

A Gene Regulatory Program in Human Breast Cancer

Renhua Li,^{*,†,1} John Campos,^{*} and Joji Iida^{*}^{*}Windber Research Institute, Windber, Pennsylvania 15963, and [†]The Jackson Laboratory, Bar Harbor, Maine 04609

ORCID ID: 0000-0003-4676-3524 (R.L.)

ABSTRACT Molecular heterogeneity in human breast cancer has challenged diagnosis, prognosis, and clinical treatment. It is well known that molecular subtypes of breast tumors are associated with significant differences in prognosis and survival. Assuming that the differences are attributed to subtype-specific pathways, we then suspect that there might be gene regulatory mechanisms that modulate the behavior of the pathways and their interactions. In this study, we proposed an integrated methodology, including machine learning and information theory, to explore the mechanisms. Using existing data from three large cohorts of human breast cancer populations, we have identified an ensemble of 16 master regulator genes (or MR16) that can discriminate breast tumor samples into four major subtypes. Evidence from gene expression across the three cohorts has consistently indicated that the MR16 can be divided into two groups that demonstrate subtype-specific gene expression patterns. For example, group 1 MRs, including *ESR1*, *FOXA1*, and *GATA3*, are overexpressed in luminal A and luminal B subtypes, but lowly expressed in HER2-enriched and basal-like subtypes. In contrast, group 2 MRs, including *FOXM1*, *EZH2*, *MYBL2*, and *ZNF695*, display an opposite pattern. Furthermore, evidence from mutual information modeling has congruently indicated that the two groups of MRs either up- or down-regulate cancer driver-related genes in opposite directions. Furthermore, integration of somatic mutations with pathway changes leads to identification of canonical genomic alternations in a subtype-specific fashion. Taken together, these studies have implicated a gene regulatory program for breast tumor progression.

KEYWORDS breast cancer; master regulator; regulator–regulon interactions; gene regulatory program

HUMAN breast cancer is the most common malignancy in women, with >200,000 new cases diagnosed each year in the United States (Tolaney and Winer 2007). Breast cancer is a complex disease that is caused by multiple genetic and environmental factors. Molecular heterogeneity in breast tumors has challenged diagnosis, prognosis, and clinical treatment.

Human breast cancer has significant intra- and intertumor molecular heterogeneity. Regarding intratumor heterogeneity, evidence from next-generation sequencing has indicated that different tumor subclones display diversified DNA mutation profiles (Tolaney and Winer 2007; Nik-Zainal *et al.* 2012; Yates *et al.* 2015). Furthermore, the mutation profiles are subject to change over time due to the fact that tumor cells can adapt to the selective pressure of therapies (Klein 2013).

On the other hand, intertumor variation is manifested by molecular subtypes that represent significant differences in prognosis and survival (Parker *et al.* 2009). Evidence from gene expression profiling and unsupervised clustering analysis has indicated four major subtypes: luminal A, luminal B, HER2-enriched, and basal-like (Perou *et al.* 2000). While the majority of the luminal tumors are estrogen receptor (ER)-positive, the other two subtypes, especially the basal-like, are mainly ER-negative tumors. A set of 50 genes (PAM50) has been proposed to classify breast tumor samples into the subtypes (Parker *et al.* 2009). Six (*ESR1*, *PGR*, *FOXA1*, *FOXC1*, *MYC*, and *MYBL2*) of the 50 genes are transcriptional factor genes (Parker *et al.* 2009). Recent studies on genomics and transcriptomics of large patient populations have identified additional subtypes (Curtis *et al.* 2012; Guedj *et al.* 2012). In the present study, we will focus on the four major subtypes.

High-throughput technologies, including SNP array and next-generation sequencing studies on large cohorts of patient and control populations, have helped identify both common and rare DNA alternations impacting cancer etiology and development (Cancer Genome Atlas Network 2012; Michailidou *et al.* 2015). Genome-wide association studies

Copyright © 2015 by the Genetics Society of America

doi: 10.1534/genetics.115.180125

Manuscript received July 2, 2015; accepted for publication October 20, 2015; published Early Online October 26, 2015.

Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.180125/-/DC1.

¹Corresponding author: 12101 Village Square, Rockville, MD 20852.

E-mail: renhua.li10@gmail.com

have accumulatively identified ~90 loci that are associated with predisposition to breast cancer risk (Michailidou *et al.* 2015). The germline mutations collectively account for 16% of the genetic variation in breast cancer (Michailidou *et al.* 2015). On the other hand, somatic mutations have been identified in tens of breast cancer driver genes, which lead to protein functional changes in tumors (Cancer Genome Atlas Network 2012). However, connections of both the genetic and genomic alternations to the subtypes are largely unknown.

Assuming that there are intrinsic connections between various biological pathways and different tumor subtypes that represent significant differences in prognosis and survival, we suspect that there might be gene regulatory mechanisms that modulate the behavior of the pathways. Thus we hypothesize that identification of master regulators (MRs) and gene regulatory pathways can bridge the gap between genotype and phenotype and shed light on tumor progression.

MRs are transcriptional factor genes that play a pivotal role in modulating downstream pathways or gene networks. Studies have established that ER is such a MR that regulates multiple pathways related to ER-positive breast tumors (Fletcher *et al.* 2013). However, a complete set of MRs for each tumor subtype, as well as the MR–MR interactions, are little known. Therefore, it is crucial to perform a systematic search for master regulators.

Systems genetics that integrates approaches of genetics, genomics, and multi-level phenotype characterization is powerful in understanding genotype–phenotype dependency (Li *et al.* 2006; Bjorkegren *et al.* 2015). Gene–gene interactions are the core of systems genetics (Li *et al.* 2006; Li and Churchill 2010). But they are challenging to explore in human populations. In this study, using existing data from three large cohorts of human breast cancer populations, including *The Cancer Genome Atlas* (TCGA) breast tumor cohort (Cancer Genome Atlas Network 2012), the Curtis cohort (Curtis *et al.* 2012), and the Guedj cohort (Guedj *et al.* 2012), we systematically scrutinize MR–MR and MR–regulon interactions. Our studies have revealed a gene regulatory program that impacts tumor progression in human breast cancer.

Materials and Methods

Three cohorts of human breast cancer populations

In this study, we used existing data generated from three cohorts of human breast cancer populations to identify and cross-validate MRs and gene regulatory pathways. These data include the TCGA breast tumor cohort (Cancer Genome Atlas Network 2012), the Curtis cohort (Curtis *et al.* 2012), and the Guedj cohort (Guedj *et al.* 2012). The TCGA breast tumor samples represent a large patient population from the United States, with 90% white women and 10% African American women. In addition to the DNA-Seq data and clinical data, we downloaded RNA-Seq data of 646 primary breast tumor cases (samples) from the TCGA data portal (<https://tcga-data.nci.nih.gov/>). For cross-validation studies, we also

downloaded the microarray-based gene expression data and clinical data from the Curtis cohort (Curtis *et al.* 2012) and from the Guedj cohort (Guedj *et al.* 2012). The Curtis cohort encompasses ~2000 breast tumor samples from the United Kingdom and Canada (Curtis *et al.* 2012). The Guedj cohort has 537 breast tumor samples from France (Guedj *et al.* 2012). Among the three cohorts, the breast cancer cases are all females. Regarding age at diagnosis, it is comparable between the TCGA and Curtis cohorts, with a median age of ~60 (Supporting Information, Figure S1). However, in the Guedj cohort the median age shifts to ~30. This may partially be explained by the different clinical record systems because age at initial diagnosis was recorded in the Guedj cohort. Regarding pathological phenotypes, such as tumor stage and grade, it is challenging to compare because different pathological systems were used across the cohorts.

Regarding the gene expression data, we processed the data, followed by a data quality check. The TCGA RNA-Seq data downloaded are already processed data (level 3), in which gene expression is quantified as fragments per kilobase transcript per million mapped reads (FPKM). Using the Mbatch tool (<http://bioinformatics.mdanderson.org/tcgambatch/>), we checked the data for batch effect. The gene expression data from the Curtis and Guedj cohorts are probe-level measurements based on the microarray platforms. Using corresponding human annotation files downloaded from the Bioconductor (<http://bioconductor.org>), we performed probe-to-gene annotations. Gene-level expression was then quantified as the maximum measurement across the corresponding probes. We also applied the Mbatch tool to check data quality. In addition, the gene expression data were transformed using van der Waerden scores (Lehmann and D’Abrera 1988) to center the mean of each gene to 0 and compare only variance and covariance between genes.

Machine learning to select an assemble of master regulator genes

Supervised machine learning is a powerful method for gene selection and tumor sample classification. Schematic framework of machine learning is shown in Figure S2. Since RNA-Seq provides a better measurement of gene expression compared to microarray, we used the downloaded TCGA RNA-Seq data for feature (or gene) selection. But findings will be validated in the other two cohorts mentioned above. We split the TCGA primary breast tumor samples ($n = 646$) into training (3/4) and test (1/4) subsets. Subtype of each sample in the training subset was assigned by use of the PAM50 method (Parker *et al.* 2009). The subtype information was kindly provided by K. A. Horsley at the University of North Carolina.

Since the first step is to systematically identify MRs for each subtype, we focused on two gene sets for feature selection: (1) a set of ~1400 well-annotated transcriptional factor (TF) genes (Vaquerizas *et al.* 2009) and (2) another set of 138 cancer driver genes (Vogelstein *et al.* 2013). There are only 23 overlapping genes between the two gene sets. As the majority of the cancer driver genes are non-TF genes, the second gene set

is used as a control for feature selection. Between the combined gene panel (from gene sets 1 and 2) and the PAM50, only 10 genes overlap, including *FOXC1*, *ESR1*, *FOXA1*, *ERBB2*, *MYC*, *MDM2*, *PGR*, *BCL2*, *EGFR*, and *MYBL2*.

Regarding feature (or gene) selection, we employed the Random Forest algorithm (Breiman 2001) coupled with recursive gene elimination. Random Forest is one of the state-of-the-art algorithms in supervised machine learning. It is a decision-tree-based method for selection of an ensemble of features that are able to best discriminate the four subtypes of breast tumors. See [SI Methods](#) for more details.

Once the feature set is identified by the algorithm, we then train a classifier (or model), using the training data. There are multiple algorithms that can be applied to train a classifier. For example, one algorithm develops a probabilistic model that is based on the Bayesian method (Berger 1985). Application of the classifier to a new tumor sample will generate the posterior probability that the sample belongs to a subtype, given the data and the model. We used this method to train a classifier.

We need a reference for prediction assessment. Using the PAM50 (Parker *et al.* 2009) method, we called subtypes for the samples in each of the test data sets ([Figure S2](#)). Prediction (or classification) accuracy is defined as the percentage of the samples (in each of the test data sets) that have been correctly predicted by the MR16 classifier. This is a measurement of concordance in subtype calling between the PAM50 and the MR16.

Disease-free survival prediction

Disease-free survival prediction of a classifier is of clinical interest. Since the Guedj cohort has a long follow-up time in metastasis relapse-free survival (Guedj *et al.* 2012), we focus on this cohort for survival prediction. Using the MR16 from the TCGA training data, we developed a classifier. Application of the classifier to the Guedj cohort can predict the samples into the four major subtypes. We then fit a COX proportional hazards regression model (Cox and Oakes 1984), where the disease-free survival time was used as the response variable, the predicted subtype information as an explanatory variable, plus tumor grade as a covariate. Regarding other cofactors, the effect of age at diagnosis is captured by the subtypes, as the basal-like tumors are frequently detected in young women. Other pathological traits, such as tumor stage and size, are correlated with tumor grade.

In contrast, we used the PAM50 method (Parker *et al.* 2009) to directly call subtypes for the same samples from the Guedj cohort (Guedj *et al.* 2012). The PAM50 method often predicts five subtypes, including a normal-like subtype in addition to the four major subtypes. We then fit a similar model for disease-free survival prediction.

Mutual information for identification of target genes

Mutual information is based on the concept of entropy (Shannon 1948). When a population size is large, mutual information is able to capture the nonlinear regulator–regulon interactions (Basso *et al.* 2005; Carro *et al.* 2010). See [SI Methods](#) for

more details. To stringently control false-positive regulator–regulon interactions, we need to take several filtering steps: (1) permutation tests to adjust for multiple hypothesis testing; (2) bootstrap resampling to obtain a consensus network; and (3) data-processing inequality analysis to identify the most likely paths of information flow (Margolin *et al.* 2006). Detailed methods for the steps were reported previously (Margolin *et al.* 2006).

Identification of canonical pathway changes in tumors

To associate DNA somatic mutations with gene regulatory pathway changes in tumor subtypes, we focus on the connection between subtype-enriched somatic mutations and subtype-specific pathway amplification, comparing corresponding gene expression between the TCGA tumor samples and the matched normal samples.

Data Available

Data are available at <https://tcga-data.nci.nih.gov/>.

Results

Identification and validation of master regulator genes

Since genotype–phenotype association is complex in humans, characterization of gene regulatory pathways may bridge the gap between genotype and phenotype.

Using the supervised machine learning algorithm as described in *Materials and Methods*, we have identified an ensemble of 16 genes ([Figure 1A](#)). Although the genes are ordered by their ranks in terms of importance ([Figure S3](#)), they are an entity that gives rise to the maximum accuracy in classification of the training samples ([Figure S4](#)). Furthermore, the 16 genes are among the top ones that are most frequently selected in the repeated computational experiments ([Figure S5](#)). Since the gene set plays a significant role in regulating cancer-related genes (will be addressed below), we refer to the 16 genes as master regulators or MR16.

Using the TCGA data, we then trained a classifier based on the MR16. Application of the classifier to the Curtis (Curtis *et al.* 2012) and the Guedj (Guedj *et al.* 2012) data indicated that the luminal A and luminal B samples can be classified with concordance rates of 87 and 85%, respectively, compared to the direct calling by the PAM50 method (Parker *et al.* 2009). Similarly, HER2-enriched and basal-like tumors can be classified with concordance rates of 90 and 93%, respectively. These intrigued us to predict disease-free survival.

Is the MR16 classifier able to predict disease-free survival in the independent cohorts of breast cancer populations? Since the Guedj cohort has a long follow-up time for metastasis relapse-free survival (Guedj *et al.* 2012), we focus on this cohort for survival prediction. Corresponding Kaplan–Meier curves are shown in [Figure 2A](#). In contrast, we used the PAM50 method (Parker *et al.* 2009) to directly call subtypes for the same samples, and corresponding Kaplan–Meier curves are shown in [Figure 2B](#). Comparisons of the two sets

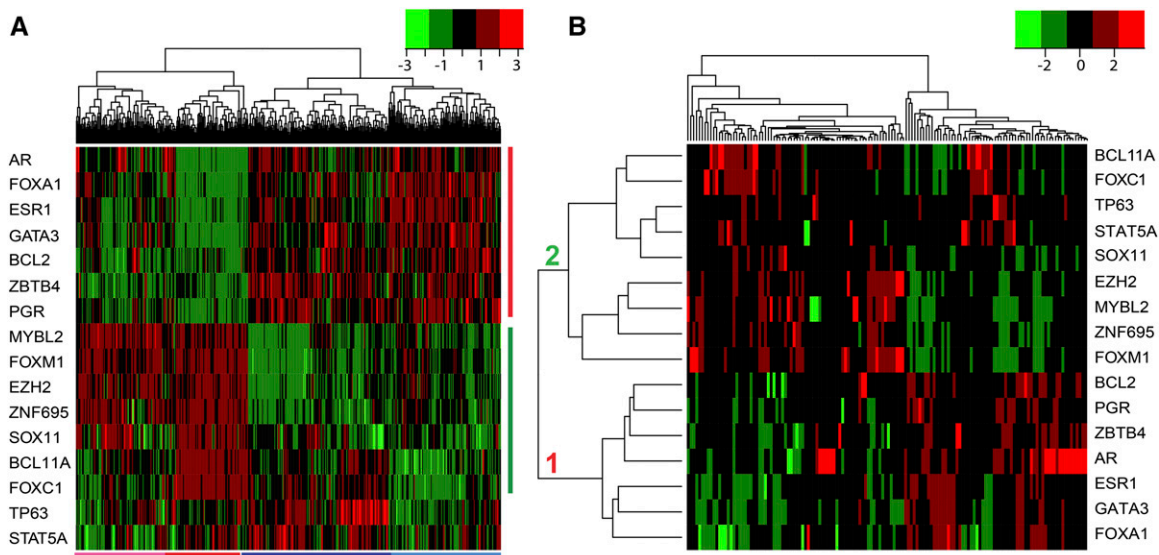


Figure 1 Patterns of MR gene expression and MR–cancer driver gene interactions in the TCGA breast tumor samples. (A) Heatmap of the MR16 gene expression across the TCGA primary tumor samples. Using the RNA-Seq data and a machine-learning algorithm, we identified the MR16 that can classify the tumor samples into two large clusters: ER-negative and ER-positive. Two subclusters are observed within each of the clusters. The four clusters correspond well to the four tumor subtypes: HER2-enriched (pink bar at the bottom), basal-like (red), luminal A (dark blue), and B (light blue). Vertical red and green bars represent the two groups of the MRs that display distinct expression patterns across the samples. Green and red colors in the heatmap indicate low and high gene expression, respectively. (B) Heatmap of the MR16–cancer driver-related gene interactions. Using the TCGA RNA-Seq data, we computed a mutual information matrix between the MR16 and the cancer driver-related genes. The MR–regulon interaction patterns indicate that two MR groups either up- or down-regulate (in red and green colors, respectively) the cancer driver-related genes in opposite directions.

of Kaplan–Meier curves indicate that the four major subtypes share similar disease-free survival, indicating that the MR16 has a cross-cohort prediction power similar to the direct prediction by the PAM50. These prediction results are consistent with clinical observations that luminal A tumors display a significantly ($\log \text{rank } P = 6.25 \times 10^{-5}$) better prognosis, compared to the other subtypes.

Gene expression patterns of the MR16

Gene expression of the MR16 can accurately separate the TCGA primary breast tumor samples into four clusters that correspond to the four major subtypes (Figure 1A). The two large clusters represent luminal (or ER-positive) and nonluminal (or ER-negative) tumors, respectively, indicating that ER status is the major separating factor between the tumor samples. Within each of the main clusters, two subclusters have been classified. In the ER-positive cluster, luminal A and B tumors can be separated by the expression pattern of a combination of the genes, including *BCL11A*, *FOXC1*, *TP63*, and *STAT5A* (Figure 1A). In the ER-negative cluster, basal-like tumors differ from HER2-enriched tumors mainly by *BCL11A* and *FOXC1* gene expression (Figure 1A). These four clusters correspond well to the four breast tumor subtypes: luminal A, luminal B, HER2-enriched, and basal-like.

Expression of the 16 genes across the TCGA samples has displayed distinct patterns (Figure 1A; Figure S3). In general, the MR16 can be divided into two groups. Group 1 genes, including *ESR1*, *FOXA1*, *GATA3*, *PGR*, *AR*, *BCL2*, and *ZBTB4*, are highly expressed in luminal tumors, but lowly expressed

in nonluminal tumors (Figure 1A). Conversely, group 2 genes, including *FOXM1*, *EZH2*, *ZNF695*, *MYBL2*, *SOX11*, *BCL11A*, and *FOXC1*, are highly expressed in the nonluminal, especially the basal-like subtype, but lowly expressed in the luminal tumors (Figure 1A).

As expected, similar gene expression patterns have been observed in the Curtis and Guedj cohorts (Figure S6). The same group 1 MRs as defined in the TCGA cohort (Figure 1A) are overexpressed in the luminal tumors, but lowly expressed in the nonluminal tumors. In contrast, the same group 2 MRs are overexpressed in the nonluminal tumors, but lowly expressed in the luminal tumors. The samples in the Curtis validation data display two large clusters that correspond to the luminal and nonluminal tumors (Figure S6A). In each of the two clusters, there are two subclusters. In general, the four clusters correspond to the four major subtypes. The samples in the Guedj cohort also exhibit two large clusters that correspond to the luminal and nonluminal tumors (Figure S6B). In each of the two clusters, the subclusters are more complicated, compared to the TCGA and the Curtis cohorts. This may suggest that there exist more subtypes in the Guedj cohort. The consistent gene expression patterns of the MR16 across different cohorts of breast cancer populations call for further investigation.

Interaction patterns between the MR16 and cancer driver-related genes

What are the relationships between the MR16 and the cancer driver genes? We can approach this question either by

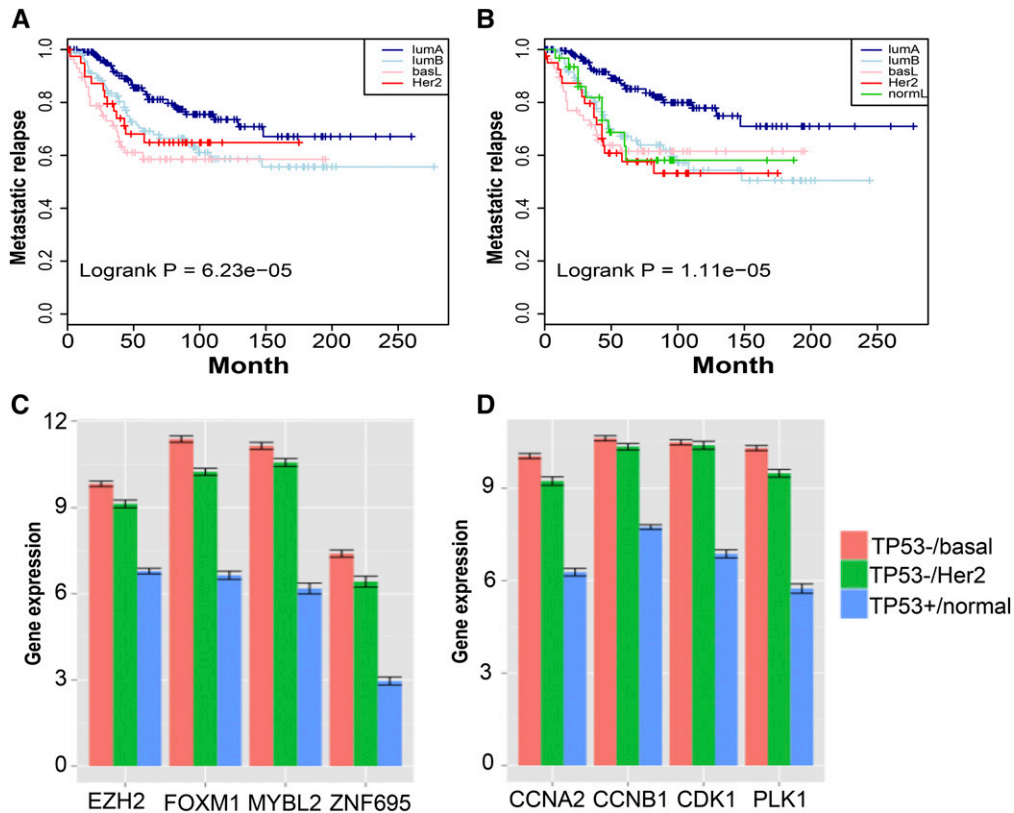


Figure 2 Prediction of disease-free survival by the MR16 and cell cycle pathway changes in ER-negative breast tumors. (A and B) Kaplan–Meier curves of metastatic relapse-free survival. (A) Using the MR16 and TCGA RNA-Seq data, we trained a classifier that was then applied to the Guedj data (Guedj *et al.* 2012) for subtype classification. We then fit a COX proportional hazards regression model (Cox and Oakes 1984) with tumor grade as a covariate. Kaplan–Meier curves of disease-free survival are shown by subtype. (B) Using the same set of samples from the Guedj data (Guedj *et al.* 2012), we applied the PAM50 method (Parker *et al.* 2009) to directly call subtypes, which resulted in five subtypes including a normal-like subtype. Corresponding Kaplan–Meier curves are plotted. Comparisons of the two sets of Kaplan–Meier curves indicate that luminal A tumors are predicted to have better prognosis. (C and D) Association of *TP53* somatic mutations with significant overexpression of cell cycle pathways in ER-negative

tumors. Using matched tumor samples with *TP53* mutations and matched normal samples without *TP53* mutations, we compare the expression of representative MRs and cell cycle genes. The y-axis is the averaged gene expression by RNA-Seq in log₂_FPKM. Standard errors are plotted on top of the bars. Red, basal-like tumor samples; green, HER2-enriched tumors; blue, normal samples.

performing either simple correlation analysis, which is based on the linearity assumption, or by providing mutual information that is a generalized correlation without such an assumption. Results from Pearson correlation analysis, based on a subset of the MR16 from the TCGA RNA-Seq data, are shown in Figure S7. It appears that there exist moderate positive correlations for the intragroup MRs. Regarding the intergroup MRs, however, there are weak or no correlations between them. In luminal B and basal-like subtypes, the correlation coefficients between the intergroup MRs are negative.

Since the sample sizes are large in the three cohorts of patient populations, we then applied mutual information to model the interactions between the MR16 and their target genes (regulons). The gene panel for mutual information modeling includes (1) the MR16; (2) the 138 cancer driver genes (Vogelstein *et al.* 2013); and (3) additional cell cycle genes that are known regulons of *FOXM1*—*PLK1*, *CCNB1*, *CCNB2*, *CDK1*, *CENPF*, and *UBE2C* (Chen *et al.* 2013). Only four genes (*i.e.*, *AR*, *EZH2*, *GATA3*, and *BCL2*) overlap between gene sets 1 and 2. The joint gene sets 2 and 3 are referred to as cancer driver-related genes.

Evidence from mutual information modeling indicates that interactions of the MR16 and the cancer driver-related genes display distinct patterns (Figure 1B). The majority of the cancer driver-related genes are found to be either up- or down-regulated by the MR16. This implicates subtype-specific

gene regulations in breast tumors, which will be scrutinized below.

Two groups of the MR16 agonistically regulate cell cycle genes

Tumor growth and progression is a hallmark for aggressiveness. The four main breast tumor subtypes have significant differences in tumor prognosis and survival (Parker *et al.* 2009). Since cell cycle genes play a significant role in regulating tumor growth and progression, we will focus on the interactions of the MR16 and the cell cycle genes.

Examination of the MR–cell cycle gene interactions across the four subtypes gave rise to noting distinct patterns. Many cell cycle genes, including *PLK1* and *CDK1*, are down-regulated by the group 1 MRs, but up-regulated by the group 2 MRs, especially *FOXM1*, *MYBL2*, *EZH2*, and *ZNF695* (Figure 3). It is noteworthy that the two group of MRs are clearly defined in the TCGA cohort and validated in both the Curtis and the Guedj cohorts (Figure 1A; Figure S6), based on the gene expression patterns instead of the MR–regulon interaction patterns.

We then scrutinized subtype-specific networks in the TCGA cohort. The MR–cell cycle gene interaction patterns are consistent with that observed across the subtypes (Figure S8). Furthermore, the group 2 MRs, including *FOXM1*, *MYBL2*, *EZH2*, and *ZNF695*, cooperatively up-regulate the cell cycle

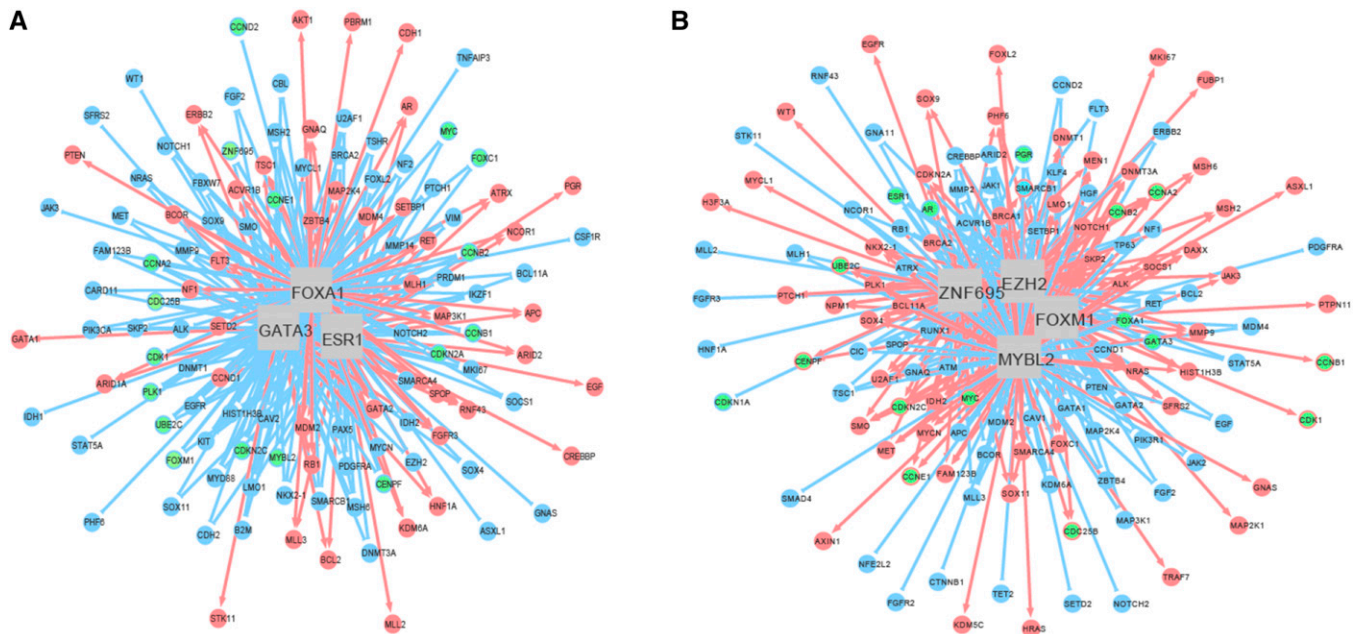


Figure 3 Cross-subtype gene regulatory networks. The networks are derived from the TCGA RNA-Seq data with 646 primary breast tumor samples across the four major subtypes. The nodes in the circle represent target genes (regulons), and the edges connecting the MRs and the target genes indicate interactions between them. Red and blue colors denote up- and down-regulation, respectively. (A) A gene regulatory network illustrating the interactions between three representative group 1 MRs and regulons. Green nodes are either highlighted cell cycle genes or group 2 MRs. Other group 1 MRs, such as *PGR*, are up-regulated by the central regulators. However, group 2 MRs, such as *FOXM1* and *ZNF695*, and cell cycle genes, such as *CCNB1* and *CDK1*, are down-regulated by the central MRs. (B) A gene regulatory network illustrating the interactions between four group 2 MRs and regulons. Green nodes are highlighted for either cell cycle genes or group 1 MRs. Other group 2 MRs, such as *SOX11* and the cell cycle genes, are up-regulated by the central MRs. However, group 1 MRs, such as *ESR1* and *AR*, are down-regulated by the central MRs.

genes in the basal-like tumors. Similar results are observed in the HER2-enriched tumors (data not shown). On the contrary, the group 1 MRs, such as *ESR1*, *FOXA1*, and *GATA3*, cooperatively down-regulate the cell cycle genes in the luminal tumors. These examples clearly demonstrate that MR–cell cycle gene interactions behave in a subtype-specific fashion.

Although the cell cycle genes are known target genes for *FOXM1* and *MYBL2* (Chen *et al.* 2013), the emergence of at least two other partners (*EZH2* and *ZNF695*) in regulating the cell cycle genes is new.

Validation of the regulator–regulon interactions

We then used the Curtis and the Guedj cohorts to validate the MR–cell cycle gene interactions. We focus on validation of the regulator–regulon interaction patterns, for example, the interactions between the two groups of MRs and the cell cycle genes. We are not testing a specific regulator–regulon interaction, since the different population sizes can affect the filtering steps, which may result in a discrepant report of the interaction.

Using the same methods, we modeled the MR–regulon interactions in both the Curtis and the Guedj cohorts. As expected, the MR–cell cycle gene interaction patterns are congruent; that is, the group 1 MRs down-regulate expression of the cell cycle genes in ER-positive tumors; in contrast, the group 2 MRs up-regulate expression of the genes in ER-negative tumors (Figure S9 and Figure S10). The subtype-specific interactions

imply that the cell cycle pathways play a critical role in promoting progression of ER-negative tumors.

Association of DNA somatic mutations with gene regulatory pathways

Since breast cancer is a heterogeneous disease in which various subclones from a tumor can exhibit different mutation profiles (Cancer Genome Atlas Network 2012), identification of canonical pathway changes in each subtype has therapeutic implications.

It is noteworthy that DNA somatic mutations in cancer driver genes are unevenly distributed across tumor subtypes (Cancer Genome Atlas Network 2012). For example, *TP53* is one of the top three genes (*TP53*, *PIK3CA*, and *GATA3*) that are most frequently mutated in breast cancer (Cancer Genome Atlas Network 2012). Regarding mutation types, 60% of the *TP53* mutations are missense mutations that lead to dysfunctional proteins (Table S1). Interestingly, the *TP53* mutations are overwhelmingly enriched in HER2-enriched and basal-like subtypes (Cancer Genome Atlas Network 2012).

Since the *TP53* gene is known to play a key role in regulating cell cycle arrest and apoptosis (Amundson *et al.* 1998), we then managed to associate the subtype-specific *TP53* somatic mutations with the overexpression of the cell cycle pathways in the ER-negative breast tumors. Using breast tumor samples with *TP53* mutations and matched normal samples without *TP53* mutations, we examined expression of

representative group 2 MRs and the cell cycle genes as well. Results from TCGA RNA-Seq data indicate that the four group 2 MRs (*FOXM1*, *MYBL2*, *EZH2*, and *ZNF695*) are significantly ($P = 0.002$) overexpressed in both HER2-enriched and basal-like subtypes, compared to the normal samples (Figure 2C). Similarly, representative cell cycle genes are significantly overexpressed in the two subtypes (Figure 2D). In contrast, these are not observed in luminal A and B subtypes (data not shown). These clearly indicate the association of *TP53* somatic mutations with up-regulated cell cycle pathways in ER-negative tumors. Furthermore, the association reflects canonical genomic alternations in the ER-negative breast tumors.

Discussion

We have identified an ensemble of 16 master regulators as an entity that has the maximum prediction power in the training data. In terms of subtype classification and survival prediction, the MR16 is comparable to the PAM50 (Parker *et al.* 2009). Compared to the PAM50, the MR16 has only six overlapping genes (*ESR1*, *PGR*, *FOXA1*, *FOXC1*, *BCL2*, and *MYBL2*). Among the top four genes (*i.e.*, *ESR1*, *GATA3*, *FOXM1*, and *EZH2*) in the MR16, three (*GATA3*, *FOXM1*, and *EZH2*) are nonoverlapping genes. The *GATA3* gene is the key regulator of luminal progenitor cell differentiation (Carr *et al.* 2012). Previous studies have indicated that *FOXM1* is a key regulator of cell proliferation and is overexpressed in many cancer types including breast cancer (Kwok *et al.* 2010). In the present study, *FOXM1* and *EZH2* are critical genes that play a significant role in regulation of the cell cycle gene expression in ER-negative tumors. However, the three genes are missed in the PAM50. Furthermore, results from mutual information modeling indicate that the majority of the PAM50 genes are predicted to be target genes of the MR16 (Figure S11).

Identification of gene regulatory pathways is critical for mechanistic understanding of genotype–phenotype dependency. The molecular subtypes in human breast cancer are associated with significant differences in prognosis and survival (Parker *et al.* 2009). Our studies have indicated that there exist intrinsic connections between specific molecular pathways and the subtypes. Furthermore, the pathways are regulated by the two groups of the MR16. Therefore, gene regulatory mechanisms play a significant role in shaping the subtypes.

Statistical evidence from information theory, based on three large cohorts of breast cancer populations, has implicated a gene regulatory program related to breast tumor progression. The program is characteristic of two types of gene–gene interactions: MR–MR and MR–regulon interactions.

Regarding MR–MR interactions, the intragroup MRs interact in a hierarchical and synergistic fashion. Recent studies using co-immunoprecipitation and CHIP-Seq technologies have indicated that group 1 MRs, including *FOXA1*, *ESR1*, and *GATA3*, are involved in the same protein complex (Theodorou *et al.* 2013). However, interactions of the four main group 2

MRs (*FOXM1*, *MYBL2*, *EZH2*, and *ZNF695*) are less known, although *FOXM1* and *MYBL2* are known to cooperatively bind to the promoter regions of the cell cycle genes (Wang and Gartel 2011; Chen *et al.* 2013). *EZH2* is a member of the polycomb repressive complex 2 (PRC2) that methylates lysine 27 of histone H3 (H3K27). The canonical function of *EZH2* is repression of tumor suppressor genes through H3K27me3 (Yamaguchi and Hung 2014). On the other hand, accumulating evidence has indicated that *EZH2* is able to activate target genes by directly binding to the regulatory regions (Lee *et al.* 2011; Gonzalez *et al.* 2014). *ZNF695* is a gene with little function known to date. Interestingly, the MR16 are all TF genes, although 115 non-TF cancer driver genes (Vogelstein *et al.* 2013), including *HER2* (or *ERBB2*) and *PTEN*, were incorporated in the gene panel for feature selection.

Contrary to the synergistic intragroup MR–MR interactions, the intergroup MR–MR interactions operate in an agnostic fashion. An interesting question is, “What is the biological mechanism underlying the inhibition between the two groups of MRs?” A recent study by Carr *et al.* (2012) indicated that *FOXM1* represses *GATA3* gene expression in the mammary gland by recruiting the methyltransferase DNMT3b to the binding sites within the *GATA3* promoter, thereby leading to methylation-induced gene silencing and the undifferentiated state of luminal progenitors. More in-depth studies are needed to elucidate the intra- and intergroup MR–MR interactions.

Regarding MR–regulon interactions, the two groups of MRs either up- or down-regulate different subsets of cancer driver-related genes in a subtype-specific fashion. This may help us understand pathway interactions in each tumor subtype. For example, the group 2 MRs are highly expressed in HER2-enriched and basal-like tumors. Correspondingly, the cell cycle genes are cooperatively up-regulated by the four major MRs in group 2. These two lines of evidence indicate that cell cycle pathways are overexpressed in the two subtypes, but not in the luminal A and B subtypes. Furthermore, the subtype-specific amplification of the cell cycle pathways coincides with the subtype-specific *TP53* mutations. Therefore, it is suggested that *TP53* mutations and corresponding cell cycle pathway amplification represent canonical genomic alternations in HER2-enriched and basal-like tumors.

Since tumor heterogeneity on the DNA level has challenged clinical breast cancer care (Yates *et al.* 2015), the present study has illustrated the benefit of combining both DNA mutations and gene expression in the identification of canonical pathway changes in human breast cancer. Identification of canonical genomic alternations is critical for precision medicine.

Acknowledgments

We thank *The Cancer Genome Atlas* network; C. Curtis and the collaborative consortium; and M. Guedj and colleagues for generating the data.

Author contributions: R.L. provided study design, data analysis, and paper writing; J.C. executed data process and management; and J.I. collaborated in cancer biology research.

Literature Cited

- Amundson, S. A., T. G. Myers, and A. J. Fornace, Jr., 1998 Roles for p53 in growth arrest and apoptosis: putting on the brakes after genotoxic stress. *Oncogene* 17: 3287–3299.
- Basso, K., A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera *et al.*, 2005 Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37: 382–390.
- Berger, J., 1985 *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics, Ed. 2. Springer-Verlag, Berlin; Heidelberg, Germany; New York.
- Bjorkegren, J. L., J. C. Kovacic, J. T. Dudley, and E. E. Schadt, 2015 Genome-wide significant loci: How important are they? Systems genetics to understand heritability of coronary artery disease and other common complex disorders. *J. Am. Coll. Cardiol.* 65: 830–845.
- Breiman, L., 2001 Random Forest. *Mach. Learn.* 45: 5–32.
- Cancer Genome Atlas Network, 2012 Comprehensive molecular portraits of human breast tumours. *Nature* 490: 61–70.
- Carr, J. R., M. M. Kiefer, H. J. Park, J. Li, Z. Wang *et al.*, 2012 FoxM1 regulates mammary luminal cell fate. *Cell Reports* 1: 715–729.
- Carro, M. S., W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao *et al.*, 2010 The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 463: 318–325.
- Chen, X., G. A. Muller, M. Quaas, M. Fischer, N. Han *et al.*, 2013 The forkhead transcription factor FOXM1 controls cell cycle-dependent gene expression through an atypical chromatin binding mechanism. *Mol. Cell. Biol.* 33: 227–236.
- Cox, D., and D. Oakes, 1984 *Analysis of Survival Data*. Chapman & Hall, New York.
- Curtis, C., S. P. Shah, S. F. Chin, G. Turashvili, O. M. Rueda *et al.*, 2012 The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486: 346–352.
- Fletcher, M. N., M. A. Castro, X. Wang, I. de Santiago, M. O'Reilly *et al.*, 2013 Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.* 4: 2464.
- Gonzalez, M., H. Moore, X. Li, K. Toy, W. Huang *et al.*, 2014 EZH2 expands breast stem cells through activation of NOTCH1 signaling. *Proc. Natl. Acad. Sci. USA* 111: 3098–3103.
- Guedj, M., L. Marisa, A. de Reynies, B. Orsetti, R. Schiappa *et al.*, 2012 A refined molecular taxonomy of breast cancer. *Oncogene* 31: 1196–1206.
- Klein, C., 2013 Selection and adaptation during metastatic cancer progression. *Nature* 501: 365–372.
- Kwok, J., B. Peck, L. Monteiro, H. D. Schwenen, J. Millour *et al.*, 2010 FOXM1 confers acquired cisplatin resistance in breast cancer cells. *Mol. Cancer Res.* 8: 24–34.
- Lee, S., Z. Li, Z. Wu, M. Aau, P. Guan *et al.*, 2011 Context-specific regulation of NF- κ B target gene expression by EZH2 in breast cancers. *Mol. Cell* 43: 798–810.
- Lehmann, E., and H. J. M. D'Abbrera, 1988 *Nonparametrics: Statistical Methods Based on Ranks*. New York: McGraw-Hill.
- Li, R., and G. Churchill, 2010 Epistasis contributes to the genetic buffering of plasma HDL cholesterol in mice. *Physiol. Genomics* 42A: 228–234.
- Li, R., S. W. Tsaih, K. Shockley, I. M. Stylianou, J. Wergedal *et al.*, 2006 Structural model analysis of multiple quantitative traits. *PLoS Genet.* 2: e114.
- Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky *et al.*, 2006 ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl 1): S7.
- Michailidou, K., J. Beesley, S. Lindstrom, S. Canisius, J. Dennis *et al.*, 2015 Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* 47: 373–380.
- Nik-Zainal, S., P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman *et al.*, 2012 The life history of 21 breast cancers. *Cell* 149: 994–1007.
- Parker, J. S., M. Mullins, M. C. Cheang, S. Leung, D. Voduc *et al.*, 2009 Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27: 1160–1167.
- Perou, C. M., T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey *et al.*, 2000 Molecular portraits of human breast tumours. *Nature* 406: 747–752.
- Shannon, C., 1948 A mathematical theory for communication. *Bell Syst. Tech. J.* 27: 379–423.
- Theodorou, V., R. Stark, S. Menon, and J. S. Carroll, 2013 GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Res.* 23: 12–22.
- Tolaney, S.M., and E. P. Winer, 2007 Follow-up care of patients with breast cancer. *Breast* 16(Suppl 2): S45–S50.
- Vaquerizas, J. M., S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, 2009 A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10: 252–263.
- Vogelstein, B., N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, Jr. *et al.*, 2013 Cancer genome landscapes. *Science* 339: 1546–1558.
- Wang, M., and A. L. Gartel, 2011 The suppression of FOXM1 and its targets in breast cancer xenograft tumors by siRNA. *Oncotarget* 2: 1218–1226.
- Yamaguchi, H., and M. C. Hung, 2014 Regulation and role of EZH2 in cancer. *Cancer Res. Treat.* 46: 209–222.
- Yates, L. R., M. Gerstung, S. Knappskog, C. Desmedt, G. Gundem *et al.*, 2015 Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* 21: 751–759.

Communicating editor: S. K. Sharan

GENETICS

Supporting Information

www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.180125/-/DC1

A Gene Regulatory Program in Human Breast Cancer

Renhua Li, John Campos, and Joji Iida

Figure S1. Distributions of age at diagnosis in the three cohorts of breast cancer populations. For the Guedj cohort, the clinical record system may be different from the other two cohorts, as age at initial diagnosis was recorded (GUEDJ *et al.* 2012).

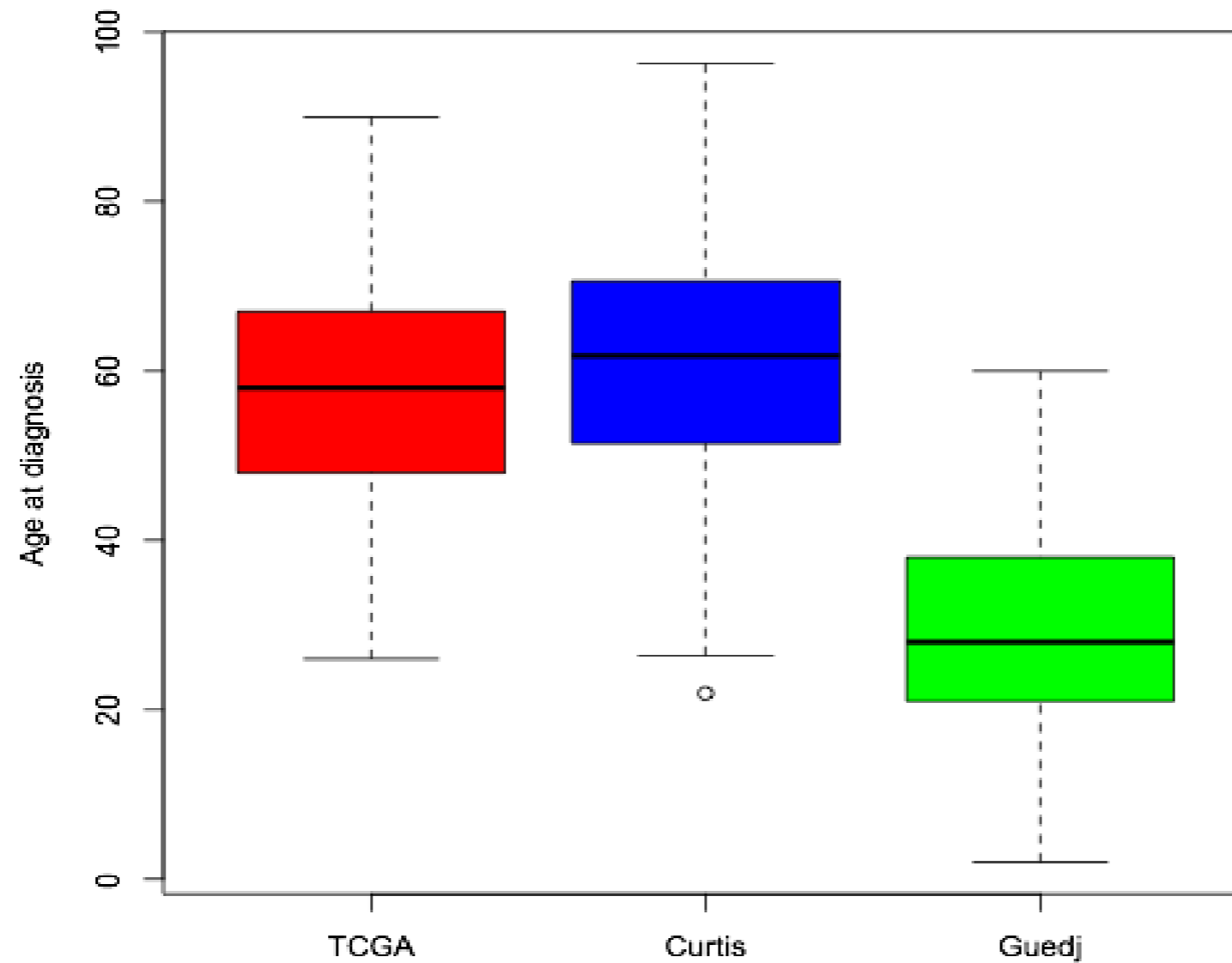


Figure S2. A schematic framework for machine learning. We used the downloaded TCGA RNA-Seq data for gene selection. In machine learning, we need to split the total samples (N=646) into training (3/4) and test (1/4) subsets. Subtype assignment of the training samples was based on the PAM50 method (PARKER et al. 2009). The gene panel used for gene search includes: 1) a set of well annotated 1,400 transcriptional factor genes (VAQUERIZAS et al. 2009); and 2) a set of 138 cancer driver genes (VOGELSTEIN et al. 2013). Once we have identified a small ensemble of genes that can best discriminate the four subtypes of breast tumors, we cross-validated the classifier using both the TCGA test data and the data from independent cohorts.

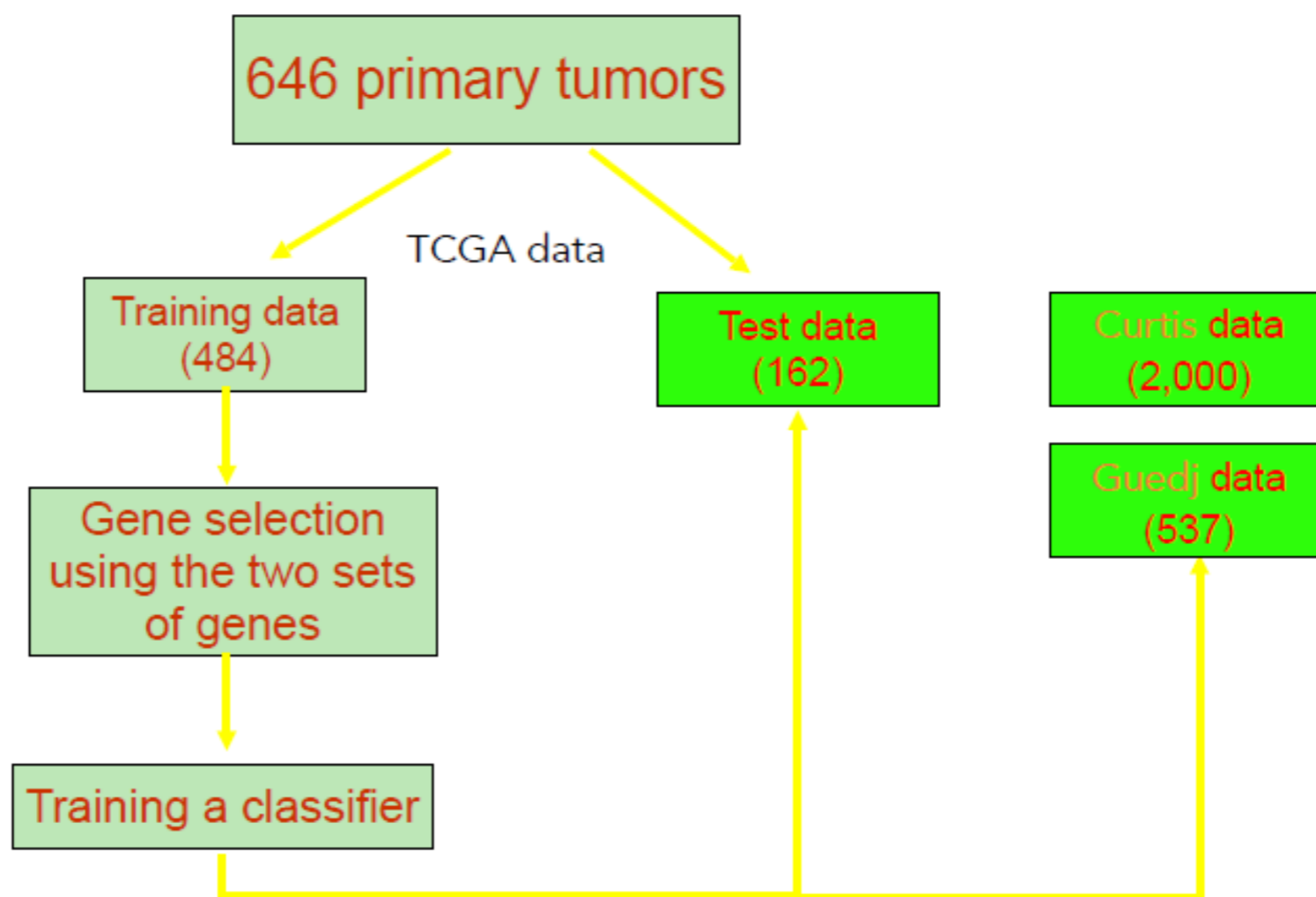
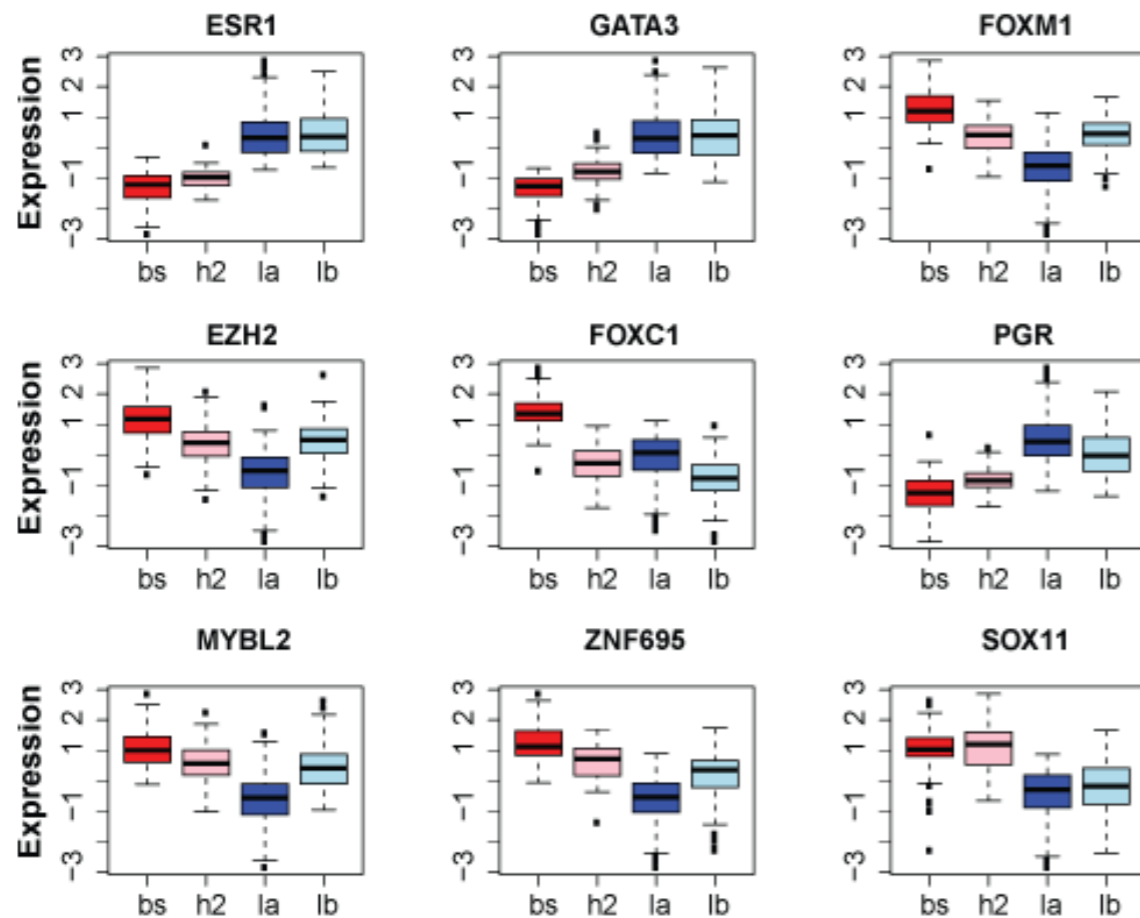


Figure S3. Gene expression patterns of the MR16 genes selected by machine learning. The box plots show gene expression across subtypes. The y-axis is the gene expression adjusted by van der Waerden scores (LEHMANN AND HJM 1988). The x-axis represents the four subtypes; that is, bs = basal-like; h2 = HER2 enriched; la = luminal A; lb = luminal B.

Top 1-9 genes



Top 10-16 genes

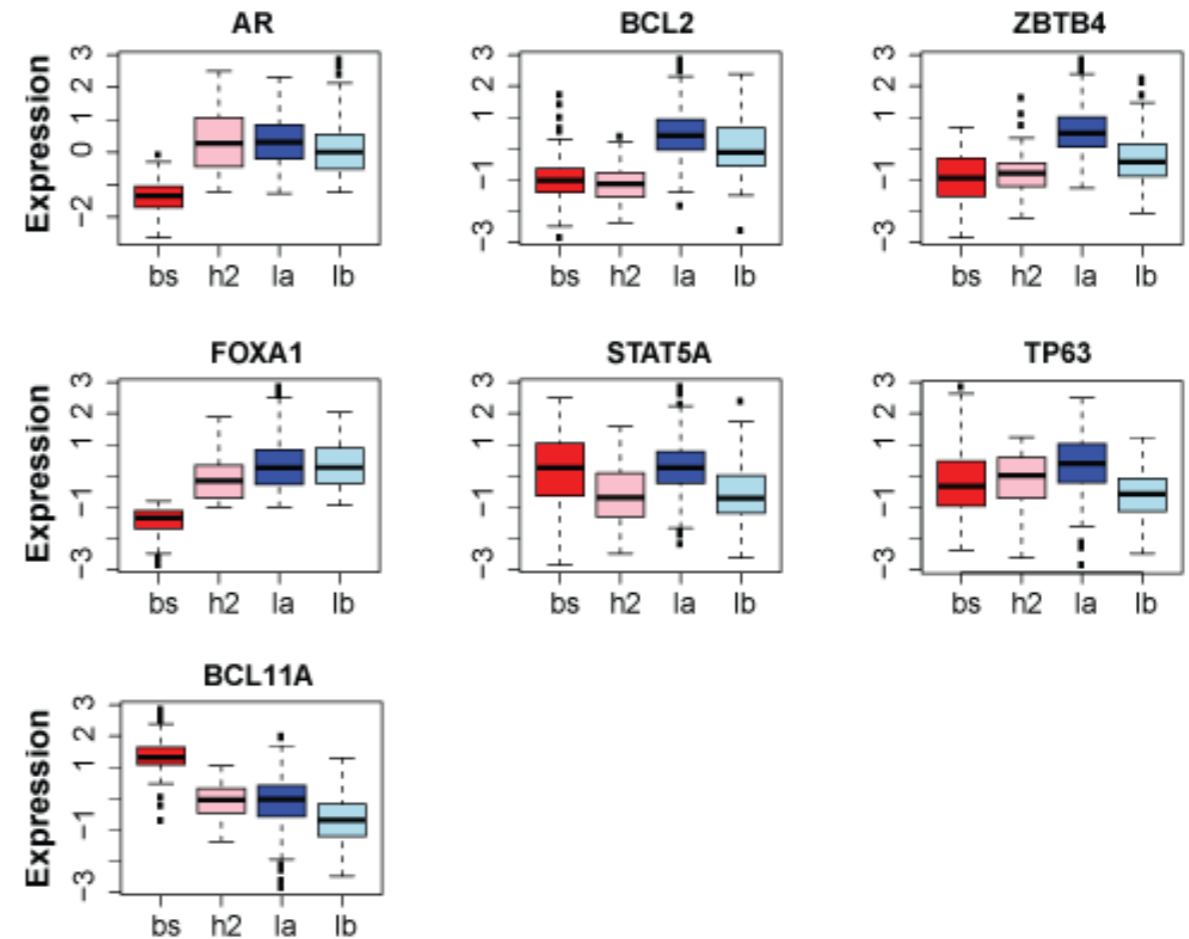


Figure S4. Classification accuracy as a function of number of genes selected.

We used an algorithm named “random-forest coupled with recursive gene elimination” to select a small number of genes that can best discriminate the training samples into the four subtypes. In order to avoid sample selection bias, bootstrap resampling within the training data was implemented. The x-axis is the number of genes to be selected. The y-axis is the averaged gene classification accuracy across the bootstrap samples.

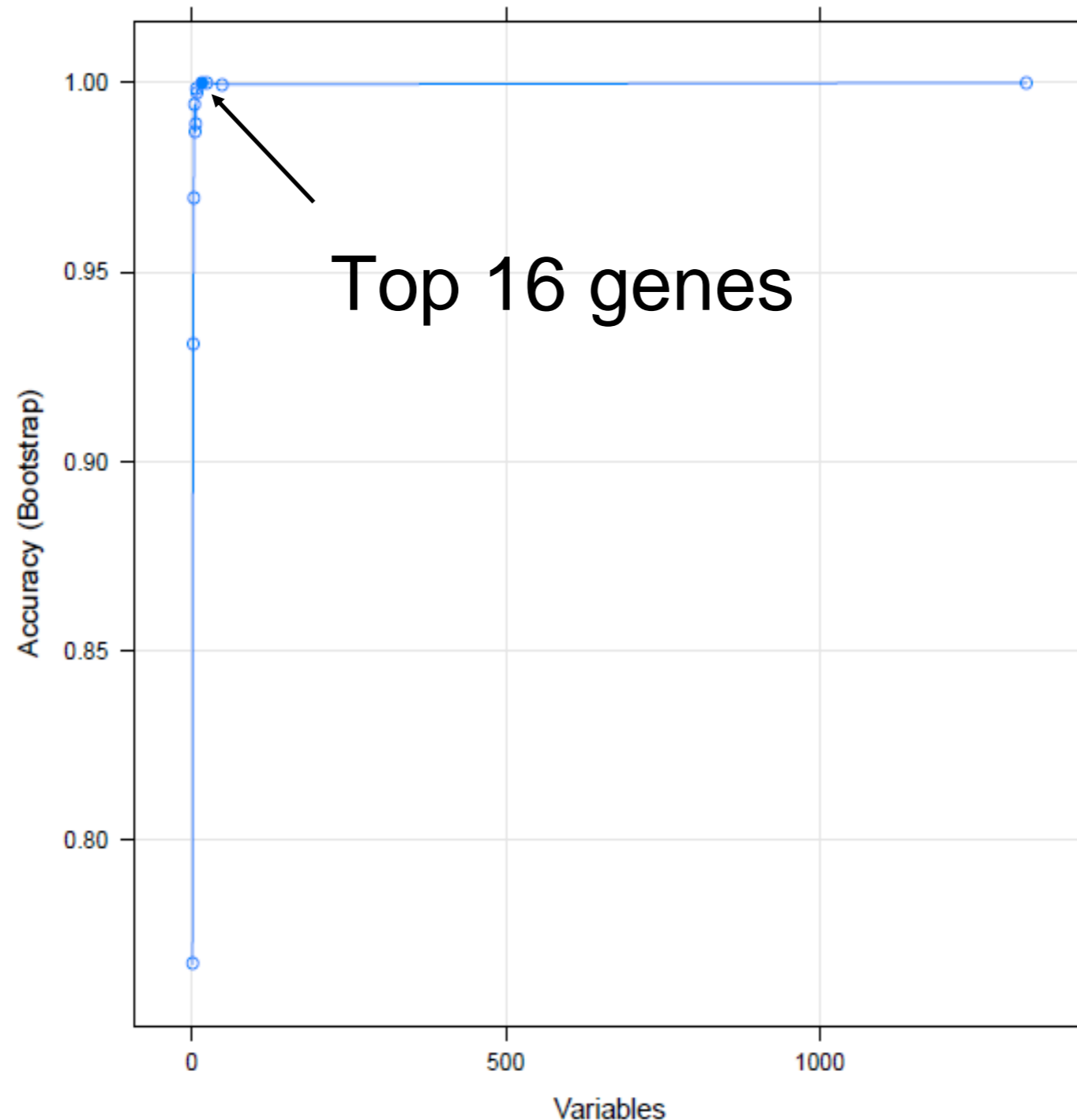


Figure S5. Gene selection frequency across different partitions of the TCGA samples.

In machine learning, gene selection instability is an issue, which means different sets of genes can be selected based on different subsets of training samples used. We partitioned the entire TCGA samples downloaded into the training and test subsets 100 times. In each iteration, we retained the top genes up to 40 that were selected by the algorithm. (In case that less than 40 genes were selected, we retained the actual genes.) Gene frequency was then computed across the 100 iterations. The 16 genes are among the top ones that are most frequently selected.

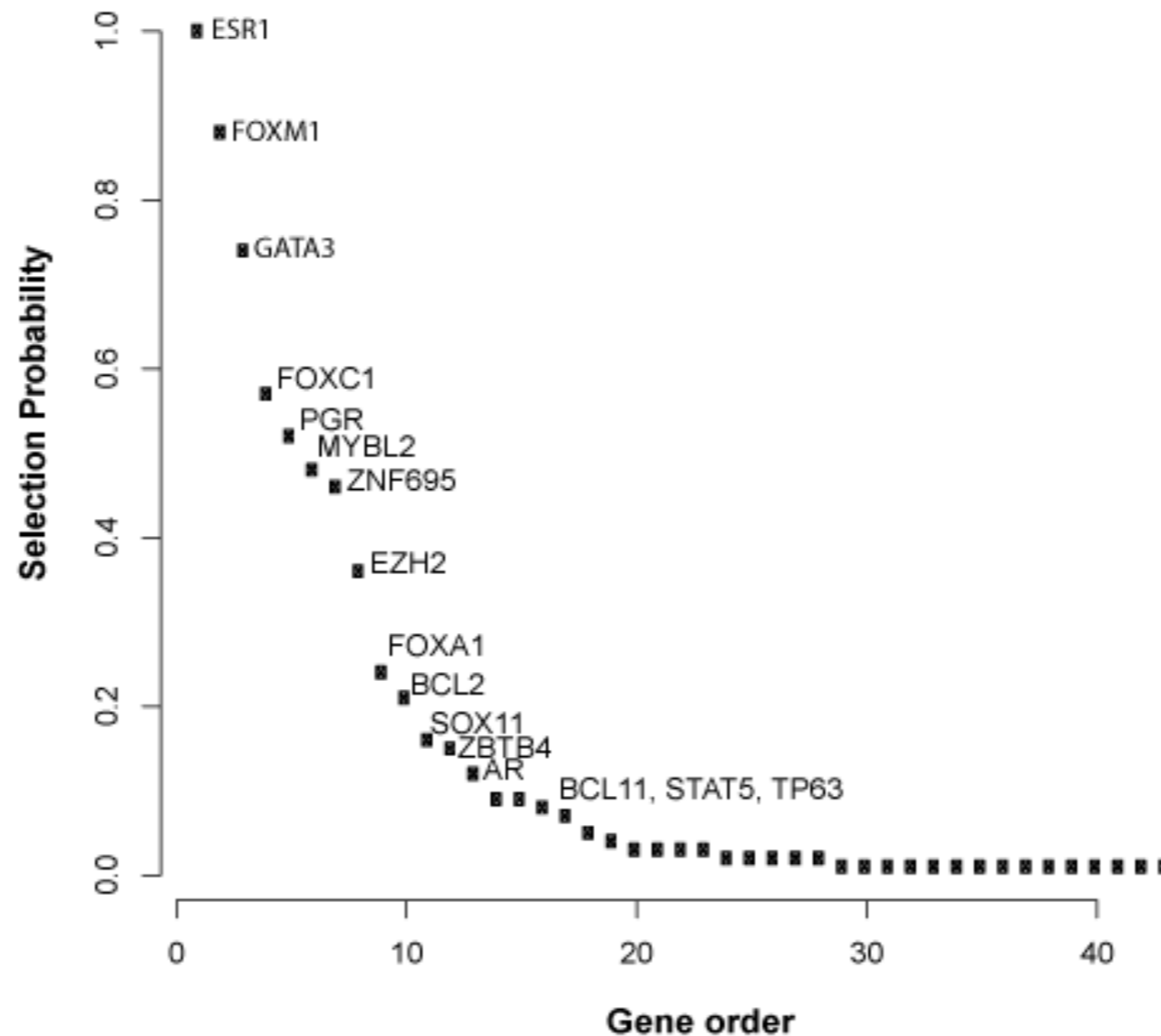


Figure S6. Heatmaps of MR16 gene expression in the Curtis and Guedj cohorts. Green and red colors indicate low and high gene expression, respectively. The same Group 1 MRs, as defined in the TCGA data (Figure 1A), are over-expressed in the ER-positive tumors (blue bar at the bottom), but lowly expressed in the ER-negative tumors (red bar at the bottom). In contrast, the same Group 2 MRs are over-expressed in the ER-negative tumors, but lowly expressed in the ER-positive tumors. **A:** The MR16 gene expression patterns in the Curtis cohort. Each column is a tumor sample. Due to the too large sample size, we used only the validation data that is approximately half of the 2000 samples. (Another half is the discovery data and results are similar.) Among the MR16 genes, *TP63* gene failed to the annotation as described in the Methods section, thus 15 genes are present in the heatmap. The samples display two large clusters that correspond to the ER-positive and ER-negative tumors. In each of the two clusters, there exist two sub-clusters. In general, the four clusters correspond well to the four major subtypes. **B:** The MR16 gene expression patterns in the Guedj cohort. Each column is a tumor sample. The samples also exhibit two large clusters that correspond to the ER-positive and ER-negative tumors. In each of the two clusters, the sub-clusters are more complicated, compared to the TCGA and the Curtis cohorts. This may suggest more subtypes in this cohort.

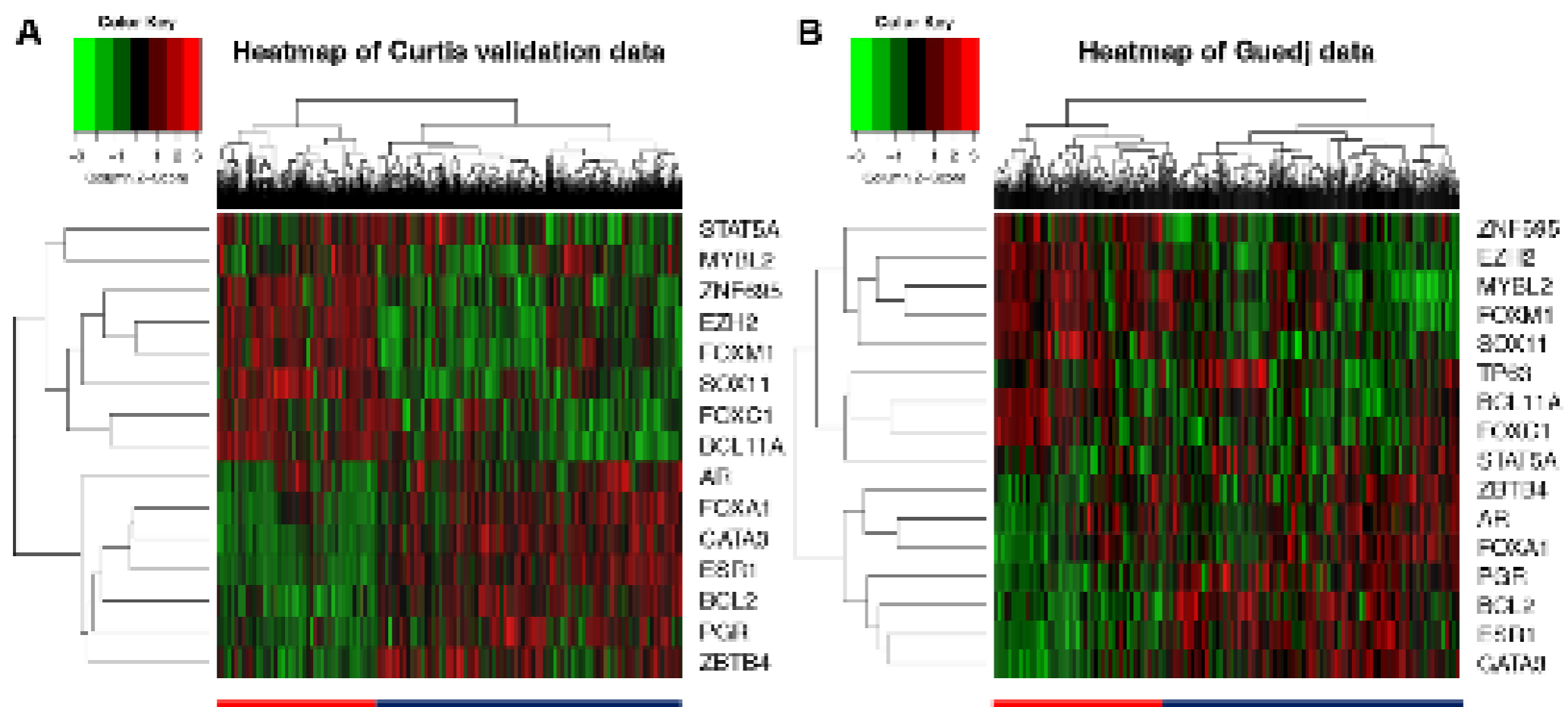
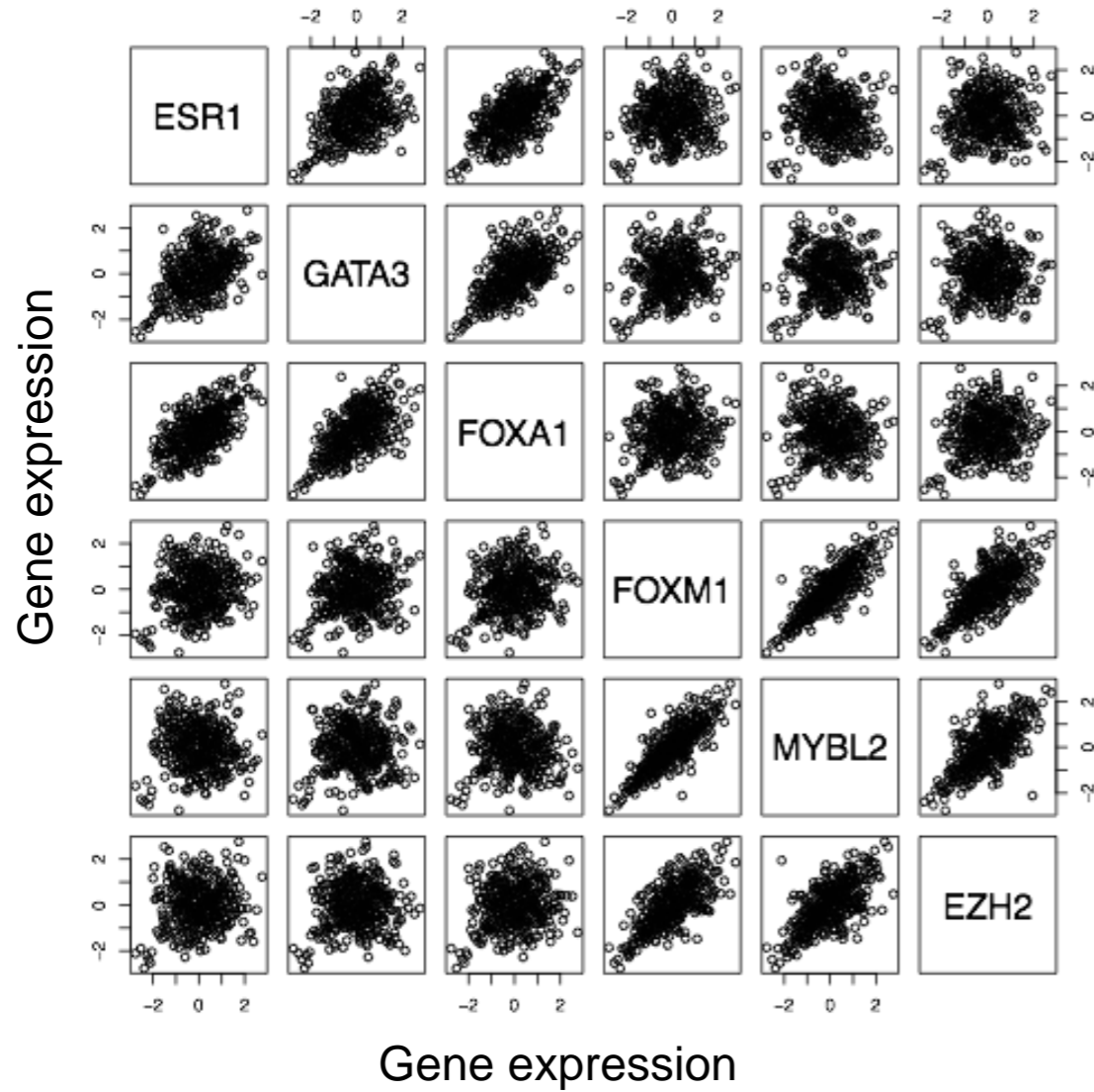


Figure S7. Correlation analyses between representative master regulator genes by subtype. We used Pearson correlation to explore the relationships among the representative six MRs. It appears that there exist moderate positive correlations for the intra-group MRs. Regarding the inter-group MRs, however, there are weak or no correlations between them. In luminal B and basal-like subtypes, the correlations of the inter-group MRs are negative.

A

Luminal A



Correlation coefficients

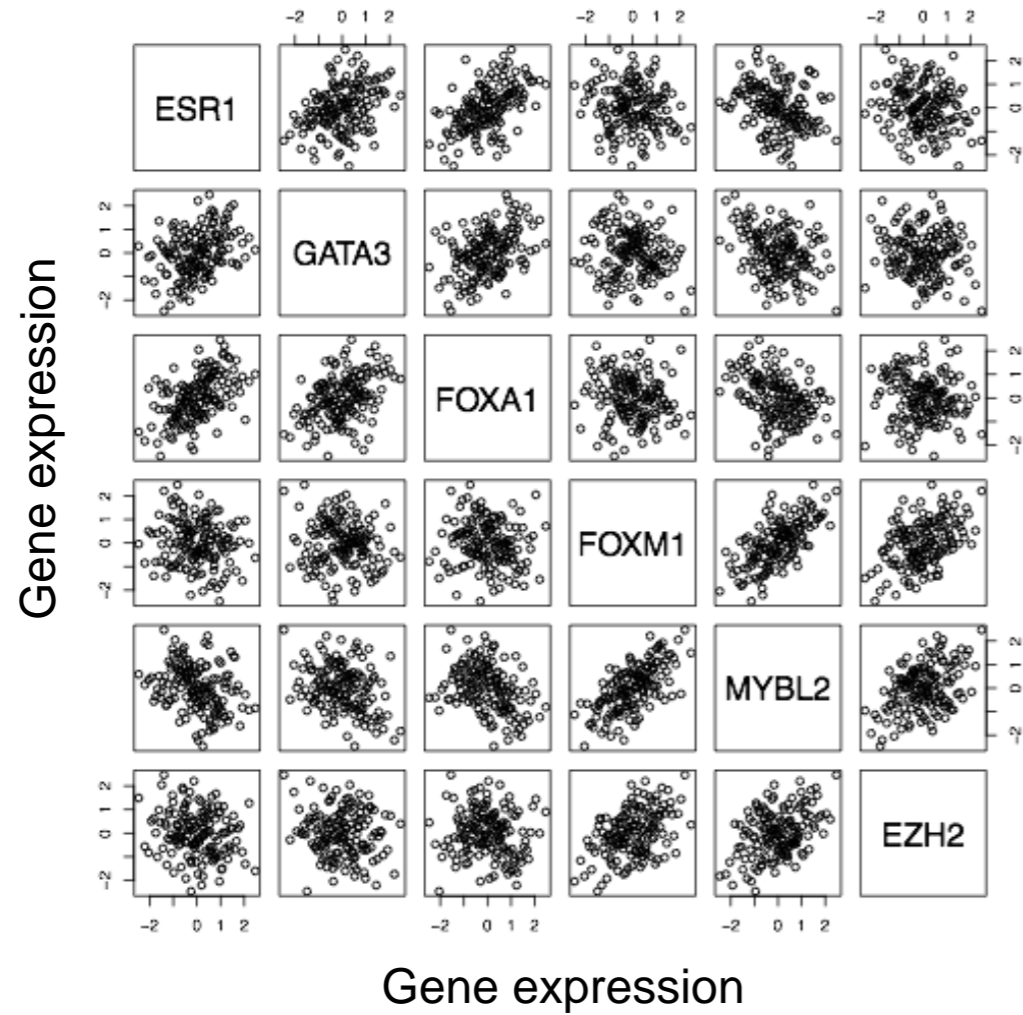
	ESR1	GATA3	FOXA1	FOXM1	MYBL2	EZH2
ESR1	1					
GATA3	0.45	1				
FOXA1	0.64	0.61	1			
FOXM1	0.14	0.18	0.17	1		
MYBL2	-0.04	0.04	-0.04	0.83	1	
EZH2	0.12	0.09	0.17	0.70	0.66	1

p-values

	ESR1	GATA3	FOXA1	FOXM1	MYBL2	EZH2
ESR1	0					
GATA3	0	0				
FOXA1	0	0	0			
FOXM1	0.01	0.00	0.00	0		
MYBL2	0.46	0.50	0.51	0	0	
EZH2	0.03	0.11	0.00	0	0	0

B

Luminal B



Correlation coefficients

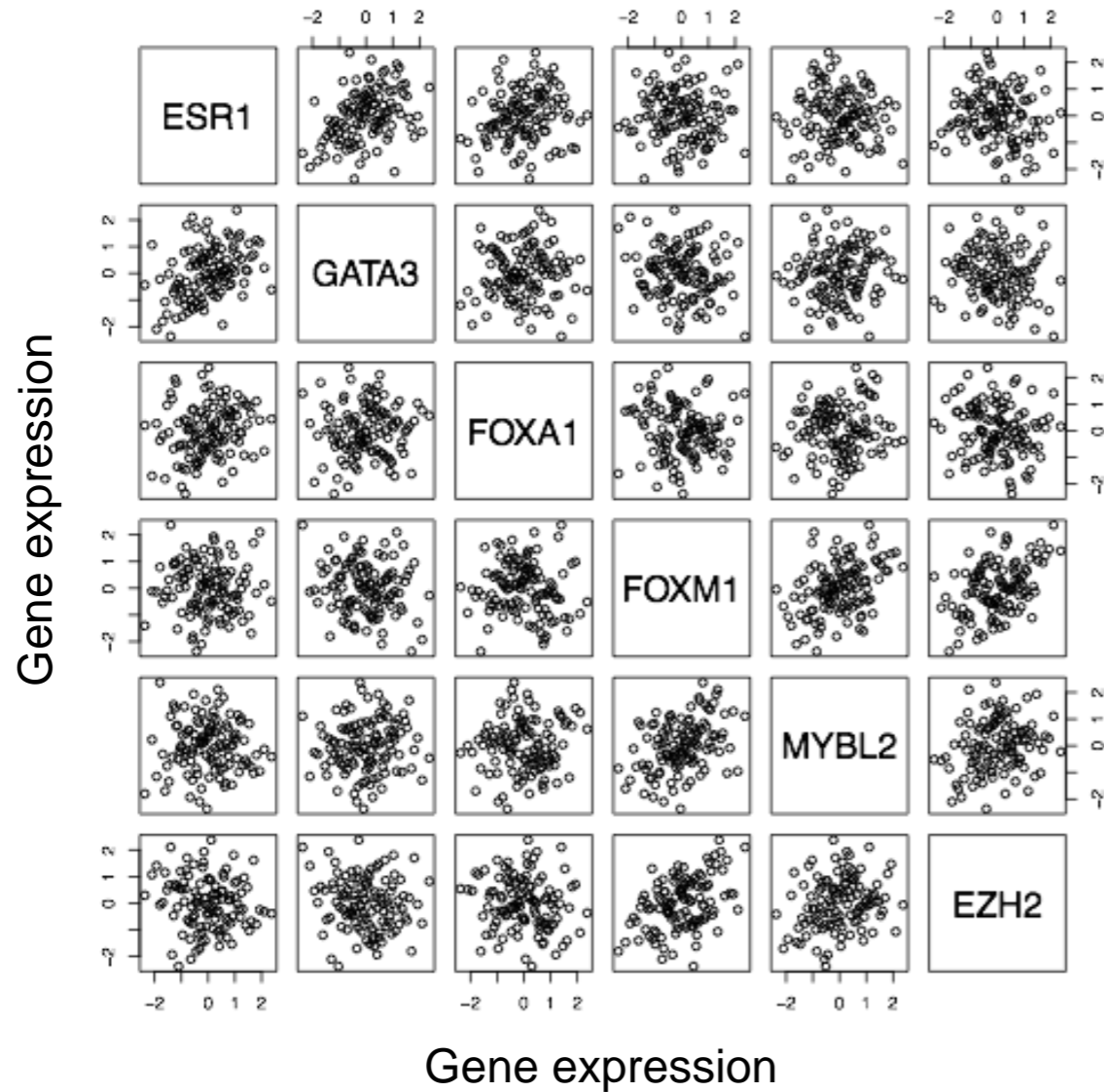
	ESR1	GATA3	FOXA1	FOXM1	MYBL2	EZH2
ESR1	1					
GATA3	0.35	1				
FOXA1	0.52	0.41	1			
FOXM1	-0.13	-0.11	-0.11	1		
MYBL2	-0.28	-0.32	-0.36	0.66	1	
EZH2	-0.17	-0.12	-0.08	0.45	0.42	1

p-values

	ESR1	GATA3	FOXA1	FOXM1	MYBL2	EZH2
ESR1	0					
GATA3	1.24e-05	0				
FOXA1	2.54e-11	2.65e-07	0			
FOXM1	0.13	0.17	0.18	0		
MYBL2	0.00	7.53e-05	6.31e-06	0	0	
EZH2	0.04	0.13	0.35	1.24e-08	9.33e-08	0

C

Basal-like



Correlation coefficients

	ESR1	GATA3	FOXA1	FOXM1	MYBL2	EZH2
ESR1	1					
GATA3	0.39	1				
FOXA1	0.17	0.06	1			
FOXM1	-0.03	-0.15	-0.08	1		
MYBL2	-0.02	0.08	0.07	0.39	1	
EZH2	-0.11	-0.12	-0.12	0.43	0.22	1

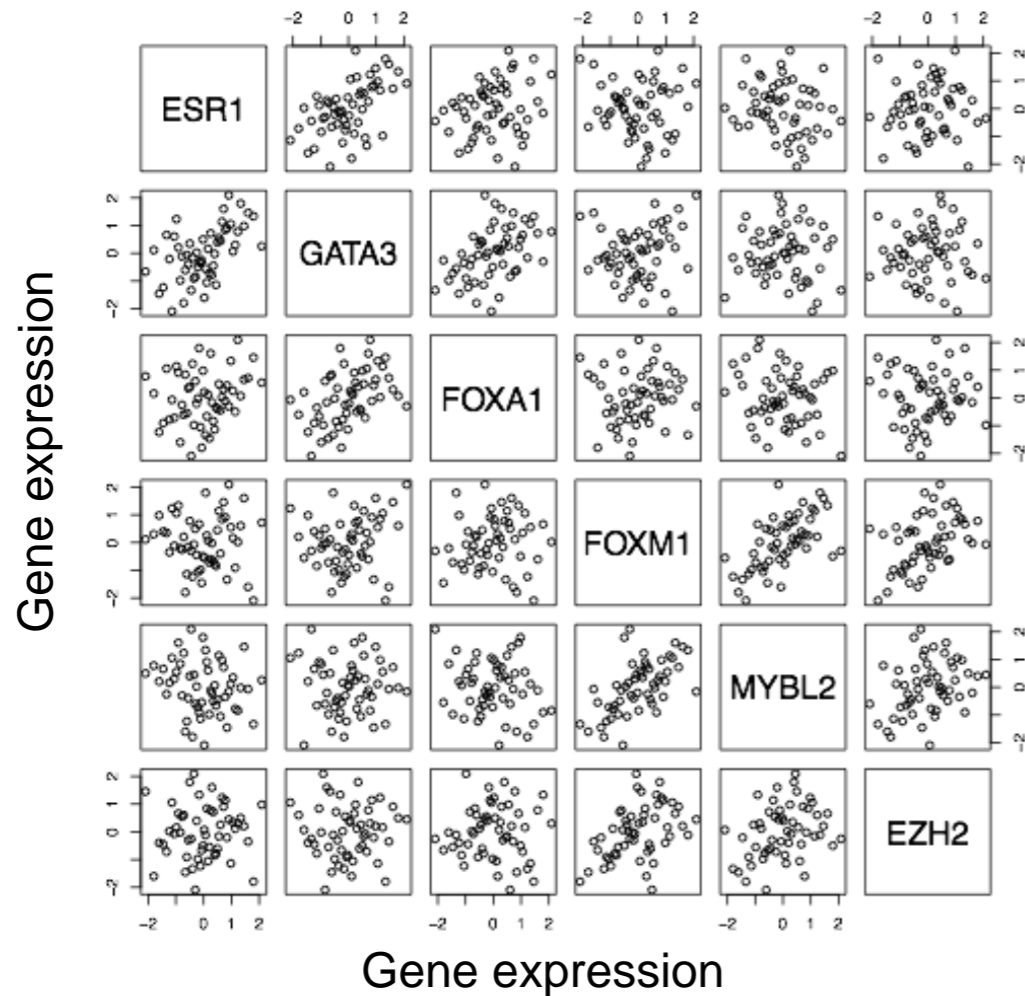
p-values

	ESR1	GATA3	FOXA1	FOXM1	MYBL2	EZH2
ESR1	0					
GATA3	1.97e-05	0				
FOXA1	0.06	0.55	0			
FOXM1	0.75	0.11	0.41	0		
MYBL2	0.81	0.40	0.44	2.31e-05	0	
EZH2	0.25	0.20	0.20	2.06e-06	0.02	0

D

HER2 enriched

Correlation coefficients



	ESR1	GATA3	FOXA1	FOXM1	MYBL2	EZH2
ESR1	1					
GATA3	0.54	1				
FOXA1	0.22	0.46	1			
FOXM1	-0.07	0.14	0.06	1		
MYBL2	-0.19	-0.04	-0.18	0.59	1	
EZH2	0.04	0.00	-0.00	0.44	0.32	1

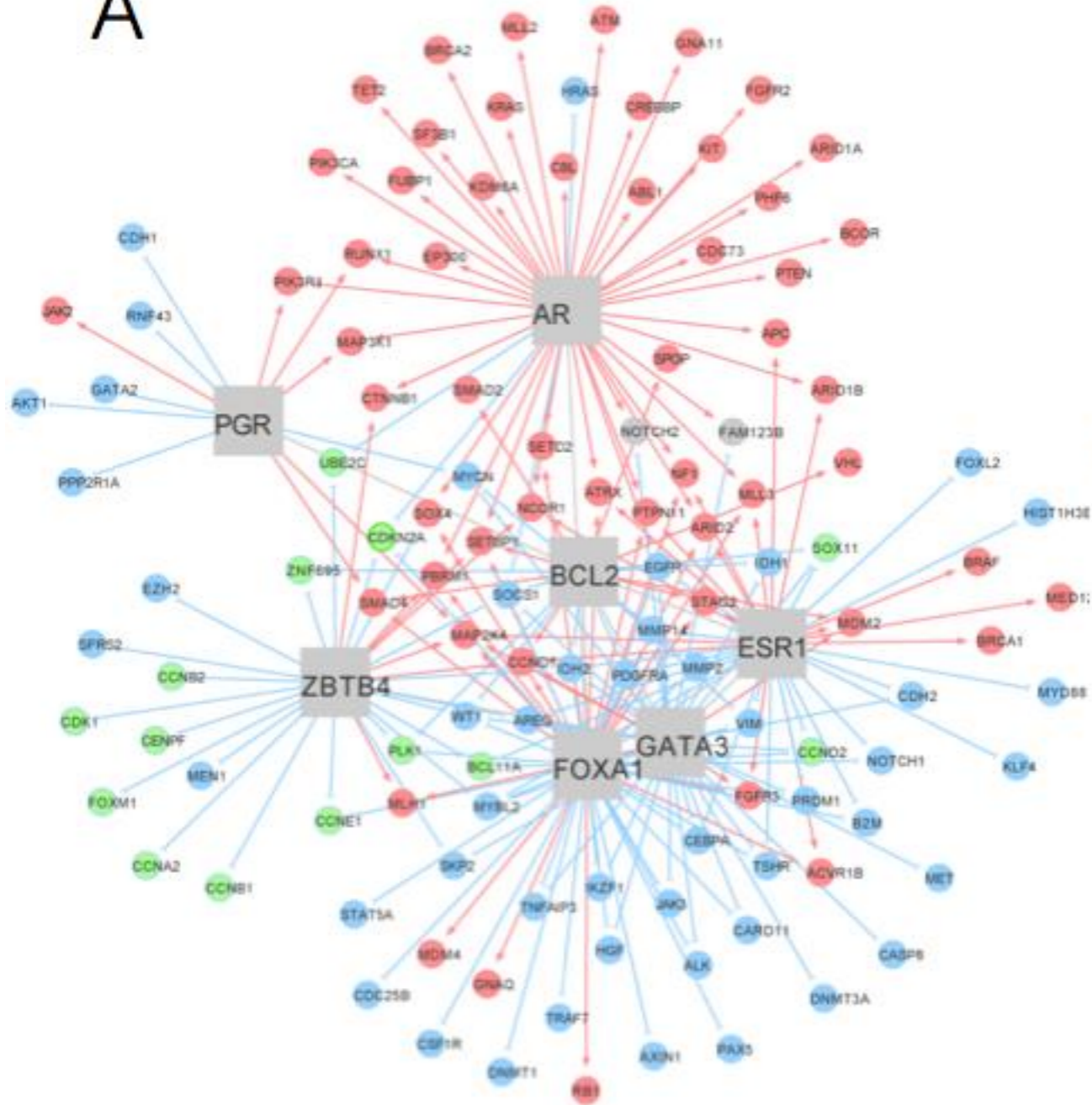
p-values

	ESR1	GATA3	FOXA1	FOXM1	MYBL2	EZH2
ESR1	0					
GATA3	2.52e-05	0				
FOXA1	0.11	0.00	0			
FOXM1	0.64	0.32	0.69	0		
MYBL2	0.17	0.77	0.18	2.18e-06	0	
EZH2	0.75	0.99	0.99	0.00	0.02	0

Figure S8. Subtype-specific gene regulatory networks in the TCGA cohort. We used two subtypes (luminal B and basal-like) to illustrate Subtype-specific gene networks. The circular nodes represent regulons and the edges connecting the MRs and the regulons indicate interactions between them. Red and blue colors denote up- and down-regulation, respectively. **A:** A network specific to the luminal B subtype. Green nodes are highlighted for either cell cycle genes or Group 2 MRs. Group 2 MRs, such as *ZNF695* and *SOX11*, and cell cycle genes, such as *CCNB1* and *CDK1*, are down-regulated by the central Group 1 MRs. **B:** A network specific to the basal-like subtype. The cell cycle genes, such as *PLK1* and *CDK1*, are up-regulated by the central Group 2 MRs.

Figure S8

A



B

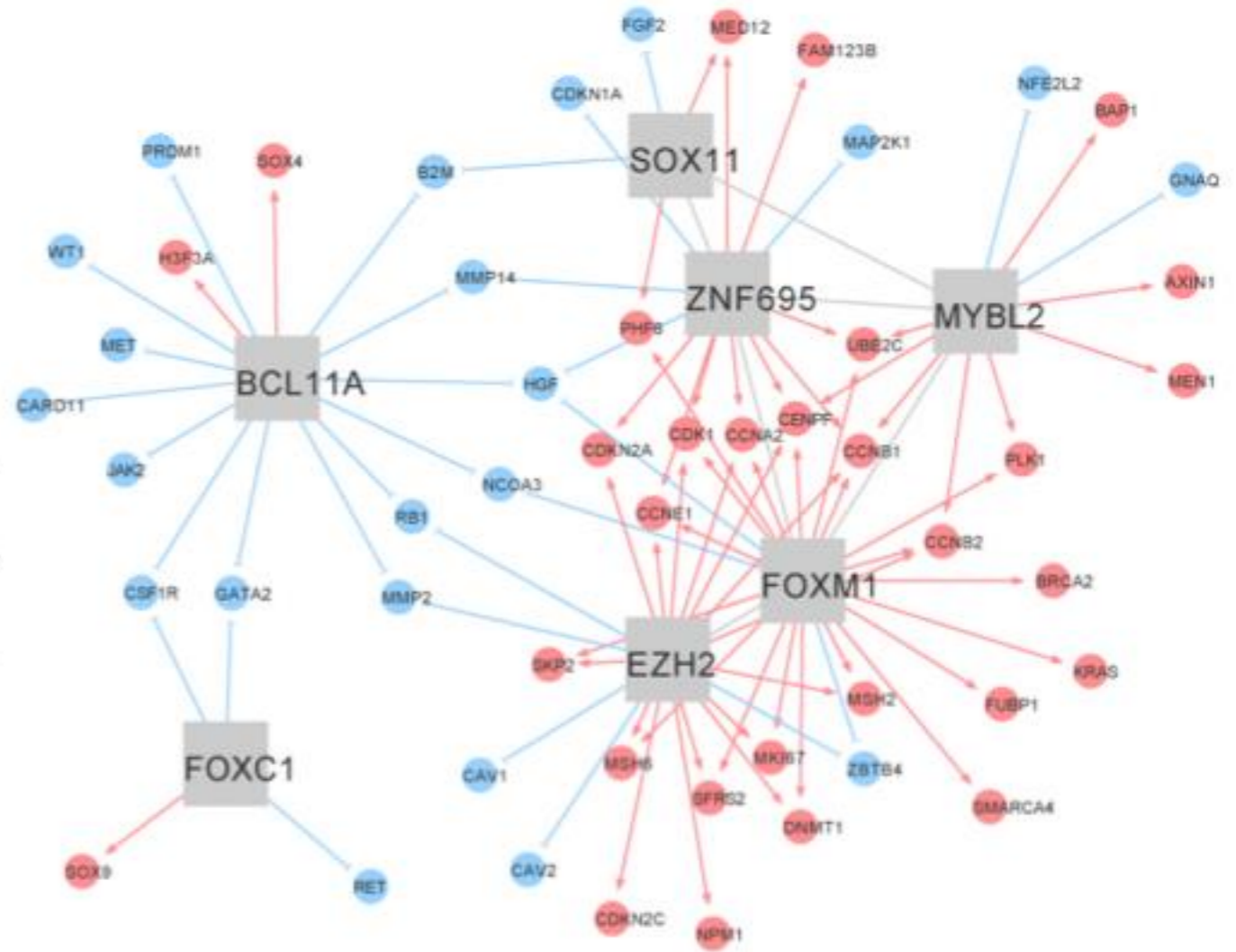
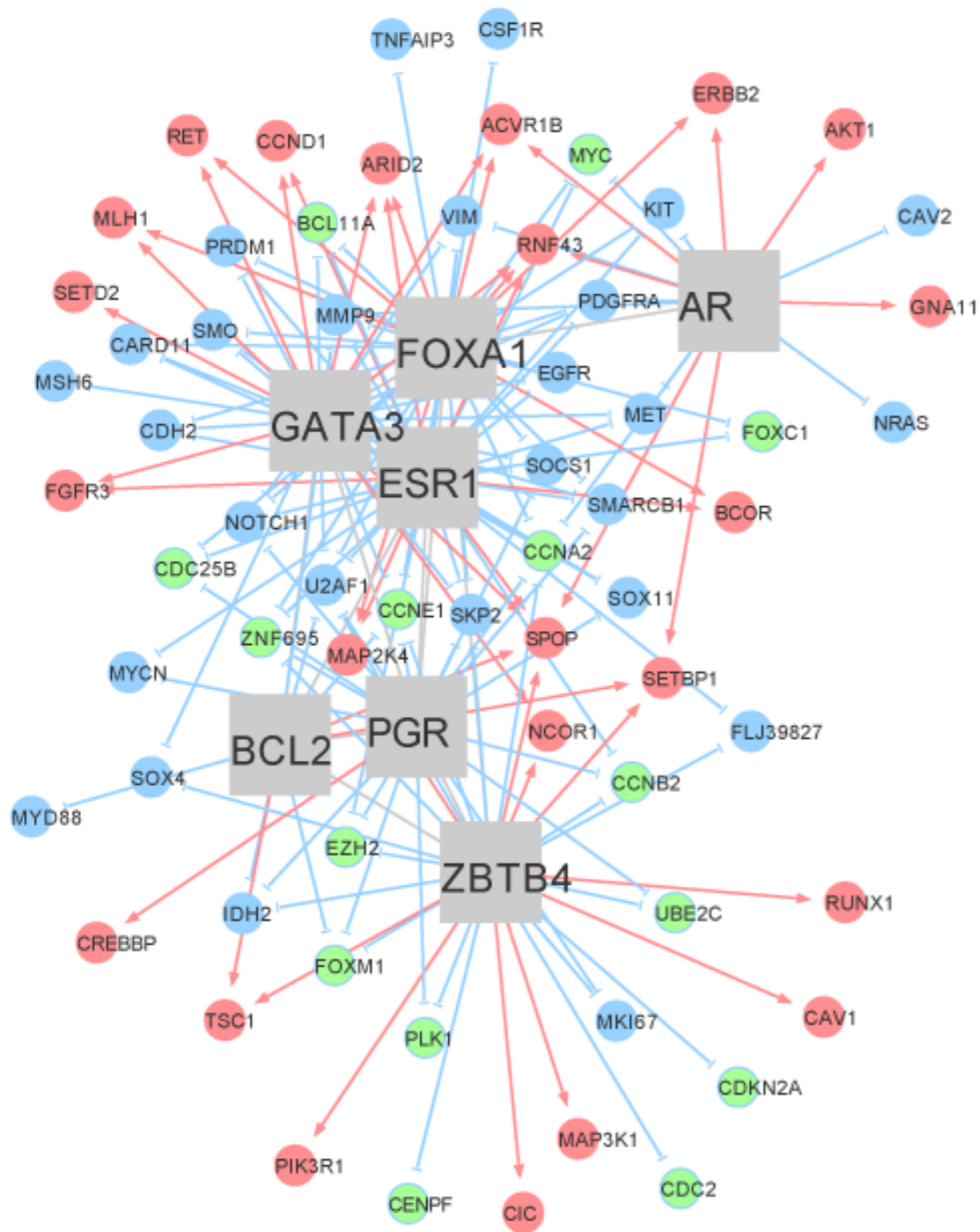


Figure S9. Validation of regulator-regulon interaction patterns in the Curtis cohort. See Figure S8 for nodes and edges. **A:** a gene regulatory network specific to the luminal B subtype. The central genes are the Group 1MRs. The genes highlighted in green are either cell cycle genes or Group 2 MRs. **B:** a gene regulatory network specific to the basal-like subtype. The central genes are the Group 2 MRs. The genes highlighted in green are either cell cycle genes or Group 1 MRs.

Figure S9

A



B

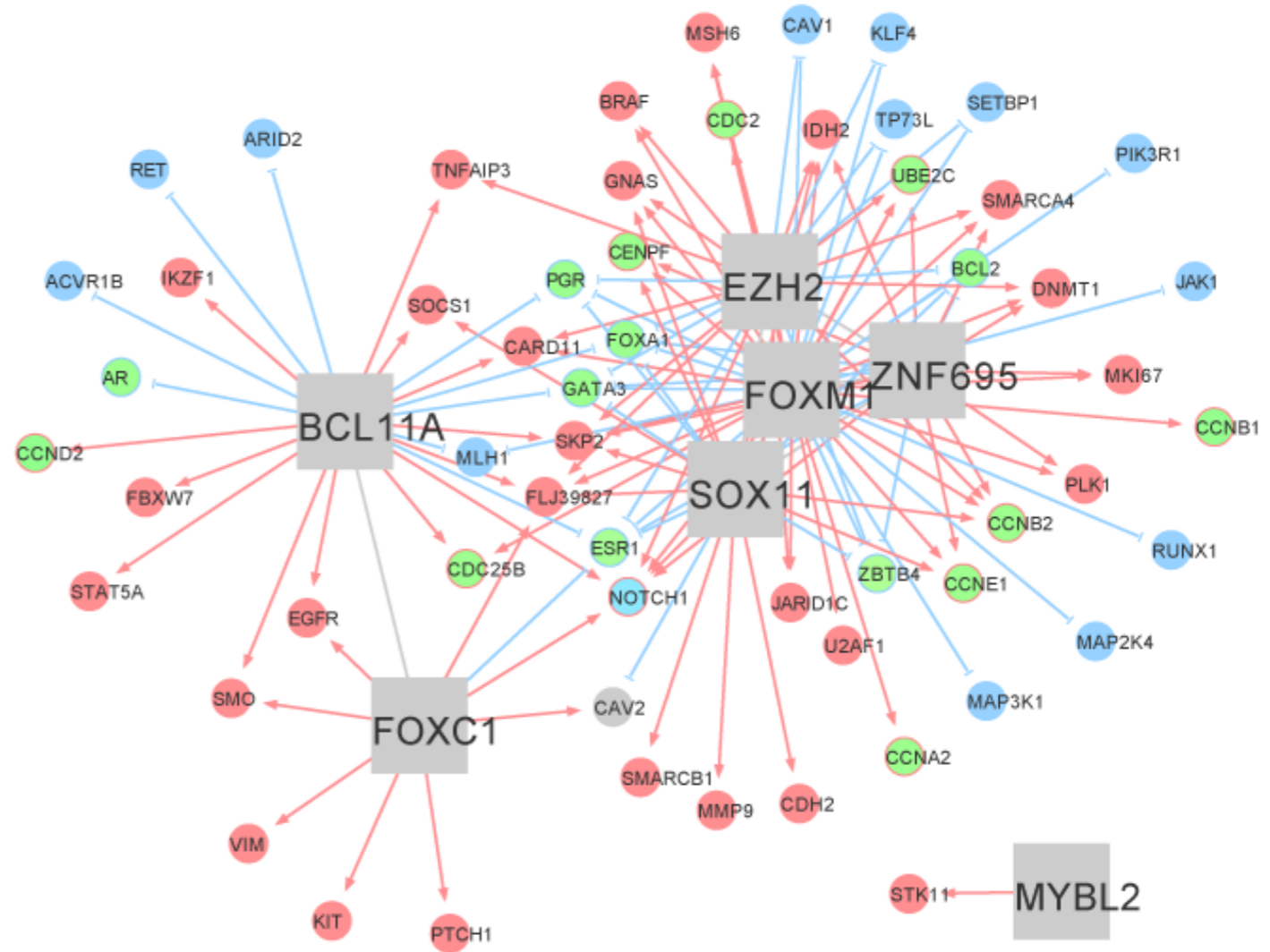
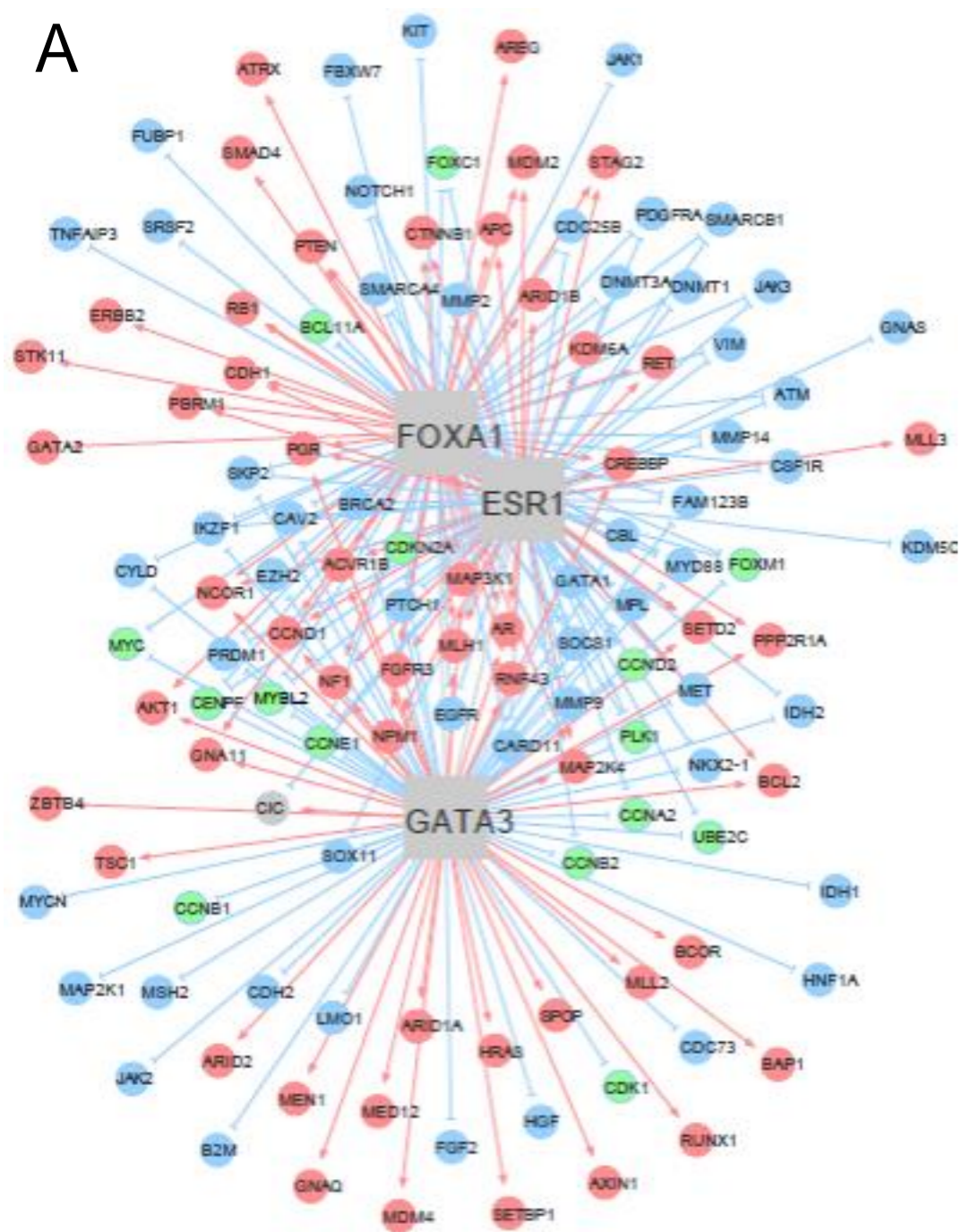


Figure S10. Validation of regulator-regulon interaction patterns in the Guedj cohort. Due to the smaller population size in the Guedj cohort, compared to the other two cohorts, we focus on the cross-subtype gene regulatory networks. See Figure S8 for nodes and edges. **A:** a gene regulatory network illustrating the interactions between three representative Group 1 MRs and cancer driver/related genes. Green nodes are highlighted for either cell cycle genes or Group 2 MRs. Other Group 1 MRs, such as *AR*, are up-regulated by the central regulators. However, Group 2 MRs, such as *FOXM1*, and the cell cycle genes, such as *CCNA2* and *CCNB2* are down-regulated by the central MRs. **B:** A gene regulatory network illustrating the interactions between the four Group 2-MRs and the regulons. Green nodes are highlighted for either the cell cycle genes or the Group 1 MRs. The cell cycle genes, such as *CDk1* and *CCNA2*, are up-regulated by the central MRs. However, the Group 1 MRs, such as *GATA3* and *FOXA1*, are down-regulated by the central MRs.

Figure S10

A



B

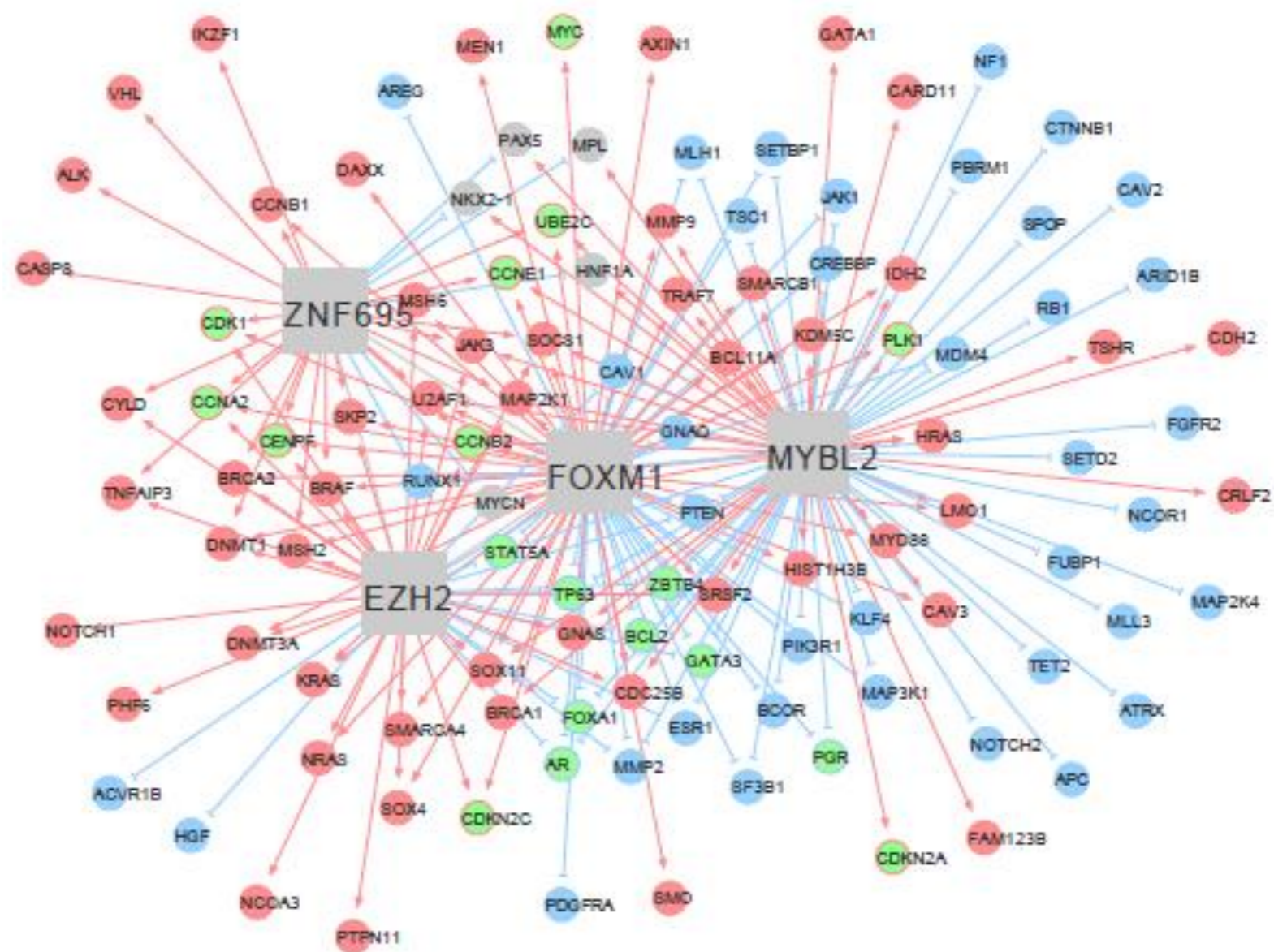


Figure S11. The relationships between MR16 and the PAM50. The central genes are the two MR groups: purple color = Group 1; green color = Group 2. See Figure S8 for nodes and edges. Red and blue colors denote up- and down-regulation, respectively. Results from mutual information modeling indicate that the majority of PAM50 genes are target genes regulated by the MR16, except for the overlapping genes.

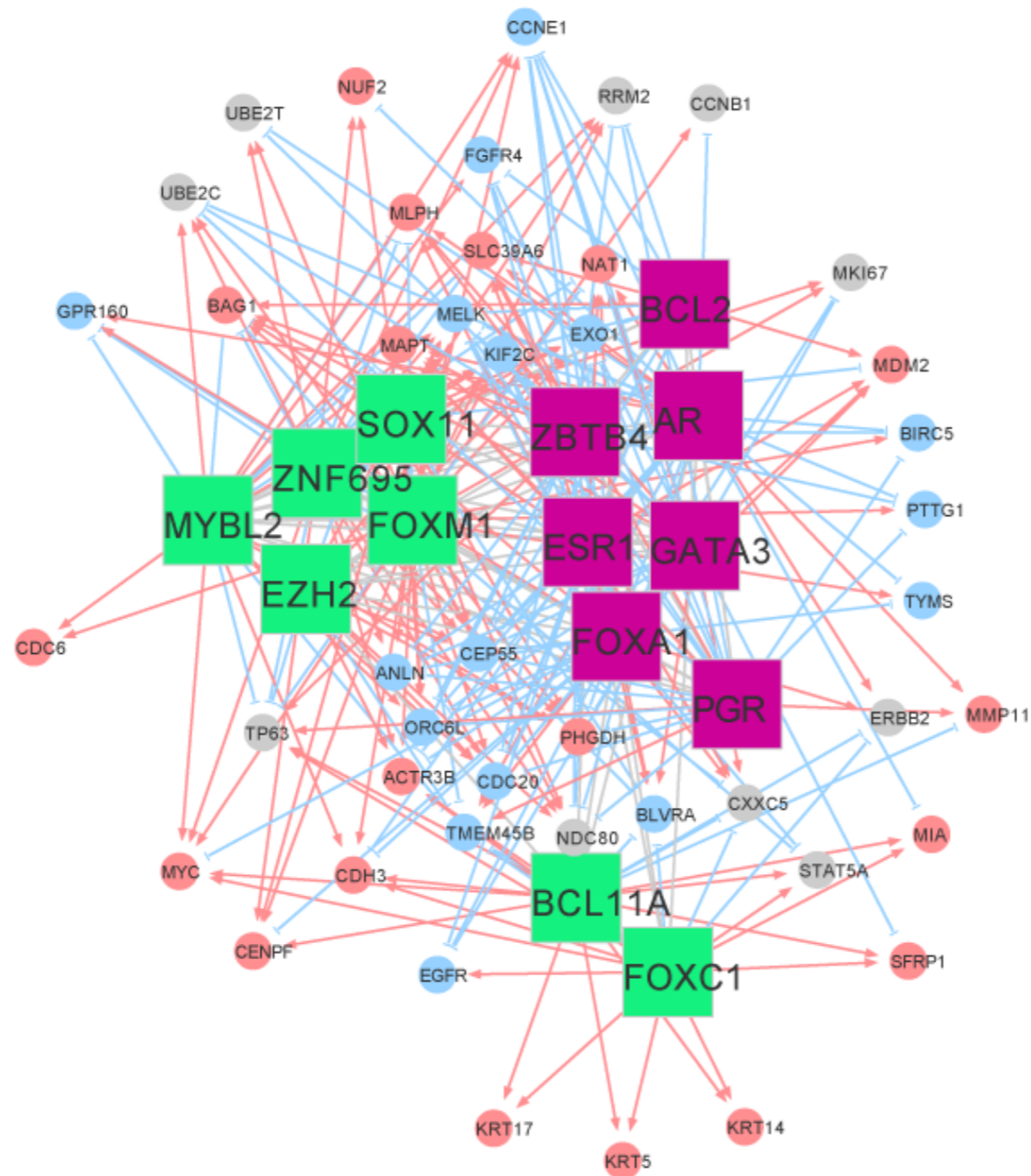


Table S1. Mutation types in the top three genes*

<i>Type</i>	<i>PIK3CA</i>	<i>TP53</i>	<i>GATA3</i>
Frame_Shift_Del	0	33	14
Frame_Shift_Ins	2	13	45
In_Frame_Del	13	5	0
In_Frame_Ins	1	0	0
Missense	273	158	6
Nonsense	0	35	0
Silent	4	3	1
Splice_Site	0	18	20
<i>Total</i>	<i>293</i>	<i>265</i>	<i>86</i>

* The mutation types are summarized here based on the 720 TCGA breast cancer cases.

* The original data in variant call format are downloaded from the TCGA data portal (<https://tcga-data.nci.nih.gov/>).