A Gene Regulatory Program in Human Breast Cancer

Renhua Li 1, 2 *, John Campos 1, Joji Iida 1

1. Windber Research Institute, Windber, PA
2. The Jackson Laboratory, Bar Harbor, ME

* Corresponding author:

Renhua Li
12101 Village Square
Rockville, MD 20852
The United States

Tel: (301) 828-8394
Email: renhua.li10@gmail.com
Abstract

Molecular heterogeneity in human breast cancer has challenged diagnosis, prognosis, and clinical treatment. It is well-known that the molecular subtypes of human 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, in order to explore the mechanisms. Using existing data from three large cohorts of human breast cancer patient populations, including the TCGA breast cancer data, 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 data 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 \textit{ESR1}, \textit{FOXA1}, and \textit{GATA3}, are over-expressed in luminal A and luminal B subtypes, but lowly expressed in HER2 enriched and basal-like subtypes. In contrast, Group 2 MRs, including \textit{FOXM1}, \textit{EZH2}, \textit{MYBL2}, and \textit{ZNF695}, display an opposite pattern. Further, 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.

**Key words**: breast cancer; master regulator; regulator-regulon interactions; gene regulatory program

**Introduction**

Human breast cancer is the most common malignancy in women, with over 200,000 new cases diagnosed each year in the US (TOLANEY SM AND EP 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 inter-tumor molecular heterogeneity. Regarding intra-tumor heterogeneity, evidence from next generation sequencing has indicated that different tumor sub-clones exhibit diversified DNA mutation profiles (TOLANEY SM AND EP 2007; Nik-Zainal *et al.* 2012; Yates *et al.* 2015b). Further, 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, inter-tumor 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 of breast
cancer; that is, 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 of breast tumors
(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 (NETWORK 2012; MICHAILIDOU et al. 2015). Genome-
wide association studies have accumulatively identified approximately 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 (NETWORK 2012). However, the 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 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 estrogen receptor (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 to understand 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
(NETWORK 2012), the Curtis cohort (CURTIS et al. 2012), and the Guedj cohort (GUEDJ et al. 2012), we systematically scrutinized 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

1. Three cohorts of human breast cancer patient populations

In this study, we used existing data generated from three cohorts of human breast cancer patient populations, in order to identify and cross-validate master regulators and gene regulatory pathways in breast tumor subtypes. The data include the TCGA breast tumor cohort (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 US, 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 approximately 2,000 breast tumor samples from the UK 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 approximate 60 (Figure S1). However, in the Guedj cohort the median age shifts to approximate 30. This may partially be explained by the different clinical record systems, as age at initial diagnosis was recorded in this cohort. Regarding pathological phenotypes, such as tumor stage and grade, it is challenging to compare, as different pathological systems were used across the cohorts.

Regarding the gene expression data, we processed data, followed by data quality check. The TCGA RNA-Seq data downloaded are already processed data (level 3), in which gene expression is quantified as fragments per kilo-base 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 HJM 1988), in order to center the mean of each gene to 0 and compare only variance and covariance between genes.

2. Machine learning to select an assemble of master regulator genes
Supervised machine learning is a powerful method for feature (or gene) selection and tumor sample classification. Schematic framework of machine leaning is shown in Supplemental Figure 2. Since RNA-Seq provides a better measurement of gene expression, compared to microarray, we used the downloaded TCGA RNA-Seq data for feature selection. But findings will be validated in the other two cohorts mentioned above. We split the TCGA primary 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 Dr. K. Horsley in the University of North Carolina.

Since the first step is to systematically identify master regulators (MRs) for each subtype, we focused on two gene sets for feature selection: 1) a set of well annotated 1,400 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 gens, the 2\textsuperscript{nd} 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 \textit{FOXC1, ESR1, FOXA1, ERBB2, MYC, MDM2, PGR, BCL2, EGFR,} and \textit{MYBL2}.

Regarding feature 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 Supplemental Methods for more details.

Once the feature set is identified by the algorithm, we then train a classifier (or statistical model) based on the training data. There are multiple algorithms that can be applied to train a classifier. For example, One of them is to develop 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 will use 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 has been correctly predicted by the MR16 classifier. This is a measurement of concordance in subtype calling between the PAM50 and the MR16.

3. 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 TCGA
training data, we developed a classifier, based on the ensemble of genes selected. 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 woman cases. 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, based on 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.

4. 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 non-linear regulator-regulon interactions (BASSO et al. 2005; CARRO et al. 2010). See Supplemental Methods for more details. In order 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
(DPI) 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).

5. Identification of canonical pathway changes in tumors

In order 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.

Results

1. 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 the Method section, 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 into the four subtypes (Figure S4). Further, 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 classifiers based on the MR16. Application of the classifier to both 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). 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 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 are consistent with clinical observations that
luminal A tumors display a significantly (log rank p = 6.25x10^-5) better prognosis, compared to the other subtypes.

2. 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 non-luminal (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 sub-clusters have been observed. 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; that is, 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 non-luminal tumors (Figure 1A). Conversely, Group 2 genes, including FOXM1, EZH2, ZNF695, MYBL2, SOX11, BCL11A, and FOXC1, are highly expressed in
the non-luminal, especially the basal-like subtype, but lowly expressed in the luminal tumors (Figure 1A).

As expected, similar gene expression patterns have been observed in both the Curtis and Guedj cohorts (Figure S6). The same Group 1 MRs as defined in the TCGA cohort (Figure 1A) are over-expressed in the luminal tumors, but lowly expressed in the non-luminal tumors. In contrast, the same Group 2 MRs are over-expressed in the non-luminal tumors, but lowly expressed in the luminal tumors. The samples in the Curtis validation data display two large clusters that correspond to the ER-positive and ER-negative tumors (Figure S6A). In each of the two clusters, there are two sub-clusters. 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 non-luminal tumors (Figure S6B). In each of the two clusters, the sub-clusters 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 intrigue further investigation.

4. 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 by performing either simple correlation analysis, which is based on the linearity assumption, or 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 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 between the inter-group MRs are negative.

Since the sample sizes are large in the three cohorts of patient populations, we then applied mutation 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 indicates that interactions of the MRs 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. These implicate subtype-specific gene regulations in breast tumors, which will be scrutinized below.

5. The 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 modulating tumor growth and progression, we will focus on illustrating the interactions of the MR16 and the cell cycle genes.

Examination of the MR-cell cycle gene interactions across the four subtypes reveals interesting 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). Further, the Group 2 MRs, including FOXM1, MYBL2, EZH2, and ZNF695, cooperatively up-regulate the cell cycle 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 clearly demonstrate that MR-cell cycle genes interactions behavior 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.

6. 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 in population sizes can affect the filtering steps, which may result in 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 (Figures S9 & S10). The subtype-specific interactions implicate that the cell cycle pathways play a critical role in promoting progression of ER-negative tumors.

7. Association of DNA somatic mutations with gene regulatory pathways
Since breast cancer is a heterogeneous disease in which various sub-clones from a tumor can display different mutation profiles (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 (NETWORK 2012). For example, TP53 is one of the top three genes (TP53, PIK3CA, and GATA3) that are most frequently mutated in breast cancer (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 (NETWORK 2012).

Since 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 over-expression 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) over-expressed in both HER2 enriched and basal-like subtypes, compared to the normal samples (Figure 2C). Similarly, representative cell cycle genes are significantly over-expressed 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 HER2 enriched and basal-like subtypes, implicating canonical genomic alternations in the ER-negative breast tumors.

**Discussion**

We have identified an ensemble of 16 master regulators (MR16) as an entity that has the maximum prediction power in the training data. In terms of subtype classification and survival prediction, it 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 non-overlapping genes. 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 over-expressed 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 the HER2 enriched and basal like tumors. However, the three genes are missed in the PAM50. Furthermore, results from mutual information analyses 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. Further, 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; that is, MR-MR interactions and MR-regulon interactions.

Regarding MR-MR interactions, the intra-group MRs interact in a hierarchical and synergistic fashion. Recent studies from 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 transcription factor (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 intra-group MR-MR interactions, the inter-group 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. (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 undifferentiated state of luminal progenitors.

More in-depth studies are deserved to elucidate the intra- and inter-group 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 instance, 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 indicate that cell cycle pathways are over-expressed in the two subtypes, but not in the luminal A and B subtypes. Further, the subtype-specific amplification of the cell cycle pathways coincides with the subtype-specific TP53 mutations. Therefore, it is implicated 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. 2015a), 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.

Contributions of authors: RL – study design, data analysis, and paper writing; JC - data process and management; JI - collaboration in cancer biology.

Acknowledgments: The authors thank the TCGA network, Dr. C. Curtis and the collaborative consortium, and Dr. M. Guedj and colleagues for generating the data.

Figure Legend
Figure 1. Patterns of master regulator 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: the ER-negative and ER-positive. Two sub-clusters are observed within each of the clusters. The four clusters correspond well to the four 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 MR16 that display distinct gene 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.
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 tumor sample 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 each of the four subtypes. B: On the basis of the same set of samples from the Guedj data (Guedj et al. 2012), we used the PAM50 method (Parker et al. 2009) to directly call subtypes, which resulted in five subtypes. Corresponding Kaplan-Meier curves are plotted. Comparisons of the two sets of Kaplan-Meier curves indicate that the MR16 has a cross-cohort prediction power similar to the direct prediction by the PAM50. C and D: Association of TP53 gene mutations with significant over-expression of cell cycle pathway genes in ER-negative tumors.
Using tumor samples with TP53 mutations and matched normal samples without TP53 mutations, we compare expression of the representative MR 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 color = Basal-like tumor samples; Green = HER2 enriched tumors; Blue = normal samples.

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 circle represent target genes (regulons) and the edges connecting the master regulators (MR) 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 the 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 master
regulators 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, such as *CCNB1* and *CDK1*, are up-regulated by the central MRs. However, Group 1 MRs, such as *ESR1* and *AR*, are down-regulated by the central MRs.

**Supplemental Methods**

1. **Random Forest algorithm**
   Random Forest is a decision tree-based supervised machine learning algorithm. Each tree is constructed using a different bootstrap sample from the training data. In each time, one-third of the samples (cases) are left out of the bootstrap sample and not used in the construction of the kth tree. This oob (out-of-bag) data is used to get a running unbiased estimate of the classification error as trees are added to the forest. This is a built-in step for cross-validation. It is also used to get estimates of feature (or gene) importance.

   Regarding feature importance, in every tree grown in the forest, we checked the oob samples and counted the number of votes cast for the correct class. Each tree in the forest has a vote. We then randomly permuted the values of feature m in the oob samples and put these samples down the tree. By subtracting the number of votes for the correct class in the feature-m-permuted oob data from the number of votes for the correct class in the untouched oob data, the average of this number over all trees in the forest is the raw importance score for feature m.

2. **Feature selection instability**
   In machine learning, feature (gene) selection instability is an issue, which means different sets of features can be selected based on different subsets of the samples in the training data. We then partitioned the entire TCGA samples downloaded into the training and test subsets repeatedly (100 times). Between
the runs of the algorithm on the different training data, various numbers of genes may be selected. The algorithm often selects less than 20 genes, rarely more than 20 genes, and very rarely more than 40 genes. Therefore, in each run, we retained the top genes up to 40. Gene frequency was then computed, based on the collection of genes being selected across the 100 runs.

Since the Random Forest algorithm has a built-in step for cross-validation, there is no need to perform additional cross-validation on the same data. However, cross-validation using independent data sets is critical.

3. Mutual information
Since gene expression levels are continuous variables, the range of each gene expression across a set of samples can be divided into multiple bins. The differential entropy, which averages the log-probability density, will then be applied (Steuer et al. 2002). We take a rank transformation on the data and then project the measurements of each gene into equispaced real numbers in the interval [0, 1], preserving their original order. This transformation has the advantage of transforming the probability density of the individual genes (x and y) into a constant, \( p(x') = p(y') = 1 \). Under this transformation, the entropy for both genes, \( S(x') \) and \( S(y') \), becomes constant and equals to zero. As a result, only the joint entropy \( S(x', y') \) needs to be estimated. A Gaussian Kernel estimator (Steuer et al. 2002) will be used to compute the mutual information between genes x and y:

\[
I(x', y') = \frac{1}{M} \sum \log \left| \frac{p(x', y')}{p(x')p(y')} \right|
\]

where M is the sample size.
Supplemental Table 1

<table>
<thead>
<tr>
<th>Type</th>
<th>PIK3CA</th>
<th>TP53</th>
<th>GATA3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame_Shift_Del</td>
<td>0</td>
<td>33</td>
<td>14</td>
</tr>
<tr>
<td>Frame_Shift_Ins</td>
<td>2</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>In_FRAME_Del</td>
<td>13</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>In_FRAME_Ins</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missense</td>
<td>273</td>
<td>158</td>
<td>6</td>
</tr>
<tr>
<td>Nonsense</td>
<td>0</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>Silent</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Splice_Site</td>
<td>0</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>293</td>
<td>265</td>
<td>86</td>
</tr>
</tbody>
</table>

* 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/).

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. The gene panel used for gene search in this study 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). We used the downloaded TCGA RNA-Seq data for gene selection. In machine learning, we need to split the entire 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). Once we have identified a small ensemble of genes that can best discriminate the four subtypes of breast tumors, we cross-validate 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.
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 in the heatmaps 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: MR16 gene expression pattern 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 to the four major subtypes. B: MR16 gene expression pattern 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. Pearson correlation analyses between representative master regulator genes by subtype. We used simple correlations to explore the relationships among representative six master regulator genes.
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 nodes in circle represent regulons and the edges connecting the master regulators and the regulons indicate interactions between them. Red and blue colors denote up- and down-regulation, respectively. A: This network is based on the TCGA RNA-Seq data from 145 primary tumor samples of the luminal B subtype. Green nodes are highlighted for either cell cycle genes or Group 2 MR. Group 2 MR, such as ZNF695 and SOX11, are down-regulated by the central group 1 MRs, as are many cell cycle genes, such as CCNB1 and CDK1. B: This network is derived from the TCGA RNA-Seq data from 111 primary tumor samples of the basal-like subtype. The cell cycle genes, including PLK1 and CDK1, are up-regulated by the central Group 2 MR genes; that is, FOXM1, EZH2, ZNF695, and MYBL2.
Figure S9. Validation of regulator-regulon interaction patterns in the Curtis cohort. The networks are based on the Curtis validation data (CURTIS et al. 2012). See Figure S8 for nodes and edges. **A**: a gene regulatory network specific to the luminal B subtype. The central genes are the Group 1 MRs. 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 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 regulons. Green nodes are highlighted for either cell cycle genes or Group 2 MRs. Other Group 1 MR, such as AR, are up-regulated by the central regulators. However, Group 2 MR, such as FOXM1, and the cell cycle genes, such as CCNA2 and CCNB2, are down-regulated by the central MR.

**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. The cell cycle genes, such as CDK1 and CCNA2, are up-regulated by the central MRs. However, Group 1 MRs, such as GATA3 and FOXA1, are down-regulated by the central MRs.
Figure S11. The relationships between MR16 and the PAM50.

The central genes are the two MR groups, as indicated by purple (Group 1) and green (Group 2). Results from mutual information indicate that the majority of PAM50 genes are target genes regulated by the MR16, except for the overlapping genes. See Figure S8 for nodes and edges.
References


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. Molecular and Cellular Biology 33: 227-236.


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.


Kwok, J., B. Peck, L. Monteiro, Schwenen HD, J. Millour et al., 2010 FOXM1 confers acquired cisplatin resistance in breast cancer cells. Mol Cancer Res. 8: 24-34.


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.