a methylation activity is gene specific.
(C) Immunoblotting analysis of G9a abundance in the protein mixture captured by UNC0965 or the G9a-negative probe UNC125 in the adjacent non-malignant (“N”) tissue versus tumor tissues of either luminal (Lum “LP”) or basal (BLBC “BP”) patients.
(D) Gene-specific G9a/GLP epiproteomes are differently assembled in different BC PAM50 subtypes. BCC, breast cancer cells; BCT, breast cancer tissues. (Left) A plot of basal/luminal LFQ ratios of UNC0965-captured proteins from the BCCs versus BCTs of similar PAM50 subtypes. (Middle and right) The volcano plots of the interactors identified in a series of BCC lines and BCTs showing either BLBC/basal- or luminal-specific UNC0965/G9a affinities correlated with the log2 LFQ ratios (basal-like/luminal) normalized against the ratios from adjacent non-malignant tissues. Proteins that show log2 LFQ ratios > 1 or < -1 and -log10 p value > 1 are putative BC subtypic interactors of statistical significance.
See also Figures S1 and S2; Tables S1–S4.
identify personal proteo-transcriptomic and proteo-genomic links (Wang et al., 2018) for the genes PAM50-subtypic G9a/GLP interactor genes. PAM50-subtypic drivers of proliferative or invasive breast tumors were identified as interactor genes that showed correlated multi-omics aberrations in, e.g., mRNA expression, DNA copy numbers, and/or proliferation score in patient subsets with poor prognosis.

The G9a ChaC Probe Captures BC Subtypic G9a Interactors

We first determined the genomic occupancy of the G9a/GLP-interacting epiproteomes in different BC PAM50 subtypes. We used UNC0965 (Konze et al., 2014) immobilized on NeutrAvidin agarose beads to capture the G9a epiproteomes and their bound DNA elements from non-malignant (HMEC or MCF10A), the PAM50 luminal-subtype (T47D or MCF7), and the basal-like subtype (MDA231) BLBC cell lines. We did not detect any differences for capture using a G9a/GLP-negative probe (UNC125). However, compared with non-malignant mammary cells, PCR amplification of the UNC0965-captured genes revealed that G9a was specifically enriched within the chromatin of tumor suppressor genes, such as DSC3 and maspin, in the malignant/tumor cells, and more so in the basal subtype (Figure 1B). Thus, the methylation/enzymatic activity of G9a/GLP differs within the chromatin of genes that show differential expression in either non-malignant versus malignant cells or in different BC subtypes. This result agrees with previous findings. DSC3 is an anti-metastatic tumor suppressor with significantly diminished expression in breast tumor cells (Oshiro et al., 2005), as its associated chromatin had increased CpG DNA methylation and H3K9me2 catalyzed by the heterodimer G9a/GLP complex (Vedadi et al., 2011; Wozniak et al., 2007). Maspin inhibits invasiveness and motility of mammary carcinoma (Wu et al., 2010). The maspin gene also carried pronounced H3K9me2 in MDA231 but not in the non-malignant HuMEC (Wozniak et al., 2007). These results indicated that (1) in malignant or invasive BC subtype, G9a/GLP is more oncogenically active or more enriched in the transcriptionally silenced chromatin of tumor suppressor genes and (2) UNC0965 can pull down the oncogenically active G9a/GLP and associated protein complexes specifically from the chromatin harboring aberrantly silenced genes in distinct tumor phenotypes; ignored are the same genes in other chromatin states in non-malignant cells or in other cell types.

We then compared the composition of the UNC0965-captured epiproteomic complexes from a cell line series containing homogeneous representatives of nonmalignant mammary epithelial cells (HuMEC or MCF10A), PAM50-luminal (BT474, T47D, and MCF7), and PAM50-BLBC (MDA231, HCC1806, and SUM159) subtypes. To ensure that our ChaC scheme is readily applicable to any clinical specimen, we coupled on-beads sampling/processing with LFQ, so that all ChaC experiments were performed on a clinical sample scale of 60–100 µg total protein mass per nLC-MS/MS run. Following repeated washing to remove non-specific proteins, and on-beads tryptic digestion, we performed nLC-MS/MS experiments on each UNC0965 pull-down for two to three biological replicates, each with three technical replicate runs. In each UNC0965 pull-down from a PAM50 subtype, G9a and GLP were unambiguously identified by 30–50 unique MS/MS-sequenced peptides. To distinguish genuine G9a-interacting epiproteomic components, we filtered non-specific associates by comparing the proteins captured by UNC0965 with proteins acquired in UNC125 or mock pull-downs (empty NeutrAvidin beads). The LFQ ratios (peak intensities) of proteins in the UNC0965-captured epiproteomes from any paired cell line correlated with the relative G9a/GLP affinities of these proteins in these cell types. Based on their affinity-correlated LFQ ratios, we identified major clusters of G9a/GLP-associated proteins that had significantly increased affinities to UNC0965 and no binding to empty beads (Figure S1A and Table S1). With respect to their associations with UNC0965 in the non-malignant cells, the majority of UNC0965-captured proteins showed quantitatively correlated G9a/GLP affinities in similar luminal- or BLBC-subtypic cell lines, i.e., the LFQ ratios of these proteins were correlated in MCF7 versus T47D or in MDA-MB-231 versus SUM159 (Pearson correlation coefficient 0.73), indicating that the G9a/GLP-interacting epiproteomes are assembled in a BC-subtypic manner (Figure S1B, Table S2). Among these ChaC-identified proteins, we found many known components of multiple epiproteomic complexes (Ooi and Wood, 2007, Shi et al., 2003, Wilson and Roberts, 2011, Ahringer, 2000). Also, in agreement with a previous report (Si et al., 2015), we found that GATA3 enhanced its interaction with G9a in luminal BC cells, whereas MTA2 was associated with G9a more specifically in BLBC cells. By identifying these chromatin regulatory complexes known to be associated with G9a, we demonstrated the specificity of UNC0965 in capturing endogenous G9a epiproteomes assembled into gene-specific, oncogenically silenced chromatin characteristic of different BC subtypes. Technically, the gene specificity of UNC0965-captured epiproteomic complexes renders ChaC superior to any abundance-based IP-MS approaches (Huttlin et al., 2015; Malovannaya et al., 2011).
ChaC-MS Identifies In Situ Oncogenic-active G9a Interactors Critical for Tumor Cell Viability

To confirm the unique ability of ChaC to sort and characterize a pathologically relevant, G9a-associated epiproteome whose composition defines the predominant tumor phenotype in a heterogeneous tissue, we performed similar cross-referencing LFQ experiments (Wang et al., 2018) with ChaC and various PAM50-subtypes from different sample types, i.e., cell lines and tissues. These samples included the aforementioned cell line series plus four frozen clinical specimens including tumor tissues with molecular classifications of similar PAM50 subtypes (and adjacent non-malignant tissue).

In the UNC0965-captured protein complexes we observed higher G9a abundance in luminal or basal patients, whereas no G9a was detected by immunoblotting against G9a antibody in either adjacent non-malignant tissue or the protein mixtures that were pulled down by the G9a-negative probe UNC125 (Figure 1C). These results confirmed that UNC0965 pulls down specifically G9a/GLP and its activity-based interactors in the proliferative or the invasive tumor cells, avoiding other cell types or non-malignant cells in which G9a is less enzymatically active.

LFQ ratios are proportional to the G9a/GLP affinities of the captured proteins. Thus, the UNC0965-captured proteins that show cross-referenced LFQ ratios among cell lines versus tissues, or among tissues of similar PAM50 subtypes (Wang et al., 2018), should be the G9a/GLP-interacting proteins characteristic of tumor cells that are the predominant population. Accordingly, we used the Perseus method (Tyanova et al., 2016) to analyze the PAM50-subtypic compositions of UNC0965-captured epiproteomes. Among 294 proteins in common in all cell lines (n = 6) (Table S3) and tissues (n = 4) (Table S4, Figure S2A), with three liquid chromatography-tandem mass spectrometry (LC-MS/MS) technical replicates for each of two biological replicates, 154 UNC0965-captured proteins showed similar, statistically significant G9a/GLP affinities in the tissues versus the cell lines of similar PAM50 subtypes, with respect to either non-malignant MCF10A or non-malignant tissues. As indicated by the LFQ ratios that correlated across both cell line and tissue origin (Pearson correlation coefficient at 0.66) (Figure 1D, left), one luminal-specific cluster of eighty-four proteins (Figure 1D, left, lower-left quadrant) showed increased UNC0965 affinity specifically in the luminal subtype but decreased affinity in the BLBC subtype; another BLBC-specific cluster of seventy proteins showed an opposite UNC0965-binding pattern (Figure 1D, left, upper-right quadrant). In a high confidence with false discovery rate of <5%, we identified fifty-eight luminal (Figure 1D, left, blue dots) and forty-four BLBC (Figure 1D, left, red dots) G9a/GLP interactors. Thus, the PAM50-subtypic G9a affinities of many epiproteomic components were consistently preserved, representing the predominant tumor phenotype, with minimum sample-to-sample variability, in both homogeneous cell lines and heterogeneous clinical tissues. Also, by LFQ-assisted LC-MS/MS, the compositions of G9a/GLP interactomes were identified with PAM50-subtypic characteristics, i.e., G9a/GLP interacts with subtype-unique chromatin proteins in different BC subtypes.

Bypassing tedious LCM tumor cell sorting or xenotransplantation, UNC0965 can sort/enrich in situ the oncogenically active G9a/GLP interactors that represent a single, predominant tumor phenotype in a heterogeneous tissue. Thus, ChaC exhibits high pathological accuracy in identifying candidate oncogenic drivers. Based on statistically significant LFQ ratios that correlated with increased UNC0965 binding of MS-identified proteins, we identified 138 BLBC/basal and 129 luminal G9a interactors (Figure 1D). Functional category/network analysis revealed that these BC-subtypic G9a interactors over-represented major pathways/sub-networks associated with RNA processing, translation initiation and elongation, ribosome biogenesis, RNA splicing, and RNA metabolism, consistent with the view that these interactors are candidate drivers of tumor growth and proliferation/invasion (Figure S2). Lastly, results of immunoblotting experiments were consistent with MS/MS experiments, i.e., we compared the level of UNC0965-captured G9a/GLP from different BC subtype cell lines versus HuMEC. UNC0965 pull-downs from equal amounts of each cell line had higher amounts of G9a(EHMT2)/GLP(EHMT1) in either the basal (long or short forms of G9a/GLP) or the luminal subtypes compared with HuMEC cells for which no G9a/GLP was detected (Figure S2B). Also, no UNC0965/G9a-interacting proteins were detected in the UNC125-captured protein mixtures.

Identification of Oncogenic-active G9a Interactors with Multi-omics Aberrations

Using TCGA (Ciriello et al., 2015) and METABRIC (Curtis et al., 2012; Pereira et al., 2016) databases, we analyzed the individualized frequency of various types of patient-specific aberrations of genes that encode PAM50-subtypic G9a interactors. These datasets contain two large independent cohorts of >2,600 patients with five subtypes classified by PAM50 (Parker et al., 2009) and information about mutations, copy-number
aberrations, mRNA expression, and associated clinical/pathological data (stages/grades and relapse status). The TCGA project analyzed 816, mainly American, patients with BC with 590 luminal A/B subtypes and 136 BLBC/TNBC subtypes. The METABRIC database contains 1866 Canadian and European patients with BC with 1,127 luminal A/B subtypes and 198 BLBC/TNBC subtypes.

First, we found that the mutation frequencies for most interactor genes were extremely low in the TCGA patient pool, i.e., less than 5% of patients with BC had somatic mutations in G9a interactor genes, except for **GATA3** (Table S5). Clearly, somatic mutations had little impact on the oncogenic activities of G9a interactors.

Next, we compared the mRNA expression of interactor encoding genes (downloaded as Z score values from the cBioPortal for Cancer Genomics: [http://www.cbioportal.org/](http://www.cbioportal.org/)) between BLBC/basal and luminal A/B TCGA patients (Ciriello et al., 2015) by performing a Mann-Whitney-Wilcoxon Test to judge the expression differences between the two PAM50 subtypic populations. This test was performed on all genes in the dataset for comparisons, and we used the Benjamini Hochberg procedure to adjust p values for multiple testing. By this multi-testing scheme, an interactor was classified as BLBC/basal if its expression level was
Figure 3. Genes that Encode PAM50-subtypic G9a Interactors Show Statistically Significant Overexpressed mRNA

The expression values for the PAM50-subtypic G9a interactor genes were obtained from the cBioPortal (mRNA expression Z scores compared with diploid tumors) for all TCGA BRCA patients.

(A) Heatmap for G9a interactor genes (rows) with mRNA expression from the TCGA sample set (columns). The TCGA samples and G9a interactor genes were clustered using unsupervised hierarchical clustering ("euclidean" distance and "ward" clustering method). The columns are annotated with PAM50 subtype of each TCGA patient. The rows are annotated with the proteomic subtype of G9a interactors.

(B and C) Box plots showing the statistically significant altered mRNA expression (x axis) for a sample set of basal G9a interactor genes (B) and a sample set of luminal G9a interactor genes (C). For each interactor gene, the distribution of mRNA expression for PAM50 basal-like TCGA patients (N = 136) is shown on the left in red, and for PAM50 luminal A and B patients (N = 560) it is shown on the right in blue. Outliers are indicated as circles. The median is indicated by the black line inside each box. The mean is indicated by the cyan diamond. The mRNA expression as a Z score is displayed on the y axis. A Mann-Whitney-Wilcoxon Test was performed on each pair to judge differences in expression levels.
significantly greater (adjusted p value < 0.05) for the basal patients. Likewise, an interactor was classified as luminal if its mRNA expression level was statistically greater for the luminal patients (adjusted p value < 0.05).

From a systems view, hierarchical clustering resulted in patient-specific mRNA expression patterns for all PAM50-subtypic G9a interactors in TCGA patients (Figure 3A). We found a similar statistically significant interaction-correlated expression pattern (iCEP) (Wang et al., 2018) wherein the PAM50-subtypic G9a/GLP binding of some UNC0965-captured proteins showed cis-mRNA expression of the genes encoding these interactors in patients with the corresponding PAM50 subtypes (Table S6). Briefly, 94 of 138 BLBC/basal G9a interactors showed basal iCEPs in the TCGA patients, including 77 interactors (82%) that displayed basal iCEPs in both TCGA and METABRIC patients (Figures 3B and S3). Conversely, among 129 luminal G9a interactors, only twenty-nine showed iCEPs across various luminal subtypes (Figures 3C and S4), whereas sixty-eight had higher expression in patients with BLBC. Notably, it is not surprising to observe fewer luminal interactor genes with statistically significant iCEPs because, compared with BLBC/basal subtypes, there are more diverse receptor-based luminal subtypes, including luminal-A and luminal-B. Nevertheless, the patient mRNA overexpression pattern of G9a/GLP interactor genes automatically clustered patients with BC (TCGA and METABRIC) based on the PAM50-classified subtypes for which these interactors were identified, confirming the pathological accuracy of UNC0965 in sorting the predominant tumor phenotype in a tissue. Importantly, our iCEP findings demonstrated that ChaC discovers the interactor-encoding genes as new BC-subtypic classifiers, i.e., these genes are fully translated into oncogenically active G9a/GLP interactors in a PAM50-subtypic manner.

We then examined the PAM50 subtype distribution of copy-number alterations in the G9a interactor genes in the TCGA patient sets. Particularly, we performed a correlation analysis comparing increased mRNA expression with copy-number amplification of all G9a interactor genes. The mRNA expression (as Z score values) and copy-number (as GISTIC values) data were downloaded from the cBioPortal for Cancer Genomics (http://www.cbioportal.org/) (Cerami et al., 2012; Gao et al., 2013). Based on copy-number status, we separated the TCGA samples into a deletion/diploid group (GISTIC values of 0, -1, or -2) and a gain/amplification group (GISTIC values of 1 or 2). We performed a Mann-Whitney-Wilcoxon Test to judge whether the mRNA expression values were greater for the gain group compared with the deletion/diploid group. This test was performed on over 18,000 genes (with both mRNA expression and GISTIC data), and p values were adjusted for multiple testing using the Benjamini Hochberg procedure. We identified 110 luminal and 117 BLBC/basal interactor genes that showed significant correlations between mRNA overexpression and copy-number gain/amplification in TCGA luminal A/B (n = 590) or basal samples (n = 136) (adjusted p value < 0.05, Table S7).

Identification of Tumor-proliferative G9a Interactors with Multi-omics Aberrations

Because proliferation is a luminal BC characteristic, we next searched for G9a interactors bearing mRNA-expression or DNA copy-number aberrations that are characteristic of proliferative tumors. Using the proliferation scores of individual patients recruited for the TCGA BRCA study (Ciriello et al., 2015) we performed a correlation analysis of PAM50-subtypic interactors, comparing mRNA expression with the tumor proliferation scores for luminal A/B (n = 590) and basal patients (n = 136). We separated the TCGA samples into high-proliferation-score (top 25%) and low-proliferation-score (bottom 75%) groups. A Mann-Whitney-Wilcoxon Test was used to judge whether the mRNA expression values were greater for the high-proliferation-score group compared with the low-proliferation-score group. This test was performed on over 18,000 genes, and p values were adjusted for multiple testing using the Benjamin Hochberg procedure. We identified forty-five luminal and twenty-three basal interactor genes that showed significant correlations between increased mRNA expression and a high proliferation score (adjusted p value < 0.05, Table S8).
expression (Figures 5B, 6B, and S9–S11, Tables S11 and S12). mRNA overexpression of individual genes have poor survival for METABRIC patients with correlations between CNA and mRNA overexpression, which indicate that these genes are essential genes of proliferative luminal breast tumors (Gatza et al., 2014). To identify new driver genes that are amplified uniquely in proliferative luminal tumors, we investigated which interactor genes are located in genomic regions with increased amplification frequency and have a coordinate increase in mRNA expression in patients with poor prognosis. For example, seven luminal interactors with multiple correlations (Figures 5A and S6) have poor overall survival. mRNA overexpression of individual genes with proliferation-multi-omics correlations (Figures 5A and S6) have poor survival. EIF6 and EIF6 are essential genes of proliferative luminal breast tumors (Gatza et al., 2014). To identify new driver genes that are amplified uniquely in proliferative luminal tumors, we investigated which interactor genes are located in genomic regions with increased amplification frequency and have a coordinate increase in mRNA expression in patients with poor prognosis. For example, seven luminal interactors with multiple correlations (Figures 5A and S6) were found on the q arm of chromosome 8 with copy-number amplification frequencies of 51%–56% (Figure 4B, Table S10). Specifically, C8orf33, EIF2S2, EIF6, and RPL8 are located at 8q24.3, RPL7 and STA2U at 8q21.11, RRS1 at 8q13.1, and RPL30 at 8q22.2. RPL7 and STA2U are near each other, separated by one gene (Figure S5). Patients with copy-number amplifications for all of these individual genes with proliferation-multi-omics correlations (Figures 5A and S6) have poor overall survival in TCGA and/or METABRIC (p values ranging from 0.001 to 0.068) (Figures 6A and S7, Table S11). Patients with STA2U mRNA overexpression in both TCGA and METABRIC (Figure 6A) have poor survival. mRNA overexpression of C8orf33 in TCGA (no data in METABRIC) indicates poor survival. Patients with mRNA overexpression of RRS1, RPL8, and BOP1 have poor survival in METABRIC (Figure S8, Table S12).

Drivers of Proliferative Luminal Tumors Have Multi-omics Aberrations in Patients with Poor Prognosis

Because patients with proliferative or invasive tumors often have poor prognosis, we performed the KM survival analysis to determine prognosis for patients whose luminal G9a interactor genes showed both CNA (a GISTIC score of 1 or 2) and mRNA overexpression (a Z score of mRNA expression in the top 50%) that correlated with high proliferation scores. The statistically significant overall survival was determined by a p value and hazard ratio. We identified (Figure 4A) twenty-nine luminal interactor genes with proliferation-correlated CNA and mRNA over-expression. Among these genes, C8orf33, EIF2S2, and EIF6 were also found in the RNAi dataset of genes essential for luminal cell line viability. EIF2S2 and EIF6 are essential genes of proliferative luminal breast tumors (Gatza et al., 2014). To identify new driver genes that are amplified uniquely in proliferative luminal tumors, we investigated which interactor genes are located in genomic regions with increased amplification frequency and have a coordinate increase in mRNA expression in patients with poor prognosis. For example, seven luminal interactors with multiple correlations (Figures 5A and S6) were found on the q arm of chromosome 8 with copy-number amplification frequencies of 51%–56% (Figure 4B, Table S10). Specifically, C8orf33, BOP1, and RPL8 are located at 8q24.3, RPL7 and STA2U at 8q21.11, RRS1 at 8q13.1, and RPL30 at 8q22.2. RPL7 and STA2U are near each other, separated by one gene (Figure S5). Patients with copy-number amplifications for all of these individual genes with proliferation-multi-omics correlations (Figures 5A and S6) have poor overall survival in TCGA and/or METABRIC (p values ranging from 0.001 to 0.068) (Figures 6A and S7, Table S11). Patients with STA2U mRNA overexpression in both TCGA and METABRIC (Figure 6A) have poor survival. mRNA overexpression of C8orf33 in TCGA (no data in METABRIC) indicates poor survival. Patients with mRNA overexpression of RRS1, RPL8, and BOP1 have poor survival in METABRIC (Figure S8, Table S12).

EIF2S2 and EIF6 are located at 20q11.22 with copy-number amplification frequencies of 40% and 43%, whereas STAU1 is located at 20q13.13 with an amplification frequency of 46% (Figure 4C, Table S10). Of particular interest, both EIF2S2 and EIF6 are essential genes of proliferative luminal breast tumors (Gatza et al., 2014). These individual genes have poor survival for METABRIC patients with correlations between CNA and mRNA overexpression (Figures 5B, 6B, and S9–S11, Tables S11 and S12). mRNA overexpression of EIF2S2 also showed a statistically significant (p value = 0.046) poor prognosis for TCGA patients. Also, METABRIC patients with poor prognosis associated with copy-number amplifications (p value ranging from 0.049 to 0.059) possessed three luminal interactor genes (RPL38, LLGL2, P4HB) that had all correlations and were located at 17q25 with CNA frequencies of 35%–36% (Figures 4D, 5C, 6C, and S12–S14, Tables S10, S11, and S12). In addition, poor prognosis was associated with LLGL2 copy-number amplifications in TCGA patients (p value = 0.068) and mRNA overexpression in METABRIC patients (p value = 0.011). Four luminal interactor genes (DDX23, ZNF385A, WIBG, RPS26) located at 12q13, when amplified, showed poor survival in both TCGA and METABRIC patients (p values < 0.001, with the exception of WIBG, which has no data in METABRIC) (Figures 4E, 5D, and S15–S17, Tables S10, S11, and S12).

Crucially, the functions of some interactor genes have been individually linked to tumor survival and invasion. BOP1 was implicated in dysregulated ribosome biogenesis that promotes metastatic breast cancer cells to the
Drivers of Invasive Tumors Show Multi-omics Aberrations in Patients with Poor-prognosis BLBC

Distinct from most of luminal G9a interactors, the majority of BLBC/basal interactor genes showed iCEP, which indicated strong correlations between the oncogenic consequences represented by enhanced G9a binding and their mRNA overexpression in basal patients (Figures 3 and S3). Figure 7A summarizes the distribution of basal interactors with correlations in iCEP, and/or mRNA overexpression, CNA, proliferation score; and/or a high CNA frequency (>= 40%). For example, fifty-one basal G9a interactors were found with high percentage of CNAs (Table S10). The chromosomal location of BLBC/basal G9a interactor genes in each panel was found in basal patient subsets with poor prognosis (Figure S19).

Among the interactors that showed at least two correlations, we searched for new invasion driver genes with proliferation-correlated multi-omics aberrations in basal patients with poor prognosis. Unlike luminal G9a luminal interactors, only a few basal G9a interactors individually were associated with poor prognosis (Tables S13 and S14). For example, in the overall survival (OS) analysis of the TCGA (Ciriello et al., 2015) (Figures S18A and S18B) and METABRIC (Curtis et al., 2012) (Figure S8B, p 0.288) patient data, overexpression of G9a (EHMT2) alone was not significantly prognostic. We therefore performed KM survival analysis to identify which combinations of multiple basal G9a interactor genes showed co-aberrations in mRNA expression and/or CNAs in basal patients having invasive tumors. We defined the altered group as patients with an mRNA expression Z score in the top 50% for all the genes in the combination. For this analysis, we examined basal patient samples from both TCGA (n = 136) and METABRIC (n = 198).

First, co-overexpression of G9a and its interactor genes PRPF6, XRN2, and YBX3 (Figures 8C and 8D) or DDX50, ILF2, and SURF6 (Figures 8E and 8F) was observed in both TCGA and METABRIC basal patient subsets with poor survival. XRN2(20p11.22) and PRPF6(20q13.33), located on different arms of chromosome 20, showed an mRNA expression versus copy-number correlation (Table S7) with a high CNA frequency (Table S10). Overexpression of PRPF6 has a role in colorectal cancer tumor growth by promoting alternative splicing of genes involved in cell proliferation (Adler et al., 2014). Copy-number amplifications of PRPF6 correlated with this overexpression of PRPF6 in colon tumors. DDX50, ILF2, and SURF6 are basal iCEP genes (Figure S3 and Table S6) showing a correlation between mRNA overexpression and CNA (Table S7).

Also, BLBC/basal patients with the worst prognosis possessed co-overexpression of multiple interactor genes with highly frequent CNA that correlated with mRNA overexpression, including HNRNPF, ILF3, and MRPL39 (Figures S18A and S18B); EHMT1, EHMT2, and ILF3 (Figures S18C and S18D); and DDX56, FBL, and HNRNPF (Figures S18E and S18F). HNRNPF and MRPL39 also showed an mRNA overexpression versus proliferation correlation. Each of these genes in a prognostically significant combination is located on different chromosomes. HNRNPF has a role in cell proliferation (Goh et al., 2010) and is a potential biomarker for colorectal cancer progression (Balasubramani et al., 2004). The transcription factor ILF3 promotes sustained urokinase-type plasminogen activator expression in metastatic breast cancer cells (Hu et al., 2013). Furthermore, we identified ten panels of multiple G9a interactor genes that showed at least two correlations (Figure 7A) and co-overexpression of interactor genes in each panel was found in basal patient subsets with poor prognosis (Figure S19).
Figure 6. Survival Analysis of Proliferative Luminal Drivers

Example Kaplan-Meier survival analysis plots for luminal G9a interactor genes in the following rows: (A) STAU2, (B) EIF2S2, (C) LLGL2, (D) DDX23. Columns 1 and 2 show Kaplan-Meier survival analysis plots for luminal samples with copy-number amplification (GISTIC score of 1 or 2) (red line) versus copy-number deletion or diploid (GISTIC score of 0, -1, or -2) (blue line) in luminal TCGA samples (column 1) and luminal METABRIC samples (column 2). Columns 3 and 4 show Kaplan-Meier survival analysis plots for luminal samples with mRNA overexpression (Z score > median Z score) (red line) compared with samples not showing mRNA overexpression (Z score < median Z score) (blue line) in luminal TCGA samples (column 3) and luminal METABRIC samples (column 4). For the Kaplan-Meier plots, “N” refers to “Number of samples” and “NE” refers to “Number of Events.” The number of events is for OS status = “DECEASED.” “Mdn” indicates the median months survival. The log rank p value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot.

See also Figures S7, S8, S10, S11, S13, S14, S16, and S17; Tables S11 and S12.
Many basal interactor genes with iCEP are located in chromosomal regions that show high CNA frequencies in TCGA patients with BLBC (Figures 7B and 7C). Of special interest, certain pairs of G9a interactors are directly adjacent to one another in the genome (Figure S20). Particularly interesting is the clustering of iCEP interactors located at genomic loci 1q21 (MRPS21, PRPF3, SPRR1A, S100A6, ILF2, GPATCH4) and 1q23.

Figure 7. Multi-omics Correlated Identification of Drivers of Invasive Basal Tumor
(A) Venn diagram showing overlap of basal G9a interactors with mRNA versus copy number, mRNA versus proliferation, and basal iCEP correlations, and an amplification (GISTIC values of 1 or 2) frequency greater or equal to 40%.
(B) UCSC Genome Browser view showing location of basal G9a interactors (MRPS21, PRPF3, SPRR1A, S100A6, ILF2, GPATCH4) located in the 1q21-23 region of chromosome 1.
(C) UCSC Genome Browser view showing location of basal G9a interactors (NOL7, DEK, HIST1H1A, HIST1H1C, HIST1H1E, ABT1, EHMT2, HMGA1, RPS10) located in an amplified region on the p arm of chromosome 6. For each view, the frequency of basal TCGA samples with copy-number gain (segmentation mean >0.1) (red) and copy-number loss (segmentation mean <0.1) (blue) is indicated.

See also Figures S20 and S21; Tables S7, S8, S9, and S10.
Figure 8. Identification of Driver Genes Showing Co-overexpression/Co-amplification in Patients with Poor Prognosis

Kaplan-Meier survival analysis indicates that altered mRNA expression of G9a interactor genes is prognostically significant in marking distinct patient subpopulations within single PAM50 subtypes.

(A, C, and E) Overall survival (OS) plots of TCGA patients with the BLBC subtype (n = 136).

(B, D, and F) Overall survival (OS) plots of METABRIC patients with the BLBC subtype (n = 198).

For (A and B) the altered group (red line) is samples with mRNA overexpression (Z score > median Z score) for EHMT2. The non-altered group (blue line) is samples not showing mRNA overexpression (Z score < median Z score) for EHMT2. For the other panels, the altered group (red line) is samples with mRNA co-overexpression (Z score > median Z score for all genes
GPATCH4, MRPS21, and PRPF3 (1q21.2) are immediate neighbors with 89% CNA frequency (Figures 7Ba and S20A). These interactors have a consistently high CNA level in the TCGA patients compared with most other interactors (Figure S21). The 1q21-23 region is amplified uniquely in human BLBC, and it contains breast cancer driver genes (Silva et al., 2015). Also, in amplified regions of chromosome 6, a pair of interactor genes, HMGA1 and RPS10 (6p21.31), showed about a 50% CNA frequency and are separated by only two genes (Figure S20E).

Similar to luminal tumor proliferation drivers, the functions of some interactor panel genes have been individually characterized as drivers of proliferative or invasive tumor. The secreted protein CYR61 promotes BLBC/TNBC metastasis (Huang et al., 2017; Sanchez-Bailon et al., 2015). LYAR, a ribosome-associated protein, is a key regulator of the migration and invasion of human CRC cells (Wu et al., 2015) and a regulator of translation associated with cell proliferation (Yonezawa et al., 2014). Recently, using CRISPR/Cas9-mediated knockdown of LYAR in the invasive BLBC MDA-MB-231 cells, we characterized LYAR as a driver of invasive breast tumor (Wang et al., 2018). DEK promotes tumor growth, cellular motility, and invasion in breast cancer (Privette Vinnedge et al., 2011). NIFK has a role in lung cancer progression, and increased NIFK expression was associated with poor prognosis in patients with lung cancer (Lin et al., 2016). DDX21 is involved in breast cancer proliferation by promoting AP-1 activity and rRNA processing (Zhang et al., 2014). DDX3X promotes breast cell proliferation by inhibiting the expression of KLF4, a cell cycle repressor (Cannizzaro et al., 2018). Overexpression of rRNA methyl-transferase FBL causes altered rRNA methylation patterns leading to decreased translation fidelity and translation of cancer-related genes from internal ribosome initiation sites (Marcel et al., 2013). Uniquely, ChaC identified these proteins as G9a interactors in the networked pathways that drive invasive BLBC (Figure 2A).

DISCUSSION

Genome-wide studies of patients with breast cancer have generated enormous amounts of genomic and transcriptomic data. Still, sizable additional efforts, such as xenotransplantation amplification of tumor fractions and predictive score modeling (Ng et al., 2016) are required to determine which genomic/transcriptomic aberrations are of oncogenic significance. Also, invasive tumors often evolve heterogeneously, producing subclones different from the earlier specimen from which a treatment decision was made. Either because of low phenotypic accuracy of genomics and transcriptomics or low MS proteomic phenotypic coverage of individual patient’s alterations, single-omics approaches usually fail to dissect this intra-tumor heterogeneity. Thus, for optimized therapeutic decisions, we require new multi-omics methods that enable real-time diagnostic identification of drivers or druggable targets in evolving tumor tissue.

We demonstrate that ChaC is a simple, robust method enabling in situ, efficient dissection of intra-tumor heterogeneity to identify the genetic and transcriptomic alterations that manifest as oncogenically active proteins. Because it specifically binds with antibody-like affinity (IC50 < 2.5 nM) to the enzymatically (oncogenically) active form of G9a, the G9a ChaC probe UNC0965 can capture and enrich endogenous, oncogenically active G9a/GLP-interacting epiproteomes for subsequent, focused MS/MS characterization. This front-end epiproteomic enrichment avoids interfering signals from non-tumor-related proteins in other cell types, especially neighboring non-malignant cells in which G9a/GLP is less enzymatically active. Accordingly, our cross-referencing LFQ data confirmed that the G9a/GLP affinities of UNC0965-captured proteins were quantitatively correlated or conserved in similar tumor phenotypes across different sample types, from homogeneous cell lines to heterogeneous tumor tissues (Figure 1D). This quantitative consistency confirms that UNC0965 can directly sort the G9a/GLP-interacting proteins that represent the predominant tumor phenotype in the tissue microenvironment.

ChaC is a generic technique for the rapid in situ characterization of endogenous, tumor-phenotypic epigenetic regulatory protein complexes. The small molecule chemoprobe substitution for an antibody...
facilitates the integration of a simple, on-bead sample processing with LFQ LC-MS/MS characterization. Unlike antibodies, a ChaC probe does not disrupt protein complexes, and ChaC probes eliminate concerns about antibody cross-reactivity. Also, the tight association between biotin and NeutrAvidin minimizes sample loss and enables extensive washing to remove MS-incompatible contaminants, e.g., detergents, salts, and nonspecific proteins, so retained proteins can be immediately subjected to reduction, alkylation, and on-bead tryptic digestion. Because tryptic digests of large antibodies (>150 kDa) may suppress the peptide signals from low-abundance proteins, the smaller NeutrAvidin (60 kDa) is particularly suitable for on-bead digestion. Concurrently, the specificity and sensitivity of MS is fully utilized for focused analysis of the low-abundance, tumor-phenotypic G9a epiproteomes that are thereby “teased out” from thousands of other proteins in a cancer proteome, making ChaC applicable to clinical isolates of a few million “sorted” cells. Thus, ChaC eliminates tedious, clinically incompatible steps and yields rapid and precise identification of oncogenically active proteins in clinical isolates without diluting information about tumor invasiveness.

ChaC-identified G9a/GLP interactors over-represented the regulatory pathways/networks associated with tumor cell viability such as RNA processing, translation, ribosomal biogenesis, and protein synthesis, the major hallmarks of tumor growth, survival, and invasion. These findings are consistent with the fact that G9a/GLP small molecule inhibitors suppressed BC cell growth and survival in vitro (Casciello et al., 2017) and inhibited epithelial-mesenchymal transition-mediated invasion of aggressive BLBC cells (Liu et al., 2018). Thus, ChaC can measure in vivo oncogenic pathway activity by identifying in situ particular chromatin proteins with tumor-phenotypic G9a/GLP interactions, as opposed to extrapolations of pathway activity based on gene-expression analyses (Gatza et al., 2014).

Because of low MS sensitivity, only small portions of a proteome can be sequenced. Thus, quantitative profiling of global protein expression differences in TCGA samples has produced MS/MS-sequenced peptides that match less than 5% of the BC-related, genomic/transcriptomic variants (Mertins et al., 2016; Ruggles et al., 2016). This low coverage considerably limits information about individualized tumor-phenotypic alterations. Our iC-MAP method has overcome this problem by linking tumor-phenotypic data (epiproteomes) to tumor-genotypic data with comprehensive coverage of individualized alterations. In this regard, without cost- and labor-intensive sampling and genome-wide fishing exploration, akin to “finding a needle in a haystack,” ChaC pinpoints particular tumor-phenotypic driver genes with prognostically significant multi-omics alterations; these driver genes would otherwise be indistinguishable from non-tumor-related genes in transcriptomic or genetic profiles. Specifically, the retrospective population-based analyses of TCGA and METABRIC patient data revealed iC-MAP wherein select ChaC-identified G9a interactors had proliferation scores that aligned with highly frequent copy-number amplifications and/or mRNA overexpression. This connection indicates that these interactor genes are fully translated into the oncogenically active proteins in patients with proliferative or invasive tumor. Furthermore, iC-MAP enables a multi-omics dissection of the intra-tumor or interpatient heterogeneity within single PAM50 subtypes that is far more complicated than might be predicted from genetic or transcriptomic aberrations alone. iC-MAP also identified the mRNA-overexpression or CNA profiles of interactor genes that are prognostically meaningful in distinguishing the subsets of luminal or BLBC patients with distinct clinical outcomes in multiple patient databases.

Notably, the goal of the current study was not to build a “perfect” prediction model for personal prognosis but rather to identify candidate genes that serve as interacting partners to G9a that function as drivers of proliferation or invasion and which may represent therapeutically actionable targets. Thus, the prognostic capacity of these genes, while noted, serves as a filtering step and these analyses have not been subjected to multiple comparison correction. Multivariate analysis has been previously reported for EIF2S2 and EIF6 (Gatza et al., 2014) in combination with standard clinical parameters including proliferation. Given that the identified genes correlate with proliferation (Figure 4), which is probably one of the strongest known prognostic feature for breast cancers, we would not expect that these genes, or combination of genes, would appreciably add to the overall prognostic model when proliferation was included within the multivariate analysis model; this observation was true for both EIF2S2 and EIF6 in the previously cited study. However, as we have noted, many of the identified genes are co-amplified or co-expressed. As illustrated in Figure S19, accounting for co-overexpression does in fact improve the prognostic capacity of some sets of genes.

Ultimately, not all identified genes will be able to function as prognosis predictive biomarkers at both the mRNA and/or DNA copy-number level. As would be expected, differences in prognostic capacity
between mRNA and DNA copy-number status for a given gene may reflect different mechanisms by which each gene is activated/regulated in addition to DNA copy-number status. We will note, however, that the DNA copy-number status and mRNA levels of the vast majority of the identified genes were prognostic in the METABRIC study, whereas less robust results were observed in the TCGA cohort. These differences likely reflect differences in the completeness of the clinical data in each dataset—the TCGA clinical follow-up time is ~2 years (24.3 months), whereas the METABRIC dataset has a more clinically relevant 7.2-year follow-up.

In the clinic, quantitative PCR assays on our identified interactor genes could stratify patient treatment regimes, i.e., how will individual patients respond, will they develop resistance, and what are the most appropriate, optimized drug or dose? Thus, clinicians will use these prognostically meaningful biomarkers to select, within the early window of opportunity for cure, personalized therapeutic strategies with the highest likelihood of success for each luminal or BLBC subpopulation, to assess therapeutic effectiveness and the risk of disease recurrence during treatment.

In sum, conceptually innovative, our iC-MAP results also revealed that the tumor heterogeneity resulting from complex interactions between genetic susceptibility and environmental perturbation is not driven by genomic aberrations alone but, instead, by dynamic, coordinated, multi-omics alterations. Guided by our ChaC breakthrough technology that enables in situ identification of driver proteins of subclonal heterogeneity in real time, iC-MAP further characterizes drivers of evolving proliferative or invasive subclones by continual acquisition of new genetic aberrations or punctuated clonal expansion. Because patients with proliferative luminal and invasive basal tumors are the clinical subsets for which most available therapeutic options are ineffective, our iC-MAP findings represent new diagnostic/prognostic markers to identify patient subsets with metastatic disease. Our findings also form the basis for precision therapeutic strategies that are matched to the proliferative or invasive potential of individual tumors.

Limitations of the Study
To obtain statistically significant correlations for driver determination, iC-MAP requires both mRNA expression and copy-number aberration data from large patient cohorts. In addition, ChaC analysis is limited to freshly collected or frozen tissues and peripheral blood cells.

METHODS
All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION
Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.07.001.

ACKNOWLEDGMENTS
This work was supported in part by grants NC Tracs TTSA Phase I TTSA021P1, NIH 1U19AI109965, 1U24CA160035-01 from the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) (to X.C.) and R01GM122749 (to J.J.). We thank Dr. Howard Fried for editorial assistance. This invention of iC-MAP is protected by United States Provisional Patent Application Serial No. 62/608,992 that was filed by the University of North Carolina-Chapel Hill.

AUTHOR CONTRIBUTIONS
J.A.W. developed the software, performed proteogenomic analysis of clinical data, and wrote the report. L.X. and L.W. performed sample preparation and processing of some cell lines and clinical tissues for MS/MS experimental analysis, analyzed data, and conducted functional characterization. C.L. performed chem-precipitation with UNC0965 for PCR. N.R., M.L.G., and K.K.G. assisted clinicopathological data analysis. Y.X., K.D.K., and J.J. provided UNC0965 and UNC125. X.C. conceived and designed the project and experiments, analyzed and interpreted data, and wrote the manuscript.

DECLARATION OF INTERESTS
The invention of ChaC is protected by US Patent Application Serial No. 15/118,061 that was filed by the University of North Carolina-Chapel Hill.
REFERENCES

Adler, A.S., McCleland, M.L., Yee, S., Yalagadlu, M., Hussain, S., Cosimo, E., Quinones, G., Modrusan, Z., Seshagiri, S., Torres, E., et al. (2014). An integrative analysis of colon cancer identifies an essential function for PRPF6 in tumor growth. Genes Dev. 28, 1068–1084.

Ahringer, J. (2000). NuRD and Sin3 histone deacetylase complexes in development. Trends Genet. 16, 351–356.

Baker, M. (2011). Making sense of chromatin states. Nat. Methods 8, 717–722.

Balasubramanii, M., Day, B.W., Schoen, R.E., and Getzenberg, R.H. (2006). Altered expression and localization of creatine kinase B, heterogeneous nuclear ribonucleoprotein F, and high mobility group box 1 protein in the nuclear matrix associated with colon cancer. Cancer Res. 66, 763–769.

Cannizzaro, E., Bannister, A.J., Han, N., Alendrar, A., and Kouzarides, T. (2018). DDX3X RNA helicase affects breast cancer cell cycle progression by regulating expression of KLF4. FEBS Lett. 592, 2308–2322.

Casciello, F., Al-Ejhel, F., Kelly, G., Brennan, D.J., Ngow, S.F., Young, A., Stoel, T., Windloch, K., Hill, M.M., Smyth, M.J., et al. (2017). G9A drives hypoacetylated-mediated repression for breast cancer cell survival and tumorigenesis. Proc. Natl. Acad. Sci. U S A 114, 7077–7082.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksay, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2, 401–404.

Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. Cell 163, 506–519.

Curtis, C., Shah, S.P., Chin, S.F., Tarushivi, 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 2000 breast tumours reveals novel subgroups. Nature 486, 346–352.

Dong, C., Wu, Y., Yao, J., Wang, Y., Yu, Y., Rychahou, P.G., Evers, B.M., and Zhou, B.P. (2012). G9A interacts with Snail and is critical for Snail-mediated E-cadherin repression in human breast cancer. J. Clin. Invest. 122, 1449–1468.

Easwaran, H., Tsai, H.C., and Baylin, S.B. (2014). Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. Mol. Cell 54, 716–727.

Gao, J., Aksay, B.A., Dogrusoz, U., Drescher, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6, pl1.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483, 570–575.

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.

Goh, E.T., Pardo, O.E., Michael, N., Niewiarowski, A., Totty, N., Volkova, D., Tsaneva, I.R., Seckl, M.J., and Gout, I. (2010). Involvement of heterogeneous ribonucleoprotein F in the regulation of cell proliferation via the mammalian target of rapamycin/56 kinase 2 pathway. J. Biol. Chem. 285, 17065–17076.

Green, E.D., and Guyer, M.S. (2011). Charting a course for genomic medicine from base pairs to bedside. Nature 470, 204–213.

Gruber, S.B., Moreno, V., Rozeck, L.S., Rennerts, H.S., Lejkiwicz, F., Bonner, J.D., Greenman, J.K., Helin, K., and Dhanak, D. (2013). Chromatin: systematic exploration of the human interactome. Cell 152, 685–689.

Laing, J.Y., Stroman, I., and Kouzarides, T. (2018). DDX3X RNA helicase affects breast cancer cell cycle progression by regulating expression of KLF4. FEBS Lett. 592, 2308–2322.

Lee, J.H., Horak, C.E., Khanna, C., Meng, Z., Yu, L.R., Veenstra, T.D., and Steeg, P.S. (2008). Alterations in Gemin5 expression contribute to alternative mRNA splicing patterns and tumor cell motility. Cancer Res. 68, 639–644.

Lin, T.C., Su, C.Y., Wu, P.Y., Lai, T.C., Pan, W.A., Jan, Y.H., Chang, Y.C., Yeh, C.T., Chen, C.L., Ger, L.P., et al. (2016). The nuclear protein NIFK promotes cancer progression via CK1alpha/beta-catenin in metastasis and Ki-67-dependent cell proliferation. Elife 5, e11288.

Liu, C., Yu, Y., Liu, F., Wei, X., Wrobel, J.A., Gunawardena, H.P., Zhou, L., Jin, J., and Chen, X. (2014a). A chromatin activity-based chemoproteomic approach reveals a transcriptional repression for gene-specific silencing. Nat. Commun. 5, 5735.

Liu, N.Q., Stingl, C., Look, M.P., Smid, M., Breauxman, R.S., De Marini, T., Sieweert, A.M., Span, P.N., Sweep, F.C., Linderholm, B.K., et al. (2014b). Comparative proteome analysis revealing an 11-protein signature for aggressive triple-negative breast cancer. J. Natl. Cancer Inst. 106, dju576.

Liu, X.R., Zhou, L.H., Hu, J.X., Liu, L.M., Wan, H.P., and Zhang, X.C. (2018). UNC0638, a G9a inhibitor, suppresses epithelial mesenchymal transition mediated cellular migration and invasion in triple negative breast cancer. Mol. Med. Rep. 17, 2239–2244.

Ma, C., Chen, H.I., Flores, M., Huang, Y., and Chen, Y. (2013). BRCA-Monet: a breast cancer specific drug treatment mode-of-action network for treatment effective prediction using large scale microarray database. BMC Syst. Biol. 7 (Suppl 5), SS.

Maier, V.K., Feeney, C.M., Taylor, J.E., Creech, A.L., Qiao, J.W., Santato, A., Das, P.P., Chevrier, N., Cifuentes-Rojas, C., Orkin, S.H., et al. (2015). Functional proteomic analysis of repressive histone methyltransferase complexes reveals ZNF518B as a G9A regulator. Mol. Cell. Proteomics 14, 1435–1446.

Malovannaya, A., Lanz, R.B., Jung, S.Y., Bulyonko, Y., Le, N.T., Chan, D.W., Ding, C., Shi, Y., Yucer, N., Krichevtsev, G., et al. (2011). Analysis of the human endogenous coregulator complexome. Cell 145, 787–799.

Marcel, V., Ghayed, S.E., Belin, S., Therziols, G., Morel, A.P., Solano-Gonzalez, E., Vendrell, J.A., Hacot, S., Mertani, H.C., Albaret, M.A., et al. (2013). pS3 acts as a safeguard of translational control by regulating fibrillin and RNA methylation in cancer. Cancer Cell 24, 318–330.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauer, K.R., Wang, P., Wang, X., Qiao, K., and Bentires-Alj, M. (2015). Breast tumor heterogeneity: source of fitness, hurdle for therapy. Mol. Cell 60, 537–546.

Koren, S., and Bajer, M. (2015). An integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6, pl1.

Kouzarides, T. (2007). Genetic variation in 8q24 associated with breast tumours reveals novel subgroups. Nature 483, 716–722.
Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 55–62.

Ng, S.W., Mitchell, A., Kennedy, J.A., Chen, W.C., McLeod, J., Ibrahimova, N., Arruda, A., Popescu, A., Gupta, V., Schimmer, A.D., et al. (2016). A 17-gene stromness score for rapid determination of risk in acute leukaemia. Nature 540, 433–437.

Ooi, L., and Wood, I.C. (2007). Chromatin crosstalk in development and disease: lessons from REST. Nat. Rev. Genet. 8, 544–554.

Oshiro, M.M., Kim, C.J., Wozniak, R.J., Junk, D.J., Munoz-Rodriguez, J.L., Burr, J.A., Fitzgerald, M., Pawar, S.C., Cres, A.E., Domann, F.E., et al. (2005). Epigenetic silencing of DSC3 is a common event in human breast cancer. Breast Cancer Res. 7, R669–R680.

Parker, J.S., Proctor, B., Ellis, P.J., Sgroi, D.,ータク, H., Hughes, S., Morrison, V., Dai, Y., Cao, Y., ... & 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.

Pereira, B., Chin, S.F., Rueda, O.M., Vollan, H.K., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.J., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat. Commun. 7, 11479.

Privette Vinnedge, L.M., McClare, R., Wagh, P.K., Wilkenheiser-Brockamp, K.A., Waltz, S.E., and Wells, S.I. (2011). The human DEK oncogene stimulates beta-catenin signaling, invasion and mammosphere formation in breast cancer. Oncogene 30, 2741–2752.

Ruggles, K.V., Tang, Z., Wang, X., Grover, H., Askenazi, M., Teubl, J., Cao, S., McLellan, M.D., Clauser, K.R., Tabb, D.L., et al. (2016). An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. Mol. Cell Proteomics 15, 106–1071.

Sanchez-Bailon, M.P., Calcabrini, A., Mayoral-Varo, V., Molinari, A., Wagner, K.U., Losada, J.P., Ciordia, S., Albar, J.P., and Martin-Perez, J. (2015). Cyr61 as mediator of Src signaling in triple negative breast cancer cells. Oncotarget 6, 13520–13538.

Shah, S.N., Cope, L., Poh, W., Belton, A., Roy, S., Talbot, C.C., Jr., Sukumar, S., Huso, D.L., and Resar, L.M. (2013). HMGA1: a master regulator of tumor progression in triple-negative breast cancer cells. PLoS One 8, e63419.

Shi, Y., Sawada, J., Sui, G., Affar el, B., Whetstone, J.R., Lan, F., Ogawa, H., Luke, M.P., Nakatani, Y., and Shi, Y. (2003). Coordinated histone modifications mediated by a CBP co-repressor complex. Nature 422, 735–738.

Si, W., Huang, W., Zheng, Y., Yang, Y., Liu, X., Shan, L., Zhou, X., Wang, Y., Su, D., Gao, J., et al. (2015). Dysfunction of the reciprocal feedback loop between GATA3- and ZEB2-nucleated repression programs contributes to breast cancer metastasis. Cancer Cell 27, 822–836.

Silva, G.O., He, X., Parker, J.S., Gatza, M.L., Carey, L.A., Hou, J.P., Moulder, S.L., Marcom, P.K., Ma, J., Rosen, J.M., et al. (2015). Cross-species DNA copy number analyses identifies multiple 1q21-q23 subtype-specific driver genes for breast cancer. Breast Cancer Res. Treat. 152, 347–356.

Swoboda, R.K., Samasundaram, R., Caputo, L., Ochoa, E.M., Gimotty, P.A., Marincola, F.M., Van Belle, P., Barth, S., Elder, D., Guerry, D., et al. (2007). Shared MHC class II-dependent melanoma ribosomal protein L8 identified by phage display. Cancer Res. 67, 3553–3559.

Torga, G., and Pienta, K.J. (2017). Patient-paired sample congruence between 2 commercial liquid biopsy tests. JAMA Oncol. 4, 868–870.

Tu, W.B., Shah, Y.J., Lourenco, C., Mullen, P.J., Dingar, D., Redel, C., Tamachi, A., Ba-Alawi, W., Ding, D., Redel, C., Tamachi, A., Ba-Alawi, W., et al. (2018). LYAR promotes colorectal cancer cell mobility by activating galecin-1 expression. Oncotarget 6, 32890–32901.

Yates, L.R., Gerstung, M., Knopp, A., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L.B., Lassimon, D., Davies, H., et al. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat. Med. 21, 751–759.

Yonezawa, K., Sugihara, Y., Oshima, K., Matsuda, T., and Nadano, D. (2014). Lyar, a cell growth-regulating zinc finger protein, was identified to be associated with cytoplasmic ribosomes in male germ and cancer cells. Mol. Cell. Biochem. 395, 221–229.

Zhang, Y., Bayasc, K.C., Yee, L.F., Saporita, A.J., and Weber, J.D. (2014). Elevated DDX21 regulates c-Jun activity and RNA processing in human breast cancers. Breast Cancer Res. 16, 449.
Supplemental Information

Multi-omic Dissection of Oncogenically Active Epiproteomes Identifies Drivers of Proliferative and Invasive Breast Tumors

John A. Wrobel, Ling Xie, Li Wang, Cui Liu, Naim Rashid, Kristalyn K. Gallagher, Yan Xiong, Kyle D. Konze, Jian Jin, Michael L. Gatza, and Xian Chen
Figure S1

(A) Heat map displays the LFQ ratios of UNC0965-captured proteins in different BC cell lines versus the normal cell line (MCF10A). ‘G9a PD’ refers to ‘UNC0965 pull-down’, and ‘Ctrl PD’ is the pull-downs from empty beads. Based on the changes of LFQ ratios, the G9a interactors were in four clusters, including luminal- or BLBC/basal- or BC-specific as well as the core components of G9a-interacting epiproteomes. (B) LFQ ratios show that the G9a bindings of UNC0965-caprtured proteins are correlated between two luminal-subtype cell lines (MCF7, T47D) versus MCF10A (left), or between two basal subtype cell lines (MDA231, SUM159) versus MCF10A (right). The proteins showing log2 LFQ ratios > 1 or < -1 and -log10 p value > 1 are putative BC subtypic interactors of statistical significance.
Figure S2

A

BCT pair 2, Log₂(basal/luminal)

BCT pair 1, Log₂(basal/luminal)

Pearson correlation 0.494

Pearson correlation 0.635

B

|            | Input       | UNC965      | Control     |
|------------|-------------|-------------|-------------|
| HuMEC      | MCF7        | T4D7        | HCC1806     | MDA231      |
| HuMEC      | MCF7        | T4D7        | HCC1806     | MDA231      |
| MCF7       | T4D7        | HCC1806     | MDA231      |             |
| MCF7       | T4D7        | HCC1806     | MDA231      |             |
| MCF7       | T4D7        | HCC1806     | MDA231      |             |
| HCC1806    | MCF7        | T4D7        | HCC1806     | MDA231      |
| MDA231     |             |             |             |             |
| EHMT2      | G9a         |             |             |             |
|           |             |             |             |             |
| RBM47      |             |             |             |             |
|           |             |             |             |             |
| NOP2       |             |             |             |             |
|           |             |             |             |             |
| EBP2 (EBNA1BP2) |         |             |             |             |
|           |             |             |             |             |
| CTCF       |             |             |             |             |
|           |             |             |             |             |
| DDX21      |             |             |             |             |
|           |             |             |             |             |
| SF3B2      |             |             |             |             |
|           |             |             |             |             |
| BRIX1/BXD2 |             |             |             |             |
|           |             |             |             |             |
| PTBP1      |             |             |             |             |
|           |             |             |             |             |
| LYAR       |             |             |             |             |
|           |             |             |             |             |
| SMARC1     |             |             |             |             |
|           |             |             |             |             |
| Tubulin    |             |             |             |             |
Figure S2. ChaC sorted and characterized the G9a-interacting epiproteome representing single predominate tumor phenotype in heterogenous tissues, Related to Figure 1.

(A) Plots of basal/luminal LFQ ratios of UNC0965-captured proteins from human breast cancer tissues (BCT). The proteins showing log2 LFQ ratios > 1 or < -1 and -log10 p value > 1 are putative BC subtypic interactors of statistical significance, including ‘purple’ for the proteins showing increased binding to G9a in the BLBC patients and ‘blue’ for the G9a interactors identified in the luminal patients. (B) Immunoblotting analysis of the indicated proteins that were captured by UNC0965 or UNC125 (G9a-negative probe) respectively in indicated cell lines. Expect for RBM47 that was identified as a luminal G9a interactor all others are the BLBC-specific G9a interactors.
Figure S3. iCEP box plots for complete set of G9a basal interactors, Related to Figure 3.

Box plots showing the statistically significant altered mRNA expression (x-axis) for the complete set of basal G9a interactor genes. For each interactor gene, the distribution of mRNA expression for PAM50 basal-like TCGA patients (N = 136) is shown on the left in red, and for PAM50 luminal TCGA patients (N = 590) is shown on the right colored in blue. Outliers are indicated as circles. The median is indicated by the black bar inside each box. The mean is indicated by the cyan diamond. The mRNA expression as a Z-score is displayed on the y-axis. A Mann-Whitney-Wilcoxon Test was performed on each pair to judge differences in expression levels. This Mann-Whitney Wilcoxon Test was performed on all genes in the TCGA data set (n = ~18,000) and the resulting p-values were adjusted for multiple comparisons. Adjusted p-values are displayed above the x-axis for basal-like samples having greater expression (top line) and luminal samples having greater expression (bottom line). Adjusted p-values < 0.05 are colored red if expression is higher for basal-like and blue if expression is higher for luminal samples.
Figure S4

PAM50 Basal-like vs. Luminal A/B mRNA expression

G9a luminal interactors (TCGA Cell 2015)
Figure S4. iCEP box plots for complete set of G9a luminal interactors, Related to Figure 3.

Box plots showing the statistically significant altered mRNA expression (x-axis) for the complete set of luminal G9a interactor genes. For each interactor gene, the distribution of mRNA expression for PAM50 basal-like TCGA patients (N = 136) is shown on the left in red, and for PAM50 luminal TCGA patients (N = 590) is shown on the right colored in blue. Outliers are indicated as circles. The median is indicated by the black bar inside each box. The mean is indicated by the cyan diamond. The mRNA expression as a Z-score is displayed on the y-axis. A Mann-Whitney-Wilcoxon Test was performed on each pair to judge differences in expression levels. This Mann-Whitney Wilcoxon Test was performed on all genes in the TCGA data set (n = ~18,000) and the resulting p-values were adjusted for multiple comparisons. Adjusted p-values are displayed above the x-axis for basal-like samples having greater expression (top line) and luminal samples having greater expression (bottom line). Adjusted p-values < 0.05 are colored red if expression is higher for basal-like and blue if expression is higher for luminal samples.
Figure S5
Figure S5. UCSC Genome Browser view of luminal G9a interactors RPL7 and STAU2 whose encoding genes are located near one another, Related to Figure 4.

The names of the G9a interactor genes are displayed in red. The gene annotation track is shown for the “Basic Gene Annotation Set from GENCODE Version 19” filtered to show protein coding transcripts with transcription support level 1 (tsl1). For each view the frequency of luminal TCGA samples with copy number gain (segmentation mean > 0.1) (red) and copy number loss (segmentation mean < -0.1) (blue) is indicated.
Figure S6

A. RRS1

B. RPL7

C. RPL30

D. BOP1

E. RPL8

F. C8orf33

Putative copy-number alterations from GISTIC

mRNA Expression z-Scores

Amplification Frequency

Proliferation Score

mRNA Expression z-Scores

Amplification Frequency

Proliferation Score

mRNA Expression z-Scores

Amplification Frequency

Proliferation Score

mRNA Expression z-Scores

Amplification Frequency

Proliferation Score

mRNA Expression z-Scores

Amplification Frequency

Proliferation Score

mRNA Expression z-Scores

Amplification Frequency

Proliferation Score

mRNA Expression z-Scores

Amplification Frequency

Proliferation Score
Figure S6. Correlation plots for luminal G9a interactor genes located in a region of high amplification frequency on the q arm of chromosome 8 in luminal TCGA samples, Related to Figure 5.

Luminal G9a interactor genes: (A) RRS1, (B) RPL7, (C) RPL30, (D) BOP1, (E) RPL8, and (F) C8orf33 on the q arm of chromosome 8. Each row includes the correlation plots for a specific gene. (left) box plots showing the distribution of mRNA expression for luminal samples with GISTIC values of 1 or 2 (AMP) compared to GISTIC values of 0, -1, or -2 (DEL/diploid). The percent of luminal TCGA samples in each group is indicated. The adjusted p-value as determined by a Mann-Whitney-Wilcoxon Test is displayed to indicate the significance of the AMP group having higher mRNA expression. (middle) box plots showing the distribution of mRNA expression for luminal samples with a proliferation score in the top quartile compared to all the other samples. The adjusted p-value as determined by a Mann-Whitney-Wilcoxon Test is displayed to indicate the significance of the Top Quartile group having higher mRNA expression. (right) bar plots showing the frequency of samples with an amplification (GISTIC values of 1 or 2) for the indicated gene for luminal samples with a proliferation score in the top quartile compared to all the other samples. The adjusted p-value as determined by a Fisher’s Exact Test for Count Data is displayed to indicate the significance of the Top Quartile group having a greater amplification frequency. See Figure 5 for the plots for STAU2.
Figure S7

A. alter in luminal (N=299, NE=43, Mdn=112.3)
not altered in luminal (N=277, NE=33, Mdn=216.6)
p−value: 0.006
HR: 1.806
95% CI: 1.157 − 2.82

B. alter in luminal (N=312, NE=45, Mdn=113.7)
not altered in luminal (N=276, NE=35, Mdn=146.4)
p−value: 0.068
HR: 1.505
95% CI: 0.967 − 2.343

C. alter in luminal (N=410, NE=129, Mdn=278.4)
not altered in luminal (N=717, NE=194, Mdn=300.7)
p−value: 0.061
HR: 1.237
95% CI: 0.99 − 1.547

D. alter in luminal (N=433, NE=139, Mdn=263.7)
not altered in luminal (N=694, NE=184, Mdn=300.7)
p−value: 0.016
HR: 1.31
95% CI: 1.051 − 1.634

E. alter in luminal (N=328, NE=53, Mdn=112.3)
not altered in luminal (N=260, NE=27, Mdn=216.6)
p−value: 0.003
HR: 1.976
95% CI: 1.242 − 3.143

F. alter in luminal (N=311, NE=47, Mdn=113.7)
not altered in luminal (N=277, NE=33, Mdn=129.6)
p−value: 0.053
HR: 1.547
95% CI: 0.991 − 2.416
Figure S7. Kaplan-Meier survival analysis plots for luminal samples with copy-number amplification for luminal G9a interactor genes located in a region of high amplification frequency on the q arm of chromosome 8 in luminal TCGA samples, Related to Figure 6.

Luminal G9a interactor genes: (A) RRS1, (B) RPL7, (C) RPL30, (D) BOP1, (E) RPL8, and (F) C8orf33 on the q arm of chromosome 8. Each row includes the Kaplan-Meier survival plots for a specific gene for luminal samples with copy-number amplification (GISTIC score of 1 or 2) (red line) vs. copy-number deletion or diploid (GISTIC score of 0, -1, or -2) (blue line) in luminal TCGA samples (left) and luminal METABRIC samples (right). For the Kaplan-Meier plots, "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot. See Figure 6 for the plots for STAU2.
Overall Survival
altered in luminal (N=294, NE=42, Mdn=122.7)
not altered in luminal (N=294, NE=38, Mdn=216.6)
p−value: 0.235
HR:        1.304
95% CI:  0.84 − 2.025

Overall Survival
altered in luminal (N=294, NE=40, Mdn=129.5)
not altered in luminal (N=294, NE=40, Mdn=114.7)
p−value: 0.926
HR:        0.98
95% CI:  0.632 − 1.519

Overall Survival
altered in luminal (N=563, NE=173, Mdn=278.4)
not altered in luminal (N=564, NE=150, Mdn=300.7)
p−value: 0.061
HR:        1.233
95% CI:  0.99 − 1.535

Overall Survival
altered in luminal (N=563, NE=154, Mdn=285.7)
not altered in luminal (N=564, NE=169, Mdn=300.7)
p−value: 0.247
HR:        0.879
95% CI:  0.706 − 1.094

Overall Survival
altered in luminal (N=294, NE=40, Mdn=129.6)
not altered in luminal (N=294, NE=40, Mdn=127.2)
p−value: 0.255
HR:        0.774
95% CI:  0.498 − 1.204

Overall Survival
altered in luminal (N=563, NE=156, Mdn=NA)
not altered in luminal (N=564, NE=167, Mdn=285.7)
p−value: 0.4
HR:        0.911
95% CI:  0.732 − 1.133

Overall Survival
altered in luminal (N=294, NE=38, Mdn=129.5)
not altered in luminal (N=294, NE=42, Mdn=127.2)
p−value: 0.665
HR:        0.907
95% CI:  0.584 − 1.41

Overall Survival
altered in luminal (N=563, NE=186, Mdn=278.4)
not altered in luminal (N=564, NE=137, Mdn=300.7)
p−value: 8.8e−06
HR:        1.644
95% CI:  1.318 − 2.052

Overall Survival
altered in luminal (N=294, NE=47, Mdn=113.7)
not altered in luminal (N=294, NE=33, Mdn=129.6)
p−value: 0.003
HR:        1.963
95% CI:  1.249 − 3.084
Figure S8. Kaplan-Meier survival analysis plots for luminal samples with mRNA overexpression for luminal G9a interactor genes located in a region of high amplification frequency on the q arm of chromosome 8 in luminal TCGA samples, Related to Figure 6.

Luminal G9a interactor genes: (A) RRS1, (B) RPL7, (C) RPL30, (D) BOP1, (E) RPL8, and (F) C8orf33 (no data in METABRIC) on the q arm of chromosome 8. Each row includes the Kaplan-Meier survival plots for a specific gene for luminal samples with mRNA overexpression (z-score > median z-score) (red line) compared to samples not showing mRNA overexpression (z-score < median z-score) (blue line) in luminal TCGA samples (left) and luminal METABRIC samples (right). For the Kaplan-Meier plots, "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot. See Figure 6 for the plots for STAU2.
**Figure S9**

A.  
- **EIF2S2**: mRNA Expression z-Scores for DEL/diploid (60%) and AMP (40%). 
  - Putative copy-number alterations from GISTIC. 
  - p < 0.0001
  - Proliferation Score

B.  
- **EIF6**: mRNA Expression z-Scores for DEL/diploid (57%) and AMP (43%). 
  - Putative copy-number alterations from GISTIC. 
  - p < 0.0001
  - Proliferation Score

C.  
- **STAU1**: mRNA Expression z-Scores for DEL/diploid (54%) and AMP (46%). 
  - Putative copy-number alterations from GISTIC. 
  - p < 0.0001
  - Proliferation Score
Figure S9. Correlation plots for luminal G9a interactor genes located in a region of high amplification frequency on the q arm of chromosome 20 in luminal TCGA samples, Related to Figure 5.

Luminal G9a interactor genes: (A) EIF2S2, (B) EIF6, and (C) STAU1 on the q arm of chromosome 20. Each row includes the correlation plots for a specific gene. (left) box plots showing the distribution of mRNA expression for luminal samples with GISTIC values of 1 or 2 (AMP) compared to GISTIC values of 0, -1, or -2 (DEL/diploid). The percent of luminal TCGA samples in each group is indicated. The adjusted p-value as determined by a Mann-Whitney-Wilcoxon Test is displayed to indicate the significance of the AMP group having higher mRNA expression. (middle) box plots showing the distribution of mRNA expression for luminal samples with a proliferation score in the top quartile compared to all the other samples. The adjusted p-value as determined by a Mann-Whitney-Wilcoxon Test is displayed to indicate the significance of the Top Quartile group having higher mRNA expression. (right) bar plots showing the frequency of samples with an amplification (GISTIC values of 1 or 2) for the indicated gene for luminal samples with a proliferation score in the top quartile compared to all the other samples. The adjusted p-value as determined by a Fisher’s Exact Test for Count Data is displayed to indicate the significance of the Top Quartile group having a greater amplification frequency.
Overall Survival

A

amp = EIF2S2
study = TCGA (Cell 2015)

p-value: 0.138
HR: 1.403
95% CI: 0.895 – 2.199

altered in luminal (N=234, NE=32, Mdn=112.3)
not altered in luminal (N=354, NE=48, Mdn=129.6)

B

amp = EIF6
study = TCGA (Cell 2015)

p-value: 0.193
HR: 1.345
95% CI: 0.86 – 2.105

altered in luminal (N=251, NE=33, Mdn=113.7)
not altered in luminal (N=337, NE=47, Mdn=129.6)

C

amp = STAU1
study = TCGA (Cell 2015)

p-value: 0.267
HR: 1.286
95% CI: 0.824 – 2.008

altered in luminal (N=274, NE=35, Mdn=113.7)
not altered in luminal (N=314, NE=45, Mdn=129.6)

D

amp = EIF2S2
study = METABRIC

p-value: 0.025
HR: 1.336
95% CI: 1.035 – 1.724

altered in luminal (N=232, NE=78, Mdn=263.7)
not altered in luminal (N=895, NE=245, Mdn=300.7)

amp = EIF6
study = METABRIC

p-value: 0.001
HR: 1.484
95% CI: 1.163 – 1.893

altered in luminal (N=253, NE=90, Mdn=263.7)
not altered in luminal (N=874, NE=233, Mdn=300.7)

amp = STAU1
study = METABRIC

p-value: 0.001
HR: 1.466
95% CI: 1.164 – 1.846

altered in luminal (N=313, NE=109, Mdn=263.7)
not altered in luminal (N=814, NE=214, Mdn=300.7)
Figure S10. Kaplan-Meier survival analysis plots for luminal samples with copy-number amplification for luminal G9a interactor genes located in a region of high amplification frequency on the q arm of chromosome 20 in luminal TCGA samples, Related to Figure 6.

Luminal G9a interactor genes: (A) EIF2S2, (B) EIF6, and (C) STAU1 on the q arm of chromosome 20. Each row includes the Kaplan-Meier survival plots for a specific gene for luminal samples with copy-number amplification (GISTIC score of 1 or 2) (red line) vs. copy-number deletion or diploid (GISTIC score of 0, -1, or -2) (blue line) in luminal TCGA samples (left) and luminal METABRIC samples (right). For the Kaplan-Meier plots, "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot.
Overall Survival

overexpress = EIF2S2
study = TCGA (Cell 2015)

- altered in luminal (N=294, NE=43, Mdn=122.7)
- not altered in luminal (N=294, NE=37, Mdn=129)

p−value: 0.046
HR: 1.562
95% CI: 1.005 − 2.428

overexpress = EIF6
study = TCGA (Cell 2015)

- altered in luminal (N=294, NE=34, Mdn=146.4)
- not altered in luminal (N=294, NE=46, Mdn=112.3)

p−value: 0.16
HR: 0.727
95% CI: 0.466 − 1.136

overexpress = STAU1
study = TCGA (Cell 2015)

- altered in luminal (N=294, NE=41, Mdn=113.7)
- not altered in luminal (N=294, NE=39, Mdn=129)

p−value: 0.215
HR: 1.319
95% CI: 0.85 − 2.046

overexpress = EIF2S2
study = METABRIC

- altered in luminal (N=562, NE=193, Mdn=256.9)
- not altered in luminal (N=565, NE=130, Mdn=300.7)

p−value: 6e−06
HR: 1.663
95% CI: 1.331 − 2.078

overexpress = EIF6
study = METABRIC

- altered in luminal (N=563, NE=183, Mdn=282.6)
- not altered in luminal (N=564, NE=140, Mdn=300.7)

p−value: 0.005
HR: 1.372
95% CI: 1.101 − 1.71

overexpress = STAU1
study = METABRIC

- altered in luminal (N=563, NE=173, Mdn=285.7)
- not altered in luminal (N=564, NE=150, Mdn=282.6)

p−value: 0.026
HR: 1.281
95% CI: 1.029 − 1.595
Figure S11. Kaplan-Meier survival analysis plots for luminal samples with mRNA overexpression for luminal G9a interactor genes located in a region of high amplification frequency on the q arm of chromosome 20 in luminal TCGA samples, Related to Figure 6.

Luminal G9a interactor genes: (A) EIF2S2, (B) EIF6, and (C) STAU1 on the q arm of chromosome 20. Each row includes the Kaplan-Meier survival plots for a specific gene for luminal samples with mRNA overexpression (z-score > median z-score) (red line) compared to samples not showing mRNA overexpression (z-score < median z-score) (blue line) in luminal TCGA samples (left) and luminal METABRIC samples (right). For the Kaplan-Meier plots, "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot.
Figure S12

A

RPL38

mRNA Expression z-Scores

DEL/diploid 64% AMP 36%
p < 0.0001
Putative copy-number alterations from GISTIC

RPL38

mRNA Expression z-Scores

DEL/diploid 64% AMP 36%
p < 0.0001
Putative copy-number alterations from GISTIC

RPL38

Amplification Frequency

p < 0.0001
Proliferation Score

B

LLGL2

mRNA Expression z-Scores

DEL/diploid 64% AMP 36%
p < 0.0001
Putative copy-number alterations from GISTIC

LLGL2

mRNA Expression z-Scores

DEL/diploid 64% AMP 36%
p < 0.0001
Putative copy-number alterations from GISTIC

LLGL2

Amplification Frequency

p < 0.0001
Proliferation Score

C

P4HB

mRNA Expression z-Scores

DEL/diploid 65% AMP 36%
p < 0.0001
Putative copy-number alterations from GISTIC

P4HB

mRNA Expression z-Scores

DEL/diploid 65% AMP 36%
p < 0.0001
Putative copy-number alterations from GISTIC

P4HB

Amplification Frequency

p < 0.0001
Proliferation Score
Figure S12. Correlation plots for luminal G9a interactor genes located in regions of high amplification frequency on the q arm of chromosome 17 in luminal TCGA samples, Related to Figure 5.

Luminal G9a interactor genes: (A) RPL38, (B) LLGL2, and (C) P4HB on the q arm of chromosome 17. Each row includes the correlation plots for a specific gene. (left) box plots showing the distribution of mRNA expression for luminal samples with GISTIC values of 1 or 2 (AMP) compared to GISTIC values of 0, -1, or -2 (DEL/diploid). The percent of luminal TCGA samples in each group is indicated. The adjusted p-value as determined by a Mann-Whitney-Wilcoxon Test is displayed to indicate the significance of the AMP group having higher mRNA expression. (middle) box plots showing the distribution of mRNA expression for luminal samples with a proliferation score in the top quartile compared to all the other samples. The adjusted p-value as determined by a Mann-Whitney-Wilcoxon Test is displayed to indicate the significance of the Top Quartile group having higher mRNA expression. (right) bar plots showing the frequency of samples with an amplification (GISTIC values of 1 or 2) for the indicated gene for luminal samples with a proliferation score in the top quartile compared to all the other samples. The adjusted p-value as determined by a Fisher’s Exact Test for Count Data is displayed to indicate the significance of the Top Quartile group having a greater amplification frequency.
Figure S13

A

amp = RPL38
study = TCGA (Cell 2015)

Overall Survival

p-value: 0.068
HR: 1.522
95% CI: 0.966 – 2.398

B

amp = LLGL2
study = TCGA (Cell 2015)

Overall Survival

p-value: 0.068
HR: 1.522
95% CI: 0.966 – 2.398

C

amp = P4HB
study = TCGA (Cell 2015)

Overall Survival

p-value: 0.268
HR: 1.298
95% CI: 0.817 – 2.062

Overall Survival

p-value: 0.058
HR: 1.279
95% CI: 0.991 – 1.651

Overall Survival

p-value: 0.059
HR: 1.289
95% CI: 1.001 – 1.66
Figure S13. Kaplan-Meier survival analysis plots for luminal samples with copy-number amplification for luminal G9a interactor genes located in a region of high amplification frequency on the q arm of chromosome 17 in luminal TCGA samples, Related to Figure 6.

Luminal G9a interactor genes: (A) RPL38, (B) LLGL2, and (C) P4HB on the q arm of chromosome 17. Each row includes the Kaplan-Meier survival plots for a specific gene for luminal samples with copy-number amplification (GISTIC score of 1 or 2) (red line) vs. copy-number deletion or diploid (GISTIC score of 0, -1, or -2) (blue line) in luminal TCGA samples (left) and luminal METABRIC samples (right). For the Kaplan-Meier plots, "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot.
Figure S14

A) overexpress = RPL38  
study = TCGA (Cell 2015)  

- Overall Survival  
  - p-value: 0.06  
  - HR: 0.654  
  - 95% CI: 0.419 – 1.021  

B) overexpress = LLGL2  
study = TCGA (Cell 2015)  

- Overall Survival  
  - p-value: 0.845  
  - HR: 1.045  
  - 95% CI: 0.673 – 1.623  

C) overexpress = P4HB  
study = TCGA (Cell 2015)  

- Overall Survival  
  - p-value: 0.435  
  - HR: 0.837  
  - 95% CI: 0.534 – 1.31

D) overexpress = RPL38  
study = METABRIC  

- Overall Survival  
  - p-value: 0.73  
  - HR: 1.039  
  - 95% CI: 0.835 – 1.293

E) overexpress = LLGL2  
study = METABRIC  

- Overall Survival  
  - p-value: 0.011  
  - HR: 1.328  
  - 95% CI: 1.066 – 1.654

F) overexpress = P4HB  
study = METABRIC  

- Overall Survival  
  - p-value: 0.341  
  - HR: 1.112  
  - 95% CI: 0.894 – 1.383
Figure S14. Kaplan-Meier survival analysis plots for luminal samples with mRNA overexpression for luminal G9a interactor genes located in a region of high amplification frequency on the q arm of chromosome 17 in luminal TCGA samples, Related to Figure 6.

Luminal G9a interactor genes: (A) RPL38, (B) LLGL2, and (C) P4HB on the q arm of chromosome 17. Each row includes the Kaplan-Meier survival plots for a specific gene for luminal samples with mRNA overexpression (z-score > median z-score) (red line) compared to samples not showing mRNA overexpression (z-score < median z-score) (blue line) in luminal TCGA samples (left) and luminal METABRIC samples (right). For the Kaplan-Meier plots, "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot.
Figure S15

DDX23

A

mRNA Expression z-Scores

DEL/diploid 78%  AMP 22%

p < 0.0001

Putative copy-number alterations from GISTIC

DDX23

mRNA Expression z-Scores

DEL/diploid 78%  AMP 22%

p < 0.0001

Putative copy-number alterations from GISTIC

DDX23

mRNA Expression z-Scores

DEL/diploid 78%  AMP 22%

p < 0.0001

Putative copy-number alterations from GISTIC

ZNF385A

B

mRNA Expression z-Scores

DEL/diploid 78%  AMP 22%

p = 1

Putative copy-number alterations from GISTIC

WIBG

C

mRNA Expression z-Scores

DEL/diploid 77%  AMP 23%

p < 0.0001

Putative copy-number alterations from GISTIC

RPS26

D

mRNA Expression z-Scores

DEL/diploid 77%  AMP 23%

p = 0.0071

Putative copy-number alterations from GISTIC
Figure S15. Correlation plots for luminal G9a interactor genes located in a region on the q arm of chromosome 12 in luminal TCGA samples, Related to Figure 5.

Luminal G9a interactor genes: (A) DDX23, (B) ZNF385A, (C) WIBG, and (D) RPS26 on the q arm of chromosome 12. Each row includes the correlation plots for a specific gene. (left) box plots showing the distribution of mRNA expression for luminal samples with GISTIC values of 1 or 2 (AMP) compared to GISTIC values of 0, -1, or -2 (DEL/diploid). The percent of luminal TCGA samples in each group is indicated. The adjusted p-value as determined by a Mann-Whitney-Wilcoxon Test is displayed to indicate the significance of the AMP group having higher mRNA expression. (middle) box plots showing the distribution of mRNA expression for luminal samples with a proliferation score in the top quartile compared to all the other samples. The adjusted p-value as determined by a Mann-Whitney-Wilcoxon Test is displayed to indicate the significance of the Top Quartile group having higher mRNA expression. (right) bar plots showing the frequency of samples with an amplification (GISTIC values of 1 or 2) for the indicated gene for luminal samples with a proliferation score in the top quartile compared to all the other samples. The adjusted p-value as determined by a Fisher’s Exact Test for Count Data is displayed to indicate the significance of the Top Quartile group having a greater amplification frequency.
**Figure S16**

### A

**amp = DDX23**
**study = TCGA (Cell 2015)**

- Overall Survival
  - altered in luminal (N=129, NE=23, Mdn=79.4)
  - not altered in luminal (N=459, NE=57, Mdn=129.6)

- **p-value:** 3.3e−04
- **HR:** 2.396
- **95% CI:** 1.465−3.918

### B

**amp = ZNF385A**
**study = TCGA (Cell 2015)**

- Overall Survival
  - altered in luminal (N=132, NE=24, Mdn=79.4)
  - not altered in luminal (N=456, NE=56, Mdn=130.1)

- **p-value:** 1.8e−04
- **HR:** 2.462
- **95% CI:** 1.512−4.01

### C

**amp = WIBG**
**study = TCGA (Cell 2015)**

- Overall Survival
  - altered in luminal (N=135, NE=24, Mdn=79.4)
  - not altered in luminal (N=453, NE=56, Mdn=129.6)

- **p-value:** 1.9e−04
- **HR:** 2.453
- **95% CI:** 1.511−3.999

### D

**amp = RPS26**
**study = TCGA (Cell 2015)**

- Overall Survival
  - altered in luminal (N=136, NE=24, Mdn=79.4)
  - not altered in luminal (N=452, NE=56, Mdn=129.6)

- **p-value:** 5.5e−06
- **HR:** 2.075
- **95% CI:** 1.504−2.862

---

**amp = DDX23**
**study = METABRIC**

- Overall Survival
  - altered in luminal (N=81, NE=39, Mdn=148.9)
  - not altered in luminal (N=1046, NE=284, Mdn=285.7)

- **p-value:** 1e−05
- **HR:** 2.09
- **95% CI:** 1.495−2.922

---

**amp = ZNF385A**
**study = METABRIC**

- Overall Survival
  - altered in luminal (N=87, NE=37, Mdn=176)
  - not altered in luminal (N=1040, NE=286, Mdn=285.7)

- **p-value:** 0.001
- **HR:** 1.764
- **95% CI:** 1.252−2.485
Figure S16. Kaplan-Meier survival analysis plots for luminal samples with copy-number amplification for luminal G9a interactor genes located in a region on the q arm of chromosome 12 in luminal TCGA samples, Related to Figure 6.

Luminal G9a interactor genes: (A) DDX23, (B) ZNF385A, (C) WIBG (no data in METABRIC), and (D) RPS26 on the q arm of chromosome 12. Each row includes the Kaplan-Meier survival plots for a specific gene for luminal samples with copy-number amplification (GISTIC score of 1 or 2) (red line) vs. copy-number deletion or diploid (GISTIC score of 0, -1, or -2) (blue line) in luminal TCGA samples (left) and luminal METABRIC samples (right). For the Kaplan-Meier plots, "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot.
**Figure S17**

**A**
- Overexpression = DDX23
- Study = TCGA (Cell 2015)
- Overall Survival
- Altered in luminal (N=294, NE=39, Mdn=127.2)
- Not altered in luminal (N=294, NE=41, Mdn=129.5)
- $p$-value: 0.216
- HR: 1.319
- 95% CI: 0.849 - 2.049

**B**
- Overexpression = ZNF385A
- Study = TCGA (Cell 2015)
- Overall Survival
- Altered in luminal (N=294, NE=33, Mdn=129.5)
- Not altered in luminal (N=294, NE=47, Mdn=112.3)
- $p$-value: 0.232
- HR: 0.762
- 95% CI: 0.487 - 1.192

**C**
- Overexpression = WIBG
- Study = TCGA (Cell 2015)
- Overall Survival
- Altered in luminal (N=294, NE=36, Mdn=130.1)
- Not altered in luminal (N=294, NE=44, Mdn=112)
- $p$-value: 0.064
- HR: 0.657
- 95% CI: 0.42 - 1.028

**D**
- Overexpression = RPS26
- Study = TCGA (Cell 2015)
- Overall Survival
- Altered in luminal (N=294, NE=41, Mdn=129.5)
- Not altered in luminal (N=294, NE=39, Mdn=129.5)
- $p$-value: 0.464
- HR: 0.847
- 95% CI: 0.543 - 1.321

**E**
- Overexpression = DDX23
- Study = METABRIC
- Overall Survival
- Altered in luminal (N=563, NE=193, Mdn=282.6)
- Not altered in luminal (N=564, NE=130, Mdn=300.7)
- $p$-value: 5.8e-05
- HR: 1.573
- 95% CI: 1.259 - 1.965

**F**
- Overexpression = ZNF385A
- Study = METABRIC
- Overall Survival
- Altered in luminal (N=563, NE=161, Mdn=300.7)
- Not altered in luminal (N=564, NE=162, Mdn=285.7)
- $p$-value: 0.781
- HR: 1.031
- 95% CI: 0.829 - 1.283

**G**
- Overexpression = RPS26
- Study = METABRIC
- Overall Survival
- Altered in luminal (N=563, NE=144, Mdn=285.7)
- Not altered in luminal (N=564, NE=179, Mdn=282.6)
- $p$-value: 0.044
- HR: 0.798
- 95% CI: 0.641 - 0.994
Figure S17. Kaplan-Meier survival analysis plots for luminal samples with mRNA overexpression for luminal G9a interactor genes located in a region on the q arm of chromosome 12 in luminal TCGA samples, Related to Figure 6.

Luminal G9a interactor genes: (A) DDX23, (B) ZNF385A, (C) WIBG (no data in METABRIC), and (D) RPS26 on the q arm of chromosome 12. Each row includes the Kaplan-Meier survival plots for a specific gene for luminal samples with mRNA overexpression (z-score > median z-score) (red line) compared to samples not showing mRNA overexpression (z-score < median z-score) (blue line) in luminal TCGA samples (left) and luminal METABRIC samples (right). For the Kaplan-Meier plots, "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot.
**Figure S18**

A. co-overexpress = HNRNPF, ILF3, MRPL39  
study = TCGA (Cell 2015)  

- **Overall Survival**  
- **p-value:** 0.04  
- **HR:** 2.553  
- **95% CI:** 1.012 - 6.44

B. co-overexpress = HNRNPF, ILF3, MRPL39  
study = METABRIC  

- **Overall Survival**  
- **p-value:** 0.014  
- **HR:** 2.249  
- **95% CI:** 1.157 - 4.375

C. co-overexpress = EHMT1, EHMT2, ILF3  
study = TCGA (Cell 2015)  

- **Overall Survival**  
- **p-value:** 0.016  
- **HR:** 2.885  
- **95% CI:** 1.171 - 7.105

D. co-overexpress = EHMT1, EHMT2, ILF3  
study = METABRIC  

- **Overall Survival**  
- **p-value:** 0.028  
- **HR:** 1.858  
- **95% CI:** 1.059 - 3.26

E. co-overexpress = DDX56, FBL, HNRNPF  
study = TCGA (Cell 2015)  

- **Overall Survival**  
- **p-value:** 0.016  
- **HR:** 2.934  
- **95% CI:** 1.17 - 7.355

F. co-overexpress = DDX56, FBL, HNRNPF  
study = METABRIC  

- **Overall Survival**  
- **p-value:** 0.014  
- **HR:** 2.029  
- **95% CI:** 1.137 - 3.619
Figure S18. Kaplan-Meier survival analysis for combinations of basal G9a interactors genes showing poor prognosis when co-overexpressed (mRNA expression Z score > median z score for all genes in the combination) in basal patients in both the TCGA and METABRIC datasets, Related to Figure 8.

Kaplan-Meier survival analysis indicates that altered mRNA expression of G9a interactor genes is prognostically significant in marking distinct patient subpopulations within single PAM50 subtypes. (A, C, & E) Overall survival (OS) plots of TCGA patients with the BLBC subtype (n=136). (B, D, & F) Overall survival (OS) plots of METABRIC patients with the BLBC subtype (n=198). The altered group (red line) is samples with mRNA co-overexpression (z-score > median z-score for all genes in the combination) for HNRNPF, ILF3, and MRPL39 (panels A & B), EHMT1, EHMT2, and ILF3 (panels C & D), and DDX56, FBL, and HNRNPF (panels E & F). The non-altered group (blue line) is the remaining samples not showing mRNA co-overexpression for the same genes. "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot.
Figure S19. Additional examples of Kaplan-Meier survival analysis for combinations of basal G9a interactors genes showing poor prognosis when co-overexpressed (mRNA expression Z score > median z score for all genes in the combination) in basal patients in both the TCGA and METABRIC datasets, Related to Figure 8.

Each panel represents a combination of basal G9a interactors with overall survival (OS) plots of TCGA patients with the BLBC subtype (N=136) displayed on the left and overall survival (OS) plots of METABRIC patients with the BLBC subtype (N=198) displayed on the right. The altered group (red line) is samples with mRNA co-overexpression (z-score > median z-score for all genes in the combination) for the indicated G9a interactor genes: (A) HNRNPF, MRPL39, and UTP23, (B) ILF2, MRPL39, and UTP23, (C) HNRNPF, ILF2, MRPL39, and UTP23, (D) ILF3, MRPL39, NONO, and PRPF3, (E) ILF2, LYAR, MRPL39, and NOLC1, (F) DEK, ILF2, MKI67, and MRPL39, (G) EHMT2, MRPL39, NONO, and PRPF3, (H) EHMT2, HNRNPF, and ILF3, (I) ILF3, MYBBP1A, PRPF6, and XRN2, (J) DDX21, DDX3X, NIFK, and SF3B1. The non-altered group (blue line) is the remaining samples not showing co-overexpression for the same genes. "N" refers to "Number of samples" and "NE" refers to "Number of Events". The number of events is for OS status = "DECEASED". "Mdn" indicates the median months survival. The log-rank p-value and Hazard Ratio (HR) with 95% Confidence Interval (CI) between the two groups are indicated in each plot.
Figure S20. UCSC Genome Browser views of basal G9a interactors whose encoding genes are located directly adjacent to one another or nearby, Related to Figure 7.

(A) MRPS21 directly adjacent to PRPF3 (1q21.2 - 1q21.3), (B) PBRM1 directly adjacent to GNL3 (3p21.1), (C) DDX50 directly adjacent to DDX21 (10q22.1), (D) NOP2 directly adjacent to CHD4, with ZNF384 located nearby (12p13.31), (E) HMGA1 located near to RPS10 (6p21.31). The G9a interactor genes are displayed in red. Where included the gene annotation track is shown for the “Basic Gene Annotation Set from GENCODE Version 19” filtered to show protein coding transcripts with transcription support level 1 (tsl1). For each view the frequency of basal TCGA samples with copy number gain (segmentation mean > 0.1) (red) and copy number loss (segmentation mean < -0.1) (blue) is indicated.
Figure S21. GISTIC copy number heatmap, Related to Figure 7.
Heatmap for basal G9a interactor genes (rows) with GISTIC copy number values from the TCGA sample set (columns). The TCGA samples and G9a interactor genes were clustered using unsupervised hierarchical clustering ("euclidean" distance and "ward" clustering method). The columns are annotated with PAM50 subtype of each TCGA patient. The rows are annotated with the proteomic subtype of G9a interactors.
TRANSPARENT METHODS

Chemicals and reagents. Antibodies against G9a (EHMT2) were from Abcam; antibodies against RBM47, NOP2, EBP2 (EBNA1BP2), TCOF1, ZNF326, DDX21, SF3B2, MKI67IP, BRIX1 (BXDV2), PTBP1, LYAR, and SMARCC1 were from Proteintech. Cell lines MCF10A, MCF7, MDA-MB-231, T47D, BT474, SUM159, and HCC1806 were purchased form ATCC (Manassas, VA). Human breast cancer frozen tissues with different receptor status were obtained from the Tissue Procurement Core Facility (TPF) at University of North Carolina.

Protein extraction. Cells were harvested at about 90% confluency. Each pellet of 10 million cells was homogenized in 10x cell pellet volume of hypotonic buffer (10 mM HEPES, pH 7.9, 1.5 mM MgCl2, 10 mM KCl). After centrifugation for 10 min at 12,000 x g at 4°C, the supernatant (cytosol) was discarded. The nuclear pellet was re-suspended in lysis buffer (50 mM Tris-HCl pH8.0, 150 mM NaCl, 0.5% CA630, protease inhibitor cocktail (Sigma-Aldrich), and 0.5 mM PMSF, sonicated and cleared by centrifugation for 20 min at 12000 x g. Protein concentration was determined by BCA assay. Frozen human tissue (approximately 0.1-0.2 g) was cut into small pieces and protein extraction was performed as above.

UNC0965 pull-down and ChaC sample processing. 1 mg nuclear protein extracted from either cell lines or clinical tissues was incubated overnight at 4°C with 2 nmole UNC0965 pre-coupled to 50 μl neutravidin-agarose (Thermo-Fisher), and washed three times with 1 ml lysis buffer to remove non-specific proteins. For on-beads sampling and processing, five additional washes with 50 mM Tris-HCl pH8.0, 150 mM NaCl were used to remove residual detergents. On-beads tryptic digestion was performed with 125 μl buffer containing 2 M urea, 50 mM Tris-HCl pH8.0, 1 mM DTT, 500 ng trypsin (Promega) for 30 min at room temperature on a mixer (Eppendorf). The tryptic digests were eluted twice with a 100 μl elution buffer containing 2 M urea, 50 mM Tris-HCl pH8.0, 5 mM iodoacetamide. Combined eluates were acidified with
trifluoroacetic acid at final concentration of 1% (TFA, mass spec grade, Thermo-Fisher) and desalted with a C18 stage tip.

**LC-MS/MS analysis.** Desalted peptide mixtures were dissolved in 30 μl 0.1% formic acid (Thermo-Fisher), of which 4 μl containing the peptides from 60-100 μg total protein was injected and analyzed by a ultra2D nanoLC system (Eksigent) coupled to a Velos LTQ Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose, CA) or an Easy nanoLC 1000 coupled to a Q-Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose, CA). In the nanoLC-Velos setup, peptides were first loaded on to a 2 mm × 0.5 mm reverse-phase (RP) C18 trap column (Eksigent) at a flow rate of 1 μl/min, then eluted, and fractionated on a 25 cm C18 RP column (25 cm × 75 μm ID, C18, 3 μm) with a gradient of 5-40% buffer B (ACN and 0.1% formic acid) at a constant flow rate of 250 nl/min over 180 min. In the Easy nanoLC- Q-Exactive setup, peptides were loaded on to a 15 cm C18 RP column (15 cm × 75 μm ID, C18, 2 μm, Acclaim Pepmap RSLC, Thermo-Fisher) and eluted with a gradient of 2-30% buffer B at a constant flow rate of 300 nl/min for 70 min followed by 30% to 80% B in 5 min and 80% B for 10 min. The Velos LTQ Orbitrap was operated in the positive-ion mode with a data-dependent automatic switch between survey Full-MS scan (m/z 300-1800) (externally calibrated to a mass accuracy of <5 ppm and a resolution of 60,000 at m/z 400) and CID MS/MS acquisition of the top 15 most intense ions. The Q-Exactive was also operated in the positive-ion mode but with a data-dependent top 20 method. Survey scans were acquired at a resolution of 70,000 at m/z 200. Up to the top 20 most abundant isotope patterns with charge ≥ 2 from the survey scan were selected with an isolation window of 2.0 m/z and fragmented by HCD with normalized collision energies of 27. The maximum ion injection time for the survey scan and the MS/MS scans was 250 ms and 120 ms, respectively, and the ion target values were set to 1e6 and 2e5, respectively. Selected sequenced ions were dynamically excluded for 20 seconds.
**Mass spec data and LFQ analysis.** Mass spectral processing and peptide identification were performed on the Andromeda search engine in MaxQuant software (Version 1.5.2.8) against a human UniProt database. All searches were performed with a defined modification of cysteine carbamidomethylation, with methionine oxidation and protein amino-terminal acetylation as dynamic modifications. Peptides were confidently identified using a target-decoy approach with a peptide false discovery rate (FDR) of 1% and a protein FDR of 5%. A minimum peptide length of 7 amino acids was required, maximally two missed cleavages were allowed, initial mass deviation for precursor ion was up to 7 ppm, and the maximum allowed mass deviation for fragment ions was 0.5 Da.

LFQ-based LC-MS/MS experiments were performed with 2-3 biological replicates, each with three technical replicates, on two sets of seven BC cell lines of distinct PAM50-subtypes (luminal or BLBC/basal), or two tissue samples of the corresponding BC PAM50 subtypes paired with their adjacent non-malignant tissues. For LFQ analysis, a match between runs option was enabled and time window at 0.7 minutes. Data processing and statistical analysis were performed on Perseus (Version 1.5.1.6) (Cox and Mann, 2011). Protein quantitation was performed on biological replicate runs and a two sample t-test statistics was used with a p-value of 5% to report statistically significant expression fold-changes.

**Analysis of functional category and networks of ISE-Interacting proteins.** Similar to our recent report (Liu et al., 2014), the biological processes and molecular functions of the G9a-interacting proteins were categorized by IPA (http://www.ingenuity.com/), DAVID (http://david.abcc.ncifcrf.gov/), and STRING (http://string-db.org/).

**TCGA and METABRIC data sets.** All TCGA (Ciriello et al., 2015) and METABRIC (Curtis et al., 2012; Pereira et al., 2016) data used in our study were downloaded from the cBioPortal (Cerami et al., 2012; Gao et al., 2013) as files brca_tcg.a_pub2015.tar.gz for TCGA and
brca_metabric.tar.gz for METABRIC or retrieved by direct programmatic access from the cBioPortal using the ‘cgdsr’ R package (Jacobsen, 2015). For our analysis of TCGA data, we used complete samples with mutation, copy-number, and mRNA expression data provided (n = 816) from the TCGA cancer study brca_tcga_pub2015 (Breast Invasive Carcinoma) (Ciriello et al., 2015). The mRNA expression data set for G9a interactor genes was obtained from the data_RNA_Seq_v2_mRNA_median_Zscores.txt file (found in brca_tcga_pub2015.tar.gz), containing mRNA expression Z-scores compared with diploid tumors (diploid for each gene). The copy-number data set for G9a interactor genes was obtained from the data_CNA.txt file (found in brca_tcga_pub2015.tar.gz) for putative copy-number from GISTIC 2.0 for each gene. TCGA clinical data were obtained from the data_bcr_clinical_data_patient.txt and data_bcr_clinical_data_sample.txt files (found in brca_tcga_pub2015.tar.gz), which included overall survival data. Additional TCGA clinical information including the PAM50 subtype assigned to each patient was obtained from “Table S1” in the TCGA publication (Ciriello et al., 2015). For the METABRIC (Curtis et al., 2012; Pereira et al., 2016) data, we used: case list id = brca_metabric_cnaseq (samples with mRNA, GISTIC, mutational data) and gene profile = brca_metabric_mrna_U133_Zscores to retrieve clinical data and mRNA expression data programmatically from the cBioPortal. The METABRIC GISTIC data set was obtained from the data_CNA.txt file (found in brca_metabric.tar.gz). We used 1866 of the 2051 samples in the METABRIC case list for which there was survival data.

Heatmap construction. The heatmap of mRNA expression levels for G9a interactor genes from the TCGA BRCA datasets was constructed using the ‘ComplexHeatmap’ R package (Gu et al., 2016). Hierarchical clustering was performed using the euclidean distance method and the Ward clustering method (option ward.D2 in R’s hclust function).
**Statistical analyses.** We used the Mann-Whitney-Wilcoxon Test with the `wilcox.test` function in R to determine significance for iCEP, mRNA vs. copy-number correlation, and mRNA vs. proliferation correlation. We used the Fisher’s Exact Test for Count Data with the `fisher.test` function in R to determine significance for the copy-number amplification vs. proliferation correlation. For these tests, we included all genes in the TCGA data set (n = ~18,000) to obtain an adjusted p-value for multiple comparisons for the G9a interactor genes. The adjusted p-values were calculated by submitting all the p-values determined for each gene from the individual test to the `p.adjust` function in R using the ‘fdr’ method. We considered an adjusted p-value < 0.05 to be significant.

**Kaplan-Meier survival analysis.** Kaplan-Meier curve plotting and statistical analysis for differences of overall survival (OS) based on mRNA expression or GISTIC score of G9a interactor genes among TCGA and METABRIC BRCA samples were performed using the ‘survival’ R package (Therneau and Grambsch, 2000). Kaplan-Meier estimator and log-rank tests were performed using the survival functions `Surv`, `survfit`, and `survdiff`. Cox proportional hazard survival analysis was performed using the survival function `coxph`.

**DATA AND SOFTWARE AVAILABILITY**

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al., 2019) partner repository with the dataset identifier PXD013795. The files used for the analysis of the TCGA and METABRIC datasets are available online at Mendeley Data.
Supplemental References

Cox, J., and Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. Annual review of biochemistry 80, 273-299.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics (Oxford, England).

Jacobsen, A. (2015). cgdsr: R-Based API for Accessing the MSKCC Cancer Genomics Data Server (CGDS). R package version 1.2.5.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic acids research 47, D442-D450.

Therneau, T.M., and Grambsch, P.M. (2000). Modeling survival data: extending the Cox model (New York: Springer).