doi:10.1038/nature11412

Comprehensive molecular portraits of human breast tumors

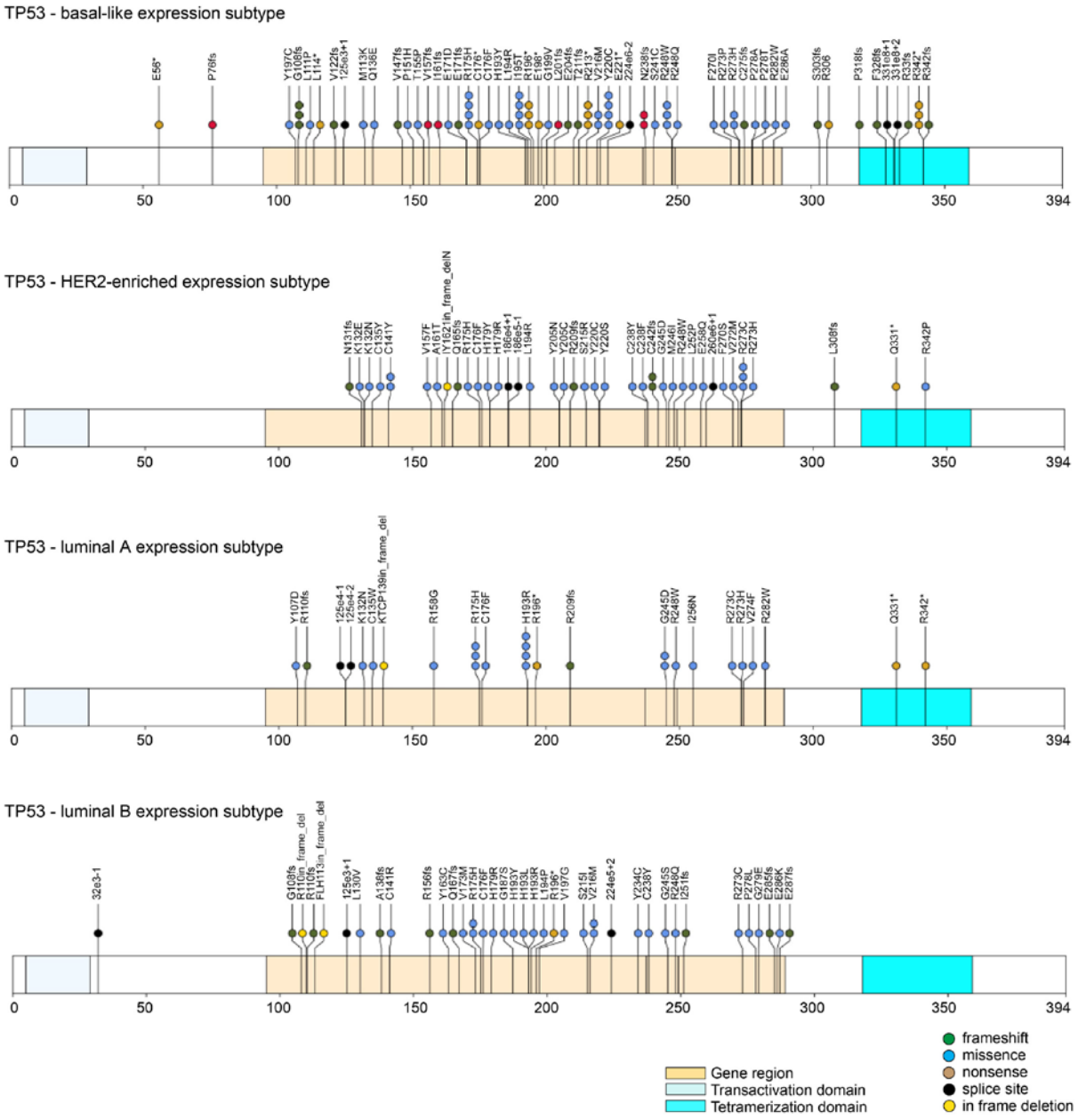
The Cancer Genome Atlas Network

I. Supplemental Figures

II. Supplemental Tables (additional xls files)

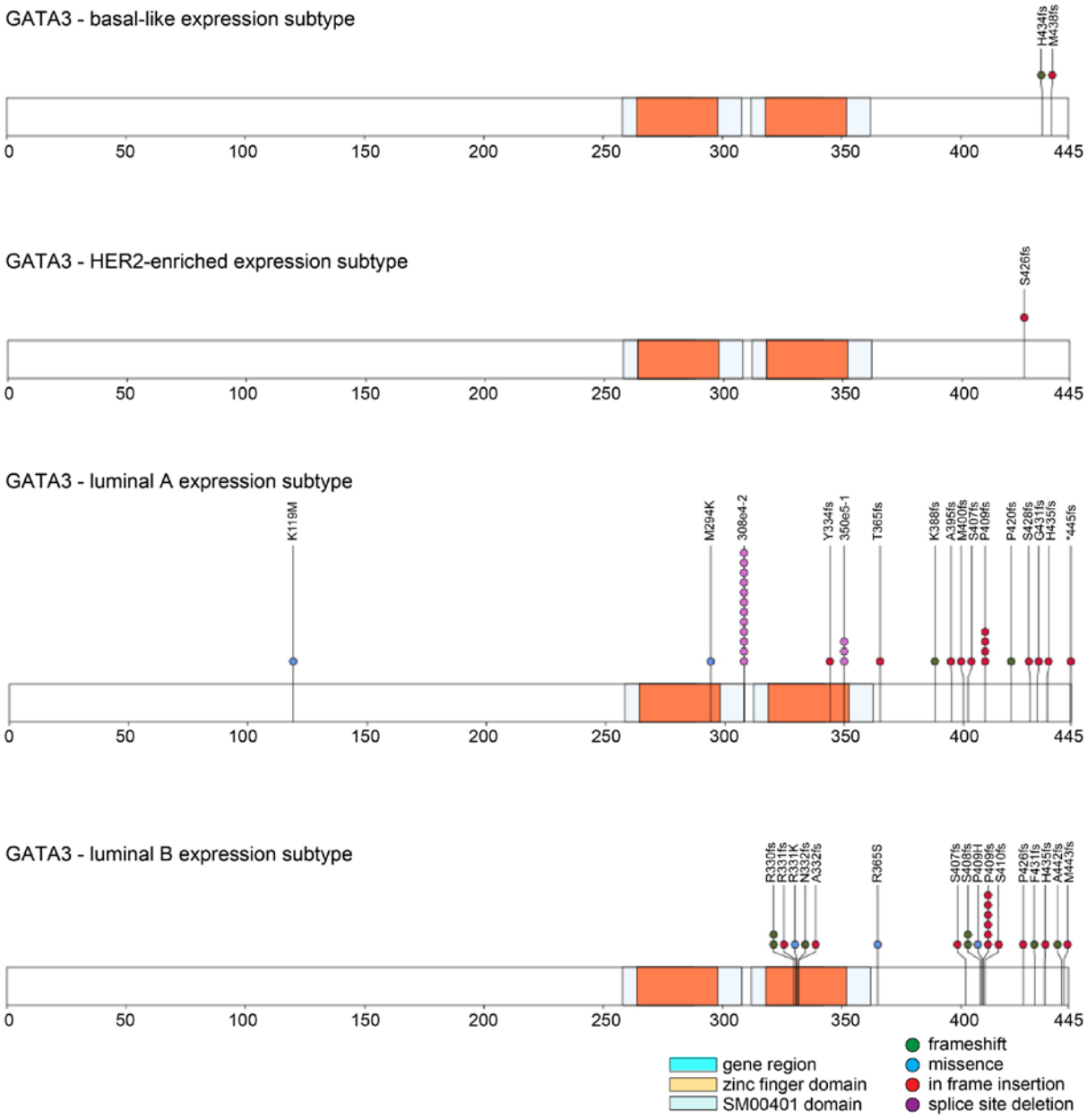
III. Supplemental Methods

Supplemental Figure 1



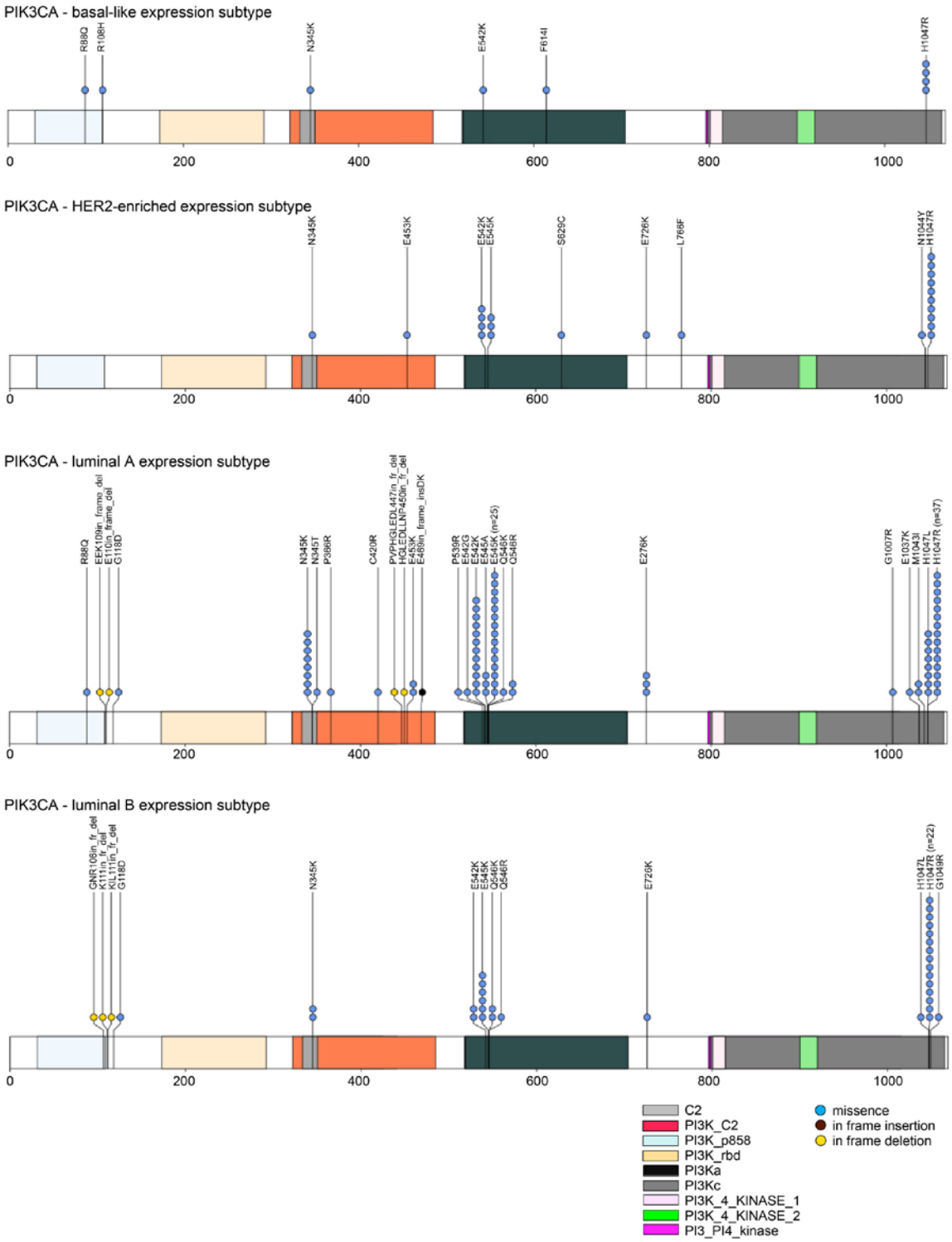
Supplemental Figure 1. TP53 mutation spectrum according to mRNA expression subtype. Lollipop mutation plots of nonsilent TP53 somatic mutations in breast cancers, by mRNA expression–based subtype. For reference, the distribution of TP53 mutation types (missense, nonsense, frame shift, splice site, in frame deletion) is shown as pie charts for both high-grade Serous Ovarian cancers and the Breast cancer subtypes.

Supplemental Figure 2



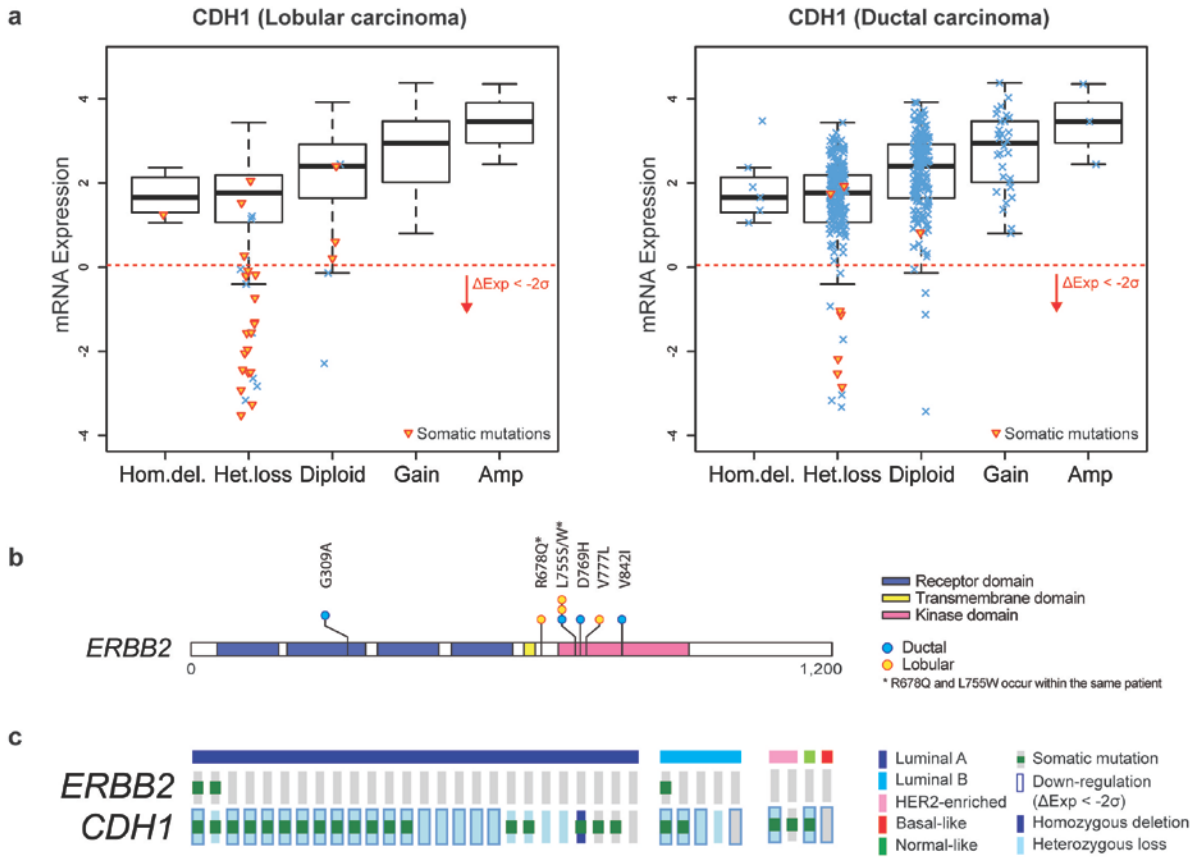
Supplemental Figure 2. GATA3 mutational spectrum according to mRNA expression subtype. Lollipop mutation plots of nonsilent GATA3 somatic mutations in breast cancer, by mRNA expression–based subtype.

Supplemental Figure 3

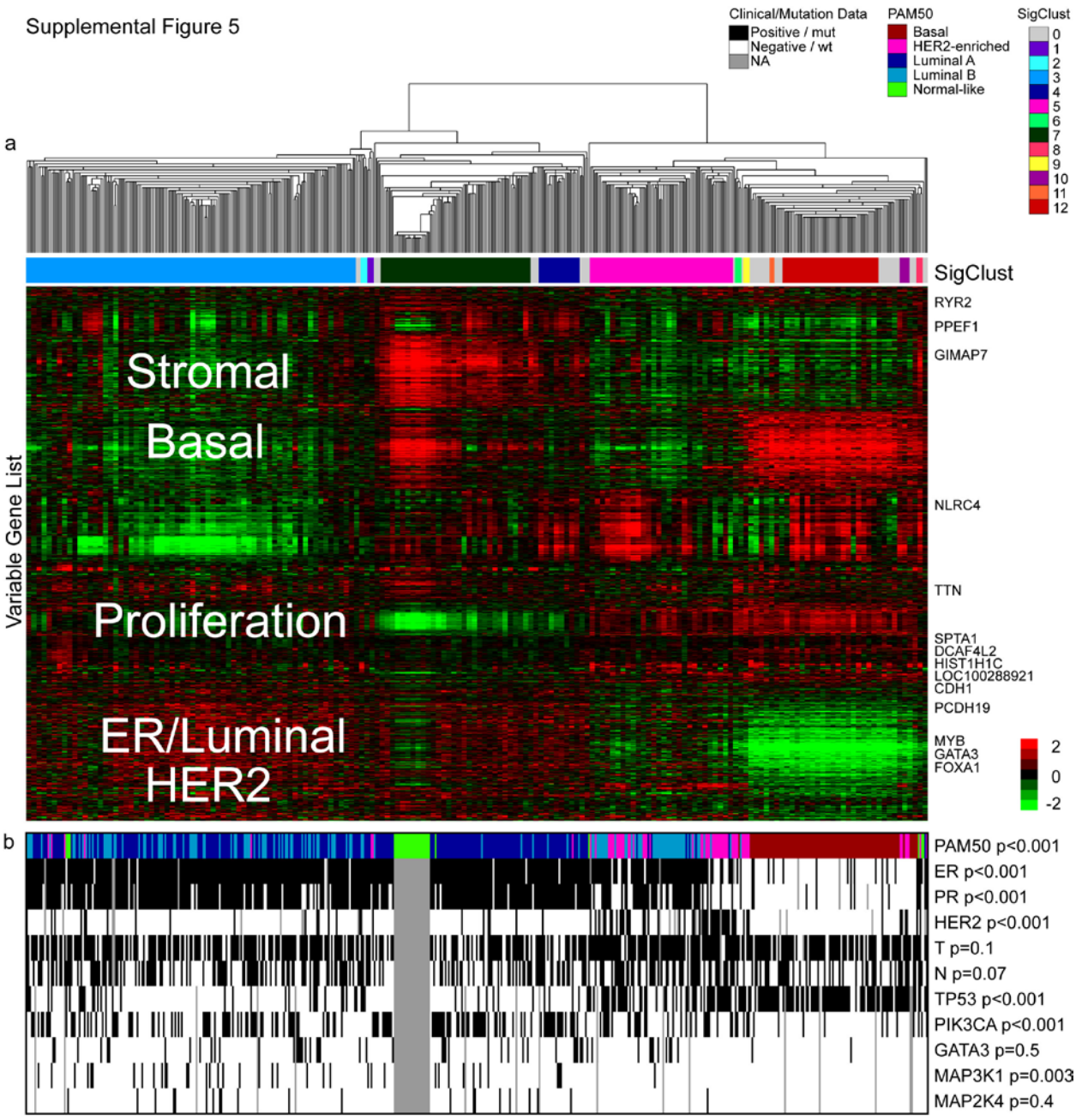


Supplemental Figure 3. PIK3CA mutational spectrum according to mRNA expression subtype. Lollipop mutation plots of nonsilent PIK3CA somatic mutations in breast cancer, by mRNA expression–based subtype.

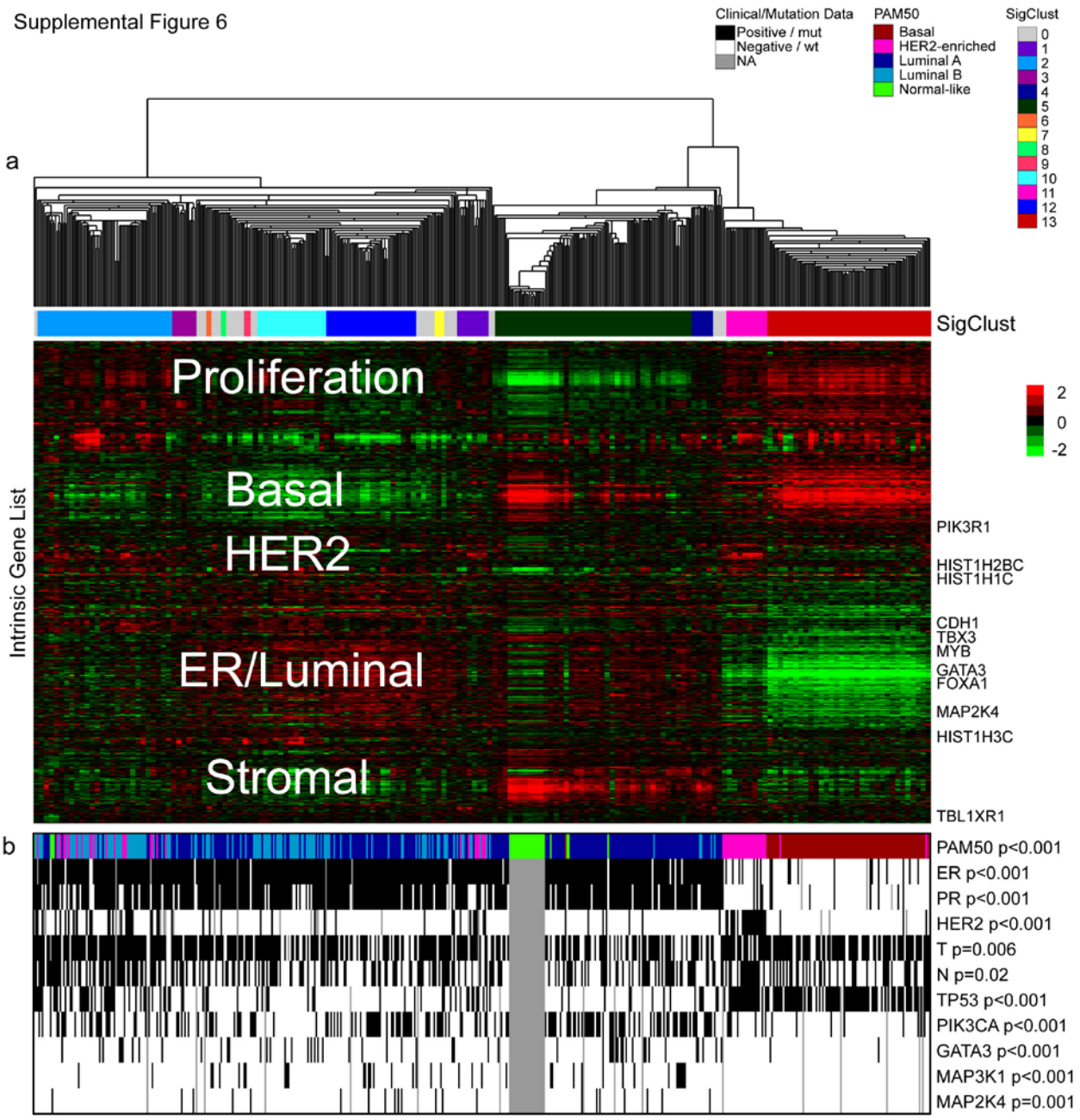
Supplemental Figure 4



Supplemental Figure 4. Lobular breast cancer enriched alterations. a) CDH1 mutation and down-regulation is a marker event of lobular breast tumors. Mutations almost always occur together with loss of one allele. b) By SEA analysis (see Supplemental Methods) HER/ERBB2 mutations were enriched in lobular tumors (p=0.0007). Notably, HER2/ERBB2 mutations occurred within the kinase domain in lobular breast cancer patients. c) Overview of alterations targeting CDH1 and HER2/ERBB2 in lobular tumors. Lobular cases were almost exclusively Luminal and preferentially Luminal A.

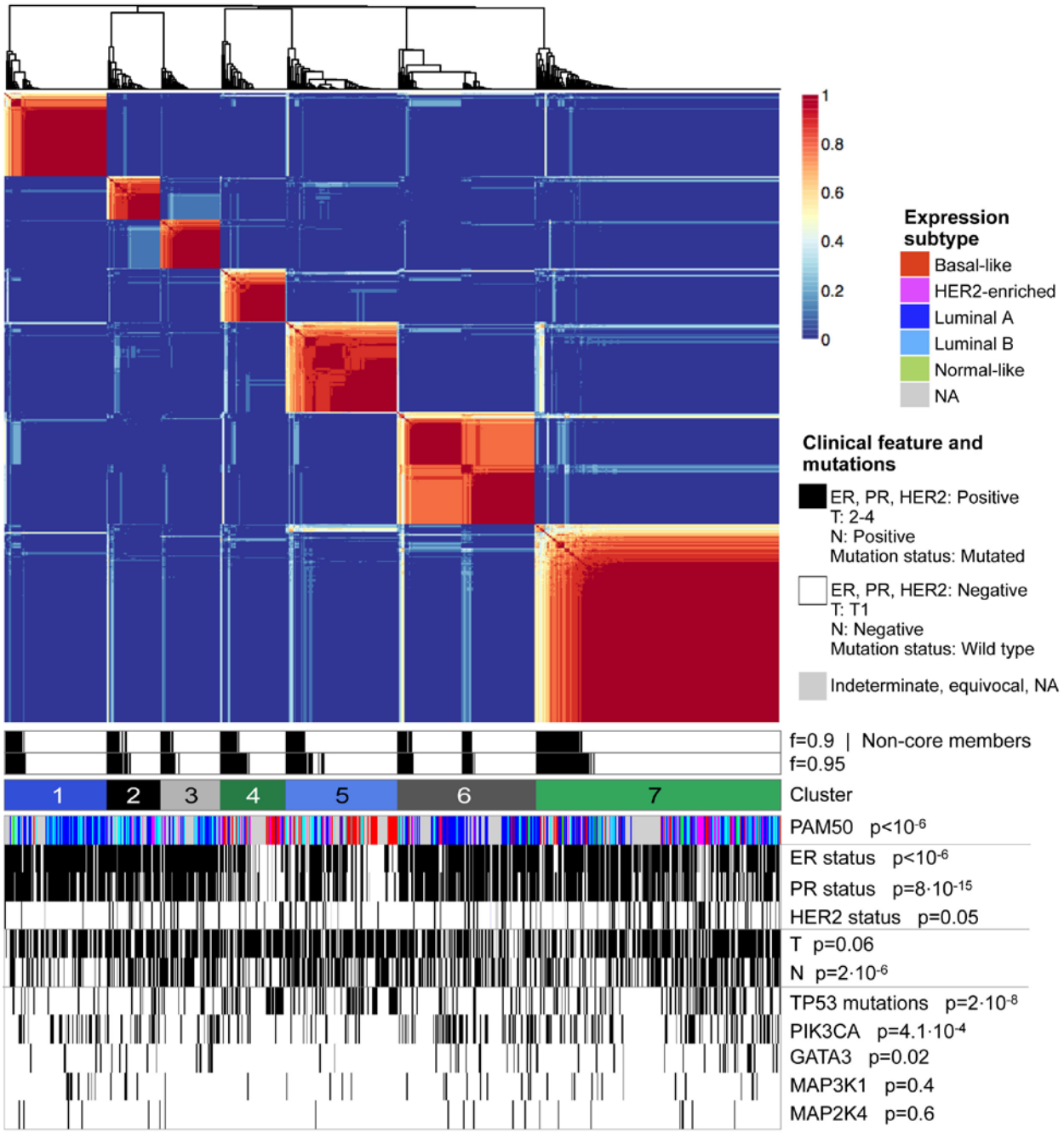


Supplemental Figure 5. Unsupervised mRNA expression analyses. 522 primary breast tumors, 3 metastatic tumors and 22 normal samples were used for two different hierarchical clustering analyses. a) a unsupervised hierarchical clustering analysis using the ~3600 most variably expressed genes was performed with the resulting data tested by SigClust to identify twelve possible groups. b) The unsupervised mRNA expression subtypes showed a high correlation to the previously defined PAM50 subtypes, as well as high correlation to ER, PR, and HER2 status, and to TP53, GATA3, MAP3K1 and MAP2K4 mutation status.



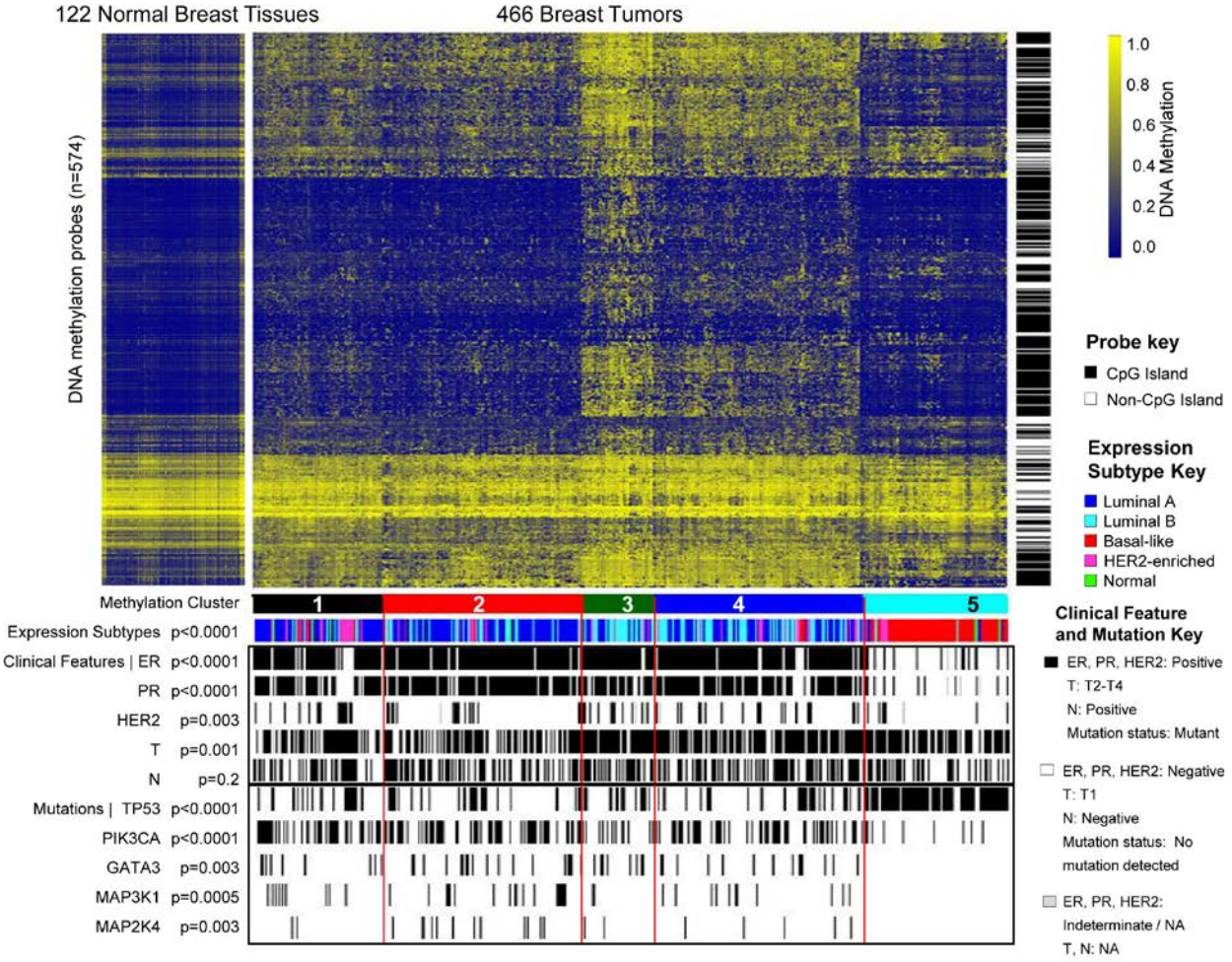
Supplemental Figure 6. Semi-supervised mRNA expression analysis. a) A semi-supervised hierarchical clustering analysis using the 1900 intrinsic gene set was performed with the resulting data tested by SigClust to identify thirteen possible groups. SMGs were mapped to the ~1900 gene "intrinsic list" where the majority of the overlap is seen within the ER+/luminal cluster, which contains the SMGs CDH1, TBX3, cMYB, GATA3, FOXA1 and MAP2K4. b) The semi-supervised mRNA expression subtypes showed a high correlation to the previously defined PAM50 subtypes, as well as high correlation to ER, PR, and HER2 status, and to TP53, GATA3, MAP3K1 and MAP2K4 mutation status.

Supplemental Figure 7



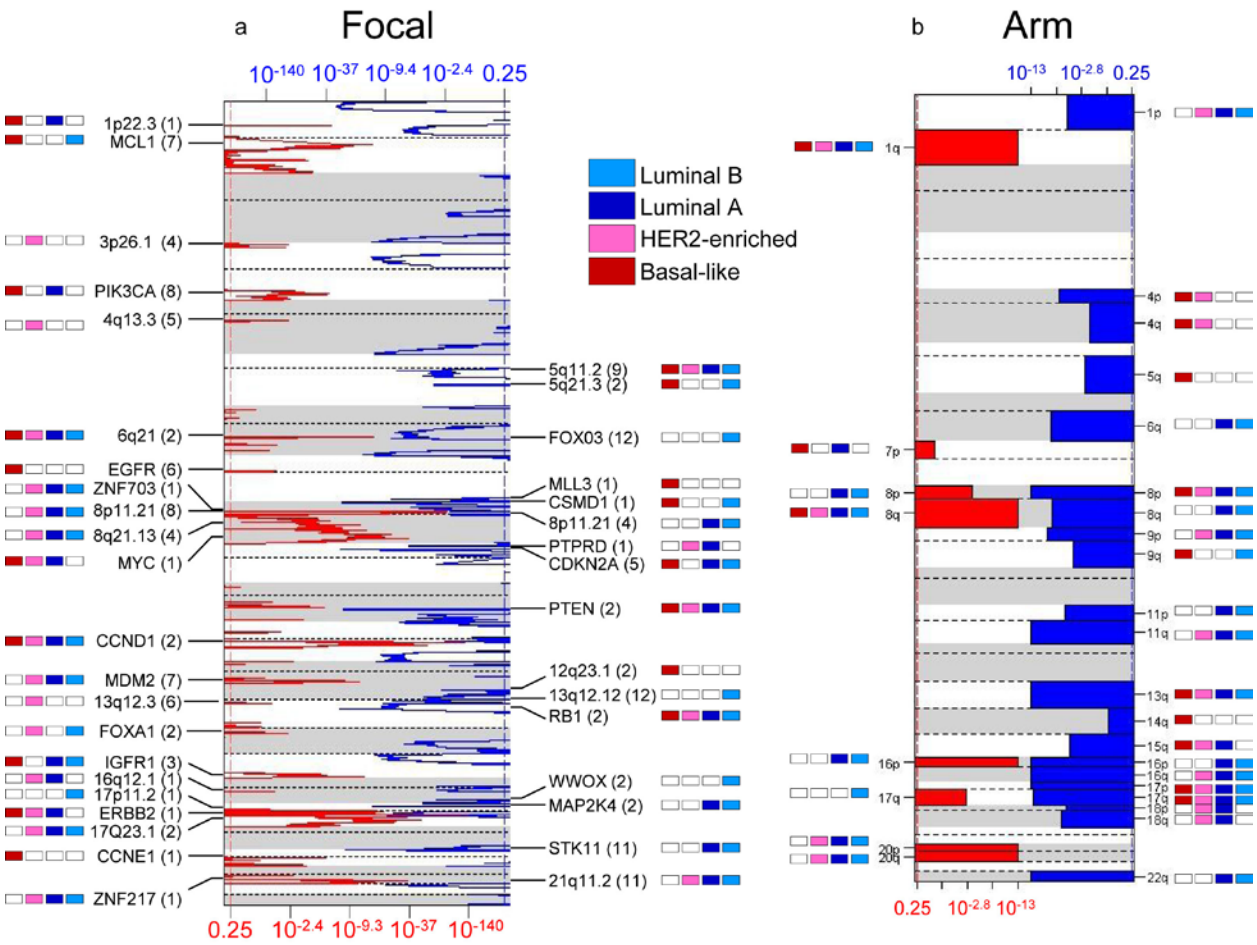
Supplemental Figure 7. MicroRNA expression analysis. Sample groups identified by NMF consensus clustering of microRNA-seq abundance profiles for 697 tumor samples. Consensus membership heatmap for seven clusters. Horizontal bands below the heatmap show (top to bottom) atypical members of each cluster (black) based on two silhouette width thresholds, PAM50 expression subtypes, then a subset of clinical covariates and mutation calls for five genes, with associated chi-square p-values. The legend indicates how colors should be interpreted for expression subtypes, covariates and mutations.

Supplemental Figure 8



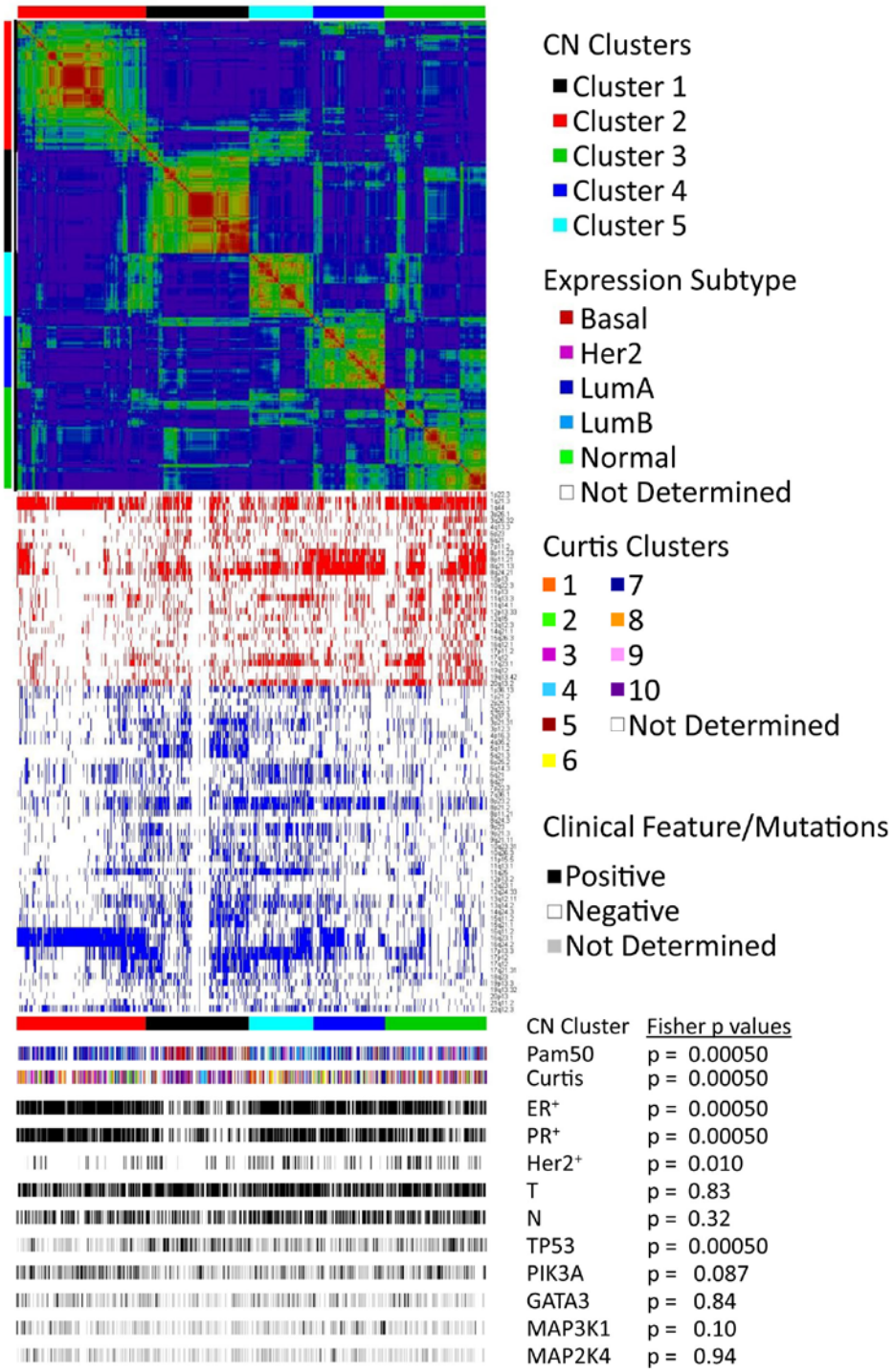
Supplemental Figure 8. DNA methylation subtypes and comparison to normal breast tissues. DNA methylation cluster membership was determined by a Recursively Partitioned Mixture Model (RPMM) for 466 breast tumors using 574 selected probes and compared to 122 breast normal samples in the same probe order. DNA methylation levels (beta value) are shown with a color spectrum; blue, no methylation to yellow, full methylation. Cluster memberships are indicated by the horizontal color bar: black Cluster 1 (n=80); red Cluster 2 (n=123); green Cluster 3 (n=44) blue Cluster 4 (n=128); cyan Cluster 5 (n=91). Molecular and clinical features as indicated in the color key. P-values for association with molecular and clinical features were calculated using a Chi-square test or Fisher's exact test, wherever applicable.

Supplemental Figure 9



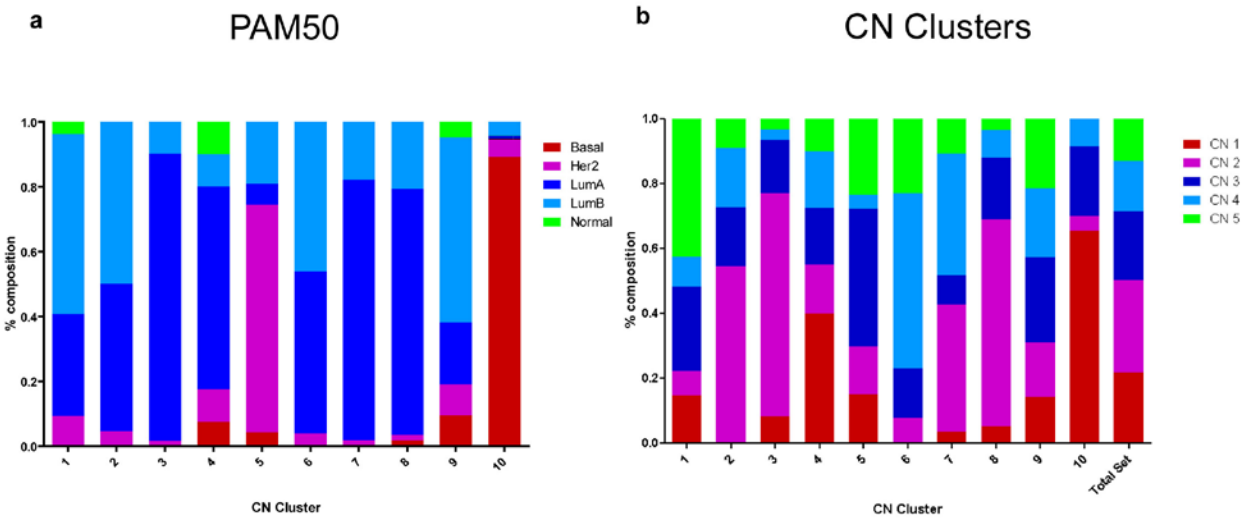
Supplemental Figure 9. DNA copy number analysis of breast tumors. a) GISTIC 2.0 analysis of Affymetrix SNP 6.0 copy number profiles from 773 tumors. Significant focally amplified (red) and deleted (blue) regions are plotted along the genome by false-discovery rates. Annotations include significant amplified and deleted regions, well-localized regions with 12 or fewer genes, and regions with known cancer genes or genes identified by genome-wide loss-of-function screens. The number of genes included in each region is given in brackets. Colored bars next to annotations indicate if overlapping peaks are present in GISTIC 2.0 analyses run on tumors with specific PAM50 expression subtypes. b) Significantly amplified (red) and deleted (blue) chromosome arms.

Supplemental Figure 10

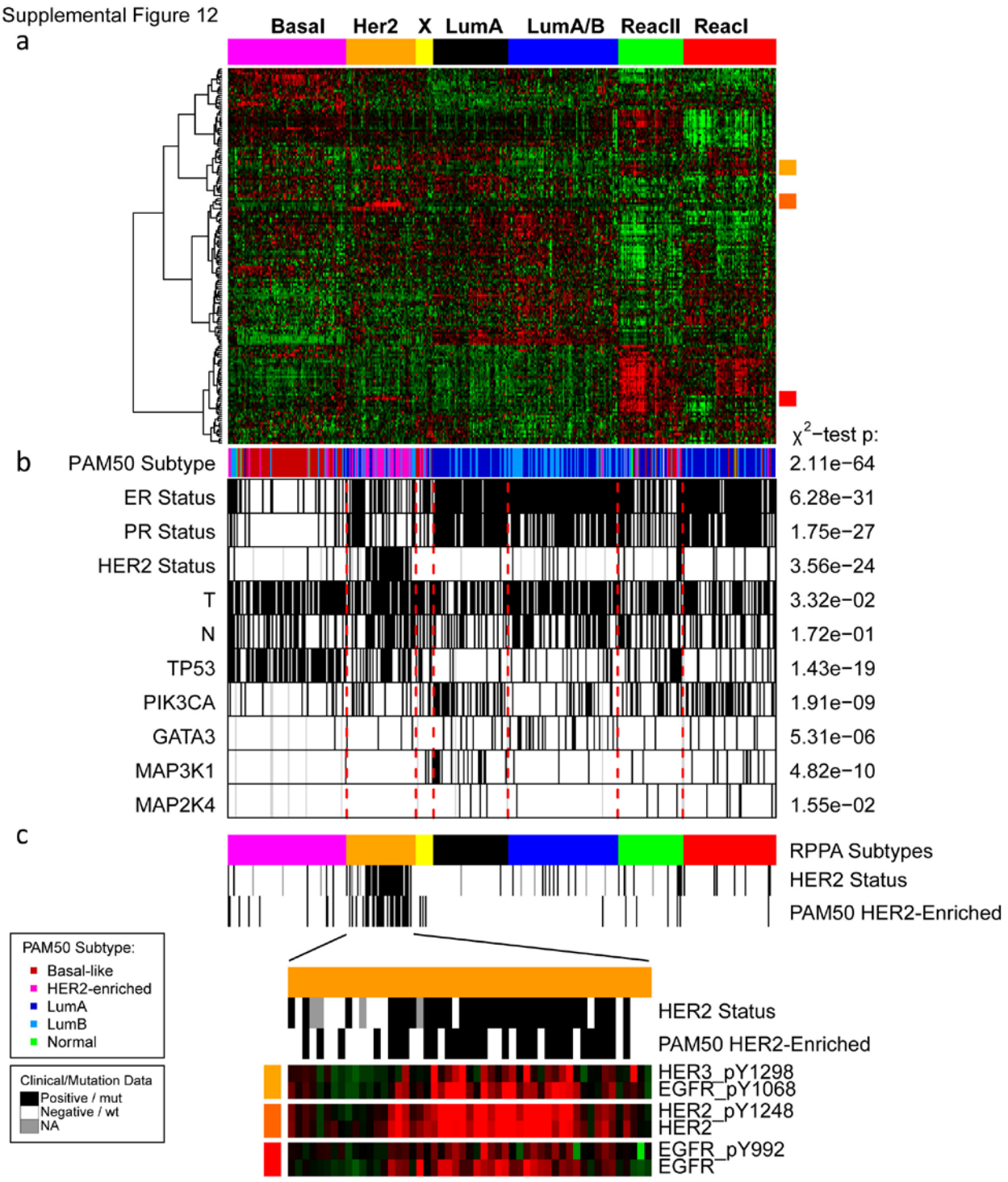


Supplemental Figure 10. Breast cancer disease subtypes identified using DNA copy number phenotypes. NMF consensus clustering was run on 773 tumors and the peaks identified by GISTIC 2.0 focal analysis. The top heatmap shows the correlation of all tumors to each other as determined by NMF clustering. The middle heatmap is colored by the presence of amplifications (red) or deletions (blue) in focal alteration (rows) in each tumor (columns). The bottom stripes show the distribution of PAM50 subtypes, clinical characteristics and mutations in each cluster. Significance values were generated by comparing the distribution of these characteristics in copy number cluster using the Fisher's exact test.

Supplemental Figure 11

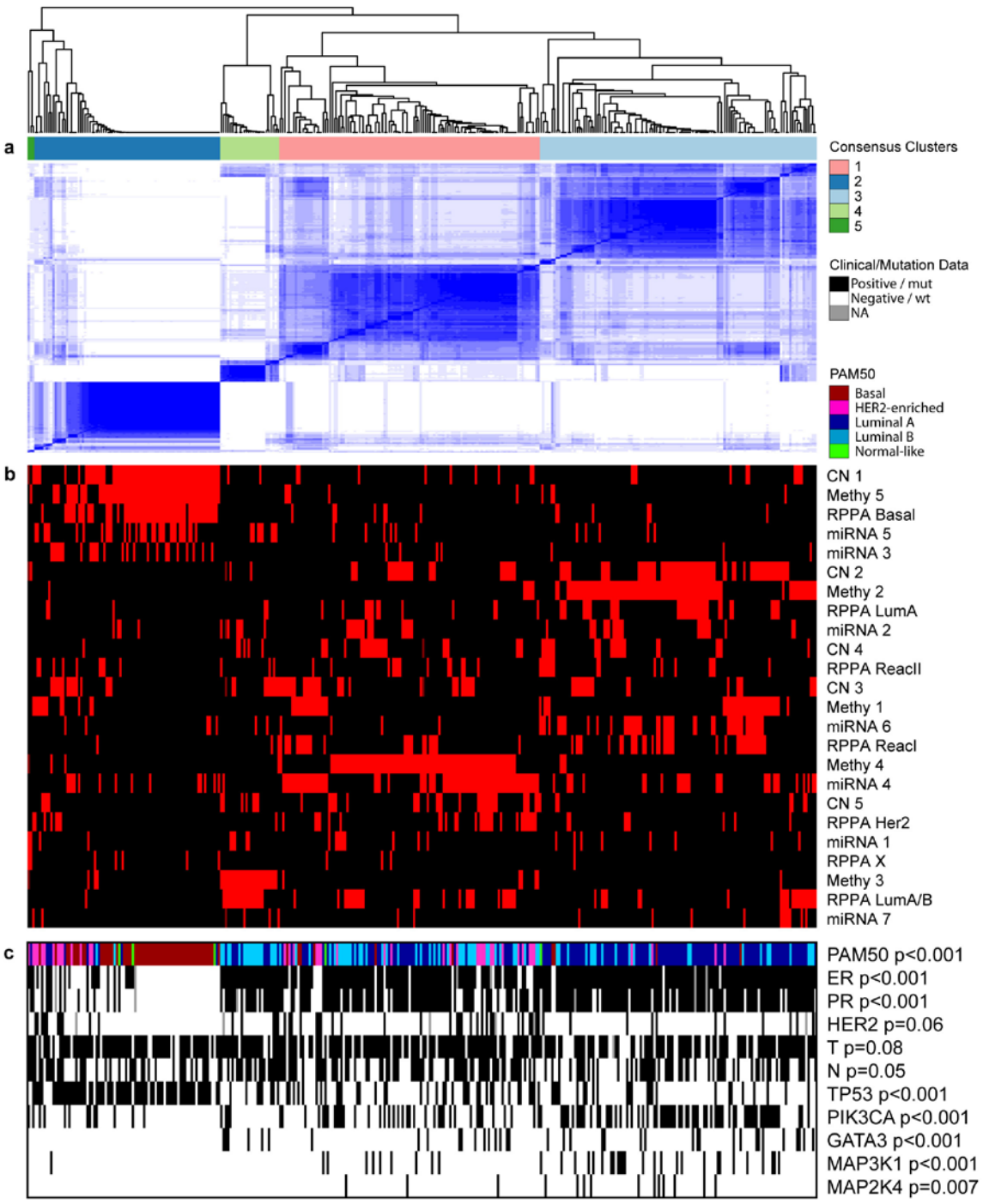


Supplemental Figure 11. Comparison of PAM50 mRNA and 5-class aCGH subtypes versus the Curtis et al. 10-class aCGH subtypes. a) PAM50 mRNA subtypes compared to Curtis et al. 10-classes, $p<0.0001$. b) TCGA aCGH subtypes compared to Curtis et al. 10-classes, $p=0.0005$.



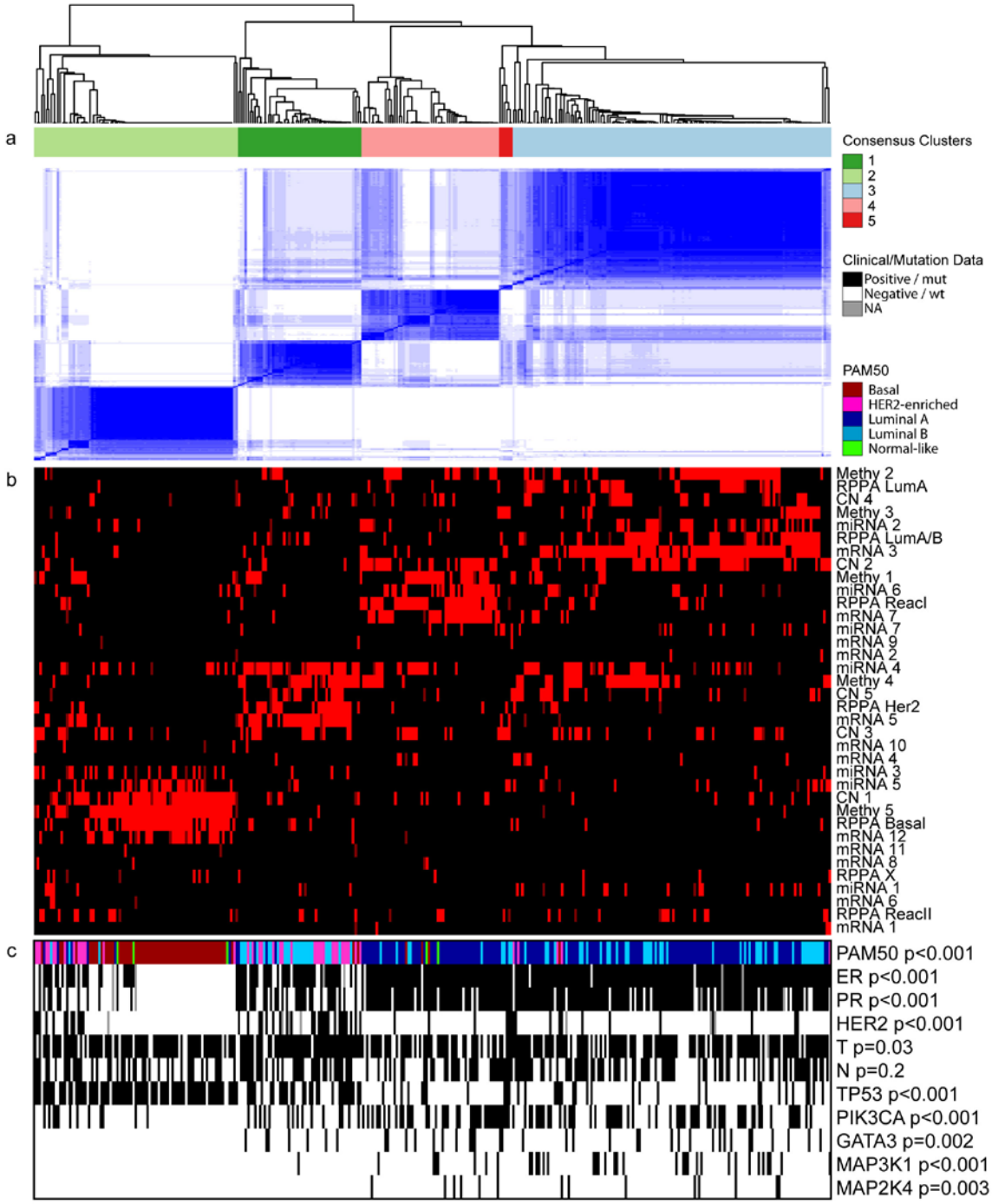
Supplemental Figure 12. Disease subtypes defined using the expression patterns of proteins and phospho-proteins. a) Samples (n=403) were ordered by unsupervised clustering using non-negative matrix factorization (NMF) with 7 clusters. Proteins were ordered using hierarchical clustering. Colored squares on right of heatmap represent regions expanded in C. b) Sample annotation ordered as in a. ER, PR, HER2, T (tumor size), and N (node) status: white, negative or T1; black, positive or T2-4; gray, unknown. Mutation status of TP53, PIK3CA, GATA3, MAP3K1 and MAP3K4: white, wild type; black, mutated; gray, unknown. c) Expanded view of the HER2 group with expression shown for EGFR, HER2, and HER3 antibodies.

Supplemental Figure 13



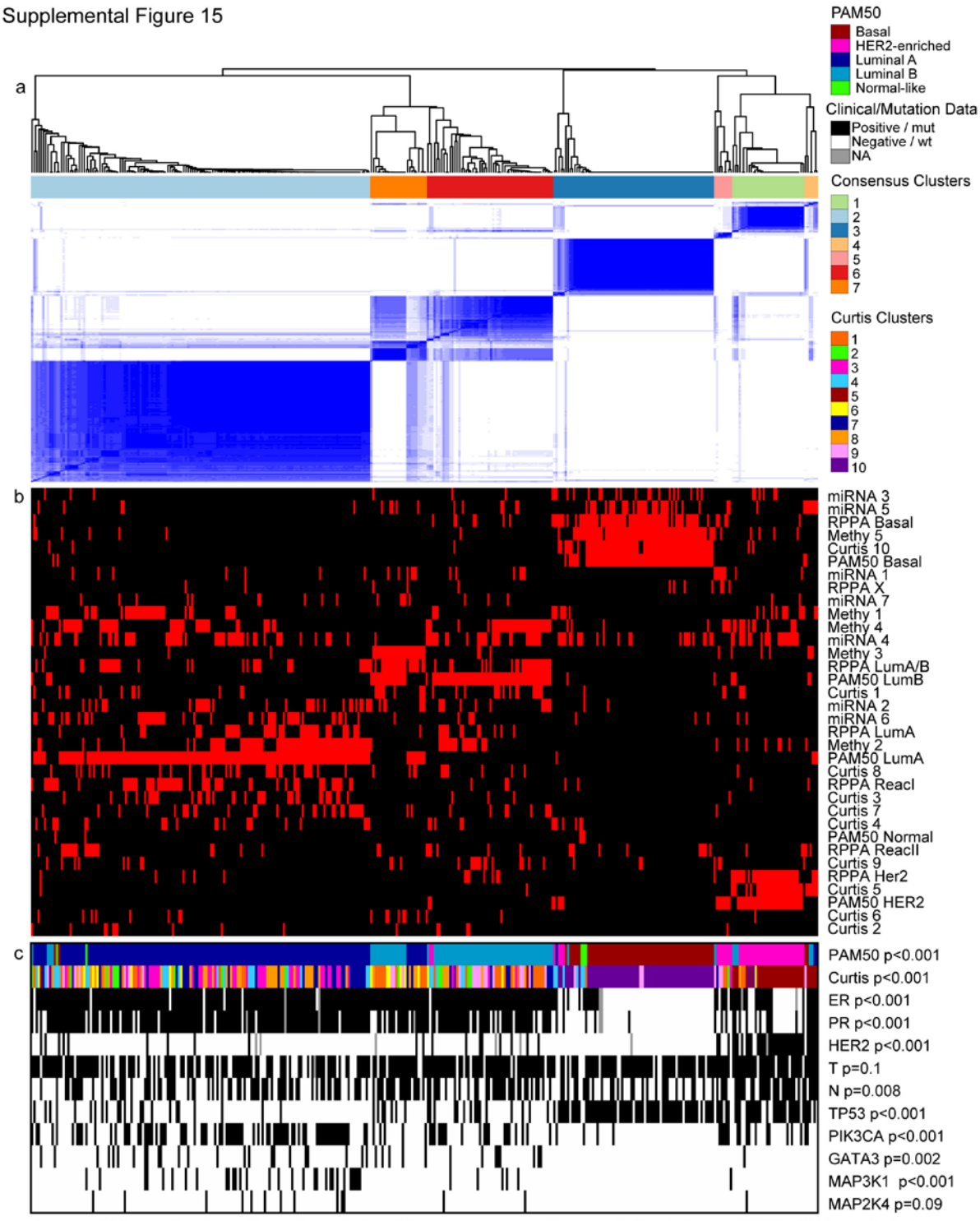
Supplemental Figure 13. Coordinated analysis of breast cancer subtypes defined from four different genomic/proteomic platforms evaluated excluding mRNA subclasses. a) Consensus Clustering (CC) analysis of the disease subtypes defined using four different technologies (excluding mRNA expression) identifies 4 groups (samples, n=348). The blue and white heatmap displays sample consensus. b) Heatmap display of the disease subtypes as defined independently by microRNAs, DNA methylation, copy number, and RPPA expression. Red bar indicates membership of a cluster type. c) Associations of the four CC-defined groups with molecular and clinical features. P-values were calculated using a Chi-square or Fisher's Exact test.

Supplemental Figure 14



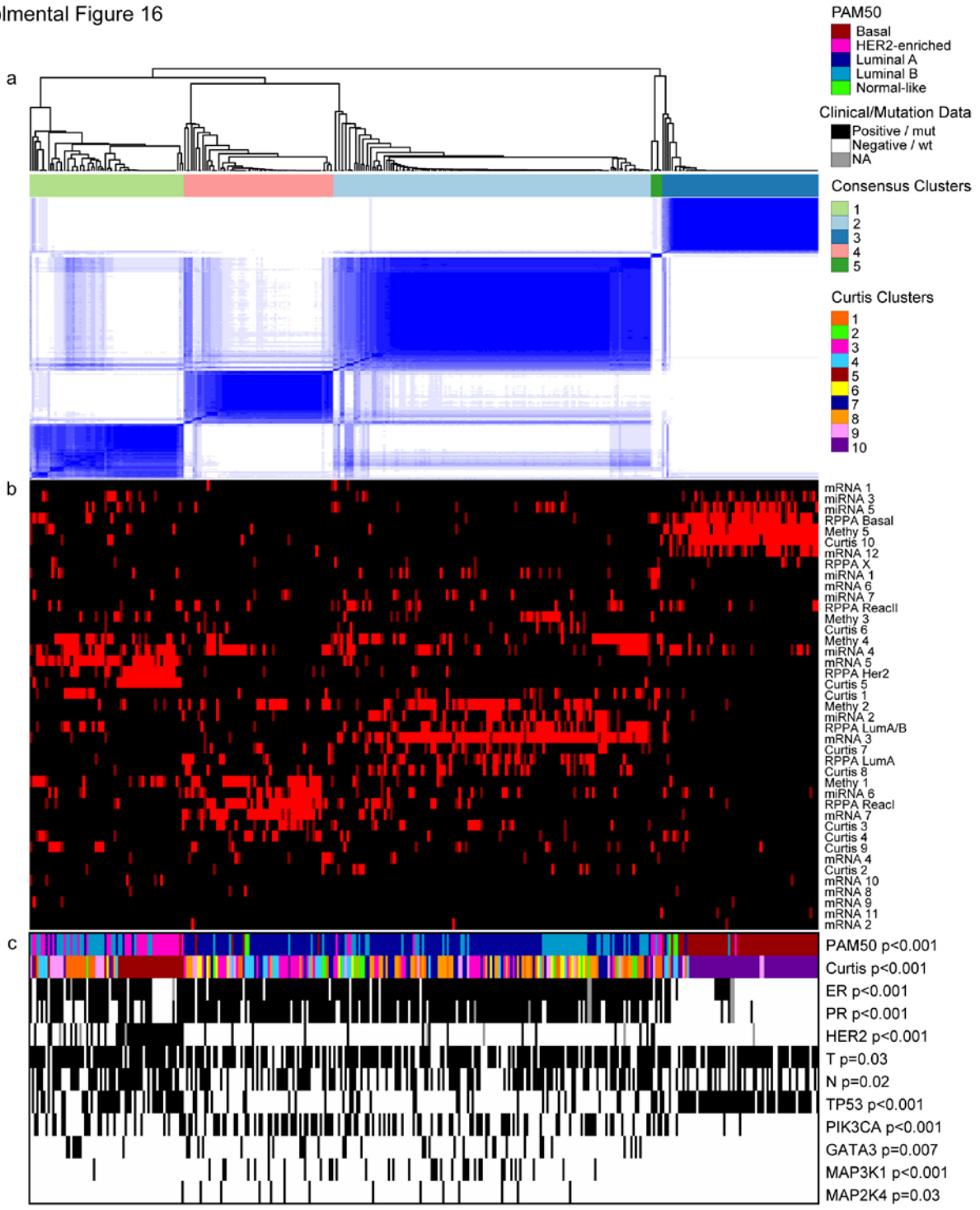
Supplemental Figure 14. Coordinated analysis of breast cancer subtypes defined from miRNA, DNA methylation, DNA copy number, protein, and 12-class unsupervised mRNA inputs. a) Consensus clustering analysis of the disease subtypes defined using five different technologies identifies at k=5, 4 main groups (samples, n=348). The blue and white heatmap displays sample consensus. b) Heatmap display of the disease subtypes as defined independently by microRNAs, DNA methylation, copy number, 12-class unsupervised mRNA expression, and RPPA expression. Red bar indicates membership of a cluster type. c) Associations of the four CC-defined groups with molecular and clinical features. P-values were calculated using a Chi-square or Fisher's Exact test.

Supplemental Figure 15



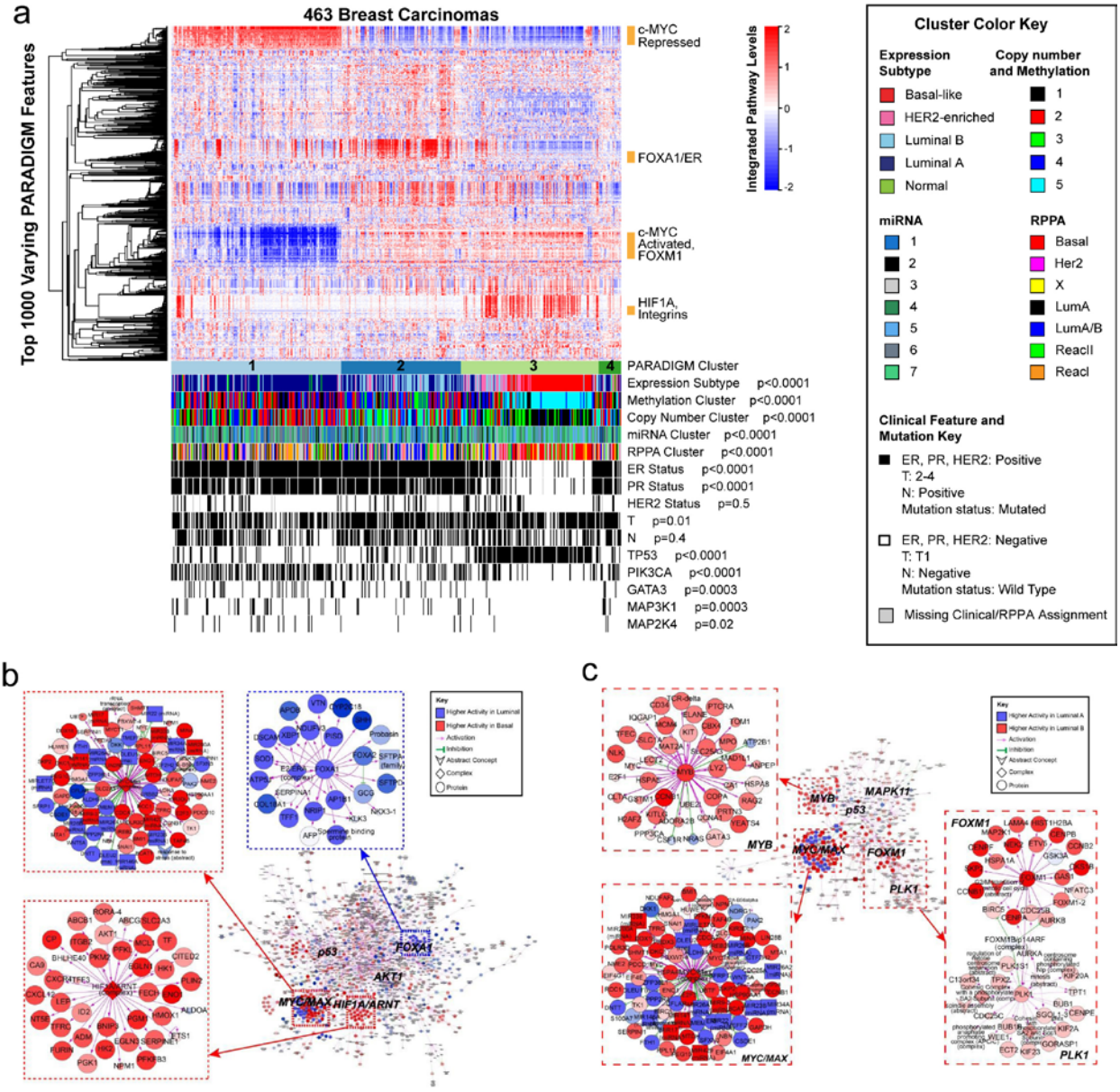
Supplemental Figure 15. Coordinated analysis of breast cancer subtypes defined from miRNA, DNA methylation, protein, PAM50 mRNA, and Curtis 10-class copy number. a) Consensus clustering analysis of the disease subtypes defined using five different technologies identifies 4-5 main groups at k=7 (samples, n=348). The blue and white heatmap displays sample consensus. b) Heatmap display of the disease subtypes as defined independently by microRNAs, DNA methylation, copy number, mRNA expression, and RPPA expression. Red bar indicates membership of a cluster type. c) Associations of the four CC-defined groups with molecular and clinical features. P-values were calculated using a Chi-square or Fisher's Exact test.

Supplmental Figure 16



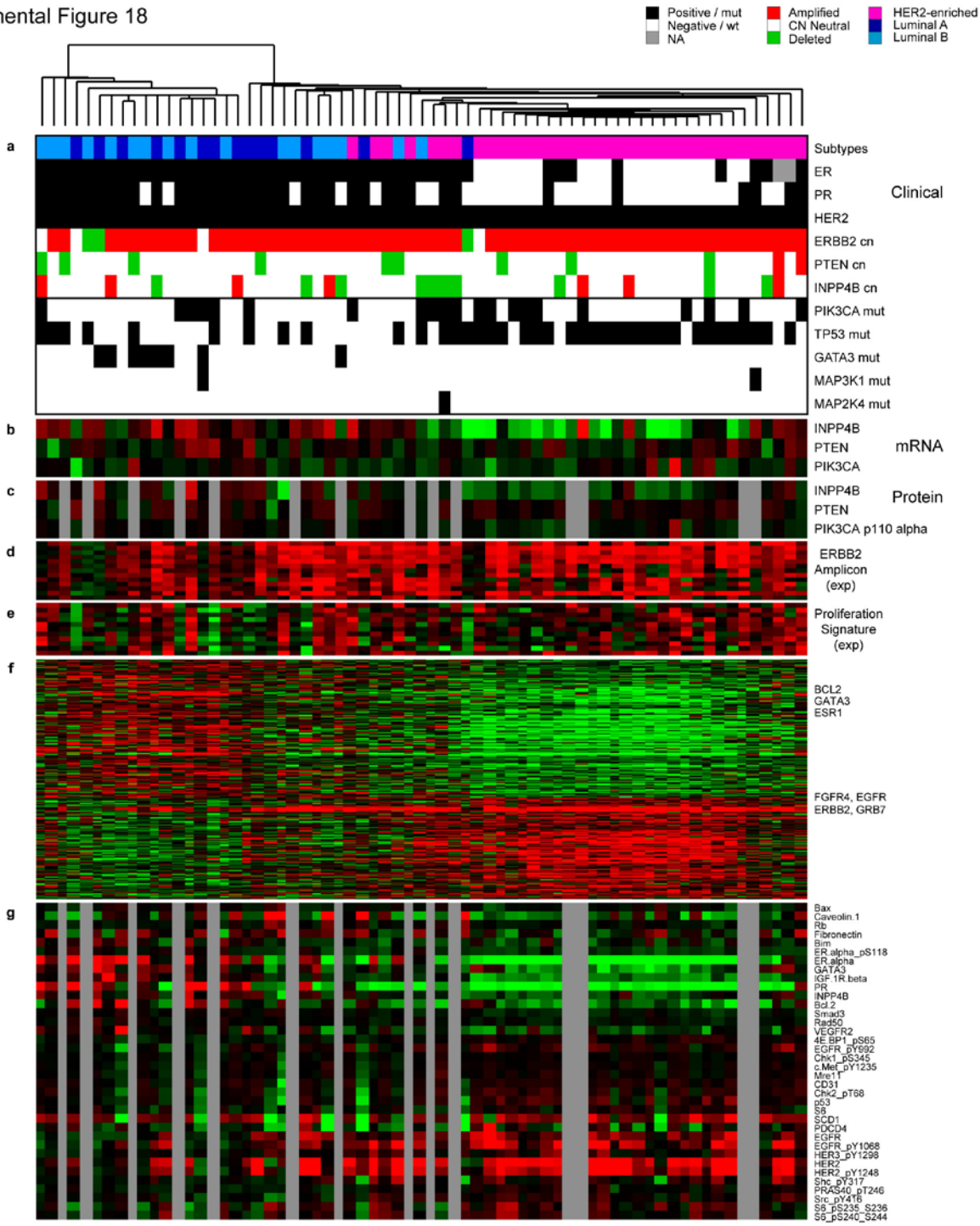
Supplemental Figure 16. Coordinated analysis of breast cancer subtypes defined from miRNA, DNA methylation, protein, 12-class unsupervised mRNA, and Curtis 10-class copy number. a) Consensus clustering analysis of the disease subtypes defined using five different technologies (with gene expression classes coming from an unsupervised analysis) identifies 4 main groups (samples, n=348). The blue and white heatmap displays sample consensus. b) Heatmap display of the disease subtypes as defined independently by microRNAs, DNA methylation, copy number, mRNA expression, and RPPA expression. Red bar indicates membership of a cluster type. c) Associations of the four CC-defined groups with molecular and clinical features. P-values were calculated using a Chi-square or Fisher's Exact test.

Supplemental Figure 17



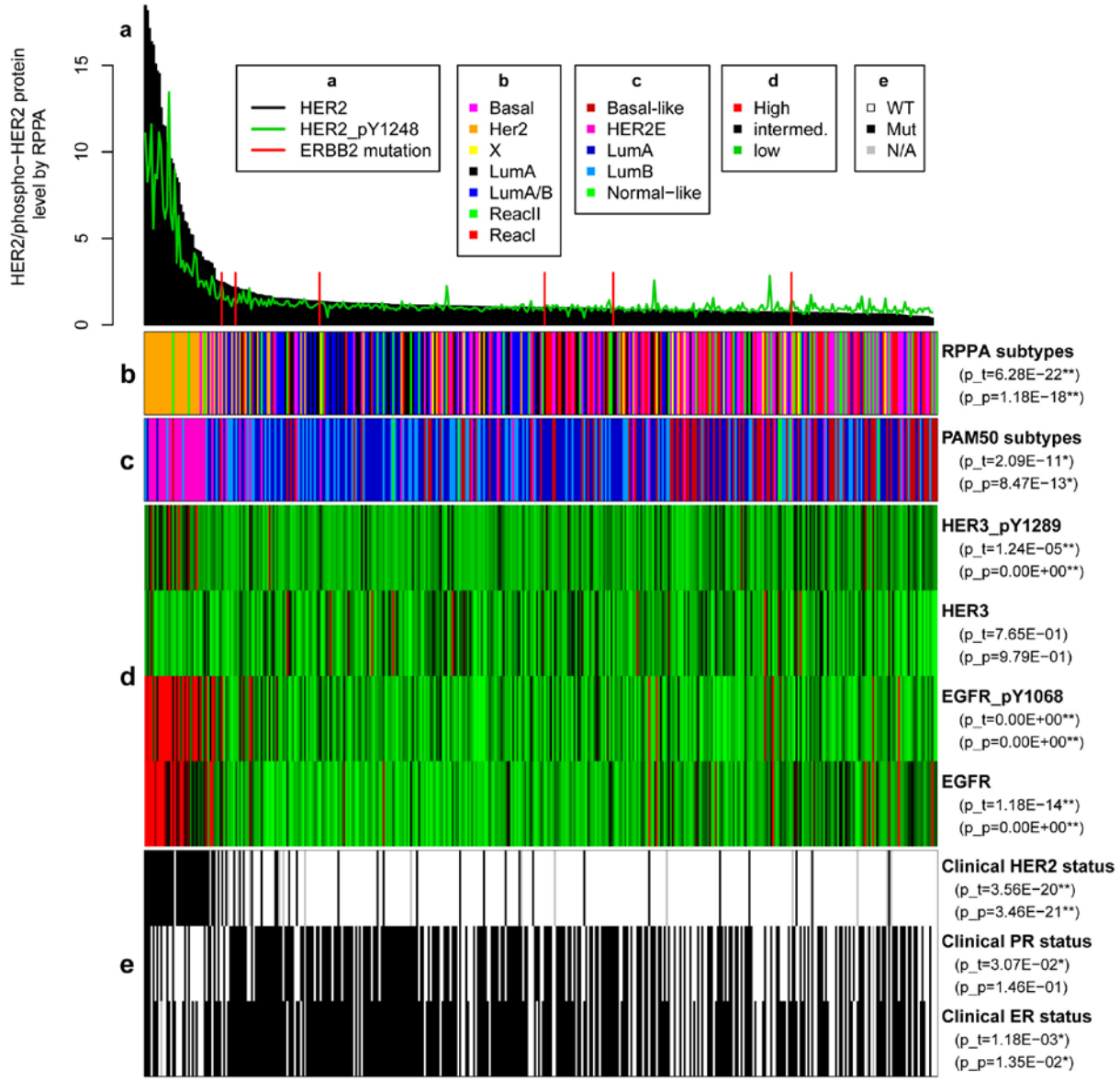
Supplemental Figure 17. Integrated pathway analyses identified using PARADIGM. a) Heatmap display of top 1000 varying pathway features within PARADIGM consensus clusters. Samples are arranged in order of their consensus cluster membership and PAM50 subtype, methylation, copy number, microRNA and RPPA cluster membership assignments, ER, PR and HER2 status, tumor size, node status and mutation status for *TP53*, *PIK3CA*, *GATA3*, *MAP3K1* and *MAPK2K4*, for each sample is displayed below. For each variable, the p-value from the $\chi 2$ test of associations with consensus clusters was displayed. Selected pathways showing distinct activation patterns among the consensus clusters were labeled (vertical orange bar). b) Differentially activated pathway features between Basal-like and Luminal (A+B) breast cancers. Largest interconnected regulatory subnetwork of differentially activated IPLs is displayed, with network hubs showing interconnectivity > 20 edges labeled. Larger views of the HIF1 □/ARNT, MYC/MAX, and FOXA1/ER hubs are shown as insets. Color intensity reflects activity differences between subtype (red: higher in basal, blue: higher in luminals). Purple arrows denote activation. Green tees represent inhibition. Node shapes reflects pathway concept type (inverted v: abstract concept, diamond: complex, circle: protein). Node size is scaled to the significance of differential activation. c) Differentially activated pathway features between Luminal A and Luminal B breast cancers. Largest interconnected regulatory subnetwork of differentially activated IPLs is displayed, with network hubs showing interconnectivity > 15 edges labeled. A larger view of the MYC/MAX, MYB, FOXM1 and PLK1 hubs are shown as insets. Color intensity reflects activity differences between subtype (red: higher in Luminal B, blue: higher in Luminal A). Purple arrows denote activation. Green tees represent inhibition. Node shapes reflects pathway concept type (inverted v: abstract concept, diamond: complex, circle: protein). Node size is scaled to the significance of differential activation.

Supplemental Figure 18



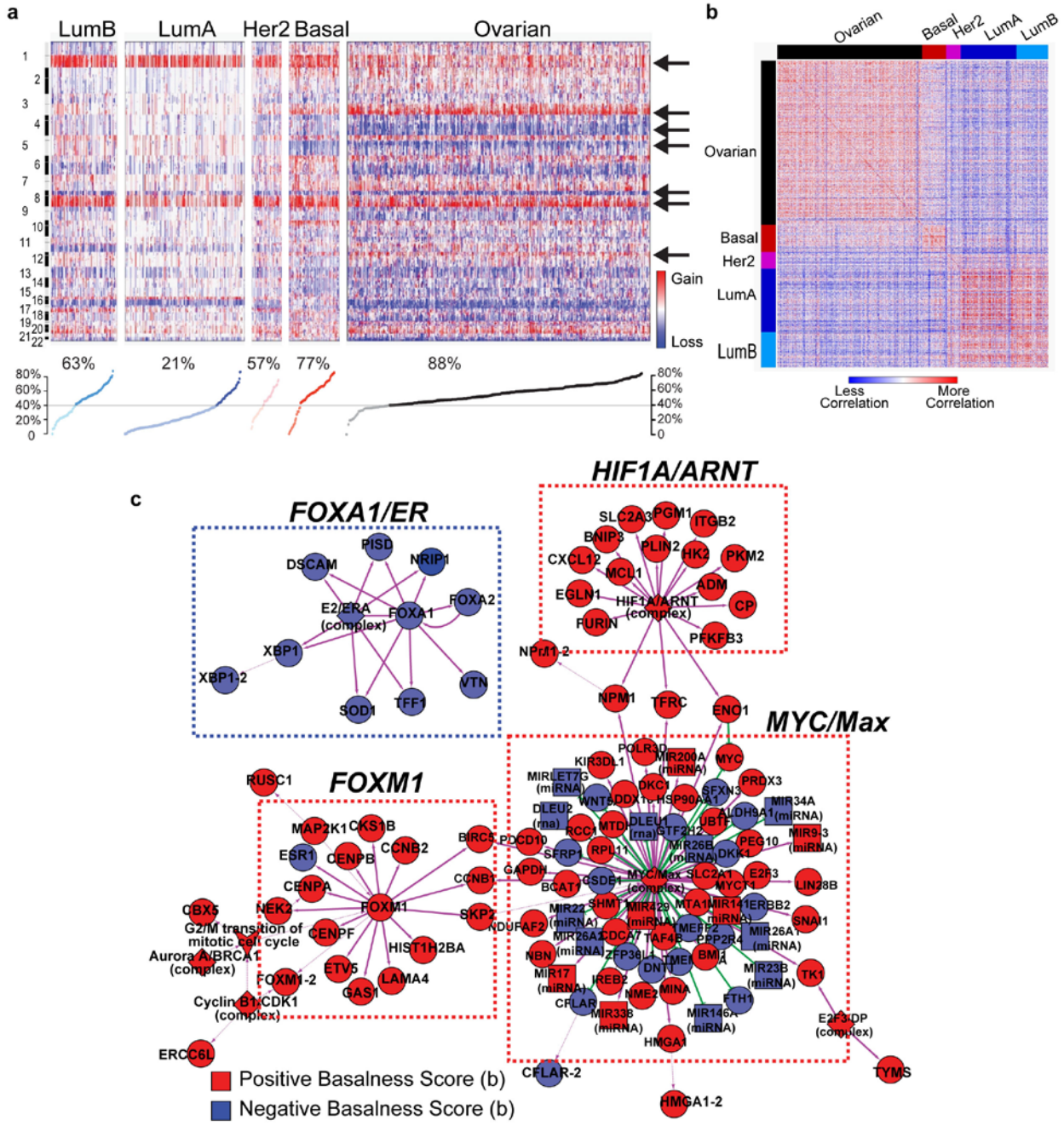
Supplemental Figure 18. Comparison of Luminal and HER2E phenotypes in clinically HER2-positive tumors. A supervised analyses of the mRNA expression data was performed to find genes that differed between 36 clinically HER2-positive, HER2E samples and 31 clinically HER2-positive, Luminal subtype samples (14 Luminal A, 17 Luminal B), and to find proteins that differed in a subset of this list (29 HER2E versus 12 Luminal A and 12 Luminal B tumors). a) Tracks for PAM50 subtypes, clinical ER, PR, and HER2, copy number and mutation status. b) mRNA expression for INPP4B, PTEN, and PIK3CA. c) Protein expression for INPP4B, PTEN, and PIK3CA. d) Gene expression surrogate for the HER2/ERBB2 amplicon. e) 11-gene proliferation signature from the PAM50 assay. f) 302 genes significantly altered between HER2E versus Luminal subtype samples. g) 36 RPPA proteins significantly different between HER2E versus Luminal subtype samples.

Supplemental Fiure 19



Supplemental Figure 19. Comparison RPPA-defined HER2 protein levels with clinical, genomic, and other proteomic features. a) Using the RPPA data, the tumors were placed into rank order based upon the quantitative protein expression levels of HER2, with p-HER2 shown as a green line, and samples with somatic HER2 mutations as a vertical red line. Comparison of HER2-protein levels versus b) NMF RPPA cluster subtypes, c) PAM50 mRNA expression subtypes, d) p-HER3, HER3, p-EGFR/HER1 and total EGFR/HER1 protein levels, e) clinical biomarker status for HER2, PR and ER.

Supplemental Figure 20



Supplemental Figure 20. Comparison of Breast and Serous Ovarian carcinoma. a) Copy number landscapes of Breast cancers and Serous Ovarian carcinomas. Copy number alterations (gains, red; losses, blue) shown along the genome (rows) across tumor samples (columns) that include from left to right: 220 Luminal A, 122 Luminal B, 56 HER2E, 93 Basal-like, and 558 ovarian tumor samples. The arrows on the right identify regions of change in common between breast Basal-like and Serous Ovarian carcinoma, while the display below the heatmap identifies the fraction of the genome that is copy number altered according to tumor subtype. b) The heatmap shows Pearson R correlation values derived by comparing thresholded copy number of 20,631 genes and miRNAs. The color scale is set so that that correlations where R > .5 have the maximum red color and those where R < -0.1 have the maximum blue color. Tumors with no gene level copy number changes were excluded from the analysis. Total number of tumors in each group: Ovarian = 557, Breast Basal-like = 97, Breast HER2E = 92, Breast Luminal A = 215 and Breast Luminal B = 121. c) PARADIGM analysis reveals common networks in Basal-like and Serous Ovarian. Ovarian/Basal-like to luminal associations were assessed using a "Basalness" score and only significant features were retained. The two largest sub-networks linking significant features through regulatory interactions are shown as a Cytoscape plot: positive values (red) indicate higher activity in Basal-like breast cancers and negative values (blue) indicate lower activity. Node shapes correspond to complexes (diamonds), proteins (circles) microRNAs (squares), and cellular processes (inverted v shapes). Network hubs (> 5 connections) are highlighted in boxes and labeled by gene name.

Supplemental Methods

I. Biospecimen Collection and Processing

Sample inclusion criteria

Biospecimens were collected from newly diagnosed patients with invasive breast adenocarcinoma undergoing surgical resection and had received no prior treatment for their disease (chemotherapy or radiotherapy). The targeted accrual was 800 ductal, 200 lobular, and a mixture of 100 other breast cancer subtypes. Institutional review boards at each tissue source site reviewed protocols and consent documentation and approved submission of cases to TCGA. Cases were staged according to the American Joint Committee on Cancer (AJCC) staging system. Each frozen primary tumor specimen had a companion normal tissue specimen which could be blood/blood components (including DNA extracted at the tissue source site) (n=684), adjacent normal tissue taken from greater than 2 cm from the tumor (n=76), or both (n=65). Three cases had a qualifying metastatic tumor in addition to the primary tumor. Each tumor specimen weighed at least 60 mg. Specimens were shipped overnight from 18 tissue source sites (TSS) using a cryoport that maintained an average temperature of less than -180°C. Each tumor and adjacent normal tissue specimen (if available) were embedded in optimal cutting temperature (OCT) medium and a histologic section was obtained for review. Each H&E stained case was reviewed by a board-certified pathologist to confirm that the tumor specimen was histologically consistent with breast adenocarcinoma and the adjacent normal specimen contained no tumor cells. The tumor sections were required to contain an average of 60% tumor cell nuclei with less than 20% necrosis for inclusion in the study per TCGA protocol requirements.

Sample Processing

RNA and DNA were extracted from tumor and adjacent normal tissue specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a *mir*Vana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp blood midi kit (Qiagen).

Each specimen was quantified by measuring Abs_{260} with a UV spectrophotometer or by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify tumor DNA and germline DNA were derived from the same patient. Five hundred nanograms of each tumor and normal DNA were sent to Qiagen for REPLI-g whole genome amplification using a 100 µg reaction scale. Only specimens yielding a minimum of 6.9 µg of tumor DNA, 5.15 µg RNA, and 4.9 µg of germline DNA were included in this study. RNA was analyzed via the RNA6000 nano assay (Agilent) for determination of an RNA Integrity Number (RIN), and only the cases with RIN >7.0 were included in this study.

At the time of the data freeze, 1,377 breast adenocarcinoma cases were received by the BCR and 72% passed quality control. The biospecimens included in this report come from 825 breast carcinoma cases included in batches 47, 56, 61, 72, 74, 80, 85, 93, 96, 103, 109, 117, 120, 124, 136, 142, 147, 155, 167, 177, 185 and 202.

II. Clinical data and quality improvement

Additional quality assurance (QA) was performed on selected data elements, in addition to the QA already performed by the TCGA Biospecimen Core Resource (BCR). The clinical data was taken from the November 2011 data archive. The clinical calls of biomarkers were supplemented with molecular data, and the stages of the cancer cases were mapped to the current staging standard. When clinical calls for ER, PR, and HER2 were available, they were used. Molecular data was only used when HER2 clinical calls were not available. In addition, certain data fields were made binary to facilitate cross-analysis with molecular data. These results are shown in Supplemental Table 1.

Clinical data QA of biomarkers and cancer stage

Clinical guidelines for determining estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) statuses for breast cancers have been established. According to the current clinical guideline jointly issued by the American Society of Clinical Oncology (ASCO) and the College of American Pathology (CAP)[1], effective January 2010, a breast tumor is called ER or PR positive if the corresponding nuclear staining is >= 1%. Prior to 2010 there was no universal standard and local hospitals used their own thresholds such as 5% or 10% for their clinical practices. Since the year of diagnosis of all breast cancer cases collected in this study ranged from 1988 to 2011, determining the clinical status for ER and PR followed a mixture of the thresholds. The situation is further complicated by the subjective nature of the percentage calls, adding additional discrepancies among different pathologists.

For HER2, following the current ASCO/CAP guideline[2], a breast tumor with an immunohistochemistry (IHC) value of 0 or 1+ is called “Negative”, IHC level 3+ is called “Positive”, and level 2+ is called “Equivocal”. Florescence in situ hybridization (FISH) is used to determine the final status of “Equivocal” cases where a case is called “Positive” if the FISH ratio is >=2.2, and “Negative” if the FISH ratio is <=1.8. Studies have shown a high concordance of the negative calls by IHC (0 and 1+) and FISH (97% or 99%) and the concordance of the positive calls by IHC (3+) and FISH is also high, at about 90%[3,4].

A number of issues were identified for ER and PR status from the original clinical dataset *clinical_patient_public_brca.txt*. Note that here an issue does not necessarily mean a problem; 1) for ER/PR nuclear staining percentage level of <10%, there were cases called “Positive” (10 for ER and 33 for PR), or “Negative” (41 for ER and 61 for PR); 2) there were cases with higher staining level that were not called “Positive” (6 for ER and 7 for PR); 3) hundreds of cases did not have staining levels but had “Positive” or “Negative” clinical statuses (506 for ER and 519 for PR); 4) dozens of cases did not have a clinical status (38 for ER and 41 for PR).

The following issues were identified for HER2: 1) 199 cases had IHC_Level of 1+, with consistent IHC_Status of “Negative”. Most of them had a consistent FISH_Status value or null, but two of them were FISH “Positive”; 2) 68 cases had IHC_Level 3+ with matching IHC_Status of “Positive”, but 2 of the 10 with FISH_Status values were FISH “Negative”; 3) 7 cases were scored IHC_Level 1+, contradicting IHC_Status of “Positive”; 4) 29 cases were scored IHC_Level 2+ but were called IHC_Status “Positive” (3 cases) or “Negative” (26 cases), and only 10 of them had FISH results; 5) 302 cases were IHC_Level “Not Available”, but had values for IHC_Status, some also had FISH_Status.

Cancer staging standards have also evolved. Currently the AJCC (American Joint Committee on Cancer) Cancer Staging Manual, 7th Edition (2010) is followed[5]. Prior to this, the 6th Edition was released in 2002, thus the breast cancer cases collected for this study were staged

using a mix of standards. For AJCC pathologic stage fields of T, N, M, Stage, and AJCC_Edition, the main issues were: 1) Multiple AJCC_Editions were used due to the time span of the cases for study (1988-2011); 2) 278 cases did not report AJCC_Edition; 3) For some cases with an AJCC_Edition, the T, N, and M did not match the stage definition of that edition; 4) There was at least one case of inconsistent N value and positive node count. The identified issues were reported to BCR for possible follow-up corrections. Some errors were subsequently corrected, when and only when the corresponding TSS was able to identify the source of the error and issued a correction. For many cases, the TSS chose to retain the IHC_Status. This is an approach consistent with our analysis below.

There were issues that could not be readily resolved due to the data collection criteria. There were also issues due to the data form design and their correction requires a revision of the data form. For example, the mixed ER/PR thresholds could be fixed by applying a consistent threshold for research purposes, if the numerical nuclear staining level (IHC_Level) values were available. However, the corresponding data fields were designed for bracketed values, with the lowest range of <10%, followed by 10-19%, 20-29%, and so on. Mapping to the 1% threshold, the current standard starting from 2010, is therefore impossible. A case of 7% staining level could have been called "Positive" or "Negative" depending on whether a 5% or a 10% threshold was used. In addition, we do not know how reliable the ranges checked on the forms were since the clinical focus is always whether ER/PR is positive which impacts the clinical intervention. Therefore, we made a decision not to use the *ER/PR IHC_Level* field for the current analysis.

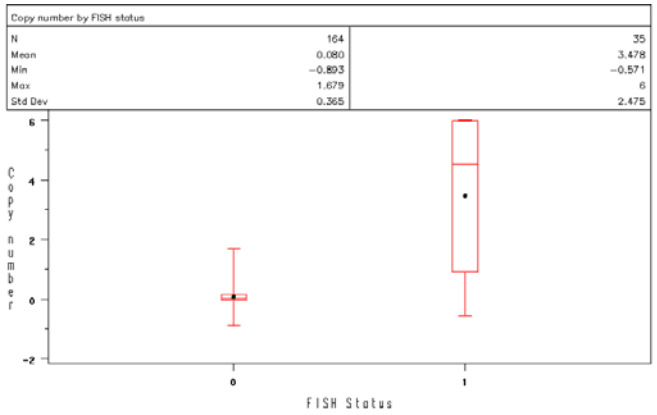
Improvement of clinical calls of HER2 using molecular data

In contrast to ER and PR, HER2 IHC status calls did not highly correlate with gene expression (AUC=0.8, *data not shown*), and this lower correlation has been previously reported[6]. Given the results from the QA analysis, we attempted to systematically determine HER2 clinical statuses based on the ASCO/CAP guideline, and explored the possibility of supplementing these calls with DNA copy number data (at 17q12.2064) derived from SNP chips (Affymetrix 6.0)

A standard logistic model was developed for prediction of FISH status by copy number data using SAS. A total of 199 patients with FISH status (Negative=164, Positive=35) also had relative copy number values. The descriptive statistics on copy number by FISH calls are shown in Figure II.1.A. There was a significant prediction of FISH status by the predictor, $\chi^2(1) = 105.3318$, $p < 0.001$, Negelkerke $R^2 = 0.679$. The prediction of FISH by Copy_Number followed, logit(Pr(Event = 1)) = -3.2 + 2.3405 * (Copy_Number), with B = 2.3405, exp (B) = 10.386, $\chi^2 (1) = 18.8537$, $p < 0.0001$. There was no significant difference between the observed and the predicted FISH status, Hosmer-Lemeshow $\chi^2 (8) = 4.8$, $p = 0.8$. The overall classification rate was good with ROC curve shown in Figure II.1.B, with an AUC = 0.92.

DNA copy number derived from SNP chips was a good predictor of FISH status, and this is consistent with a previous study where copy number was derived from aCGH[7]. We also developed a logistic model to predict HER2 clinical calls strictly following the ASCO/CAP guideline, and the results were of similar quality (AUC=0.912, *data not shown*), which suggests its high predictability by copy number as well. In addition, we modeled copy number prediction of the combined IHC_FISH calls following the ASCO/CAP guideline, supplemented by FISH_Status when IHC_Level was unavailable, which was further supplemented by IHC_Status when neither IHC_Level nor FISH_Status was available. The model was not as successful (data not shown), which was probably due to the less predictability by copy number for IHC_Status. We also tried to model FISH prediction by copy number from cases with IHC_Level = 2+ only, however the model was not as good probably due to the small sample size (*data not shown*).

A



B

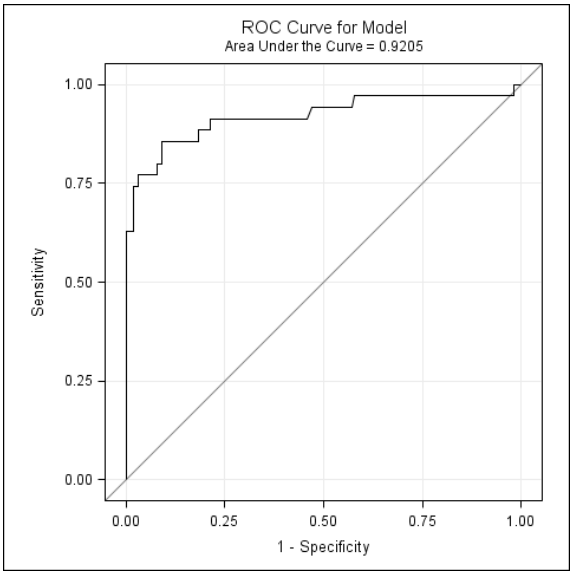


Figure II.1. Prediction of FISH status by HER2 copy number. (A) Descriptive statistics of HER2 copy number by FISH status. (B) ROC curve with AUC=0.92.

HER2 final clinical status

Based on the results of all these analyses, we derived the HER2_Final_Status in the following three steps. This field is shown as "HER2 Final Status" in Supplemental Table 1.

Step 1. *HER2 calls following the ASCO/CAP guideline.* HER2 calls were made strictly following the ASCO/CAP guideline, resulting in 372 "Negative", 85 "Positive", 32 "Equivocal", and 302 "NA" calls.

Step 2. *Supplementing with FISH results.* For cases lacking of the IHC_Level, FISH calls were used resulting in 42 new calls (33 "Negative" and 9 "Positive").

Step 3. *Supplementing with copy number calls.* In using copy number to predict FISH status, we did not use established methods to find a single optimal cut-point for minimizing both false positive and false negative calls; instead, we took a more conservative approach using two cut-points, because our goal was to supplement FISH calls with copy number calls to achieve an overall low false-prediction rate which we chose to be 5%. Thus a cut-point for positive calls of Copy_Number >= 1.69, and a cut-point for negative calls of Copy_Number <= 0.55 were chosen, which yielded a 95.7% (22/23) prediction accuracy for positive calls and a 96.8% (149/154) prediction accuracy for negative calls, respectively.

Using this approach, we only supplemented "NA" or "Equivocal" calls from the preceding steps with FISH or copy number information but not correcting any inconsistent "Positive" or "Negative" calls. The final HER2 clinical calls for cases that have copy number data were: 642 "Negative", 114 "Positive", 10 "Equivocal", and 15 "Not Performed". In the 466 data freeze set, there were only 10 cases that did not have a "Positive" or "Negative" HER2 call.

AJCC Stages mapping to 7th Edition

A script was developed to map the mixed editions to the current AJCC 7th Edition so that analyses could be performed based on a self-consistent staging system. Here are the main results: 1) Out of the 278 cases for which no edition information was available, 108 were recognizable as 6th or 7th Edition and were converted to 7th Edition; 2) There were 40 5th Edition cases; 15 of

which were converted to 7th edition. Specifically, those cases with no positive nodes (N0), positive ipsilateral internal mammary nodes (N3), or tumor stage T3 (stage not affected by nodal status) could be converted; 3). The remaining samples were either 6th Edition which were all converted, or already the 7th Edition.

Categorization of T and N

Clinical data fields of pathologic stages T and N were made binary for use in molecular data analysis. T was coded as T1 and T_Other, corresponding to smaller tumor size (<= 2 cm) and larger tumor size (> 2 cm), respectively, and TX was coded as null for missing value. N was coded as Negative corresponding to N0, and Positive corresponding to N1-N3, respectively. M was coded as Positive for M1, and Negative for others, respectively, and missing values were allowed.

Summary

For ER and PR, we found a limited number of cases with inconsistencies between clinical statuses and nuclear staining levels, and about 5% cases with missing clinical calls. For HER2, there were also a limited number of cases showing inconsistencies between IHC statuses, IHC levels, and/or FISH statuses, but most importantly, about 40% of the cases could not be directly mapped to the current ASCO/CAP guideline due to missing values, likely because of historical reasons. We were able to correct some of the inconsistencies.

Tissue Source Site provided calls were used for clinical ER and PR status. For HER2, copy number can predict FISH status very well. Therefore, we were able to strictly follow the current ASCO/CAP guideline for a consistent new clinical status call, supplementing with FISH calls, and further supplementing with FISH calls predicted by the copy number data. Such adjustments were only made to cases that were previously either Equivocal or NA from the preceding steps. For AJCC stages, we focused on converting staging calls to be consistent with AJCC Edition 7 where possible (Supplemental Table 1).

Clinical data can be found in Supplemental Table 1 and from the DCC Data portal (http://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp) or from the manuscript associated webpage (http://tcga-data.nci.nih.gov/docs/publications/brca_2012/).

III. Exome Sequencing

Whole genome amplified or genomic DNA provided by the TCGA BCR was used for exome sequencing at the Washington University Genome Institute. Libraries were constructed using ligation of Illumina adaptors to sheared whole genome amplified DNA. A Solid Phase Reversible Immobilization (SPRI) bead cleanup procedure was conducted to select size fractions between 300 and 500bp. Hybridizations were performed using customized versions of the Agilent SureSelect All Exome v2.0 kit or Nimblegen SeqCap EZ Human Exome v2.0. qPCR was used to determine the quantity of captured library necessary for loading on an Illumina Hi-Seq 2000 in order to produce greater than 10Gbp of sequence. As the throughput of the sequencing instrument increased, sample barcoding and multiplexing techniques were employed to efficiently utilize the output of sequencers. At least 70% coverage at 20x depth of the ~34 Mbp CDS target[8] and at least 90% genotype concordance with SNP array data were required for each sample to pass quality control.

Alignment, De-duplication, and BAM file generation.
Tumor and normal samples for 507 breast cancer cases were sequenced and aligned independently. For each sample, filter-passed reads were aligned to the NCBI build 37 (hg19) human reference sequence (GRCh37-lite) using BWA[9] v0.5.9. BAM files were generated using SAMtools[10] r963; duplicates were marked with Picard (http://picard.sourceforge.net) v1.46. BAM files for QC-passed samples were submitted to the dbGAP database.

Mutation Detection and Annotation
Mutation detection and annotation were performed at the Washington University Genome Institute as follows. Somatic single nucleotide variants (SNVs) detected by VarScan 2[11] and SomaticSniper[12]; somatic insertion/deletions (indels) were detected by VarScan 2 and GATK[13] IndelGenotyper v2.0. Putative somatic mutations were filtered to remove sequencing and alignment artifacts and manually reviewed in IGV. Review-passed variants were annotated using transcripts from NCBI build 37c and Ensembl release 58. In the event that multiple transcripts could be used to annotate a variant, the transcript with the greatest effect was used. Only "tier 1" coding variants in exons, noncoding RNA genes, or splice sites were reported.

Background Mutation Rate Calculation
Somatic mutations from the MAF file were filtered to remove (1) sites in dbSNP[14] build 132, which likely represent mis-called germline variants; (2) mutations in noncoding RNA genes; and (3) redundant mutations in double-normal samples that were reported for both tumor-normal and tumor-adjacent comparisons. The background mutation rate was computed as the number of non-silent SNVs per covered megabase. A base was considered sufficiently covered if it had at least 8 reads in the tumor sample and at least 6 reads in the normal sample with mapping quality of 20 or higher. By these calculations, 22,689 non-silent SNVs were detected in 16,055 megabases of covered sequence, for a background mutation rate of 1.413 mutations per megabase.

Identification of Significantly Mutated Genes
Significantly mutated genes were identified using the MuSiC package (http://gmt.genome.wustl.edu/genome-music). Briefly, the non-silent mutation rate in each gene is compared to the background mutation rate using three tests: a convolution test (CT), a Fisher's combined P-value test (FCPT), and a likelihood ratio test (LRT). Fisher's combined p-value test (FCPT): FCPT combines P-values from different mutation types into a statistic, χ_c , according to Fisher's method:

$$\chi_c = -2\sum_{i=1}^{k} \log(p_i)$$

where p_i is the p-value obtained via binomial distribution for the i-th mutation type, and k the number of mutation types for a gene. The final p-value for the entire gene is calculated as the probability of observing a value no less than χ_c, based on a χ^2 distribution with $2k$ degrees of freedom.

The LRT constructs a likelihood ratio-based statistic (χ_l) for a gene,

$$\chi_l = 2\sum_{i=1}^{k} \log\left(\frac{L(M_i, C_i \mid r_i)}{L(M_i, C_i \mid R_i)}\right)$$

where M_i, C_i, R_i and r_i are mutation number, coverage, BMR, and maximum likelihood estimate (MLE) of MR, respectively, for the i-th mutation type of a gene, k is the number of mutation types, and $L()$ is the likelihood of observed mutation number for the i-th mutation types, defined as the point probability of observing M_i mutations given a coverage of C_i and a MR of R_i or r_i. The final P-value for the entire gene is calculated as the probability of observing a value no less than χ_l, based on an approximate χ^2 distribution with k degrees of freedom.
The CT calculates a summarized log statistic of joint binomial point probability

$$S_g = -\sum_{i=1}^{k} \log(L(M_i, C_i \mid R_i))$$

where M_i, C_i, R_i, k and $L()$ are referred to as the same as in the LRT method.

To include a gene on the SMG list, a maximum FDR of 5% by both CT and LRT was required. In supplemental table 2, FDR values for CT and LRT are provided if this criterion was met for a given gene in a given subgroup.

IV. mRNA Gene Expression Profiling

Microarray processing
Agilent custom 244K whole genome microarrays were hybridized and processed as previously described[8]. Raw data (level 1), probe-level data (level 2), and gene-level data (level 3) were deposited at the DCC.

Identification of the intrinsic gene expression-based subtypes. Agilent microarray data for 522 tumors, 3 metastatic tumors, and 22 tumor-adjacent normal were combined and gene-median centered. The matrix was hierarchically clustered with an intrinsic subtype list compiled from four previous breast microarray studies[15-18]. Using this cluster, we analyzed the sample relationships by SigClust[19] and identified 13 classes. Samples were also subtyped by the 50-gene PAM50 predictor[20]. High concordance was seen between the SigClust and PAM50 subtypes calls. For simplicity, the PAM50 subtype calls were used for all analyses.

V. miRNA Expression Profiling

Library construction and sequencing
Two micrograms of total RNA per sample are arrayed into 96-well plates, with controls as described below. RNA entering library construction is required to have at least a minimum quality on the BCR submission documentation. Total RNA is mixed with oligo(dT) MicroBeads and loaded into a 96-well MACS column, which is then placed on a MultiMACS separator (Miltenyi Biotec, Germany); the separator's strong magnetic field allows beads to be captured during washes. From the flow-through, small RNAs, including miRNAs, are recovered by

ethanol precipitation. Flow-through RNA quality is checked for a subset of 12 samples using an Agilent Bioanalyzer RNA Nano chip.

miRNA-Seq libraries are constructed using an plate-based protocol developed at the British Columbia Genome Sciences Centre (BCGSC). Negative controls are added at three stages: elution buffer is added to one well when the total RNA is loaded onto the plate, water to another well just before ligating the 3' adapter, and PCR brew mix to a final well just before PCR. A 3' adapter is ligated using a truncated T4 RNA ligase2 (NEB Canada, cat. M0242L) with an incubation of 1 hour at 22^{o}C. This adapter is adenylated, single-strand DNA with the sequence 5' /5rApp/ ATCTCGTATGCCGTCTTCTGCTTGT /3ddC/, which selectively ligates miRNAs. An RNA 5' adapter is then added, using a T4 RNA ligase (Ambion USA, cat. AM2141) and ATP, and is incubated at 37^{o}C for 1 hour. The sequence of the single strand RNA adapter is 5'GUUCAGAGUUCUACAGUCCGACGAUCUGGUCAA3'.

When ligation is completed, 1st strand cDNA is synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat.18064 014) and RT primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'). This is the template for the final library PCR, into which we introduce index sequences to enable libraries to be identified from a sequenced pool that contains multiple libraries. Briefly, a PCR brew mix is made with the 3' PCR primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'), Phusion Hot Start High Fidelity DNA polymerase (NEB Canada, cat. F-540L), buffer, dNTPs and DMSO. The mix is distributed evenly into a new 96-well plate. A Biomek FX (Beckman Coulter, USA) is used to transfer the PCR template (1st strand cDNA) and indexed 5' PCR primers into the brew mix plate. Each indexed 5' PCR primer, 5'-AATGATACGGCGACCACCGACAGNNNNNNGTTCAGAGTTCTACAGTCCGA-3', contains a unique six-nucleotide 'index' (shown here as N's), and is added to each well of the 96-well PCR brew plate. PCR is run at 98°C for 30 sec, followed by 15 cycles of 98°C for 15 sec, 62°C for 30 sec and 72°C for 15 sec, and finally a 5 min incubation at 72^{o}C. Quality is then checked across the whole plate using a Caliper LabChipGX DNA chip. PCR products are pooled, then are size selected to remove larger cDNA fragments and smaller adapter contaminants, using a 96-channel automated size selection robot that was developed at the BCGSC. After size selection, each pool is ethanol precipitated, quality checked using an Agilent Bioanalyzer DNA1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. Q32854). Each pool is then diluted to a target concentration for cluster generation and loaded into a single lane of an Illumina GAIIx or HiSeq 2000 flow cell. Clusters are generated, and lanes are sequenced with a 31-bp main read for the insert and a 7-bp read for the index.

Preprocessing, alignment and annotation

Briefly, the sequence data are separated into individual samples based on the index read sequences, and the reads undergo an initial QC assessment. Adapter sequence is then trimmed off, and the trimmed reads for each sample are aligned to the NCBI GRCh37-lite reference genome. Below we describe these steps in more detail.

Routine QC assesses a subset of raw sequences from each pooled lane for the abundance of reads from each indexed sample in the pool, the proportion of reads that possibly originate from adapter dimers (i.e. a 5' adapter joined to a 3' adapter with no intervening biological sequence) and for the proportion of reads that map to human miRNAs. Sequencing error is estimated by a method originally developed for SAGE[21].

Libraries that pass this QC stage are preprocessed for alignment. While the size-selected miRNAs vary somewhat in length, typically they are ~21 bp long, and so are shorter than the 31-bp read length. Given this, each read sequence extends some distance into the 3' sequencing adapter. Because this non-biological sequence can interfere with aligning the read to the reference genome, 3' adapter sequence is identified and removed (trimmed) from a read. The adapter-trimming algorithm identifies as long an adapter sequence as possible, allowing a number of mismatches that depends on the adapter length found. A typical sequencing run yields several million reads; using only the first (5') 15 bases of the 3' adapter in trimming makes processing efficient, while minimizing the chance that an miRNA read will match the adapter sequence.

The algorithm first determines whether a read sequence should be discarded as an adapter dimer by checking whether the 3' adapter sequence occurs at the start of the read. For reads passing this stage, the algorithm then tries to identify an exact 15-bp match anywhere within the read sequence. If it cannot, it then retries, starting from the 3' end, and allowing up to 2 mismatches. If the full 15bp is not found, decreasing lengths of adapter are checked, down to the first 8 bases, allowing one mismatch. If a match is still not found, from 7 bases down to 1 base is checked, with an exact match required. Finally, the algorithm will trim 1 base off the 3' end of a read if it happens to match the first base of the adapter. This is based on two considerations. First, it is preferable to get a perfect alignment than an alignment that has a potential one-base mismatch. Second, if only 1 base of adapter was found in the read sequence, the read is likely too long to be from a miRNA and the effect of the trimming on its alignment would not affect this sample's overall miRNA profiling result.

After each read has been processed, a summary report is generated containing the number of reads at each read length. Because the shortest mature miRNA in miRBase v16 is 15 bp, any trimmed read that is shorter than 15bp is discarded; remaining reads are submitted for alignment to the reference genome. BWA[9] alignment(s) for each read are checked with a series of three filters. A read with more than 3 alignments is discarded as too ambiguous. For TCGA quantification reports, only perfect alignments with no mismatches are used. Based on comparing expression profiles of test libraries (data not shown), reads that fail the Illumina basecalling chastity filter are retained, while reads that have soft-clipped CIGAR strings are discarded.

For reads retained after filtering, each coordinate for each read alignment is annotated using the reference databases (Table V.1), and requiring a minimum 3-bp overlap between the alignment and an annotation. In annotating reads we address two potential issues. First, a single read alignment can overlap feature annotations of different types; second, a read can have up to three alignment locations, and each alignment location can overlap a different type of feature annotation. By considering heuristically determined priorities (Table V.1), we resolve the first issue by giving each alignment a single annotation. We resolve the second by collapsing multiple annotations to a single annotation, as follows.

If a read has more than one alignment location, and the annotations for these are different, we use the priorities from Table V.1 to assign a single annotation to the read, as long as only one alignment is to a miRNA. When there are multiple alignments to different miRNAs, the read is flagged as cross-mapped[22], and all of its miRNA annotations are preserved, while all of its non-miRNA annotations are discarded. This ensures that all annotation information about ambiguously mapped miRNAs is retained, and allows annotation ambiguity to be addressed in downstream analyses. Note that we consider miRNAs to be cross-mapped only if they map to

different miRNAs, not to functionally identical miRNAs that are expressed from different locations in the genome. Such cases are indicated by miRNA miRBase names, which can have up to 4 separate sections separated by "-", e.g. hsa-mir-26a-1. A difference in the final (e.g. '-1') section denotes functionally equivalent miRNAs expressed from different regions of the genome, and we consider only the first 3 sections (e.g. 'hsa-mir-26a') when comparing names. As long as a read maps to multiple miRNAs for which the first 3 sections of the name are identical (e.g. hsa-mir-26a-1 and hsa-mir-26a-2), it is treated as if it maps to only one miRNA, and is not flagged as cross-mapped.

From the profiling results for a tumor type, for a minimum of approximately 100 samples, we identify the depth of sequencing required to detect the miRNAs that are expressed in a sample by considering a graph of the number of miRNAs detected in a sample as a function of the number of reads aligned to miRNAs. For the current work, a library from a sequenced pool was required to have at least 750,000 reads mapped to miRBase annotations. For any sequencing run that fails to meet this threshold, we sequence the sample again to achieve at least the minimum number of miRNA-aligned reads.

Finally, for each sample, the reads that correspond to particular miRNAs are summed and normalized to a million miRNA-aligned reads to generate the quantification files that are submitted to the DCC. Quantification files include information on variable 5' and 3' read alignment locations, which can reflect isoforms, adapter trimming and RNA degradation.

Unsupervised consensus clustering

Normalized read count data for 697 tumor samples were extracted from Level 3 data archives on the TCGA Data Portal website (http://tcga-data.nci.nih.gov/tcga/). The set of isoform.quantification.txt files, which give read counts at base pair resolution, was processed to sum up read counts at mature and star strand resolution (corresponding to miRBase v16 MIMAT accessions). Read counts for each sample were normalized to RPM, i.e. to reads per million reads aligned to miRBase mature or star strands. Mature and star strands were ranked by RPM variance across the samples, and the most variant 25% (306 MIMATs) were used as input to unsupervised consensus clustering with the NMF v0.5.02 package[23] in R v2.12.0, using the default Brunet algorithm and 50 and 200 iterations respectively for the rank survey and the main run. A seven-cluster result was selected by considering profiles of cophenetic score and average silhouette width for clustering solutions having between 3 and 15 clusters; comparing silhouette plots for favorable solutions; comparing core/non-core cluster members with clinical covariate tracks (below); and preferring a result set with fewer, larger clusters. Silhouette results were generated from the consensus membership matrix using the 'cluster' v1.14.1 R package. Silhouette width profiles were generated for samples ordered as in the NMF heatmap, and atypical, or 'non-core' members in each cluster were identified using a silhouette width threshold set to a fraction (e.g. 0.90) of the maximum width in each cluster. Asymptotic association p-values for covariate contingency tables were calculated using R's chi-square test.

Table V.1. Priorities for resolving annotation ambiguities for aligned reads.

Priority	Annotation type	Database
1	mature strand	miRBase v16
2	star strand	
3	precursor miRNA	
4	stemloop, from 1 to 6 bases outside the mature strand, between the	

5	mature and star strands "unannotated", any region other than the mature strand in miRNAs where no star strand is annotated	
6 7 8 9 10 11 12	snoRNA tRNA rRNA snRNA scRNA srpRNA Other RNA repeats	UCSC small RNAs, RepeatMasker
13 14 15 16 17	coding exons with zero annotated CDS region length 3' UTR 5' UTR coding exon intron	UCSC knownGenes
18 19 20 21 22 23 24 25 26	LINE SINE LTR Satellite RepeatMasker DNA RepeatMasker Low complexity RepeatMasker Simple Repeat RepeatMasker Other RepeatMasker Unknown	UCSC RepeatMasker

VI. DNA Methylation Profiling

Array-based DNA methylation assay: We used the Illumina Infinium DNA methylation platforms, HumanMethylation27 (HM27) BeadChip and HumanMethylation450 (HM450) BeadChip (Illumina, San Diego, CA) to obtain gene promoter and gene body DNA methylation profiles of 802 TCGA breast cancer samples and 122 adjacent non-tumor breast tissue samples. The Infinium HM27 array targets 27,578 CpG sites located in proximity to the transcription start sites of 14,475 consensus coding sequencing (CCDS) in the NCBI Database (Genome Build 36). The Infinium HM450 array targets 482,421 CpG sites and covers 99% of RefSeq genes, with an average of 17 CpG sites per gene region distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR. It covers 96% of CpG islands, with additional coverage in island shores and the regions flanking them. The assay probe sequences and information on each interrogated CpG site on both Infinium DNA methylation platforms can be found in the MAGE-TAB ADF (Array Design Format) file deposited on the TCGA Data Portal.

We performed bisulfite conversion on 1 µg of genomic DNA from each sample using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA) according to the manufacturer's instructions. We assessed the amount of bisulfite converted DNA and completeness of bisulfite conversion using a panel of MethyLight-based quality control (QC) reactions as previously described[24]. All the TCGA samples passed our QC tests and entered the Infinium DNA methylation assay pipeline.

Bisulfite-converted DNA was whole genome amplified (WGA) and enzymatically fragmented prior to hybridization to BeadChip arrays. With respect to the Illumina Infinium HM27 platfom, the amplified and fragmented DNA molecules anneal to a locus-specific DNA oligomers (50 mers) covalently attached to a specific bead type. The HM27 platform utilizes the Infinium I chemistry. Each interrogated CpG locus can hybridize to methylated (CpG) or unmethylated (TpG) oligo bead types. DNA methylation-specific primer annealing is followed by single-base extension using labeled nucleotides [cy5 (red) or cy3 (green)]. Both methylated (M) and unmethylated (U) bead types for a specific CpG locus incorporate the same labeled nucleotide, as determined by the base immediately preceding the cytosine being interrogated by the assay, and are subsequently detected in a single color channel. Fluorescence intensities of the M and U bead types for each CpG locus were measured using the Illumina BeadArray Reader. The mean signal intensities for replicate M and U probes for each CpG locus were extracted from Illumina GenomeStudio software.

With respect to the Illumina Infinium HM450 platform, the oligomer probe designs follow the Infinium I and II chemistries, in which locus-specific base extension follows hybridization to a methylation-specific oligomer. The Infinium I probes are used in the HM27 platform and terminate complementary to the interrogated CpG site for methylated loci, or complementary to the TpG for unmethylated alleles. The Infinium type II probes terminate immediately 3' to the interrogated CpG (or TpG) site. A matched oligomer-template DNA molecule hybrid will allow for the incorporation of a cy3- or cy5-labeled nucleotide immediately adjacent to the interrogated CpG (or TpG) site for Infinium I probes, or at the targeted CpG for the Infinium II probes. However, if the probe and template are mismatched, then primer extension will not occur. After labeled nucleotide incorporation, the intensities of both cy3 and cy5 are obtained after scanning each hybridized array. BeadArrays are scanned using the Illumina iScan technology, and the raw data are extracted using the R-based *methylumi* package[25] to calculate the beta value DNA methylation score for each probe and sample.

The level of DNA methylation at each CpG locus is scored as beta (β) value calculated as (M/(M+U)), ranging from 0 to 1, with values close to 0 indicating low levels of DNA methylation and beta values close to 1 indicating high levels of DNA methylation.

The detection P values provide an indication of the quality of DNA methylation measurement and are calculated as previously described. We determined that data points with a detection P value >0.05 are not significantly different from background measurements, and therefore were masked as "NA" in the Level 2 and 3 in HM27 and Level 3 in HM450 data packages, as detailed below.

TCGA data packages: The data levels and the files contained in each data level package are described below and are present on the TCGA Data Portal website (http://tcga-data.nci.nih.gov/tcga/). Please note that with continuing updates of genomic databases, data archive revisions become available at the TCGA Data Portal.

HM27: *Level 1:* Level 1 data contain the non-background corrected signal intensities of the M and U probes and the mean negative control cy5 (red) and cy3 (green) signal intensities. A detection P value for each data point, the number of replicate beads for M and U probes as well as the standard error of M, U, and control probe signal intensities are also provided. It is important to note that for some CpG targets, both M and U measurements will be cy3, and for others both will be cy5. To resolve ambiguities regarding this subtlety of the Infinium DNA Methylation assay, we have labeled the cy3 and cy5 values deposited to the DCC as "Methylated Signal Intensity" and "Unmethylated Signal Intensity". The information of which dye is used for

each CpG locus is supplied in the MAGE-TAB ADF file deposited in the DCC. *Level 2:* Level 2 data files contain the β-value calculations for each probe and sample. Data points with detection P values >0.05 were not considered to be significantly different from background, and were masked as "NA". *Level 3:* Level 3 data contain β-value calculations, HUGO gene symbol, chromosome number and genomic coordinate for each targeted CpG site on the array. In addition, we masked data points with "NA" from the probes that 1) contain known single nucleotide polymorphisms (SNPs) after comparison to the dbSNP database (Build 130), 2) contain repetitive sequence elements that cover the targeted CpG locus in each 50 bp probe sequence, 3) are not uniquely aligned to the human genome (NCBI build 36.1) at 20 nucleotides at the 3' terminus of the probe sequence, 4) span known regions of small insertions and deletions (indels) in the human genome (dbSNP build 130).

HM450: *Level 1:* Level 1 data contain raw IDAT files. IDAT files are the direct output from the scanning software - Illumina iScan Control. *Level 2:* Level 2 data contain background corrected signal intensities of the M and U probes. *Level 3:* Level 3 data files contain β-value calculations and masked data points with "NA" from the probes that are annotated as having a SNP within 10 base pairs of the interrogated locus (HM27 carryover or recently discovered, dbSNP build 131). The genomic characteristics for each probe are available for download via Illumina (www.illumina.com).

The following data archives were used for the analyses described in this manuscript.

S. No	Archive Name	Platform
1	jhu-usc.edu_BRCA.HumanMethylation27.Level_3.1.0.0.tar.gz	HM27
2	jhu-usc.edu_BRCA.HumanMethylation27.Level_3.2.0.0.tar.gz	HM27
3	jhu-usc.edu_BRCA.HumanMethylation27.Level_3.3.0.0.tar.gz	HM27
4	jhu-usc.edu_BRCA.HumanMethylation27.Level_3.4.0.0.tar.gz	HM27
5	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.1.1.0.tar.gz	HM450
6	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.2.1.0.tar.gz	HM450
7	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.3.1.0.tar.gz	HM450
8	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.4.1.0.tar.gz	HM450
9	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.5.1.0.tar.gz	HM450
10	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.6.1.0.tar.gz	HM450
11	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.7.1.0.tar.gz	HM450
12	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.8.1.0.tar.gz	HM450
13	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.9.1.0.tar.gz	HM450
14	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.10.1.0.tar.gz	HM450
15	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.11.1.0.tar.gz	HM450
16	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.12.1.0.tar.gz	HM450
17	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.13.1.0.tar.gz	HM450
18	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.14.1.0.tar.gz	HM450
19	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.15.1.0.tar.gz	HM450
20	jhu-usc.edu_BRCA.HumanMethylation450.Level_3.16.1.0.tar.gz	HM450

Unsupervised clustering analysis of DNA methylation data

Statistical analysis and data visualization were carried out using the R/Biocoductor software packages (http://www.bioconductor.org). The following filtering steps were adopted for selection of probes for unsupervised clustering analysis. We first removed probes that are designed for the

sequences on X and Y chromosomes. As we observed batch and platform specific methylation patterns, we filtered the probes based on the following criteria. 1. We selected probes that were represented on both HM27 and HM450 platforms (N=25,014). 2. We removed probes that contain any "NA"-masked data points (probes selected, N= 20,847). 3. In order to filter-out probes with high batch effects, we applied ANOVA (logit (Beta)~Batch) and removed probes with above median F statistics. In this step, we tried to retain probes with high biological effect by simultaneously selecting for probes with high within batch standard deviation (probes selected, N= 10,422). 4. Further, we filtered-out probes with platform specific methylation patterns by applying t-tests (logit(Beta)~platform) and removed 90% of probes with highest t-statistics. As in the previous step, we simultaneously selected probes with high within platform standard deviation (probes selected, N= 1,042). Finally, we selected a union of the probes (batch and platform effect filtered) with an above median standard deviation of betas calculated separately for each platform (N=574).

We used recursively partitioned mixture model (RPMM) for the identification of breast cancer subgroups based on the Illumina Infinium DNA methylation datasets. RPMM is a model-based unsupervised clustering approach well-suited for beta-distributed DNA methylation measurements which lie between 0 and 1, and implemented as the RPMM R/Bioconductor package[26].

We performed RPMM clustering on 802 breast cancer samples and the above-mentioned 574 probes. A fanny algorithm (a fuzzy clustering algorithm) was used for initialization and level-weighted version of Bayesian information criterion (BIC) as a split criterion for an existing cluster as implemented in the RPMM package. The DNA methylation β-values for 466 data-freeze breast cancer samples and 122 adjacent normals were represented graphically using a heatmap, generated by the R package *Heatplus*. Ordering of the samples within a RPMM class in the heatmaps was obtained by using the R package *seriation*. A non-parametric Pearson $\chi 2$ test with Yates continuity correction was used to assess the significance of association of various categorical covariates with DNA methylation clusters. Fisher's exact test was performed on covariates in which samples were less than five in any cell in the contingency table.

VII. SNP Based Copy Number Analysis

DNA from each tumor or germline-derived sample was hybridized to the Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute[27]. From raw .CEL files, Birdseed was used to infer a preliminary copy-number at each probe locus[28]. For each tumor, genome-wide copy number estimates were refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor[8] (and Tabak B and Beroukhim R. Manuscript in preparation). This linear combination of normal samples tends to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. Individual copy-number estimates then undergo segmentation using Circular Binary Segmentation[29]. As part of this process of copy-number assessment and segmentation, regions corresponding to germline copy-number alterations were removed by applying filters generated from either the TCGA germline samples from the ovarian cancer analysis or from samples from this collection.

Segmented copy number profiles for tumor and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude

to the set of inferred copy number changes underlying each segmented copy number profile[29]. Analysis of broad copy number alterations was then conducted as previously described[28]. Significant focal copy number alterations were identified from segmented data using GISTIC 2.0[30]. NMF consensus clustering of copy number data was performed using the presence or absence of amplifications or deletions in regions identified by GISTIC 2.0 analysis.

For correlations between copy number alterations in breast and ovarian tumors, thresholded gene level copy number values were calculated using GISTIC 2.0[31]. These values were calculated using the maximum copy number change in each gene or miRNA. Log2 copy number values were thresholded as followed: values < -1 are set to -2, values between -1 and -0.3 are set to -1, values between -0.3 and 0.3 are set to 0, values between -1 and -0.3 are set to 1, and values > 1 are set to 2. Tumors where all gene level thresholded copy number changes equaled 0 were excluded from the analysis. Pearson R correlations were calculated for all possible pairs of ovarian and breast tumors.

Altered Genome Fraction

Basal-like tumors are characterized by high genomic instability reflected by extensive genomic re-arrangements. Copy number DNA alterations are more frequent and typically smaller than in any other subtypes. Similar pattern of alterations have been observed before in serous ovarian carcinoma, and this again remarks the similarity between these two cancer types.

To quantify this similarity in terms of number of copy number alterations, we compute the fraction of the genome that is altered for each sample, and compare the distribution of these values across breast cancer subtypes and ovarian cancer.

Given the segmented data for each sample, such that probes within the same segment have the same copy number level (log2-ratios), we determined the percentage of mega-bases (MB) that have copy number level above a given threshold **T** (gain) or below a given threshold **–T** (loss). Genomic regions are thus defined *altered* if the corresponding level is above **T** or below **-T**, by contrast regions with level in the interval **[-T,T]** are considered diploid. For this analysis we select **T** = 0.15. Altered genome fraction distributions show remarkably similarities between Basal-like and serous ovarian tumors with both tumor types having over 70% of the samples with more than 40% of the genome in a non-diploid status (Fig. VII.1).

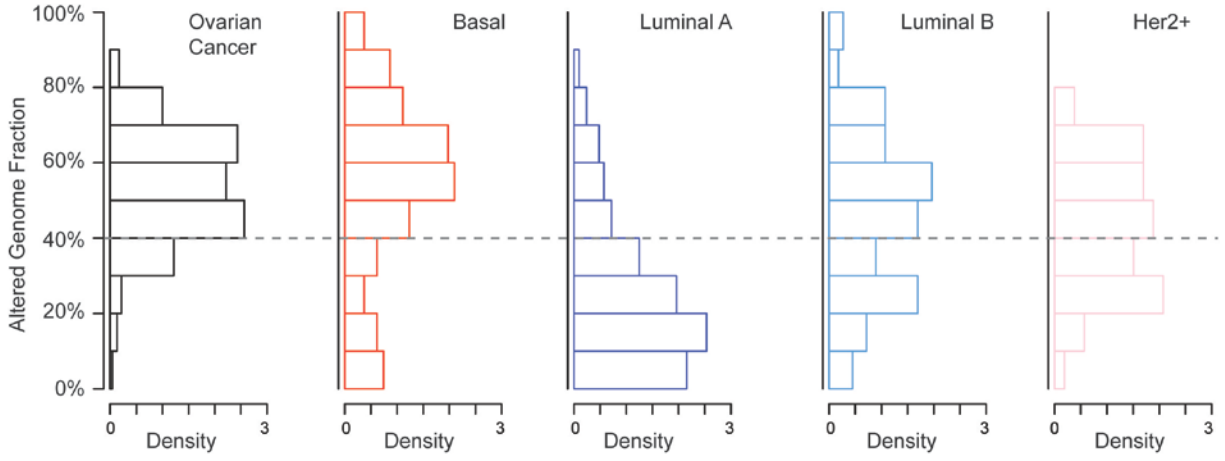


Fig. VII.1: Altered Genome Fraction distribution across breast cancer subtypes and ovarian carcinoma.

Curtis et al Copy Number Classifications

For assigning Curtis et al. subtype (1 through 10) to each TCGA sample, we first obtained the set of 754 features (39 gene copy, 715 gene mRNA), originally used in the Curtis et al. study to classify the validation samples (Table S42[32]). Each feature in the Curtis classifier had previously been assigned a scaled weighting for each of the ten subtypes. After first standardizing our TCGA datasets (mRNA features standardized to standard deviations from the median across the cancer sample profiles; gene copy features standardized to log2 fold relative to normal), we computed the Pearson's correlation between each TCGA sample profile (copy+mRNA) and each of the ten Curtis subtype profiles. For a given TCGA sample, the Curtis subtype having the highest correlation was assigned to that sample.

VIII. Reverse Phase Protein Array (RPPA)

Protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 nmol/L Hepes (pH 7.4), 150 nmol/L NaCl, 1.5 nmol/L MgCl2, 1 mmol/L EGTA, 100 nmol/L NaF, 10 nmol/L NaPPi, 10% glycerol, 1 nmol/L phenylmethylsulfonyl fluoride, 1 nmol/L Na3VO4, and aprotinin 10 Ag/mL) from human tumors and RPPA was performed as described previously[33-37]. Lysis buffer was used to lyse frozen tumors by Precellys homogenization. Tumor lysates were adjusted to 1 µg/µL concentration as assessed by bicinchoninic acid assay (BCA) and boiled with 1% SDS. Tumor lysates were manually diluted in five-fold serial dilutions with lysis buffer. An Aushon Biosystems 2470 arrayer (Burlington, MA) printed 1,056 samples on nitrocellulose-coated slides (Grace Bio-Labs). Slides were probed with 171 validated primary antibodies (Table VIII.1) followed by corresponding secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG or Rabbit anti-Goat IgG). Signal was captured using a DakoCytomation-catalyzed system and DAB colorimetric reaction. Slides were scanned in CanoScan 9000F. Spot intensities were analyzed and quantified using Microvigene software (VigeneTech Inc., Carlisle, MA), to generate spot signal intensities (Level 1 data). The software SuperCurveGUI[35,37], available at http://bioinformatics.mdanderson.org/Software/supercurve/, was used to estimate the EC50 values of the proteins in each dilution series (in log2 scale). Briefly, a fitted curve ("supercurve") was plotted with the signal intensities on the Y-axis and the relative log2 concentration of each protein on the X-axis using the non-parametric, monotone increasing B-spline model[33]. During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric[37] was returned for each slide to help determine the quality of the slide: if the score is less than 0.8 on a 0-1 scale, the slide was dropped. In most cases, the staining was repeated to obtain a high quality score. If more than one slide was stained for an antibody, the slide with the highest QC score was used for analysis (Level 2 data). Protein measurements were corrected for loading as described[35,37,38] using median centering across antibodies (level 3 data). In total, 171 antibodies and 410 samples were used. For the selection of the 171 antibody set, we focused on markers currently used for breast cancer classification due to their value in treatment decisions (ER, PR and HER2), markers implicated in breast cancer pathophysiology and markers implicated in the pathophysiology of other cancer lineages. Final selection of antibodies was also driven by the availability of high quality antibodies that consistently pass a strict validation process as previously described[39]. Antibodies are labeled as validated and use with caution based on degree of validation by criteria previously described[39]. Seven samples were redacted during sample collection and were excluded from further analysis.

Raw data (level 1), SuperCurve nonparameteric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the DCC.

RPPA Subtype discovery

We used non-negative Matrix Factorization (NMF) to identify robust subtypes of breast cancer based on protein expression[31]. Through visual inspection of consensus and correlation matrices and based on the cophenetic coefficient plot, we identified clustering with seven clusters as the optimal solution (Figure VIII.1). The 403 samples in the RPPA data set were classified into seven groups after performing NMF using 100 iterations and using the divergence error function. With the sample order returned by NMF, we plotted a semi-supervised hierarchical clustering using the same data sets, with the antibodies clustered using a Pearson correlation coefficient based distance matrix (distance=(1-R)/2 where R is the Pearson correlation coefficient between two rows in the data) and Ward's minimum variance based agglomeration algorithm. The PAM50 gene based subtypes, the clinical status of ER, PR, HER2 (based on IHC), tumor size, node status and the mutation status of five selected genes with abundant mutations, were also plotted with the samples ordered by NMF groups. Chi-square test p-values were used to compare frequencies of the mutations or other events across among the 7 groups.

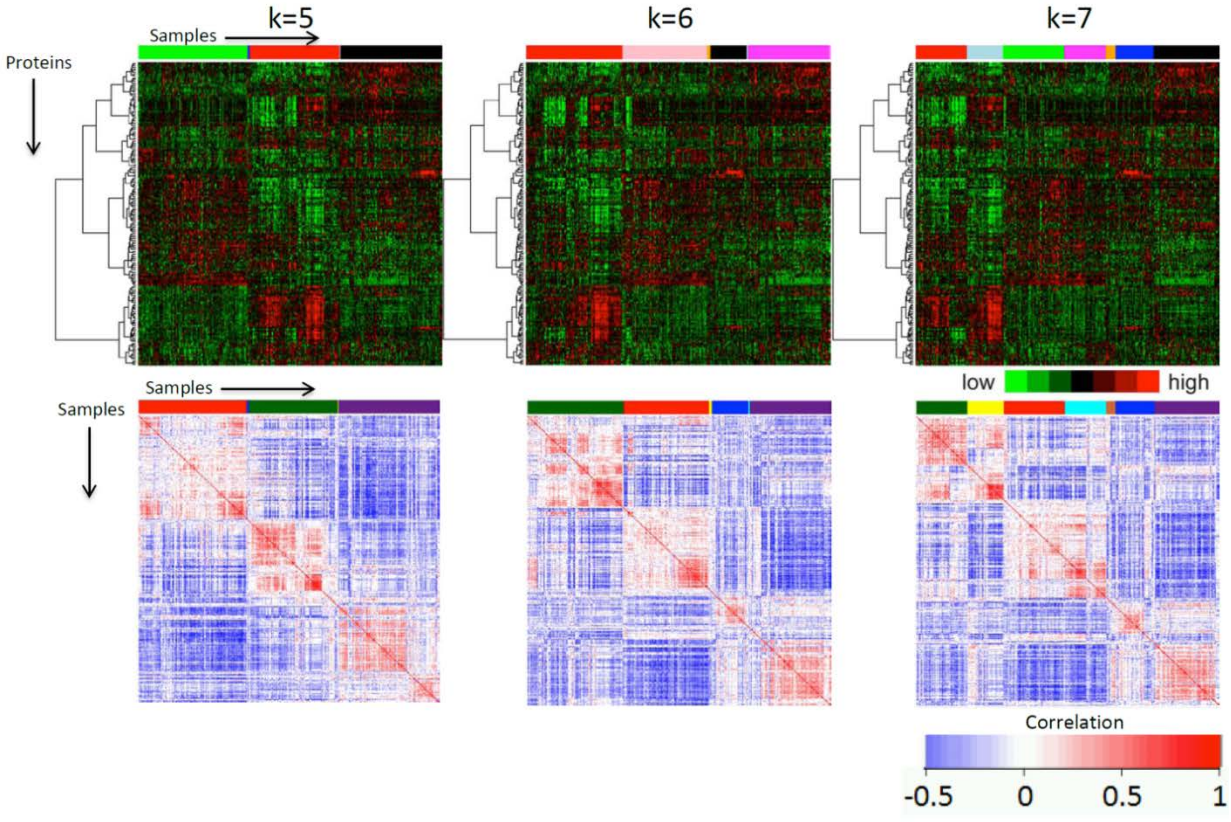


Figure VIII.1. Expression and correlation heatmaps of RPPA clustering data. Protein expression data for 403 samples and 171 proteins was clustered using non-negative matrix factorization and different numbers of k.

Table VIII.1. Antibody information for all proteins included on the RPPA platform.

Full Slide Name (Antibody Name + Slide ID)	Protein Name	Gene Name	Antibody validation status	Antibody Origin	Antibody Source (Company)	Catalog Number
14-3-3_epsilon-M-C_GBL9013486	14-3-3_epsilon	YWHAE	Use with Caution	Mouse	Santa Cruz	sc-2395
4E-BP1_pS65-R-V_GBL9013201	4E-BP1_pS65	EIF4EBP1	Validated	Rabbit	CST	9456
4E-BP1_pT37-R-V_GBL9013202	4E-BP1_pT37	EIF4EBP1	Validated	Rabbit	CST	9459
4E-BP1_pT70-R-V_GBL9013203	4E-BP1_pT70	EIF4EBP1	Validated	Rabbit	CST	9455
4E-BP1-R-V_GBL9013200	4E-BP1	EIF4EBP1	Validated	Rabbit	CST	9452
53BP1-R-C_GBL9013204	53BP1	TP53BP1	Use with Caution	Rabbit	CST	4937
A-Raf_pS299-R-NA_GBL9013439	A-Raf_pS299	ARAF	NA	Rabbit	CST	4431
ACC_pS79-R-V_GBL9013205	ACC_pS79	ACACA ACACB	Validated	Rabbit	CST	3661
ACC1-R-C_GBL9013206	ACC1	ACACA	Use with Caution	Rabbit	Epitomics	1768-1
AIB1-M-V_GBL9013344	AIB1	NCOA3	Validated	Mouse	BD Biosciences	611105
Akt_pS473-R-V_GBL9013208	Akt_pS473	AKT1 AKT2 AKT3	Validated	Rabbit	CST	9271
Akt_pT308-R-V_GBL9013209	Akt_pT308	AKT1 AKT2 AKT3	Validated	Rabbit	CST	9275
Akt-R-V_GBL9013465	Akt	AKT1 AKT2 AKT3	Validated	Rabbit	CST	9272
alpha-Catenin-M-V_GBL9013471	alpha-Catenin	CTNNA1	Validated	Mouse	Calbiochem	CA1030
AMPK_alpha-R-C_GBL9013210	AMPK_alpha	PRKAA1	Use with Caution	Rabbit	CST	2532
AMPK_pT172-R-V_GBL9013211	AMPK_pT172	PRKAA1	Validated	Rabbit	CST	2535
ANLN-M-NA_GBL9013453	ANLN	ANLN	Validated	Mouse	Atlas	CAB036211
Annexin_I-R-V_GBL9013410	Annexin_I	ANXA1	Validated	Rabbit	Invitrogen	71-3400
AR-R-V_GBL9013213	AR	AR	Validated	Rabbit	Epitomics	1852-1
ARID1A-M-V_GBL9013497	ARID1A	ARID1A	Validated	Mouse	Abgent	AT1188a
ATM-R-NA_GBL9013214	ATM	ATM	NA	Rabbit	Abcam	ab32420
B-Raf-M-NA_GBL9013421	B-Raf	BRAF	NA	Mouse	Santa Cruz	sc-5284
Bak-R-C_GBL9013429	Bak	BAK1	Use with Caution	Rabbit	Epitomics	1542-1
Bax-R-V_GBL9013216	Bax	BAX	Validated	Rabbit	CST	2772
Bcl-2-M-V_GBL9013345	Bcl-2	BCL2	Validated	Mouse	Dako	Dako M0887
Bcl-X-R-C_GBL9013218	Bcl-X	BCL2L1	Use with Caution	Rabbit	Epitomics	1018-1
Bcl-xL-R-C_GBL9013219	Bcl-xL	BCL2L1	Use with Caution	Rabbit	CST	2762
Beclin-G-V_GBL9013414	Beclin	BECN1	Validated	Goat	Santa Cruz	sc-10086
beta-Catenin-R-V_GBL9013217	beta-Catenin	CTNNB1	Validated	Rabbit	CST	9562
Bid-R-C_GBL9013220	Bid	BID	Use with Caution	Rabbit	Epitomics	1008-1
Bim-R-V_GBL9013221	Bim	BCL2L11	Validated	Rabbit	Epitomics	1036-1

c-Jun_pS73-R-C_GBL9013232	c-Jun_pS73	JUN	Use with Caution	Rabbit	CST	9164
c-Kit-R-V_GBL9013262	c-Kit	KIT	Validated	Rabbit	Epitomics	1522
c-Met_pY1235-R-C_GBL9013234	c-Met_pY1235	MET	Use with Caution	Rabbit	CST	3129
c-Met-M-C_GBL9013381	c-Met	MET	Use with Caution	Mouse	CST	3127
c-Myc-R-C_GBL9013271	c-Myc	MYC	Use with Caution	Rabbit	CST	9402
C-Raf_pS338-R-C_GBL9013300	C-Raf_pS338	RAF1	Use with Caution	Rabbit	CST	9427
C-Raf-R-V_GBL9013299	C-Raf	RAF1	Validated	Rabbit	Millipore	05-739
Caspase-7_cleavedD198-R-C_GBL9013224	Caspase-7_cleavedD198	CASP7	Use with Caution	Rabbit	CST	9491
Caspase-8-M-C_GBL9013487	Caspase-8	CASP8	Use with Caution	Mouse	CST	9746
Caspase-9_cleavedD330-R-C_GBL9013478	Caspase-9_cleavedD330	CASP9	Use with Caution	Rabbit	CST	9501
Caveolin-1-R-V_GBL9013227	Caveolin-1	CAV1	Validated	Rabbit	CST	3238
CD31-M-V_GBL9013423	CD31	PECAM1	Validated	Mouse	Dako	M0823
CD49b-M-V_GBL9013489	CD49b	ITGA2	Validated	Mouse	BD	611016
CDK1-R-V_GBL9013228	CDK1	CDC2	Validated	Rabbit	CST	9112
Chk1_pS345-R-C_GBL9013230	Chk1_pS345	CHEK1	Use with Caution	Rabbit	CST	2348
Chk1-R-V_GBL9013431	Chk1	CHEK1	Validated	Rabbit	CST	2345
Chk2_pT68-R-C_GBL9013231	Chk2_pT68	CHEK2	Use with Caution	Rabbit	CST	2197
Chk2-M-C_GBL9013347	Chk2	CHEK2	Use with Caution	Mouse	CST	3440
cIAP-R-V_GBL9013479	cIAP	BIRC2	Validated	Rabbit	Millipore	07-759
Claudin-7-R-V_GBL9013233	Claudin-7	CLDN7	Validated	Rabbit	Novus	NB100-91714
Collagen_VI-R-V_GBL9013235	Collagen_VI	COL6A1	Validated	Rabbit	Santa Cruz	SC-20649
COX-2-R-C_GBL9013249	COX-2	PTGS2	Use with Caution	Rabbit	Epitomics	2169-1
Cyclin_B1-R-V_GBL9013248	Cyclin_B1	CCNB1	Validated	Rabbit	Epitomics	1495-1
Cyclin_D1-R-V_GBL9013247	Cyclin_D1	CCND1	Validated	Rabbit	Santa Cruz	SC-718
Cyclin_E1-M-V_GBL9013348	Cyclin_E1	CCNE1	Validated	Mouse	Santa Cruz	SC-247
DJ-1-R-C_GBL9013245	DJ-1	PARK7	Use with Caution	Rabbit	Abcam	ab76008
Dvl3-R-V_GBL9013457	Dvl3	DVL3	Validated	Rabbit	CST	3218
E-Cadherin-R-V_GBL9013244	E-Cadherin	CDH1	Validated	Rabbit	CST	4065
eEF2-R-V_GBL9013243	eEF2	EEF2	Validated	Rabbit	CST	2332
eEF2K-R-V_GBL9013242	eEF2K	EEF2K	Validated	Rabbit	CST	3692
EGFR_pY1068-R-V_GBL9013480	EGFR_pY1068	EGFR	Validated	Rabbit	CST	2234
EGFR_pY1173-R-C_GBL9013240	EGFR_pY1173	EGFR	Use with Caution	Rabbit	Epitomics	1124
EGFR_pY992-R-V_GBL9013246	EGFR_pY992	EGFR	Validated	Rabbit	CST	2235
EGFR-R-C_GBL9013241	EGFR	EGFR	Use with Caution	Rabbit	Santa Cruz	SC-03
eIF4E-R-V_GBL9013238	eIF4E	EIF4E	Validated	Rabbit	CST	9742

ER-alpha_pS118-R-V_GBL9013237	ER-alpha_pS118	ESR1	Validated	Rabbit	Epitomics	1091-1
ER-alpha-R-V_GBL9013236	ER-alpha	ESR1	Validated	Rabbit	Lab Vision	RM-9101-S
ERCC1-M-C_GBL9013425	ERCC1	ERCC1	Use with Caution	Mouse	Lab Vision	MS-671-PO
ERK2-R-NA_GBL9013250	ERK2	MAPK1	NA	Rabbit	Santa Cruz	sc-154
FAK-R-C_GBL9013251	FAK	PTK2	Use with Caution	Rabbit	Epitomics	1700-1
Fibronectin-R-C_GBL9013445	Fibronectin	FN1	Use with Caution	Rabbit	Epitomics	1574-1
FOXO3a_pS318_S321-R-C_GBL9013253	FOXO3a_pS318_S321	FOXO3	Use with Caution	Rabbit	CST	9465
FOXO3a-R-C_GBL9013252	FOXO3a	FOXO3	Use with Caution	Rabbit	CST	9467
GAB2-R-V_GBL9013459	GAB2	GAB2	Validated	Rabbit	CST	3239
GATA3-M-V_GBL9013350	GATA3	GATA3	Validated	Mouse	BD Biosciences	558686
GSK3-alpha-beta_pS21_S9-R-V_GBL9013254	GSK3-alpha-beta_pS21_S9	GSK3A GSK3B	Validated	Rabbit	CST	9331
GSK3-alpha-beta-M-V_GBL9013417	GSK3-alpha-beta	GSK3A GSK3B	Validated	Mouse	Santa Cruz	SC-7291
HER2_pY1248-R-V_GBL9013467	HER2_pY1248	ERBB2	Validated	Rabbit	Upstate (Millipore)	06-229
HER2-M-V_GBL9013468	HER2	ERBB2	Validated	Mouse	Lab Vision	MS-325-P1
HER3_pY1289-R-V_GBL9013462	HER3_pY1289	ERBB3	Validated	Rabbit	CST	4791
HER3-M-C_GBL9013491	HER3	ERBB3	Use with Caution	Mouse	Lab Vision	MS-201-P1ABX
HSP70-R-C_GBL9013257	HSP70	HSPA1A	Use with Caution	Rabbit	CST	4872
IGF-1R-beta-R-C_GBL9013259	IGF-1R-beta	IGF1R	Use with Caution	Rabbit	CST	3027
IGFBP2-R-V_GBL9013258	IGFBP2	IGFBP2	Validated	Rabbit	CST	3922
INPP4B-G-C_GBL9013415	INPP4B	INPP4B	Use with Caution	Goat	Santa Cruz	SC-12318
IRS1-R-V_GBL9013260	IRS1	IRS1	Validated	Rabbit	Upstate (Millipore)	06-248
JNK_pT183_pT185-R-C_GBL9013501	JNK_pT183_pT185	MAPK8	Validated	Rabbit	CST	4668
JNK2-R-C_GBL9013261	JNK2	MAPK9	Use with Caution	Rabbit	CST	4672
K-Ras-M-C_GBL9013379	K-Ras	KRAS	Use with Caution	Mouse	Santa Cruz	sc-30 (F234)
Ku80-R-C_GBL9013263	Ku80	XRCC5	Use with Caution	Rabbit	CST	2180
LBK1-M-NA_GBL9013424	LBK1	STK11	NA	Mouse	Abcam	ab15095
Lck-R-V_GBL9013483	Lck	LCK	Validated	Rabbit	CST	2752
MAPK_pT202_Y204-R-V_GBL9013265	MAPK_pT202_Y204	MAPK1 MAPK3	Validated	Rabbit	CST	4377
MEK1_pS217_S221-R-V_GBL9013267	MEK1_pS217_S221	MAP2K1	Validated	Rabbit	CST	9154
MEK1-R-V_GBL9013266	MEK1	MAP2K1	Validated	Rabbit	Epitomics	1235-1
MIG-6-M-V_GBL9013383	MIG-6	ERRFI1	Validated	Mouse	Sigma	WH0054206M1
Mre11-R-C_GBL9013268	Mre11	MRE11A	Use with Caution	Rabbit	CST	4847
MSH2-M-C_GBL9013419	MSH2	MSH2	Use with	Mouse	CST	2850

			Caution			
MSH6-R-C_GBL9013269	MSH6	MSH6	Use with Caution	Rabbit	SDI	2203.00.02
mTOR_pS2448-R-C_GBL9013504	mTOR_pS2448	FRAP1	Validated	Rabbit	CST	2971
mTOR-R-V_GBL9013270	mTOR	FRAP1	Validated	Rabbit	CST	2983
N-Cadherin-R-V_GBL9013432	N-Cadherin	CDH2	Validated	Rabbit	CST	4061
NF-kB-p65_pS536-R-C_GBL9013273	NF-kB-p65_pS536	NFKB1	Use with Caution	Rabbit	CST	3033
NF2-R-C_GBL9013272	NF2	NF2	Use with Caution	Rabbit	SDI	2271.00.02
Notch1-R-V_GBL9013274	Notch1	NOTCH1	Validated	Rabbit	CST	3268
Notch3-R-C_GBL9013275	Notch3	NOTCH3	Use with Caution	Rabbit	Santa Cruz	sc-5593
P-Cadherin-R-C_GBL9013223	P-Cadherin	CDH3	Use with Caution	Rabbit	CST	2130
p21-R-C_GBL9013276	p21	CDKN1A	Use with Caution	Rabbit	Santa Cruz	SC-397
p27_pT157-R-C_GBL9013280	p27_pT157	CDKN1B	Use with Caution	Rabbit	R&D	AF1555
p27_pT198-R-V_GBL9013278	p27_pT198	CDKN1B	Validated	Rabbit	Abcam	ab64949
p27-R-V_GBL9013279	p27	CDKN1B	Validated	Rabbit	Epitomics	1591-1
p38_MAPK-R-C_GBL9013281	p38_MAPK	MAPK14	Use with Caution	Rabbit	CST	9212
p38_pT180_Y182-R-V_GBL9013282	p38_pT180_Y182	MAPK14	Validated	Rabbit	CST	9211
p53-R-V_GBL9013437	p53	TP53	Validated	Rabbit	CST	9282
p70S6K_pT389-R-V_GBL9013285	p70S6K_pT389	RPS6KB1	Validated	Rabbit	CST	9205
p70S6K-R-V_GBL9013284	p70S6K	RPS6KB1	Validated	Rabbit	Epitomics	1494-1
p90RSK_pT359_S363-R-C_GBL9013438	p90RSK_pT359_S363	RPS6KA1	Use with Caution	Rabbit	CST	9344
PAI-1-M-NA_GBL9013500	PAI-1	SERPINE1	NA	Mouse	BD Biosciences	612024
PARP_cleaved-M-C_GBL9013420	PARP_cleaved	PARP1	Use with Caution	Mouse	CST	9546
Paxillin-R-V_GBL9013288	Paxillin	PXN	Validated	Rabbit	Epitomics	1500-1
PCNA-M-V_GBL9013360	PCNA	PCNA	Validated	Mouse	Abcam	ab29
PDCD4-R-NA_GBL9012498	PDCD4	PDCD4	NA	Rabbit	Rockland	600-401-965
PDK1_pS241-R-V_GBL9013289	PDK1_pS241	PDK1	Validated	Rabbit	CST	3061
Pea-15-R-V_GBL9013290	Pea-15	PEA15	Validated	Rabbit	CST	2780
PI3K-p110-alpha-R-C_GBL9013291	PI3K-p110-alpha	PIK3CA	Use with Caution	Rabbit	CST	4255
PKC-alpha_pS657-R-V_GBL9013293	PKC-alpha_pS657	PRKCA	Validated	Rabbit	Upstate (Millipore)	06-822
PKC-alpha-M-V_GBL9013374	PKC-alpha	PRKCA	Validated	Mouse	Upstate (Millipore)	05-154
PKC-delta_pS664-R-V_GBL9013484	PKC-delta_pS664	PRKCD	Validated	Rabbit	Millipore	07-875
PR-R-V_GBL9013294	PR	PGR	Validated	Rabbit	Epitomics	1483-1
PRAS40_pT246-R-V_GBL9013295	PRAS40_pT246	AKT1S1	Validated	Rabbit	Biosource	441100G

PRDX1-R-NA_GBL9013449	PRDX1	PRDX1	NA	Rabbit	Sigma/Atlas	HPA007730
PTCH-R-C_GBL9013296	PTCH	PTCH1	Use with Caution	Rabbit	SDI	2113.00.02
PTEN-R-V_GBL9013297	PTEN	PTEN	Validated	Rabbit	CST	9552
Rab25-R-C_GBL9013298	Rab25	RAB25	Use with Caution	Rabbit	Covance Custom	Covance Custom
Rad50-M-C_GBL9013362	Rad50	RAD50	Use with Caution	Mouse	Millipore	05-525
Rad51-M-C_GBL9013385	Rad51	RAD51	Use with Caution	Mouse	Chem Biotech	na 71
Rb_pS807_S811-R-V_GBL9013301	Rb_pS807_S811	RB1	Validated	Rabbit	CST	9308
Rb-M-V_GBL9013387	Rb	RB1	Validated	Mouse	CST	9309
RBM3-M-NA_GBL9013452	RBM3	RBM3	Validated	Mouse	Atlas	CAB030038
S6_pS235_S236-R-V_GBL9013303	S6_pS235_S236	RPS6	Validated	Rabbit	CST	2211
S6_pS240_S244-R-V_GBL9013411	S6_pS240_S244	RPS6	Validated	Rabbit	CST	2215
S6-R-NA_GBL9013302	S6	RPS6	NA	Rabbit	CST	2217
SCD1-M-NA_GBL9013502	SCD1	SCD1	Validated	Mouse	Santa Cruz	sc-58420
SETD2-R-NA_GBL9013256	SETD2	SETD2	NA	Rabbit	Abcam	ab69836
Shc_pY317-R-NA_GBL9013304	Shc_pY317	SHC1	NA	Rabbit	CST	2431
Smac-M-V_GBL9013476	Smac	DIABLO	Validated	Mouse	CST	2954
Smad1-R-V_GBL9013485	Smad1	SMAD1	Validated	Rabbit	Epitomics	1649-1
Smad3-R-V_GBL9013305	Smad3	SMAD3	Validated	Rabbit	Epitomics	1735-1
Smad4-M-C_GBL9013389	Smad4	SMAD4	Use with Caution	Mouse	Santa Cruz	sc-7866
Snail-M-C_GBL9013426	Snail	SNAI2	Use with Caution	Mouse	CST	3895
Src_pY416-R-C_GBL9013307	Src_pY416	SRC	Use with Caution	Rabbit	CST	2101
Src_pY527-R-V_GBL9013306	Src_pY527	SRC	Validated	Rabbit	CST	2105
Src-M-V_GBL9013358	Src	SRC	Validated	Mouse	Upstate (Millipore)	05-184
STAT3_pY705-R-V_GBL9013308	STAT3_pY705	STAT3	Validated	Rabbit	CST	9131
STAT5-alpha-R-V_GBL9013309	STAT5-alpha	STAT5A	Validated	Rabbit	Epitomics	1289-1
Stathmin-R-V_GBL9013310	Stathmin	STMN1	Validated	Rabbit	Epitomics	1972-1
Syk-M-V_GBL9013477	Syk	SYK	Validated	Mouse	Santa Cruz	sc-1240
Tau-M-C_GBL9013349	Tau	MAPT	Use with Caution	Mouse	Upstate (Millipore)	05-348
TAZ_pS89-R-C_GBL9013444	TAZ_pS89	WWTR1	Use with Caution	Rabbit	Santa Cruz	sc-17610
Transglutaminase-M-V_GBL9013418	Transglutaminase	TGM2	Validated	Mouse	Lab Vision	MS-224
Tuberin-R-C_GBL9013314	Tuberin	TSC2	Use with Caution	Rabbit	Epitomics	1613-1
VASP-R-C_GBL9013315	VASP	VASP	Use with Caution	Rabbit	CST	3112
VEGFR2-R-C_GBL9013446	VEGFR2	KDR	Use with Caution	Rabbit	CST	2479
XBP1-G-C_GBL9013416	XBP1	XBP1	Use with Caution	Goat	Santa Cruz	sc-32136

XIAP-R-C_GBL9013317	XIAP	XIAP	Use with Caution	Rabbit	CST	2042
XRCC1-R-C_GBL9013403	XRCC1	XRCC1	Use with Caution	Rabbit	CST	2735
YAP_pS127-R-C_GBL9013442	YAP_pS127	YAP1	Use with Caution	Rabbit	CST	4911
YAP-R-V_GBL9013441	YAP	YAP1	Validated	Rabbit	Santa Cruz	sc-15407
YB-1_pS102-R-V_GBL9013443	YB-1_pS102	YBX1	Validated	Rabbit	CST	2900
YB-1-R-V_GBL9013448	YB-1	YBX1	Validated	Rabbit	SDI	1725.00.02

IX. Batch effects analysis for TCGA breast cancer data sets

We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the breast cancer data sets. Five different data sets were analyzed: mRNA expression (Agilent G4502A microarray), mRNA expression (RNA-seq Illumina GA), miRNA expression (RNA-seq Illumina GA), DNA methylation (Infinium HM27 microarray), and SNPs (GW SNP 6). All of the data sets were at TCGA level 3, the level which of the analyses in the paper are based. We assessed batch effects with respect to two variables; batch ID and TSS.

For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure. We clustered the samples and then annotated them with colored bars at the bottom. Each color corresponded to a batch ID or a TSS. For PCA, we plotted the first four principal components, but only plots of the first two components are shown here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) were connected to the batch centroid by lines. The centroids were computed by taking the mean across all samples in the batch. That procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches. The results for the five data sets follow.

mRNA Expression (Agilent G4502A microarray)

Figures IX.1-3 show clustering and PCA plots for the Agilent G4502A mRNA expression platform. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.

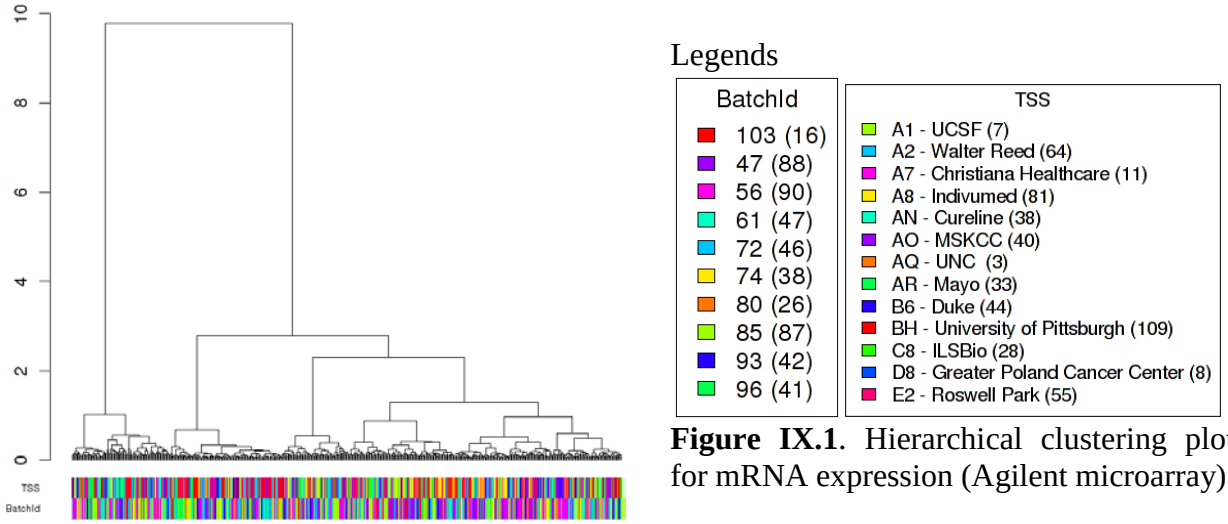


Figure IX.1. Hierarchical clustering plot for mRNA expression (Agilent microarray)

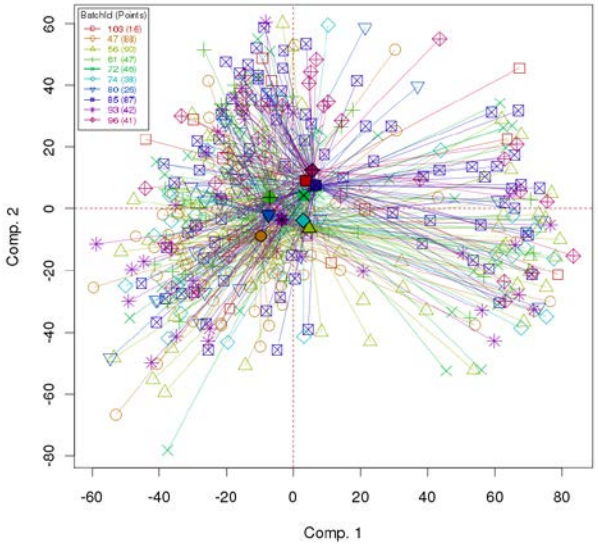


Figure IX.2. First two principal components for mRNA microarray expression, with samples connected by centroids according to batch ID.

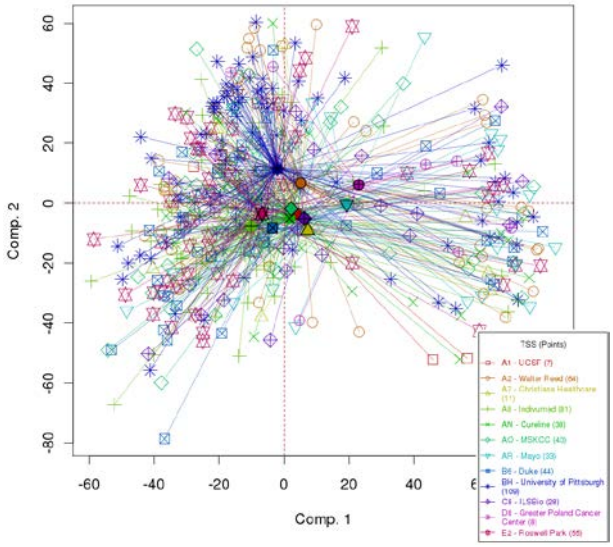
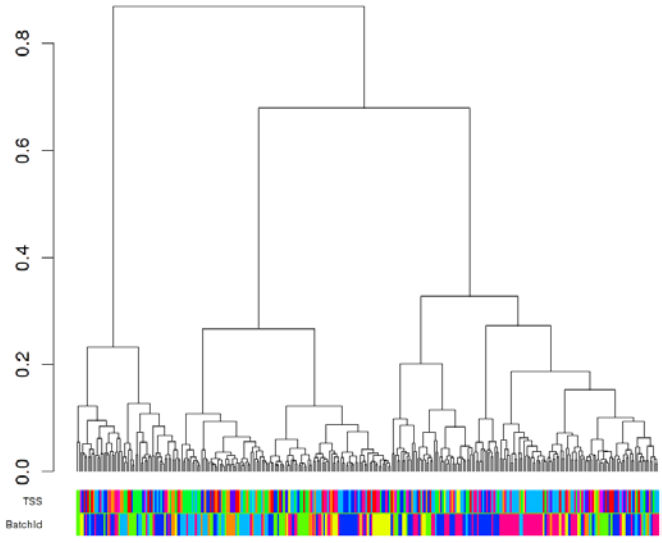


Figure IX.3. First two principal components for mRNA microarray expression, with samples connected by centroids according to TSS.

miRNA Expression (RNA-seq Illumina GA)

The following figures show clustering and PCA plots for RNA-seq miRNA data (Figures IX.5-7). Genes with zero values were removed and the read counts were $\log_2$-transformed before generating the figures. Unlike the other data types, miRNA expression does show small amounts of clustering by batch ID in the hierarchical clustering plot. However, the PCA plots do not show any batches that significantly stand out from the others. For that reason, we didn't consider batch effects to be so strong as to warrant any batch effects correction. The trade off with batch effects correction algorithms is the possibility of losing important biological variation in the data, along with the technical variation.



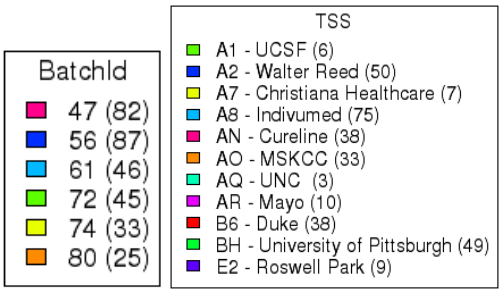


Figure IX.4. Hierarchical clustering of samples for miRNA expression from RNA-seq data.

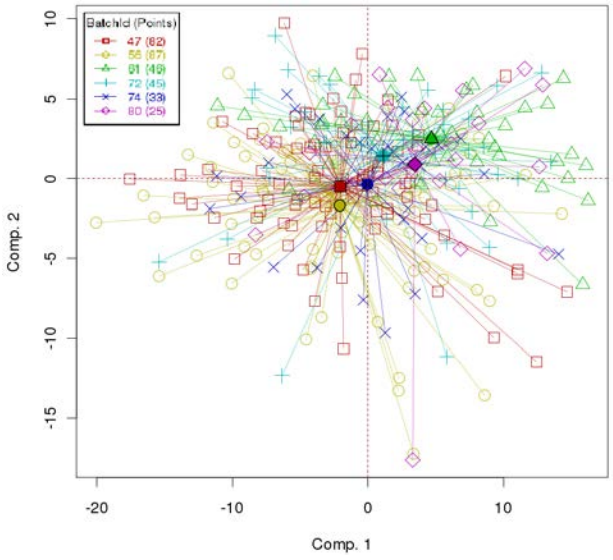


Fig. IX.5. PCA: First two principal components for miRNA expression from RNA-seq data, with samples connected by centroids according to batch ID.

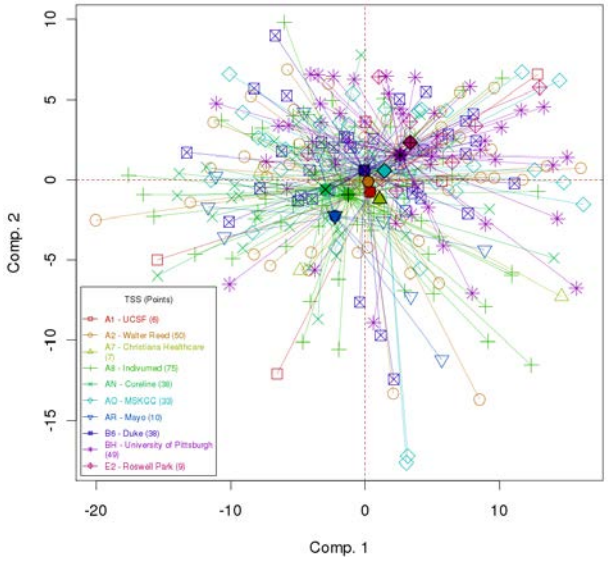


Fig. IX.6. PCA: First two principal components for miRNA expression from RNA-seq data, with samples connected by centroids according to TSS.

DNA Methylation (Infinium HM27 microarray)

The following figures show clustering and PCA plots for the Infinium DNA methylation platform (Figures IX.7-9). None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.

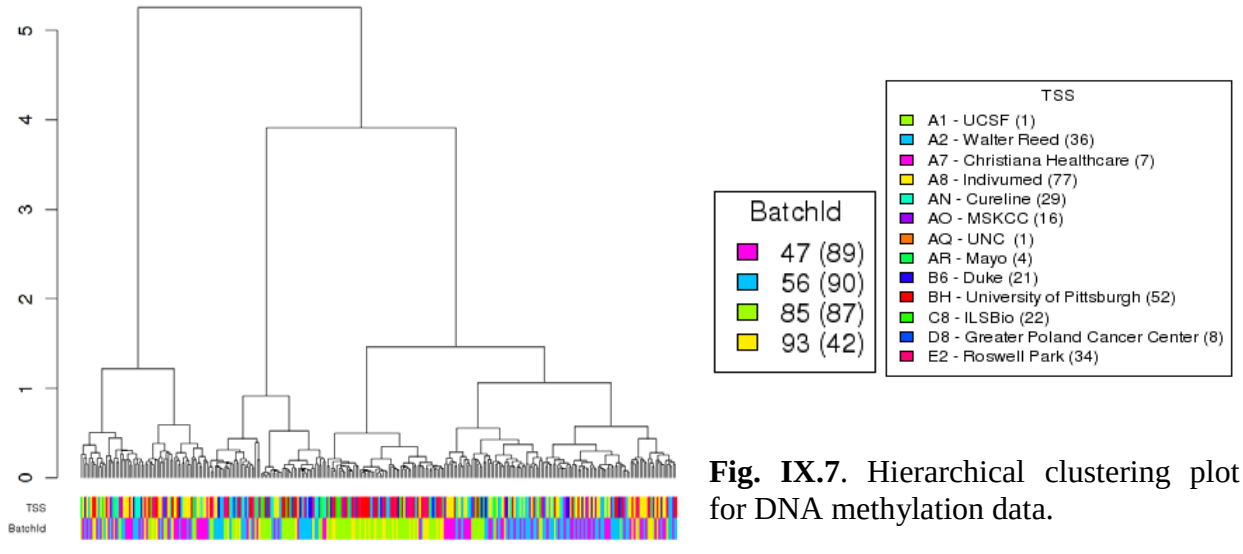


Fig. IX.7. Hierarchical clustering plot for DNA methylation data.

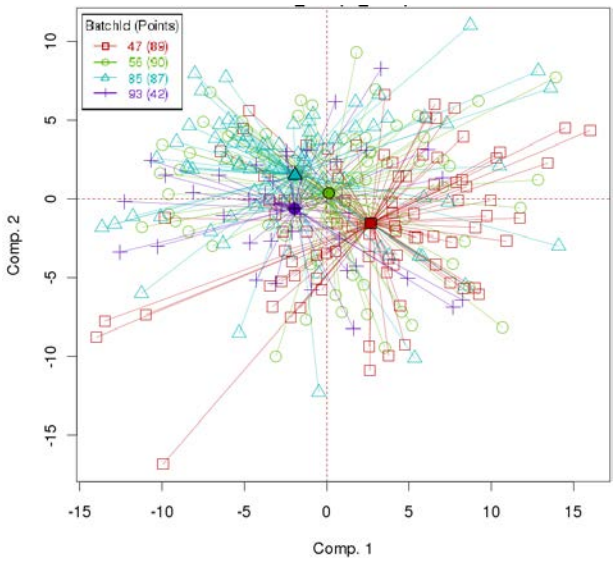


Fig. IX.8. PCA for DNA methylation, with samples connected by centroids according to batch ID.

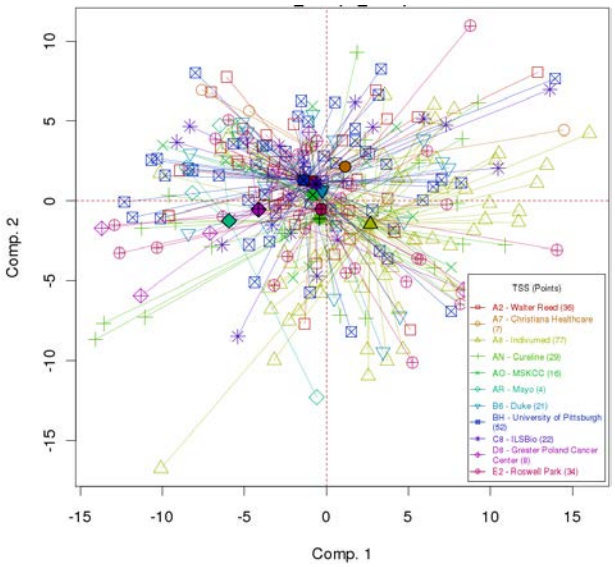
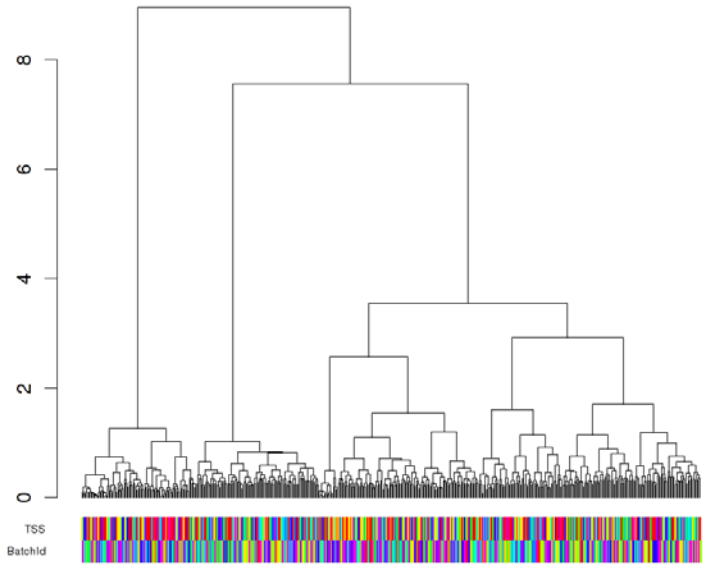


Fig. IX.9. PCA for DNA methylation, with samples connected by centroids according to TSS.

SNPs (GW SNP 6)

The following figures show clustering and PCA plots for the SNP platform (Figures IX.10-12). At level 3, the TCGA SNP data resemble copy number data when we use chromosomal segment counts (rather than actual SNPs). We mapped the chromosomal segments to genes (using build hg18) and then used them to construct the plots shown in Figs. 13-15. Once again, none of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.



Legends

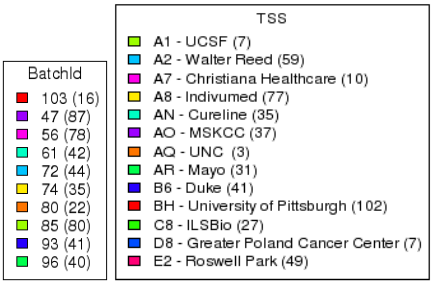


Fig. IX.10. Hierarchical clustering plot for SNP data.

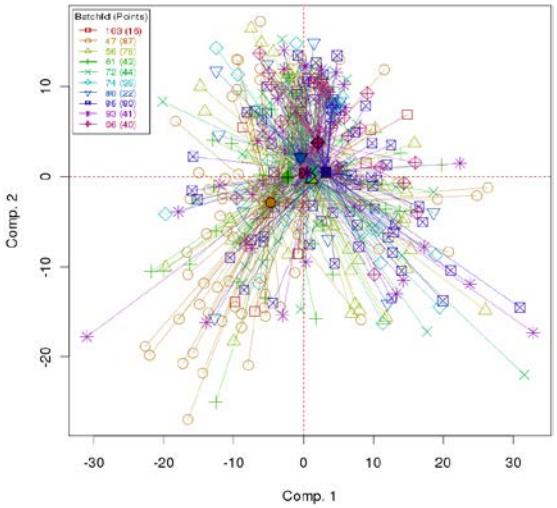


Fig. IX.11. PCA for SNPs, with samples connected by centroids according to batch ID.

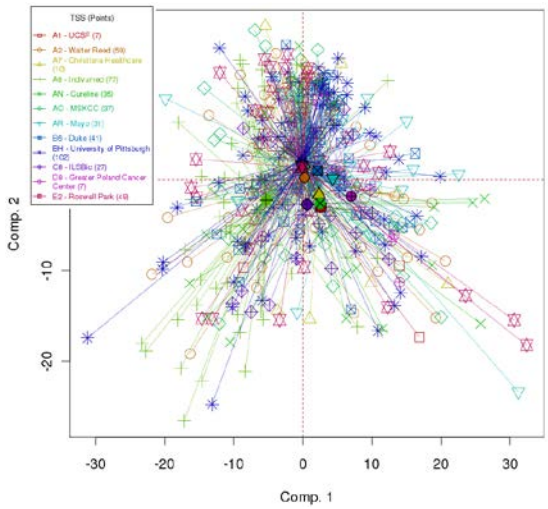


Fig. IX.12. PCA for SNPs, with samples connected by centroids according to TSS.

Conclusions

Overall, the TCGA batch effects in the breast cancer data sets are minimal. In 4 out of 5 data sets, no major batch effects were observed by either clustering or PCA plots. In miRNA expression, some small clusters of batches were observed in the hierarchical clustering plot, but not in the PCA plots. We didn't consider those effects to be strong enough to warrant batch effects correction. Batch effects correction algorithms run the risk of removing important biological variation along with technical variation. Based on the above figures, we believe overall that technical batch effects in the data sets are reasonably small and unlikely to influence high-level analyses in a major way.

X. Cross Platform Subtype Analysis

Subtype calls from each of the 5 platforms analyzed for subtypes within each data type were used to identify relationships between the different classifications. Subtypes defined from each platform were coded into a series of indicator variables for each subtype. The matrix of 1 and 0s was used in ConsensusClusterPlus R-package[40,41], to identify structure and relationship of the samples. Parameters for Consensus cluster were 80% sample resampling with 1000 iterations of hierarchical clustering based on a Pearson correlation distance metric. Correlation of subtypes with clinical features and mutations was done using a Pearson's Chi-Squared test or Fisher's Exact test in R.

XI. PARADIGM Analyses

PARADIGM integrated pathway analysis of copy number and expression data

Integration of copy number, mRNA expression and pathway interaction data was performed on the 463 samples using the PARADIGM software[42]. Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes using pathway interactions and genomic and functional genomic data from a single patient sample. The mRNA data was converted to relative mRNA expression levels by subtracting each gene's median computed over

22 tumor-adjacent normal controls from its level observed in each patient sample. Level 3 copy number data (segmented and normalized to reflect the difference in copy number between a gene's level detected in tumor versus normal blood) was mapped to the genome using the UCSC hg19 Knowngenes track. Gene-level copy number estimates were then derived by taking the median of all segments falling within the length of the gene. Both expression and gene-level copy number data were then rank transformed before use by the PARADIGM analysis.

Pathways were obtained in BioPax Level 3 format, and included the NCIPID and BioCarta databases from http://pid.nci.nih.gov and the Reactome database from http://reactome.org. Gene identifiers were unified by UniProt ID then converted to Human Genome Nomenclature Committee's HUGO symbol using mappings provided by HGNC (http://www.genenames.org/). Interactions from all of these sources were then combined into a merged Superimposed Pathway (SuperPathway). Genes, complexes, and abstract processes (e.g. "cell cycle" and "apoptosis") were retained and henceforth referred to collectively as pathway concepts. Before merging gene concepts, all gene identifiers were translated into HUGO standard identifiers wherever possible. The belief propagation algorithm employed by PARADIGM can be run with cycles and contradictory interactions. Therefore, for the sake of completeness and simplicity, all interactions were included and no attempt was made to resolve conflicting influences if they existed in the resulting SuperPathway. A breadth-first traversal starting from the concept with the highest number of interactions was performed to build one single component. The resulting pathway structure contained a total of 16352 concepts, representing 6906 proteins, 7345 complexes, 1449 families, 55 RNAs, 15 miRNAs and 582 processes.

The PARADIGM algorithm infers an integrated pathway level (IPL) for each gene that reflects a gene's activity in a tumor sample relative to the normal controls. Including only pathway concepts with relative activities distinguishable from normal (0.05 absolute activity) in at least one patient sample and non-zero activity in at least 10% of the samples yielded over 12,000 concepts. To identify patient subtypes implicated from shared patterns of pathway inference, we ran Consensus clustering using the median-centered IPLs implemented with the ConsensusClusterPlus package[41] in R [http://www.R-project.org] with 80% subsampling over 1000 iterations of hierarchical clustering based on a Pearson correlation distance metric. Heatmap display of the top 1000 varying IPLs was generated using the heatmap.plus package in R.

Pathway-based biomarkers of basal-like versus luminal subtypes.

IPLs differentially activated between the basal and luminal (A+B) subtypes were identified using the t-test and Wilcoxon Rank Sum test with Benjamini-Hochberg(BH) FDR correction. Only features deemed significant (FDR corrected $p<0.05$) by both tests were selected. Pathways enriched among differentially activated IPLs were assessed using the EASE score[43] with BH FDR correction; and sub-networks were constructed to identify regulatory hubs based on interconnectivity and visualized using Cytoscape[44]. Pathway-based biomarkers differentiating luminal A and B subtypes were similarly identified (Supplemental Figure 14).

Machine learning classifiers based on mRNA expression classify serous ovarian samples as basal-like

We first asked whether mRNA expression data supports the hypothesis that basal cancers share common molecular signatures with the TCGA serous ovarian samples. To address this, we asked whether machine-learning classifiers, trained to recognize basal from luminal samples using

mRNA expression also classify ovarian samples as basal even though the ovarian samples were not used during training. To this end, we implemented a Nonnegative Matrix Factorization (NMF) approach adapted for supervised learning as previously described[45] using the Weka Java library[46]. We tested the ability of the predictor to identify both luminal and basal samples held out of training in a cross-validation test. To conduct such a test, we first partitioned the data into a set of 300 samples used for cross-validation accuracy estimation and another distinct set of 80 separated out for a second-stage validation. The 80 validation samples made up of 13 basal-likes and 67 luminals represent an important set for performing a second accuracy evaluation to ensure that models selected to maximize cross-validation performance generalize to unseen cases. With the 300 samples we repeated 5 different runs of 5-fold cross-validation in which 80% of the 300 samples were separated out and used to train the model while the remaining 20% was used to measure the model's accuracy. In every case, the basal-like predictor trained on 80% of the samples correctly predicted every sample in the held-out 20%. In addition, the model was found to be 100% accurate when tested on the 80-sample validation set. In every case, the predicted samples received scores more extreme than any seen among scores computed from chimeric background samples. Thus, serous ovarian samples look indistinguishable to the NMF machine-learning classifier trained to recognize basal from luminal breast cancer supporting the hypothesis that basal-like and serous ovarian tumors share common molecular signatures.

Because our goal was to classify new samples as either basal-like or not, it was necessary to develop an appropriate background model so that we could determine if new samples are significantly similar to basal-like samples seen during training. We therefore simulated "chimeric" samples in which half of the mRNA levels were drawn from basal patients and the other half of the mRNA levels were drawn from luminal patients. While we expect the classifier scores on such chimeric samples to hover around zero (halfway between basal and luminal classification) the simulation provides an estimate of the variability of the classification scores needed when assessing significance of new classifications. Comparing the scores of the TCGA breast samples to the chimeric background did indeed reveal that most of the true TCGA breast samples were classified into their correct sub-types with scores exceeding chance expectation.

We applied the TCGA-breast classifiers to external datasets to determine if they could robustly classify serous ovarian samples collected by TCGA as well as multiple other tumor types. Indeed we found that most of the ovarian samples were classified as basal-like with 411 samples predicted as significantly basal out of the 441 that were predicted (i.e. outside the range of the background distribution). Next, data was collected from the T-GEN expO dataset that included samples on colorectal, lung squamous cell carcinoma, kidney, prostate, ovarian, and breast. Note that this test therefore also provides an external validation of the basal predictors on breast samples not included as part of this publication. To check if the chimeric distribution estimated with the TCGA basal-luminal predictors is reasonable in this setting, we also generated chimeric samples from the external data and verified that the chimeric samples do indeed fall in the range estimated using only the TCGA basal-luminal chimeras. Chimeric samples from external data were made by randomly shuffling tissue subtypes in the T-GEN cohort. Reflecting the cross-platform accuracy of the models, the classifiers trained from the TCGA samples were accurate in identifying the PAM50 subtypes in external datasets (72 out of 81 T-GEN basal-likes correctly predicted). In addition to most of the ovarian samples receiving high basal classifier scores in the T-GEN dataset (134 out of 204), two other tumor types also had a number of samples receiving significant basal-like scores including 81% of the colon (235 out of 289) and

74% of the lung (96 out of 129). These results suggest that basal cancers share a common molecular program with other cancers perhaps of epithelial origin.

Genes with shared expression patterns in basal-like and serous ovarian are significantly interconnected by known pathway interactions.

Integrated pathway levels generated from PARADIGM for the 377 TCGA ovarian samples were obtained; and the average IPL across samples was computed. Among the ~14K features present in the ovarian dataset, 5763 mapped to IPLs showing significant differential activation between basal vs. luminal (A+B) breast cancers identified as described above. Restricting to these IPLs, a linear fit of average ovarian activity onto the basal-like vs. luminal differential score was performed (Figure XI.1). A basalness score was computed as the orthogonal projection of the average ovarian activity onto the linear fit. Features with basalness scores at least two standard deviations from the mean were defined as significant; and regulatory sub-networks within the SuperPathway structure linking these features were identified and displayed using Cytoscape.

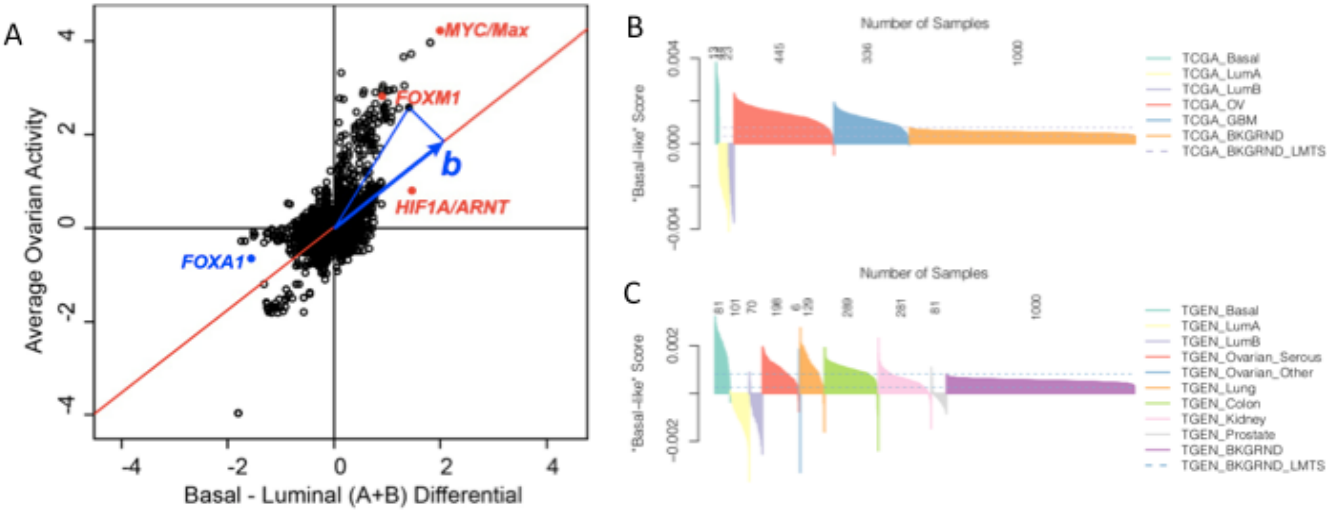


Figure XI.1. Basal-like versus ovarian comparison. A. PARADIGM inferences differential for basal-like versus luminal are highly concordant with overall inferred activity in the TCGA serous ovarian cohort. Scatterplot of average ovarian activity vs. basal-like – luminal differential scores. Average ovarian activity was computed across 377 samples; and plotted against the mean difference in activity between basal and luminal breast cancers for the 5763 features identified as differentially activated between these breast cancer subtypes. Regression line of ovarian activity on the basal-luminal differential was fit (red line) and the orthogonal projection of a given point (blue arrow) onto the linear fit was determined to calculate the basalness score (b). Highlighted in red are specific points representing important regulatory hubs (MYC/Max, HIF1A/ARNT, FOXM1) significantly activated in basal breast cancers; and highlighted in blue is the FOXA1 hub with significantly higher activity in luminal breast and lower in ovarian cancers. (B) Basal-like -Luminal predictors significantly classify serous ovarian samples as basal-like. A basal-luminal classifier was trained on 80% randomly selected TCGA BRCA data. (C) Basal-like prediction scores classify several tumors from multiple tissue types as Basal-like. Plotted in red hues are prediction scores from remaining 20%BRCA+OV TCGA. BRCA samples are colored by subtype. In blue hues are prediction scores from T-GEN GEO data. Margins of the chimeric

background distributions from 20%BRCA+OV and T-GEN GEO are shown in orange and blue dashed lines, respectively.

Assessing the significance of pathway-based biomarker network maps

We assessed the significance of the pathway marker maps using 1000 simulated cohorts each containing a different random assignment of labels (e.g. basal/luminal) to random patient samples so that the number of generated subtypes matched the number in the original data set. Random patient samples in the simulated cohort were constructed by permuting the gene names within a patient sample, which effectively scrambled the association of data tuples to the pathway topology. We then assessed significance by asking whether the identified pathway signature had more concepts interconnected together than would be expected by chance. We first collected all pathway features passing the two tests of significance above without multiple test adjustment. We then retained significant links as any regulatory connection from the SuperPathway connecting two significantly-scoring features. We then identified the largest connected component (LCC) of the graph defined as the subgraph with the most number of pathway features identified from a depth-first traversal of the resulting significant links. The procedure was repeated for each of the 1000 simulated cohorts, obtaining a distribution of the LCC sizes for use as a background distribution. The observed LCC sizes derived from SuperPathway networks for the Basal-like versus luminal (A+B), for the Basal-like comparison to serous ovarian, and the Luminal B versus Luminal A were all found to be significantly higher than expected given the background control of their respective simulated cohorts (Figure XI.2).

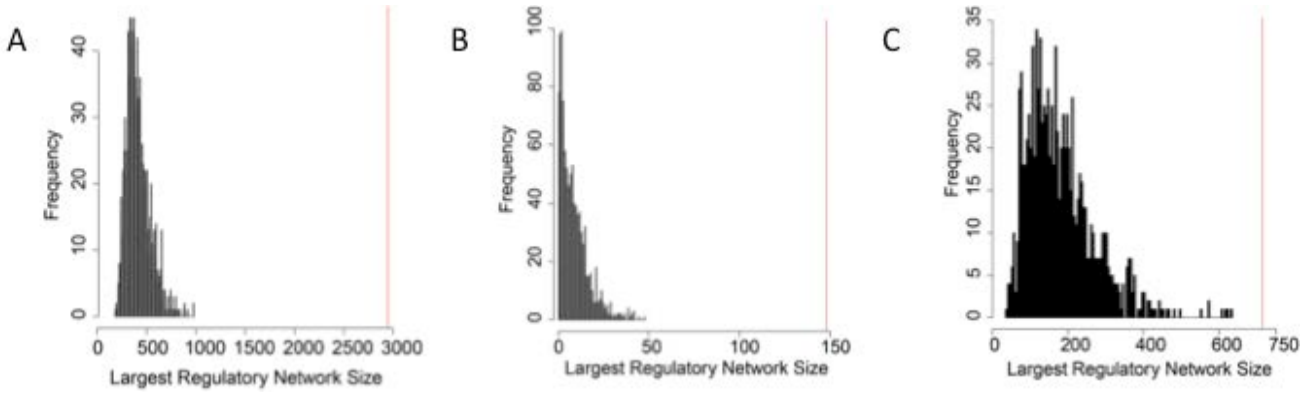


Figure XI.2. Genes differentially activated in basal-like tumors compared to luminal tumors and Luminal B's versus Luminal A's are significantly interconnected in known pathways. The collection of all pathway features from the PARADIGM SuperPathway that had significant differential activity in one subtype versus another was collected. The size of the largest connected component (LCC) in the SuperPathway (x-axis in A-C) for these features was recorded (red vertical line in A-C). A background distribution of the LCC size was determined using random re-labelings of the patient subtypes and using random patient data; the frequency of the background distribution was then plotted for quantized values of the LCC size (y-axis; black bars in A-C). (A) Significance of the LCC size for the Basal-like versus Luminal (A+B) comparison. (B) LCC size significance for the networks derived with the basalness score in which pathway features were scored according to whether they were both differential in the Basal-like versus luminal comparison and activated in serous ovarian TCGA samples. (C) LCC size significance of the Luminal B versus Luminal A comparison .

Inter-sample correlations between TCGA mRNA profiles and "TGEN expO" profiles
For the inter-sample mRNA correlations featured in main Figure 7C, the "expO" dataset from T-Gen was obtained from the Gene Expression Omnibus (GSE2109). The analysis focused on the profiles of breast, colon, lung, kidney, ovarian (including serous) and prostate, n=1337 profiles in all, including 204 ovarian and 89 serous ovarian. For each of the two mRNA profile datasets (TCGA breast and TGEN expO), we first normalized genes across samples to standard deviations from the centroid mean of the major groups. We then took the Pearson's correlation of each expO profile (using all the genes) with each TCGA breast profile.

XII. Integrated Pathway Analysis

Gene transcription signatures of pathways were defined as follows. P53 pathway: "IARC" signature, canonical bound and up-regulated p53 gene targets, as catalogued in the p53 IARC database (http://www-p53.iarc.fr/TargetGenes.html); "GSK" signature, from Glaxo-Smith-Kline (GSK) cell line database, coupled with "R14" p53 database of mutations in cell lines (N=248 cell lines with TP53 status), where a t-test of P<0.01 was used to determine genes higher in wt versus mutant cell lines; "Kannan" signature, from MSigDB ("UP" targets), http://www.broadinstitute.org/gsea/msigdb/cards/KANNAN_TP53_TARGETS_UP.html; "Troester" signature, list of genes reported repressed by TP53 knockdown in MCF7 cells (from Troester et al.[47]. RB pathway: "Lara" signature, from GEO database GSE9562, comparing mouse keratinocyte cultures with RB1 knockout versus wt (P<0.01, fold>1.5); "Chicas" signature, from GSE19864 profiles of RNAi-mediated suppression of RB in IMR90 cells (using P<0.01, fold>1.5); "Herschkowitz" signature, genes differentially expressed in breast tumors with RB1 LOH[48]. PI3K pathway: Gene signatures were described previously in Creighton et al.[49]; "Saal" PTEN loss signature, genes correlated with Pten protein levels in breast cancer; "CMap" PI3K/mTOR signature, genes modulated *in vitro* by inhibitors to PI3K or mTOR, according to CMap dataset (P<0.01, comparing PI3K/mTOR-inhibited cells with the rest of the Cmap profiles); "Majumder" Akt signature, genes modulated in a mouse model of inducible Akt (P<0.01).

For a given gene transcription signature, we extracted the expression values from the TCGA gene expression array dataset. For each gene, we normalized expression values to standard deviations from the median across tumors. For signatures with genes moving in one direction (the p53 signatures), we computed the average normalized expression of the signature genes within each tumor. For signatures with "up" and "down" genes (the Rb and PI3K signatures), we computed our previously described "t-score"[8,49]. For evaluating the significance of correlation between specific molecular features and pathway signatures, we first normalized the gene signature scores across tumors to standard deviations from the median across tumors, and a "summary score" for each pathway was then computed as the average of the individual normalized signature scores.

The PI3K RPPA proteomic signature consisted of the signature described previously[49], with the addition of p4EBP1 and INPP4B. For each tumor, the PI3K protein score was the sum of the phosphoprotein levels of Akt, mTOR, GSK3, S6K, S6 and 4EBP1, minus the total levels of pathway inhibitors PTEN and INPP4B (all proteins levels being first normalized to standard deviations from the median across tumors); in other words, PI3K score = [pAkt + pmTOR + pGSK3 + pS6K + pS6 + p4EBP1] – [INPP4b + PTEN].

XIII. Integrated Analysis

Subtype Enriched Alterations

Given the tremendous heterogeneity displayed by breast tumors, we analyzed the overall spectrum of genomic alterations across different subtypes looking for subtype-specific patterns. Our approach relies on the general abstraction of gene alteration per sample where each alteration belongs to one of the 3 categories:

- **Category 1:** Gene is altered by mutations.
- **Category 2:** Gene is primarily altered by copy number alterations, and mRNA expression levels correlate with copy number changes.
- **Category 3:** "Wild-card" events (e.g. gene shows aberrant mRNA expression and/or methylation status independent of mutations and copy number).

Categories 1 and 2 rely on two systematic approaches: for mutations we selectively analyzed the list of SMGs identified by the algorithm MuSiC (http://gmt.genome.wustl.edu/genome-music/current/), for copy number we analyzed frequently amplified and deleted Region of Interest (ROI) as identified by GISTIC (Beroukhim, 2007). Category 3 allows the user to specify a genomic event of interest. In our analysis wild cards events included:

- AKT3 over-expression (>1 Standard Deviation, SD, from the average)
- RB1 down-regulation (<3 SD)
- PTEN down-regulation (<3 SD)
- EGFR over-expression (>2 SD)
- BRCA1 hyper-methylation

All our analyses were run on the 463 samples dataset (three metastasis were not included).

To systematically look for subtype-specific genomic events, we develop a method: Subtype Enriched Alterations (SEA). Subtype enrichment is tested in two steps: (1) the distribution of alterations is compared to the expected given the number of samples that belong to each subtype by a *goodness-of-fit* test, (2) a hypergeometric p-value is derived for the subtype with highest percentage of alterations when compared against all the others. Alterations in category are tested separately and treated independently. PAM50 subtypes results are showed in Table XIII.1.

This analysis confirmed previous findings indicating TP53 mutations and MYC amplification as basal-like events, CCND1 amplification more frequent in luminal, PIK3CA and MAP3K1 mutations enriched in luminal A, and ERBB2 amplification as marker of the HER2-enriched subtype. Other interesting subtype enriched alterations include 12p amplification in basal-like tumors, 4q amplification in HER2-enriched, and KCNB2 mutations in luminal B tumors.

We noticed remarkably similarities between basal-like and serous ovarian cancer that goes beyond TP53 mutation, to include AKT3 over-expression, CCNE1 amplification, and BRCA1 mutation and hyper-methylation. BRCA1 hyper-methylation which was present in a relatively large percentage of basal-like samples never co-occurs with BRCA1/2 mutations.

Table XIII.1: Subtype enriched alterations (SEA) analysis across breast cancer intrinsic subtypes. The table shows all the events such that at least one between the FDR corrected chi-square p-value and the nominal hypergeometric p-value is < 0.05.

COPY NUMBER REGION OF INTEREST (ROI)	Chisq. test	Chisq. FDR	p-value	Enriched Set	Her2 [53]	Basal [81]	LumB [112]	LumA [209]	Genes in the ROI
chr17:35122703-35124065	7.18E-30	2.37E-28	5.35E-25	Her2	37	3	13	10	ERBB2
chr10:12118816-12131522	3.54E-06	5.84E-05	1.06E-05	Basal	1	10	2	1	UPF2
chr11:69185447-69190570	4.35E-06	5.84E-05	1.02E-06	LumB	9	2	37	30	CCND1
chr4:74353853-74353853	3.33E-05	2.75E-04	2.78E-04	Her2	6	0	2	2	ANKRD17
chr8:128824019-128837392	5.40E-05	3.56E-04	3.63E-05	Basal	11	25	16	16	MYC
chr13:47774593-47957722	5.93E-05	3.56E-04	1.63E-04	Basal	0	6	0	1	RB1
chr19:35007110-35025598	0.000	0.001	3.42E-04	Basal	0	7	0	3	CCNE1
chr1:120228172-120273351	0.000	0.001	3.42E-04	Basal	0	7	2	1	NOTCH2
chr1:148885942-148912194	0.002	0.008	0.005	Basal	8	15	12	9	GOLPH3L
chr12:67883912-68036772	0.004	0.012	0.001	LumB	3	2	13	5	[MDM2]
chr12:792653-795167	0.005	0.016	0.003	Basal	0	6	2	2	WNK1
chr6:13622618-13679779	0.015	0.042	0.011	Basal	0	4	0	2	[GFOD1,SIRT5]
chr3:14837984-14885073	0.018	0.046	0.037	Her2	3	1	3	0	FGD5
chr8:37835997-37863117	0.022	0.052	0.009	LumB	3	6	24	31	RAB11FIP1, [ZNF703,WHSC1L1]
chr1:202750134-202795870	0.043	0.095	0.027	Her2	7	2	9	9	MDM4
chr9:21953430-21986996	0.050	0.103	0.013	Basal	2	7	3	4	CDKN2A
chr11:77021940-77038715	0.092	0.179	0.025	LumB	7	4	18	19	CLNS1A,[PAK1]
chr11:32879476-32882898	0.108	0.197	0.030	Basal	1	5	1	4	QSER1

MUTATED GENES	Chisq. test	Chisq. FDR	p-value	Enriched Set	Her2 [53]	Basal [81]	LumB [112]	LumA [209]
TP53	4.45E-23	1.87E-21	7.32E-22	Basal	40	68	36	24
PIK3CA	2.42E-06	5.08E-05	2.29E-07	LumA	22	6	35	102
KCNB2	0.00036807	0.00515292	2.01E-04	LumB	0	0	6	0
MAP3K1	0.00037895	0.00515292	1.57E-05	LumA	2	0	6	30
ATP1A4	0.00055276	0.00515292	0.00156833	Her2	5	1	2	1
GATA3	0.00190507	0.01333547	0.05190173	LumB	1	1	17	29
USH2A	0.00328174	0.01969043	0.00365988	Basal	4	9	1	6
FLG	0.01123141	0.0589649	0.02001072	Basal	5	8	4	4
CDH1	0.01128181	0.0589649	0.00160036	LumA	3	0	5	22
BRCA2	0.01387822	0.0589649	0.02028732	Her2	4	3	2	1
MAP2K4	0.0175458	0.06699305	0.00163222	LumA	1	0	3	16
PTPN22	0.02257807	0.07902325	0.03726817	Her2	3	0	3	1
PREX2	0.03580291	0.11567094	0.09916299	Her2	3	4	2	1
RB1	0.03640809	0.11567094	0.0724606	Basal	0	3	3	0
RPGR	0.03720155	0.11567094	0.02131714	Basal	0	4	2	1
BRCA1	0.04002219	0.11567094	0.02131714	Basal	1	4	0	2
CTCF	0.1031446	0.25482784	0.01625014	LumA	1	0	1	9

WILD CARD EVENTS	Chisq. test	Chisq. FDR	p-value	Enriched Set	Her2 [53]	Basal [81]	LumB [112]	LumA [209]
AKT3 over-espression	3.97E-18	6.36E-17	1.86E-15	Basal	0	25	1	5
RB1 down-regulation	1.24E-09	9.90E-09	2.32E-08	Basal	1	16	5	1
BRCA1 germline mutations	9.38E-08	5.00E-07	9.55E-07	Basal	0	10	0	2
BRCA1 hyper-methylated	4.95E-07	1.98E-06	2.31E-06	Basal	1	11	2	1
PTEN down-regulation	2.57E-05	8.23E-05	7.74E-05	Basal	1	8	2	0
EGFR over-espression	0.0003716	0.00099094	0.0030862	Basal	3	6	1	0
TP53 germline mutations	0.01644679	0.03759266	0.07550029	Her2	3	4	1	1
BRIP1 germline mutations	0.0268991	0.0537982	0.01461397	LumB	0	0	3	0

Mutually Exclusivity Modules in Cancer (MEMo)

To analyze genomic alterations in a pathway context we use the algorithm MEMo[50]. MEMo (Mutual Exclusivity Modules) automatically identifies mutually exclusive alterations targeting frequently altered genes that are likely to belong to the same pathway. We first ran MEMo across all breast tumors, selecting alterations affecting at least 2% of the samples (10 samples out of 463). Again, genomic events were defined as in the previous section following the "gene alteration per sample" abstraction. On this dataset, MEMo identified up to 22 statistically significant modules (FDR-corrected p-value ≤ 0.1; Table XIII.2). Twenty of these modules highly overlapped, sometimes differing for only one gene. Nicely, these modules together recapitulated the RTK/PI(3)K signaling and p38/JNK1 signaling which are interlinked through Akt. Top-scoring modules include core components of the RTK/PI(3)K cascade: EGFR, IGF1R, ERBB2, PIK3CA, PIK3R1, PTEN, and AKT1. We referred to this module, together with AKT3,

as the *core module* (Figure 6A). Besides RTK/PI(3)K modules, MEMo identified two modules including alterations at the apoptotic pathway. The first module includes AKT1, ATM, MDM4, MDM2, and TP53 (FDR-corrected p-value = 0.02); while the second includes ATM, CHEK2, MDM4, MDM2, and TP53 (FDR-corrected p-value = 0.1). After merging the two modules together, the mutual exclusivity is preserved and still significant (p-value = 0.002). This module highlighted a broader extent of p53 signaling inactivation that goes beyond TP53 mutations. Finally, we ran MEMo exclusively on Basal-like tumors. In Basal-tumors MEMo identified only one significant module including ATM, BRCA1, BRCA2, CCNE1, and RB1 (FDR-corrected p-value < 0.01). Surprisingly, the same module (with the exception of ATM) was also found as significant in serous ovarian cancer[50]. Again, deregulation of major cell-cycle checkpoints is reflected by high genomic instability.

Table XIII.2: Mutually exclusivity modules identified by MEMo.

Module ID	Genes	Total Percent Altered Cases	p-value	p*-value
M1	AKT1 [12], EGFR [15], ERBB2 [69], IGF1R [20], PIK3CA [101], PIK3R1 [13], PTEN [34],	50.11%	0	0.02
M2	AKT1 [12], EGFR [15], IGF1R [20], PAK1 [40], PIK3CA [101], PIK3R1 [13], PTEN [34],	45.36%	0	0.02
M3	AKT1 [12], BRCA1 [32], EGFR [15], PAK1 [40], PIK3CA [101], PIK3R1 [13], PTEN [34],	46.65%	0	0.02
M4	AKT1 [12], ERBB2 [69], IGF1R [20], PAK1 [40], PIK3CA [101], PIK3R1 [13], PTEN [34],	53.13%	0	0.02
M5	AKT1 [12], ERBB2 [69], IGF1R [20], MAP3K1 [39], PAK1 [40], PIK3CA [101], PIK3R1 [13],	52.27%	0	0.02
M6	AKT1 [12], CCND1 [80], EGFR [15], IGF1R [20], MAP2K4 [31], PIK3CA [101], PIK3R1 [13],	49.89%	0	0.02
M7	AKT1 [12], ERBB2 [69], MAP3K1 [39], MYC [71], PAK1 [40], PIK3CA [101],	58.53%	0	0.02
M8	AKT1 [12], BRCA1 [32], CCND1 [80], EGFR [15], PIK3CA [101], PIK3R1 [13],	48.81%	0	0.02
M9	AKT1 [12], BRCA1 [32], CCND1 [80], MYC [71], PIK3CA [101], PIK3R1 [13],	54.86%	0	0.02
M10	AKT1 [12], CCND1 [80], ERBB2 [69], MYC [71], PIK3CA [101],	58.10%	0	0.02
M11	AKT1 [12], ATM [19], MDM2 [18], MDM4 [26], TP53 [172],	49.03%	0	0.02
M12	AKT1 [12], BRCA1 [32], IKBKB [37], PIK3CA [101], PIK3R1 [13],	39.52%	0	0.02
M13	AKT1 [12], EGFR [15], MAP3K1 [39], MYC [71], PAK1 [40], PIK3CA [101], PIK3R1 [13],	52.92%	0.001	0.02
M14	AKT1 [12], EGFR [15], ERBB2 [69], MAP3K1 [39], MYC [71], PIK3CA [101], PIK3R1 [13],	56.16%	0.001	0.02
M15	AKT1 [12], CCND1 [80], EGFR [15], MAP2K4 [31], MYC [71], PIK3CA [101], PIK3R1 [13],	55.72%	0.001	0.02
M16	AKT1 [12], AKT3 [33], CCND1 [80], IGF1R [20], MAP2K4 [31],	35.42%	0.001	0.02
M17	AKT1 [12], EGFR [15], MAP2K4 [31], MAP3K1 [39], MYC [71], PAK1 [40], PIK3R1 [13],	41.68%	0.002	0.04
M18	AKT1 [12], CCND1 [80], EGFR [15], ERBB2 [69], IGF1R [20], PIK3CA [101],	52.92%	0.002	0.04
M19	BRCA1 [32], CCND1 [80], MDM2 [18], NBN [4], RB1 [23],	32.18%	0.002	0.04
M20	AKT1 [12], BRCA1 [32], MYC [71], PAK1 [40], PIK3CA [101], PIK3R1 [13],	49.24%	0.003	0.06
M21	AKT1 [12], EGFR [15], IGF1R [20], MAP2K4 [31], MAP3K1 [39], PAK1 [40], PIK3CA [101], PIK3R1 [13],	48.81%	0.005	0.08
M22	CCND1 [80], CDKN2A [16], MDM2 [18], MYB [7], RB1 [23],	29.81%	0.005	0.08
M23	ATM [19], CHEK2 [4], MDM2 [18], MDM4 [26], TP53 [172],	47.52%	0.006	0.1

XIV. Integrated Analysis and Interactive Exploration

To gain greater insight into the underlying system-level phenomena that characterize the development and progression of breast cancer, we have integrated all of the data types produced by TCGA and described in this paper into a single "feature matrix". From this single heterogeneous dataset, significant pairwise associations have been inferred using statistical analysis and can be visually explored in a genomic context using Regulome Explorer, an interactive web application (http://explorer.cancerregulome.org). In addition to associations that are inferred directly from the TCGA data, additional sources of information and tools are integrated into the visualization for more extensive exploration (e.g., literature-based associations, molecular interaction databases, miRBase, the UCSC Genome Browser, etc.). A few examples of the types of explorations that are possible are described below.

Exploring significant associations between molecular features

Filtering for significant associations between microRNA features and gene expression (mRNA) features in which there is a negative correlation relationship results in the circular view shown in Figure XIV.1A in which each arc indicates an association between a microRNA and a gene. Hovering over a single arc allows the user to see additional feature information, and clicking on

the arc produces a scatterplot of the underlying data. The majority of the arcs are due to microRNAs hsa-mir-17/18a at chromosome 13q31, hsa-mir-190b at 1q21, and hsa-mir-210 at 11p15. While microRNAs can affect distal genes, one would expect copy-number aberrations to primarily affect the expression of proximal genes, and this can be seen in Figure XIV.1B in which the most significant associations between copy-number features (orange) and gene expression features (blue) are nearly always proximal and are concentrated in hot spots on chromosome arms 1q (e.g., PARP1) and 5q (e.g., REEP5 and IL6ST) and chromosomes 8 (e.g., BRF2 and RAD21), 16 (e.g., CENPN) and 17 (e.g., ERBB2 and MIEN1). These hotspots all include corresponding GISTIC regions, as shown in supplemental Figure 8.

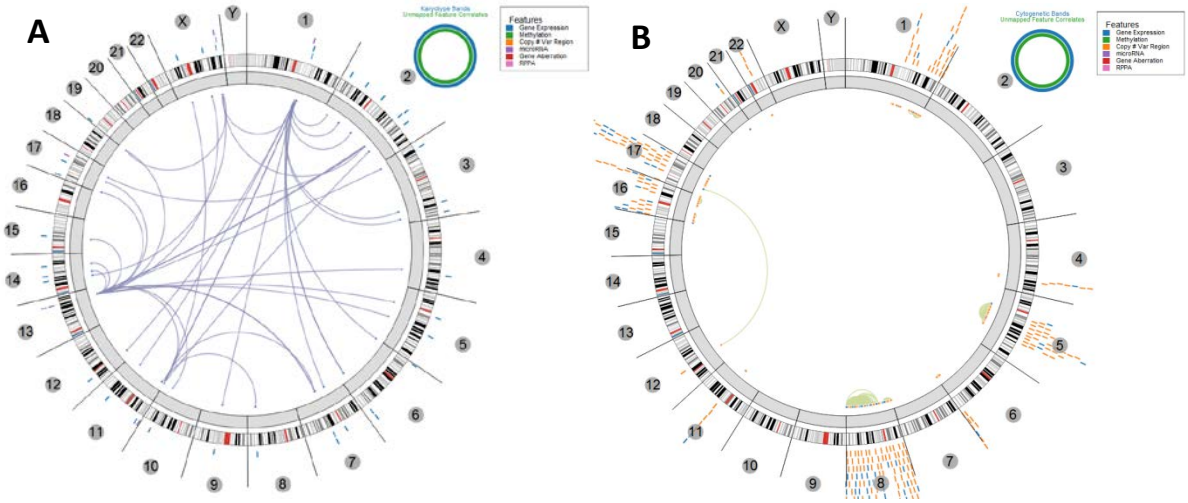


Figure XIV.1. Associations between molecular features. Statistically significant associations between features with genomic coordinates are indicated by arcs connecting pairs of dots which represent the features. Two examples are shown: significant associations between microRNA and mRNA expression levels (A), and between copy-number and mRNA expression (B).

Exploring significant associations with a subtype

The heterogeneous feature matrix also contains categorical variables including cluster assignments. Associations between molecular or other features and these categorical features can also be explored. In this case, statistically significant associations between molecular features (with genomic coordinates) and a categorical feature are shown as dots on a circular graph with a radial axis representing correlation coefficient or $\log_{10}$ (p-value). For example, CpG dinucleotides that are significantly differentially methylated in the hyper-methylated cluster shown in main Figure 2 can be identified by filtering for features associated with a binary feature that indicates membership in that cluster, as shown in Figure XIV.2A, or proteins that are significantly differentially expressed between the RPPA-inferred Reactive I and Reactive II groups shown in main Figure 3 can be identified as shown in Figure XIV.2B.

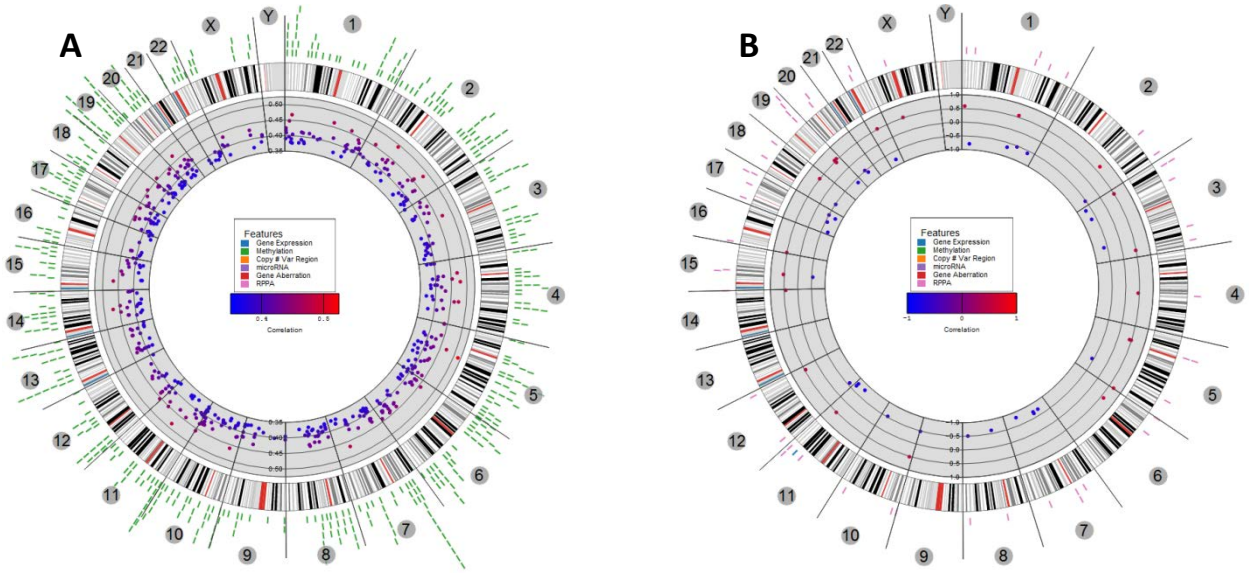


Figure XIV.2. Associations between molecular and categorical features. Methylation probes that are significantly hyper-methylated in methylation cluster 3 (A), and proteins that are significantly differentially expressed between the Reactive I and Reactive II groups (B).

Feature Matrix Construction

A feature matrix was constructed using all available clinical, sample, and molecular data for 832 unique patient/tumor samples. The clinical information includes features such as age, stage, ER/PR/HER2 status and histology; while the sample information includes features derived from molecular data such as the PAM50 subtype, single- and cross-platform cluster assignments and mutation rates. The molecular data includes mRNA and microRNA expression levels (Agilent and Illumina data respectively), protein levels (RPPA data), copy number alterations (derived from segmented Affymetrix SNP data as well as GISTIC regions of interest and arm-level values), DNA methylation levels (Illumina Infinium Methylation array), and germline and somatic mutations. For each mutated gene, several binary mutation features indicating the presence or absence of a mutation in each sample were generated, depending on the type and position of the mutations. Mutation types considered were synonymous, missense, nonsense and frameshift. Protein domains (InterPro) including any of these mutation types were annotated as such, with nonsense and frameshift annotations being propagated to all subsequent protein domains.

Pairwise Statistical Significance

The statistical significance of each pairwise association is assessed using rank-ordered data and a statistical test appropriate to each data type pair, e.g. Fisher's test (categorical-categorical), F-statistic (continuous-continuous) and ANOVA (continuous-categorical).

XV. HER2-positive analyses

Starting with the 466 freeze list, clinically HER2-positive tumors were selected. To identify genes that differ between 36 clinically HER2-positive HER2-enriched subtype and 32 clinically HER2-positive Luminal subtype, the samr package in R was used and 302 genes were found with a FDR of 0. SAM was also run on the RPPA data set of 29 clinically HER2-positive HER2-enriched subtype and 24 clinically HER2-positive Luminal subtype identifying 36 proteins with a FDR < 5. Correlation of subtypes with clinical features and mutations was done using a Pearson's Chi-Squared test for Fisher's Exact test in R. The SAM-derived gene and protein data sets were run with 10-fold cross validation to determine the smallest set of genes/proteins with the lowest cross-validation accuracy. Genes/proteins are listed in Supplemental Tables 6 and 7.

XVI. References

1 Hammond, M. E. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Arch Pathol Lab Med* **134**, 907-922, doi:10.1043/1543-2165-134.6.907 [pii] (2010).

2 Wolff, A. C. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Arch Pathol Lab Med* **131**, 18-43 (2007).

3 Persons, D. L., Borelli, K. A. & Hsu, P. H. Quantitation of HER-2/neu and c-myc gene amplification in breast carcinoma using fluorescence in situ hybridization. *Mod Pathol* **10**, 720-727 (1997).

4 Minot, D. M. *et al.* Image analysis of HER2 immunohistochemical staining. Reproducibility and concordance with fluorescence in situ hybridization of a laboratory-validated scoring technique. *Am J Clin Pathol* **137**, 270-276, doi:137/2/270 [pii] 10.1309/AJCP9MKNLHQNK2ZX.

5 Edge, S. B. *et al. AJCC Cancer Staging Manual*. Seventh Edition edn, (Springer-Verlag, 2010 (6th Printing)).

6 Dabbs, D. J. *et al.* High false-negative rate of HER2 quantitative reverse transcription polymerase chain reaction of the Oncotype DX test: an independent quality assurance study. *J Clin Oncol* **29**, 4279-4285, doi:JCO.2011.34.7963 [pii] 10.1200/JCO.2011.34.7963 (2011).

7 Staaf, J. *et al.* High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer. *Breast Cancer Res* **12**, R25, doi:bcr2568 [pii] 10.1186/bcr2568 (2010).

8 The Cancer Genome Atlas Research Network & Perou, C. M. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

9 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

10 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

11 Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:gr.129684.111 [pii] 10.1101/gr.129684.111 (2012).

12 Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311-317, doi:btr665 [pii] 10.1093/bioinformatics/btr665 (2012).

13 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:gr.107524.110 [pii] 10.1101/gr.107524.110 (2010).

14 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).

15 Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869-10874. (2001).

16 Hu, Z. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7**, 96, doi:1471-2164-7-96 [pii] 10.1186/1471-2164-7-96 (2006).

17 Perreard, L. *et al.* Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Res* **8**, R23 (2006).

18 Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418-8423 (2003).

19 Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. Statistical Significance of Clustering for High Dimensional Low Sample Size Data. *Journal of the American Statistical Association* **103** (2008).

20 Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J Clin Oncol*, doi:JCO.2008.18.1370 [pii] 10.1200/JCO.2008.18.1370 (2009).

21 Khattra, J. *et al.* Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines. *Genome Research* **17**, 108-116, doi:10.1101/gr.5488207 (2007).

22 de Hoon, M. J. L. *et al.* Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Research* **20**, 257-264, doi:10.1101/gr.095273.109.

23 Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367.

24 Campan, M., Weisenberger, D. J., Trinh, B., Laird, P. W. & Tost, J. Vol. 507 *Methods in Molecular Biology* 325-337 (Humana Press, 2009).

25 Cheung, L. W. *et al.* High Frequency of PIK3R1 and PIK3R2 Mutations in Endometrial Cancer Elucidates a Novel Mechanism for Regulation of PTEN Protein Stability. *Cancer discovery* **1**, 170-185, doi:10.1158/2159-8290.CD-11-0039 (2011).

26 Houseman, E. A. *et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9**, 365 (2008).

27 McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-1174 (2008).

28 Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-1260 (2008).

29 Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572 (2004).

30 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).

31 Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **101**, 4164-4169, doi:10.1073/pnas.0308531101 (2004).

32 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, doi:10.1038/nature10983 (2012).

33 Tibes, R. *et al.* Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular Cancer Therapeutics* **5**, 2512-2521, doi:10.1158/1535-7163.mct-06-0334 (2006).

34 Liang, J. *et al.* The energy sensing LKB1-AMPK pathway regulates p27kip1 phosphorylation mediating the decision to enter autophagy or apoptosis. *Nat Cell Biol* **9**, 218-224 (2007).

35 Hu, J. *et al.* Non-parametric quantification of protein lysate arrays. *Bioinformatics* **23**, 1986-1994, doi:10.1093/bioinformatics/btm283 (2007).

36 Hennessy, B. T. *et al.* Pharmacodynamic Markers of Perifosine Efficacy. *Clinical Cancer Research* **13**, 7421-7431, doi:10.1158/1078-0432.ccr-07-0760 (2007).

37 Coombes, K. *et al.* SuperCurve: SuperCurve Package. R package version 1.4.1. (2011).

38 Gonzalez-Angulo, A. *et al.* Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. *Clinical proteomics* **8**, 11.

39 Hennessy, B. *et al.* A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected Human Breast Cancers. *Clinical proteomics* **6**, 129-151, doi:10.1007/s12014-010-9055-y.

40 Monti, S., Tamayo, P., Mesirov, J. & Golub, T. R. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52**, 91-118 (2003).

41 Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573, doi:10.1093/bioinformatics/btq170 (2010).

42 Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237-245, doi:10.1093/bioinformatics/btq182 (2010).

43 Hosack, D. A., Dennis, G., Jr., Sherman, B. T., Lane, H. C. & Lempicki, R. A. Identifying biological themes within lists of genes with EASE. *Genome Biol* **4**, R70 (2003).

44 Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

45 Gonzalez-Perez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* **88**, 440-449, doi:10.1016/j.ajhg.2011.03.004 (2011).

46 Hall, M. *et al.* The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **11** (2009).

47 Troester, M. A. *et al.* Gene expression patterns associated with p53 status in breast cancer. *BMC Cancer* **6**, 276 (2006).

48 Herschkowitz, J. I., He, X., Fan, C. & Perou, C. M. The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. *Breast Cancer Res* **10**, R75, doi:bcr2142 [pii] 10.1186/bcr2142 (2008).

49 Creighton, C. J. *et al.* Proteomic and transcriptomic profiling reveals a link between the PI3K pathway and lower estrogen-receptor (ER) levels and activity in ER+ breast cancer. *Breast Cancer Res* **12**, R40, doi:10.1186/bcr2594 (2010).

50 Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* **22**, 398-406, doi:10.1101/gr.125567.111 (2012).