11-7-2018

Decipher Mechanisms by which Nuclear Respiratory Factor One (NRF1) Coordinates Changes in the Transcriptional and Chromatin Landscape Affecting Development and Progression of Invasive Breast Cancer

Jairo Ramos
jramo033@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

DECIPHER MECHANISMS BY WHICH NUCLEAR RESPIRATORY FACTOR ONE (NRF1) COORDINATES CHANGES IN THE TRANSCRIPTIONAL AND CHROMATIN LANDSCAPE AFFECTING DEVELOPMENT AND PROGRESSION OF INVASIVE BREAST CANCER

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PUBLIC HEALTH

by

Jairo Ramos

2018

To: Dean Tomás R. Guilarte
Robert Stempel College of Public Health and Social Work

This dissertation, written by Jairo Ramos, and entitled Decipher Mechanisms by which Nuclear Respiratory Factor One (NRF1) Coordinates Changes in the Transcriptional and Chromatin Landscape Affecting Development and Progression of Invasive Breast Cancer, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read the dissertation and recommend that it be approved.

Alok Deoraj

Juan Luizzi

Quentin Felty

Deodutta Roy, Co-Major Professor

Changwon Yoo, Co-Major Professor

Date of Defense: November 7, 2018

The dissertation of Jairo Ramos is approved.

Dean Tomás R. Guilarte
R. Stempel College of Public Health and Social Work

Andres G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2018

DEDICATION

This dissertation is dedicated to the memory of my parents, Raul Ramos and Elinor Vega de Ramos, who always encouraged me to pursue my dreams through education; to the memory of my sister, Meibis Ramos, who was always there for me through the good and bad times; and to my wife, Mile, and my two sons, Jairo and Andres, who were my inspiration to pursue my doctoral degree.

ACKNOWLEDGMENTS

I would like to thank first my major professor, Dr. Deodutta Roy, for his continuous support and patience and for sharing with me all his knowledge in the fields of genetics, epigenetics, and carcinogenesis. I have learned a great deal from him. The research and writing of this dissertation would have not been possible without his help and guidance.

I wish to thank my co-major professor, Dr. Changwon Yoo, for his assistance and support in the field of biostatistics and for facilitating the tools (hardware and software) I needed for the statistical analysis section of my research.

I also want to express my gratitude to the other members of my dissertation committee: Dr. Quentin Felty, Dr. Alok Deoraj, and Dr. Juan Liuzzi, for their encouragement, comments, and suggestions that helped me to complete successfully this dissertation.

I thank too my fellow students from the Environmental Health Science Department and from the SMLG group for the stimulating discussions and camaraderie and for sharing with me their knowledge in different fields of health sciences and statistics. My thanks also to Marie Alvarez, our former departmental secretary, who helped me greatly with all the administrative matters and the use of the university’s resources.

My sincere gratitude goes to the Starr Foundation for its generosity and to the committee members at Stempel College who selected me as one of the recipients of the STARR scholarship in 2016 and 2017.

Last but not least, I would like to thank my brother, Dr. Ariel Ramos, and my nephew, Larry Ramos, for their continuous support and encouragement while I pursued my dream of becoming a scientist.

ABSTRACT OF THE DISSERTATION

DECIPHER MECHANISMS BY WHICH NUCLEAR RESPIRATORY FACTOR ONE (NRF1) COORDINATES CHANGES IN THE TRANSCRIPTIONAL AND CHROMATIN LANDSCAPE AFFECTING DEVELOPMENT AND PROGRESSION OF INVASIVE BREAST CANCER

by

Jairo Ramos

Florida International University, 2018

Miami, Florida

Professor Deodutta Roy, Co-Major Professor

Professor Changwon Yoo, Co-Major Professor

Despite tremendous progress in the understanding of breast cancer (BC), gaps remain in our knowledge of the molecular basis underlying the aggressiveness of BC and BC disparities. Nuclear respiratory factor 1 (NRF1) is a transcription factor (TF) known to control breast cancer cell cycle progression. DNA response elements bound by NRF1 positively correlate with the progression of malignant breast cancer. Mechanistic aspects by which NRF1 contributes to susceptibility to different breast tumor subtypes are still not fully understood. Therefore, the primary objective of this dissertation was to decipher mechanisms by which NRF1 coordinates changes in the transcriptional and chromatin landscape affecting development and progression of invasive breast cancer. Our hypothesis was that NRF1 reprogramming the transcription of tumor initiating gene(s) and tumor suppressor gene(s) contribute in the development and progression of invasive breast cancer. To test this hypothesis, we proposed three specific aims: (a) Decipher regulatory landscape of NRF1 networks in breast cancer. (b) Determine the role of NRF1 gene networks in different subtypes of breast cancer. (c) Determine differential NRF1 gene network sensitivity contributing to breast cancer disparities. Our approach to test these aims consisted of a systematic integration of ChIP DNA-seq, RNA-Seq, NRF1 protein-DNA motif binding, signal pathway analysis, and Bayesian machine learning. We uncovered a novel oncogenic role for NRF1. This discovery strongly supported the supposition that NRF1 overexpression is sufficient to derive breast tumorigenesis. We also observed new roles for NRF1 in the acquisition of breast tumor initiating cells, regulation of epithelial to mesenchymal transition (EMT), and invasiveness of BC stem cells. Furthermore, through the use of Bayesian network structure learning we found that the NRF1 motif was enriched in 14 associated with HER2 amplified breast cancer. Three genes—GSK3B, E2F3, and PIK3CA—were able to predict HER2 breast tumor status with 96% to100% confidence. The findings of this study also showed the roles of NRF1 sensitivity to development of lobular A, Her2+, and TNBC in different racial/ethnic groups of breast cancer patients. In summary, our study revealed for the first time the role of NRF1 in the pathogenesis of invasive BC and BC disparities.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

Chapter I

Chapter II

Chapter III

Chapter IV

Chapter V

LIST OF FIGURES

FIGURE PAGE

Chapter II

Chapter III

Chapter IV

Chapter V

ABBREVIATIONS AND ACRONYMS

AA	African American
AUC	Area under the curve
BC	Breast cancer
BDe	Bayesian Dirichlet score
BMA	Bayesian modeling averaging
BN	Bayesian network
ChIP DNA-seq	Chromatin immunoprecipitation (ChIP) followed by DNA sequencing
ChIP on Chip	Chromatin immunoprecipitation (ChIP) with DNA microarray (chip)
ChIP-DSL	Chromatin immunoprecipitation (ChIP) with a DNA ligation and selection (DSL)
ChIP-Microarray	Chromatin immunoprecipitation (ChIP) with DNA microarray (chip)
COSMICS	The catalogue of somatic mutations in cancer
DAG	Direct acyclic graphs
DAVID	Database for Annotation, Visualization and Integrated Discovery
DE	Differentially expressed
EA	European American
EMT	Epithelial to Mesenchymal Transition
ENCODE	Encyclopedia of DNA Elements
ER	Estrogen receptor
FE	Fold enrichment

FPKM	Fragments Per Kilobase Million
FPR	False positive rate
GEO	Gene Expression Omnibus
GGOTF	Generic gene ontology term finder
GO	Gene ontology
GSE	Gene set enrichment
HER2	Human epidermal growth factor receptor 2
HMEC	Human mammary epithelial cells
IHC	Immunochemistry
KEGG	Kyoto Encyclopedia of Genes and Genomes
NCBI	National Center for Biotechnology Information
NHW	Non-Hispanic white
NRF1	Nuclear Respiratory Factor 1
OXPHOS	Oxidative phosphorylation
PR	Progesterone receptor
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic curve
ROS	Reactive oxygen species
RPKM	Reads per Kilobase Million
RR	Relative risk
SRA	Sequence Read Archive
TCGA	The Cancer Genome Atlas
TF	Transcription factor

TFTEA	Transcription factor target enrichment analysis
TMA	Tissue microarrays
TNBC	Triple negative breast cancer
TPM	Transcripts per Kilobase Million
TSS	Transcription start site

CHAPTER I

INTRODUCTION

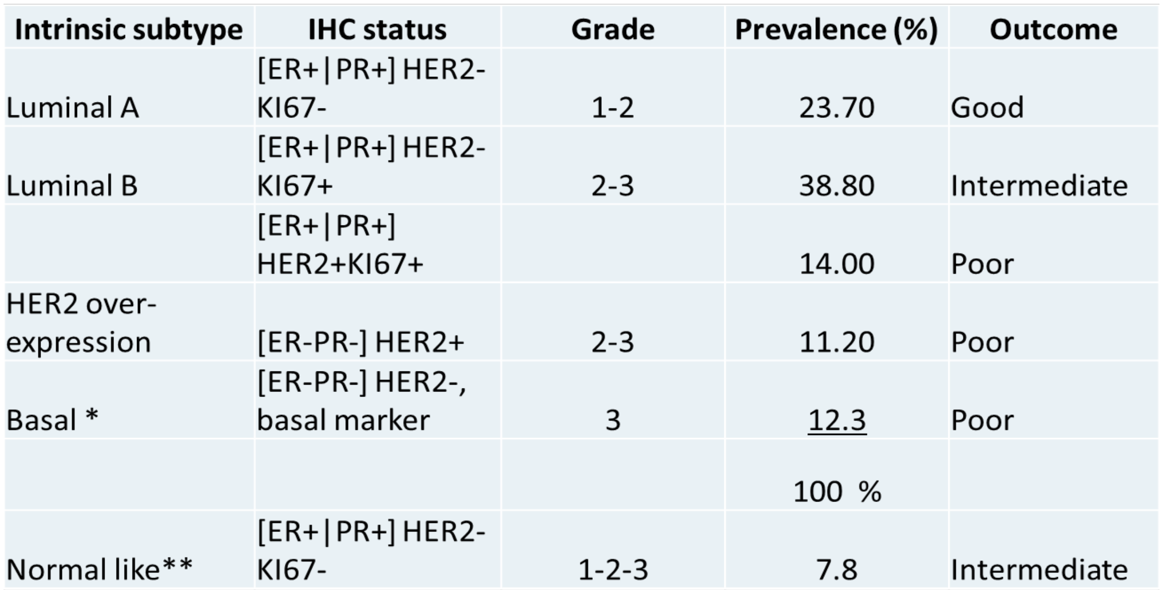
Current projections show that during 2018 approximately 878,980 women in the United States will be diagnosed with malignant tumors. The number of projected new cases of breast cancer is 266,120 (30.3%), ranking number one followed by cancer in the digestive system with 137,200 (15.61%) (Siegel, Miller, & Jemal, 2018). The estimated number of female deaths due to cancer during the same year is 286,010, with breast cancer the third most important cause with 40,920 (14.31%). This number is surpassed only by cancer of the respiratory system (larynx, lung, and bronchus, and other respiratory organs) at 71,570 (25.02%) and cancer in the digestive system organs at 66,590 (23.28%) (Siegel et al., 2018).

Breast cancer was previous thought to be homogenous; however, in the decade of 2000 to 2010, scientists realized it was a heterogeneous disease (Anders & Carey, 2009). Based on classical immunochemistry (IHC) markers [estrogen receptor (ER), progesterone receptor (PR) and Human epidermal growth factor receptor 2 (HER2)] and patterns of gene expression (DNA microarrays), five subtypes of breast cancer have been identified: Luminal A, Luminal B, Human epidermal growth factor receptor 2 (HER2) enriched, Basal-like, and Normal breast-like (Dai et al., 2015; Yuan et al., 2014). Tumors with negative hormone receptor and HER2 status (ER-, PR- HER2-) are known as triple negative (TNBC).

Although any of the subtypes can be triple negative, most (71% to 91%) are Basal-like (Huang,Liu, Chen, Liu, & Shao, 2013). Tumor response to treatment does not depend on anatomical factors but rather on detailed expression profiles (Dai et al., 2015). Her2 enriched and triple negative subtypes are the two most aggressive and have the worst prognostic (Lee, Oprea-Ilies, & Saavedra, 2015; Sorlie et al., 2001). Statistics were not consistent when we searched for prevalence of breast cancer by molecular subtype in the United States. Therefore, the figures presented here should be taken with caution. Table 1, adapted from Dai et al. (2015), summarizes our searching results for classification, prevalence, and predicted outcome of breast tumors (Cheang et al., 2009).

Table 1

Breast Cancer Intrinsic Subtype Classification, IHC Status, Prevalence, and Predicted Outcome



Intrinsic subtype	IHC status	Grade	Prevalence (%)	Outcome
Luminal A	[ER+|PR+] HER2-KI67-	1-2	23.70	Good
Luminal B	[ER+|PR+] HER2-KI67+	2-3	38.80	Intermediate
	[ER+|PR+] HER2+KI67+		14.00	Poor
HER2 over-expression	[ER-PR-] HER2+	2-3	11.20	Poor
Basal *	[ER-PR-] HER2-, basal marker	3	12.3	Poor
			100 %	
Normal like**	[ER+|PR+] HER2-KI67-	1-2-3	7.8	Intermediate

*Even though any of the subtypes can be triple negative, most (71% to 91%) are Basal-like (Huang et al., 2013).

**Normal-like is part of Luminal A as it shares similar IHC status.

Note: Table adapted from Dai et al. (2015).

Incidence, prevalence, and death rates vary depending not only on molecular subtypes but also on ethnic groups. African American women are diagnosed at younger ages with more advanced stage tumors and more aggressive histologic features than Non-Hispanic White women. Risk of recurrence is also higher and survival rates are lower after controlling for age and stage at diagnosis (Danforth, 2013; Vidal, Bursac, Miranda-Caroni, White-Means, & Starlard-Davenport, 2017). It is believed that biological and nonbiological factors may contribute to these disparities. Although nonbiological factors, such as access to health care, comorbidities, mammography and cultural issues, have been studied extensively, there is a lack of understanding of biological differences in breast tumors that may explain these disparities (Chlebowski et al., 2005; Danforth, 2013).

Genetic alterations account for only 5% to 10% of all breast cancer cases and include mutations in widely known high risk genes, such as BRAC1 and BRAC2 (Kleibl & Kristensen. 2016). In general, cancer development is a multistep process caused by alterations in the expression or biochemical functions of certain genes that lead normal human cells to a progressive transformation into malignant cells (Hanahan & Weinberg, 2000). The main objective of cancer research is to identify causal genes to create new methods of diagnosis and treatment (Furney, Higgins, Ouzounis, & Lopez-Bigas, 2006).

Transcription factors (TFs) play an important role in the regulation of gene expression in multicellular genomes. Together with microRNAs, TFs are the most abundant of all regulatory factors that affect gene expression (Hobert, 2008). Currently several transcription factors have been identified as oncogenes or tumor suppressor genes, such as the very well-known P53 linked to different type of cancers (Falco, Bleda, Carbonell-Caballero, & Dopazo, 2016; Strano et al., 2007).

Nuclear Respiratory Factor one (NRF1) is a redox-sensitive pioneer transcription factor (also known as ALPHA-PAL) that regulates several genes essentials in different cellular processes, such as mitochondrial functions, RNA degradation, cell cycle DNA repair, and apoptosis (Cam et al., 2004; Satoh, Kawana, & Yamamoto, 2013; Scarpulla, 2008). Embryonic stem cells have been shown to have approximately 33% of all active genes bound by NRF1 (ENCODE Project Consortium, 2012). NRF1 activity has been found increased in several cancers, including breast tumors (Falco et al., 2016) and also linked to metastasis and poor overall survival in breast cancer patients (Ertel et al., 2012). In spite of growing evidence of NRF1 involvement in breast cancer, the underlying mechanisms are not yet fully understood.

Overall Goal

The overall goal of this dissertation was to decipher mechanisms by which nuclear respiratory factor 1 (NRF1) coordinates changes in the transcriptional and chromatin landscape affecting development and progression of invasive breast cancer, especially in the most aggressive subtypes. These subtypes are of major concern because they are associated with increased recurrence, lower survival, and higher rates of metastasis to the brain compared to other subtypes (Wu et al., 2016). Despite tremendous progress in the understanding of breast cancer, gaps still remain in our knowledge of the molecular basis underlying these disparities in aggressiveness and outcomes associated with molecular subtypes and ethnicity. Therefore, filling these gaps may lead to discovery of novel causal genes which can be the basis for the development of new therapies for treating and preventing the most aggressive breast tumors.

Hypothesis and Specific Aims

Hypothesis: NRF1 reprogramming of the transcription of tumor initiating gene(s) and tumor suppressor gene(s) contribute to the development and progression of invasive breast cancer.

Aim 1: Decipher regulatory landscape of NRF1 networks in breast cancer.

Aim 2: Determine the role of NRF1 gene networks in different subtypes of breast cancer.

Aim 3: Determine differential NRF1 gene network sensitivity contributing to breast cancer disparities.

REFERENCES

Anders, C. K., & Carey, L. A. (2009). Biology, metastatic patterns, and treatment of patients with triple-negative breast cancer. *Clinical Breast Cancer,* 9 Suppl 2, S73-S81.

Cam, H., Balciunaite, E., Blais, A., Spektor, A., Scarpulla, R. C., Young, R., . . . Dynlacht, B. D. (2004). A common set of gene regulatory networks links metabolism and growth inhibition. *Molecular Cell, 16*(3), 399-411.

Cheang, M. C., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., . . . Perou, C. M. (2009). Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *Journal of the National Cancer Institute, 101*(10), 736-750.

Chlebowski, R. T., Chen, Z., Anderson, G. L., Rohan, T., Aragaki, A., Lane, D., . . . Adams-Campbell, L. L. (2005). Ethnicity and breast cancer: Factors influencing differences in incidence and outcome. *Journal of the National Cancer Institute, 97*(6), 439-448.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research, 5*(10), 2929-2943.

Danforth, D. N., Jr. (2013). Disparities in breast cancer outcomes between Caucasian and African American women: A model for describing the relationship of biological and nonbiological factors. *Breast Cancer Research, 15*(3), 1-13.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74.

Ertel, A., Tsirigos, A., Whitaker-Menezes, D., Birbe, R. C., Pavlides, S., Martinez-Outschoorn, U. E., . . . Lisanti, M. P. (2012). Is cancer a metabolic rebellion against host aging? in the quest for immortality, tumor cells try to save themselves by boosting mitochondrial metabolism. *Cell Cycle, 11*(2), 253-263.

Falco, M. M., Bleda, M., Carbonell-Caballero, J., & Dopazo, J. (2016). The pan-cancer pathological regulatory landscape. *Scientific Reports, 6*, 1-13.

Furney, S. J., Higgins, D. G., Ouzounis, C. A., & Lopez-Bigas, N. (2006). Structural and functional properties of genes involved in human cancer. *BMC Genomics, 7*(3), 1-11

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell, 100*(1), 57-70.

Hobert, O. (2008). Gene regulation by transcription factors and microRNAs. *Science, 319*(5871), 1785-1786.

Huang, L., Liu, Z., Chen, S., Liu, Y., & Shao, Z. (2013). A prognostic model for triple-negative breast cancer patients based on node status, cathepsin-D and ki-67 index. *PloS One, 8*(12), 1-7.

Kleibl, Z., & Kristensen, V. N. (2016). Women at high risk of breast cancer: Molecular characteristics, clinical presentation and management. *Breast, 28*, 136-144.

Lee, M., Oprea-Ilies, G., & Saavedra, H. I. (2015). Silencing of E2F3 suppresses tumor growth of Her2+ breast cancer cells by restricting mitosis. *Oncotarget, 6*(35), 37316-37334.

Satoh, J., Kawana, N., & Yamamoto, Y. (2013). Pathway analysis of ChIP-seq-based NRF1 target genes suggests a logical hypothesis of their involvement in the pathogenesis of neurodegenerative diseases. *Gene Regulation and Systems Biology, 7*, 139-152.

Scarpulla, R. C. (2008). Transcriptional paradigms in mammalian mitochondrial biogenesis and function. *Physiological Reviews, 88*(2), 611-638.

Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians, 68*(1), 7-30.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., . . . Thorsen, T. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America, 98*(19), 10869-10874.

Strano, S., Dell'Orso, S., Di Agostino, S., Fontemaggi, G., Sacchi, A., & Blandino, G. (2007). Mutant p53: An oncogenic transcription factor. *Oncogene, 26*(15), 2212-2219.

Vidal, G., Bursac, Z., Miranda-Carboni, G., White-Means, S., & Starlard-Davenport, A. (2017). Racial disparities in survival outcomes by breast tumor subtype among African American women in Memphis, Tennessee. *Cancer Medicine, 6*(7), 1776-1786.

Wu, X., Baig, A., Kasymjanova, G., Kafi, K., Holcroft, C., Mekouar, H., . . . Muanza, T. (2016). Pattern of local recurrence and distant metastasis in breast cancer by molecular subtype. *Cureus, 8*(12), e924.

Yuan, N., Meng, M., Liu, C., Feng, L., Hou, L., Ning, Q., . . . Zhao, X. (2014). Clinical characteristics and prognostic analysis of triple-negative breast cancer patients. *Molecular and Clinical Oncology, 2*(2), 245-251.

CHAPTER II

CURRENT KNOWLEDGE OF NRF1 INVOLVEMENT IN THE PATHOGENESIS OF BREAST CANCER, INCLUDING GENE ONTOLOGY AND PATHWAY ANALYSIS OF NRF1 REGULATED NETWORKS

Abstract

Current projections show that approximately 266,120 women in the United States will be diagnosed with breast cancer in 2018, the highest number among all types of cancer. Hormone therapy, advances in the identification of tumor genetic profile, and the advent of targeted therapy such as Trastuzumab have increased the overall survival of breast cancer patients. In spite of these advances, the molecular risk factors involved in the pathogenesis of breast cancer are still not completely understood. Nuclear respiratory factor 1 (NRF1), also known as Alpha-palindromic binding protein (ALPHA-PAL), is a transcription factor (TF) known to be involved in cellular processes important in cancer development. These include RNA degradation, cell cycle, DNA replication, DNA repair, mitosis, and apoptosis. NRF1 activity has been associated with breast cancer development in multiple ways and poor outcomes among breast cancer patients. We performed a literature review searching for current knowledge about mechanisms of NRF1 involvement in breast cancer, ChIP-Seq computational analysis to identify NRF1 target genes, and Gene Ontology and Pathway Analysis of NRF1 regulatory network to investigate its participation in signaling pathways and cellular processes important for cancer biology. We found growing evidence that NRF1 may be involved in breast cancer through different mechanisms, including the increase of

mitochondrial function to support proliferation of cancer cells, the increase of NRF1 activity due to estrogen-induced ROS signaling, which in turn dysregulates cell cycle genes, and epigenetic changes affecting NRF1 binding such as DNA methylation. Identification of NRF1 targets demonstrated that NRF1 network is cell-context- dependent, suggesting that these dissimilarities may help to elucidate differences in breast tumor behavior among molecular subtypes. We also found that the KEGG breast cancer pathway was enriched with NRF1 target genes. Finally, we noticed that a high percentage of the well-known breast cancer genes were directly or indirectly regulated by NRF1, including the very well-known BRCA1 that seems to be regulated by a transcriptional network formed by GABP and NRF1 (NRF-1 > GABPβ > BRCA1).

Introduction

In 2018, the projected number of women in the United States diagnosed with cancer is 878,980. Breast cancer is expected to rank number one, with 266,120 new cases accounting for 30.3% (Siegel, Miller, & Jemal, 2018). Identification of gene expression patterns in tumors has been one of the key elements for the advances in the treatment of this disease, with a corresponding increase in overall survival. In spite of these advances, the molecular risk factors involved in the pathogenesis of breast cancer are not completely understood.

Genetic and epigenetic alterations are involved in breast cancer development and progression (Campoy et al., 2016; Hanahan & Weinberg, 2011). Epigenetic alterations include DNA methylation and variations in chromatin, histone, and regulatory RNA (Campoy et al., 2016). Genetic alterations account for 5% to 10% of all breast cancer cases and include mutations in widely known high-risk genes, such as BRAC1 and BRAC2 (Kleibel & Kristensen, 2016). The human gene Nuclear respiratory factor 1 (NRF1), also known as Alpha-palindromic binding protein (ALPHA-PAL), is a transcription factor (TF). This factor regulates the expression of a number of genes involved in mitochondrial functions essentials for cellular growth and development, such as organelle biogenesis and cellular respiration (Scarpulla, 2008), as well as other cellular processes involved in cancer development, such as RNA degradation, cell cycle, DNA replication, DNA repair, mitosis. and apoptosis (Cam et al., 2004; Satoh, Kawana, & Yamamoto, 2013).

NRF1 activity has been linked to breast cancer in different ways. We performed a review of the literature searching for current knowledge about mechanisms of NRF1 involvement in breast cancer. Additionally, we searched for ChIP-Seq studies attempting to identify NRF1 targets genes. Because researchers use different protocols as well as computational analysis parameters, it is difficult to compare results. However, to arrive at a better comparative approach, we took the peak calling files and unified the gene annotation method using the same software (GREAT) and keeping the same setting parameters. Finally, we carried out Gene Ontology and Pathway Analysis of the NRF1 regulatory network to investigate its participation in signaling pathways and cellular processes important for cancer biology.

Results and Discussion

NRF1, Breast Cancer, and Mitochondrial Function

Niida et al. (2008) reported that motifs bound by NRF1 were positively correlated with tumor malignancy and progression of breast cancer. Since NFR1 regulates several nuclear-encoded mitochondrial genes and increases the respiratory capacity of mitochondria (Scarpulla. 2006), Niida et al. (2008) suggested that this activity could be an indication of hypermetabolism in aggressive breast cancer. Ertel et al. (2012) found that NRF1 activity was higher in breast cancer tissue than adjacent normal tissue. They used bioinformatics analysis to show that upregulation of NRF1 target genes was associated with metastasis, recurrence, and poor overall survival. The hypothesis of Ertel et al. (2012) was that cancer cells intended to save themselves from the aging process, characterized by significant reductions in oxidative mitochondrial function, throughout the implementation of a defensive mechanism that includes amplification of the mitochondrial oxidative metabolism (OXPHOS) and overexpression of NRF1. This overexpression of NRF1 in turn causes upregulation of NRF1 target genes.

Sotgia et al. (2012) carried out research to investigate the role of epithelial mitochondrial biogenesis in malignant breast tumors, analyzing the transcriptional profiles of epithelial cancer cells and comparing them to adjacent stromal cells. The researchers found that 39 genes encoding mitochondrial ribosomal proteins (MRPs) were involved in mitochondrial translation of OXPHOS complex components, and other transcription factors associated with mitochondrial biogenesis, including NRF1, were upregulated (twofold to fivefold). Confirming the hypothesis that increased mitochondrial function plays an important role in proliferation of breast cancer cells, Jafaar et al. (2014) induced cell death in MCF-7 and LCC9 breast cancer cells lines by inhibiting NRF-1.

NRF1, Breast Cancer and Epigenetic Factors (DNA Methylation).

We mentioned previously that gene expression is affected by epigenetic factors such as DNA methylation. Campoy et al. (2016) discovered changes in the levels of DNA methylation in breast tumors, which may be linked to LSD1, one of the main cofactors of NRF1. LSD1 (lysine-specific demethylase 1) (Benner et al., 2013) is a protein encoded by the KDM1A gene that controls the level of methylation through its demethylase activity exerted by removing the methyl groups from methylated lysine 4 of histone H3 and lysine 9 of histone H3 (Lim et al., 2010).

LSD1 was found overexpressed in breast cancer tissue, especially in clinical advanced and ER- tumors. In vitro experiments decreasing or inhibiting LSD1 resulted in growth retardation of breast cancer cells (Lim et al., 2010). One possible explanation is that LSD1 switch off reduces the demethylation activity. Therefore, NRF1 binding decreases causing cell growth delay, which is aligned with the idea that NRF1 may be important for breast cancer cell proliferation.

NRF1, Estrogen, and Breast Cancer

Genetic and epigenetic factors affect the level of exposure of a specific tissue in the body to estrogen and its metabolites. Epidemiological studies have shown that lifetime exposure to estrogens is a major risk for breast cancer development (Clemons & Goss, 2001). Recent investigations support these epidemiological findings and elucidate the mechanisms of how estrogen exposure contributes to breast cancer initiation and progression.

We have shown that estrogen or its metabolites generate reactive oxygen species (ROS), which cause damage to the genome of cells. This process may be involved in cancer development (Roy, Cai, Felty, & Narayan, 2007). Estrogen also induces changes in mitochondrial reactive oxygen species (mtROS), which play an important role as signaling molecules that may alter the cell cycle probably by modifying the expression of early cell cycle genes (Felty, Singh, & Roy, 2005; Parkash, Felty, & Roy., 2006). Further in vitro testing confirmed this hypothesis, demonstrating that estrogen-induced ROS signaling increases the binding activity of NRF1. This activity in turn increases the mRNA expression of NRF1-regulated cell cycle genes CDC2, PRC1, PCNA, cyclin B1, and CDC25C, contributing in this way to the growth of MCF-7 breast cancer cells (Okoh et al., 2015).

Identification of NRF1 Target Genes

To understand the molecular mechanisms of NRF1 involvement in the pathogenesis of breast cancer, one of the key aspects is to identify the NRF1 gene networks. Transcription factor’s binding to specific genomic sites is a complex process determined by identification of features beyond the genomic signature (sequence motif). These features include epigenetic factors, transcription cofactors, cooperative DNA binding with other transcription factors, the 3-D structures and flexibility of the transcription factors, and their DNA binding sites and the interactions between them (Slattery et al., 2014).

Identification of NRF1 target genes is essential to elucidate the mechanisms of NRF1 involvement in breast cancer; previous reports indicated that the number of NRF1-regulated genes were 690 (Cam et al., 2004), until recent studies showed that the number of NRF1 target genes were 2,470 (Satoh et al., 2013). However, these studies were not based on human mammary or breast cancer cells (Table 1).

We used published ChIP-Seq and ChIP-microarray data from MCF7, T47D, and HCC1954 breast cancer cells, normal human mammary epithelial cells-HMEC and normal blood circulating monocytes to identify candidate NRF1 target genes. Some of these studies were not aimed at identifying NRF1 target genes. However, we processed the peak calling data and used the same gene annotation webserver GREAT to unify the identification of candidate target genes with the same parameters (see Table 2 for results).

Although the great majority of genes have been successfully annotated in the human genome, our knowledge of how transcription factors and other regulatory elements control gene expression in the different cell types is nevertheless very limited, including the identification of regulatory domain, which is not clearly defined (Narlikar & Ovcharenko. 2009). GREAT version 3.0.0 assigns NRF1 binding peaks regions to genes by calculating statistics and associating genomic regions with nearby genes. The Regulatory domain is defined as 5,000 bp upstream and 1,000 bp downstream of the transcription start site (TSS). This site can be extended in both directions up to a maximum of 1,000 bp, depending on the distance of the nearby gene’s regulatory domain.

Table 1

Published Chip-Microarray or Chip-Seq Studies With NRF1 Target Genes Results Found in Our Literature Search

Cell type	**Method**	**Significance Analysis in microarray or peak calling (in ChIP Seq): p- value, FDR, peak ratio, fold enrichment (FE), etc.**	**Software used to identify ChIP Seq peaks**	**Method used for gene identification- based on peak location / enrichment**	**Number of genes with NRF1 binding sites**	**Reference**
T98G Quiescent Glioblastoma cells	ChIP Microarray (ChIP-on-Chip)	p value cutoff p < 0.005	NA	13,000 proximal promoters from −700 to +200 relative to TSS were scanned. Genes considered to be significantly enriched if the median rank of their binding ratios was greater than 94%	691	Cam et al. 2004

Cell type	**Method**	**Significance Analysis in microarray or peak calling (in ChIP Seq): p- value, FDR, peak ratio, fold enrichment (FE), etc.**	**Software used to identify ChIP Seq peaks**	**Method used for gene identification- based on peak location / enrichment**	**Number of genes with NRF1 binding sites**	**Reference**
Human peripheral blood monocytes from several healthy donors	ChIP Microarray (ChIP-on-Chip)	TF binding regions identified using a sliding window approach of five probes (maximal distance of 500 bp between two neighboring probes) and the average of two independent tests.	NA	Genomic locations based on the March 2006 human reference sequence (NCBI Build 36.1). Enriched regions assigned to closest genes.	1,474	Gebhard et al. 2010

Cell type	**Method**	**Significance Analysis in microarray or peak calling (in ChIP Seq): p- value, FDR, peak ratio, fold enrichment (FE), etc.**	**Software used to identify ChIP Seq peaks**	**Method used for gene identification- based on peak location / enrichment**	**Number of genes with NRF1 binding sites**	**Reference**
		Minimum signal intensity of 0.4 (log10)				
MCF7 Breast cancer cells	ChIP DSL	*p*-value <0.0001 and False positive rate calculated experimentally of 3%	NA	Testing of approx. 20,000 human promoters between -800 bp and +200 bp relative to TSS	1,593	Benner et al (2013)
MCF7	ChIP DSL	*p value* <0.001	NA	Same	1,936	Benner et al (2013)

Cell type	Method	Significance Analysis in microarray or peak calling (in ChIP Seq): p- value, FDR, peak ratio, fold enrichment (FE), etc.	Software used to identify ChIP Seq peaks	Method used for gene identification- based on peak location / enrichment	Number of genes with NRF1 binding sites	Reference
MCF7	ChIP DSL	*p value* <0.01	NA	Same	2,435	Benner et al (2013)
MCF7 Breast cancer cells	ChIP Seq	Not specified	Not specified	Any distance, closest gene assigned. Some of them > +10kb from TSS (intragenic)	1,081	Benner et al (2013)
SK-N-SH Neuroblastoma cells	ChIP Seq	FE>= 20 and FDR<=0.01	MACS	Neighboring gene analysis within a distance of 5,000 bp upstream or downstream from peaks to 5' or 3' ends of the genes (peaks located in	2,470	Satoh et al. 2013

Cell type	**Method**	**Significance Analysis in microarray or peak calling (in ChIP Seq): p- value, FDR, peak ratio, fold enrichment (FE), etc.**	**Software used to identify ChIP Seq peaks**	**Method used for gene identification- based on peak location / enrichment**	**Number of genes with NRF1 binding sites**	**Reference**
				non-coding and uncategorized genes were omitted)		
T47D Breast Cancer cells Under hypoxic conditions (1 % O2)	ChIP Seq	p value <0.00001	MACS	Any distance, closest gene assigned. Some of them at distance greater than +10kb from TSS (intragenic)	9,678	Zhang, Wang et al., 2015

Cell type	**Method**	**Significance Analysis in microarray or peak calling (in ChIP Seq): p- value, FDR, peak ratio, fold enrichment (FE), etc.**	**Software used to identify ChIP Seq peaks**	**Method used for gene identification- based on peak location / enrichment**	**Number of genes with NRF1 binding sites**	**Reference**
HMEC Normal breast cancer cells	ChIP Seq	Not specified	Peakzilla	Peaks assigned to the closest gene TSS	9,415	Domcke et al. 2015
HCC1954 Breast cancer cells	ChIP Seq	Not specified	Peakzilla	Peaks assigned to the closest gene TSS	9,415	Domcke et al. 2015

Table 2

Published Chip-Microarray or Chip-Seq Studies With NRF1 Target Genes Results Unifying Gene Annotation (GREAT 3.0.0)

Cell type	**Method**	**Method used for NRF1 binding peaks identification**	**Peak file dataset reference**	**Method used for target gene identification**	**Number of genes with NRF1 binding sites in regulatory domain**
T98G Quiescent Glioblastoma cells	ChIP Microarray (ChIP-on-Chip)	*p value* cutoff $p < 0.005$	Cam et al. 2004	13,000 proximal promoters from −700 to +200 relative to TSS were scanned. Genes considered to be significantly enriched if the median rank of their binding ratios was greater than 94%	691
SK-N-SH Neuroblastoma cells	ChIP Seq	FE>= 20 and FDR<=0.01	Satoh et al. 2013	Neighboring gene analysis within a distance of 5,000 bp upstream or downstream from peaks to 5' or 3' ends of	2,470

Cell type	Method	Method used for NRF1 binding peaks identification	Peak file dataset reference	Method used for target gene identification	Number of genes with NRF1 binding sites in regulatory domain
		Software: MACS		the genes (peaks located in non-coding and uncategorized genes were omitted)	
Human peripheral blood monocytes from several healthy donors	ChIP Microarray (ChIP-on-Chip) (on CpG island microarrays)	TF binding regions identified using a sliding window approach of five probes (maximal distance of 500 bp between two neighboring probes) and the average of two independent tests. Minimum signal intensity of 0.4 (log10)	Gebhard et al. 2010 GEO accession: GSE16078	5,000 bp upstream and 1,000 bp downstream of the transcription start site (TSS). This gene regulatory domain may be extended up to 1,000 in both directions until reaching the basal domain of the nearest gene. GREAT 3.0.0	2,374

Cell type	**Method**	**Method used for NRF1 binding peaks identification**	**Peak file dataset reference**	**Method used for target gene identification**	**Number of genes with NRF1 binding sites in regulatory domain**
MCF7 Breast cancer cells	ChIP Seq	Not specified	Benner et al (2013)	5,000 bp upstream and 1,000 bp downstream of the transcription start site (TSS). This gene regulatory domain may be extended up to 1,000 in both directions until reaching the basal domain of the nearest gene. GREAT 3.0.0	1,767
T47D Breast Cancer cells Under	ChIP-Seq	*p value* <0.00001 Software: MACS	Zhang, Wang et al., 2015	5,000 bp upstream and 1,000 bp downstream of the transcription start site (TSS). This gene regulatory domain may be extended up to 1,000 in both directions	9,688

Cell type	Method	Method used for NRF1 binding peaks identification	Peak file dataset reference	Method used for target gene identification	Number of genes with NRF1 binding sites in regulatory domain
hypoxic conditions (1 % O2)				until reaching the basal domain of the nearest gene. GREAT 3.0.0	
HMEC Normal breast cancer cells (unmethylated genome)	ChIP Seq	Not specified Software: Peakzilla	Domcke et al. 2015	5,000 bp upstream and 1,000 bp downstream of the transcription start site (TSS). This gene regulatory domain may be extended up to 1,000 in both directions until reaching the basal domain of the nearest gene. GREAT 3.0.0	11,205

Cell type	Method	Method used for NRF1 binding peaks identification	Peak file dataset reference	Method used for target gene identification	Number of genes with NRF1 binding sites in regulatory domain
HCC1954 Breast cancer cells (unmethylated genome)	ChIP Seq	Not specified Software: Peakzilla	Domcke et al. 2015	5,000 bp upstream and 1,000 bp downstream of the transcription start site (TSS). This gene regulatory domain may be extended up to 1,000 in both directions until reaching the basal domain of the nearest gene. GREAT 3.0.0	10,909

We found that NRF1 binding activity is cell-context dependent and also influenced by other factors, such as DNA methylation and microenvironment (ex hypoxia conditions). The number of NRF1 candidate target genes we identified was 1,767 in MCF7 cells; 2,374 in human blood monocytes (CpG island microarray); 9,688 in T47D cells (hypoxia conditions); 10,909 in HCC1954; and 11,205 in HMEC cells. We also found that the absolute distance to TSS of NRF1 binding peak region-gene association was different for each cell line (Figure 1). T47d cells show a higher proportion of TSS proximity (0 to 5 kb) with 58%, followed by HCC1954, HMEC with 56%, monocytes with 7%, and finally MCF7 with 6%.

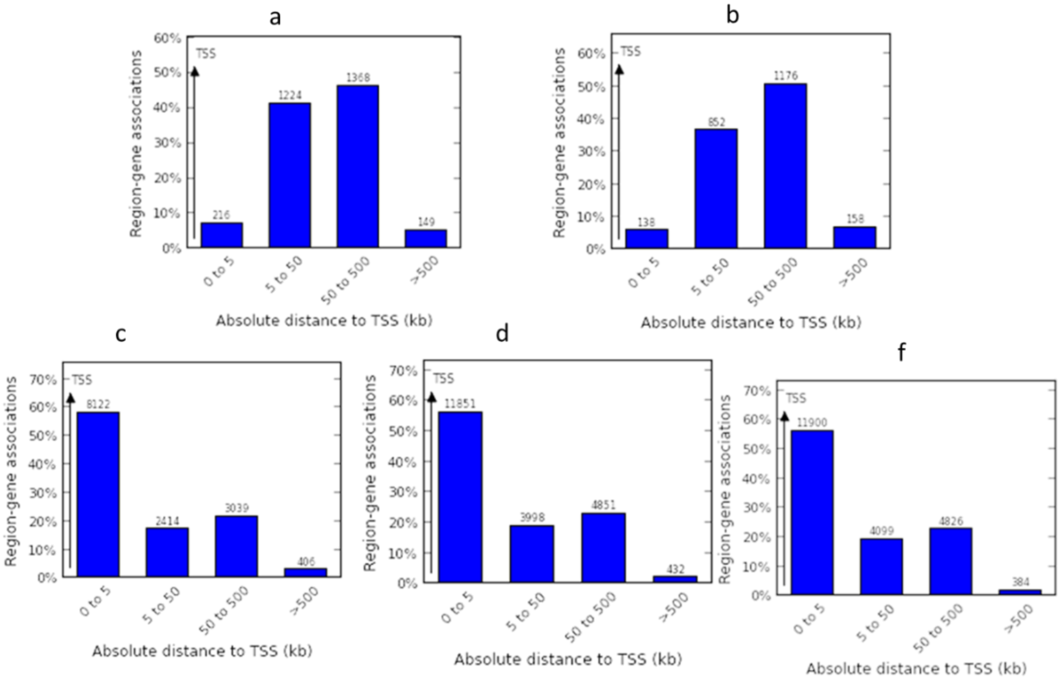


Figure 1. Absolute distance of NRF1 binding regions to TSS in different cell lines. Monocytes (a), MCF7 cells (b), T47D cells (c), HCC1954 (d), and HMEC (f).

Differences in NRF1 Network Between Breast Cancer Cell Lines and Normal Human Mammary Epithelial Cells

Molecular classification of tumors allows physicians to provide specific, targeted therapies to breast cancer patients (Eliyatkin, Yalcin, Zengel, Aktas, & Vardar, 2015). Therefore, it is important to identify differences among normal mammary cells and the different types of breast cancer. We compared the list of putative NRF1 target genes in normal Human Mammary Epithelial Cells (HMEC)—isolated from adult female breast tissue—to three breast cancer cell lines representing different molecular subtypes. These were (a) HCC1954-(breast ductal carcinoma) (ER-/PR-/HER2+) negative for expression of estrogen receptor, with amplified HER2 and high abundance of EGFR, representing well-accepted model systems of HER2-positive breast cancer (Metastatic); (b) T47D—molecularly classified as Luminal A (ER+/PR+/HER2-) with P53 mutant; and (c) MCF7-molecularly classified as Luminal A (ER+/PR+/HER2-) with P53 wild type.

We compared the four cell lines together (Figure 2, Venn diagram) and individually (Figure 3, Venn diagrams). We found 306 genes that were NRF1-regulated exclusively in MCF7 cells, 613 in T47D cells, and 395 in HCC1954 cells. These cell context differences in the NRF1 regulatory network may provide additional information of NRF1 involvement in breast cancer. Therefore, we proceeded to classify these genes using the Functional Annotation tool from DAVID (Database for Annotation, Visualization and Integrated Discovery) to find enriched KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways using a Fisher Exact P value cut off of 0.01. Results are shown in Table 3.

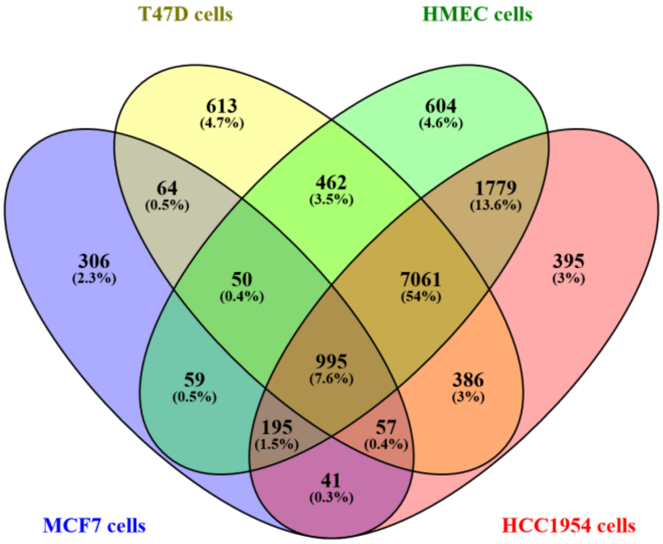


Figure 2. Venn diagram: Comparison of NRF1 network in normal Human Mammary Epithelial Cells (HMEC) with three different breast cancer cell lines. HCC1954-breast ductal carcinoma (ER-/PR-/HER2+)—negative for expression of estrogen receptor and with amplified HER2 and high abundance of EGFR- representing well-accepted model systems of HER2 positive breast cancer (Metastatic); T47D—molecularly classified as Luminal A (ER+/PR+/HER2-) with P53 mutant; and MCF7—molecularly classified as Luminal A (ER+/PR+/HER2-) with P53 wild type. This diagram was constructed with VENNY 2.1.0 (Oliveros, 2007/2015). VENNY is an interactive tool for comparing lists with Venn diagrams (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

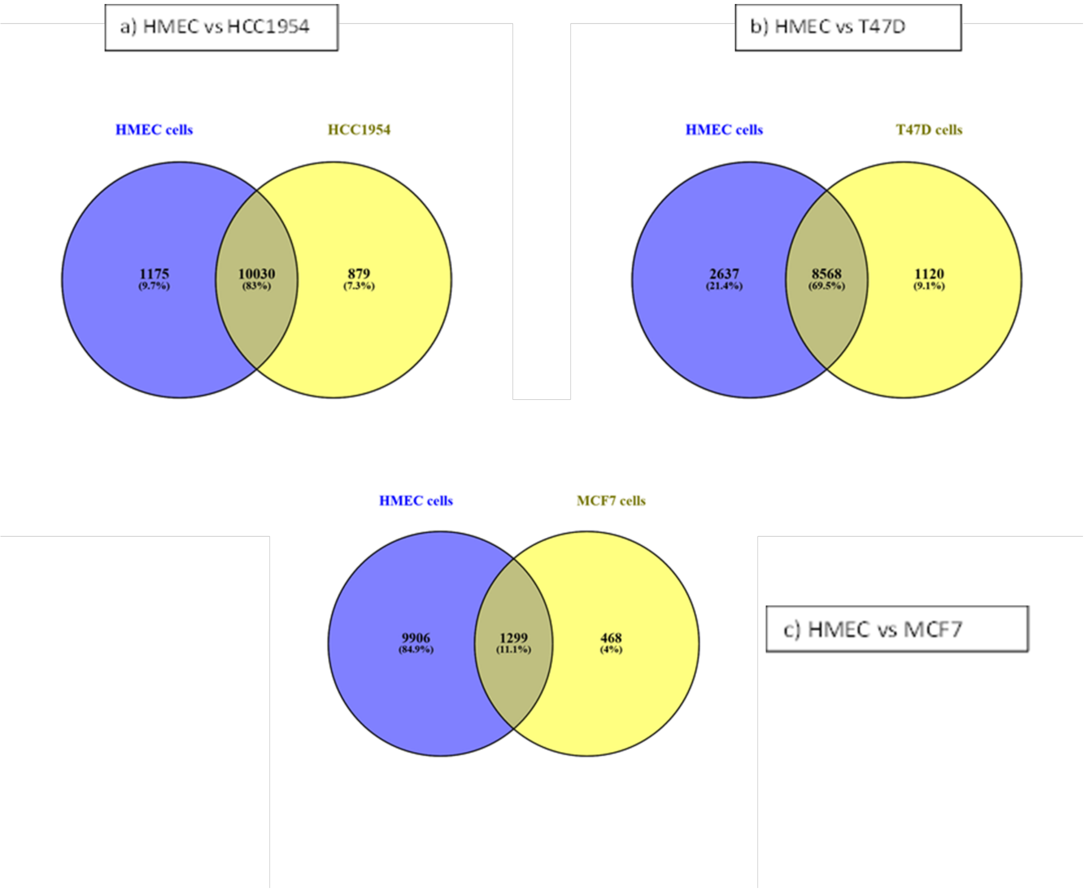


Figure 3. Venn diagrams: Individual comparison between NRF1 network in normal Human Mammary Epithelial Cells (HMEC) with three different breast cancer cell lines. HCC1954-breast ductal carcinoma (ER-/PR-/HER2+)— negative for expression of estrogen receptor and with amplified HER2 and high abundance of EGFR, representing well-accepted model systems of HER2 positive breast cancer (Metastatic); T47D—molecularly classified as Luminal A (ER+/PR+/HER2-) with P53 mutant; and MCF7—molecularly classified as Luminal A (ER+/PR+/HER2-) with P53 wild type. These diagrams were constructed using VENNY 2.1.0 (Oliveros, 2007/2015). VENNY is an interactive tool for comparing lists with Venn diagrams (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

Table 3

Functional Classification of NRF1 Target Genes Exclusively Regulated in Each One of the Different Breast Cancer Cell Lines

KEGG Pathway	**Count**	**Genes**
Pathways enriched by NRF1 target genes exclusively regulated in MCF7 cells		
Cell adhesion molecules (CAMs)	7	NRCAM, PTPRC, CD86, NFASC, CNTN2, LRRC4B, HLA-DPB1
ECM-receptor interaction	5	LAMB4, GP6, COL6A6, COL5A3, FN1
Protein digestion and absorption	5	FXYD2, COL6A6, ACE2, CPA2, COL5A3
Osteoclast differentiation	6	LILRB1, LILRA1, LILRA2, LILRA4, LILRB4, TREM2
Pathways enriched by NRF1 target genes exclusively regulated in T47D cells		
Propanoate metabolism	6	ALDH6A1, MUT, SUCLG1, ABAT, ACSS3, ACAT1
Neuroactive ligand-receptor interaction	16	GABRG3, PTGER3, GLRA2, GABBR2, VIPR2, SSTR4, AGTR1, HTR1B, P2RY6, GRM3, SSTR3, S1PR1, PRSS3, NPFFR2, ADRA1A, CALCRL
Renin-angiotensin system	4	AGTR1, KLK2, PRCP, MME
Valine, leucine and isoleucine degradation	5	ALDH6A1, MUT, ALDH2, ABAT, ACAT1
Vascular smooth muscle contraction	8	KCNU1, AGTR1, PPP1CA, BRAF, MYLK3, ADRA1A, CACNA1F, CALCRL
beta-Alanine metabolism	4	ALDH6A1, ALDH2, ABAT, DPYS
Lysosome	8	CLTB, AP1G1, AP3M1, CTSO, PPT2, NEU1, GGA1, ATP6V0D2

KEGG Pathway	**Count**	**Genes**
Aldosterone synthesis and secretion	6	PRKD1, AGTR1, KCNK9, CACNA1I, CACNA1H, CACNA1F
Pathways enriched by NRF1 target genes exclusively regulated in HCC1954 cells		
Rap1 signaling pathway	16	FGF19, FGFR4, ADCY2, DRD2, ADORA2A, GRIN2A, HGF, APBB1IP, RGS14, PRKD2, RASSF5, CNR1, ANGPT1, RAPGEF1, ANGPT2, FGF4
Complement and coagulation cascades	6	KNG1, C7, VWF, C6, BDKRB1, BDKRB2
Ras signaling pathway	11	FGF19, RASSF5, FGFR4, RASAL3, GRIN2A, ZAP70, ANGPT1, HGF, ANGPT2, PLA2G2F, FGF4
Glycine, serine and threonine metabolism	4	PGAM2, GNMT, SARDH, AGXT
Regulation of actin cytoskeleton	9	FGF19, FGFR4, DIAPH3, IQGAP3, BDKRB1, ACTN2, BDKRB2, FGF4, INSRR
Cocaine addiction	4	DRD2, PPP1R1B, TH, GRIN2A

NRF1 and Breast Cancer Genes

In our literature search, we found that several widely known hereditary genes associated with breast cancer, such as BRAC1 and BRAC2, were directly or indirectly regulated by NRF1. Mutations in BRAC1 and BRAC2 account for an important proportion of early onset breast tumors. Approximately 5% of all breast cancers are attributable to variants in these two high penetrance genes (Van der Groep, Van der Wall, & Van Diest, 2011).

There is evidence that the loss of BRCA1 can initiate a cancer stem cell that drives the formation of breast tumors. BRCA1 expression seems to be regulated by a transcriptional network formed by GABP (GA Binding Protein Transcription Factor) and NRF1 (NRF-1 > GABPβ > BRCA1). Because NRF1 and GABP also have the common role of regulating mitochondrial function, this pathway suggests a possible link between tumor initiation via disruption of stem cell maturation and the Warburg effect found in several types of tumors (dysfunctional mitochondrial metabolism) (Thompson, MacDonald, & Mueller, 2011).

Hunter et al. (2007) carried out a genome-wide association study (GWAS) in a sample of 1,145 White patients with invasive breast cancer and 1,142 controls. All patients were postmenopausal White women. The researchers genotyped 528,173 SNP's and found four variants in FGFR2(fibroblast growth factor receptor 2) that were significantly associated with breast cancer. FGFR2 is a tyrosine kinase receptor that had previously been recognized as an oncogene involved in breast tumor angiogenesis (Groose & Dickson, 2005) and was also identified as NRF1 target gene in T47D breast cancer cells by Zhang, Wang et al. (2015).

Germline BRCA1 or BRCA2 mutations are associated with a high lifetime risk of up to 60% to 85% for breast cancer (Ripperger et al., 2009). Several other genes have been identified as high-, moderate- or low-penetrance breast cancer susceptibility genes. Genes are considered to have high or moderate penetrance when at least 5% of individuals with the relevant mutations develop the disease (Ripperger, Gadzicki, Meindl, & Schlegelberger, 2009). In addition to BRCA1 and BRCA2, the list of the breast cancer high susceptibility genes includes CDH1, PTEN, STK11, and TP53 (Bonifaci et al., 2008; Kleibl & Kristensen. 2016; Pasche, 2008; Rahman, 2014; Shiovitz & Korde, 2015; van der Groep, van der Wall, & van Diest, 2011).

We found that not only hereditary breast cancer genes were NRF1 regulated but also that several other genes altered in breast cancer, such as FDXR (ferredoxin reductase), a mitochondrial flavoprotein involved in the regulation of the electron transport chain. FDXR, regulated by NRF1, and EglN2 (Prolyl Hydroxylase Domain-Containing Protein 1), have been found overexpressed in breast cancer patients compared with disease–free individuals and also positively correlated with poor prognosis in ER-positive breast tumor (Zhang, Zheng, & Zhang, 2015). The list of NRF1 targets dysregulated in breast cancer include BCL2 (Apoptosis regulator Bcl-2), an important regulator of apoptosis found overexpressed in breast tumors (Shen et al., 2005). Cancer can be described as an imbalance between cell growth and cell death; BCL2 exerts an anti-apoptotic role by encoding a protein that blocks this process (Hardwick & Soane, 2013).

Another important gene in this list that has been found to play a significant role as promoter of breast cancer metastasis is a splice variant of KLF6- denominated KLF6-SV1, whose role is the opposite of the wild KLF6 that acts as a tumor suppressor gene. KLF6-SV1 overexpression enhances cell survival, migration, and invasion and is also associated with poor survival in breast cancer patients (Hatami et al., 2013). EDN1 (ET-1), endothelin 1, has also been found overexpressed in breast carcinomas and associated with aggressiveness and invasiveness potential of premalignant breast lesions (Wulfing et al., 2004).

FOXO1 (forkhead box O1) regulates FYN, a gene overexpressed in breast cancer that promotes cell proliferation, migration, and invasion.

It is believed that upregulation of FYN induces epithelial-mesenchymal transition (EMT), a key process in cancer metastasis (Xie et al., 2016). HMGA1, the high mobility group AT–Hook 1, has been shown to drive metastatic progression in triple negative breast cancer cells (MDA-MB-231, Hs578T) by reprogramming them to stem cancer cells (Shah et al., 2013). The expression of LYN, a Src-family kinase and one of the most important gene signatures in EMT, has been associated with triple negative breast cancer and shorter overall survival in breast cancer patients during the 2- to 6-year period after surgery due to its role as a mediator of invasion and epithelial-mesenchymal transition (Choi et al., 2010).

MED1, mediator complex subunit 1, plays an important role in in breast cancer cell growth, although the exact mechanism is unknown (Hasegawa et al., 2012). SRC encodes a nonreceptor tyrosine kinase overexpressed in ductal carcinoma in situ that has been associated with tumor proliferation, invasiveness, and lower recurrence-free survival (Wilson et al., 2006). NCOA3 (SRC-3), nuclear receptor coactivator, is overexpressed in breast cancer promoting carcinogenesis through different pathways, including AKT and E2F pathways which contribute to proliferation, growth, survival , migration, invasion and metastasis of cancer cells. NCOA3 also plays a role in tumor resistance to anti hormonal therapy (Gupta et al., 2016; Johnson & O'Malley. 2012; Xu, Wu, & O'Malley, 2009).

PRDX3, peroxiredoxin 3, has been found overexpressed in breast tumor tissue compared to adjacent normal breast tissue (Karihtala, Mantyniemi, Kang, Kinnula, & Soini, 2003). It is believed that peroxiredoxin 3 plays a role in protecting breast cancer cells from cytotoxicity due to oxidative stress (McDonald, Muhlbauer, Perlmutter, Taparra, & Phelan, 2014). UBE2C, ubiquitin conjugating enzyme E2C, is essential for cell cycle progression. The levels of UBE2C mRNA have been found associated with poor disease-free survival in breast in breast cancer patients (Psyrri et al., 2012).

Several databases are available with the list of genes associated with breast cancer. We used COSMICS (the Catalogue of Somatic Mutations in Cancer), one of the world's largest database of somatic mutations in human cancer (Forbes et al., 2017) and combined the information found with the results of our literature research to create a list of 94 breast cancer genes, listed in Table 4. We found that a high percentage of them were predicted NRF1 targets (percentage depends on cell line, methods, and parameters used for target gene identification).

Table 4 shows the list of breast cancer genes classified by the Functional Annotation tool from DAVID (top 10 categories, based on key words and ranked by adjusted *p* value), and also indicates potential NRF1 regulation in different breast cancer cell lines. The top 10 categories include genes classified in important roles in cancer biology, such as tumor suppressor genes, DNA repair, apoptosis, and cell cycle.

Table 4

Functional Classification of Breast Cancer Genes and Their Classification as Potential NRF1 Targets in Three Different Breast Cancer Cell Lines (MCF7, T47D Under Hypoxia Conditions, and HCC1954), in Normal Breast Cancer Epithelial Cells (HMEC) and Blood Peripheral Monocytes of Healthy Individuals

Gene Symbol	**Gene Name**	**MCF7**	**T47D**	**HCC 1954**	**HMEC**	**Monocytes**
	Tumor suppressor					
ATM	ATM serine/threonine kinase		X	X	X	
BRCA1	BRCA1, DNA repair associated		X	X	X	
BRCA2	BRCA2, DNA repair associated			X	X	
CDKN2B	cyclin dependent kinase inhibitor 2B	X	X	X	X	
CHEK2	checkpoint kinase 2		X	X	X	
CTCF	CCCTC-binding factor		X	X	X	
DLC1	DLC1 Rho GTPase activating protein		X	X	X	
MLH1	mutL homolog 1		X	X	X	
MSH2	mutS homolog 2	X	X	X	X	
NF2	neurofibromin 2		X	X	X	
PALB2	partner and localizer of BRCA2		X			X
PBRM1	polybromo 1		X			

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
PMS1	PMS1 homolog 1, mismatch repair s. component	X	X	X	X	
PMS2	PMS1 homolog 2, mismatch repair s. component	X	X			
PTEN	phosphatase and tensin homolog		X			
RASSF1	Ras association domain family member 1	X	X	X	X	
RB1	retinoblastoma gene			X	X	
STK11	serine/threonine kinase 11		X	X	X	X
TP53	tumor protein p53					
	DNA Damage					
ATM	ATM serine/threonine kinase		X	X	X	
BRCA1	BRCA1, DNA repair associated		X	X	X	
BRCA2	BRCA2, DNA repair associated			X	X	
BRIP1	BRCA1 interacting protein C-terminal helicase 1					
CCND1	cyclin D1		X	X	X	
CHEK2	checkpoint kinase 2		X	X	X	
FANCA	Fanconi anemia complementation group A		X	X	X	
FANCC	Fanconi anemia complementation group C		X	X	X	X

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
FANCE	Fanconi anemia complementation group E		X	X	X	
FANCM	Fanconi anemia complementation group M		X	X	X	
MLH1	mutL homolog 1		X	X	X	
MSH2	mutS homolog 2	X	X	X	X	
MSH3	mutS homolog 3		X			
MSH6	mutS homolog 6		X	X	X	
NBN	Nibrin		X	X	X	
PALB2	partner and localizer of BRCA2		X			X
PMS1	PMS1 homolog 1, mismatch repair s. component	X	X	X	X	
PMS2	PMS1 homolog 2, mismatch repair s. component	X	X			
RAD50	RAD50 double strand break repair protein	X	X	X	X	
RAD51	RAD51 recombinase		X	X	X	
STK11	serine/threonine kinase 11		X	X	X	X
XRCC2	X-ray repair cross complementing 2		X	X	X	
	DNA Repair					
BRCA1	BRCA1, DNA repair associated		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
BRCA2	BRCA2, DNA repair associated			X	X	
BRIP1	BRCA1 interacting protein C-terminal helicase 1					
CHEK2	checkpoint kinase 2		X	X	X	
FANCA	Fanconi anemia complementation group A		X	X	X	
FANCC	Fanconi anemia complementation group C		X	X	X	X
FANCE	Fanconi anemia complementation group E		X	X	X	
FANCM	Fanconi anemia complementation group M		X	X	X	
MLH1	mutL homolog 1		X	X	X	
MSH2	mutS homolog 2	X	X	X	X	
MSH3	mutS homolog 3		X			
MSH6	mutS homolog 6		X	X	X	
NBN	Nibrin		X	X	X	
PALB2	partner and localizer of BRCA2		X			X
PMS1	PMS1 homolog 1, mismatch repair s. component	X	X	X	X	
PMS2	PMS1 homolog 2, mismatch repair s. component	X	X			

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HME C	Mono-cytes
RAD50	RAD50 double strand break repair protein	X	X	X	X	
RAD51	RAD51 recombinase		X	X	X	
XRCC2	X-ray repair cross complementing 2		X	X	X	
	Nucleus					
AHR	aryl hydrocarbon receptor		X	X	X	
AKT1	v-akt murine thymoma viral oncogene homolog 1			X	X	
ANG	angiogenin		X			
ATM	ATM serine/threonine kinase		X	X	X	
BAP1	BRCA1 associated protein-1		X	X	X	
BCL2	BCL2, apoptosis regulator		X	X	X	
BLM	Bloom syndrome RecQ like helicase					
BRCA1	BRCA1, DNA repair associated		X	X	X	
BRCA2	BRCA2, DNA repair associated			X	X	
BRIP1	BRCA1 interacting protein C-terminal helicase 1					
CCND1	cyclin D1		X	X	X	
CEBPG	CCAAT/enhancer binding protein gamma		X	X	X	X

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
CHEK2	checkpoint kinase 2		X	X	X	
CTCF	CCCTC-binding factor		X	X	X	
E2F1	E2F transcription factor 1		X	X	X	
EP300	300 kd E1A-Binding protein gene		X	X	X	
ERBB2	erb-b2 receptor tyrosine kinase 2		X	X	X	
ETV6	ets variant gene 6 (TEL oncogene)					
FANCA	Fanconi anemia complementation group A		X	X	X	
FANCC	Fanconi anemia complementation group C		X	X	X	X
FANCE	Fanconi anemia complementation group E		X	X	X	
FANCM	Fanconi anemia complementation group M		X	X	X	
FOXA1	forkhead box A1		X	X	X	
FOXO1	forkhead box O1		X	X	X	
GATA3	GATA binding protein 3		X	X	X	
HMGA1	high mobility group AT-hook 1		X	X	X	
HTT	huntingtin		X	X	X	
IFI16	interferon gamma inducible protein 16		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
KLF5	Kruppel like factor 5		X	X	X	
KLF6	Kruppel like factor 6		X	X	X	
LYN	LYN proto-oncogene, Src family tyrosine kinase		X	X	X	
MAP2K4	mitogen-activated protein kinase kinase 4		X	X	X	
MED1	mediator complex subunit 1		X	X	X	
MLH1	mutL homolog 1		X	X	X	
MSH2	mutS homolog 2	X	X	X	X	
MSH6	mutS homolog 6		X	X	X	
NBN	Nibrin		X	X	X	
NCOA3	nuclear receptor coactivator 3	X	X	X	X	
NF2	neurofibromin 2		X	X	X	
NFIC	nuclear factor I C		X	X	X	
OVOL2	ovo like zinc finger 2		X	X	X	
PALB2	partner and localizer of BRCA2		X			X
PBRM1	polybromo 1		X			
PMS1	PMS1 homolog 1, mismatch repair s. component	X	X	X	X	
PMS2	PMS1 homolog 2, mismatch repair s. component	X	X			

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
PTEN	phosphatase and tensin homolog		X			
RAD50	RAD50 double strand break repair protein	X	X	X	X	
RAD51	RAD51 recombinase		X	X	X	
RASSF1	Ras association domain family member 1	X	X	X	X	
RB1	retinoblastoma gene			X	X	
RECQL	RecQ like helicase					
SFN	Stratifin		X	X	X	
SRC	SRC proto-oncogene, non-receptor tyrosine kinase		X			
STK11	serine/threonine kinase 11		X	X	X	X
TOX3	TOX high mobility group box family member 3					
TP53	tumor protein p53					
TP53BP2	tumor protein p53 binding protein 2		X	X	X	
TRERF1	transcriptional regulating factor 1		X	X	X	
XRCC2	X-ray repair cross complementing 2		X	X	X	
	Phosphoprotein					
ACVR2B	activin A receptor type 2B		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
AKT1	v-akt murine thymoma viral oncogene homolog 1			X	X	
ATM	ATM serine/threonine kinase		X	X	X	
AURKA	aurora kinase A		X	X	X	
BAP1	BRCA1 associated protein-1		X	X	X	
BCL2	BCL2, apoptosis regulator		X	X	X	
BLM	Bloom syndrome RecQ like helicase					
BMPR2	bone morphogenetic protein receptor type 2		X	X	X	
BRCA1	BRCA1, DNA repair associated		X	X	X	
BRCA2	BRCA2, DNA repair associated			X	X	
BRIP1	BRCA1 interacting protein C-terminal helicase 1					
CASP8	caspase 8					
CCND1	cyclin D1		X	X	X	
CDC20	cell division cycle 20		X	X	X	X
CDH1	cadherin 1			X	X	
CHEK2	checkpoint kinase 2		X	X	X	
CTCF	CCCTC-binding factor		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HME C	Mono-cytes
DLC1	DLC1 Rho GTPase activating protein		X	X	X	
E2F1	E2F transcription factor 1		X	X	X	
EP300	300 kd E1A-Binding protein gene		X	X	X	
ERBB2	erb-b2 receptor tyrosine kinase 2		X	X	X	
ETV6	ets variant gene 6 (TEL oncogene)					
FADD	Fas associated via death domain		X	X	X	X
FANCA	Fanconi anemia complementation group A		X	X	X	
FANCE	Fanconi anemia complementation group E		X	X	X	
FANCM	Fanconi anemia complementation group M		X	X	X	
FDXR	ferredoxin reductase		X	X	X	
FGFR2	fibroblast growth factor receptor 2		X	X	X	
FOXA1	forkhead box A1		X	X	X	
FOXO1	forkhead box O1		X	X	X	
GATA3	GATA binding protein 3		X	X	X	
HMGA1	high mobility group AT-hook 1		X	X	X	
HMMR	hyaluronan mediated motility receptor	X	X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
HTT	huntingtin		X	X	X	
IFI16	interferon gamma inducible protein 16		X	X	X	
IL6ST	interleukin 6 signal transducer		X	X	X	
LGALS1	galectin 1		X	X	X	
LSP1	lymphocyte-specific protein 1					
LYN	LYN proto-oncogene, Src family tyrosine kinase		X	X	X	
MAP2K4	mitogen-activated protein kinase kinase 4		X	X	X	
MAP3K1	mitogen-activated protein kinase kinase kinase 1		X	X	X	
MAP3K5	mitogen-activated protein kinase kinase kinase 5		X	X	X	
MED1	mediator complex subunit 1		X	X	X	
MLH1	mutL homolog 1		X	X	X	
MSH2	mutS homolog 2	X	X	X	X	
MSH3	mutS homolog 3		X			
MSH6	mutS homolog 6		X	X	X	
NBN	Nibrin		X	X	X	
NCOA3	nuclear receptor coactivator 3	X	X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
NF2	neurofibromin 2		X	X	X	
NFIC	nuclear factor I C		X	X	X	
NTRK3	neurotrophic tyrosine kinase, receptor, type 3		X			
OVOL2	ovo like zinc finger 2		X	X	X	
PALB2	partner and localizer of BRCA2		X			X
PBRM1	polybromo 1		X			
PMS2	PMS1 homolog 2, mismatch repair s. component	X	X			
PRDX3	peroxiredoxin 3	X	X	X	X	
PTEN	phosphatase and tensin homolog		X			
RAD50	RAD50 double strand break repair protein	X	X	X	X	
RAD51	RAD51 recombinase		X	X	X	
RASSF1	Ras association domain family member 1	X	X	X	X	
RB1	retinoblastoma gene			X	X	
RECQL	RecQ like helicase					
SFN	Stratifin		X	X	X	
SOCS3	suppressor of cytokine signaling 3		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
SRC	SRC proto-oncogene, non-receptor tyrosine kinase		X			
STK11	serine/threonine kinase 11		X	X	X	X
TGFBR1	transforming growth factor beta receptor 1					
TP53	tumor protein p53					
TP53BP2	tumor protein p53 binding protein 2		X	X	X	
TRERF1	transcriptional regulating factor 1		X	X	X	
UBE2C	ubiquitin conjugating enzyme E2 C		X	X	X	
XRCC2	X-ray repair cross complementing 2		X	X	X	
	Apoptosis					
AKT1	v-akt murine thymoma viral oncogene homolog 1			X	X	
BCL2	BCL2, apoptosis regulator		X	X	X	
CASP8	caspase 8					
CHEK2	checkpoint kinase 2		X	X	X	
E2F1	E2F transcription factor 1		X	X	X	
FADD	Fas associated via death domain		X	X	X	X
FGFR2	fibroblast growth factor receptor 2		X	X	X	
FOXO1	forkhead box O1		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
HTT	huntingtin		X	X	X	
IFI16	interferon gamma inducible protein 16		X	X	X	
LGALS1	galectin 1		X	X	X	
MAP2K4	mitogen-activated protein kinase kinase 4		X	X	X	
MAP3K5	mitogen-activated protein kinase kinase kinase 5		X	X	X	
PTEN	phosphatase and tensin homolog		X			
STK11	serine/threonine kinase 11		X	X	X	X
TGFBR1	transforming growth factor beta receptor 1					
TP53	tumor protein p53					
TP53BP2	tumor protein p53 binding protein 2		X	X	X	
HRK	harakiri, BCL2 interacting protein		X	X	X	
TNFRSF10B	TNF receptor superfamily member 10b		X			
TOX3	TOX high mobility group box family member 3					
	Disease mutation					
ACVR2B	activin A receptor type 2B		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
AKT1	v-akt murine thymoma viral oncogene homolog 1			X	X	
ANG	angiogenin		X			
ATM	ATM serine/threonine kinase		X	X	X	
BCL2	BCL2, apoptosis regulator		X	X	X	
BLM	Bloom syndrome RecQ like helicase					
BMPR2	bone morphogenetic protein receptor type 2		X	X	X	
BRCA1	BRCA1, DNA repair associated		X	X	X	
BRCA2	BRCA2, DNA repair associated			X	X	
BRIP1	BRCA1 interacting protein C-terminal helicase 1					
CASP8	caspase 8					
CDH1	cadherin 1			X	X	
CHEK2	checkpoint kinase 2		X	X	X	
CTCF	CCCTC-binding factor		X	X	X	
ECM1	extracellular matrix protein 1		X			X
EDN1	endothelin 1		X	X	X	
EP300	300 kd E1A-Binding protein gene		X	X	X	
ETV6	ets variant gene 6 (TEL oncogene)					

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HME C	Mono-cytes
FADD	Fas associated via death domain		X	X	X	X
FANCA	Fanconi anemia complementation group A		X	X	X	
FANCC	Fanconi anemia complementation group C		X	X	X	X
FANCE	Fanconi anemia complementation group E		X	X	X	
FGFR2	fibroblast growth factor receptor 2		X	X	X	
GATA3	GATA binding protein 3		X	X	X	
HTT	huntingtin		X	X	X	
MLH1	mutL homolog 1		X	X	X	
MSH2	mutS homolog 2	X	X	X	X	
MSH6	mutS homolog 6		X	X	X	
NBN	Nibrin		X	X	X	
NF2	neurofibromin 2		X	X	X	
PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide		X	X	X	
PMS2	PMS1 homolog 2, mismatch repair s. component	X	X			
PTEN	phosphatase and tensin homolog		X			
RAD51	RAD51 recombinase		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HME C	Monocytes
RB1	retinoblastoma gene			X	X	
STK11	serine/threonine kinase 11		X	X	X	X
TGFB1	transforming growth factor beta 1					
TGFB3	transforming growth factor beta 3		X	X	X	
TGFBR1	transforming growth factor beta receptor 1					
TP53	tumor protein p53					
	Ubl conjugation					
AKT1	v-akt murine thymoma viral oncogene homolog 1			X	X	
AURKA	aurora kinase A		X	X	X	
BAP1	BRCA1 associated protein-1		X	X	X	
BCL2	BCL2, apoptosis regulator		X	X	X	
BLM	Bloom syndrome RecQ like helicase					
BRCA1	BRCA1, DNA repair associated		X	X	X	
BRCA2	BRCA2, DNA repair associated			X	X	
CCND1	cyclin D1		X	X	X	
CDC20	cell division cycle 20		X	X	X	X
CDH1	cadherin 1			X	X	
CHEK2	checkpoint kinase 2		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
CTCF	CCCTC-binding factor		X	X	X	
CUL1	cullin 1		X	X	X	
EP300	300 kd E1A-Binding protein gene		X	X	X	
ETV6	ets variant gene 6 (TEL oncogene)					
FANCE	Fanconi anemia complementation group E		X	X	X	
FGFR2	fibroblast growth factor receptor 2		X	X	X	
FOXO1	forkhead box O1		X	X	X	
HTT	huntingtin		X	X	X	
IFI16	interferon gamma inducible protein 16		X	X	X	
KLF5	Kruppel like factor 5		X	X	X	
LYN	LYN proto-oncogene, Src family tyrosine kinase		X	X	X	
MAP3K5	mitogen-activated protein kinase kinase kinase 5		X	X	X	
NF2	neurofibromin 2		X	X	X	
PBRM1	polybromo 1		X			
PTEN	phosphatase and tensin homolog		X			
RAD51	RAD51 recombinase		X	X	X	
SFN	Stratifin		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
SRC	SRC proto-oncogene, non-receptor tyrosine kinase		X			
TGFBR1	transforming growth factor beta receptor 1					
TP53	tumor protein p53					
UBE2C	ubiquitin conjugating enzyme E2 C		X	X	X	
	Cell cycle					
AHR	aryl hydrocarbon receptor		X	X	X	
ATM	ATM serine/threonine kinase		X	X	X	
AURKA	aurora kinase A		X	X	X	
BRCA1	BRCA1, DNA repair associated		X	X	X	
BRCA2	BRCA2, DNA repair associated			X	X	
CCND1	cyclin D1		X	X	X	
CDC20	cell division cycle 20		X	X	X	
CDKN2B	cyclin dependent kinase inhibitor 2B	X	X	X	X	
CHEK2	checkpoint kinase 2		X	X	X	
E2F1	E2F transcription factor 1		X	X	X	
EP300	300 kd E1A-Binding protein gene		X	X	X	
MLH1	mutL homolog 1		X	X	X	
NBN	Nibrin		X	X	X	

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
RAD50	RAD50 double strand break repair protein	X	X	X	X	
RASSF1	Ras association domain family member 1	X	X	X	X	
RB1	retinoblastoma gene			X	X	
SRC	SRC proto-oncogene, non-receptor tyrosine kinase		X			
STK11	serine/threonine kinase 11		X	X	X	
TP53	tumor protein p53					
TP53BP2	tumor protein p53 binding protein 2		X	X	X	
UBE2C	ubiquitin conjugating enzyme E2 C		X	X	X	
	DNA Binding					
AHR	aryl hydrocarbon receptor		X	X	X	
ANG	angiogenin		X			
ATM	ATM serine/threonine kinase		X	X	X	
BLM	Bloom syndrome RecQ like helicase					
BRCA1	BRCA1, DNA repair associated		X	X	X	
BRCA2	BRCA2, DNA repair associated			X	X	
CEBPG	CCAAT/enhancer binding protein gamma		X	X	X	X

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
CTCF	CCCTC-binding factor		X	X	X	
E2F1	E2F transcription factor 1		X	X	X	
ETV6	ets variant gene 6 (TEL oncogene)					
FANCM	Fanconi anemia complementation group M		X	X	X	
FOXA1	forkhead box A1		X	X	X	
FOXO1	forkhead box O1		X	X	X	
GATA3	GATA binding protein 3		X	X	X	
HMGA1	high mobility group AT-hook 1		X	X	X	
IFI16	interferon gamma inducible protein 16		X	X	X	
KLF5	Kruppel like factor 5		X	X	X	
KLF6	Kruppel like factor 6		X	X	X	
MED1	mediator complex subunit 1		X	X	X	
MSH2	mutS homolog 2	X	X	X	X	
MSH3	mutS homolog 3		X			
MSH6	mutS homolog 6		X	X	X	
NFIC	nuclear factor I C		X	X	X	
OVOL2	ovo like zinc finger 2		X	X	X	
PALB2	partner and localizer of BRCA2		X			X

Gene Symbol	Gene Name	MCF7	T47D	HCC 1954	HMEC	Monocytes
PBRM1	polybromo 1		X			
PMS1	PMS1 homolog 1, mismatch repair s. component	X	X	X	X	
RAD51	RAD51 recombinase		X	X	X	
RB1	retinoblastoma gene			X	X	
RECQL	RecQ like helicase					
TOX3	TOX high mobility group box family member 3					
TP53	tumor protein p53					
TRERF1	transcriptional regulating factor 1		X	X	X	
XRCC2	X-ray repair cross complementing 2		X	X	X	

Gene Ontology

The enrichment analysis using Generic Gene Ontology Term Finder (GGOTF) classified the dataset of NRF1 target genes into 520 Gene Ontology (GO) annotations, using a cut off *p* value<0.01. The top 20 GO annotations, ranked by *p* values and the number of genes (Table 5) were as follows: cellular component organization or biogenesis (1,138 genes), positive regulation of cellular process (781), cellular component organization (1,092), negative regulation of cellular process (743), negative regulation of biological process (780), positive regulation of biological process (843), negative regulation of macromolecule metabolic process (451), positive regulation of macromolecule metabolic process (493), negative regulation of metabolic process (475), organelle organization (660), positive regulation of metabolic process (505), negative regulation of cellular metabolic process (441), positive regulation of cellular metabolic process (482), developmental process (883), single-organism developmental process (869), anatomical structure development (819), macromolecular complex subunit organization (489), transcription from RNA polymerase II promoter (392), cell cycle (372), and regulation of macromolecule metabolic process (1069).

Table 5

The Top 20 GO Annotations, Number of NRF1 Target Genes Ranked by p Value

GO ID	Description	Adjusted *p* value	Gene count
GO:0071840	cellular component organization or biogenesis	9.23E-60	1138
GO:0048522	positive regulation of cellular process	1.77E-53	781
GO:0016043	cellular component organization	7.09E-53	1092
GO:0048523	negative regulation of cellular process	1.04E-52	743
GO:0048519	negative regulation of biological process	3.05E-51	780

GO ID	**Description**	**Adjusted *p* value**	**Gene count**
GO:0048518	positive regulation of biological process	5.34E-48	843
GO:0010605	negative regulation of macromolecule metabolic process	2.00E-42	451
GO:0010604	positive regulation of macromolecule metabolic process	2.62E-42	493
GO:0009892	negative regulation of metabolic process	1.11E-41	475
GO:0006996	organelle organization	1.62E-41	660
GO:0009893	positive regulation of metabolic process	5.82E-38	505
GO:0031324	negative regulation of cellular metabolic process	1.23E-37	441
GO:0031325	positive regulation of cellular metabolic process	1.25E-37	482
GO:0032502	developmental process	1.88E-37	883
GO:0044767	single-organism developmental process	1.72E-36	869
GO:0048856	anatomical structure development	6.00E-34	819
GO:0043933	macromolecular complex subunit organization	7.77E-33	489

GO ID	Description	Adjusted *p* value	Gene count
GO:0006366	transcription from RNA polymerase II promoter	1.88E-32	392
GO:0007049	cell cycle	1.48E-31	372
GO:0060255	regulation of macromolecule metabolic process	2.60E-31	1069

We analyzed these 20 biological processes and their subcategories (child processes) to identify the ones that might be associated with the hallmarks of cancer (Hanahan & Weinberg. 2011). To accomplish this identification, we used the mapping that links GO processes to hallmarks of cancer developed by Knijnenburg, Bismeijer, Wessels, and Shmulevich (2015). We found five biological processes (GO) that are representative of the following cancer hallmarks: activating invasion and metastasis (EMT), reprogramming energy metabolism, resisting cell death, and sustaining proliferative signaling. Results are shown in Table 6, including number of NRF1 targets and *p* value.

Table 6

Among the Top 20 Biological Processes Enriched With NRF1 Targets We Found Five Linked to Hallmarks of Cancer

Biological process (parent)	**Biological process – sub category (child process)**	**Gene count**	**Adjusted *p* value**	**Hallmark of Cancer**
GO:0016043 cellular component organization	GO:0032989 cellular component morphogenesis	210	8.62E-12	Activating invasion and metastasis (EMT)
GO:0048522 positive regulation of cellular process	GO:0031325 Positive regulation of cellular metabolic process	482	1.25E-37	Reprogramming energy metabolism
GO:0048523 negative regulation of cellular process	GO:0060548 negative regulation of cell death	140	1.75E-05	Resisting cell death
GO:0006366 transcription from RNA polymerase II promoter	GO:0045944 positive regulation of transcription from RNA polymerase II promoter	162	2.96E-08	Sustaining proliferative signaling
GO:0007049: cell cycle	GO:0045787 positive regulation of cell cycle	62	9.29E-4	Sustaining proliferative signaling

Pathway Analysis

We imported the set of NRF1 target genes into DAVID, which resulted in an output of 89 KEGG pathways. The list of the top 10 enriched pathways, based on *p* value and the corresponding genes count in parenthesis were as follows: hsa04120: Ubiquitin mediated proteolysis (85); hsa04110: Cell cycle (78); hsa05016: Huntington's disease (106); hsa04141: Protein processing in endoplasmic reticulum (95); hsa04144: Endocytosis (133); hsa03018: RNA degradation (51); hsa03040: Spliceosome (73); hsa03015: mRNA surveillance pathway (53); hsa04512: AMPK signaling pathway (66); and hsa04932: Non-alcoholic fatty liver disease (NAFLD) (78). See Table 7 for gene list and *p* values.

The top two pathways, Ubiquitin mediated proteolysis (Figure 4) and Cell cycle (Figure 5), have been reported as altered in breast cancer (Guille, Chaffanet, & Birnbaum, 2013). Ubiquitin mediated proteolysis ranked number one with 85 NRF1 target genes. This pathway plays a critical role in cell cycle regulation and includes several breast cancer genes regulated by NRF1, such as UBE2C whose overexpression has been associated with poor prognosis in breast cancer patients; and CUL1 (Cullin1) that promotes proliferation and migration of breast cancer cells. CUL1 overexpression is also associated with worse survival (Bai et al., 2013). Another gene in this pathway that plays an important role in breast cancer progression is CDC20 (Wang et al., 2015). Its encoded protein and mRNA levels have been found elevated in breast cancer cells (Yuan et al., 2006). Aligned with these results, Karra et al. (2014) found that CDC20 was overexpressed in 445 breast cancer patients and also correlated with short-term survival.

Table 7

Top 10 NRF1 Network Enriched KEGG Pathways Ranked by p Value

KEGG PATHWAY	**Count**	***p VALUE***	**GENES**
hsa04120: Ubiquitin mediated proteolysis	85	1.70E-07	UBE2G1, BTRC, UBE2G2, SAE1, CUL3, FANCL, MGRN1, WWP2, WWP1, ITCH, CUL1, ANAPC1, ANAPC2, SOCS3, ANAPC4, SOCS1, UBE2J1, HERC4, UBE2F, UBE2J2, UBE2H, UBE2C, HERC1, UBE2N, RFWD2, TRIM37, HUWE1, PIAS4, PIAS3, UBE2K, UBE2M, TRIM32, DDB2, UBE2W, MDM2, SIAH1, ANAPC7, PIAS1, UBE2S, FZR1, UBE3A, PPIL2, KEAP1, ANAPC10, ANAPC11, NHLRC1, UBE3C, STUB1, RBX1, UBE2R2, PRPF19, UBE2D4, UBE2D3, UBE2D2, FBXW8, MAP3K1, RHOBTB2, RHOBTB1, FBXO4, NEDD4L, RCHY1, UBE2D1, FBXW11, UBE4A, VHL, UBE4B, CBL, BIRC6, CDC20, PARK2, BIRC3, UBE2Q2, BIRC2, CDC27, UBE2Q1, RNF7, NEDD4, UBA1, UBA2, TCEB2, DET1, SMURF2, TCEB1, SMURF1, UBE2E2

KEGG PATHWAY	**Count**	***p VALUE***	**GENES**
hsa04110: Cell cycle	78	1.90E-07	E2F1, MAD1L1, E2F3, CDC14A, CDC14B, TGFB3, TTK, PTTG1, CCNE2, CCNE1, RAD21, CDKN2B, CDKN2C, CDKN2D, MYC, CCNA2, CUL1, STAG2, STAG1, ANAPC1, CDC7, CDK1, CDC6, ANAPC2, CCNH, ANAPC4, ESPL1, MCM2, CDK7, MCM3, CDK4, MCM4, WEE1, MCM6, CCND1, MAD2L1, GADD45G, BUB1B, MDM2, ANAPC7, GADD45B, GADD45A, FZR1, YWHAZ, PRKDC, CHEK1, ANAPC10, SFN, ANAPC11, ZBTB17, RBX1, TFDP2, BUB1, TFDP1, CREBBP, SMAD4, YWHAB, SMAD3, CDC20, ATR, CDC27, YWHAE, CDC25A, ATM, CDC25B, CDKN1C, CCNB1, YWHAG, CDKN1A, HDAC2, YWHAH, HDAC1, PLK1, GSK3B, PCNA, YWHAQ, SMC1A, ABL1

KEGG PATHWAY	Count	*p VALUE*	GENES
hsa05016: Huntington's disease	106	3.40E-06	NDUFAB1, REST, COX5A, COX5B, UQCR10, SIN3A, UQCR11, CREB3L1, TAF4B, RCOR1, DCTN4, DCTN1, COX6C, ATP5C1, DNAH11, DNAH14, COX7C, HAP1, AP2M1, HIP1, TAF4, HTT, NDUFA4L2, GRIN1, CREBBP, COX8A, VDAC2, VDAC3, VDAC1, UQCRHL, PPIF, NDUFV3, NRF1, HDAC2, HDAC1, BBC3, NDUFV1, NDUFV2, COX6A1, DNAL1, CLTCL1, ATP5D, UQCRC2, CLTA, CLTB, UQCRC1, AP2S1, CYC1, CLTC, UQCRFS1, NDUFS7, NDUFS6, CASP3, AP2B1, PLCB4, CASP9, NDUFS8, DLG4, ATP5O, ATP5H, NDUFS1, ATP5J, NDUFB11, NDUFB10, SLC25A4, CYCS, NDUFC2, COX4I1, NDUFC1, NDUFA12, NDUFA11, UQCRH, UQCRB, POLR2H, POLR2G, NDUFB4, POLR2F, POLR2E, POLR2L, NDUFB7, NDUFB8, NDUFB9, POLR2I, ATP5G2, ATP5G1, DNAH2,

KEGG PATHWAY	Count	*p VALUE*	GENES
hsa04141:Protein processing in endoplasmic reticulum	95	4.50E-06	HSP90AB1, DNAJC5B, SEC31A, SEC24A, PDIA3, UBE2G1, UBE2G2, DNAJC10, MAN1B1, DNAJB12, PDIA4, PRKCSH, UBQLN1, CANX, SSR1, OS9, BAK1, MAP3K5, BAG1, DNAJB11, ATF6B, RPN1, DNAJC5, DNAJC3, SEC24C, MAP2K7, SEC24D, CUL1, DNAJC1, HSP90AA1, MAN1A2, ERP29, UBE2J1, MOGS, UBE2J2, MAN1A1, DDIT3, EIF2AK1, EIF2S1, TXNDC5, SIL1, UGGT2, UGGT1, EIF2AK4, SEC23B, SEC61G, RAD23B, GANAB, DERL1, RAD23A, HSPA1A, EDEM3, LMAN1, EDEM2, STUB1, EDEM1, SEC63, RBX1, NGLY1, STT3B, HSPA1L, UBE2D4, UBE2D3, UBE2D2, STT3A, BCL2, DAD1, DNAJA1, UBE2D1, TRAM1, SEC61A1, HSPA8, DNAJA2, SEC61A2, P4HB, NPLOC4, RRBP1, CKAP4, UBE4B, PARK2, MARCH6, MAN1C1

KEGG PATHWAY	Count	*p VALUE*	GENES
hsa04144:Endocytosis	133	7.90E-06	HRAS, CHMP4B, CAPZA2, CAPZA1, CHMP7, TGFB3, GBF1, WWP1, VPS4B, GIT2, DNAJC6, VPS4A, ITCH, SH3GL3, PLD1, KIF5A, PSD4, HLA-E, ARPC1A, ACAP3, ARRB2, ARRB1, ACAP2, PDCD6IP, BIN1, RAB10, VPS26A, SH3GL1, FGFR2, CHMP2A, ARFGAP1, ARFGAP3, FGFR3, ERBB4, SNX5, SNX2, SNX1, ASAP1, HSPA1A, ARF6, SNX4, ARPC5, SNX3, ARFGEF2, CAPZB, SRC, CHMP2B, HSPA1L, ARPC2, IQSEC1, AP2M1, GIT1, PARD6A, PARD6B, RAB8A, SMAD3, RABEP1, ARF1, NEDD4, ARF3, GRK6, SMURF2, SMURF1, GRK5, CLTCL1, CLTA, CLTB, RAB5C, AP2S1, PIP5K1C, VPS37C, EPS15L1, VPS37D, PIP5K1A, CLTC, CDC42, AP2B1, SMAP1, CXCR4, ZFYVE16, SPG20, KIAA1033, SPG21, AGAP1, AGAP3, RAB4A, PRKCI, WAS, RAB11FIP5, ADRB2,

KEGG PATHWAY	**Count**	***p VALUE***	**GENES**
hsa03018: RNA degradation	51	6.70E-06	CNOT8, LSM8, LSM7, PABPC4, CNOT3, CNOT1, CNOT7, CNOT4, EXOSC10, PATL1, DCPS, PARN, CNOT6L, ENO2, LSM5, LSM4, LSM2, PABPC1, ENO1, HSPA9, PAN2, NUDT16, EXOSC8, EXOSC9, PAN3, EXOSC6, PFKL, EXOSC7, EXOSC4, TTC37, EXOSC5, CNOT10, EXOSC2, PAPD7, PFKP, EXOSC3, PAPD5, PFKM, EXOSC1, DDX6, DIS3, BTG2, WDR61, DCP2, BTG1, DCP1A, HSPD1, MPHOSPH6, PABPC1L, TOB2, TOB1
hsa03040:Spliceosome	73	3.40E-04	NCBP2, SRSF1, CHERP, LSM8, U2AF2, SNRPD3, LSM7, CWC15, ZMAT2, SNRPD1, SNRPD2, SART1, SMNDC1, CTNNBL1, DDX23, U2AF1, PQBP1, LSM5, LSM4, LSM2, SNRPA1, EFTUD2, PRPF3, CDC5L, HNRNPU, PRPF6, EIF4A3, SNRNP200, SNRPB, SNRPA, SLU7, SNRPF, SNRPE, THOC1, SNRPG, SRSF10, CCDC12, TRA2B, TRA2A, SNRPB2, HSPA1A, XAB2, SF3B2, PRPF19, HSPA1L, SF3B1, HNRNPM, PRPF8, USP39, DHX15, DHX16, SNRNP70, PRPF40B, HSPA8, RBM25, PRPF40A,

KEGG PATHWAY	**Count**	***p VALUE***	**GENES**
			BCAS2, DHX8, SNW1, DDX5, U2AF1L4, SRSF3, PPIE, PPIH, SRSF5, SRSF4, SRSF7, SRSF6, SRSF9, SYF2, PUF60, TXNL4A, RBM17
hsa03015:mRNA surveillance pathway	53	7.40E-04	NCBP2, PPP2R5B, PPP2R5A, PPP2R5D, PPP2R5C, HBS1L, WDR82, RNGTT, PNN, CLP1, SRRM1, WDR33, PPP2R1B, PABPN1, PPP2R1A, SYMPK, PPP1CC, CSTF2T, PPP1CB, EIF4A3, PCF11, PPP1CA, CPSF7, CPSF6, PPP2R5E, CPSF4, CPSF1, SSU72, PPP2R2A, NXT1, FIP1L1, PABPC4, DAZAP1, PPP2CA, NUDT21, MSI1, MSI2, PABPC1, UPF2, CSTF3, UPF1, CSTF2, SMG6, SMG7, SAP18, SMG1, RNPS1, NXF1, PAPOLB, PAPOLA, PABPC1L, CSTF1, PPP2R3C

KEGG PATHWAY	Count	*p VALUE*	GENES
hsa04512:AMPK signaling pathway	66	1.40E-03	PPP2R5B, PPP2R5A, PPP2R5D, PPP2R5C, PRKAG2, RPS6KB2, FOXO1, RPS6KB1, FOXO3, CAMKK1, CAMKK2, PDPK1, SLC2A4, EEF2K, CREB3L1, PIK3CA, CAB39, INSR, CCNA2, AKT2, PPP2R1B, PPP2R1A, PFKL, PIK3CB, PRKAB2, PFKP, PRKAB1, ADIPOR2, ADIPOR1, EEF2, PFKM, CCND1, RAB14, PPP2R5E, RAB10, PPP2R2A, CRTC2, CAB39L, PFKFB4, PFKFB3, STK11, PFKFB2, G6PC3, IGF1R, AKT1S1, PPP2CA, GYS1, RAB11B, FASN, PIK3R5, PIK3R3, PIK3R1, PIK3R2, RAB2A, SREBF1, RAB8A, CREB1, SCD, ACACA, STRADA, SIRT1, ADIPOQ, CPT1A, TSC2, RHEB, PPP2R3C
hsa04932: Non-alcoholic fatty liver disease (NAFLD)	78	1.90E-03	UQCRC2, UQCRC1, CYC1, PRKAG2, NDUFAB1, NFKB1, UQCRFS1, COX5A, COX5B, NDUFS7, NDUFS6, CDC42, CASP3, UQCR10, MAP3K5, UQCR11, NDUFS8, PIK3CA, FAS, ITCH, INSR, NDUFS1, AKT2, NDUFB11, NDUFB10, PIK3CB, SOCS3, RXRA, RELA, CYCS, PRKAB2, NDUFC2, ADIPOR2, PRKAB1, COX4I1, ADIPOR1, NDUFC1,

KEGG PATHWAY	Count	*p VALUE*	GENES
			CYP2E1, NDUFA12, DDIT3, BCL2L11, NDUFA11, COX6C, UQCRH, EIF2S1, MAP3K11, UQCRB, BID, NDUFB4, NDUFB7, NDUFB8, NDUFB9, COX7C, RAC1, MLXIP, PIK3R5, PIK3R3, PIK3R1, PIK3R2, SREBF1, CEBPA, NDUFA4, NDUFA5, NDUFA3, NDUFA8, NDUFA9, NDUFA4L2, NDUFA6, COX8A, ADIPOQ, UQCRHL, NDUFV3, SDHB, GSK3A, NDUFV1, GSK3B, NDUFV2, COX6A1

Cell cycle ranked number two, with 78 NRF1 target genes. Two NRF1 targets, Cyclin D1 (CCND1) and its binding partner Cyclin-dependent kinase 4 (CDK4). play a key role in cell cycle, regulating the G1 to S-phase transition (Harbour, Luo, Dei Santi, Postigo, & Dean., 1999; Lamb et al., 2013). CCND1 is considered an oncogene that has been found upregulated in 25% to 60 % and amplified in 10% to 30 % of invasive breast tumors (Courjal et al., 1996; Gillett et al., 1996; Lamb, Lehn, Rogerson, Clarke, & Landberg, 2013; McIntosh et al., 1995). A recent study by Ortiz et al. (2017) found that prognosis of CCND1 overexpression depends on molecular subtypes, and gene amplification is associated with shorter disease-free survival and poor outcome.

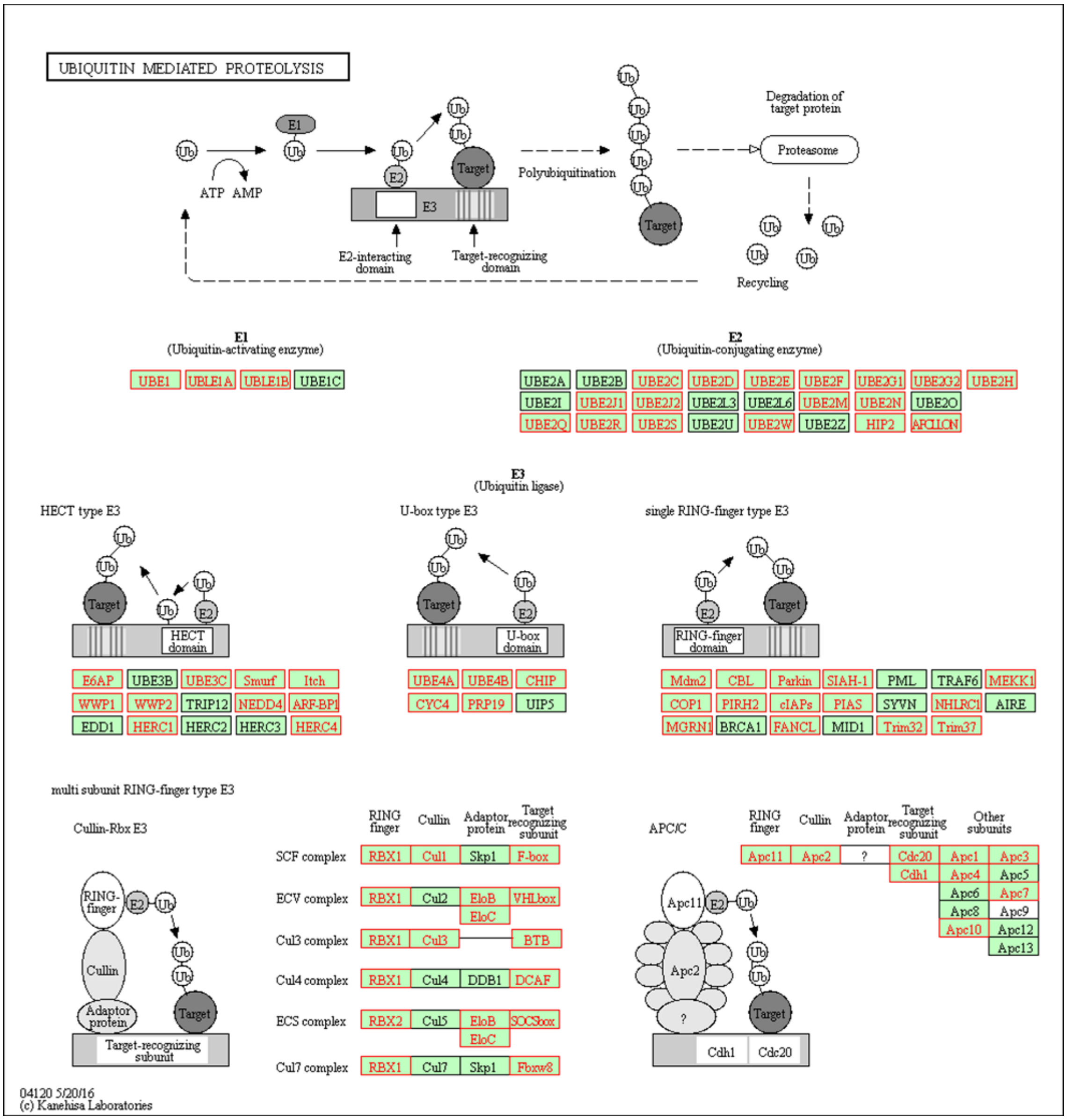


Figure 4. Ubiquitin mediated proteolysis pathway (KEGG Ref: hsa04120) showing NRF1 target genes highlighted in red.

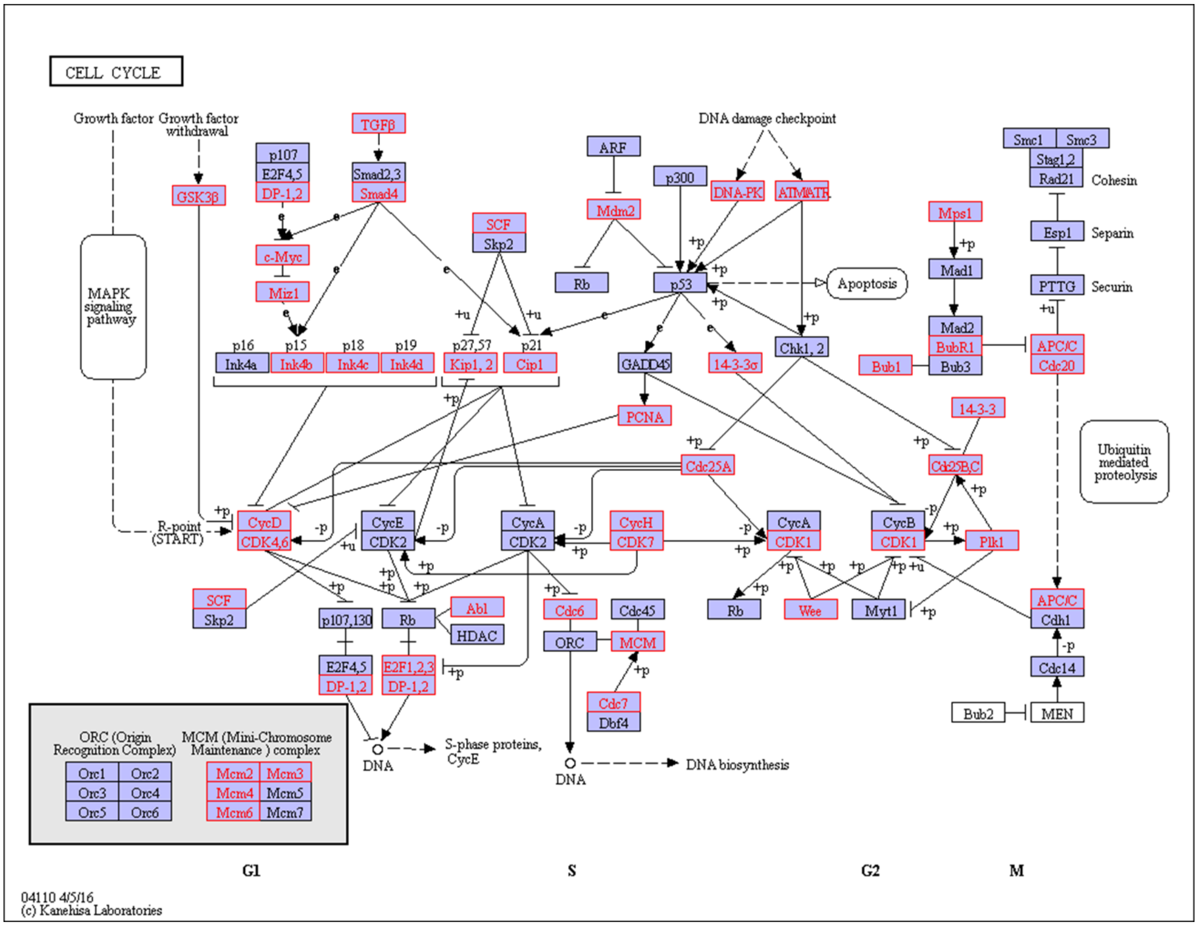


Figure 5. Cell cycle pathway (KEGG Ref: hsa04110) showing NRF1 target genes highlighted in red.

Table 8

KEGG Breast Cancer Pathway Enriched With NRF1 Targets

KEGG PATHWAY	Count	GENES
hsa05224: Breast cancer	68	AKT2, APC, APC2, ARAF, AXIN1, CCND1, CDK4, CDKN1A, CSNK1A1, CTNNB1, DLL4, DVL1, DVL2, DVL3, E2F1, E2F3, ERBB2, ESR2, FGF12, FGF21, FGF9, FGFR1, FOS, FRAT1, FRAT2, FZD1, FZD10, FZD3, FZD4, FZD8, FZD9, GRB2, GSK3B, HES1, HEY1, HEY2, HRAS, IGF1R, JAG1, KRAS, LRP6, MAP2K2, MAPK1, MYC, NCOA3, NFKB2, NOTCH1, NOTCH3, PIK3CA, PIK3CB, PIK3R1, PIK3R2, PIK3R3, PIK3R5, RAF1, RPS6KB1, RPS6KB2, SHC1, SOS1, SOS2, SP1, TCF7L1, WNT10A, WNT11, WNT16, WNT2B, WNT8B, WNT9A

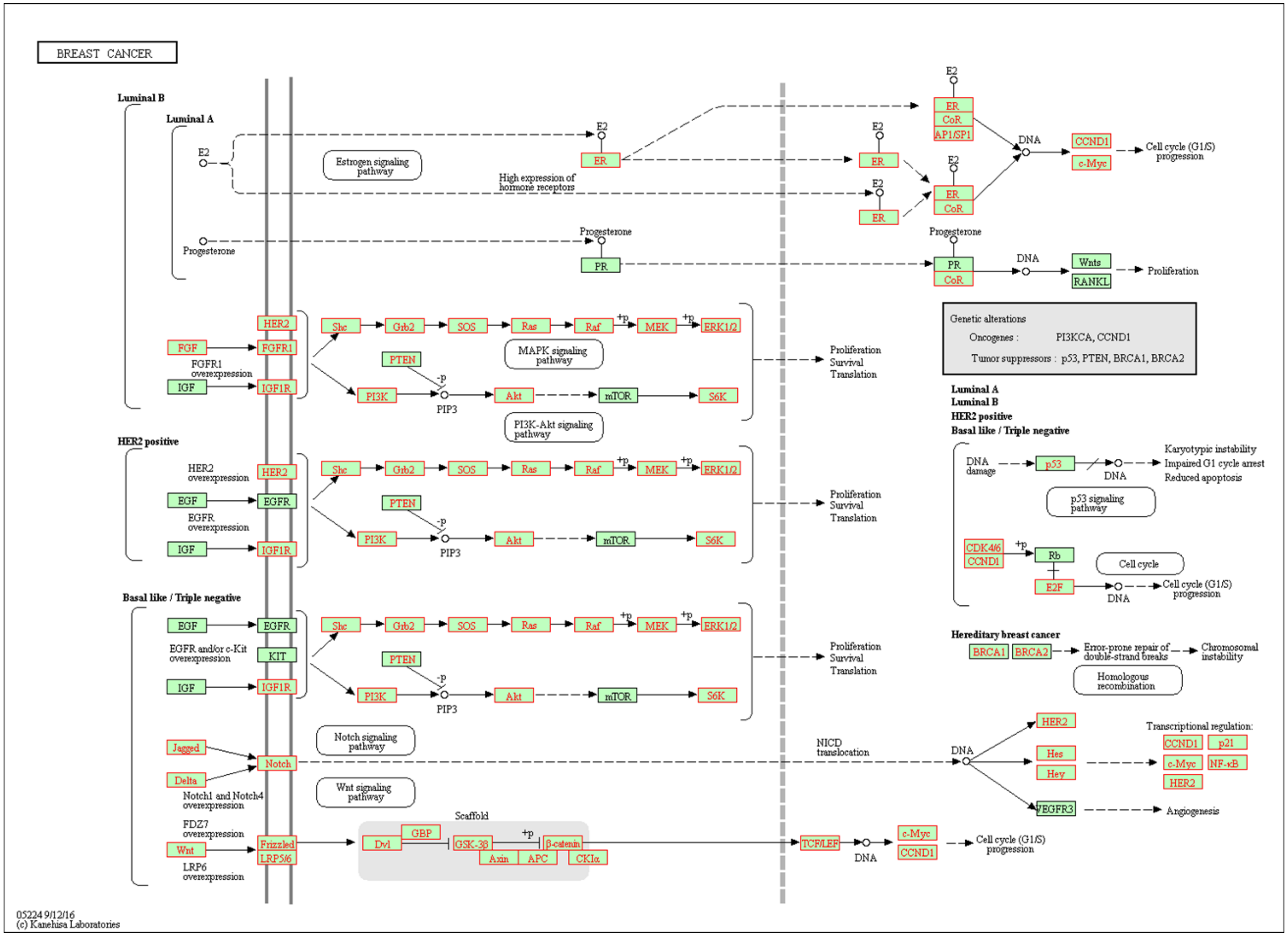


Figure 6. Breast cancer pathway (KEGG Ref: hsa05224) showing NRF1 target genes highlighted in red.

Conclusion

Increased activity or expression of one or more transcription factors might be required for the survival and growth of human cancers (Darnell, 2002). There is growing evidence in the scientific literature that the transcription factor NRF1 may be involved in breast cancer through different mechanisms, including the increase of mitochondrial function to support proliferation of cancer cells and the increase of NRF1 activity due to estrogen-induced ROS signaling. This activity in turn dysregulates cell cycle genes and epigenetic changes affecting NRF1 binding, such as DNA methylation.

Studies using ChIP microarrays or ChIP-Seq to identify NRF1 targets showed that the NRF1 network is cell-context dependent. These dissimilarities might improve our knowledge of differences in breast tumor behavior among molecular subtypes. We also found that a high percentage of the well-known breast cancer genes were directly or indirectly regulated by NRF1. Finally, Gene Ontology and Pathway Analysis confirmed the participation of NRF1-regulated genes in signaling pathways and biological processes important in cancer biology.

Methods

First, we searched the literature through PubMed for NRF1-related articles and selected those focused on associations between NRF1 and its target genes and breast cancer. The next step was to search the literature to construct a dataset of ChIP-Seq-based NRF1 target genes for use in our Gene Ontology and Pathway Analysis. The search for downstream genes regulated by NRF1 was conducted with different techniques and cell lines. Dring the last 5 years, with the use of modern ChIP-Seq methods, the list has increased considerably. We found four studies which were analyzed for overlaps and commonalities to finally produce a list of 8,022 potential NRF1 target genes.

Finally, we performed the Gene Ontology and Pathway Analysis to find genes that may be involved in breast cancer. Gene Ontology was performed using the Generic Gene Ontology Term Finder (GGOTF), a tool developed by the Lewis-Sigler Institute of Princeton University. This web server classified the dataset of 8,022 target genes (identified from MCF7 and T47D breast cancer cells) into functional categories. This classification is based on statistical testing for enriched gene functional categories defined by the Gene Ontology Consortium.

The Pathway Analysis was performed with the Functional Annotation tool of DAVID and KEGG to identify enriched pathways that may be involved in breast cancer development and progression. DAVID and KEGG are available to the general public at http://david.ncifcrf.gov and http://www.genome.jp/kegg/.

REFERENCES

Bai, J., Yong, H. M., Chen, F. F., Mei, P. J., Liu, H., Li, C., . . . Zheng, J. N. (2013). Cullin1 is a novel marker of poor prognosis and a potential therapeutic target in human breast cancer. *Annals of Oncology: Official Journal of the European Society for Medical Oncology, 24*(8), 2016-2022.

Benner, C., Konovalov, S., Mackintosh, C., Hutt, K. R., Stunnenberg, R., & Garcia-Bassets, I. (2013). Decoding a signature-based model of transcription cofactor recruitment dictated by cardinal cis-regulatory elements in proximal promoter regions. *PLoS Genetics, 9*(11), 1-18.

Bonifaci, N., Berenguer, A., Diez, J., Reina, O., Medina, I., Dopazo, J., . . . Pujana, M. A. (2008). Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes. *BMC Medical Genomics, 1*, 1-16.

Cam, H., Balciunaite, E., Blais, A., Spektor, A., Scarpulla, R. C., Young, R., . . . Dynlacht, B. D. (2004). A common set of gene regulatory networks links metabolism and growth inhibition. *Molecular Cell, 16*(3), 399-411.

Campoy, E. M., Laurito, S. R., Branham, M. T., Urrutia, G., Mathison, A., Gago, F., . . . Roqué, M. (2016). Asymmetric cancer hallmarks in breast tumors on different sides of the body. *PloS One, 11*(7), 1-20.

Choi, Y. L., Bocanegra, M., Kwon, M. J., Shin, Y. K., Nam, S. J., Yang, J. H., . . . Pollack, J. R. (2010). LYN is a mediator of epithelial-mesenchymal transition and a target of dasatinib in breast cancer. *Cancer Research, 70*(6), 2296-2306.

Clemons, M., & Goss, P. (2001). Estrogen and the risk of breast cancer. *New England Journal of Medicine, 34*4(4), 276-285.

Courjal, F., Louason, G., Speiser, P., Katsaros, D., Zeillinger, R., & Theillet, C. (1996). Cyclin gene amplification and overexpression in breast and ovarian cancers: Evidence for the selection of cyclin D1 in breast and cyclin E in ovarian tumors. *International Journal of Cancer, 69*(4), 247-253.

Darnell, J. E., Jr. (2002). Transcription factors as targets for cancer therapy. *Nature Reviews Cancer, 2,* 740-749.

Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L., & Schubeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature, 528*(7583), 575-579.

Eliyatkin, N., Yalcin, E., Zengel, B., Aktas, S., & Vardar, E. (2015). Molecular classification of breast carcinoma: From traditional, old-fashioned way to A new age, and A new way. *Journal of Breast Health*, *11*(2), 59-66.

Ertel, A., Tsirigos, A., Whitaker-Menezes, D., Birbe, R. C., Pavlides, S., Martinez-Outschoorn, U. E., . . . Lisanti, M. P. (2012). Is cancer a metabolic rebellion against host aging? in the quest for immortality, tumor cells try to save themselves by boosting mitochondrial metabolism. *Cell Cycle, 11*(2), 253-263.

Felty, Q., Singh, K. P., & Roy, D. (2005). Estrogen-induced G1/S transition of G0-arrested estrogen-dependent breast cancer cells is regulated by mitochondrial oxidant signaling. *Oncogene, 24*(31), 4883-4893.

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., . . . Stefancsik, R. (2017). COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research, 45*(D1), D777-D783.

Gebhard, C., Benner, C., Ehrich, M., Schwarzfischer, L., Schilling, E., Klug, M., . . . Rehli, M. (2010). General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells. *Cancer Research, 70*(4), 1398-1407.

Gillett, C., Smith, P., Gregory, W., Richards, M., Millis, R., Peters, G., & Barnes, D. (1996). Cyclin D1 and prognosis in human breast cancer. *International Journal of Cancer, 69*(2), 92-99.

Guille, A., Chaffanet, M., & Birnbaum, D. (2013). Signaling pathway switch in breast cancer. *Cancer Cell International, 13*(66),1-5.

Gupta, A., Hossain, M. M., Miller, N., Kerin, M., Callagy, G., & Gupta, S. (2016). NCOA3 coactivator is a transcriptional target of XBP1 and regulates PERK-eIF2alpha-ATF4 signalling in breast cancer. *Oncogene, 35*(45), 5860-5871.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646-674.

Harbour, J. W., Luo, R. X., Dei Santi, A., Postigo, A. A., & Dean, D. C. (1999). Cdk phosphorylation triggers sequential intramolecular interactions that progressively block rb functions as cells move through G1. *Cell, 98*(6), 859-869.

Hardwick, J. M., & Soane, L. (2013). Multiple functions of BCL-2 family proteins. *Cold Spring Harbor Perspectives in Biology, 5*(2), 1-22.

Hasegawa, N., Sumitomo, A., Fujita, A., Aritome, N., Mizuta, S., Matsui, K., . . . Mukohara, T. (2012). Mediator subunits MED1 and MED24 cooperatively contribute to pubertal mammary gland development and growth of breast carcinoma cells. *Molecular and Cellular Biology, 32*(8), 1483-1495.

Hatami, R., Sieuwerts, A. M., Izadmehr, S., Yao, Z., Qiao, R. F., Papa, L., . . . Germain, D. (2013). KLF6-SV1 drives breast cancer metastasis and is associated with poor survival. *Science Translational Medicine, 5*(169), 1-16.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., . . . Wang, J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics, 39*(7), 870-874.

Jafaar, Z. M., Litchfield, L. M., Ivanova, M. M., Radde, B. N., Al-Rayyan, N., & Klinge, C. M. (2014). Beta-D-glucan inhibits endocrine-resistant breast cancer cell proliferation and alters gene expression. *International Journal of Oncology, 44*(4), 1365-1375.

Johnson, A. B., & O'Malley, B. W. (2012). Steroid receptor coactivators 1, 2, and 3: Critical regulators of nuclear receptor activity and steroid receptor modulator (SRM)-based cancer therapy. *Molecular and Cellular Endocrinology, 348*(2), 430-439.

Karihtala, P., Mantyniemi, A., Kang, S. W., Kinnula, V. L., & Soini, Y. (2003). Peroxiredoxins in breast carcinoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 9*(9), 3418-3424.

Karra, H., Repo, H., Ahonen, I., Loyttyniemi, E., Pitkanen, R., Lintunen, M., . . . Kronqvist, P. (2014). Cdc20 and securin overexpression predict short-term breast cancer survival. *British Journal of Cancer, 110*(12), 2905-2913.

Kleibl, Z., & Kristensen, V. N. (2016). Women at high risk of breast cancer: Molecular characteristics, clinical presentation and management. *Breast, 28*, 136-144.

Knijnenburg, T. A., Bismeijer, T., Wessels, L. F., & Shmulevich, I. (2015). A multilevel pan-cancer map links gene mutations to cancer hallmarks. *Chinese Journal of Cancer, 34*(10), 439-449.

Lamb, R., Lehn, S., Rogerson, L., Clarke, R. B., & Landberg, G. (2013). Cell cycle regulators cyclin D1 and CDK4/6 have estrogen receptor-dependent divergent functions in breast cancer migration and stem cell-like activity. *Cell Cycle, 12*(15), 2384-2394.

Lim, S., Janzer, A., Becker, A., Zimmer, A., Schule, R., Buettner, R., & Kirfel, J. (2010). Lysine-specific demethylase 1 (LSD1) is highly expressed in ER-negative breast cancers and a biomarker predicting aggressive biology. *Carcinogenesis, 31*(3), 512-520.

McDonald, C., Muhlbauer, J., Perlmutter, G., Taparra, K., & Phelan, S. A. (2014). Peroxiredoxin proteins protect MCF-7 breast cancer cells from doxorubicin-induced toxicity. *International Journal of Oncology, 4*5(1), 219-226.

McIntosh, G. G., Anderson, J. J., Milton, I., Steward, M., Parr, A. H., Thomas, M. D., . . .Home, C. H. (1995). Determination of the prognostic value of cyclin D1 overexpression in breast cancer. *Oncogene, 11*(5), 885-891.

Narlikar, L., & Ovcharenko, I. (2009). Identifying regulatory elements in eukaryotic genomes. *Briefings in Functional Genomics & Proteomics, 8*(4), 215-230.

Niida, A., Smith, A. D., Imoto, S., Tsutsumi, S., Aburatani, H., Zhang, M. Q., & Akiyama, T. (2008). Integrative bioinformatics analysis of transcriptional regulatory programs in breast cancer cells. *BMC Bioinformatics, 9*, 1-14.

Okoh, V. O., Garba, N. A., Penney, R. B., Das, J., Deoraj, A., Singh, K. P., . . . Roy, D. (2015). Redox signalling to nuclear regulatory proteins by reactive oxygen species contributes to oestrogen-induced growth of breast cancer cells. *British Journal of Cancer, 112*(10), 1687-1702.

Oliveros, J. C. (2007/2015). *VENNY. An interactive tool for comparing lists with Venn diagrams*. Retrieved from http://bioinfogp.cnb.csic.es/tools/venny/index.html

Ortiz, A. B., Garcia, D., Vicente, Y., Palka, M., Bellas, C., & Martin, P. (2017). Prognostic significance of cyclin D1 protein expression and gene amplification in invasive breast carcinoma. *PloS One, 12*(11), 1-13.

Parkash, J., Felty, Q., & Roy, D. (2006). Estrogen exerts a spatial and temporal influence on reactive oxygen species generation that precedes calcium uptake in high-capacity mitochondria: Implications for rapid nongenomic signaling of cell growth. *Biochemistry, 45*(9), 2872-2881.

Pasche, B. (2008). Recent advances in breast cancer genetics. *Cancer Treatment and Research, 141*, 1-10.

Psyrri, A., Kalogeras, K. T., Kronenwett, R., Wirtz, R. M., Batistatou, A., Bournakis, E., . . . Makatsoris, T. (2012). Prognostic significance of UBE2C mRNA expression in high-risk early breast cancer. A hellenic cooperative oncology group (HeCOG) study. *Annals of Oncology: Official Journal of the European Society for Medical Oncology, 23*(6), 1422-1427.

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. *Nature, 505*(7483), 302-308.

Ripperger, T., Gadzicki, D., Meindl, A., & Schlegelberger, B. (2009). Breast cancer susceptibility: Current knowledge and implications for genetic counselling. *European Journal of Human Genetics, 17*(6), 722-731.

Roy, D., Cai, Q., Felty, Q., & Narayan, S. (2007). Estrogen-induced generation of reactive oxygen and nitrogen species, gene damage, and estrogen-dependent cancers. *Journal of Toxicology and Environmental Health. Part B, Critical Reviews, 10*(4), 235-257.

Satoh, J., Kawana, N., & Yamamoto, Y. (2013). Pathway analysis of ChIP-seq-based NRF1 target genes suggests a logical hypothesis of their involvement in the pathogenesis of neurodegenerative diseases. *Gene Regulation and Systems Biology, 7,* 139-152.

Scarpulla, R. C. (2006). Nuclear control of respiratory gene expression in mammalian cells. *Journal of Cellular Biochemistry, 97*(4), 673-683.

Scarpulla, R. C. (2008). Transcriptional paradigms in mammalian mitochondrial biogenesis and function. *Physiological Reviews, 88*(2), 611-638.

Shah, S. N., Cope, L., Poh, W., Belton, A., Roy, S., Talbot, C. C., Jr., . . . Resar, L. M. (2013). HMGA1: A master regulator of tumor progression in triple-negative breast cancer cells. *PloS One, 8*(5), 1-9.

Shen, D., Chang, H. R., Chen, Z., He, J., Lonsberry, V., Elshimali, Y., . . . Gombein, J. A. (2005). Loss of annexin A1 expression in human breast cancer detected by multiple high-throughput analyses. *Biochemical and Biophysical Research Communications, 326*(1), 218-227.

Shiovitz, S., & Korde, L. A. (2015). Genetics of breast cancer: A topic in evolution. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, *26*(7), 1291-1299.

Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians, 68*(1), 7-30.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordan, R., & Rohs, R. (2014). Absence of a simple code: How transcription factors read the genome. *Trends in Biochemical Sciences, 39*(9), 381-399.

Sotgia, F., Whitaker-Menezes, D., Martinez-Outschoorn, U. E., Salem, A. F., Tsirigos, A., Lamb, R., . . . Lisanti, M. P. (2012). Mitochondria "fuel" breast cancer metabolism: Fifteen markers of mitochondrial biogenesis label epithelial cancer cells, but are excluded from adjacent stromal cells. *Cell Cycle, 11*(23), 4390-4401.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., . . . Kuhn, M. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research, 43(Database issue),* D447-52.

Thompson, C., MacDonald, G., & Mueller, C. R. (2011). Decreased expression of BRCA1 in SK-BR-3 cells is the result of aberrant activation of the GABP beta promoter by an NRF-1-containing complex. *Molecular Cancer, 10*, 1-17.

van der Groep, P., van der Wall, E., & van Diest, P. J. (2011). Pathology of hereditary breast cancer. *Cellular Oncology, 34*(2), 71-88.

Wang, L., Zhang, J., Wan, L., Zhou, X., Wang, Z., & Wei, W. (2015). Targeting Cdc20 as a novel cancer therapeutic strategy. *Pharmacology & Therapeutics, 151,* 141-151.

Wilson, G. R., Cramer, A., Welman, A., Knox, F., Swindell, R., Kawakatsu, H., . . . Bundred, N. J.. (2006). Activated c-SRC in ductal carcinoma in situ correlates with high tumour grade, high proliferation and HER2 positivity. *British Journal of Cancer, 95*(10), 1410-1414.

Wulfing, P., Diallo, R., Kersting, C., Wulfing, C., Poremba, C., Greb, R. R., . . . Kiesel, L. (2004). Endothelin-1, endothelin-A- and endothelin-B-*receptor expression in preinvasive and invasive breast disease. Oncology* Reports, 11(4), 791-796.

Xie, Y. G., Yu, Y., Hou, L. K., Wang, X., Zhang, B., & Cao, X. C. (2016). FYN promotes breast cancer progression through epithelial-mesenchymal transition. *Oncology Reports, 36*(2), 1000-1006.

Xu, J., Wu, R. C., & O'Malley, B. W. (2009). Normal and cancer-related functions of the p160 steroid receptor co-activator (SRC) family. *Nature Reviews Cancer, 9*(9), 615-630.

Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., & Jarvis, E. D. (2004). Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics, 20*(18), 3594-3603.

Yuan, B., Xu, Y., Woo, J. H., Wang, Y., Bae, Y. K., Yoon, D. S., . . . Gabrielson, E. (2006). Increased expression of mitotic checkpoint genes in breast cancer cells with chromosomal instability. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 12(2), 405-410.

Zhang, J., Wang, C., Chen, X., Takada, M., Fan, C., Zheng, X., . . . Aird, K. M. (2015). EglN2 associates with the NRF1-PGC1alpha complex and controls mitochondrial function in breast cancer. *EMBO Journal, 34*(23), 2953-2970.

Zhang, J., Zheng, X., & Zhang, Q. (2015). EglN2 positively regulates mitochondrial function in breast cancer. *Molecular & Cellular Oncology, 3*(2), 1-3.

CHAPTER III

INTEGRATED CHIP-SEQ AND RNA-SEQ DATA ANALYSIS TO INVESTIGATE REGULATORY MECHANISMS OF NRF1 TRANSCRIPTION FACTOR ON TARGET GENES IN HER2+ BREAST CANCER CELLS

Abstract

Nuclear respiratory factor 1 (NRF1) is a redox sensitive transcription factor involved in breast cancer development and progression. Recent studies have reported increased activity of NRF1 target genes in breast cancer compared to normal adjacent tissue; however, the underlying mechanisms of NRF1 involvement in mammary tumors have not been elucidated. In this paper, we show how, by the use of integrative data analysis of NRF1 ChIP-Seq and RNA seq in human epidermal growth factor receptor 2 positive (ER/PR -ve Her2+ve +) breast tumor cells, we discovered a set of predicted NRF1 targets with significant differential expression and NRF1 motifs that can be potentially considered as disease drivers. We also found that some of these genes had already been reported as associated with breast cancer, therapeutic resistance, and poor prognosis. A good portion of the paper is devoted to describing and discussing the importance of selecting the appropriate computational analysis methods, software, and parameters for the processing of NRF1 ChIP-Seq and RNA-Seq raw data as well as for their integrative target analysis in order to obtain accurate results.

Introduction

Nuclear respiratory factor 1 (NRF1) is a redox sensitive transcription factor that has been associated with breast cancer. NRF1 activity was found higher in breast cancer compared to adjacent normal tissue, and upregulation of its target genes was found to be associated with metastasis and lower overall survival in breast cancer patients. These results were obtained with bioinformatics analysis (Ertel et al., 2012). In vitro testing also confirmed this association by demonstrating that estrogen-induced reactive oxygen species (ROS) increased the binding activity of NRF1, which causes the upregulation of NRF1 regulated cell cycle genes contributing to the growth of MCF-7 breast cancer cells (Okoh et al., 2015).

Chromatin regulators and transcription factors (TFs) play two of the most important roles among numerous mechanisms involved in the regulation of gene expression (Wang et al., 2013). While TFs bind to DNA, chromatin regulators can modify the chromatin structure by catalyzing or binding to histone modifications; these actions affect the access of proteins to DNA. Frequently, chromatin regulators may also act as transcription cofactors (Dai, 2007; Wang et al., 2013). Chromatin immunoprecipitation followed by next-generation sequencing (ChIP-Seq) is an effective tool that is being widely used by researchers to study transcription factors binding to DNA and locations of histone modifications (Bailey & Machanick, 2012; Furey, 2012).

To investigate the effect of chromatin regulators and TFs in the regulation of gene expression, it is necessary to integrate the ChIP-Seq results with the transcriptome data measured under different conditions, such as transcription factor binding and nonbinding states (Wang et al., 2013). RNA-Seq is currently the preferred method of measuring transcriptome data, replacing microarrays. Both assays, ChIP-Seq and RNA-Seq, are based on next-generation sequencing (NGS) (Finotello & Di Camillo, 2015).

This chapter focuses on describing methods for bioinformatics processing of NRF1 ChIP-Seq and RNA-Seq data with the use of raw sequencing datasets from breast cancer and normal human mammary epithelial cells. Subsequently, we show how NRF1 ChIP-Seq and RNA-Seq results can be integrated toward understanding of the regulatory mechanisms of NRF1 on gene expression and how these discoveries can be used to identify genes involved in breast cancer.

Methods

Datasets

Raw NRF1 ChIP-Seq dataset in HCC1954 breast cancer cells was retrieved from NCBI-Gene expression omnibus (GEO) with accession numbers GSM1891658 and GSM1891659 (replicates 1 and 2. respectively). NRF1 ChIP-Seq dataset in HMEC-Human mammary epithelial cells was retrieved from GSM1891655 and GSM1891656. NRF1 input in HMEC was retrieved from GSM1891657 and NRF1 input in HCC1954 from GSM1891660. Sequencing was done using Illumina machine HiSeq 2500 at 50 bp read length single end and in accordance with Illumina standards (Domcke et al., 2015). Raw RNA-Seq dataset in HCC1954 breast cancer cells was retrieved from GSM721140 (SRR201983 and SRR201984) and RNA-Seq dataset in HMEC-Human mammary epithelial cells was retrieved from GSM721141 (SRR201985 / SRR201986) (Hon et al., 2012).

ChIP-Seq

Overview of ChIP-Seq. The main goal of ChIP-Seq experiments is the mapping of transcription factor binding sites, histone modifications, and other DNA binding proteins on a genome-wide scale (Feng et al., 2012; Park, 2009). ChIP assays are performed in several steps. First, cells are treated with a chemical agent (frequently formaldehyde) to crosslink the protein under study to DNA. This procedure is followed by a process of sonication to divide the chromatin into 100 to 300 bp fragment sizes. Subsequently, the protein of interest together with its bound DNA is immunoprecipitated with an antibody specific to the protein. After immuno-enrichment, the crosslinks are reversed in order to release the DNA, which is then purified and prepared for high-throughput sequencing, also called next generation sequencing (NGS) (Landt et al., 2012; Park, 2009). ChIP-Seq data needs to be processed and analyzed to produce meaningful results other than sequencing files.

Replication and sequencing depth. Before ChIP-Seq data analysis is begun, it is necessary to make sure replication and sequencing depth requirements have been met. These requirements or guidelines are issued by the Encyclopedia of DNA Elements Consortium (ENCODE), an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). One of ENCODE’s aims is to construct a comprehensive list of the regulatory elements that control gene expression. ENCODE requires that all ChIP assays should be carried out in two independent biological replicates.

It has been reported that additional replicates do not have any significant effect in protein-DNA site discovery. With regard to sequencing depth for transcription factors in mammalian cells, a minimum of 10,000,000 unique mapped readings per replicate are required. This number totals 20,000,000 per transcription factor (Landt et al., 2012; Rozowsky et al., 2009).

Control (reference) sample. Another feature requiring confirmation before ChIP data analysis is the existence of an appropriate control sample. There are two reasons why the control sample is so important. First, when the sonication process takes place, regions of DNA with open chromatin are more prone to break and therefore are overrepresented. Second, different platforms currently used in ChIP sequencing, including the Illumina, the most popular platform, have their own biases (Auerbach et al., 2009; Dohm, Lottaz, Borodina, & Himmelbauer, 2008). Some algorithms have been developed to correct these biases (Cheung, Down, Latorre, & Ahringer, 2011) for ChIP-Seq data analysis based upon peak calling results, such as identification of transcription factor binding sites. However, the control sample is a logical approach to overcome biases, assuming that ChIP and control samples have the same sequencing biases when the same instruments are used for the assay.

There are two types of control DNA: input DNA and IgG control. Input DNA is obtained by isolation of DNA after crosslinking and fragmentation of the same cells used for the ChIP assay, following the same procedures but without immunoprecipitation. IgG control is obtained by simulating a ChIP reaction using an antibody specific to a non-nuclear antigen (Landt et al., 2012). Once the ChIP seq experiment has been successfully completed, assurances must be made that the computational analysis of sequencing reads is properly performed to obtain reliable results.

Computational analysis of ChIP-Seq. Computational analysis of ChIP-Seq includes several steps, shown in the flowchart of Figure 1. The steps are Quality control of sequencing reads, Alignment (mapping sequencing reads to the genome), Peak calling (identifying binding sites), and Data visualization and Annotation (identifying transcription factor target genes). Galaxy (https://usegalaxy.org) is a web-based platform open to the public with many tools to analyze large biomedical data, including ChIP-Seq datasets. Through Galaxy, a great number of computing resources can be accessed to carry out each one of the steps in ChIP data analysis.

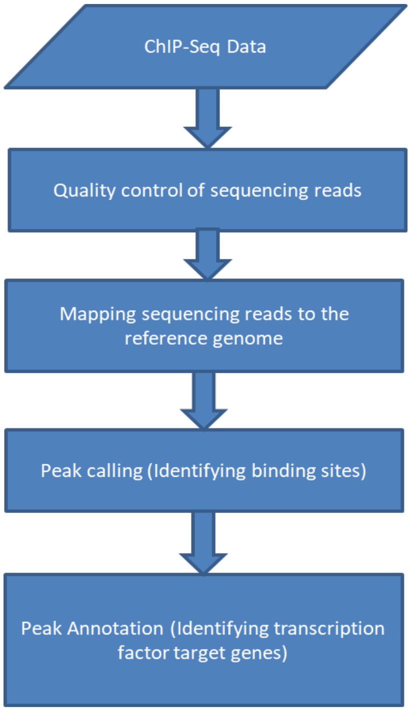

ChIP-Seq Data
Quality control of sequencing reads
Mapping sequencing reads to the reference genome
Peak calling (Identifying binding sites)
Peak Annotation (Identifying transcription factor target genes)

Figure 1. Flow chart of the steps in computational analysis of ChIP-Seq data.

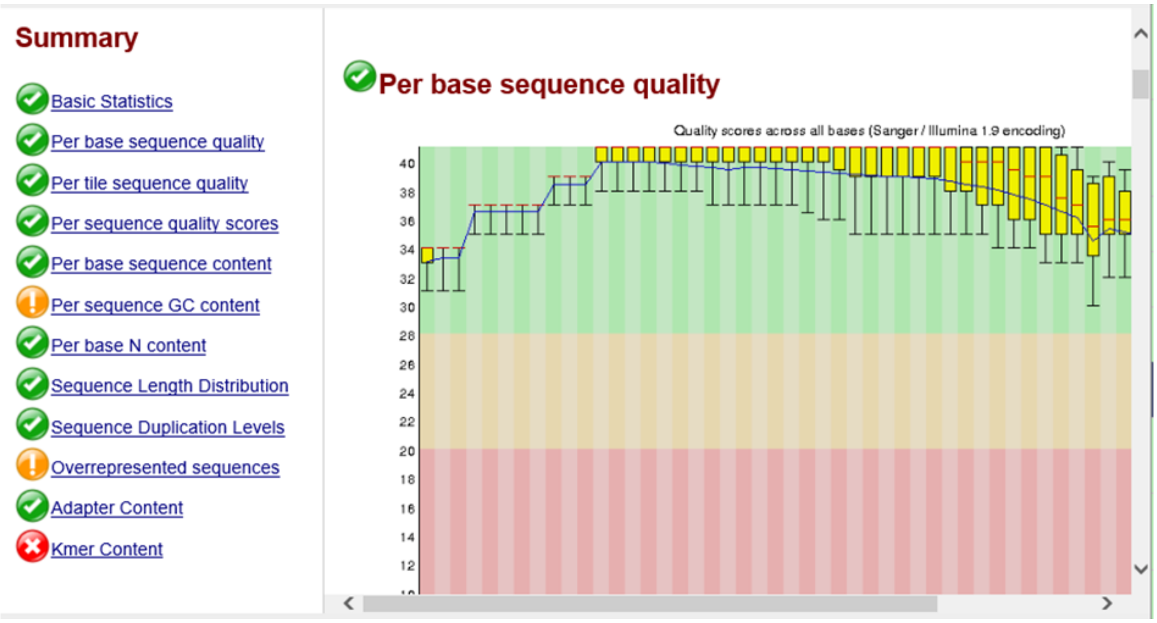
Galaxy has been available online since 2007 and currently has over 124,000 registered users who run approximately 245,000 analyses every month (Afgan et al., 2018). All software used for ChIP-Seq data analysis in this chapter was accessed through Galaxy. Tutorials for learning how to use Galaxy and all accessible tools are available and can be accessed through the same webserver. The first step in ChIP-Seq is quality control of sequencing reads.

Quality control of sequencing reads. Next-generation sequencing (NGS) produces millions of short sequences, also called short reads, ranging between 25 and 75 bp (Feng et al., 2012). Technology of next-generation sequencing involves the use of optical sensors and software that analyze the sensor data to determine the individual bases. This final step is known as base calling (Ledergerber & Dessimoz, 2011). Sequencing files for short reads generally appear in FASTQ format. Each entry in a FASTQ file is composed of four lines: the identifier, the sequence, the quality score line identifier (only the + sign), and the quality score of each base call. Below is an example of FASTQ file entry (Illumina, 2011-2014, FASTQ Files section, para. 10):

```
@SIM:1: FCX: 1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

The last line is the Phred quality score (Q) for each nucleotide, representing the level of confidence during the base calling process: $Q = -10 \log10 (P)$; therefore, $P = 10^{-Q/10}$. P = is the error probability for the base call. For example, if the estimated probability of error for a specific base call is 0.001, then the assigned Phred score Q= 30 (Ewing & Green, 1998). In this case, the Phred score is reported in ASCII characters, which can be converted into Q scores. For instance, the symbol < is equivalent to a quality score Q= 27, which in turn represents an estimated Probability of P = 0.002 that the base is incorrect.

FastQC is one of the most widely used software programs to perform quality control checks on raw sequence data, including the Phred quality score of base calling. The reports generated by FastQC include a text file with the following information: basic statistics, per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, and Kmer Content. Figure 2 shows a screenshot with a partial view of the report generated by FastQC. This software is available through Galaxy in the tools tab to the left, under "NGS: QC and manipulation" and also at http://www.bioinformatics.babraham.ac.uk/ projects/fastqc/ (Andrews, 2010).



Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

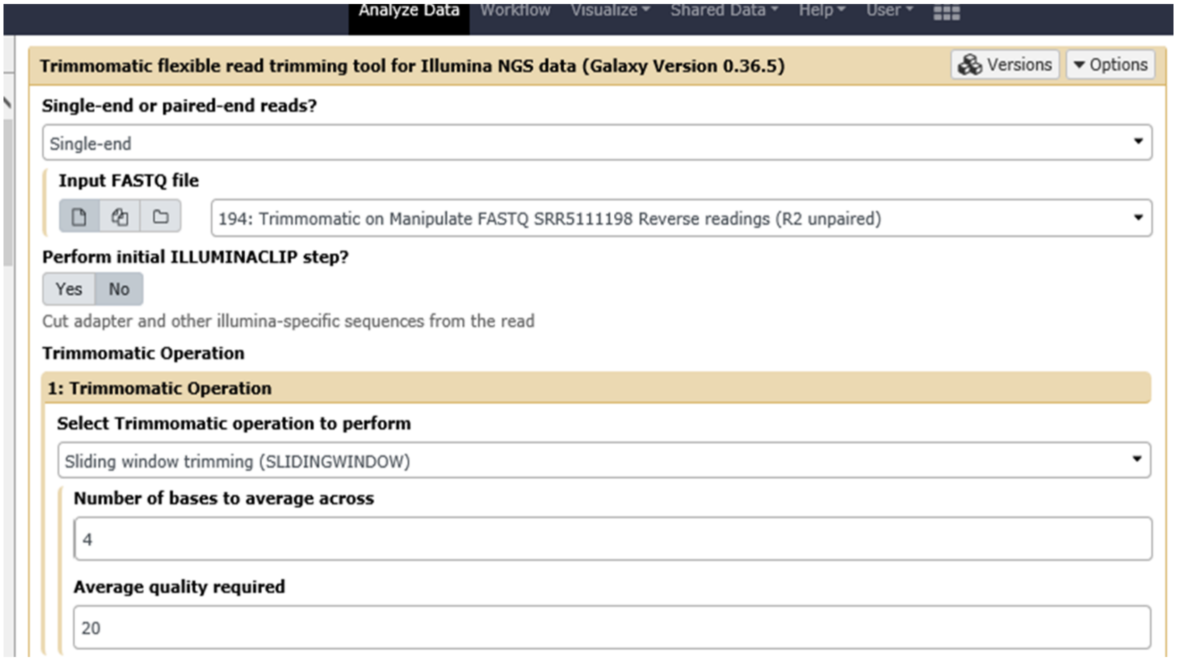
✓ **Per base sequence quality**

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12

Figure 2. Screenshot with a partial view of the report generated by FastQC in Galaxy.

Based on the results generated by FastQC, a threshold can be established to discard all readings with quality score below that number. Trimmomatic (Bolger, Lohse, & Usadel, 2014) is one of the software programs that allows us to do this and can be also accessed through Galaxy under NGS: QC and manipulation. Trimmomatic can be used to perform different operations, but our present focus is concentration on dropping the readings if the average quality (number of bases to average across = 4) is below a Phred score of 20. A Phred score of 20 is equivalent to a probability of 1 in 100 that the base is called wrong (99 % accuracy of the base call).

Figure 3 shows a screenshot of the Galaxy/Trimmomatic step before execution; the minimum Phred score has been set up to 20 to drop base calls with Q < 20. After quality control, readings are ready for mapping into the genome.



Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.5)

Single-end or paired-end reads?
Single-end

Input FASTQ file
194: Trimmomatic on Manipulate FASTQ SRR5111198 Reverse readings (R2 unpaired)

Perform initial ILLUMINACLIP step?
Yes [✓ No]
Cut adapter and other illumina-specific sequences from the read

Trimmomatic Operation

1: Trimmomatic Operation

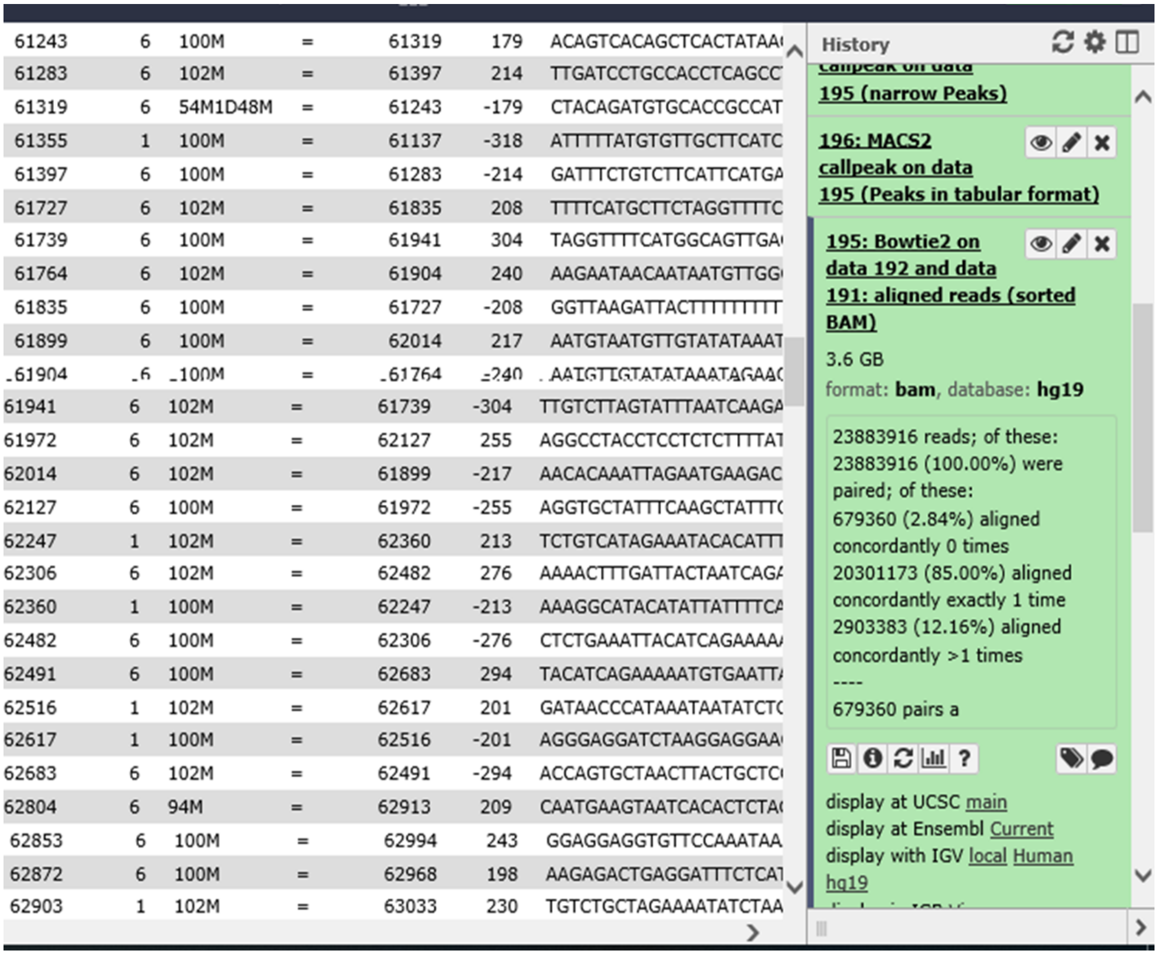
Select Trimmomatic operation to perform
Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across
4

Average quality required
20

Figure 3. Screenshot of the Galaxy/Trimmomatic step before execution. The minimum Phred score has been set up to 20 to drop base calls with Q < 20.

Mapping sequencing reads to the genome (Alignment). For mapping readings into the referenced human genome, we used BOWTIE2 (Langmead & Salzberg, 2012), available through Galaxy in the tools tab to the left, under NGS: Mapping. Reads are mapped to a reference genome that we need to select. In our case, we used the human genome reference hg19 because we needed to compare results of NRF1 target genes with previous ChIP-Seq experiments that had used this reference. However, a new Homo sapiens genome reference hg38 is currently being used in all new ChIP-Seq experiments. Results are in BAM format, as can be seen in Figure 4. The screenshot also shows the statistics of alignments.



```
61243   6   100M       =   61319    179   ACAGTCACAGCTCACTATAA
61283   6   102M       =   61397    214   TTGATCCTGCCACCTCAGCC
61319   6   54M1D48M   =   61243   -179   CTACAGATGTGCACCGCCAT
61355   1   100M       =   61137   -318   ATTTTTATGTGTTGCTTCATC
61397   6   100M       =   61283   -214   GATTTCTGTCTTCATTCATGA
61727   6   102M       =   61835    208   TTTTCATGCTTCTAGGTTTTC
61739   6   100M       =   61941    304   TAGGTTTTCATGGCAGTTGA
61764   6   102M       =   61904    240   AAGAATAACAATAATGTTGG
61835   6   100M       =   61727   -208   GGTTAAGATTACTTTTTTTTT
61899   6   100M       =   62014    217   AATGTAATGTTGTATATAAAT
61904   6   100M       =   61764   -240   AATGTTGTATATAAATAGAAC
61941   6   102M       =   61739   -304   TTGTCTTAGTATTTAATCAAGA
61972   6   102M       =   62127    255   AGGCCTACCTCCTCTCTTTTAT
62014   6   102M       =   61899   -217   AACACAAATTAGAATGAAGAC
62127   6   100M       =   61972   -255   AGGTGCTATTTCAAGCTATTTC
62247   1   102M       =   62360    213   TCTGTCATAGAAATACACATTT
62306   6   102M       =   62482    276   AAAACTTTGATTACTAATCAGA
62360   1   100M       =   62247   -213   AAAGGCATACATATTATTTTCA
62482   6   100M       =   62306   -276   CTCTGAAATTACATCAGAAAAA
62491   6   100M       =   62683    294   TACATCAGAAAAATGTGAATTA
62516   1   102M       =   62617    201   GATAACCCATAAATAATATCTC
62617   1   100M       =   62516   -201   AGGGAGGATCTAAGGAGGAA
62683   6   102M       =   62491   -294   ACCAGTGCTAACTTACTGCTC
62804   6   94M        =   62913    209   CAATGAAGTAATCACACTCTA
62853   6   100M       =   62994    243   GGAGGAGGTGTTCCAAATAA
62872   6   100M       =   62968    198   AAGAGACTGAGGATTTCTCAT
62903   1   102M       =   63033    230   TGTCTGCTAGAAAATATCTAA
```

History

callpeak on data
195 (narrow Peaks)

196: MACS2
callpeak on data
195 (Peaks in tabular format)

195: Bowtie2 on
data 192 and data
191: aligned reads (sorted
BAM)

3.6 GB
format: **bam**, database: **hg19**

```
23883916 reads; of these:
23883916 (100.00%) were
paired; of these:
679360 (2.84%) aligned
concordantly 0 times
20301173 (85.00%) aligned
concordantly exactly 1 time
2903383 (12.16%) aligned
concordantly >1 times
----
679360 pairs a
```

display at UCSC main
display at Ensembl Current
display with IGV local Human
hg19

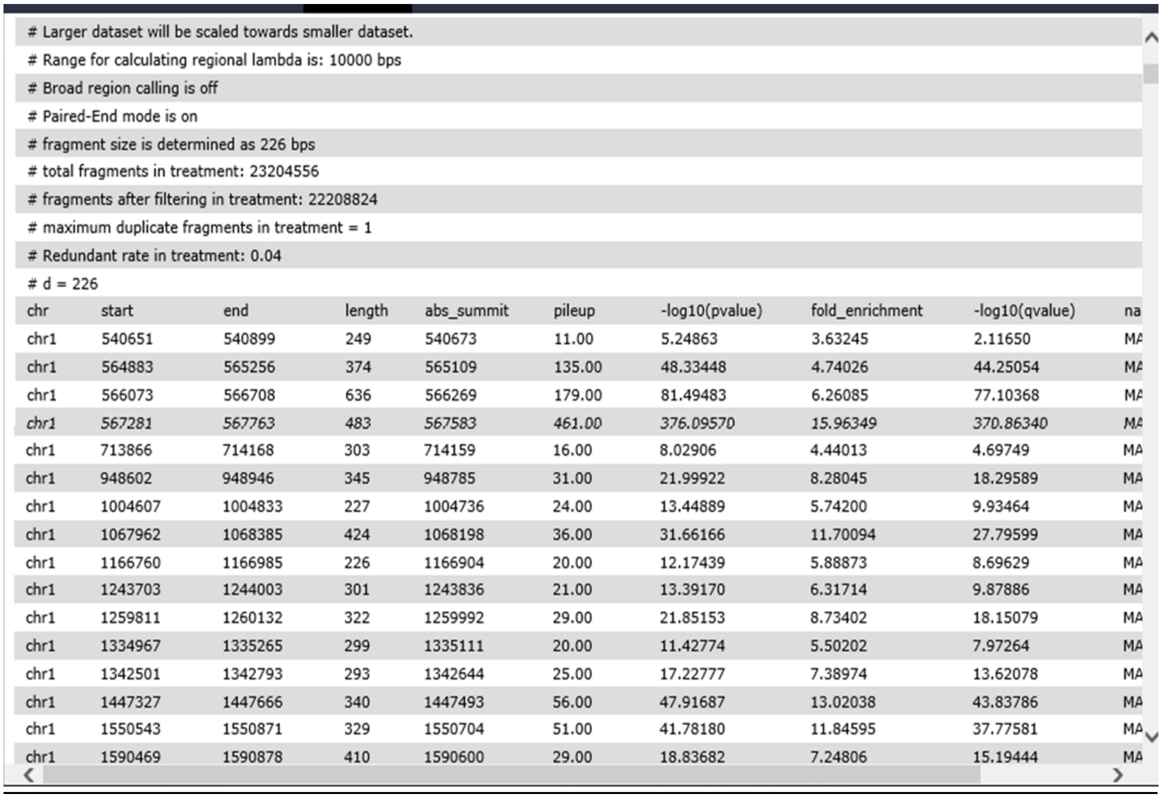
Figure 4. Partial view of Galaxy screenshot showing the alignment results generated by BOWTIE. Also on the right hand side, the statistics of alignment show that the percentage of reads aligned concordantly exactly 1 time was 85%.

One of the important statistics to examine is the percentage of reads aligned concordantly exactly 1 time. In this case, the result was 85%, as can be observed in the righthand side of Figure 4. ENCODE guidelines require that the Non-Redundant Fraction (NRF)—“Number of distinct uniquely mapping reads (i.e., after removing duplicates) / Total number of reads”—should be greater than or equal to

80% (Landt et al., 2012). The next step after alignment is identification of binding sites through Peak calling.

Peak calling. MACS2 (Zhang et al., 2008) is the software we used for peak calling. This software can be accessed through Galaxy in the tools tab to the left, under NGS: Peak Calling. Prior to execution, some parameters must be defined, including the false discovery rate (*q* value) which we set up to 0.01. Figure 5 shows a screen shot of results from MACS2 which are provided by the software in tabular format. The column fold enrichment can be used to establish a cutoff point to filter these results. We discarded peaks with fold enrichment value below 5 following ENCODE's guidelines for point-source transcription factors (Landt et al., 2012). In general, the final list of peaks depends on the parameter settings (for example, *p* value and false discovery rate), the software used to identify them, the selected control (reference sample), and the quality of the ChIP-Seq experiment (Landt et al., 2012).

Usually, once the binding sites have been established, the next step is to identify the TF target genes though a process called Gene Annotation, described in the next section. However, in this research we did not use this method because we were interested in finding target genes integrating ChIP–Seq and RNA-Seq, as described later in the section Integration of ChIP-Seq and RNA-Seq Data Analysis.



```
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 10000 bps
# Broad region calling is off
# Paired-End mode is on
# fragment size is determined as 226 bps
# total fragments in treatment: 23204556
# fragments after filtering in treatment: 22208824
# maximum duplicate fragments in treatment = 1
# Redundant rate in treatment: 0.04
# d = 226
```

chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	na
chr1	540651	540899	249	540673	11.00	5.24863	3.63245	2.11650	MA
chr1	564883	565256	374	565109	135.00	48.33448	4.74026	44.25054	MA
chr1	566073	566708	636	566269	179.00	81.49483	6.26085	77.10368	MA
chr1	567281	567763	483	567583	461.00	376.09570	15.96349	370.86340	MA
chr1	713866	714168	303	714159	16.00	8.02906	4.44013	4.69749	MA
chr1	948602	948946	345	948785	31.00	21.99922	8.28045	18.29589	MA
chr1	1004607	1004833	227	1004736	24.00	13.44889	5.74200	9.93464	MA
chr1	1067962	1068385	424	1068198	36.00	31.66166	11.70094	27.79599	MA
chr1	1166760	1166985	226	1166904	20.00	12.17439	5.88873	8.69629	MA
chr1	1243703	1244003	301	1243836	21.00	13.39170	6.31714	9.87886	MA
chr1	1259811	1260132	322	1259992	29.00	21.85153	8.73402	18.15079	MA
chr1	1334967	1335265	299	1335111	20.00	11.42774	5.50202	7.97264	MA
chr1	1342501	1342793	293	1342644	25.00	17.22777	7.38974	13.62078	MA
chr1	1447327	1447666	340	1447493	56.00	47.91687	13.02038	43.83786	MA
chr1	1550543	1550871	329	1550704	51.00	41.78180	11.84595	37.77581	MA
chr1	1590469	1590878	410	1590600	29.00	18.83682	7.24806	15.19444	MA

Figure 5. Screenshot of Galaxy showing the results of peak calling from MACS2. Columns 1, 2, and 3 display NRF1 protein peak location and column 8 shows fold enrichment (FE). We discarded values below 5 FE.

Gene Annotation (Identifying target genes). GREAT (Genomic Regions Enrichment of Annotations Tool), a webserver available at http://great.stanford.edu/public/html/, is one of several tools available to identify NRF1 target genes based on ChIP-Seq results. GREAT (McLean et al., 2010) associates the TF's binding sites identified during peak calling with TF's putative target genes by assigning to each gene those peaks that fall within a previously defined gene regulatory domain. By default, the software establishes the gene regulatory domain as 5,000 bp upstream and 1,000 bp downstream of the TSS; however, the user can modify these parameters. GREAT also takes into account distal binding sites found in the literature as curated domains; however, the user has the option of modifying the settings to exclude them.

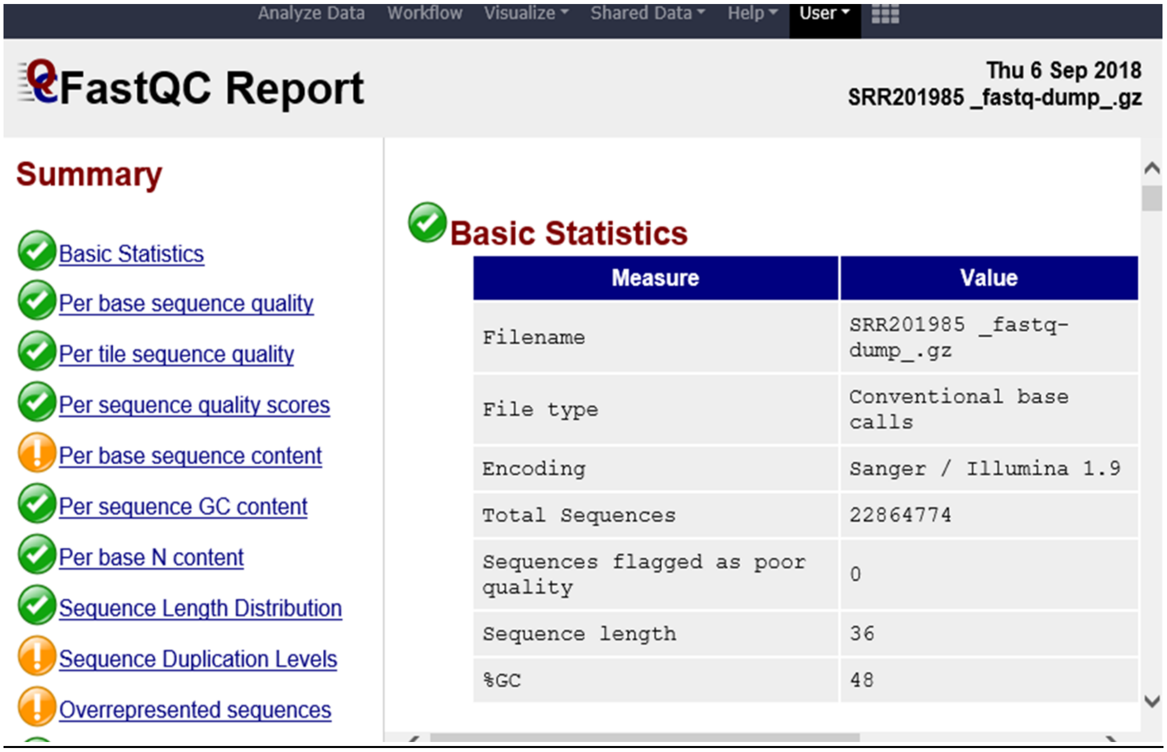
RNA-Seq

Overview of RNA-Seq. RNA-Seq is widely used by the life science scientific community, among other procedures, because it allows combining two pieces into one experiment—the sequence discovery and the quantification of RNA, the key intermediary between DNA and proteasome. Design of a good experiment is the first step in successful completion of an RNA-Seq study. Experimental design includes selection of the appropriate library type, number of replicates, and sequencing depth. There is abundant scientific literature and many variants of RNA-Seq protocols for conducting RNA-Seq assays and computational analysis tools to process the results. Choosing the right features is therefore not an easy task, especially for new users. The right choices depend on the biological system under study and the research question being addressed (Conesa et al. 2016).

Sequencing depth (or library size) is the number of sequence reads for a single sample. If the sequencing is deeper, the number of transcripts detected will be larger and the quantification will be more accurate (Conesa et al., 2016; Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008). It is difficult to establish an optimal level; although some scientists consider five million reads enough to precisely quantify genes with medium to high expression levels in eukaryotic RNA-Seq experiments, other researchers recommend sequences up to one hundred million reads to accurately measure genes with low expression levels (Conesa et al., 2016; Sims, Sudbery, Ilott, Heger, & Ponting, 2014).

Another important aspect of RNA-Seq experimental design is the number of replicates. A minimum number of three replicates is recommended, or use of software to calculate the number of replicates based on the desired statistical power (Conesa et al., 2016). Once the experiment is completed, the first step in analysis of the RNA-Seq data is to evaluate the quality of sequencing reads.

Quality control of Next Generation Sequence (NGS). Quality control (QC) of raw RNA-Seq data (sequence reads in FastQC file) retrieved from GEO was performed following the same procedure and using the same program (FastQC) utilized for QC of ChIP-Seq. (See previous section quality control of sequencing reads for details.) Figure 6 shows a screenshot with a partial view of the report generated by FastQC for RNA-Seq of human mammary epithelial cells (HMEC) GSM721141-SRR201985. FastQC, as noted, is available through Galaxy in the tools tab to the left, under “NGS: QC and manipulation” and also at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Andrews, 2010).



FastQC Report

Thu 6 Sep 2018
SRR201985 _fastq-dump_.gz

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences

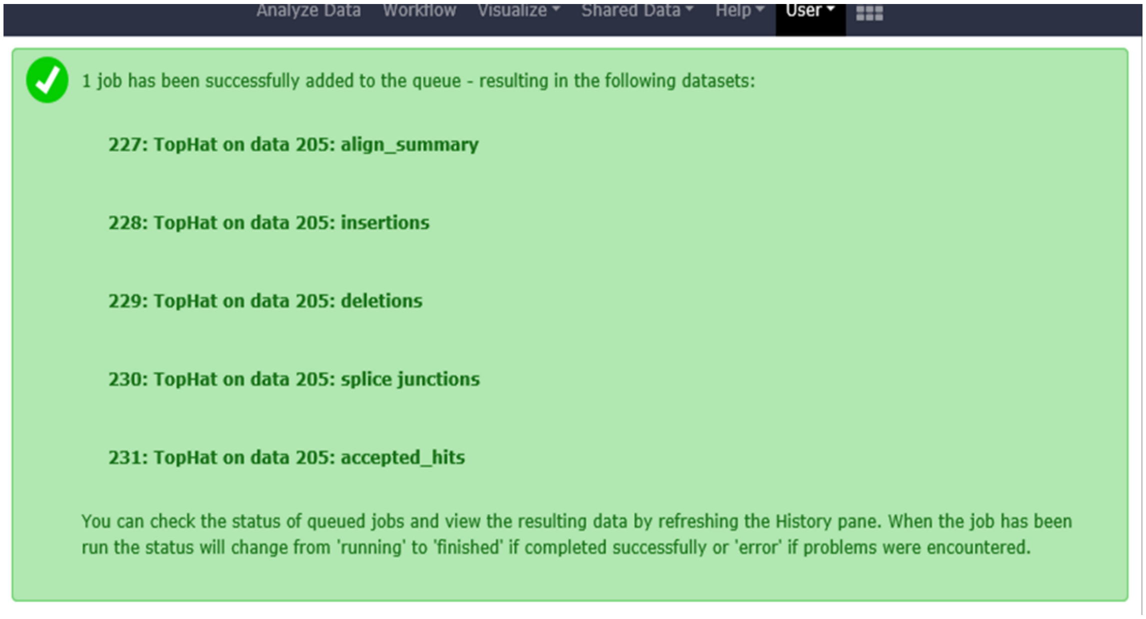
✓ Basic Statistics

Measure	Value
Filename	SRR201985 _fastq-dump_.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	22864774
Sequences flagged as poor quality	0
Sequence length	36
%GC	48

Figure 6. Screenshot with a partial view of the Galaxy report generated by FastQC for RNA-Seq of Human Mammary Epithelial Cells (HMEC). GEO accession reference GSM721141-SRR201985.

Mapping (Generation of Alignments). After evaluation of the quality of RNA sequencing readings, the next step is identification of transcripts by mapping RNA-Seq reads onto the genome. We mapped the raw RNA–Seq files retrieved from GEO against Homo sapiens genome reference GRCh37/hg19 using the TopHat program (Galaxy Version 2.1.1) (Kim et al., 2013). TopHat can be accessed through Galaxy under the submenu entitled NGS: RNA Analysis, located on the left side window. One of the difficulties of mapping RNA-Seq reads compared to DNA-Seq alignment is that genes contains introns while sequencing readings come from mature mRNA transcripts where introns have been removed (spliced).

A second challenge of the software is the presence of pseudogenes in the genome that are very similar (sequentially speaking) to functional genes which can causes incorrect alignment. Different software programs use different algorithms to deal with these challenges. The new version of TopHat (TopHat2) incorporates new features to ensure that reads are also aligned with true insertions and deletions (Kim et al., 2013). TopHat output contains much information distributed in five different files: align summary, insertions, deletions, splice junctions and accepted hits. Figure 7 shows a screenshot of Galaxy listing TopHat output files. Although all files are important, for our specific goal we focused on accepted hits, the file that contains all the valid alignments to be used in transcript quantification, our next step.



1 job has been successfully added to the queue - resulting in the following datasets:

227: TopHat on data 205: align_summary

228: TopHat on data 205: insertions

229: TopHat on data 205: deletions

230: TopHat on data 205: splice junctions

231: TopHat on data 205: accepted_hits

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

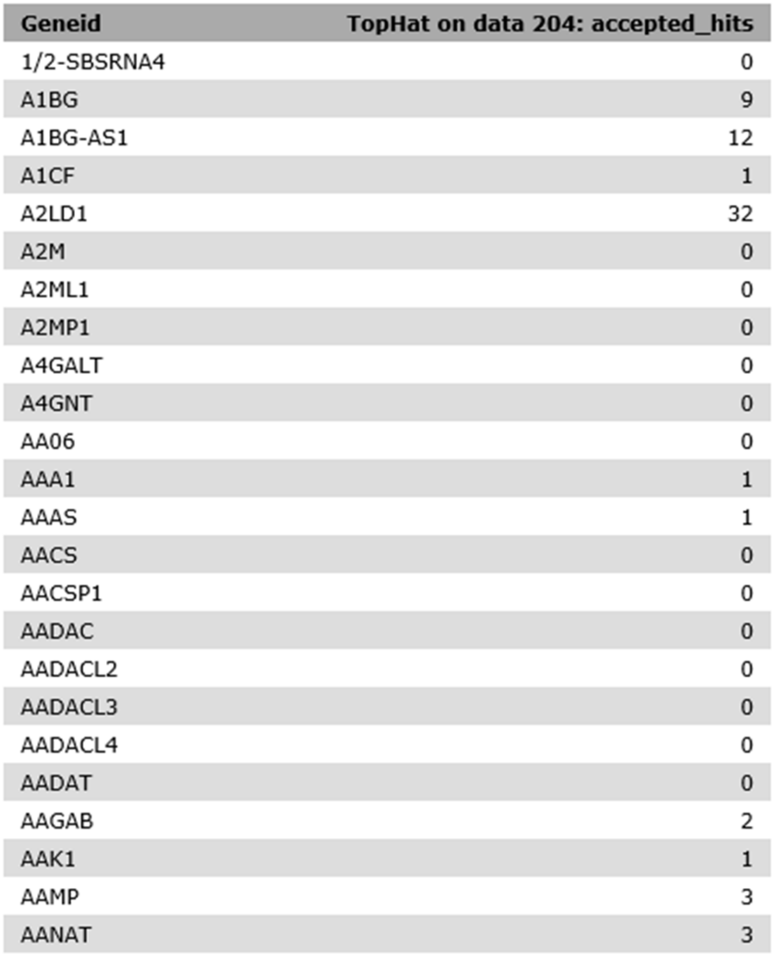
Figure 7. Screenshot of Galaxy listing the five TopHat output files from reads alignment: align summary, insertions, deletions, splice junctions, and accepted hits. Although al files are important, depending on the specific research question, for our purpose of quantifying number of reads to find differentially expressed genes, accepted hits was our file of interest.

Transcript quantification. There are several programs to quantify the number of readings that have been mapped to each transcript sequence. We used Htseq-count program (Anders, Pyl & Huber, 2015) to aggregate raw counts and assign them to genes. Two files are needed as input in the Htseq-count program: the alignments file labeled “accepted hits” (BAM format) generated by TopHat, and the annotated reference transcriptome in GTF format. For the latter, we used the human reference gene dataset GRCh37/hg19 generated by iGenomes (https://support.illumina.com/sequencing/sequencing_software/igenome.html).

Illumina iGenomes contains data downloaded from UCSC, NCBI, or Ensembl, and the GRCh37/hg19 file can be downloaded directly into Galaxy with use of the shared data option in the menu bar. Other available sources that can be used to access the Human reference dataset containing the gene locations in the appropriate GTF format are UCSC Genome Browser (http://genome.ucsc.edu/) and Gencode (https://www.gencodegenes.org/). Figure 8 is a partial view of Htseq-count output showing the number of reads assigned to each gene.

Identification of Differential Expressed (DE) Genes

For comparison of expression levels, raw counts must be normalized to address different aspects, such as sequencing biases, transcript length, and total number of reads. Reads per kilobase per million reads (RPKM) or FPKM (Fragments per Kilobase Million) are the units used to account for transcript length and library size factors. TPM (Transcripts per Kilobase Million) is now also widely used; however, a formula can easily convert FPKM into TPM (Conesa et al., 2016; Pachter, 2011).

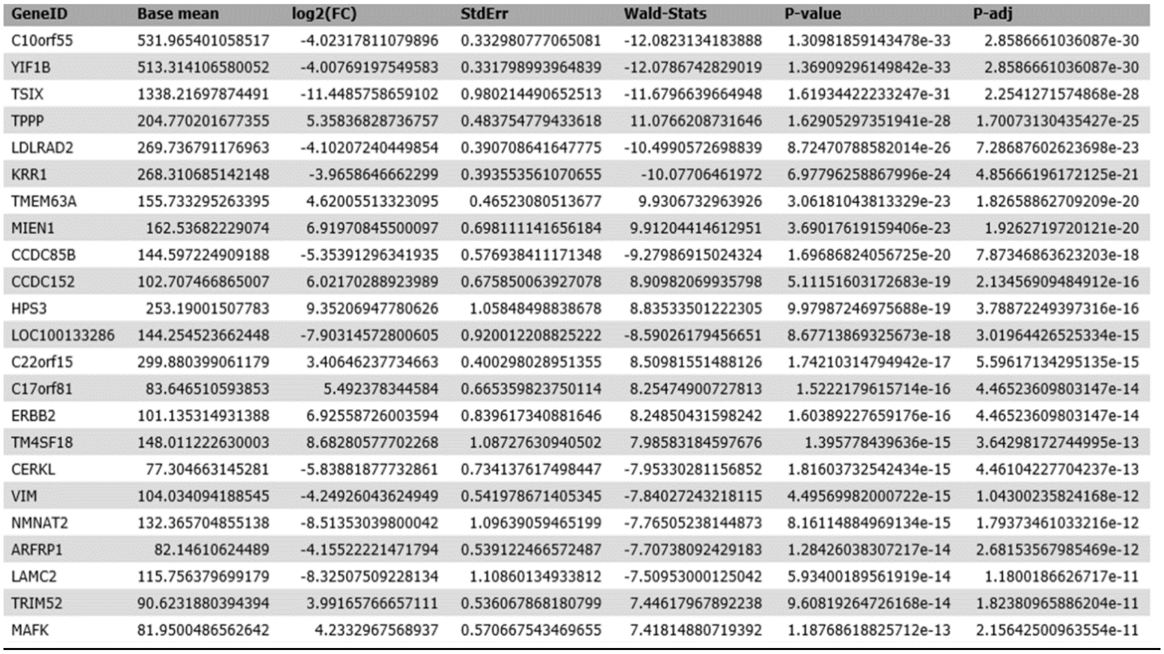


Geneid	TopHat on data 204: accepted_hits
1/2-SBSRNA4	0
A1BG	9
A1BG-AS1	12
A1CF	1
A2LD1	32
A2M	0
A2ML1	0
A2MP1	0
A4GALT	0
A4GNT	0
AA06	0
AAA1	1
AAAS	1
AACS	0
AACSP1	0
AADAC	0
AADACL2	0
AADACL3	0
AADACL4	0
AADAT	0
AAGAB	2
AAK1	1
AAMP	3
AANAT	3

Figure 8. Screenshot with a partial view of Galaxy window showing the number of reads assigned to each gene generated by Htseq-count. For comparison of expression levels to find differentially expressed genes, these raw counts need to be normalized.

With the aim of identifying genes in breast cancer cells with statistically significant differential expression (DE), compared to normal mammary epithelial cells, we used DESeq2 program (Love, Huber, & Anders, 2014) accessed within Galaxy platform. DESeq2 carries out the normalization and quantitative analysis of count data (previously generated by Htseq-count) based primarily on statistical strength rather than on the amount of differential expression (Love et al., 2014).

For illustration purposes, Figure 9 shows a screenshot of DESeq2 output displaying the top DE genes ranked by adjusted *p* value.



GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
C10orf55	531.965401058517	-4.02317811079896	0.332980777065081	-12.0823134183888	1.30981859143478e-33	2.8586661036087e-30
YIF1B	513.314106580052	-4.00769197549583	0.331798993964839	-12.0786742829019	1.36909296149842e-33	2.8586661036087e-30
TSIX	1338.21697874491	-11.4485758659102	0.980214490652513	-11.6796639664948	1.61934422233247e-31	2.2541271574868e-28
TPPP	204.770201677355	5.35836828736757	0.483754779433618	11.0766208731646	1.62905297351941e-28	1.70073130435427e-25
LDLRAD2	269.736791176963	-4.10207240449854	0.390708641647775	-10.4990572698839	8.72470788582014e-26	7.28687602623698e-23
KRR1	268.310685142148	-3.9658646662299	0.393553561070655	-10.07706461972	6.97796258867996e-24	4.85666196172125e-21
TMEM63A	155.733295263395	4.62005513323095	0.46523080513677	9.9306732963926	3.06181043813329e-23	1.82658862709209e-20
MIEN1	162.53682229074	6.91970845500097	0.698111141656184	9.91204414612951	3.69017619159406e-23	1.9262719720121e-20
CCDC85B	144.597224909188	-5.35391296341935	0.576938411171348	-9.27986915024324	1.69686824056725e-20	7.87346863623203e-18
CCDC152	102.707466865007	6.02170288923989	0.675850063927078	8.90982069935798	5.11151603172683e-19	2.13456909484912e-16
HPS3	253.19001507783	9.35206947780626	1.05848498838678	8.83533501222305	9.97987246975688e-19	3.78872249397316e-16
LOC100133286	144.254523662448	-7.90314572800605	0.920012208825222	-8.59026179456651	8.67713869325673e-18	3.01964426525334e-15
C22orf15	299.880399061179	3.40646237734663	0.400298028951355	8.50981551488126	1.74210314794942e-17	5.59617134295135e-15
C17orf81	83.646510593853	5.492378344584	0.665359823750114	8.25474900727813	1.5222179615714e-16	4.46523609803147e-14
ERBB2	101.135314931388	6.92558726003594	0.839617340881646	8.24850431598242	1.60389227659176e-16	4.46523609803147e-14
TM4SF18	148.011222630003	8.68280577702268	1.08727630940502	7.98583184597676	1.395778439636e-15	3.64298172744995e-13
CERKL	77.304663145281	-5.83881877732861	0.734137617498447	-7.95330281156852	1.81603732542434e-15	4.46104227704237e-13
VIM	104.034094188545	-4.24926043624949	0.541978671405345	-7.84027243218115	4.49569982000722e-15	1.04300235824168e-12
NMNAT2	132.365704855138	-8.51353039800042	1.09639059465199	-7.76505238144873	8.16114884969134e-15	1.79373461033216e-12
ARFRP1	82.14610624489	-4.15522221471794	0.539122466572487	-7.70738092429183	1.28426038307217e-14	2.68153567985469e-12
LAMC2	115.756379699179	-8.32507509228134	1.10860134933812	-7.50953000125042	5.93400189561919e-14	1.1800186626717e-11
TRIM52	90.6231880394394	3.99165766657111	0.536067868180799	7.44617967892238	9.60819264726168e-14	1.82380965886204e-11
MAFK	81.9500486562642	4.2332967568937	0.570667543469655	7.41814880719392	1.18768618825712e-13	2.15642500963554e-11

Figure 9. Screenshot of DESeq2 output displaying the top DE genes ranked by adjusted *p* value.

Integration of ChIP-Seq and RNA-Seq Data Analysis

We used BETA (Wang et al., 2013), a software that integrates TF binding and differential expression, to identify target genes displaying significant statistical changes in gene expression that may be attributable to TF regulation activity. BETA requires two different dataset inputs: ChIP-Seq and DE expression. With this information, BETA calculates two scores, a binding potential rank (Rgb) and a differential expression rank (Rge). The first score (Rgb) measures the regulatory potential of the TF on the expression of the target gene. Rgb is calculated by modeling the influence of all binding sites falling within an established distance from the Transcription Start Site (TSS; default is 100 kb), using a monotonically decreasing function based on the distance from each peak to TSS. The gene with the highest regulatory potential is scored Rgb = 1. The second score (Rge) is calculated based on differential expression assigning Rge = 1 to the gene with the strongest differential expression based on *p* value.

BETA calculates the rank product (RP), which is the multiplication of Rgb and Rge, that can be considered as the *p* value showing the probability of a gene regulatory potential and significant differentially expressed (Wang et al., 2013). For consideration of a gene to be very likely regulated by NRF1, we established as cutoff point RP = 10^{-3}. Finally, BETA conducts motif analysis to identify enriched DNA sequences in the ChIP-Seq binding summits, representing them in position-specific scoring matrices (PSSM) (Wang et al., 2013). BETA is available to the public at http://cistrome.org/BETA/.

Results and Discussion

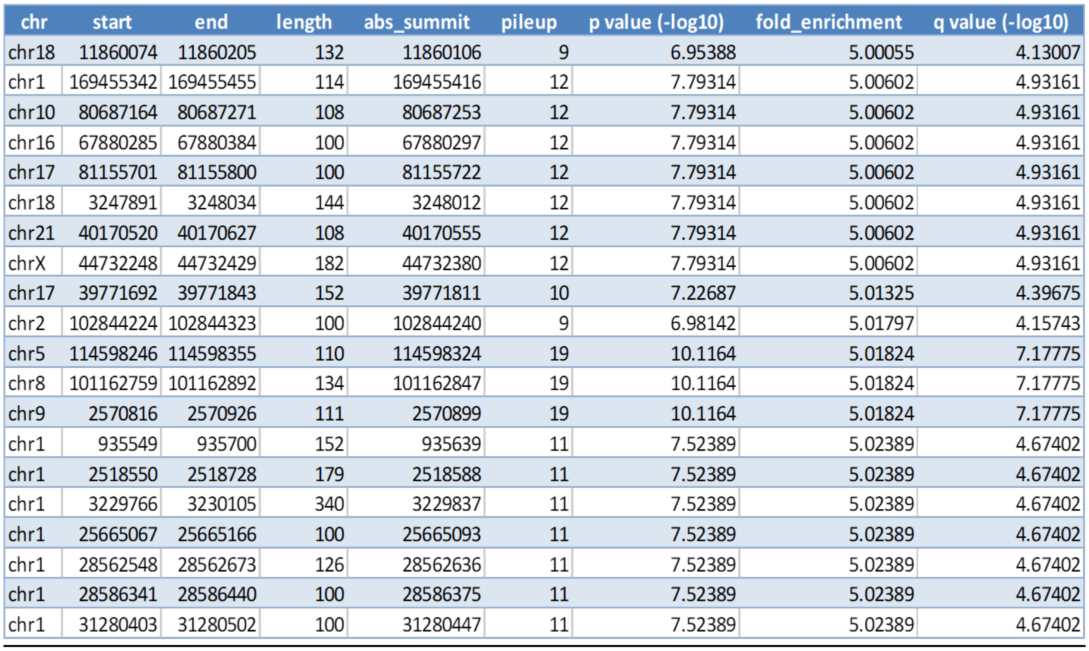
New in vitro studies have proven that binding of some specific transcription factors and histone modifications can predict gene expression; conversely, changes in gene expression are correlated to chromatin marks and changes in transcription factor binding (Cheng et al., 2012; Klein et al., 2014; Ouyang, Zhou, & Wong, 2009). Computational analysis of NRF1 ChIP-Seq data provides important evidence about location of NRF1-DNA binding sites, including the relative amount of NRF1 protein, but this information is not sufficient to determine NRF1 regulation. To establish possible NRF1 regulation, we further analyzed the relationship between gene expression of NRF1 target genes and factor-binding sites, which is not a simple one-to-one relationship (Wang et al., 2013). Therefore, in addition to ChIP-Seq data, we used the transcriptome (gene expression) data obtained from the computational analysis of RNA-Seq data in normal mammary cells (HMEC) and breast cancer cells (HCC1954) to investigate how changes in NRF1 activity affect the expression of NRF1 targets genes.

HCC1954 cells represent a good model of breast ductal carcinoma (ER/PR-ve Her2 +ve) with amplified HER2 and high abundance of EGFR. It has been reported that NRF1 activity, measured based on the activity of target genes, is increased in different malignant tumors, including breast cancer (Falco, Bleda, Carbonell-Caballero, & Dopazo, 2016). Computational analysis of NRF1 ChIP-Seq in HCC1954 breast cancer cells resulted in 21,400 binding sites with fold enrichment greater than 5. Table 1 shows the location of the lowest 20 peaks ranked by fold enrichment.

The lowest peak is in accordance with the established fold enrichment cutoff value of 5. Differential expression analysis of RNA–Seq in breast cancer cells (HCC1954) were compared to normal mammary epithelial cells (HMEC) using DESeq2. This comparison revealed 390 genes with statistically significant DE (adjusted *p* value < 0.05).

Table 1

Output From MACS2 Listing the Lowest 20 Peaks Ranked by Fold Enrichment



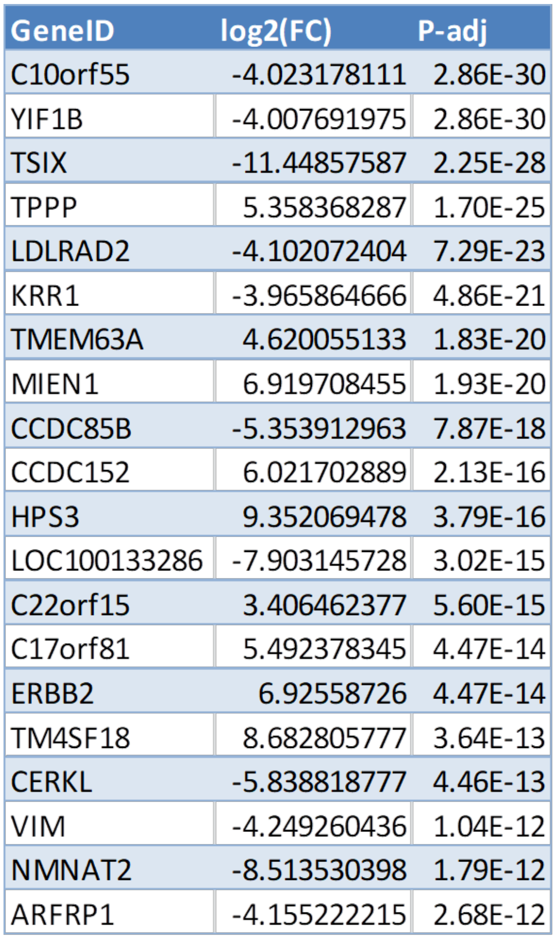
chr	start	end	length	abs_summit	pileup	p value (-log10)	fold_enrichment	q value (-log10)
chr18	11860074	11860205	132	11860106	9	6.95388	5.00055	4.13007
chr1	169455342	169455455	114	169455416	12	7.79314	5.00602	4.93161
chr10	80687164	80687271	108	80687253	12	7.79314	5.00602	4.93161
chr16	67880285	67880384	100	67880297	12	7.79314	5.00602	4.93161
chr17	81155701	81155800	100	81155722	12	7.79314	5.00602	4.93161
chr18	3247891	3248034	144	3248012	12	7.79314	5.00602	4.93161
chr21	40170520	40170627	108	40170555	12	7.79314	5.00602	4.93161
chrX	44732248	44732429	182	44732380	12	7.79314	5.00602	4.93161
chr17	39771692	39771843	152	39771811	10	7.22687	5.01325	4.39675
chr2	102844224	102844323	100	102844240	9	6.98142	5.01797	4.15743
chr5	114598246	114598355	110	114598324	19	10.1164	5.01824	7.17775
chr8	101162759	101162892	134	101162847	19	10.1164	5.01824	7.17775
chr9	2570816	2570926	111	2570899	19	10.1164	5.01824	7.17775
chr1	935549	935700	152	935639	11	7.52389	5.02389	4.67402
chr1	2518550	2518728	179	2518588	11	7.52389	5.02389	4.67402
chr1	3229766	3230105	340	3229837	11	7.52389	5.02389	4.67402
chr1	25665067	25665166	100	25665093	11	7.52389	5.02389	4.67402
chr1	28562548	28562673	126	28562636	11	7.52389	5.02389	4.67402
chr1	28586341	28586440	100	28586375	11	7.52389	5.02389	4.67402
chr1	31280403	31280502	100	31280447	11	7.52389	5.02389	4.67402

Note. It should be noticed how these binding sites meet the established fold enrichment cutoff value of 5.0.

Table 2 show the top 20 DE genes ranked by adjusted *p* value. To identify among all DE genes the up- or downregulated NRF1 targets, we used BETA in the cistrome Galaxy platform. The list of differentially genes in tab-delimited text files and the list of NRF1 binding sites in BED format were input into BETA to obtain the following outputs: the activating and repressive function prediction of NRF1, the list of inferred NRF1 upregulated and downregulated targets, and the results of NRF1 motif analysis.

Table 2

Results of Differential Expression (DE) Analysis Of RNA-Seq in Breast Cancer Cells (HCC1954) Compared to Normal Mammary Epithelial Cells (HMEC) Using Deseq2 Revealed 390 Genes With Statistically Significant DE (Adjusted p Value < 0.05)



GeneID	log2(FC)	P-adj
C10orf55	-4.023178111	2.86E-30
YIF1B	-4.007691975	2.86E-30
TSIX	-11.44857587	2.25E-28
TPPP	5.358368287	1.70E-25
LDLRAD2	-4.102072404	7.29E-23
KRR1	-3.965864666	4.86E-21
TMEM63A	4.620055133	1.83E-20
MIEN1	6.919708455	1.93E-20
CCDC85B	-5.353912963	7.87E-18
CCDC152	6.021702889	2.13E-16
HPS3	9.352069478	3.79E-16
LOC100133286	-7.903145728	3.02E-15
C22orf15	3.406462377	5.60E-15
C17orf81	5.492378345	4.47E-14
ERBB2	6.92558726	4.47E-14
TM4SF18	8.682805777	3.64E-13
CERKL	-5.838818777	4.46E-13
VIM	-4.249260436	1.04E-12
NMNAT2	-8.513530398	1.79E-12
ARFRP1	-4.155222215	2.68E-12

Note. These are the top 20 DE genes ranked by adjusted *p* value.

Activating and Repressive Function Prediction

The likelihood of a gene to be regulated by NRF1 (regulatory potential) is individually estimated by BETA for each gene and calculated as Sg =

$\sum_{i=1}^{k} e^{-(0.5+4\Delta i)}$, the sum of the regulatory potential of all NRF1 binding peaks (*k*) within a specified distance from TSS (+/- 5 Kb in our case). Δi is the distance between the binding site *i* and the TSS divided by 100 kb (for example, for 5 kb, Δi = 0.05). This equation is a function that decreases monotonically as the distance of each binding site from TSS increases. The shape of the equation is a good approximation of empirical data of the binding site's distance to TSS and differentially expressed genes obtained from many ChIP-Seq tests (Tang et al., 2011).

Based on regulatory potential and the DE list, genes are divided into three groups: upregulated, downregulated, and unchanged, as shown in Figure 10. Dotted lines represent the genes with no changes. Red lines represent the upregulated, and blue line the downregulated groups.

Among the top 15% ranked up- and downregulated genes, there is a slightly higher NRF1 regulatory potential in upregulated genes compared to the downregulated genes. That is, among this group of genes the ones with a gain in gene expression tend to have also a higher enrichment of NRF1 biding sites (red line in Table 10). It can also be observed that after the 15% ranked up- and downregulated genes, the NRF1 enrichment pattern changes and the downregulated genes tend to have a higher enrichment of NRF1 binding sites. Values listed at the top of Figure 10 are the *p* values of the Kolmogorov-Smirnov test used to determine the significance of the difference in NRF1 binding between the up- and downregulated genes, compared to the unchanged genes. These results (0.995 and 1) indicate that there was no significant difference.

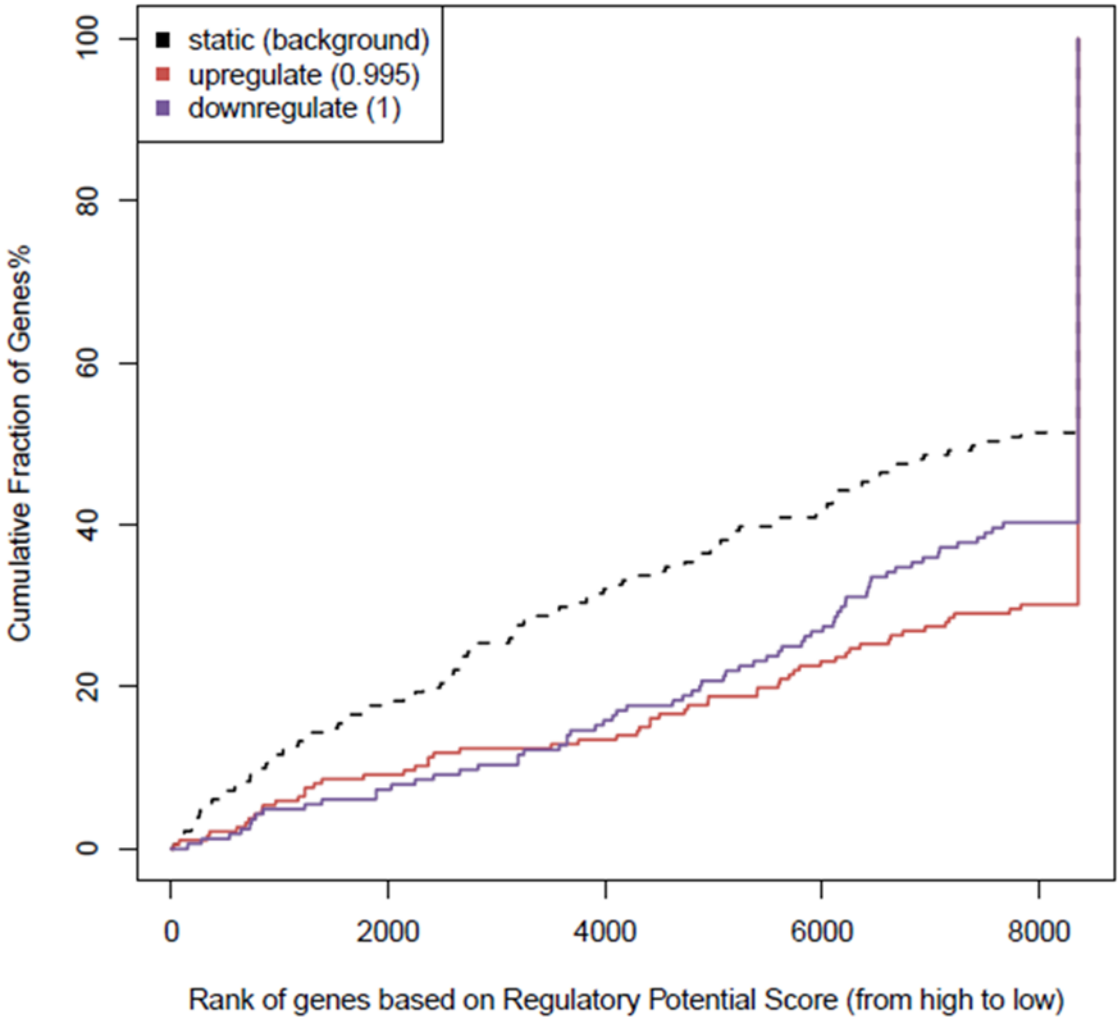


Figure 10. BETA output of activating / repressive function prediction of NRF1 in HER2 enriched breast cancer cells HCC1954.

Note. The dotted lines represent the genes with no changes, the red line the upregulated, and purple line the downregulated genes. Within the top 15% ranked up- and downregulated genes, there is a slightly higher NRF1 regulatory potential in the upregulated ones. This difference means that among this group of genes the ones with a gain in gene expression tend to have also a higher enrichment in NRF1 biding sites (the red is above purple line). Values listed on top are the *p* values of the Kolmogorov-Smirnov test used to determine the significance of the difference in NRF1 binding between the up- and downregulated genes compared to the unchanged genes.

Overall, the activating and repressive function prediction generated by the BETA algorithm shows that the increase of NRF1 activity in breast cancer cells can be either an activator or repressor of target genes. There is no established pattern of NRF1 enrichment for predicting a specific role. This outcome suggests that additional elements, cofactors, or combination of them may play an important role in explaining the repressive or activating role of NRF1 in the changes of expression (DE) of target genes in HCC1954 breast cancer cells.

Direct NRF1 Target Prediction

Prediction of NRF1 target genes is accomplished by BETA combining the binding potential rank (Rgb) with the differential expression rank (Rge) and calculating the rank product (RP). This combination is the basis for consideration of a gene as predicted NRF1 target, provided the established cutoff value is met (we used RP = 10^{-3}). Results indicate that out of 390 genes with significant differential expression 63 were predicted NRF1 upregulated targets (Table 3) and 73 were NRF1 downregulated targets (Table 4).

Table 3

Predicted NRF1 Target Upregulated Genes in HCC1954 Breast Cancer Cells Ranked by Rank Product (RP)

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr5	659976	693510	NM_007030	2.85E-05	-	TPPP
chr1	244998638	245008359	NM_198076	7.97E-05	+	COX20
chr1	226033232	226070420	NM_014698	1.10E-04	-	TMEM63A
chr16	30007529	30017111	NM_173618	2.02E-04	+	INO80E
chr2	148687965	148778316	NM_001190879	3.97E-04	-	ORC4
chr19	46268042	46272497	NM_175875	4.33E-04	-	SIX5
chr8	119201694	119634184	NM_001101676	4.93E-04	-	SAMD12
chr5	180683385	180688119	NM_032765	6.81E-04	-	TRIM52
chr1	225997835	226033262	NM_001291163	8.19E-04	+	EPHX1
chr5	470624	473080	NR_024158	8.77E-04	-	PP7080

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr19	3750770	3761673	NM_004886	1.19E-03	-	APBA3
chr16	3074031	3077756	NM_024339	1.30E-03	+	THOC6
chr4	699572	764427	NM_006315	1.31E-03	+	PCGF3
chr14	37667117	38020464	NM_0011195296	1.36E-03	+	MIPOL1
chr20	32077927	32237837	NM_0010322999	1.58E-03	+	CBFA2T2
chr11	102267055	102323775	NM_052932	1.67E-03	-	TMEM123
chr16	69151911	69166493	NR_033227	1.74E-03	-	CHTF8
chr4	56294067	56413076	NM_004898	1.74E-03	-	CLOCK
chr3	113367232	113415493	NR_111981	2.13E-03	-	KIAA2018
chr7	27210209	27213955	NM_018951	2.18E-03	-	HOXA10
chr9	97872507	98079991	NM_0012433744	2.20E-03	-	FANCC
chr5	112357795	112824527	NM_0010855377	2.24E-03	-	MCC

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr17	29109701	29151778	NM_015986	2.32E-03	-	CRLF3
chr9	15464064	15511003	NM_001128217	2.47E-03	-	PSIP1
chr22	22051825	22090123	NM_013313	2.89E-03	-	YPEL1
chr4	175411327	175444044	NM_001256301	3.12E-03	-	HPGD
chrX	30845558	30907511	NM_152787	3.33E-03	-	TAB3
chr1	110276553	110283660	NM_000849	3.37E-03	-	GSTM3
chr2	85832375	85839179	NM_001013649	3.40E-03	-	C2orf68
chr16	67562719	67580691	NM_001193522	3.45E-03	+	FAM65A
chr21	18965967	18985268	NM_006806	3.49E-03	-	BTG3
chr2	97481990	97501121	NM_017623	3.52E-03	+	CNNM3
chr22	31318294	31322640	NR_026920	3.65E-03	+	MORC2-AS1
chr7	134850531	134855578	NM_001243754	3.78E-03	-	C7orf49

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr22	28315363	28320951	NR_026962	3.82E-03	+	TTC28-AS1
chr6	34433837	34503000	NM_020804	3.95E-03	+	PACSIN1
chr1	6281252	6296044	NM_012405	3.96E-03	-	ICMT
chr19	33699569	33716756	NM_019849	4.24E-03	-	SLC7A10
chr9	5784571	5833081	NM_024896	4.25E-03	-	ERMP1
chrX	13053735	13062917	NM_174901	4.30E-03	-	FAM9C
chr5	108670409	108745675	NM_014819	4.50E-03	-	PJA2
chr1	56960418	57045257	NM_003713	4.76E-03	-	PPAP2B
chr8	74332603	74659162	NM_001164380	4.77E-03	-	STAU2
chr17	29158987	29222883	NM_024857	4.86E-03	+	ATAD5
chr12	1100373	1605099	NR_027946	5.23E-03	+	ERC1
chr16	67313426	67323403	NM_001129731	5.43E-03	+	PLEKHG4

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr16	69599868	69738569	NM_006599	5.44E-03	+	NFAT5
chr6	43543877	43588260	NM_006502	5.58E-03	+	POLH
chr17	4442190	4458681	NM_014520	5.68E-03	-	MYBBP1A
chr22	35937351	35950045	NM_014310	5.72E-03	+	RASD2
chr17	46018888	46026674	NM_018129	5.80E-03	+	PNPO
chr17	7761063	7765600	NM_144607	6.13E-03	+	CYB5D1
chr17	37219555	37307902	NM_020405	6.33E-03	-	PLXDC1
chr12	6571403	6580065	NM_016830	7.23E-03	-	VAMP1
chr17	43224683	43229468	NM_006460	7.78E-03	+	HEXIM1
chr19	14247963	14282075	NR_045214	7.86E-03	+	LOC100507373
chr8	95892452	95907482	NM_057749	7.93E-03	-	CCNE2
chr11	65265232	65273939	NR_002819	8.30E-03	+	MALAT1

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr8	117886662	117889107	NR_033886	1.09E-02	+	RAD21-AS1
chr19	10982252	11033448	NM_199141	1.11E-02	+	CARM1
chr6	28317690	28336954	NM_024493	1.20E-02	+	ZKSCAN3
chr7	99647416	99662663	NM_145914	1.32E-02	+	ZSCAN21
chr1	228395830	228548951	NM_052843	1.36E-02	+	OBSCN

Note. These results can be interpreted as the *p* value of the likelihood of being NRF1 regulated based on integrative analysis of NRF1 binding peaks next to TSS (+/-5.0 kb) and differential expression.

Table 4

Predicted NRF1 Target Downregulated Genes in HCC1954 Breast Cancer Cells Ranked by Rank Product (RP)

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr19	38794199	38806445	NM_001039671	3.86E-05	-	YIF1B
chr12	75891418	75905418	NM_007043	2.31E-04	-	KRR1
chr18	12328942	12377275	NM_006796	2.74E-04	-	AFG3L2
chr3	156390959	156393502	NR_027954	3.49E-04	-	TIPARP-AS1
chr7	36363758	36429734	NM_001100425	4.55E-04	-	KIAA0895
chr11	65657874	65659106	NM_006848	5.26E-04	+	CCDC85B
chr4	57844805	57897328	NM_000938	7.01E-04	+	POLR2B
chr4	186320693	186347139	NM_018359	7.67E-04	-	UFSP2
chr20	62329994	62339365	NM_001267546	9.04E-04	-	ARFRP1
chr10	17270257	17279592	NM_003380	9.09E-04	+	VIM
chr14	67827033	67853233	NM_004094	9.17E-04	+	EIF2S1
chr1	53392900	53517289	NM_001193617	1.19E-03	+	SCP2
chr18	11883471	11908796	NM_001242904	1.30E-03	-	MPPE1

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr8	144989314	145018126	NM_201382	1.39E-03	-	PLEC
chr16	56642477	56643409	NM_005953	1.46E-03	+	MT2A
chr7	134464163	134655480	NM_033138	1.47E-03	+	CALD1
chr10	33189245	33247293	NM_002211	1.59E-03	-	ITGB1
chr10	99400442	99436189	NM_018425	1.60E-03	+	PI4K2A
chr12	50017196	50038452	NM_001031698	1.69E-03	+	PRPF40B
chr19	48828628	48833810	NM_001425	1.86E-03	+	EMP3
chr7	5632435	5646287	NM_003088	1.94E-03	+	FSCN1
chr4	145915726	146019371	NM_001256706	1.95E-03	-	ANAPC10
chr12	76419226	76425556	NM_007350	2.50E-03	-	PHLDA1
chr12	28111016	28122894	NM_198966	2.51E-03	-	PTHLH
chr10	101468504	101492423	NM_078470	2.67E-03	-	COX15
chr2	173292313	173371181	NM_000210	2.67E-03	+	ITGA6
chr14	71189242	71275888	NM_033141	2.77E-03	-	MAP3K9
chr18	21452983	21535029	NM_000227	2.84E-03	+	LAMA3
chr22	38339056	38349654	NM_032561	2.89E-03	-	C22orf23
chr1	153533584	153538306	NM_005978	3.41E-03	-	S100A2

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr1	157963062	158070052	NM_001286349	3.54E-03	+	KIRREL
chr1	33116748	33151812	NM_001135255	3.61E-03	+	RBBP4
chr16	87863628	87903100	NM_003486	3.74E-03	-	SLC7A5
chr7	116164838	116201239	NM_001753	4.13E-03	+	CAV1
chr12	57482676	57489259	NM_005967	4.43E-03	+	NAB2
chr2	235401685	235405693	NM_005737	4.52E-03	-	ARL4C
chr1	11714913	11723384	NM_183413	4.76E-03	+	FBXO44
chr11	18415935	18429765	NM_001165415	4.78E-03	+	LDHA
chr7	27179982	27195547	NR_038832	4.93E-03	+	HOXA-AS3
chr9	130267616	130331396	NM_022833	5.22E-03	-	FAM129B
chr13	98795351	99102027	NM_005766	5.37E-03	+	FARP1
chr16	57496550	57505921	NM_032940	6.01E-03	+	POLR2C
chr10	112631552	112659764	NM_014456	6.20E-03	+	PDCD4
chr1	25071759	25170815	NM_013943	6.29E-03	+	CLIC4
chr20	5100231	5100615	NR_028370	6.30E-03	+	PCNA-AS1
chr17	2207236	2228558	NM_021947	6.32E-03	+	SRR
chr17	7486964	7491527	NR_024603	6.59E-03	+	MPDU1

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr1	85742040	85743771	NR_045484	6.64E-03	+	LOC646626
chr11	62201013	62314332	NM_024060	6.66E-03	-	AHNAK
chr19	45909466	45914024	NM_001297590	6.69E-03	+	CD3EAP
chr19	38794803	38795646	NM_033520	6.93E-03	+	C19orf33
chr15	39873279	39889668	NM_003246	7.08E-03	+	THBS1
chr1	38268613	38273865	NM_024640	7.09E-03	-	YRDC
chr4	122722471	122738176	NM_001034194	7.15E-03	+	EXOSC9
chr19	5691844	5720463	NM_001276480	7.18E-03	-	LONP1
chr2	219081816	219119071	NM_152862	7.75E-03	+	ARPC2
chr14	70232999	70234430	NR_029378	7.89E-03	-	LOC100289511
chr1	45271581	45272957	NM_001013632	8.28E-03	-	TCTEX1D4
chr11	105921824	105948465	NM_152433	8.58E-03	-	KBTBD3
chr15	66782665	66790146	NM_006049	8.65E-03	-	SNAPC5
chr2	231577556	231685790	NM_016289	8.71E-03	+	CAB39
chr8	38268655	38326352	NM_001174064	8.76E-03	-	FGFR1

Chroms	txStart	txEnd	refseqID	rank product	Strands	Gene Symbol
chr4	159045731	159093718	NM_016613	9.35E-03	-	FAM198B
chr1	243419306	243663393	NM_006642	9.62E-03	+	SDCCAG8
chr1	109472129	109506121	NM_001048210	9.80E-03	-	CLCC1
chr19	4045215	4066816	NM_015898	1.03E-02	-	ZBTB7A
chr17	40554466	40575338	NM_012232	1.09E-02	-	PTRF
chr1	27189632	27190947	NM_006142	1.13E-02	+	SFN
chr11	65686727	65689048	NM_006442	1.14E-02	+	DRAP1
chr1	38273472	38275126	NM_001142726	1.17E-02	+	C1orf122
chr4	7032280	7047958	NR_033828	1.58E-02	-	LOC100129931
chr12	58118075	58135944	NM_014770	1.70E-02	-	AGAP2
chr1	156084460	156107657	NM_005572	1.93E-02	+	LMNA

Note. These results can be interpreted as the *p* value of the likelihood of being NRF1 regulated based on integrative analysis of NRF1 binding peaks next to TSS (+/-5.0 kb) and differential expression.

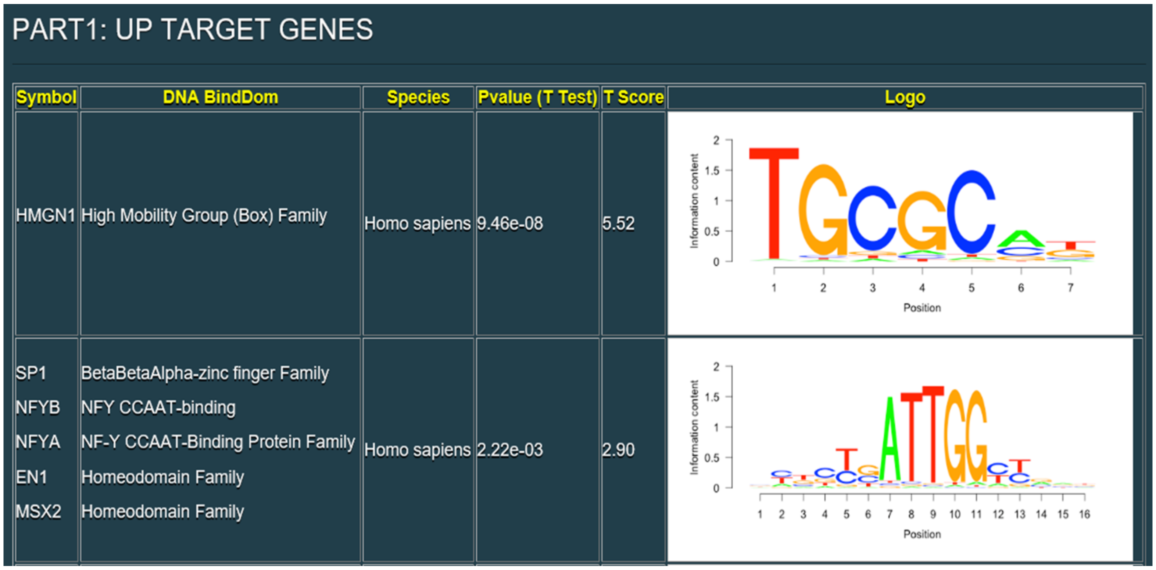
The list of upregulated and downregulated NRF1 targets includes several genes previously reported as connected to breast cancer, such as CCNE2, HPGD, FGFR1, ITGA6, LAMA3, and PDCD4. Cyclin E2 (CCNE2) overexpression has been linked with endocrine resistance in breast cancer, found overexpressed in

Her2 enriched and luminal B breast cancers and also associated with shorter distant metastasis-free survival among breast cancer patients after endocrine therapy (Caldon et al., 2012). HPGD (15-hydroxyprostaglandin dehydrogenase) was reported to promote epithelial mesenchymal transition (EMT) in aggressive breast tumors, and its upregulation was associated with poor prognosis in a subset of breast cancer patients (Lehtinen et al., 2012).

FGFR1 (fibroblast growth factor receptor 1) belongs to the FGFR gene family, a group of tyrosine kinase receptors that play an important role in the development and differentiation of the human mammary gland (Pond et al., 2013). In vitro essays of FGFR1 have shown that its activation resulted in cellular transformation of nontransformed MC10A human mammary cells, cell proliferation, survival, loss of cell polarity, and EMT (Xian et al., 2009). FGFR1 is also part of the KEGG (Kyoto Encyclopedia of Genes and Genomes) breast cancer pathway. ITGA6 (integrin subunit alpha 6) is part of Pathways in cancer (KEGG) and found downregulated in breast cancer tissue samples of HER2+ patients (Zubor et al., 2015). LAMA3 downregulation due to epigenetic changes in breast cancer has been found to be associated with increased tumor stage and tumor size (Sathyanarayana et al., 2003). PDCD4 is a tumor suppressor gene whose downregulation promotes antiapoptosis and chemotherapy resistance in the breast tumor cell line MCF-7 (Bourguignon, Spevak, Wong, Zia, & Gilad, 2009).

Binding motif analysis

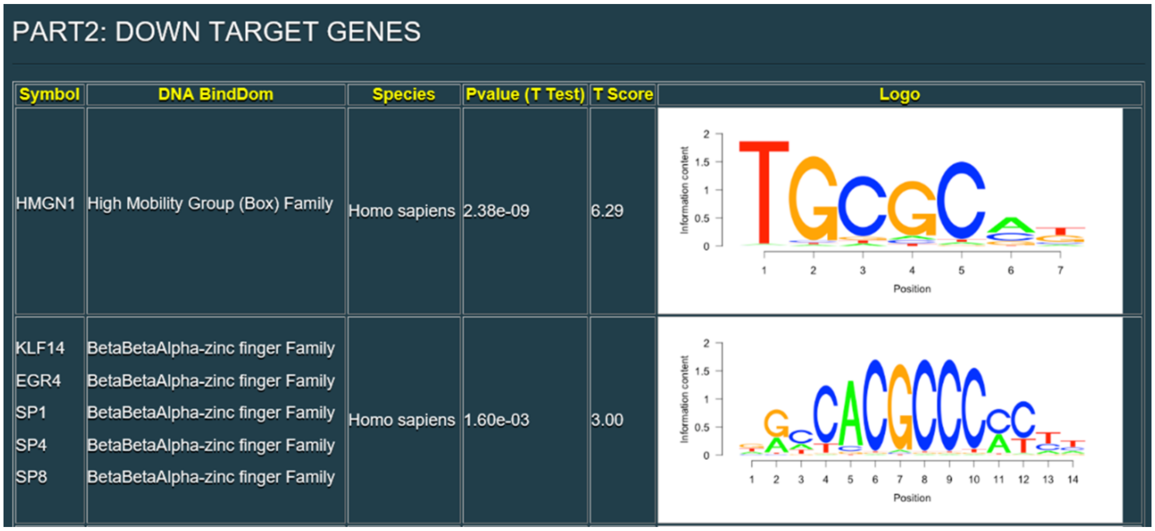
Figures 11 and 12 are screenshots from BETA outputs showing the results of binding motif analysis in NRF1 upregulated and downregulated target genes, respectively. T scores and *p* values are the statistics of the enrichment. The most significant binding motif in both groups upregulated and downregulated targets was the same—TGCGCAT (Figure 13)—confirming the NRF1 motif in breast cancer cells reported by Zhang et al. (2015).



PART1: UP TARGET GENES

Symbol	DNA BindDom	Species	Pvalue (T Test)	T Score	Logo
HMGN1	High Mobility Group (Box) Family	Homo sapiens	9.46e-08	5.52	
SP1 NFYB NFYA EN1 MSX2	BetaBetaAlpha-zinc finger Family NFY CCAAT-binding NF-Y CCAAT-Binding Protein Family Homeodomain Family Homeodomain Family	Homo sapiens	2.22e-03	2.90	

Figure 11. Screenshot of BETA output showing the results of binding motif analysis in the group of NRF1 upregulated targets.



PART2: DOWN TARGET GENES

Symbol	DNA BindDom	Species	Pvalue (T Test)	T Score	Logo
HMGN1	High Mobility Group (Box) Family	Homo sapiens	2.38e-09	6.29	
KLF14 EGR4 SP1 SP4 SP8	BetaBetaAlpha-zinc finger Family BetaBetaAlpha-zinc finger Family BetaBetaAlpha-zinc finger Family BetaBetaAlpha-zinc finger Family BetaBetaAlpha-zinc finger Family	Homo sapiens	1.60e-03	3.00	

Figure 12. BETA output showing the results of binding motif analysis in the group of NRF1 downregulated targets.

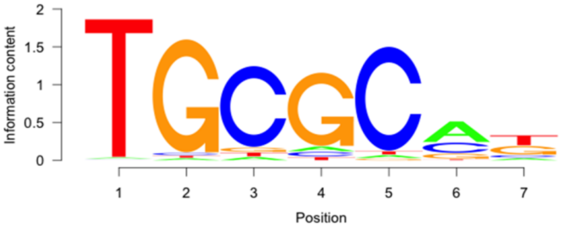


Figure 13. BETA output showing the most significant binding motif in both groups upregulated and downregulated targets which resulted in being the same.

Conclusions

Integrated data analysis of NRF1 ChIP-Seq and RNA-Seq data in HER2 positive breast cancer (HCC1954) and normal human mammary epithelial cells (HMEC) revealed a set of 63 upregulated and 73 downregulated genes that are very likely NRF1 regulated targets with binding sites within +/- 5.0 kb from TSS. Twenty-five (25) genes were upregulated more than 4 log2 fold change (SAMD12, MIPOL1,HOXA10, KIAA2018, TPPP, TAB3, RASD2, PACSIN1, SIX5, FAM9C, SLC7A10, APBA3, TMEM63A, GSTM3, PLXDC1, PJA2, ZKSCAN3, CBFA2T2 ,PPAP2B, HPGD, PLEKHG4, C7orf49, PNPO, TTC28-AS1, and STAU2) and twenty-one (21) were downregulated more than 4 log2 fold change (PTHLH, MT2A, EMP3, KIRREL, LAMA3, CCDC85B, FSCN1, POLR2B , PHLDA1 , S100A2 , SDCCAG8 , EIF2S1 , HOXA-AS3 , MAP3K9 , PI4K2A, CALD1, FGFR1, VIM, ARFRP1, TIPARP-AS1, and YIF1B).

These genes can be considered candidate drivers of HER2+ breast cancer. Binding motif analysis confirmed the presence of the preferred NRF1 motif TGCGCAT in the summit of NRF1 peaks. Our results were aligned with other studies that had found some of these genes such as CCNE2, HPGD, FGFR1, ITGA6, LAMA3 and PDCD4 associated with the development and progression of breast tumors.

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., . . . & Guerler, A. (2018). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537-W544.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A python framework to work with high-throughput sequencing data. *Bioinformatics, 31*(2), 166-169.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/ projects/fastqc/

Auerbach, R. K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrancois, P., . . . Snyder, M. (2009). Mapping accessible chromatin regions using sono-seq. *Proceedings of the National Academy of Sciences of the United States of America, 106*(35), 14926-14931.

Bailey, T. L., & Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research, 40*(17), 1-10.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for llumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Bourguignon, L. Y., Spevak, C. C., Wong, G., Xia, W., & Gilad, E. (2009). Hyaluronan-CD44 interaction with protein kinase C(epsilon) promotes oncogenic signaling by the stem cell marker nanog and the production of microRNA-21, leading to down-regulation of the tumor suppressor protein PDCD4, anti-apoptosis, and chemotherapy resistance in breast tumor cells. *Journal of Biological Chemistry, 284*(39), 26533-26546.

Caldon, C. E., Sergio, C. M., Kang, J., Muthukaruppan, A., Boersma, M. N., Stone, A., . . . Gee, J, M, (2012). Cyclin E2 overexpression is associated with endocrine resistance but not insensitivity to CDK2 inhibition in human breast cancer cells. *Molecular Cancer Therapeutics, 11*(7), 1488-1499.

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., . . . Davis, C. A. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Research, 22*(9), 1658-1667.

Cheung, M. S., Down, T. A., Latorre, I., & Ahringer, J. (2011). Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Research, 39*(15), 1-9.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., . . . Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*(13), 1-9.

Dai, Q. (2007). Chromatin regulators and transcriptional control of drosophila (Unpublished Doctoral dissertation). Wenner-Grens institut för experimentell biologi, Stockholm University, Stockholm, Sweden.

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read datasets from high-throughput DNA sequencing. *Nucleic Acids Research, 36*(16), 1-10.

Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L., & Schubeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature, 528*(7583), 575-579.

Ertel, A., Tsirigos, A., Whitaker-Menezes, D., Birbe, R. C., Pavlides, S., Martinez-Outschoorn, U. E., . . . Lisanti, M P. (2012). Is cancer a metabolic rebellion against host aging? in the quest for immortality, tumor cells try to save themselves by boosting mitochondrial metabolism. *Cell Cycle, 11*(2), 253-263.

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Research, 8*(3), 186-194.

Falco, M. M., Bleda. M., Carbonell-Caballero, J., & Dopazo, J. (2016). The pan-cancer pathological regulatory landscape. *Scientific Reports, 6*(39709), 1-13.

Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nature Protocols, 7*(9), 1728-1740.

Finotello, F., & Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. *Briefings in Functional Genomics, 14*(2), 130-142.

Furey, T. S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics, 13*(12), 840-852.

Hon, G. C., Hawkins, R. D., Caballero, O. L., Lo, C., Lister, R., Pelizzola, M., . . . (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Research, 22(*2), 246-258.

Illumina. (2011-2014). Retrieved from http://support.illumina.com/ content/ dam/ illumina-support/help/BaseSpaceHelp_v2/Content/Vault/ Informatics/ Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14*(4), 1-13.

Klein, H. U., Schafer, M., Porse, B. T., Hasemann, M. S., Ickstadt, K., & Dugas, M. (2014). Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. *Bioinformatics, 30*(8), 1154-1162.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., . . . Chen, Y. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research, 22*(9), 1813-1831.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods, 9*(4), 357-359.

Ledergerber, C., & Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics, 12*(5), 489-497.

Lehtinen, L., Vainio, P., Wikman, H., Reemts, J., Hilvo, M., Issa, R., . . . Kallioniemi, O. (2012). 15-hydroxyprostaglandin dehydrogenase associates with poor prognosis in breast cancer, induces epithelial-mesenchymal transition, and promotes cell migration in cultured breast cancer cells. *Journal of Pathology, 226*(4), 674-686.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(550), 1-21.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., . . . Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology, 28*(5), 495-501.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods, 5*(7), 621-628.

Okoh, V. O., Garba, N. A., Penney, R. B., Das, J., Deoraj, A., Singh, K. P., . . . Roy, D. (2015). Redox signalling to nuclear regulatory proteins by reactive oxygen species contributes to oestrogen-induced growth of breast cancer cells. *British Journal of Cancer, 112*(10), 1687-1702.

Ouyang, Z., Zhou, Q., & Wong, W. H. (2009). ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America, 106*(51), 21521-21526.

Pachter, L. (2011). Models for transcript quantification from RNA-seq. Cornell University Library, arXiv:1104.3889, 1-28.

Park, P. J. (2009). ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics, 10*(10), 669-680.

Pond, A. C., Bin, X., Batts, T., Roarty, K., Hilsenbeck, S., & Rosen, J. M. (2013). Fibroblast growth factor receptor signaling is essential for normal mammary gland development and stem cell function. *Stem Cells, 31*(1), 178-189.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., . . . Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology, 27*(1), 66-75.

Sathyanarayana, U. G., Padar, A., Huang, C. X., Suzuki, M., Shigematsu, H., Bekele, B. N., & Gazdar, A. F. (2003). Aberrant promoter methylation and silencing of laminin-5-encoding genes in breast carcinoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 9*(17), 6389-6394.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics, 15*(2), 121-132.

Tang, Q., Chen, Y., Meyer, C., Geistlinger, T., Lupien, M., Wang, Q., . . . Liu, X. S. (2011). A comprehensive view of nuclear receptor cancer cistromes. *Cancer Research, 71*(22), 6940-6947.

Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., . . . Liu, X. S. (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nature Protocols, 8*(12), 2502-2515.

Xian, W., Pappas, L., Pandya, D., Selfors, L. M., Derksen, P. W., de Bruin, M., . . . Brugge, J. S. (2009). Fibroblast growth factor receptor 1-transformed mammary epithelial cells are dependent on RSK activity for growth and survival. *Cancer Research, 69*(6), 2244-2251.

Zhang, J., Wang, C., Chen, X., Takada, M., Fan, C., Zheng, X., . . . Aird, K. M. (2015). EglN2 associates with the NRF1-PGC1alpha complex and controls mitochondrial function in breast cancer. *EMBO Journal, 34*(23), 2953-2970.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based analysis of ChIP-seq (MACS). *Genome Biology, 9*(9), R137.1-R137.9.

Zubor, P., Hatok, J., Moricova, P., Kapustova, I., Kajo, K., Mendelova, A., . . . Danko, J. (2015). Gene expression profiling of histologically normal breast tissue in females with human epidermal growth factor receptor 2positive breast cancer. *Molecular Medicine Reports, 11*(2), 1421-1427.

CHAPTER IV

NRF1 MOTIF SEQUENCE-ENRICHED GENES INVOLVED IN ER-PR-HER2+ BREAST CANCER SIGNALING PATHWAYS

Note. This chapter was previously published: Ramos, J., Das, J., Felty, Q., Yoo, C., Poppiti, R., Murrell, D., . . . Roy, D. (2018). NRF1 motif sequence-enriched genes involved in ER/PR -ve HER2 +ve breast cancer signaling pathways. *Breast Cancer Research and Treatment*, *8,* 1-17.

Abstract

Nuclear respiratory factor 1 (NRF1) transcription factor has recently been shown to control breast cancer progression. However, mechanistic aspects by which NRF1 may contribute to susceptibility to different breast tumor subtypes are still not fully understood. Since transcriptional control of NRF1 seems to be dependent on epidermal growth factor receptor signaling, herein we investigated the role of NRF1 in estrogen receptor/progesterone receptor negative, but human, epidermal growth factor receptor 2 positive (ER-PR-HER2+) breast cancer. We found that both mRNA and protein levels of NRF1, and its transcriptional activity, were significantly higher in ER-PR-HER2+ breast cancer samples compared to normal breast tissues. This result was consistent with our observation of higher NRF1 protein expression in the experimental model of HER2+ breast cancer brain metastasis. To identify network-based pathways involved in the susceptibility to the ER-PR-HER2+ breast cancer subtype, the NRF1 transcriptional regulatory genome-wide landscape was analyzed using the approach consisting of a systematic integration of ChIP DNA-seq, RNA-Microarray, NRF1 protein-DNA

motif binding, signal pathway analysis, and Bayesian machine learning. Our findings showed that a high percentage of known HER2+ breast cancer susceptibility genes, including EGFR, IGFR, and E2F1, are under transcriptional control of NRF1. Promoters of several genes from the KEGG (Kyoto Encyclopedia of Genes and Genomes) HER2+ breast cancer pathway and 11 signaling pathways linked to 6 hallmarks of cancer contain the NRF1 motif. By pathway analysis, key breast cancer hallmark genes of epithelial-mesenchymal transition, stemness, cell apoptosis, cell cycle regulation, chromosomal integrity, and DNA damage/repair were highly enriched with NRF1 motifs. In addition, we found using Bayesian network-based machine learning that 30 NRF1 motif-enriched genes—including growth factor receptors FGFR1, IGF1R; E2Fs transcription factor family-E2F1, E2F3; MAPK pathway-SHC2, GRB2, MAPK1; PI3K-AKT-mTOR signaling pathway-PIK3CD, PIK3R1, PIK3R3, RPS6KB2; WNT signaling pathway-WNT7B, DLV1, DLV2, GSK3B, NRF1, and DDB2, known for its role in DNA repair and involvement in early events associated with metastatic progression of breast cancer cells—were associated with HER2 amplified breast cancer. Machine learning search further revealed that the likelihood of HER2 positive breast cancer was almost 100% in a patient with high NRF1 expression combined with expression patterns of high E2F3, GSK3B, and MAPK1; low or no change in E2F1 and FGFR1; and high or no change in PIK3R3. In summary, our findings suggest novel roles of NRF1 and its regulatory networks in susceptibility to the ER-PR-HER2+ aggressive breast cancer subtype. Clinical confirmation of our machine-learned Bayesian networks will have significant impact on our understanding of the role of NRF1 as a valuable biomarker for breast cancer diagnosis and prognosis as well as provide strong rationale for future studies to develop NRF1 signaling-based therapeutics to target HER2+ breast cancer.

Introduction

Nuclear respiratory factor 1 (NRF1) [alpha-palindromic binding protein (α-PAL)], is a redox-sensitive transcription factor (Okoh, Deoraj, & Roy, 2011; Piantadosi & Suliman, 2006; Roy & Tamuli, 2009; Scarpulla, 2006, 2008). The role of NRF1 in breast cancer (BC) has remained largely unexplored. We have recently shown that reactive oxygen species (ROS) contribute to estrogen-induced growth of BC cells through a NRF1 signaling pathway (Okoh et al., 2015). Several cell cycle genes, including CDC2, PRC1, PCNA, cyclin B1, and CDC25C, are regulated by NRF1 and implicated in estrogen-induced breast carcinogenesis (Okoh et al., 2011). A bioinformatics study showed that NRF1 is one of the principal regulatory motifs significantly associated with worsening histological grades and poor breast cancer prognosis (Niida et al., 2008).

NRF1 activity is higher in breast cancer tissue compared to adjacent normal tissue (Ertel et al., 2012). NRF1 activity correlates significantly with histological grades and prognosis of BC (Falco, Bleda, Marbonell-Caballero, & Dopazo, 2016). A recent report showed that NRF1 expression is significantly higher in breast cancer tissue of Chinese patients compared with adjacent normal tissues (Gao et al., 2018). Despite these studies, the impact of NRF1- regulated gene networks on aggressive growth and metastasis of BC is still unknown.

NRF1 is one of the transcription factors with the highest enrichment scores in mutated epidermal growth factor receptor (EGFR, L858R; T790M mutations)-derived lung tumors; treatment of mice with an irreversible EGFR/HER2 tyrosine kinase inhibitor- afatinib drug significantly down-regulates the expression of this gene in tumors (Weaver et al., 2012). These data suggest that transcriptional control of NRF1 depends on EGFR signaling. Therefore, in this study we examined the role of NRF1 in human epidermal growth factor receptor 2 (HER2) positive breast tumors, one of the two most aggressive breast cancer subtypes with poor prognosis (Lee, Oprea-Ilies, & Saavedra, 2015; Sorlie et al., 2001). Here, we report higher NRF1 expression in HER2 positive breast tumors. To further understand the role of NRF1 in HER2+ breast cancer, we also deciphered the regulatory landscape of NRF1 networks in a HER2+ breast cancer line and HER2+ breast cancer samples. Our findings revealed novel roles of NRF1 and its regulatory network associated with ER− PR− HER2+ breast cancer.

Results

Higher NRF1 Expression in HER2+ Breast Cancer

As a first step in discovering the association between NRF1 transcription factor and ER-PR-HER2+ breast cancer, we used TCGA NRF1 microarray data (log2 normalized (cy5/cy3) from 61 normal samples and 22 HER2+ breast tumor samples. We performed the SAS PROCTTEST, which estimates for the equality of means for a two-sample (independent group) *t* test, to compare the mean NRF1 expression values of these two groups. The results are summarized using boxplots and histograms in Figure 1. The comparison of the relative NRF1 distribution in normal and HER2+ breast tumors showed that a higher proportion of breast cancer specimens possessed elevated levels of NRF1 compared to normal breast tissue specimens. Statistical analysis of mRNA expression showed that NRFI was significantly overexpressed in ER-PR-HER2+ breast cancer tissue compared to normal tissues (Figure 1, $p < 0.0014$).

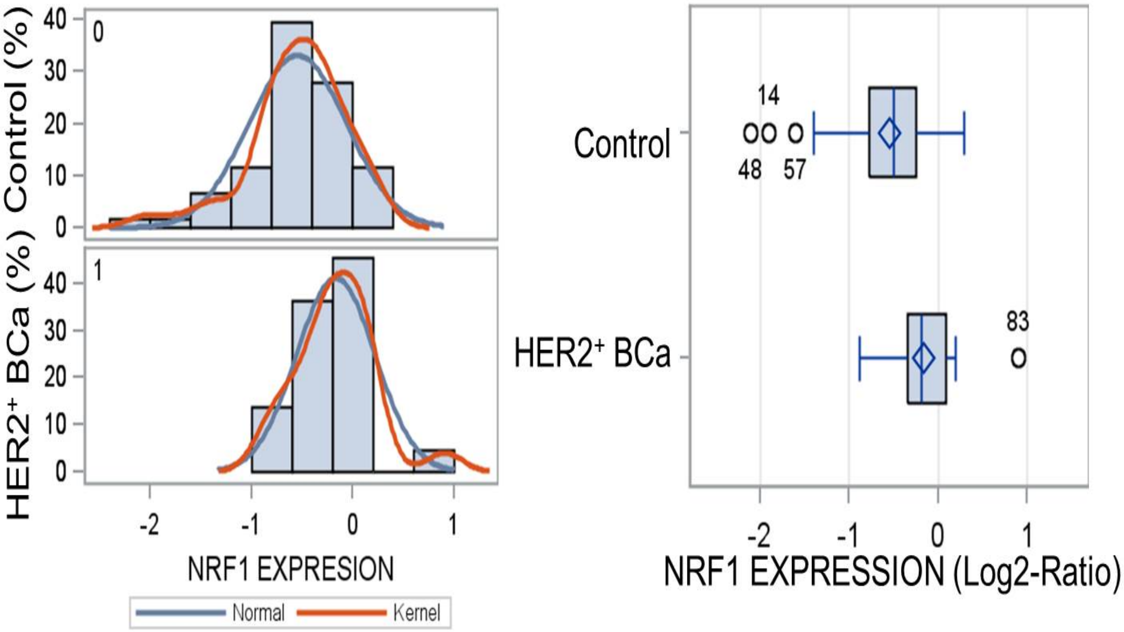


Figure 1. Shows a histogram and boxplot of NRF1 expression in ER-PR-HER2+ breast cancer TCGA samples compared to normal samples.

To corroborate this finding, we investigated NRF1 protein expression in a breast cancer tissue microarray (TMA). TMA stained with antibodies specific for NRF1 were analyzed by confocal immunofluorescence microscopy. The representative confocal TMA immunofluorescence analysis showed increased expression levels of NRF1 in ER-PR-HER2+ breast cancer specimens compared to normal breast specimens. Tumor cells overexpressing HER2 showed moderate to high nuclear staining of NRF1. A majority of normal breast cells showed weak to moderate NRF1 nuclear immunoreactivity (Figure 2).

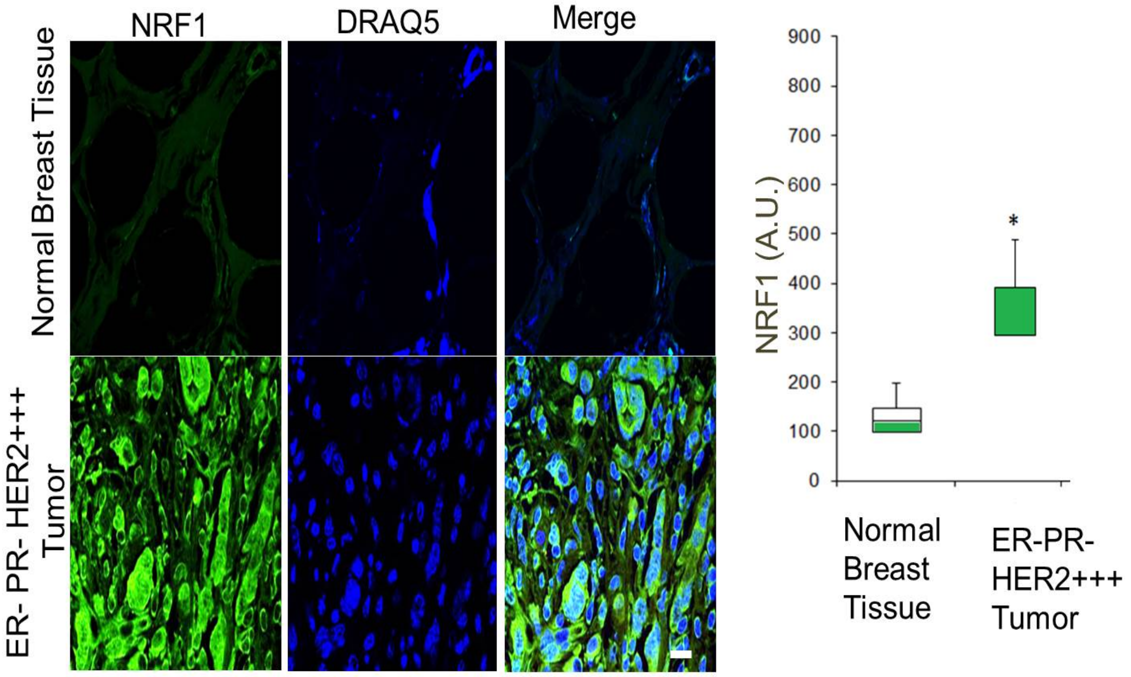


Figure 2. NRF1 protein expression was higher in ER- PR- HER2+ breast cancer tissue sections compared to normal breast tissue section. Shown in (A) Representative immunoreactivity of NRF1 antibodies and (B) the box plot distribution of intensity scores for NRF1 immunoreactivity (arbitrary unit =A.U.). *$p < 0.05$.

We also measured the transcription activity of NRF1 by estimating modulation of mRNA levels of its target genes in a coordinated way in normal and ER-PR-HER2+ breast cancer TCGA samples. NRF1 transcription activity was significantly upregulated in ER-PR-HER2+ breast cancer compared to normal breast tissues (Figure 3). The increase in NRF1 activity was consistent with our observations of higher NRF1 mRNA and protein levels in in ER-PR-HER2+ breast cancer.

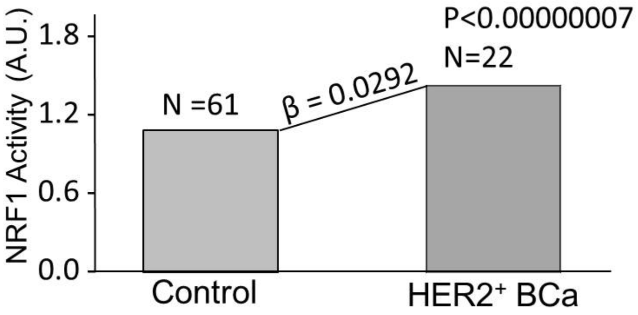


Figure 3. Shows an increased NRF1 activity in ER-PR-HER2+ breast cancer TCGA samples compared to normal samples.

We also evaluated NRF1 protein expression in the experimental model of HER2+ breast cancer metastasis. Statistical analysis of NRF1 immuno-reactivity showed that NRF1 was significantly higher in HER2+ breast cancer cells derived from brain tumors compared to MDA-MB-231-BR (231-BR)-vector cell–derived brain metastases (Figure 4, $p < 0.01$). Our finding of NRF1 overexpression in HER2+ breast cancer brain metastatic tumors was consistent with the observation of a previous report of EGFR-derived lung tumors (Weaver et al., 2012). In summary, these data suggest NRF1 expression is significantly associated with HER2+ breast cancer in the preclinical model and clinical breast cancer human samples.

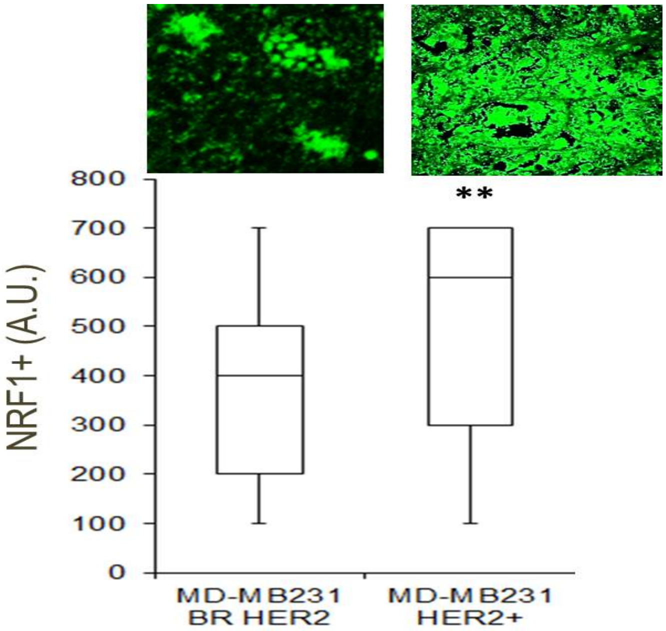


Figure 4. The representative confocal immunofluorescence microscopy image of NRF1 protein expression and box plot showing relative quantitative value o NRF1 intensity in brain tumors overexpressing HER2. Mice were injected with MDA-MB-231-BR (231-BR)-vector (*n* = 4 mice) or HER2 overexpressing MDA-MB-231-BR brain tumor sections (*n* = 6).

Discovery of NRF1 Bound DNA Regions in ER-PR-HER2+ Cells

To understand NRF1's role in ER-PR-HER2+ breast cancer, it is critical that we identify NRF1 transcriptional regulation of target genes. As a first step, we identified the actual occupancy of NRF1 protein to the DNA motif site(s) of the different regions of the genome and the distance of NRF1 protein binding sites from the transcription start site (TSS) in breast cancer cells that are ER-PR-HER2+ (HCC1954); and normal breast epithelial cells (HMEC). To accomplish this comparison, we used archived NRF1 ChIP DNA-seq data of HCC1954 breast cancer cells from Gene Expression Omnibus (GEO) (Domcke et al., 2015), aligned them into the human genome using BOWTIE2, and subsequently used MACS2 to identify enriched NRF1 peaks with fold enrichment (FE) greater or equal to 5. After peak identification, we determined genes associated with NRF1 binding sites using GREAT 3.0.0. GEO accession numbers for NRF1. ChIP-Seq data and details of software and setting parameters used for alignment, peak calling, and gene identification can be found in the Method section.

We identified NRF1 bound target genes that had binding activity localized in promoter proximal regions (+/- 2,000 bp from the TSS) in both normal human mammary epithelial cells (HMECs) isolated from adult female breast tissue and the breast cancer cell line HCC1954 that represents a breast ductal carcinoma (ER-PR-HER2+) with amplified HER2 and high abundance of EGFR. This cell line is a well-accepted model of metastatic HER2+ breast cancer (Henjes et al., 2012). We found 1,283 genes that were NRF1 targets exclusively in HMEC cells, 1,225 exclusively in HCC1954, and 10,911 NRF1 targets common in both cell lines (Figure 5). We could not compare our observations with a previous report from HCC1954 breast cancer cells to investigate the effect of DNA methylation to NRF1 binding because the number of target genes was not reported.

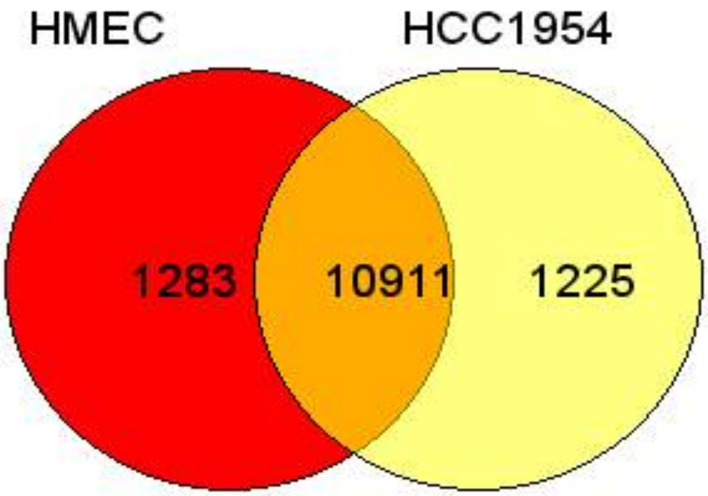


Figure 5. Venn diagram showing the number of common and unique NRF1 target genes in HER2+ breast cancer cells (HCC1954) compared to normal breast epithelial cells (HMEC).

An increase in NRF1 activity has been previously reported in breast cancer compared to normal tissue; therefore, we also compared the NRF1 network in HMEC normal cells with HCC1954 HER2+ breast cancer cells. We used the Jaccard coefficient (JC) to measure the intersection between the two sets of genes: JC = (A∩B) / (A∪B). A very high level of similarity was found in the HCC1954 cells (JC=81.3 %).

NRF1 Motif-Enriched Target Genes Are Part of the Breast Cancer Hallmark Pathway

To identify the pathways in HER2 amplified breast cancer that may be regulated by NRF1, we used DAVID (Database for Annotation, Visualization and Integrated Discovery) and KEGG to map NRF1 identified target genes in HCC1954 with hallmark genes of cancer and breast cancer signaling pathways (Hanahan &

Weinberg. 2011). Figure 6 and Table 1 show NRF1 target genes identified in each signaling pathway as well as the associated hallmarks of cancer.

We found 11 critical signaling pathways enriched with NRF1 target genes—PI3K-Akt signaling, MAP-kinase pathway, mTOR pathway, cellular senescence, p53 signaling, apoptosis, TGF-beta signaling, autophagy, VEGF signaling, T cell receptor signaling, and B cell receptor signaling. These signaling pathways when altered are involved in the following hallmarks of cancer: sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and evading immune destruction. NRF1 target enriched signaling pathways ranked by number of genes, starting with mTOR (48 genes) and ending with TGF-beta signaling (5 genes) are shown in Figure 6.

Sustaining proliferative signaling which allows cancer cells to maintain continuous growth shows the maximum number of associated signaling pathways (4: PI3K-Akt signaling, MAP-kinase pathway, mTOR pathway, and cellular senescence) followed by evading growth suppressors (3: p53 signaling, apoptosis, and TGF-beta signaling) and resisting cell death (3: p53 signaling, apoptosis, and autophagy). PI3K-Akt signaling contained 46 NRF1 target genes, including five genes—FGF13, FGF19, FGF3, FGF4, and FLT4 that were present only in HER2+ breast cancer cells. MAP-kinase Pathway contained 34 NRF1 target genes. mTOR Pathway contained 48 NRF1 target genes, including FZD10 and WNT1 only present in HER2+ breast cancer cells.

NRF1 motif was present in 28 genes of cellular senescence as part of the sustaining proliferative signaling hallmark. NRF1 motifs were present in 10 genes of the p53 signaling pathway including BAX only found in HER2+ cells. NRF1 motifs were found in 23 genes in apoptosis, 19 genes in autophagy, and 5 genes in TGF-beta signaling. Figure 6 and Table 1 provide detailed information of target genes classified by signaling pathway.

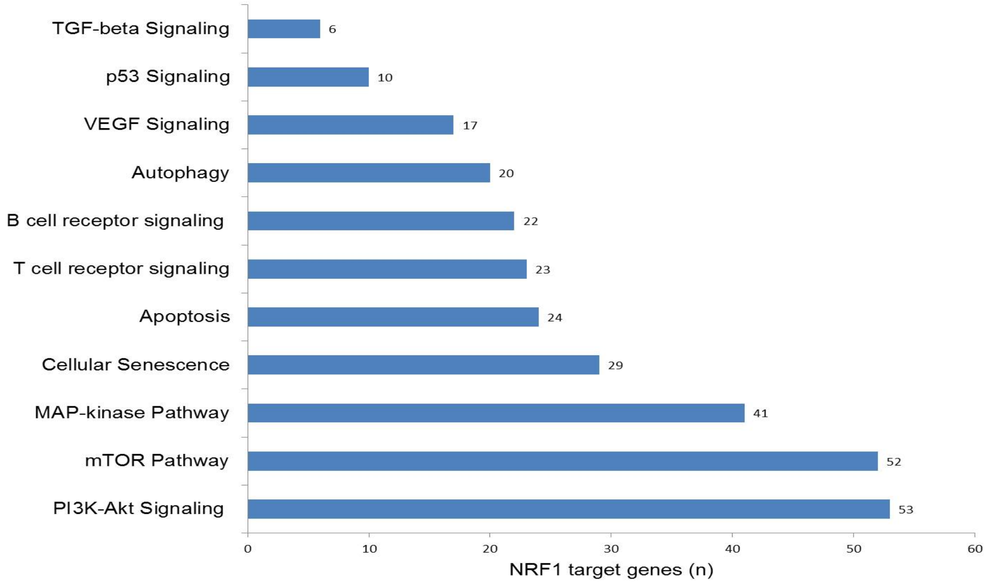


Figure 6. Number of genes containing NRF1 motif discovered in cancer hallmark signaling pathways.

Table 1

Signaling Pathway Enriched With NRF1 Target Genes in the BC and Hallmark of Cancer Pathway

	Hallmarks of cancer						
KEGG Signaling pathway / NRF1 target genes	Sustaining proliferative signaling	Evading growth suppressor	Resisting cell death	Enabling replicative immortality	Angiogenesis	Evading immune destruction	*p*
PI3K-Akt Signaling: AKT1 , AKT2 , BRCA1 , CCND1 , CDK4 , CDK6 , CDKN1A , EGFR , FGF1 , FGF10 , FGF11 , FGF12 , FGF18 , FGF21 , FGF22 , FGF7 , FGF9 , FGFR1 , GRB2 , GSK3B , HRAS , IGF1R , KIT , KRAS , MAP2K1 , MAP2K2 , MAPK1 , MAPK3 , MYC , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , PTEN , RAF1 , RPS6KB1 , RPS6KB2 , SOS1 , SOS2 , FGF13 , FGF19** , FGF3** , FGF4** , FLT4****	X						1.0E-5
MAP-kinase Pathway: AKT1 , AKT2 , EGFR , FGF1 , FGF10 , FGF11 , FGF12 , FGF18 , FGF21 , FGF22 , FGF7 , FGF9 , FGFR1 , FOS , GADD45A , GADD45B , GADD45G , GRB2 , HRAS , JUN , KRAS , MAP2K1 , MAP2K2 , MAPK1 , MAPK3 , MYC , NFKB2 , RAF1 , SOS1 , SOS2 , FGF13 , FGF19** , FGF3** , FGF4****	X						4.6E-240
mTOR Pathway: AKT1 , AKT2 , DVL1 , DVL2 , DVL3 , FZD1 , FZD2 , FZD3 , FZD4 , FZD5 , FZD6 , FZD8 , FZD9 , GRB2 , GSK3B , HRAS , IGF1R , KRAS , LRP5 , LRP6 , MAP2K1 , MAP2K2 , MAPK1 , MAPK3 , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , PTEN , RAF1 , RPS6KB1 , RPS6KB2 , SOS1 , SOS2 , WNT10A , WNT10B , WNT11 ,	X						1.0E-44

WNT3A , WNT4 , WNT5A , WNT7A , WNT7B , WNT8B , WNT9A , FZD10** , WNT1**							
Cellular Senescence: AKT1 , AKT2 , CCND1 , CDK4 , CDK6 , CDKN1A , E2F1 , E2F2 , E2F3 , GADD45A , GADD45B , GADD45G , HRAS , KRAS , MAP2K1 , MAP2K2 , MAPK1 , MAPK3 , MYC , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , PTEN , RAF1 , RB1	X						1.0 E-5
p53 Signaling: CCND1 , CDK4 , CDK6 , CDKN1A , DDB2 , GADD45A , GADD45B , GADD45G , PTEN , BAX**		X	X	X			3.5 E-58
Apoptosis: AKT1 , AKT2 , BAK1 , FOS , GADD45A , GADD45B , GADD45G , HRAS , JUN , KRAS , MAP2K1 , MAP2K2 , MAP3K5 , MAPK1 , MAPK3 , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , RAF1 , BAX**		X	X				1.2 E-71
TGF-beta Signaling: MAPK1 , MAPK3 , MYC , RPS6KB1 , RPS6KB2		X					9.5 E-74
Autophagy: AKT1 , AKT2 , HRAS , IGF1R , KRAS , MAP2K1 , MAP2K2 , MAPK1 , MAPK3 , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , PTEN , RAF1 , RPS6KB1 , RPS6KB2			X				1.0 E-5
VEGF Signaling: AKT1 , AKT2 , HRAS , KRAS , MAP2K1 , MAP2K2 , MAPK1 , MAPK3 , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , RAF1 , SHC2					X		2.5 E-62
T cell receptor signaling: AKT1 , AKT2 , CDK4 , FOS , GRB2 , GSK3B , HRAS , JUN , KRAS , MAP2K1 , MAP2K2 , MAPK1 , MAPK3 , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , RAF1 , SOS1 , SOS2						X	1.7 E-92
B cell receptor signaling: AKT1 , AKT2 , FOS , GRB2 , GSK3B , HRAS , JUN , KRAS , MAP2K1 , MAP2K2 , MAPK1 , MAPK3 , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , RAF1 , SOS1 , SOS2						X	2.7 E-61

*Note. ** Indicates only present In HCC1954 cells.*

We compared the 78 candidate identified genes in the signaling pathway analysis against the list of NRF1 target genes reported in ENCODE by Harmonizome—a collection of 125 unique processed datasets (Rouillard et al., 2016). This comparison resulted in 59 overlapping genes (Table 2).

Table 2

NRF1 Motif Present in Genes Involved in Development of Malignant Breast Tumors

AKT1*	AKT serine/threonine kinase 1	HRAS*	HRas proto-oncogene, GTPase
AKT2*	AKT serine/threonine kinase 2	IGF1R*	insulin like growth factor 1 receptor
BAK1	BCL2 antagonist/killer 1	JUN*	Jun proto-oncogene, AP-1 transcription factor subunit
BRCA1*	BRCA1 DNA repair associated	LRP6*	LDL receptor related protein 6
CDK4*	cyclin dependent kinase 4	MAP2K1*	mitogen-activated protein kinase kinase 1
CDK6	cyclin dependent kinase 6	MAP2K2*	mitogen-activated protein kinase kinase 2
DDB2*	damage specific DNA binding protein 2	MAP3K5*	mitogen-activated protein kinase kinase kinase 5
DVL1*	dishevelled segment polarity protein 1	MAPK1*	mitogen-activated protein kinase 1
DVL2*	dishevelled segment polarity protein 2	MYC	MYC proto-oncogene, bHLH transcription factor
E2F1*	E2F transcription factor 1	NFKB2*	nuclear factor kappa B subunit 2
E2F3*	E2F transcription factor 3	PIK3CA*	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
EGFR*	epidermal growth factor receptor	PIK3CD*	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta
FGF1	fibroblast growth factor 1	PIK3R1*	phosphoinositide-3-kinase regulatory subunit 1

FGF12	fibroblast growth factor 12	PIK3R2	phosphoinositide-3-kinase regulatory subunit 2
FGF18	fibroblast growth factor 18	PIK3R3*	phosphoinositide-3-kinase regulatory subunit 3
FGF22*	fibroblast growth factor 22	PTEN*	phosphatase and tensin homolog
FGF9*	fibroblast growth factor 9	RAF1*	Raf-1 proto-oncogene, serine/threonine kinase
FGFR1*	fibroblast growth factor receptor 1	RB1*	RB transcriptional corepressor 1
FZD1*	frizzled class receptor 1	RPS6KB1*	ribosomal protein S6 kinase B1
FZD2	frizzled class receptor 2	RPS6KB2*	ribosomal protein S6 kinase B2
FZD3*	frizzled class receptor 3	SHC2*	SHC adaptor protein 2
FZD4*	frizzled class receptor 4	SOS1*	SOS Ras/Rac guanine nucleotide exchange factor 1
FZD5*	frizzled class receptor 5	SOS2*	SOS Ras/Rho guanine nucleotide exchange factor 2
FZD8*	frizzled class receptor 8	WNT10A	Wnt family member 10A
FZD9*	frizzled class receptor 9	WNT10B*	Wnt family member 10B
GADD45A*	growth arrest and DNA damage inducible alpha	WNT5A*	Wnt family member 5A
GADD45B*	growth arrest and DNA damage inducible beta	WNT7A	Wnt family member 7A
GADD45G*	growth arrest and DNA damage inducible gamma	WNT7B*	Wnt family member 7B
GRB2*	growth factor receptor bound protein 2	WNT9A*	Wnt family member 9A
GSK3B*	glycogen synthase kinase 3 beta		

Note. Genes with an asterisk have NRF1 binding activity in proximal promoter regions (+/- 2,000 bp from TSS).

A protein–DNA interaction network of these genes was developed using CYTOSCAPE (Shannon et al., 2003) to visualize NRF1 regulation (Figure 7).

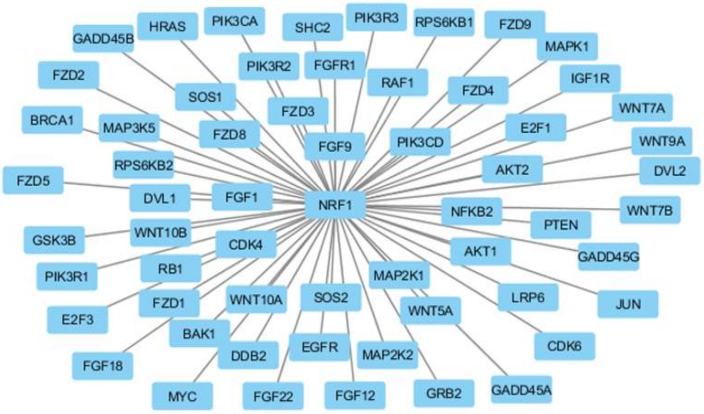


Figure 7. Protein-DNA interaction of 59 NRF1 target genes that contribute to acquiring hallmarks of cancer in breast neoplasms.

To investigate whether NRF1 target genes interact among themselves, we also constructed a protein-protein interaction network using STRING (Szklarczyk et al., 2015) with direct (physical) as well as indirect (functional) associations. The resulting network formed 59 nodes and 546 edges (interactions) (Figure 8).

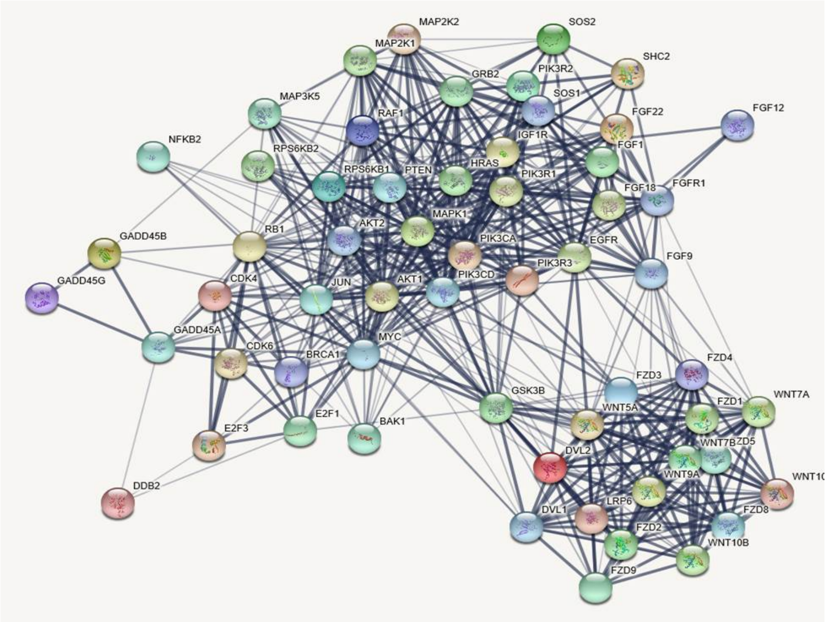


Figure 8. Protein-protein interaction of 59 NRF1 target genes that may contribute to cells acquiring hallmarks of cancer in breast neoplasms.

The expected number of interactions, for a random set of similar number of proteins drawn from the genome is 130; therefore, this high enrichment (546 vs. 130) indicates that these proteins are at least partially biologically connected, as a group (PPI enrichment *p*-value = 0). These 59 genes were also input into KEGG, and 21 of them enriched the section of HER2+ in the KEGG breast cancer pathway-hsa05224 (Figure 9, genes highlighted in yellow). In summary, pathway analysis showed that several NRF1 may regulate many genes that are part of the hallmarks of cancer in ER-PR-HER2 + type breast tumors.

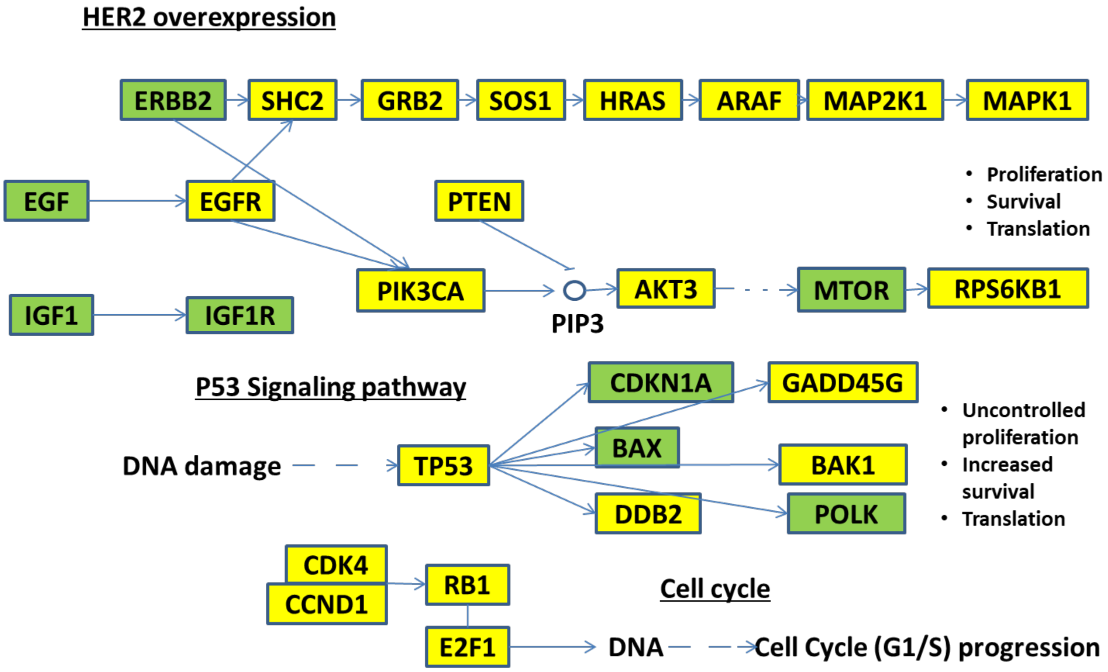


Figure 9. NRF1 target genes highlighted in yellow in KEGG HER2+ breast cancer pathway.

NRF1 Motif-Enriched Genes Correlated With Breast Cancer

To investigate whether the binding activity of NRF1 was correlated with breast cancer, we searched for NRF1 target genes in HCC1954 and HMEC cell lines with peaks (NRF1 binding regions from MACS2) located in promoter region (+/- 2,000 bp from TSS) and showing a minimum Fold Enrichment (FE) equal to 5. This was the same cutoff we had established for peak detection. This screening resulted in 7,663 genes meeting these criteria (out of more than 10,000 NRF1 targets).

We then performed a point biserial correlation analysis using breast cancer (BC) as the dependent categorical variable (coded BC = 0 for HMEC cells and BC = 1 for HCC1954 cells), and the average FE of the peaks as the independent variables, to identify the genes that had a different level of NRF1 binding when comparing breast cancer to normal cells (Table 3). We selected the top 2,000 correlated genes, 1,000 positive correlated and 1,000 negative correlated (absolute correlation ranging from 0.866 to 0.790) to visualize the difference in binding activity.

NRF1 binding activity in breast cancer cell line HCC1954 compared to normal mammary epithelial cells (HMEC) shows that NRF1 binding to 2,000 genes was differentially correlated when compared to normal breast epithelial cells. Furthermore, we observed the NRF1 motif (+/- 2,000 bp from TSS) in 49 genes involved in the development of malignant breast tumors (Table 2, genes with asterisk). Subsequently, we used these genes for NRF1 mediated transcriptome analysis in HER2+ breast tumors.

Table 3

Top 10 Genes With Changes in NRF1 Binding Correlated With HER2+ Breast Cancer

GENE		r_{pb}	HMEC-1*	HMEC-2*	HCC1954-1*	HCC1954-2*
BC (status)		1.00000	0.00000	0.00000	1.00000	1.00000
SERBP1		0.86599	9.95728	10.02096	16.74631	16.80865
PPP1R2		0.86597	12.45139	12.65966	28.54475	28.70575
ACSL3		0.86592	10.13207	10.02223	23.02297	23.28911
UQCC		0.86592	37.66980	37.28456	71.46165	72.12209
SF3B1		0.86583	36.03093	35.40324	64.47330	65.08048
LRFN3		0.86577	17.18126	16.47726	38.27592	38.06927
BMF		0.86574	9.38154	9.68616	17.90063	17.87512
CHERP		0.86572	25.38617	26.48630	61.58603	62.39032
SLC35G1		0.86570	40.94544	39.72945	105.82870	108.11758
CEBPA		0.86570	11.42519	11.18819	20.95348	20.66899

Note. *-1 = replicate 1 *-2 = replicate 2 in ChIP-Seq analysis. r_{pb} = point biserial correlation coefficient.

Machine Learning of NRF1 Target Genes Involved in HER2+ Breast Cancer

Bayesian network structure learning was used to discover gene-gene interactions and identify putative causal interactions with HER2+ breast cancer. We used TCGA microarray data to obtain the gene expression of the 49 NRF1 enriched genes in 61 normal breast tissue samples and 22 samples with the ER- PR- HER2+ breast cancer subtype. We calculated the mean expression of each gene and standard deviation. We then defined the cutoff values for up and down regulation to be the mean plus or minus two standard deviations.

The best network (BDe score = -1172.9652) was reached after examination of 20,699 million of networks during 8 hours in the second run of the software

Banjo. The methods section provides a more detailed explanation of how the scoring function BDe (Bayesian metrics with Dirichlet priors and equivalent) measures the probability of each searched structure G given the data D (P (G/D) to evaluate different structures (Le et al., 2013). Best network structure is shown in Figure 10. Table 4 shows the 30 genes that formed the Markov Blanket of the HER2+ breast cancer node. These genes are also shown color-coded in red in Figure 10.

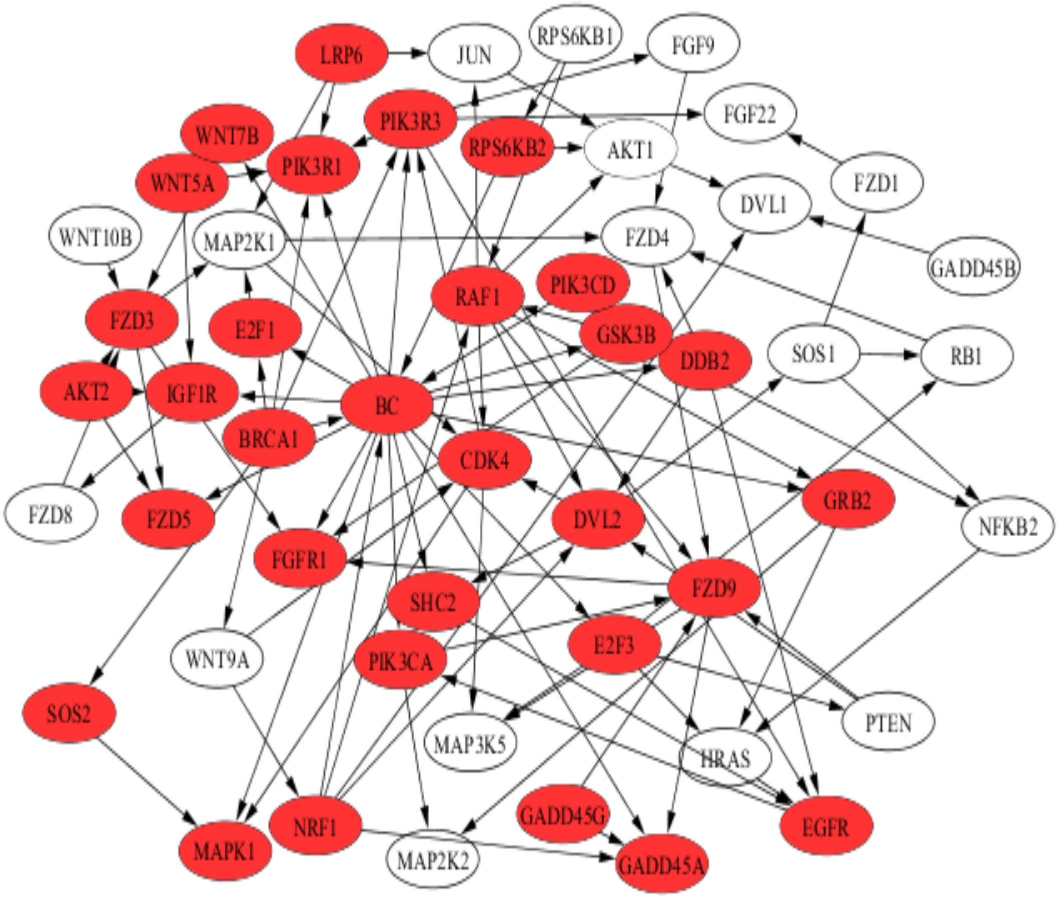


Figure 10. Bayesian network showing genes associated with ER- PR- HER2+ breast cancer.

The machine learned NRF1 motif-enriched genes included growth factor receptors—FGFR1, IGF1R; E2Fs transcription factor family—E2F1, E2F3; MAPK pathway-SHC2, GRB2, MAPK1; PI3K-AKT-mTOR signaling pathway—PIK3CD, PIK3R1, PIK3R3, RPS6KB2; WNT signaling pathway—WNT7B, DLV1, DLV2, GSK3B, NRF1, and DDB2, known for its role in DNA repair and involvement in early events associated with metastatic progression of breast cancer cells, were associated with HER2 amplified breast cancer. Consequently, we used these genes to estimate susceptibility to HER2+ breast cancer.

Table 4

Markov Blanket Genes of HER2+ BC in the Structure With the Best BDe Score

PARENTS	CHILDREN	OTHER CHILDREN'S PARENTS
BRCA1	DDB2	BRCA1*
PIK3CD	E2F1	FZD3
NRF1	E2F3	FZD9
RPS6KB2	FGFR1	GSK3B*
	FZD5	AKT2
	GADD45A	GADD45G
	GRB2	NRF1*
	GSK3B	RAF1
	IGF1R	WNT5A
	MAPK1	CDK4
	PIK3CA	SOS2
	PIK3R1	EGFR
	PIK3R3	LRP6
	SHC2	PIK3R3*
	WNT7B	DVL2

Note. Network created by the software BANJO after Bayesian network learning. Genes with asterisk are repeated.

Impact of NRF1 Target Genes on the Probability of Risk for ER-PR-HER2+ Breast Cancer

We used GeNIe, (software developed by the University of Pittsburg) to estimate the conditional and marginal probability distributions of HER2+ breast cancer as a result of modifications in 30 NRF1 target genes (Figure 11). A sensitivity analysis modifying NRF1 status to up-regulated in all subjects increased the marginal probability of HER2+ breast cancer from 30 % to 67 %. Similarly, when we modified the status of either PIK3R3 or WNT7B to up-regulated in all subjects, we observed an increased marginal probability of HER2+ breast cancer from 30% to 92% and 30% to 88.5%, respectively.

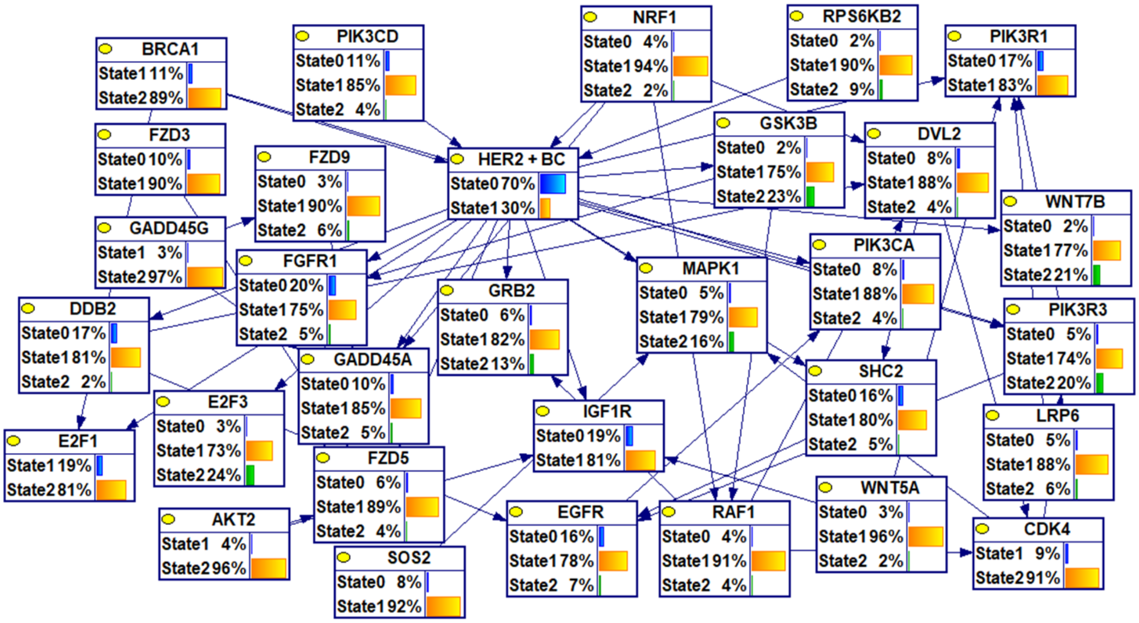


Figure 11. Bayesian probabilistic graphic model using Markov blanket genes of ER-PR-HER2+ breast cancer.

To validate the key Markov blanket genes as causal/signature genes for possible ER-PR-HER2+ breast cancer targets or biomarkers, it is important to analyze its predictive capability to distinguish between normal healthy and ER-PR-HER2+ breast tumor cases. Genie's "learn parameters" (Figure 11) function analysis of the 30 genes associated with the HER2+ breast cancer network showed that 10 Markov blanket genes were able to consistently distinguish between nontumor and tumor cases.

The prediction accuracy to distinguish normal healthy or HER2+ breast tumor cases was alternatively verified by expression patterns of the combination of genes. Table 5 summarizes the top 12 maximum relative risk (RR) of the minimum set of combination of gene expression patterns in predicting HER2 BC. The likelihood of HER2 BC is almost 100% in a patient with the expression pattern of the [high] NRF1 combined with E2F1 [low or no change], E2F3 [high], FGFR1 [low or no change], GSK3B [high], MAPK1 [high], and PIK3R3 [high or no change]. Whereas a subject that has low NRF1 expression combined with E2F1 [no change], E2F3 [low or no change], FGFR1 [no change], GSK3B [no change], MAPK1 [low], and PIK3R3 [low] expression has almost 0% probability of HER2 BC. This finding confirmed the association of high NRF1 combined with its target genes showed high probability of HER2+ breast cancer.

Discussion

Major advances in HER2 targeted therapies have been made; nevertheless, there are many women with hormone receptor negative, HER2+ metastatic breast cancer, who do not experience the same success with these therapies. This subtype of breast cancer, along with triple negative breast cancer (TNBC), are of major concern because they are associated with increased recurrence, lower survival rates, and higher rates of metastasis to the brain compared to other breast cancer subtypes (Wu et al., 2016). Despite tremendous progress in the understanding of breast cancer, gaps remain in our knowledge of the molecular basis underlying the disparity in aggressiveness of HER2+ breast cancer and its metastasis to the brain. Thus, knowledge of the molecular basis underlying the disparity in aggressiveness and resistance to therapy, and new molecular drug targets, are urgently needed for women diagnosed with this breast cancer subtype.

NRF1 is a redox-sensitive pioneer transcription factor. Embryonic stem cells have been shown to have roughly 33% of all active genes bound by NRF1 (ENCODE Project Consortium, 2012). NRF1 appears to be involved in several human cancers, including breast cancer (Ertel et al. 2012; Falco et al., 2016). NRF1 activity correlated significantly with histological grades and prognosis of BC (Gao et al., 2018; Niida et al., 2008). This study revealed that both mRNA and protein expression of NRF1 were significantly higher in ER-PR-HER2+ breast cancer samples compared to normal breast tissues. This is consistent with previous and our current observations showing higher expression of NRF1 in EGFR/HER2+ tumors in the experimental model (Weaver et al., 2012).

Table 5

Summary of the Top 12 Maximum Relative Risk (RR) of the Minimum Set of Combination of Gene Expression Patterns

Gene Expression Patterns							Probability of HER2 BC	RR
E2F1[0]	E2F3[2]	FGFR1[0]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[2])	0.999998	1.52E+04
E2F1[1]	E2F3[1]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	6.58E-05	
E2F1[0]	E2F3[2]	FGFR1[0]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[1])	0.999996	1.52E+04
E2F1[1]	E2F3[1]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	6.58E-05	
E2F1[0]	E2F3[2]	FGFR1[1]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[2])	0.999996	1.52E+04
E2F1[1]	E2F3[1]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	6.58E-05	
E2F1[0]	E2F3[2]	FGFR1[0]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[2])	0.999994	1.52E+04
E2F1[1]	E2F3[1]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	6.58E-05	
E2F1[1]	E2F3[2]	FGFR1[0]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[2])	0.999994	1.52E+04
E2F1[1]	E2F3[1]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	6.58E-05	
E2F1[0]	E2F3[2]	FGFR1[1]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[1])	0.999991	1.52E+04
E2F1[1]	E2F3[1]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	6.58E-05	
E2F1[0]	E2F3[2]	FGFR1[0]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[2])	0.999998	1.18E+04

Gene Expression Patterns							Probability of HER2 BC	RR
E2F1[1]	E2F3[0]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	8.50E-05	
E2F1[0]	E2F3[2]	FGFR1[0]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[1])	0.999996	1.18E+04
E2F1[1]	E2F3[0]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	8.50E-05	
E2F1[0]	E2F3[2]	FGFR1[1]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[2])	0.999996	1.18E+04
E2F1[1]	E2F3[0]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	8.50E-05	
E2F1[0]	E2F3[2]	FGFR1[0]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[2])	0.999994	1.18E+04
E2F1[1]	E2F3[0]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	8.50E-05	
E2F1[1]	E2F3[2]	FGFR1[0]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[2])	0.999994	1.18E+04
E2F1[1]	E2F3[0]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	8.50E-05	
E2F1[0]	E2F3[2]	FGFR1[1]	GSK3B[2]	MAPK1[2]	NRF1[2]	PIK3R3[1])	0.999991	1.18E+04
E2F1[1]	E2F3[0]	FGFR1[1]	GSK3B[1]	MAPK1[0]	NRF1[0]	PIK3R3[0])	8.50E-05	

Note. [0] = Low expression, [1] = No Change in the expression, [2] = High expression.

These findings further provide support to the concept that transcription control of NRF1 seems to be dependent on EGFR signaling. Discovery of NRF1 localization to several thousand sites in the human genome may indicate they occupy up to 15% of the promoter regions. NRF1 binding activity was higher in HER2 amplified HCC1954 breast cancer cells compared to normal mammary epithelial cells. Here, we observed new roles of NRF1 in contributing to critical pathways involved in the transformation of normal cells to cancerous cells. These roles included PI3K-Akt, MAPK, mTOR, and Wnt signaling pathways controlling cellular senescence, sustaining proliferative signaling; p53 and TGF-beta signaling evading growth suppressors; apoptosis and autophagy resisting cell death; enabling replicative immortality hallmark; VEGF signaling inducing angiogenesis; and finally, the pathways T and B cell receptor signaling evading immune destruction.

HCC1954 is known to be trastuzumab resistant due to a hotspot PIK3CA mutation (H1047R, PI3K gain-of-function) (Kataoka et al., 2010; von der Heyde et al., 2015). Interestingly, NRF1 transcriptional control involving the PI3K-AKT pathway was observed in this study, which fits the PI3K gain-of-function in this resistant HCC1954 cell line. Our results may also point to an important role of NRF1 in driving trastuzumab resistance via regulating PI3K-AKT pathway. This finding may open a new direction of NRF1's role in HER2+ breast cancer resistance to therapy.

To understand the mechanistic aspects of the contribution of NRF1 in susceptibility to the HER2+ breast cancer subtype, we focused our efforts on NRF1 motif-enriched 59 genes, including AKT1, BRCA1, EGFR, which are implicated in breast cancer. The majority of these genes, which participate in the process of cells acquiring characteristics of malignancy, contain NRF1 binding sites in the region located +/- 2,000 bp from TSS. To further understand how these NRF1 target genes may contribute in HER2 amplified breast cancer, we conducted a Bayesian network analysis of NRF1 target genes. In addition to known genes involved in KEGG HER2+ breast cancer pathway, such as growth factor receptor genes- FGFR1, IGF1R; MAPK pathway genes—SHC2, GRB2, MAPK1; and E2Fs—E2F1 and E2F3, we observed mTOR signaling pathway genes-PIK3CD, PIK3R1, PIK3R3, RPS6KB2; NRF1 motif-enriched WNT signaling pathway genes—WNT7B, DLV1, DLV2, GSK3B—and the damaged DNA binding protein 2 (DDB2), known for its role in DNA repair, were strongly associated with HER2 amplified breast cancer. WNT7B is known to be associated with angiogenesis, invasion, and metastasis of breast cancer (Yeo et al., 2014).

There is a direct relationship between an increase in gene expression of NRF1, WNT7B, and PIK3R3 and the probability of HER2 amplified breast cancer. DDB2 has been recently shown to be involved in early events associated with metastatic progression of breast cancer cells (Barbieux et al., 2016). Both E2F1 and E2F3 are important mediators of HER2/Neu-initiated mammary tumorigenesis (Andrechek, 2015). Deregulation of E2Fs (E2F1 and E2F3) contributes in centrosome amplification in HER2+ HCC1954 cells (Lee, Moreno, & Saavedra, 2014), and deregulated expression of the E2Fs in breast cancers influences outcome of survival and chemotherapeutic responses, including resistance to the

Cdk4/Cdk6 inhibitor PD-0332991 (Lee et al., 2014). Thus, further study of NRF1-regulated breast cancer hallmark pathways may provide clues not only to understanding of how HER2+ breast tumors initiate and progress but also may help to explain how HER2+ breast cancer cells fail to respond to common therapies.

NRF1 may regulate target gene expression in HER2+ breast cancer cells either alone or in combination with additional factors. NRF1 is a “pioneer transcription factor” and its binding to DNA in condensed chromatin allows access to “settler transcription factor” to bind to its motif sequences (Sherwood et al., 2014, p. 174). When methylation prevents NRF1 binding to its motif sequence, it acts as a “settler transcription factor,” and requires other factors, such as a demethylase, to remove methylated residues from its motif for binding (Domcke et al., 2015, p. 578). Promoters containing the nuclear respiratory factor 1 (NRF1) motif are pervasively associated with lysine-specific demethylase 1 (LSD1/KDM) (Benner et al., 2013). Recently, Campoy et al. (2016) discovered changes in the levels of DNA methylation in breast tumors are linked to LSD1, one of the main cofactors of NRF1. NRF1 also interacts with histone variants such as mH2A1s that promote or repress target gene activities through chromatin modifications (Lavigne et al., 2015). Work is currently under way in our laboratory to determine how NRF1 in concert with additional factors may regulate target gene expression in HER2+ breast cancer.

In conclusion, we applied the ChIP DNA-seq and RNA-microarray coupled with identification of signaling pathways and functional enrichment analysis to identify differentially regulated NRF1 target genes involved in ER-PR-HER2+ breast cancer and Bayesian machine learning method to understand their role in this disease. The findings of our study suggest that the gain of NRF1 function may contribute to the susceptibility of ER-PR-HER2+ breast cancer subtype via perturbation of regulation of diverse growth factor receptors, PI3K-Akt-mTOR, MAPK, E2Fs, and Wnt pathways. Clinical confirmation of our study will have a significant impact on understanding the role of NRF1 as a valuable additional biomarker for assessing resistance to therapeutic response in HER2+ breast cancer and will provide a strong rationale for the future studies to further develop NRF1 signaling-based therapy for HER2+ breast cancer.

Methods

Analysis of NRF1 mRNA Expression in HER2+ Breast Tumor Samples

The Cancer Genome Atlas (TCGA) microarray data from 61 normal samples and 22 HER2+ breast tumor samples [log2 lowess normalized (cy5/cy3)] was downloaded using Broad Institute's Firehose tool -Version: std. data 2016-01-28 (Cancer Genome Atlas Network, 2012). The SAS PROC TTEST was used to compare the means of the two groups.

NRF1 Activity

Transcription factor activity of NRF1 was assessed as a function of the collective mRNA levels of its target genes in normal and ER-PR-HER2+ breast cancer TCGA samples using the limma R package (Falco et al., 2016).

Analysis of NRF1 Protein Expression

Human breast cancer tissue arrays were purchased from US Biomax, Rockville, Maryland. Sections of experimental model of breast cancer brain metastasis from MDA-MB-231-BR-HER2 cells were kindly provided by Dr. Donna Murrell (Murrell et al., 2015) and MDA-MB-231-BR-vector cells by Dr. Brunilde Gril (Gril et al., 2008) injected intracardially into mice. NRF1 was measured by immunofluorescence confocal microscopy (IFC) using NRF1 specific antibodies paired with Alexafluor 488 and DRAQ5 for nuclear stain. Expression was scored as low (<the median intensity value), and high (>the median intensity value) levels per cancer cells based on immunofluorescence tissue staining intensity.

Identification of NRF1 Target Genes

We retrieved the following NRF1 ChIP sequence dataset from NCBI-Gene expression omnibus (GEO) and uploaded the Sequence Read Archive (SRA) files directly into GALAXY using the NCBI SRA Tool under GALAXY's menu (Domcke et al. 2015): NRF1 input in HMEC: SRR2500899 - GSM1891657, NRF1 input in HCC1954: SRR2500902 - GSM1891660, NRF1 ChIP in HCC1954 replicate #1: SRR2500900 - GSM1891658, NRF1 ChIP in HCC1954 - replicate # 2: SRR2500901 - GSM1891659, NRF1 ChIP in HMEC- replicate # 1: SRR2500897 - GSM1891655,NRF1 ChIP in HMEC- replicate # 2: SRR2500898 - GSM1891656.

ChIP-Seq experiments were conducted at Friedrich Miescher Institute for Biomedical Research in Switzerland using ChIP antibody NRF1 ABCm, ab55744. Sequencing was performed using Illumina machine HiSeq 2500 at 50 bp read length single end, following Illumina standards (Domcke et al., 2015). Initially we evaluated the quality of the data using the FASTQC software accessed through GALAXY, available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

We then used Trimmomatic (http://www.usadellab.org/cms/index .php?page=trimmomatic) to cut the adapters, drop readings with an average quality Phred score below 20 (number of bases to average across = 4) and discard sequences below 36 and 70 bases for 50 and 100 bp length readings (minimum read length). Phred score of 20 is equivalent to a probability of 1 in 100 that the base is called wrong (99% accuracy of the base call).

The next step was mapping the readings on the human genome reference Hg19 using Bowtie2. Subsequently, we used MACS2 (Galaxy Version 2.1.1.20160309.0) to identify peaks from alignment results, setting up a false discovery rate (*q* value) cutoff of 0.01%. Peaks were filtered discarding those with fold enrichment lower than 5, following ENCODE guidelines for point-source transcription factors (Landt et al., 2012). We then performed Gene annotation using GREAT 3.0.0, available at http://bejerano.stanford.edu/great/public/html /index.php to discover the list of candidate NRF1 target genes. GREAT predicts functions of cis-regulatory regions using different settings.

Initially we used the Basal plus extension option in which each gene is assigned a regulatory domain region 5,000 bp upstream and 1,000 bp downstream of the TSS. This gene regulatory domain is extended in both directions to the nearest gene's regulatory domain but no more than a maximum extension of 1,000 bp. Additionally, the software also considers peaks falling in other literature curated regulatory domains with experimental evidence of being regulatory elements for a specific gene, regardless of its location. The intersection was found using VENNY 2.1.0. VENNY is an interactive tool used to compare lists with Venn diagrams, developed by J. C. Oliveros that can be found at http://bioinfogp.cnb.csic.es/tools/venny/index.html

Analysis of Similarity in the NRF1 Network of Different Cell Lines

To measure the overlap between NRF1 target genes in human epithelial mammary cells and the HCC1954 breast cancer cell line used in this study, we calculated the Jaccard coefficient (JC), which is defined as the intersection (common genes) divided by the union of the sample sets. The formula used was JC = $\frac{A \cap B}{A \cup B}$, in which A is the list of NRF1 target genes in normal mammary epithelial cells (HMEC) and B is the list of NRF1 target genes in HCC1954 breast cancer cell line.

Identification of Signaling Pathways and Functional Enrichment Analysis to Select Genes of Interest

The functional annotation tool from DAVID and KEGG were utilized to identify NRF1 target genes in each one of the signaling pathways of interest. To further investigate the mechanisms of breast cancer development, we found the overlap of these genes with the genes in the breast cancer pathway (KEGG). This last step was conducted with an Excel sheet using formulas to select only those genes that were in the signaling pathway of interest and also were part of the breast cancer pathway. Seventy-eight genes were selected in this process. DAVID and KEGG are available to the general public at http://david.ncifcrf.gov and http://www.genome.jp/kegg/.

Protein-Protein Interaction and Protein-DNA Networks Among Selected Genes

After selection of the 78 genes associated with hallmarks of cancer development, we compared this list to the list of more than 11,000 NRF1 target genes reported in ENCODE by Harmonizome (Rouillard et al., 2016) to filter again those 78 genes and make the list more selective. This process resulted in 59 genes overlapping. To investigate the protein-protein interaction among these 59 genes, we used the Search Tool for the Retrieval of Interacting Genes (STRING) database (Version 10.5), found at: https://string-db.org. We also used Cytoscape software (Version 3.4.0) found at http://www.cytoscape.org/ to manually construct the protein-DNA network.

Changes in NRF1 Binding in Proximal Promoter Regions of Target and Selected Breast Cancer Genes in HMEC vs. HCC1954 Cell Lines and Correlation with Breast Cancer

Out of the more than 10,000 NRF1 target genes we found in the ChIP DNA sequence analysis, we selected those genes with NRF1 binding peaks located on or near the promoter regions (+/- 2,000 bp from TSS). We then calculated the average Fold Enrichment (FE) of the peaks for each one of the 7,663 genes that met the criterion. This average FE measures the relative amount of NRF1 protein bound to the DNA. Subsequently, we ran a point biserial correlation to measure the strength of the relationship between breast cancer, which we coded as a binary variable (0 in HMEC and 1 in HCC1954), and the average NRF1 peak in the promoter region. We selected the top 2,000 correlated genes (1,000 positive correlated and 1,000 negative correlated) (absolute r_{pb} was between a maximum of 0.866 to a minimum of 0.790) to draw a bar graph. We used Excel to better visualize the difference in NRF1 binding for the two cell lines in these specific set of genes.

We then compared the list of 59 selected candidate genes (see previous section) against the list of 7,663 with NRF1 binding sites in the promoter regions to select only those overlapping. The new list of 49 genes plus NRF1 was used as the list of variables to perform the Bayesian analysis and develop the probabilistic graphic model described in the next section. TCGA dataset of normal mammary tissue and breast tumor classified as ER-,PR- and HER2+ (similar genetic profile to HCC1954 cells) were collected for this analysis.

Bayesian Modelling of NRF1 Target Genes in HER2+ Breast Cancer

We performed a Bayesian data analysis with the 49 candidate genes including two additional variables, NRF1 for obvious reasons and the disease status (HER2+ BC), for a total of 51 variables. The goal of this analysis was to create a proposed network showing the interaction among these variables in Her2+ (Her2+ER−PR−) breast cancer subtype to identify possible drivers. The nodes represented the expression microarray data collected from TCGA. We were able to identify 22 primary solid tumors with the genetic profile Her2+ER−PR− and 61 normal tissues for a total of 83 samples.

The software that carries out the Bayesian network learning process generates a series of probabilistic graphical networks known as directed acyclic graphs (DAG) that represent a set of random variables and their conditional dependencies. The nodes of the networks represent the expression of the genes and the clinical variables (in this case we only included one clinical variable the disease—HER2+ BC). All of these were assumed to be variables conditionally independent from each other. Resulting networks are graphic representation sof the causal hypothesis (Friedman, Linial, Nachman, & Pe'er, 2000; Kunkle, Yoo, & Roy, 2013).

We used the software Banjo developed by Duke University to find the best Bayesian network. Since Banjo needed the variables to be categorized, we used the gene expression values of the normal tissue samples and calculated the mean +/-2 standard deviations as the cutoff points for low and upregulation. Banjo works by performing structure inference scoring metrics for discrete variables. The scoring metric used is called Bayesian Dirichlet equivalence (BDe). The program keeps making incremental changes in the structure to improve the score of the network. The final DAG shows regulation between genes and their possible involvement on the outcome (disease) (Kunkle et al., 2013). Additional explanation of Bayesian network learning is given in the results section.

Bayesian Parameter Estimation of Proposed Network and Sensitivity Analysis of HER2 + Breast Cancer Probability

To calculate the probabilities of the variables in the proposed probabilistic graphic model and to test the sensitivity of breast cancer status to changes in the gene expression of Markov genes, we recreated the best Banjo network structure using the software GeNIe. This tool was developed by the Decision Systems Laboratory of the University of Pittsburgh, available at http://genie.sis.pitt.edu. The

TCGA microarray dataset was uploaded. The GeNIe performed the estimation of the parameters for each node. After the structure and parameters were assembled, a sensitivity analysis was conducted modifying the evidence (marginal probability) of different nodes (gene expression stages) and observing the effect on the probability of the breast cancer node BC- (probability of HER2 breast cancer).

Estimation of the Minimum Set of Combination of Gene Expression Patterns That Yield a Maximum Relative Risk (RR)

We used 15 genes and used the model presented by BANJO and calculated all 4,251,528 $(= 2^3 \times 3^{12})$ different gene configurations *g* with the collected dataset and using the SMILE (2018) library (https://www.bayesfusion.com/smile-engine) and C++ program and calculate the following:

$$P(D|R = g)$$

where *D* represents a subject has HER2 breast cancer and *R* = (DDB2, E2F1, E2F3, FGFR1, GRB2, GSK3B, IGF1R, MAPK1, NRF1, PIK3CD, PIK3R1, PIK3R3, RPS6KB2, SHC2, WNT7B). Among the gene configurations *g* that predicts HER2 breast cancer with high or low probability (i.e., $P(D|R = g) > 0.99999$ or $P(D|R = g) < 1.0 \times 10^{-4}$), we focused on *g* where NRF1 was either expressed high or expressed low.

To find the minimum set of combination of gene expression patterns that give us a maximum Relative Risk (RR), we calculated the following:

$$S = \underset{Q}{argmax} \frac{P(D|Q = q)}{P(D|Q = q')}$$

where *Q* represents any subset of *R*, *S* represents a set of the minimum number of genes that maximizes the RR term, $q = \underset{g}{argmax} P(D|Q = g)$ and $q' = \underset{g}{argmin} P(D|Q = g)$. Note that *q* and *q'* represents two different gene expression patterns among the genes in *S* that maximize and minimize $P(D|Q),$ respectively. We report the top 12 RR that we calculated from the dataset.

REFERENCES

Andrechek, E. R. (2015). HER2/Neu tu morigenesis and metastasis is regulated by E2F activator transcription factors. *Oncogene, 34*(2), 217-225.

Barbieux, C., Bacharouche, J., Soussen, C., Hupont, S., Razafitianamaharavo, A., Klotz, R., . . . Grandemange, S. (2016). DDB2 (damaged-DNA binding 2) protein: A new modulator of nanomechanical properties and cell adhesion of breast cancer cells. *Nanoscale, 8*(9), 5268-5279.

Benner, C., Konovalov, S., Mackintosh, C., Hutt, K. R., Stunnenberg, R., & Garcia-Bassets, I. (2013). Decoding a signature-based model of transcription cofactor recruitment dictated by cardinal cis-regulatory elements in proximal promoter regions. *PLoS Genetics, 9*(11), 1-18.

Campoy, E. M., Laurito, S. R., Branham, M. T., Urrutia, G., Mathison, A., Gago, F., . . . Roqué, M. (2016). Asymmetric cancer hallmarks in breast tumors on different sides of the body. *PloS One, 11*(7), 1-20.

Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*(7418), 61-70.

Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L., & Schubeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature, 528*(7583), 575-579.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74.

Ertel, A., Tsirigos, A., Whitaker-Menezes, D., Birbe, R. C., Pavlides, S., Martinez-Outschoorn, U. E., . . . Lisanti, M. P. (2012). Is cancer a metabolic rebellion against host aging? in the quest for immortality, tumor cells try to save themselves by boosting mitochondrial metabolism. *Cell Cycle, 11*(2), 253-263.

Falco, M. M., Bleda, M., Carbonell-Caballero, J., & Dopazo, J. (2016). The pan-cancer pathological regulatory landscape. *Scientific Reports, 6,* 39709, 1-13.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 7*(3-4), 601-620.

Gao, W., Wu, M., Wang, N., Zhang, Y., Hua, J., Tang, G., & Wang, Y. (2018). Increased expression of mitochondrial transcription factor A and nuclear respiratory factor-1 predicts a poor clinical outcome of breast cancer. *Oncology Letters, 15*(2), 1449-1458.

Gril, B., Palmieri, D., Bronder, J. L., Herring, J. M., Vega-Valle, E., Feigenbaum, L., . . . Steeg, P. S. (2008). Effect of lapatinib on the outgrowth of metastatic breast cancer cells to the brain. *Journal of the National Cancer Institute, 100*(15), 1092-1103.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646-674.

Henjes, F., Bender, C., von der Heyde, S., Braun, L., Mannsperger, H. A., Schmidt, C., . . . Korf, U. (2012). Strong EGFR signaling in cell line models of ERBB2-amplified breast cancer attenuates response towards ERBB2-targeting drugs. *Oncogenesis, 1,* 1-9.

Kataoka, Y., Mukohara, T., Shimada, H., Saijo, N., Hirai, M., & Minami, H. (2010). Association between gain-of-function mutations in PIK3CA and resistance to HER2-targeted agents in HER2-amplified breast cancer cell lines. *Annals of Oncology: Official Journal of the European Society for Medical Oncology, 21*(2), 255-262.

Kunkle, B. W., Yoo, C., & Roy, D. (2013). Reverse engineering of modified genes by bayesian network analysis defines molecular determinants critical to the development of glioblastoma. *PloS One, 8*(5), 1-23**.**

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., . . . Chen, Y. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research, 22*(9), 1813-1831.

Lavigne, M. D., Vatsellas, G., Polyzos, A., Mantouvalou, E., Sianidis, G., Maraziotis, I., . . . Thanos, D. (2015). Composite macroH2A/NRF-1 nucleosomes suppress noise and generate robustness in gene expression. *Cell Reports, 11*(7), 1090-1101.

Le, T. D., Liu, L., Liu, B., Tsykin, A., Goodall, G. J., Satou, K., & Li, J. (2013). Inferring microRNA and transcription factor regulatory networks in heterogeneous data. *BMC Bioinformatics, 14*(1), 1-13.

Lee, M., Oprea-Ilies, G., & Saavedra, H. I. (2015). Silencing of E2F3 suppresses tumor growth of Her2+ breast cancer cells by restricting mitosis. *Oncotarget, 6*(35), 37316-37334.

Lee, M. Y., Moreno, C. S., & Saavedra, H. I. (2014). E2F activators signal and maintain centrosome amplification in breast cancer cells. *Molecular and Cellular Biology, 34*(14), 2581-2599.

Murrell, D. H., Hamilton, A. M., Mallett, C. L., van Gorkum, R., Chambers, A. F., & Foster, P. J. (2015). Understanding heterogeneity and permeability of brain metastases in murine models of HER2-positive breast cancer through magnetic resonance imaging: Implications for detection and therapy. *Translational Oncology, 8*(3), 176-184.

Niida, A., Smith, A. D., Imoto, S., Tsutsumi, S., Aburatani, H., Zhang, M. Q., & Akiyama, T. (2008). Integrative bioinformatics analysis of transcriptional regulatory programs in breast cancer cells. *BMC Bioinformatics, 9*(1),1-13.

Okoh, V., Deoraj, A., & Roy, D. (2011). Estrogen-induced reactive oxygen species-mediated signalings contribute to breast cancer. *Biochimica et Biophysica Acta, 1815*(1), 115-133.

Okoh, V. O., Garba, N. A., Penney, R. B., Das, J., Deoraj, A., Singh, K. P. . . . Roy, D. (2015). Redox signalling to nuclear regulatory proteins by reactive oxygen species contributes to oestrogen-induced growth of breast cancer cells. *British Journal of Cancer, 112*(10), 1687-1702.

Piantadosi, C. A., & Suliman, H. B. (2006). Mitochondrial transcription factor A induction by redox activation of nuclear respiratory factor 1. *Journal of Biological Chemistry, 281*(1), 324-333.

Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G., & Ma'ayan, A. (2016). The harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database, 2016*, 1-16.

Roy, D., & Tamuli, R. (2009). NRF1 (nuclear respiratory factor 1). *Atlas of Genetics and Cytogenetics in Oncology and Haematolology. 13*(11), 861-864.

Scarpulla, R. C. (2006). Nuclear control of respiratory gene expression in mammalian cells. *Journal of Cellular Biochemistry, 97*(4), 673-683.

Scarpulla, R. C. (2008). Transcriptional paradigms in mammalian mitochondrial biogenesis and function. *Physiological Reviews, 88*(2), 611-638.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*(11), 2498-2504.

Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., . . . Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology, 32*(2), 171-178.

SMILE: Structural Modeling, Inference, and Learning Engine. BayesFusion, LLC, 2018. Retrieved from https://www.bayesfusion.com/smile-engine

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., . . . Thorsen, T. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America, 98*(19), 10869-10874.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., . . . Kuhn, M. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research, 43(Database issue),* D447-D452.

von der Heyde, S., Wagner, S., Czerny, A., Nietert, M., Ludewig, F., Salinas-Riester, G., . . . Beißbarth, T. (2015). mRNA profiling reveals determinants of trastuzumab efficiency in HER2-positive breast cancer. *PloS One, 10*(2), 1-27.

Weaver, Z., Difilippantonio, S., Carretero, J., Martin, P. L., El Meskini, R., Iacovelli, A. J., . . . Baran, M. (2012). Temporal molecular and biological assessment of an erlotinib-resistant lung adenocarcinoma model reveals markers of tumor progression and treatment response. *Cancer Research, 72*(22), 5921-5933.

Wu, X., Baig, A., Kasymjanova, G., Kafi, K., Holcroft, C., Mekouar, H. . . . Muanza, T. (2016). Pattern of local recurrence and distant metastasis in breast cancer by molecular subtype. *Cureus, 8*(12), 1-15.

Yeo, E. J., Cassetta, L., Qian, B. Z., Lewkowich, I., Li, J. F., Stefater, J. A. 3rd, . . . Lang, R. A. (2014). Myeloid WNT7b mediates the angiogenic switch and metastasis in breast cancer. *Cancer Research, 74*(11), 2962-2973.

CHAPTER V

DIFFERENTIAL NRF1 GENE NETWORK SENSITIVITY CONTRIBUTING TO BREAST CANCER DISPARITIES

Abstract

This study investigated a novel molecular mechanism to help explain the higher invasive breast cancer disparity in African Americans by examining contribution of the differences in the nuclear respiratory factor 1 (NRF1) sensitivity to the racial/ethnic disparity of invasive breast cancer. The significance of this clinically translational knowledge will be in predicting the clinical outcomes of African American (AA) and non-Hispanic Whites (NHW) who are most susceptible to invasive breast cancer. This is a topic of high relevance to breast cancer disparities. Invasive breast cancer, particularly triple-negative, is both aggressive and nonresponsive to existing therapies. AA patients have higher breast cancer mortality in part due to the three times higher proportion of triple-negative breast cancer (TNBC) cases among AA patients compared to European American (EA) women even though the incidence is lower in AA women. It is crucial to understand the racial differences in molecular signatures to develop targeted therapy, and subsequently, increase the survival rate of AA women with TNBC. A lack of effective molecular targets as well as limited therapeutic options, particularly for AA breast cancer patients, leads to high morbidity and poor survival. Our novel research has shown that NRF1 overexpression drives estrogen-dependent breast tumorigenesis. However, the impact of the NRF1 pathway on breast cancer metastasis is unknown. Herein, our objective was to examine an untested and

highly innovative hypothesis in breast cancer disparities research, i.e., that differential NRF1 sensitivity contributes to disparities in susceptibility to basal triple-negative breast cancer in racial/ethnic groups of breast cancer patients, AA and NHW women. The findings of this study will elucidate the roles of NRF1 sensitivity to develop TNBC in different racial/ethnic groups of breast cancer patients. This elucidation could provide new strategies to delay or even to prevent this important clinical problem. Such strategies may allow personalized intervention and treatment.

Introduction

Despite tremendous progress in the understanding breast cancer (BC), gaps remain in our knowledge of the molecular basis underlying the disparity in aggressiveness of BC and the metastasis to the different organs. Thus far, we have not made a major leap in our understanding of the molecular causes of racial disparity in BC. Earlier molecular epidemiological population studies were primarily focused on socioeconomic factors, health care access, and Mendelian genetics-based ancestral heredity to explain breast cancer disparity.

These studies successfully showed that differences in environment, economic factors, and lifestyle contribute to the disparity in the incidence and mortality of breast cancer. However, the studies did not take into account the contribution of stochastic reprogramming resulting in multiple lineages of human breast cancer stem/progenitor cells, gene-environment interactions, and gene-gene interactions to explain breast cancer disparity. Indeed, emerging data now suggest that, in addition to socioeconomic factors and lifestyle differences, biological factors, such as differences at the genetic and epigenetic levels, are crucial for understanding the pathogenesis of breast cancer in the United States general population of diverse ancestral lineages accounting for individual variability in genes, environment, and lifestyle for each person.

This recognition has resulted into an initiative towards translational basic research for establishment and precise understanding of the involved molecular mechanisms and identification of the causal elements in gene regulatory networks driving the etiology of breast cancer in the individual patient as well as the general population to address racial inequalities in breast cancer incidence and clinical outcomes. However, such information is emerging based on precision genomics, but in most molecular epidemiological studies, breast cancer patients are often stratified as White (Caucasian or European) American, African American, or Latino/Hispanic American based on race/ethnicity. Each patient group is considered as a single race/ethnicity possessing a distinctive "breast cancer phenotype."

On the contrary, the majority of Caucasian/European Americans, African Americans, and Latino/Hispanic Americans are genetically mixed and with several distinct racial types among them. Genetic evidence shows many distinct ancestries in the Caucasian/European American, African American. and Latino/Hispanic American populations (Bryc, Durand, Macpherson, Reich, & Mountain, 2015). A comprehensive research effort is needed to address the existing gap in the understanding of breast cancer disparity by accounting for individual variability in genes, environment, and lifestyle. Lack of robust methodology to analyze the interaction of multiple differentially expressed up or down genes identified from RNA-Seq data. Funding sources emphasizing focused research limit the uncovering this complex knowledge for understanding a biological disparity in the risk of breast cancer. The main focus of this study was to investigate a novel molecular mechanism that deciphers racial differences in the aggressive growth of BC.

We recently discovered that NRF1-regulated gene networks in breast cancer cells from women of Indian origin seem quite different from European White women. NRF1 is associated with several human cancers, including breast cancer. Genes from the KEGG HER2+ breast cancer pathway and 11 signaling pathways linked to six hallmarks of cancer seem to be under transcription control of NRF1 (Ramos et al., 2018). In this study, we have expanded our efforts to identify the causal elements in the NRF1 gene regulatory networks driving etiology of breast cancer disparities.

Results

Transcription Factor Target Enrichment Analysis (TFTEA) Reveals Upregulation of NRF1 Activity Across Different Breast Cancer Subtypes Clustered by Patient’s Race and Ethnicity

TCGA breast cancer tumor samples were classified based on molecular subtypes, race, and ethnicity (Table 1). Some of the subclassifications did not have enough number of samples for application of the statistical tests and categorization, as explained in the Methods sections. Eight groups were selected for the study of changes in NRF1 activity compared to normal samples (Table 2).

These groups were four Luminal A (Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, and Hispanic White); two triple-negative (Non-Hispanic Black and Non-Hispanic White); one Luminal B (Non-Hispanic White); and one HER2 enriched (Non-Hispanic White). (African American and Black are used interchangeably in this study). A significant number of normal samples (79) were available only for Non-Hispanic White. Therefore, these normal samples were used as counterparts for calculating differential expression (DE) of all eight breast cancer clusters.

Table 1

Number of Breast Cancer and Normal Samples in TCGA Dataset Classified by Molecular Subtypes, Race, and Ethnicity

ETHNICITY	**RACE**	**Luminal A**	**Luminal B**	**HER2 enriched**	Triple-negative	**NA**	**NORMAL**	**TOT**
		ER+ and /or PR+ / HER2-	**ER+ and /or PR+ / HER2+**	**ER-/ PR- / HER2+**	**ER-/ PR- / HER2-**			
Hispanic or Latino	**Asian**	1						1
	Black or African American				1	1		2
	white	21	4		7	2		34
	NA	1	1					2
		23	**5**	**0**	**8**	**3**	**0**	**39**
Non-Hispanic or Latino	**Asian**	22	6	8	8	14	1	59
	Black or African American	69	14	7	48	29	6	173
	White	373	83	17	74	112	79	738
	NA	1			1			2
	American Indian or			1				1

	Alaska native						
		465	**103**	**33**	**131**	**155**	**86**
NA	**Asian**	2					
	Black or African American	7	3		4		
	White	46	9	1	9	2	25
	NA	57	25	3	6		1
		112	**37**	**4**	**19**	**2**	**26**
		600		**37**	**158**	**160**	**112**

Table 2

NRF1 Activity in Breast Cancer Based on Differential Expression of Target Genes

	Breast Cancer samples			**Normal samples**			**TFTEA**		
MOL SUBT	**ETHNICITY AND RACE**		**# of samples**	**ETHNICITY AND RACE**		**# of samples**	**# of NRF1 target genes with DE p<0.05**	**DIRECTION**	**P-VALUE**
HER2 Enriched	Non-Hispanic	White	17	Non-Hispanic	White	79	2,252	Upregulation	3.32E-07
Luminal A	Non-Hispanic	Asian	22	Non-Hispanic	White	79	2,739	Upregulation	2.21E-06
Luminal A	Non-Hispanic	White	373	Non-Hispanic	White	79	3,103	Upregulation	5.32E-06
Luminal B	Non-Hispanic	White	83	Non-Hispanic	White	79	2,683	Upregulation	5.47E-06
Luminal A	Hispanic	White	21	Non-Hispanic	White	79	2,793	Upregulation	6.01E-06

Triple-negative	Non-Hispanic	White	74	Non-Hispanic	White	79	2,760	Upregulation	9.48E-06
Luminal A	Non-Hispanic	Black	69	Non-Hispanic	White	79	2,740	Upregulation	2.56E-04
Triple-negative	Non-Hispanic	Black	48	Non-Hispanic	White	79	2,696	Upregulation	4.90E-03

Normalized RNA-Seq gene expression of 20,501 genes in breast cancer samples were compared to their counterparts in normal samples to obtain the average DE for each of the genes and for each group of breast cancer tumors. The R/Bioconductor software package limma was used for this task. Features of differential expression analysis using limma are explained in the methods section.

Transcription Factor Target Enrichment Analysis (TFTEA) was then applied to the DE gene lists of each cluster to determine changes in NRF1 activity. A logistic regression approach using the gene set enrichment application called LRpath (Sartor, Leikaur, & Medvedovic, 2009) was utilized for this purpose. This application measures the enrichment and direction (upregulation or downregulation) of a set of biologically related genes (NRF1 target genes in this case) using the list of differentially expressed genes.

The TFTEA results shown in Table 2 indicate that NRF1 activity was significantly increased (upregulated) in all eight groups of breast cancer samples compared to normal tissue samples (*p* values for the logistic regression coefficients were under 0.05). These results suggest that NRF1 plays a role in breast cancer development. Table 2 also provides information on the number of NRF1 target genes found with statistically significant differential expression in breast cancer tissues ($p < 0.05$). This number ranges from a maximum of 3,103 genes for Luminal A tumors in Non-Hispanic Whites to a minimum of 2,252 for HER2 enriched tumors in patients within the same ethnicity and race. This range means that 26% to 37% of NRF1 targets were differentially expressed.

Bayesian Networks Learning Aimed to Discover Breast Cancer-Causal Hypothesis Genes Shows Differences Among Different Subtypes Grouped by Race and Ethnicity

The lists of differentially expressed genes identified during TFTEA were screened using the list shown in Table 3 to select the genes involved in hallmarks of cancer signaling pathways. Subsequently, the input data matrix of gene expression (RNA-Seq) data was created with selected genes on the rows and samples (BC and normal) in columns. Banjo is a software package developed by Duke University used in this research for learning Bayesian networks from data.

The graphical representation of the best Bayesian network (probabilistic model) for Non-Hispanic White Luminal A breast cancer is shown in Figure 1. Nodes represent the variables (genes symbols and disease–BC). This is a partial view because the entire network is too large for display here. The partial view is presented as an example of all networks obtained for each cluster.

Table 3

NRF1 Target Genes in Signaling Pathways Linked to Hallmarks of Cancer

HALLMARKS OF CANCER	**SIGNALING PATHWAYS ASSOCIATED**	**NRF1 TARGET GENES (Identified by ChIP-seq using HCC1954-breast cancer cells and HMEC-normal mammary epithelial cells)**
Sustaining proliferative signaling	PI3K-Akt Signaling	ANGPT1, ANGPT2, EIF4EBP1, FGF13, FGF19, FGF3, FGF4, FGFR3, FGFR4, FLT4, GNG10, GNG3, GNG8, HGF, IBSP, IL4, IL7R, LPAR1, PCK1, PPP2R2C, PPP2R2D, PRLR, VTN, YWHAB, AKT1, AKT2, ANGPT4, ATF2, ATF4, ATF6B, BAD, BCL2, BCL2L1, BCL2L11, BRCA1, C8orf44-SGK3, CASP9, CCND1, CCND2, CCND3, CCNE1, CCNE2, CDC37, CDK4, CDK6, CDKN1A, CDKN1B, CHRM1, CHUK, COL1A1, COL2A1, COL4A1, COL4A2, COL4A3, COL4A4, COL6A3, COL6A5, COL9A3, COMP, CREB1, CREB3, CREB3L1, CREB3L2, CREB3L4, CREB5, CRTC2, CSF1, CSF1R, CSH2, DDIT4, EFNA1, EFNA2, EFNA3, EFNA4, EFNA5, EGFR, EIF4B, EIF4E, EIF4E2, EPHA2, EPO, EPOR, FGF1, FGF10, FGF11, FGF12, FGF18, FGF21, FGF22, FGF7, FGF9, FGFR1, FGFR2, FOXO3, G6PC3, GNB1, GNB3, GNB4, GNB5, GNG13, GNG2, GNG5, GNG7, GNGT2, GRB2, GSK3B, GYS1, HRAS, HSP90AA1, HSP90AB1, IFNAR1, IFNAR2, IGF1R, IKBKB, IL3RA, IL4R, IL6, IL6R, IL7, INSR, IRS1, ITGA10, ITGA11, ITGA2, ITGA2B, ITGA3, ITGA6, ITGA9, ITGAV, ITGB1, ITGB3, ITGB4, ITGB5, ITGB6, ITGB7, ITGB8, JAK1, JAK2, JAK3, KIT, KITLG, KRAS, LAMA1, LAMA2, LAMA3, LAMA5, LAMB1, LAMB2, LAMC1, LAMC2, LAMC3, LPAR3, MAP2K1, MAP2K2, MAPK1, MAPK3, MCL1, MDM2, MET, MLST8, MTCP1, MYB, MYC, NFKB1, NGF, NGFR, NOS3, NR4A1, OSM, OSMR, PCK2, PDGFA, PDGFB, PDGFC, PDGFRA, PDPK1, PHLPP1, PHLPP2, PIK3AP1, PIK3CA, PIK3CB, PIK3CD, PIK3CG, PIK3R1, PIK3R2, PIK3R3, PIK3R5, PKN1, PKN2, PKN3, PPP2CA, PPP2R1A, PPP2R2A, PPP2R3A, PPP2R3B, PPP2R3C, PPP2R5A, PPP2R5B, PPP2R5C, PPP2R5D, PPP2R5E, PRKAA1, PRKCA, PRL, PTEN, PTK2, RAC1, RAF1, RBL2, RELA, RHEB, RPS6KB1, RPS6KB2, RPTOR, RXRA, SGK1, SGK3, SOS1, SOS2, STK11, SYK, THBS1, THBS2, THBS3, THEM4, TLR2, TLR4, TSC1, TSC2, VEGFA, VEGFC, YWHAE, YWHAG, YWHAH, YWHAQ, YWHAZ, AKT3, COL9A1, COL9A2, EGF, FGF14, FGF2, FGF20, FGF5, FGF8, FN1, GHR, GNB2, GNG12, IFNA2, IFNA8, ITGA4, LAMB3, PDGFD, PDGFRB, PGF, PIK3R6, PPP2R1B, PPP2R2B, PRKAA2, TNN, TNR, VEGFB

HALLMARKS OF CANCER	SIGNALING PATHWAYS ASSOCIATED	NRF1 TARGET GENES (Identified by ChIP-seq using HCC1954-breast cancer cells and HMEC-normal mammary epithelial cells)
Sustaining proliferative signaling	MAP-kinase Pathway	CACNA1F , CACNG2 , CACNG4 , CACNG7 , ELK1 , FGF13 , FGF19 , FGF3 , FGF4 , FGFR3 , FGFR4 , MAPK10 , MAPK8IP1 , MAPKAPK2 , NTF3 , PLA2G4D , PLA2G4E , PRKCG , AKT1 , AKT2 , ARRB1 , ARRB2 , ATF2 , ATF4 , CACNA1A , CACNA1B , CACNA1C , CACNA1D , CACNA1E , CACNA1G , CACNA1H , CACNA1I , CACNA1S , CACNA2D1 , CACNA2D2 , CACNA2D3 , CACNA2D4 , CACNB1 , CACNG5 , CACNG6 , CACNG8 , CASP3 , CDC25B , CDC42 , CHUK , CRK , CRKL , DAXX , DDIT3 , DUSP1 , DUSP10 , DUSP16 , DUSP2 , DUSP3 , DUSP4 , DUSP5 , DUSP6 , DUSP7 , DUSP8 , DUSP9 , ECSIT , EGFR , ELK4 , FAS , FGF1 , FGF10 , FGF11 , FGF12 , FGF18 , FGF21 , FGF22 , FGF7 , FGF9 , FGFR1 , FGFR2 , FLNA , FLNB , FOS , GADD45A , GADD45B , GADD45G , GNA12 , GRB2 , HRAS , HSPA1A , HSPA1B , HSPA1L , HSPA2 , HSPA6 , HSPA8 , HSPB1 , IKBKB , IL1R1 , IL1R2 , JUN , JUND , KRAS , LAMTOR3 , MAP2K1 , MAP2K2 , MAP2K3 , MAP2K4 , MAP2K5 , MAP2K6 , MAP2K7 , MAP3K1 , MAP3K11 , MAP3K12 , MAP3K13 , MAP3K2 , MAP3K3 , MAP3K4 , MAP3K5 , MAP3K6 , MAP3K8 , MAP4K1 , MAP4K2 , MAP4K3 , MAP4K4 , MAPK1 , MAPK11 , MAPK12 , MAPK13 , MAPK14 , MAPK3 , MAPK7 , MAPK8 , MAPK8IP2 , MAPK8IP3 , MAPK9 , MAPKAPK3 , MAPKAPK5 , MAPT , MAX , MECOM , MEF2C , MKNK1 , MKNK2 , MRAS , MYC , NF1 , NFATC3 , NFKB1 , NFKB2 , NGF , NLK , NR4A1 , NTRK2 , PAK1 , PAK2 , PDGFA , PDGFB , PDGFRA , PLA2G4B , PLA2G4C , PPM1A , PPM1B , PPP3CA , PPP3CB , PPP3CC , PPP3R1 , PPP5C , PPP5D1 , PRKACA , PRKACB , PRKACG , PRKCA , RAC1 , RAC2 , RAC3 , RAF1 , RAP1A , RAP1B , RAPGEF2 , RASA1 , RASA2 , RASGRF1 , RASGRP1 , RASGRP3 , RASGRP4 , RELA , RELB , RPS6KA1 , RPS6KA2 , RPS6KA3 , RPS6KA4 , RPS6KA5 , RRAS , RRAS2 , SOS1 , SOS2 , SRF , STK3 , STK4 , STMN1 , TAB1 , TAB2 , TAOK1 , TAOK2 , TAOK3 , TGFB1 , TGFB2 , TGFB3 , TGFBR1 , TGFBR2 , TNFRSF1A , TRAF2 , TRAF6 , AKT3 , BDNF , CACNB2 , CACNB4 , CACNG3 , CD14 , EGF , FGF14 , FGF2 , FGF20 , FGF5 , FGF8 , GNG12 , MOS , NFATC1 , PDGFRB , PLA2G4A , PRKCB , RPS6KA6

HALLMARKS OF CANCER	SIGNALING PATHWAYS ASSOCIATED	NRF1 TARGET GENES (Identified by ChIP-seq using HCC1954-breast cancer cells and HMEC-normal mammary epithelial cells)
Sustaining proliferative signaling	mTOR Pathway	ATP6V1G2 , EIF4EBP1 , FZD10 , PRKCG , WNT1 , AKT1 , AKT1S1 , AKT2 , ATP6V1A , ATP6V1B2 , ATP6V1C1 , ATP6V1C2 , ATP6V1D , ATP6V1E1 , ATP6V1F , ATP6V1G1 , ATP6V1G3 , ATP6V1H , CAB39 , CHUK , CLIP1 , DDIT4 , DEPDC5 , DEPTOR , DVL1 , DVL2 , DVL3 , EIF4B , EIF4E , EIF4E2 , FLCN , FNIP1 , FNIP2 , FZD1 , FZD2 , FZD3 , FZD4 , FZD5 , FZD6 , FZD8 , FZD9 , GRB10 , GRB2 , GSK3B , HRAS , IGF1R , IKBKB , INSR , IRS1 , KRAS , LAMTOR3 , LAMTOR4 , LPIN1 , LRP5 , LRP6 , MAP2K1 , MAP2K2 , MAPK1 , MAPK3 , MAPKAP1 , MIOS , MLST8 , NPRL2 , NPRL3 , PDPK1 , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , PRKAA1 , PRKCA , PRR5 , PTEN , RAF1 , RHEB , RHOA , RICTOR , RNF152 , RPS6KA1 , RPS6KA2 , RPS6KA3 , RPS6KB1 , RPS6KB2 , RPTOR , RRAGA , RRAGC , RRAGD , SEH1L , SESN2 , SGK1 , SKP2 , SLC3A2 , SLC7A5 , SOS1 , SOS2 , STK11 , STRADA , STRADB , TBC1D7 , TELO2 , TNFRSF1A , TSC1 , TSC2 , TTI1 , ULK1 , ULK2 , WDR59 , WNT10A , WNT10B , WNT11 , WNT3A , WNT4 , WNT5A , WNT7A , WNT7B , WNT8B , WNT9A , AKT3 , ATP6V1B1 , PRKAA2 , PRKCB , RPS6KA6 , WNT2 , WNT2B , WNT3 , WNT5B
Sustaining proliferative signaling	Cellular Senescence	CALML6, CAPN1, E2F4, EIF4EBP1, MAPKAPK2, NFATC4, PPP1CA, RAD9A, RBL1, TRAF3IP2, AKT1, AKT2, ATM, ATR, BTRC, CACNA1D, CALM1, CALM2, CALM3, CALML3, CALML5, CAPN2, CCNA1, CCNA2, CCNB1, CCNB2, CCND1, CCND2, CCND3, CCNE1, CCNE2, CDK1, CDK4, CDK6, CDKN1A, CDKN2A, CDKN2B, CHEK1, CHEK2, E2F1, E2F2, E2F3, ETS1, FBXW11, FOXO1, FOXO3, GADD45A, GADD45B, GADD45G, GATA4, HIPK2, HIPK3, HIPK4, HRAS, HUS1, IGFBP3, IL6, ITPR1, ITPR3, KRAS, LIN37, LIN52, LIN54, LIN9, MAP2K1, MAP2K2, MAP2K3, MAP2K6, MAPK1, MAPK11, MAPK12, MAPK13, MAPK14, MAPK3, MCU, MDM2, MRAS, MYBL2, MYC, NBN, NFATC2, NFATC3, NFKB1, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3, PPP1CB, PPP1CC, PPP3CA, PPP3CB, PPP3CC, PPP3R1, PTEN, RAD50, RAF1, RASSF5, RB1, RBBP4, RBL2, RELA, RHEB, RRAS, RRAS2, SIRT1, SLC25A4, SLC25A5, SLC25A6, SMAD2, SMAD3, SQSTM1, TGFB1, TGFB2, TGFB3, TGFBR1, TGFBR2, TRPM7, TRPV4, TSC1, TSC2, VDAC1, VDAC2, VDAC3, ZFP36L1, ZFP36L2, AKT3, E2F5, HLA-A, ITPR2, NFATC1

HALLMARKS OF CANCER	SIGNALING PATHWAYS ASSOCIATED	NRF1 TARGET GENES (Identified by ChIP-seq using HCC1954-breast cancer cells and HMEC-normal mammary epithelial cells)
Evading growth suppressors	p53 Signaling	BAX , CCNG1 , MDM4 , APAF1 , ATM , ATR , BBC3 , BID , CASP3 , CASP9 , CCNB1 , CCNB2 , CCND1 , CCND2 , CCND3 , CCNE1 , CCNE2 , CCNG2 , CDK1 , CDK4 , CDK6 , CDKN1A , CDKN2A , CHEK1 , CHEK2 , CYCS , DDB2 , EI24 , FAS , GADD45A , GADD45B , GADD45G , GTSE1 , IGFBP3 , MDM2 , PERP , PPM1D , PTEN , RCHY1 , RFWD2 , RRM2 , RRM2B , SERPINB5 , SESN1 , SESN2 , SESN3 , SFN , SHISA5 , SIAH1 , STEAP3 , THBS1 , TNFRSF10B , TP73 , TSC2 , ZMAT3 , CD82 , TP53I3
Evading growth suppressors	Apoptosis	BAX , BIRC5 , CAPN1 , CASP7 , CTSH , MAPK10 , TNFSF10 , TUBA3D , TUBA3E , ACTB , ACTG1 , AIFM1 , AKT1 , AKT2 , APAF1 , ATF4 , ATM , BAD , BAK1 , BBC3 , BCL2 , BCL2L1 , BCL2L11 , BID , BIRC2 , BIRC3 , CAPN2 , CASP10 , CASP2 , CASP3 , CASP9 , CFLAR , CHUK , CSF2RB , CTSB , CTSC , CTSD , CTSF , CTSL , CTSV , CYCS , DAB2IP , DAXX , DDIT3 , DFFB , DIABLO , EIF2AK3 , EIF2S1 , ENDOG , ERN1 , FADD , FAS , FOS , GADD45A , GADD45B , GADD45G , HRAS , HRK , HTRA2 , IKBKB , IL3RA , ITPR1 , ITPR3 , JUN , KRAS , LMNA , LMNB1 , LMNB2 , MAP2K1 , MAP2K2 , MAP3K5 , MAPK1 , MAPK3 , MAPK8 , MAPK9 , MCL1 , NFKB1 , NFKBIA , NGF , PARP1 , PARP2 , PARP3 , PARP4 , PDPK1 , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , PTPN13 , RAF1 , RELA , RIPK1 , SEPT4 , SPTAN1 , TNFRSF10A , TNFRSF10B , TNFRSF10D , TNFRSF1A , TRADD , TRAF1 , TRAF2 , TUBA1B , TUBA1C , TUBA3C , TUBA4A , XIAP , AKT3 , BCL2A1 , ITPR2 , PRF1 , TNFRSF10C , TUBA1A , TUBA8
Evading growth suppressors	TGF-beta Signaling	ACVR1, ACVR1B, ACVR1C, ACVR2A, ACVR2B, BAMBI, BMP2, BMP4, BMP6, BMP7, BMP8A, BMP8B, BMPR1A, BMPR1B, BMPR2, CDKN2B, CHRD, CREBBP, CUL1, DCN, E2F4, E2F5, EP300, FST, GDF5, GDF6, ID1, ID2, ID3, ID4, INHBA, INHBB, INHBC, LEFTY1, LTBP1, MAPK1, MAPK3, MINOS1-NBL1, MYC, NBL1, NOG, PITX2, PPP2CA, PPP2R1A, PPP2R1B, RBL1, RBX1, RHOA, ROCK1, RPS6KB1, RPS6KB2, SKP1, SMAD1, SMAD2, SMAD3, SMAD4, SMAD5, SMAD6, SMAD7, SMURF1, SMURF2, SP1, TFDP1, TGFB1, TGFB2, TGFB3, TGFBR1, TGFBR2, TGIF1, THBS1, ZFYVE16, ZFYVE9

HALLMARKS OF CANCER	SIGNALING PATHWAYS ASSOCIATED	NRF1 TARGET GENES (Identified by ChIP-seq using HCC1954-breast cancer cells and HMEC-normal mammary epithelial cells)
Resisting cell death	p53 Signaling	BAX , CCNG1 , MDM4 , APAF1 , ATM , ATR , BBC3 , BID , CASP3 , CASP9 , CCNB1 , CCNB2 , CCND1 , CCND2 , CCND3 , CCNE1 , CCNE2 , CCNG2 , CDK1 , CDK4 , CDK6 , CDKN1A , CDKN2A , CHEK1 , CHEK2 , CYCS , DDB2 , EI24 , FAS , GADD45A , GADD45B , GADD45G , GTSE1 , IGFBP3 , MDM2 , PERP , PPM1D , PTEN , RCHY1 , RFWD2 , RRM2 , RRM2B , SERPINB5 , SESN1 , SESN2 , SESN3 , SFN , SHISA5 , SIAH1 , STEAP3 , THBS1 , TNFRSF10B , TP73 , TSC2 , ZMAT3 , CD82 , TP53I3
Resisting cell death	Apoptosis	BAX , BIRC5 , CAPN1 , CASP7 , CTSH , MAPK10 , TNFSF10 , TUBA3D , TUBA3E , ACTB , ACTG1 , AIFM1 , AKT1 , AKT2 , APAF1 , ATF4 , ATM , BAD , BAK1 , BBC3 , BCL2 , BCL2L1 , BCL2L11 , BID , BIRC2 , BIRC3 , CAPN2 , CASP10 , CASP2 , CASP3 , CASP9 , CFLAR , CHUK , CSF2RB , CTSB , CTSC , CTSD , CTSF , CTSL , CTSV , CYCS , DAB2IP , DAXX , DDIT3 , DFFB , DIABLO , EIF2AK3 , EIF2S1 , ENDOG , ERN1 , FADD , FAS , FOS , GADD45A , GADD45B , GADD45G , HRAS , HRK , HTRA2 , IKBKB , IL3RA , ITPR1 , ITPR3 , JUN , KRAS , LMNA , LMNB1 , LMNB2 , MAP2K1 , MAP2K2 , MAP3K5 , MAPK1 , MAPK3 , MAPK8 , MAPK9 , MCL1 , NFKB1 , NFKBIA , NGF , PARP1 , PARP2 , PARP3 , PARP4 , PDPK1 , PIK3CA , PIK3CB , PIK3CD , PIK3R1 , PIK3R2 , PIK3R3 , PTPN13 , RAF1 , RELA , RIPK1 , SEPT4 , SPTAN1 , TNFRSF10A , TNFRSF10B , TNFRSF10D , TNFRSF1A , TRADD , TRAF1 , TRAF2 , TUBA1B , TUBA1C , TUBA3C , TUBA4A , XIAP , AKT3 , BCL2A1 , ITPR2 , PRF1 , TNFRSF10C , TUBA1A , TUBA8
Resisting cell death	Autophagy	AKT1, AKT1S1, AKT2, AKT3, ATG10, ATG12, ATG13, ATG14, ATG16L1, ATG16L2, ATG2A, ATG2B, ATG3, ATG4B, ATG4C, ATG4D, ATG5, ATG7, ATG9A, ATG9B, BAD, BCL2, BCL2L1, BECN1, BNIP3, CAMKK2, CFLAR, CTSB, CTSD, CTSL, DAPK1, DAPK3, DDIT4, DEPTOR, EIF2AK3, EIF2AK4, EIF2S1, ERN1, GABARAP, GABARAPL1, HIF1A, HMGB1, HRAS, IGF1R, IRS1, IRS2, IRS4, ITPR1, KRAS, LAMP1, MAP2K1, MAP2K2, MAPK1, MAPK10, MAPK3, MAPK8, MAPK9, MLST8, MRAS, MTMR14, MTMR3, MTMR4, NRBF2, PDPK1, PIK3C3, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3, PPP2CA, PRKAA1, PRKAA2, PRKACA, PRKACB, PRKACG, PRKCD, PRKCQ, PTEN, RAB33B, RAB7A, RAF1, RB1CC1, RHEB, RPS6KB1, RPS6KB2, RPTOR, RRAGA, RRAGC, RRAGD, RRAS, RRAS2, SH3GLB1, SNAP29,

HALLMARKS OF CANCER	**SIGNALING PATHWAYS ASSOCIATED**	**NRF1 TARGET GENES (Identified by ChIP-seq using HCC1954-breast cancer cells and HMEC-normal mammary epithelial cells)**
		STK11, STX17, SUPT20H, TRAF6, TSC1, TSC2, ULK1, ULK2, UVRAG, VAMP8, WIPI1, WIPI2, ZFYVE1
Enabling replicative immortality	p53 Signaling	BAX , CCNG1 , MDM4 , APAF1 , ATM , ATR , BBC3 , BID , CASP3 , CASP9 , CCNB1 , CCNB2 , CCND1 , CCND2 , CCND3 , CCNE1 , CCNE2 , CCNG2 , CDK1 , CDK4 , CDK6 , CDKN1A , CDKN2A , CHEK1 , CHEK2 , CYCS , DDB2 , EI24 , FAS , GADD45A , GADD45B , GADD45G , GTSE1 , IGFBP3 , MDM2 , PERP , PPM1D , PTEN , RCHY1 , RFWD2 , RRM2 , RRM2B , SERPINB5 , SESN1 , SESN2 , SESN3 , SFN , SHISA5 , SIAH1 , STEAP3 , THBS1 , TNFRSF10B , TP73 , TSC2 , ZMAT3 , CD82 , TP53I3
Inducing Angiogenesis	VEGF Signaling	AKT1, AKT2, AKT3, BAD, CASP9, CDC42, HRAS, HSPB1, KRAS, MAP2K1, MAP2K2, MAPK1, MAPK11, MAPK12, MAPK13, MAPK14, MAPK3, MAPKAPK2, MAPKAPK3, NFATC2, NOS3, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3, PLA2G4A, PLA2G4B, PLA2G4C, PLA2G4D, PLA2G4E, PLCG1, PLCG2, PPP3CA, PPP3CB, PPP3CC, PPP3R1, PRKCA, PRKCB, PRKCG, PTGS2, PTK2, PXN, RAC1, RAC2, RAC3, RAF1, SHC2, SPHK1, SPHK2, SRC, VEGFA
Activating invasion and metastasis	ECM-receptor interaction	AGRN, CD44, CD47, COL1A1, COL2A1, COL4A1, COL4A2, COL4A3, COL4A4, COL6A3, COL6A5, COL9A1, COL9A2, COL9A3, COMP, DAG1, FN1, GP1BA, GP1BB, GP5, GP9, HMMR, HSPG2, IBSP, ITGA10, ITGA11, ITGA2, ITGA2B, ITGA3, ITGA4, ITGA6, ITGA9, ITGAV, ITGB1, ITGB3, ITGB4, ITGB5, ITGB6, ITGB7, ITGB8, LAMA1, LAMA2, LAMA3, LAMA5, LAMB1, LAMB2, LAMB3, LAMC1, LAMC2, LAMC3, SDC1, SDC4, SV2A, SV2B, SV2C, THBS1, THBS2, THBS3, TNN, TNR, VTN
Activating invasion and metastasis	Cell adhesion molecules (CAMs)	CADM1, CD2, CD226, CD274, CD276, CD28, CD34, CD4, CD40LG, CD58, CD6, CD8A, CD8B, CD99, CDH1, CDH15, CDH2, CDH3, CDH4, CDH5, CLDN14, CLDN15, CLDN17, CLDN19, CLDN22, CLDN23, CLDN3, CLDN4, CLDN5, CLDN6, CLDN7, CLDN9, CNTN1, CNTNAP1, ESAM, F11R, GLG1, HLA-A, HLA-DMB, HLA-DOA, ICAM1, ICAM2, ICOSLG, ITGA4, ITGA6, ITGA9, ITGAM, ITGAV, ITGB1, ITGB2, ITGB7, ITGB8, JAM3, L1CAM, LRRC4, LRRC4B, MADCAM1,

HALLMARKS OF CANCER	SIGNALING PATHWAYS ASSOCIATED	NRF1 TARGET GENES (Identified by ChIP-seq using HCC1954-breast cancer cells and HMEC-normal mammary epithelial cells)
		MPZL1, NCAM1, NEGR1, NEO1, NFASC, NLGN2, NRCAM, NRXN2, NRXN3, NTNG1, NTNG2, OCLN, PDCD1, PDCD1LG2, PTPRC, PTPRF, PTPRM, PVR, SDC1, SDC2, SDC3, SDC4, SELPLG, SPN, VCAM1, VCAN, VTCN1
Evading immune destruction	T cell receptor signaling pathway	AKT1, AKT2, AKT3, BCL10, CARD11, CBL, CBLB, CBLC, CD247, CD28, CD3D, CD3E, CD4, CD40LG, CD8A, CD8B, CDC42, CDK4, CHUK, DLG1, FOS, FYN, GRAP2, GRB2, GSK3B, HRAS, IKBKB, IL4, IL5, JUN, KRAS, LAT, LCK, MALT1, MAP2K1, MAP2K2, MAP2K7, MAP3K8, MAPK1, MAPK11, MAPK12, MAPK13, MAPK14, MAPK3, MAPK9, NCK1, NCK2, NFATC1, NFATC2, NFATC3, NFKB1, NFKBIA, NFKBIB, NFKBIE, PAK1, PAK2, PAK3, PAK4, PAK6, PDCD1, PDPK1, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3, PLCG1, PPP3CA, PPP3CB, PPP3CC, PPP3R1, PRKCQ, PTPN6, PTPRC, RAF1, RASGRP1, RELA, RHOA, SOS1, SOS2, TEC, VAV1, VAV2, VAV3, ZAP70
Evading immune destruction	B cell receptor signaling pathway	AKT1, AKT2, AKT3, BCL10, BLNK, CARD11, CD72, CD81, CHUK, CR2, FOS, GRB2, GSK3B, HRAS, IFITM1, IKBKB, INPPL1, JUN, KRAS, LYN, MALT1, MAP2K1, MAP2K2, MAPK1, MAPK3, NFATC1, NFATC2, NFATC3, NFKB1, NFKBIA, NFKBIB, NFKBIE, PIK3AP1, PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R2, PIK3R3, PLCG2, PPP3CA, PPP3CB, PPP3CC, PPP3R1, PRKCB, PTPN6, RAC1, RAC2, RAC3, RAF1, RASGRP3, RELA, SOS1, SOS2, SYK, VAV1, VAV2, VAV3

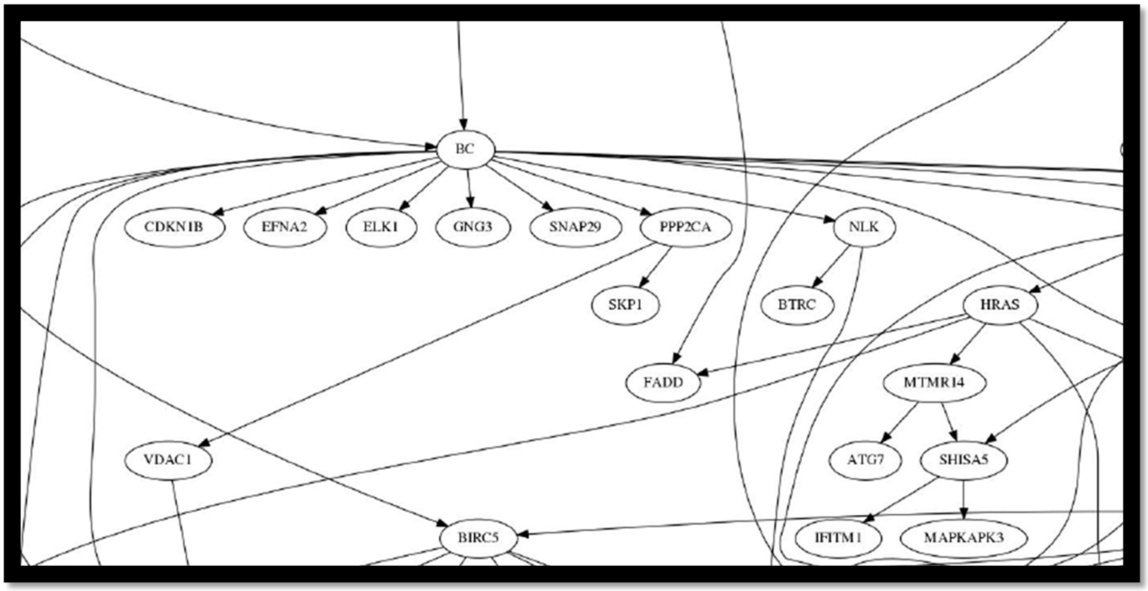


Figure 1. Partial view of the best Bayesian network for the Non-Hispanic White Luminal A cluster generated by the software Banjo. Node BC represents the variable Luminal A breast cancer and the other nodes represent the genes.

Once the best networks were selected, the causal hypothesis (Markov blanket) genes were identified. Markov blanket genes of the variable of interest (BC) is the minimal set of genes conditioned on which all the other genes in the network are independent (probabilistically speaking) of the variable of interest. Figure 2 shows the localization in the network of the Markov blanket genes for Non-Hispanic Luminal A (highlighted in blue), and Table 4 lists the results for all eight clusters.

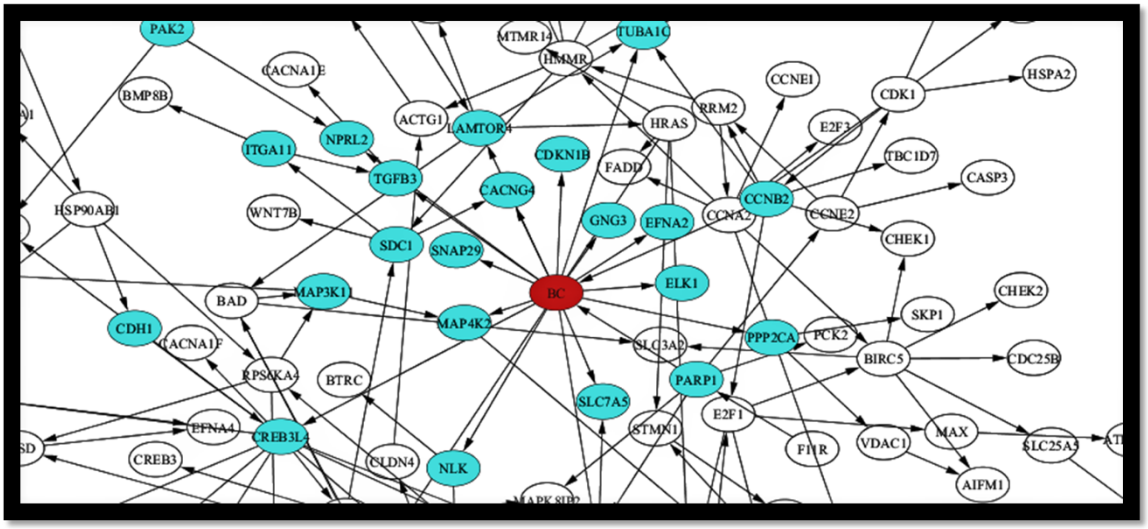


Figure 2. Bayesian network for Non-Hispanic White Luminal A cluster recreated using Cytoscape to provide a better view of the causal hypothesis genes highlighted in blue. The set of selected genes (highlighted in blue) form a substructure around the node of interest (BC) that makes all the other variables probabilistically independent of the disease. This narrows down the search for the drivers of Luminal A breast cancer among Non-Hispanic Whites to this set of genes.

The strategy followed for searching the best networks included running Banjo three times during 8 hours for each group. Genes highlighted in yellow and blue in appeared at least twice when the three network outcomes were compared, which suggest a possible involvement of these genes in the development of the disease.

Table 4

Causal Hypothesis Genes of Breast Cancer from Bayesian Networks Analysis

Luminal A				Triple-negative		HER2+	LUMINAL B
Non-Hispanic	Non-Hispanic	Non-Hispanic	Hispanic	Non-Hispanic	Non-Hispanic	Non-Hispanic	Non-Hispanic
White	Black	Asian	White	White	Black	White	White
ATP6V1G1	ACVR1B	AIFM1	BMP8A	BAK1	ATG3	ACTB	ACTG1
BAK1	ATP6V1D	CACNA1E	CACNG4	BMP8A	ATG4D	AKT1S1	BAK1
CACNG4	BAX	CACNG4	CLDN3	CADM1	ATP6V1C2	ATG16L1	BAX
CCNB2	BECN1	CBLC	ELK1	CASP2	BAK1	ATP6V1A	BMP8A
CDC37	BMP8A	CDH15	LAMTOR4	CBLC	BIRC5	BMP8A	BMP8B
CDH1	CBLC	CDK1	MAPK8IP2	CCNB1	CCNB1	CDK1	CACNA1E
CDKN1B	CCNB1	EFNA2	MAPKAPK5	CD44	COL9A3	CDK4	CTSD
CREB3L4	CCNB2	ENDOG	PPP2CA	CLDN4	DVL2	CLDN6	EIF4E
EFNA2	CDH15	NPRL3	RRM2	CLDN7	ECSIT	E2F1	FADD
ELK1	DVL3	RBX1		DAXX	GNG8	EI24	GSK3B
GNG3	F11R	STMN1		EFNA1	HSP90AB1	GTSE1	HRAS
GSK3B	GNG3	TUBA1C		EFNA2	HSPA8	HSP90AB1	ITGA11
ITGA11	HSPA2	VDAC3		ELK1	ICAM1	MAP4K2	MAPK9
LAMTOR4	HSPA6			GNB1	LMNB2	MAPK13	PARP1
MAP3K11	MAPK8IP2			HSPA8	MAP2K2	MYBL2	PIK3R3
MAP4K2	MAPK9			IFNAR2	NFKB2	PAK4	PPP2R1A
NLK	NLK			INPPL1	PARP1	SNAP29	PPP5C
NPRL2	PIK3R3			LMNB2	RHEB	TBC1D7	RAC3
PAK2	PPP1CA			PPP1CA	TELO2	VDAC1	RBL1
PARP1	PPP5C			PPP2R1A	YWHAZ		RRM2
PPP2CA	PRLR			PPP2R5D			SHISA5
SDC1	RPS6KB2			PTK2			SLC7A5
SLC7A5				RELB			SNAP29
SNAP29				RRM2			SPHK2
TGFB3				STK4			STK3
TUBA1C				STMN1			TRAF2
				TELO2			TUBA1C
				VAMP8			VAV2
				VAV2			VDAC3
				VDAC1			YWHAH
				YWHAZ			

Note. Genes in red were common in all three Bayesian networks outcomes (Banjo software).

Note. Genes in blue were common in all three Bayesian networks outcomes (Banjo software).

Parametric Learning and Validation of Proposed Bayesian Networks Verify They Are Good Prediction Models of Breast Cancer With Accuracy Levels Above 96%

Having the structure is the first of two tasks for Bayesian network learning. Once the structures were obtained, the second step was parametric learning or estimation of conditional probabilities (Fuster-Parra et al., 2016). Parameters were obtained by recreating the substructure of the Markov blanket genes around the breast cancer node with the software GeNIe Modeler, a tool for modeling and learning with Bayesian network developed by BayesFusion. GeNIe generates a conditional distribution probability table for each node as well as the joint probability.

Figure 3 shows the structure of causal hypothesis genes in Luminal A breast cancer for the group of patients classified as Non-Hispanic Whites with the corresponding joint probability for each node. All eight proposed BNs were validated using a 10-fold crossvalidation method (see Methods section for details).

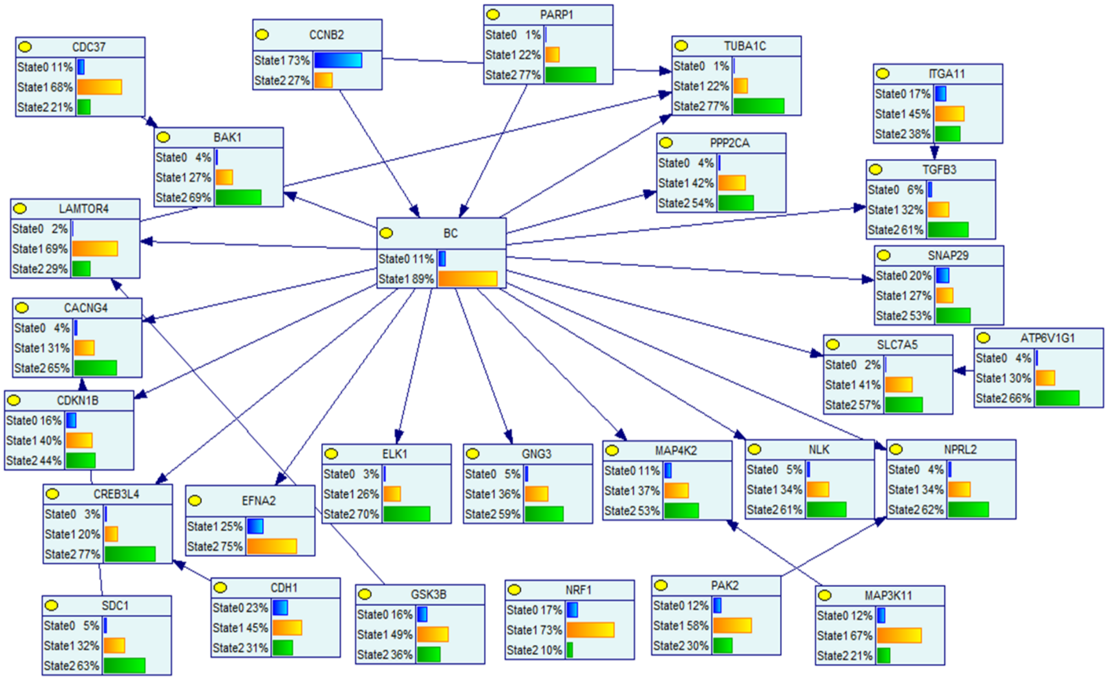


Figure 3. Bayesian network (BN) of causal hypothesis genes with the corresponding initial joint probabilities estimated from the dataset of Non-Hispanic White-Luminal A breast cancer (*n* = 373) and normal samples (*n* = 79). Gene expression was categorized as follows: 0 = down, 1 = Normal and 2 = upregulated. BC was categorized 1 for breast cancer samples and 0 for normal. The model shows a high probability of Luminal A breast cancer (89%), given the initial evidence for the current expression levels of all genes (example: high probability - 77% - of PARP1 to be up regulated). GeNIe allows sensitivity analysis by changing the expression levels (evidence) of one or several genes and the software recalculates the estimated probability of BC.

Table 5 is an example (partial view) of model validation results in the group of Non-Hispanic White Luminal A. Overall results for this model showed 98% accuracy in predicting BC status (443 out of 452 samples).

Table 5

Validation of the Bn Mode

SAMPLE	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**
BC STATUS	State1	State1	State1	State1	State1
Probability of BC_State0	6.20E-07	1.28E-11	5.07E-11	6.55E-10	2.57E-12
Probability of BC_State1	0.999999	1	1	1	1
BC_predicted STATUS	State1	State1	State1	State1	State1
Prediction was correct?	YES	YES	YES	YES	YES
GENES	**EXPRESSION LEVELS (0=DOWN, 1=NORMAL, 2=UP)**				
ATP6V1G1	1	2	2	1	2
BAK1	2	2	2	2	2
LAMTOR4	1	2	2	1	1
CACNG4	2	2	2	2	2
CCNB2	2	2	2	2	2
CDC37	1	2	2	1	1
CDH1	1	0	1	2	2
CDKN1B	1	2	2	0	1
CREB3L4	2	2	2	2	2
EFNA2	1	1	1	1	1
ELK1	2	2	2	2	2
GNG3	1	1	2	1	2
GSK3B	2	0	1	2	1
ITGA11	2	0	2	1	2
MAP3K11	1	1	2	1	1
MAP4K2	1	2	2	1	2
NLK	1	2	2	2	2
NPRL2	1	2	2	2	2
PAK2	2	1	1	2	1
PARP1	2	2	2	2	2
PPP2CA	2	2	2	2	2
SDC1	2	1	2	1	2
SLC7A5	1	2	1	1	2
SNAP29	2	2	0	2	2
TGFB3	2	2	2	2	2
TUBA1C	2	1	2	2	2

Note. Partial view of the output file during validation process (5 samples) of the model generated by Banjo and GeNIe for the group of Non-Hispanic White Luminal A. Accuracy of predicting BC status = 98.01%.

Results of crossvalidation for all BN models are shown in Table 6, including specificity, accuracy, and the area under the receiver operating characteristic (ROC) curve (AUC). Notice how sensitivity, the percentage of correctly predicted samples positive for breast cancer, is greater than 95% for all the models and accuracy is always above 96%. The area under receiver operating characteristic curve (AUC), a metric that combines sensitivity with false positive rate (FPR), is always above 0.99, demonstrating these are very good prediction models of BC.

Table 6

Results of Crossvalidation for all Eight Bayesian Network Models

Cluster	**Description**	**Number of samples**	**Correctly predicted**	**Percentage**	**ROC (AUC)**	**Accuracy of the model**
Non-Hispanic White Luminal A	**Breast Cancer**	373	369	98.93%	0.996	
	Normal	79	74	93.67%		
	Totals	452	443			98.01%
Non-Hispanic BLack Luminal A	**Breast Cancer**	69	69	100.00%	1	
	Normal	79	77	97.47%		
	Totals	148	146			98.65%
Non-Hispanic Asian Luminal A	**Breast Cancer**	22	22	100.00%	1	
	Normal	79	79	100.00%		
	Totals	101	101			100.00%
Hispanic Asian Luminal A	**Breast Cancer**	21	20	95.24%	0.995	
	Normal	79	76	96.20%		
	Totals	100	96			96.00%
Non-Hispanic White Triple-negative	**Breast Cancer**	74	73	98.65%	0.993	
	Normal	79	79	100.00%		
	Totals	153	152			99.35%
Non-Hispanic Black Triple-negative	**Breast Cancer**	48	47	97.92%	0.999	
	Normal	79	77	97.47%		
	Totals	127	124			97.64%

Non-Hispanic White HER2 +	**Breast Cancer**	48	47	97.92%		
	Normal	79	76	96.20%		
	Totals	127	123		0.998	96.85%
Non-Hispanic White Luminal B	**Breast Cancer**	83	83	100.00%		
	Normal	79	78	98.73%		
	Totals	162	161		0.999	99.38%

Note. Sensitivity, the percentage of the proportion of actual positives for breast cancer that were correctly identified, is shown under the column “Percentage” in the breast cancer row. For example, for Non-Hispanic White the sensitivity of the model was 98.93%. Notice that sensitivity is above 95% for all the models and accuracy is above 96%. The area under receiver operating characteristic curve (AUC), a metric that combines sensitivity with false positive rate (FPR), is above 0.99 in all cases, indicating these are very good prediction models of breast cancer.

Sensitivity Analysis of Bayesian Networks (Bns) Shows Differences in NRF1 Molecular Signature of Possible Disease Drivers That May Explain the Biological Differences in Breast Cancer Outcomes by Race and Ethnicity

Table 4 lists causal hypothesis genes for each cluster under study. In order to further identify possible disease drivers among those genes, we performed sensitivity analysis to discover the ones that had the highest impact on the relative risk (RR) of breast cancer when maximizing or minimizing their expression. BNs are used to estimate new probabilities when new information is incorporated into the model; therefore, the strategy we followed was to use the software GeNIe to simulate upregulation (marginal probability of state 2 = 100%), normalization (state 1) and downregulation (state 0) of candidate genes, estimate the probability of breast cancer [Pr.(BC)] for each case, and calculate the relative risk at upregulation, normalization, and downregulation using the initial marginal probability of the model as baseline. For example, given the evidence g = up (gene is upregulated), we can write the equation for RR as RR= [Pr. (BC/g=up) / Pr. (BC/g=current evidence)].

We focused the sensitivity analysis on those genes that were common in at least two of the three locally optimal networks (those highlighted in yellow and blue in Table 4), which, given the current results, have a higher probability of being part of the globally optimal network. Nevertheless, none of the causal hypothesis genes in the locally optimal BNs should be discarded as potential disease drivers. With this in mind, we also performed sensitivity analysis simulating changes in expression level of more than one gene simultaneously, especially among those identified as affecting significantly breast cancer risk. We also tested the parents of the breast cancer node regardless of their frequency in the three BNs. Here we present the results of the sensitivity analysis for each of the clusters, including comments on the most important findings.

Luminal A in Non-Hispanic Whites

Figure 3 displays the BN with marginal probabilities for all variables in this cluster, and Table 7 shows the results for the sensitivity analysis of Markov blanket genes, ordered by relative risk (RR). Notice in Table 7 how the greatest positive impact on the probability of breast cancer (lowering from 89% to 7%) was obtained through TUBA1C downregulation. TUBA1C encodes the protein Tubulin Alpha 1C, the principal component of microtubules, and has been found upregulated in breast cancer and its overexpression associated with lower overall survival (Chen et al., 2015). PARP1, ELK1 and CREB3L4 individual downregulation also lower considerably the probability of breast cancer; furthermore, downregulation of PARP1 and TUBA1C simultaneously lower Pr. (BC) to 0%.

Table 7

Results of Sensitivity Analysis for Luminal A BN Model in Non-Hispanic White Cluster

Luminal A Non-Hispanic White			
Candidate Gene	**Simulated change in gene expression**	**Probability of breast Cancer (%)**	**Relative Risk (RR)-Initial Pr=89 %**
NLK	UP	99	1.11
ELK1	UP	98	1.10
PARP1	UP	98	1.10
TGFB3	UP	98	1.10
CACNG4	UP	97	1.09
CDKN1B	UP	97	1.09
CREB3L4	UP	97	1.09
TUBA1C	UP	97	1.09
CDKN1B	Down	94	1.06
CACNG4	Down	92	1.03
CDKN1B	Normal	80	0.90
TGFB3	Down	78	0.88
NLK	Normal	77	0.87
TGFB3	Normal	75	0.84
CACNG4	Normal	73	0.82
ELK1	Normal	73	0.82
CREB3L4	Normal	71	0.80
PARP1	Down	68	0.76
NLK	Down	67	0.75
TUBA1C	Normal	67	0.75
PARP1	Normal	62	0.70
ELK1	Down	45	0.51
CREB3L4	Down	31	0.35
TUBA1C	Down	7	0.08

Note. Notice how Relative Risk figures indicate that the biggest impact in reducing Pr. (BC)—initially 89%--is achieved by downregulation of TUBA1C. Other genes with substantial impact in lowering breast cancer risk are PARP1, ELK1, and CREB3L4.

Luminal A in Non-Hispanic Blacks

Table 8 shows the results of the sensitivity analysis. Notice how Relative Risk figures indicate that the biggest impact in reducing Pr. (BC) from 47% to 3% was achieved by switching CCNB1 to 100% normal (initial probability of normal expression level in cluster sample was 44%). Genes BMP8A, CBLC, and MAPK8IP2 appeared in all three local BNs. Since BMP8A and MAPK8IP2 were directly connected to the BC node in the network (Figure 4), we began the analysis with the overexpression of these two genes to 100%, resulting in an increase of the probability of BC [Pr(BC)] from the initial 47% to 93% (Relative Risk = 1.98).

Table 8

Results of Sensitivity Analysis for Luminal A BN model in Non-Hispanic Black Cluster

Luminal A Non-Hispanic Black			
Candidate Gene	**Simulated change in gene expression**	**Probability of breast Cancer (%)**	**Relative Risk (RR)compared to Initial Pr=47 %**
CCNB2	Normal	90	1.91
PPP1CA	UP	89	1.89
NLK	UP	88	1.87
CCNB1	UP	86	1.83
MAPK9	UP	86	1.83
BMP8A	UP	82	1.74
DVL3	UP	80	1.70
ACVR1B	UP	76	1.62
ATP6V1D	UP	76	1.62
BECN1	UP	74	1.57
MAPK8IP2	UP	70	1.49
DVL3	Down	48	1.02
ATP6V1D	Down	47	1.00
ACVR1B	Down	39	0.83

Luminal A Non-Hispanic Black			
Candidate Gene	Simulated change in gene expression	Probability of breast Cancer (%)	Relative Risk (RR)compared to Initial Pr=47 %
NLK	Down	38	0.81
BECN1	Down	34	0.72
ATP6V1D	Normal	31	0.66
MAPK9	Normal	30	0.64
MAPK9	Down	30	0.64
ACVR1B	Normal	23	0.49
CCNB1	Down	23	0.49
NLK	Normal	21	0.45
MAPK8IP2	Down	20	0.43
BECN1	Normal	19	0.40
DVL3	Normal	18	0.38
MAPK8IP2	Normal	7	0.15
BMP8A	Down	6	0.13
PPP1CA	Normal	6	0.13
PPP1CA	Down	5	0.11
CCNB2	UP	5	0.11
BMP8A	Normal	4	0.09
CCNB1	Normal	3	0.06

Note. Notice how Relative Risk figures indicate that the biggest impact in reducing Pr. (BC) from 47% to 3% was achieved by switching CCNB1 to 100% normal (initial probability of normal expression level in cluster sample was 44%). Conversely, upregulation of PPP1CA increased the probability of breast cancer to 89% (RR = 1.89). The same PPP1CA showed a substantial impact in lowering the breast cancer risk when downregulated.

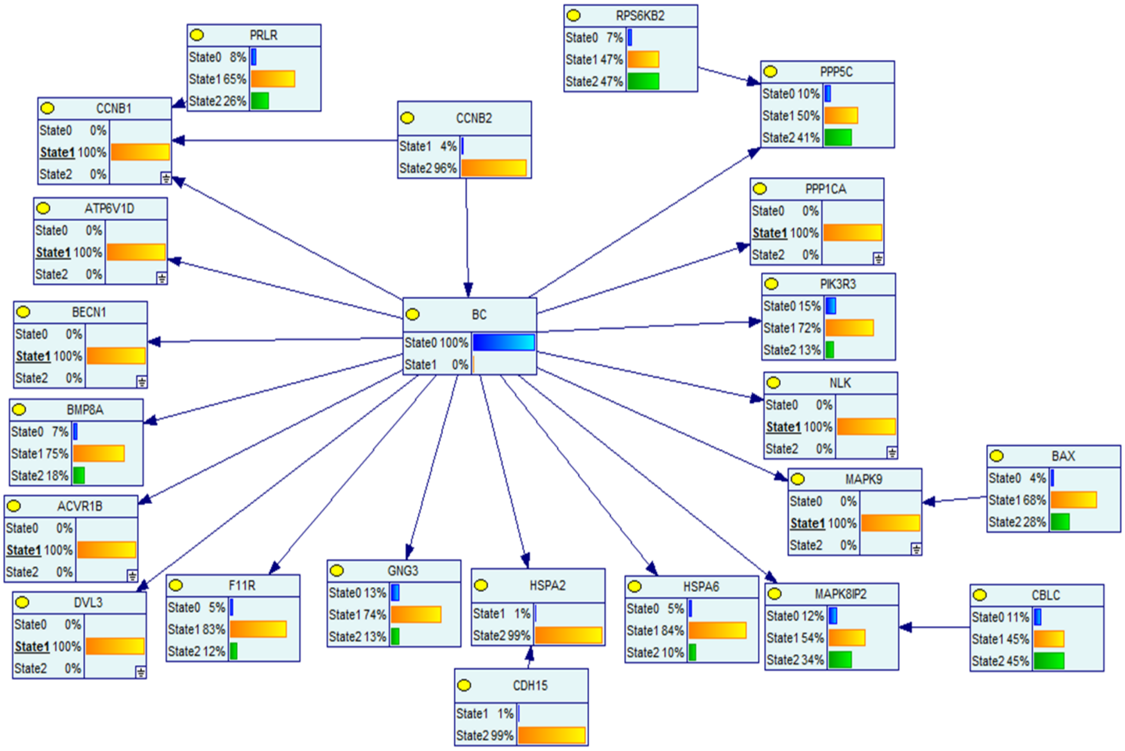


Figure 4. Bayesian network (BN) of causal hypothesis genes learned from the dataset of Non-Hispanic African American-Luminal A breast cancer cluster (*n* = 69) and normal samples (*n* = 79). Gene expression was categorized as follows: 0 = down, 1 = Normal, and 2 = upregulated. BC was categorized 1 for breast cancer samples and 0 for normal. This figure shows how the posterior probability of Luminal A breast cancer changed to 0% from the original 47% (initial evidence) after changing the expression levels for genes ACVR1B, ATP6V1D, BECN1, CCNB1, DVL3, MAPK9, NLK, and PPP1CA to normal. It was further noticed during sensitivity analysis that simply changing to normal, two of them (CCNB1 and PPPC1A) produced the same effect of lowering breast cancer probability to zero.

BMP8A presented the highest individual impact on the Pr (BC) at 100% upregulation. This gene (Bone morphogenetic protein 8A) is part of the Bone morphogenetic protein family involved in the regulation of different cellular processes, such as proliferation, differentiation, apoptosis and migration (Alarmo & Kallioniemi, 2010). We also used the model to demonstrate that when the evidence was changed simultaneously to normal expression for genes ACVR1B, ATP6V1D, BECN1, CCNB1, DVL3, MAPK9, NLK. and PPP1CA (highlighted in blue in Table 4 as common in two out of three BN's), the probability of breast cancer went down to 0%, suggesting that their dysregulation also may be involved in the development and/or progression of Luminal A breast cancer among Non-Hispanic African Americans (Figure 4).

It was also noticed that the genes with the highest initial likelihood of upregulation CCNB1 and PPP1CA (51% and 49%, respectively) were the ones with the highest impact in lowering the breast cancer posterior probability. Sensitivity analysis demonstrated that simply changing to normal CCNB1 (also known as CyclinB1) lowered the probability of breast cancer from 47% to 3%. Setting up the two of them to normal reduced BC probability to 0%, the same effect as setting up to normal all eight above-mentioned genes.

These results are aligned with recent discoveries of CCNB1 overexpression associated with poor distant metastasis free survival, overall survival, and disease-free survival of patients with Estrogen Receptor positive (ER+) breast cancer (Ding, Li, Zou, Zou, & Wang, 2014). Previously, CCNB1 had also been reported as possibly involved in the epithelial-mesenchymal transition (EMT) process (Song et al., 2008). Unlike the Non-Hispanic African American cluster, CCNB1 is not among the causal hypothesis genes of the other Luminal A groups—Non Hispanic Whites, Non-Hispanic Asian, and Hispanic Whites.

It is interesting to notice in this model that when CCNB2 (cyclin B2), direct parent of the BC node, was switched to normal (state1 = 100%), this switching caused upregulation of CCNB1 (state 2 from 51% to 93%). The combined effect resulted in almost doubling the probability of breast cancer to 90% from an initial value of 47% (RR = 1.91), the highest number in the sensitivity analysis table. High levels of cytoplasmic cyclin B2 have been found associated with short-term disease-specific survival in breast cancer patients (Shubbar et al., 2013).

Luminal A in Non-Hispanic Asians

Table 9 lists the results of sensitivity analysis. EFNA2 was the only gene that appeared in all the three local BNs. The simulation of changing the gene expression from a probability of 90% upregulation to 100% normal resulted in an increase in the probability of breast cancer from 18% to 58% (Relative Risk = 3.22), suggesting that EFNA2 (Ephrin A2) exerts a protective effect in Luminal A breast cancer among Non-Hispanic Asians. Conversely, an increase in the probability of upregulation from 90% to 100%, resulted in a decrease in the probability of breast cancer to 14% (RR = 0.77).

We also performed sensitivity analysis on the causal hypothesis genes common to two out of three BNs' outcomes from Banjo: AIFM1, CDK1, TUBA1C, and VDAC3. This analysis revealed that overexpression of CDK1 reduced the probability of breast cancer in this cluster from 18% to2 %--Relative Risk = 0.11 (Figure 5). Cyclin-dependent kinase 1 (CDK1) plays an important role in cell cycle regulation, especially in mitosis during the transition from the G2 to M phase. CDK1 also has several other functions at the molecular level that are not well understood yet (Roberts et al., 2012; Vassilev et al., 2006).

Table 9

Results of Sensitivity Analysis for Luminal A BN Model in Non-Hispanic Asian Cluster

Luminal A Non-Hispanic Asian			
Candidate Gene	**Simulated change in gene expression**	**Probability of breast Cancer (%)**	**Relative Risk (RR)compared to Initial Pr=18 %**
VDAC3	Normal	84	4.67
EFNA2	Normal	58	3.22
CDK1	Normal	53	2.94
NPRL3	UP	33	1.83
TUBA1C	UP	33	1.83
TUBA1C	Down	15	0.83
EFNA2	UP	14	0.78
NPRL3	Normal	13	0.72
TUBA1C	Normal	12	0.67
VDAC3	UP	12	0.67
NPRL3	Down	7	0.39
CDK1	UP	2	0.11

Note. Notice how Relative Risk figures indicate that the biggest impact in reducing Pr. (BC) from 18% to 2% was achieved by switching CDK1 to 100% overexpressed (initial probability of upregulation in cluster sample was 69%).

TUBA1C upregulation increased the probability of BC from 18% to 33% (RR = 1.83). TUBA1C (Tubulin alpha 1c), a component of tubulin, has been reported as significantly highly expressed in breast tumor tissues compared to normal tissue and as a negative predictor of overall survival (Chen, Li, WaNg, & Jiao, 2015). Finally, lowering the expression of VDAC3 (voltage dependent anion channel 3) from 91% upregulation (state 2) to 100% normal (state 1) had the effect of increasing the probability of BC from 18% to 84% (RR = 4.67), suggesting that VDAC3 upregulation exerts a protection effect against BC.

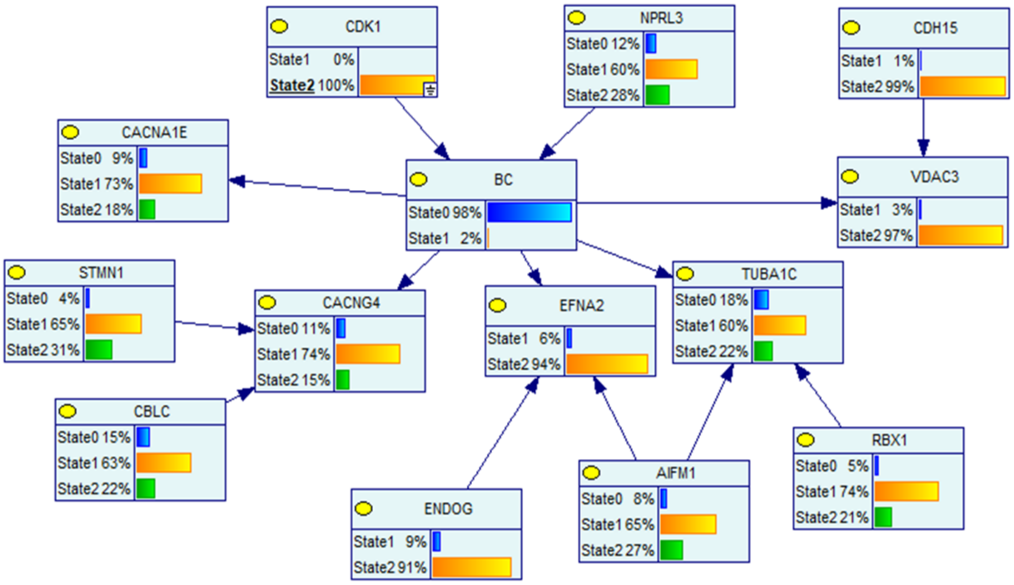


Figure 5. Bayesian network (BN) of causal hypothesis genes learned from the dataset of Non-Hispanic Asian-Luminal A breast cancer cluster (*n* = 22) and normal samples (*n* = 79). Gene expression was categorized as follows: 0 = down, 1 = Normal, and 2 = upregulated. BC was categorized 1 for breast cancer samples and 0 for normal. This figure shows how the posterior probability of Luminal A breast cancer changed to 2% from the original 18% (initial evidence) after changing the expression levels for gene CDK1 to Upregulated.

Luminal A Breast Cancer in Hispanics or Latino Whites

Table 10 shows the results of sensitivity analysis for causal hypothesis genes ranked by relative risk. Sensitivity analysis changing to 100% normal the expression levels of genes BMP8A, CACNG4, and CLDN3 (genes in all three local BNs) reduced the joint probability of breast cancer for the Hispanic or Latino Luminal A cluster from 18% to 0%. The initial marginal probabilities of upregulation for these three genes were 32%, 21%, and 23%, respectively**.** Figure 6 shows the Bayesian network (BN) of causal hypothesis genes learned from the dataset.

Table 10

Results of Sensitivity Analysis for Luminal A BN Model in Hispanic White Cluster

Luminal A Hispanic White			
Candidate Gene	**Simulated change in gene expression**	**Probability of breast Cancer (%)**	**Relative Risk (RR)compared to Initial Pr=18 %**
BMP8A	UP	54	3.00
CACNG4	UP	46	2.56
RRM2	Normal	46	2.56
MAPK8IP2	UP	43	2.39
CACNG4	Down	25	1.39
RRM2	UP	10	0.56
MAPK8IP2	Down	9	0.50
CACNG4	Normal	8	0.44
MAPK8IP2	Normal	6	0.33
BMP8A	Down	5	0.28
BMP8A	Normal	0	0.00

Note. Notice how Relative Risk figures indicate that the biggest impact in reducing Pr. (BC) from 18% to 0% was achieved by switching BMP8A to 100% probability of normal expression in the sample (initial probability of normal expression was 62%). The same gene, BMPA, showed the highest increase in breast cancer risk when upregulated, approximately 3 times the initial marginal risk of 18%.

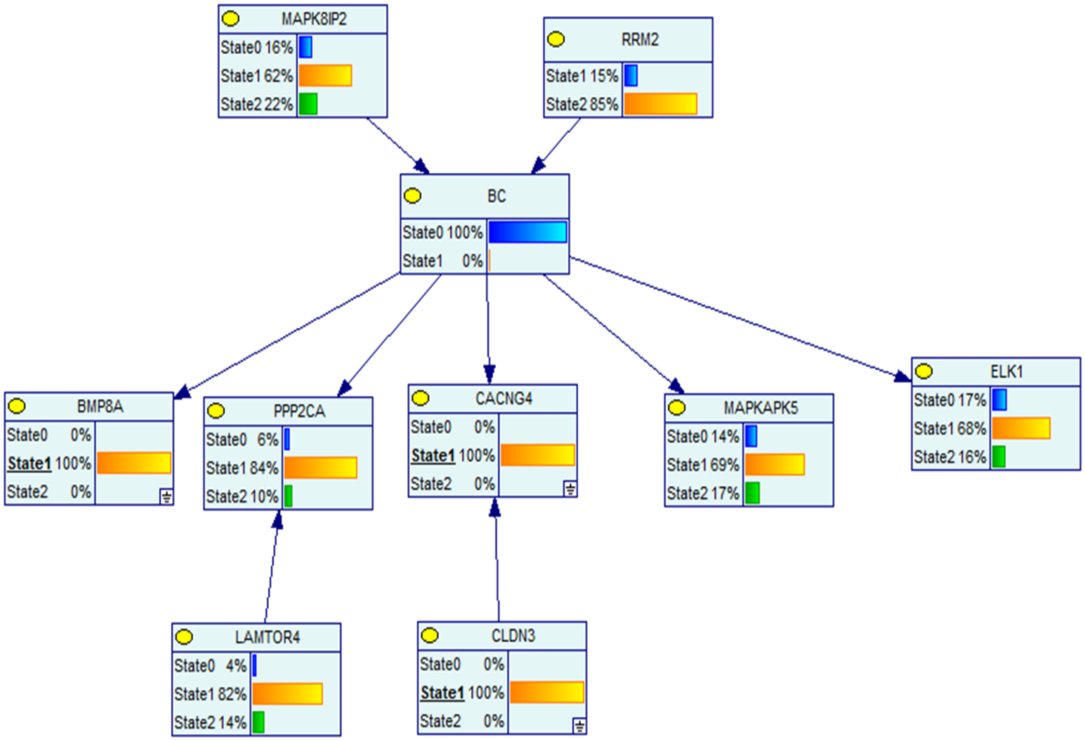


Figure 6. Bayesian network (BN) of causal hypothesis genes learned from the dataset of Hispanic or Latino White-Luminal A breast cancer cluster (*n* = 21) and normal samples (*n* = 79). Gene expression was categorized as follows: 0 = down, 1 = normal, and 2 = upregulated. BC was categorized 1 for breast cancer samples and 0 for normal. This figure shows how the posterior probability of Luminal A breast cancer changed to 0% from the original 18% (initial evidence) after changing to normal the expression levels for genes BMP8A, CACNG4, and CLDN3. These genes were common in all three best local Bayesian networks outcomes generated by Banjo. It was also noticed that simply changing BMP8A that originally had a 32% joint distribution probability of being upregulated to 100% normal yielded the same effect (Pr. BC = 0).

Continuing with the analysis, it was noticed that only setting up the marginal probability of one gene, BMP8A (Bone morphogenetic protein 8A), to normal had the same effect of reducing the probability of breast cancer to 0%. The Bone morphogenetic protein (BMP) family is a group of more than 20 growth factor proteins involved in bone formation and other developmental processes. These extracellular signaling molecules regulate various cellular functions, such as proliferation, differentiation, apoptosis, and migration (Alarmo & Kallioniemi, 2010). In fact, BMP8A is in the list of NRF1 target genes in the TGF-beta signaling pathway linked to the hallmark of cancer Evading Growth Suppressors. Aberrant expression of BMPs and BMP signaling has been reported in breast cancer and bone metastasis. Recent studies also found BMP signaling activity involved in the processes of EMT, angiogenesis, invasion, stemness, and quiescence (Zabkiewicz, Resaul, Hargest, Jiang, & Ye, 2017). However, we did not find any previous study specifically reporting BMP8A involvement in breast cancer.

Triple-Negative Breast Cancer in Non-Hispanic White Cluster

The biggest impact in reducing Pr. (BC) from 49% to 6% (RR = 0.12) was achieved by switching BMP8A to 100% probability of downregulation---the initial probability of downregulation was 4% (Table 11). The initial BN model of TNBC in the Non-Hispanic cluster showed 49% joint probability of developing the disease, given the initial evidence of the gene expression levels of the 31 causal hypothesis genes (Figure 7). This cluster was formed of 153 samples, 74 with TNBC and 79 normal.

Part of sensitivity analysis was to simulate a simultaneous change in the gene expression, to 100% normal of the three genes that had appeared consistently in all three local networks—CASP2 (initial probability of upregulation = 46%), ELK1 (51%), and PPP1CA (48%). This simulation lowered the probability of TNBC from 49% to 3%, suggesting that these three genes may play a role in the disease. When an additional change was added to the simulation, BMP8A (probability of upregulation = 53%) was switched to 100% normal, and the probability of TNBC went down to 0%. A Similar exercise was done but instead of switching to normal, the same four genes were switched to 100% upregulated, resulting as expected an increase in the probability of TNBC to 99%. This procedure confirms the possible role of these genes as drivers of TNBC in Non-Hispanic Whites.

Table 11

Results of Sensitivity Analysis for Triple-Negative BN Model in Non-Hispanic White Cluster

TNBC Non-Hispanic White			
Candidate Gene	**Simulated change in gene expression**	**Probability of breast Cancer (%)**	**Relative Risk (RR)compared to Initial Pr=49 %**
HSPA8	UP	87	1.78
ELK1	UP	85	1.73
CASP2	UP	84	1.71
BMP8A	UP	83	1.69
CCNB1	UP	76	1.55
HSPA8	Down	46	0.94
CCNB1	Down	45	0.92
HSPA8	Normal	25	0.51
CASP2	Normal	20	0.41
CCNB1	Normal	17	0.35
CASP2	Down	15	0.31
ELK1	Down	15	0.31
BMP8A	Normal	10	0.20
ELK1	Normal	9	0.18
BMP8A	Down	6	0.12

Note. Notice how Relative Risk figures indicate that the biggest impact in reducing Pr. (BC) from 49% to 6% (RR = 0.12) was achieved by switching BMP8A to 100% probability of downregulation (initial probability of downregulation was 4%).

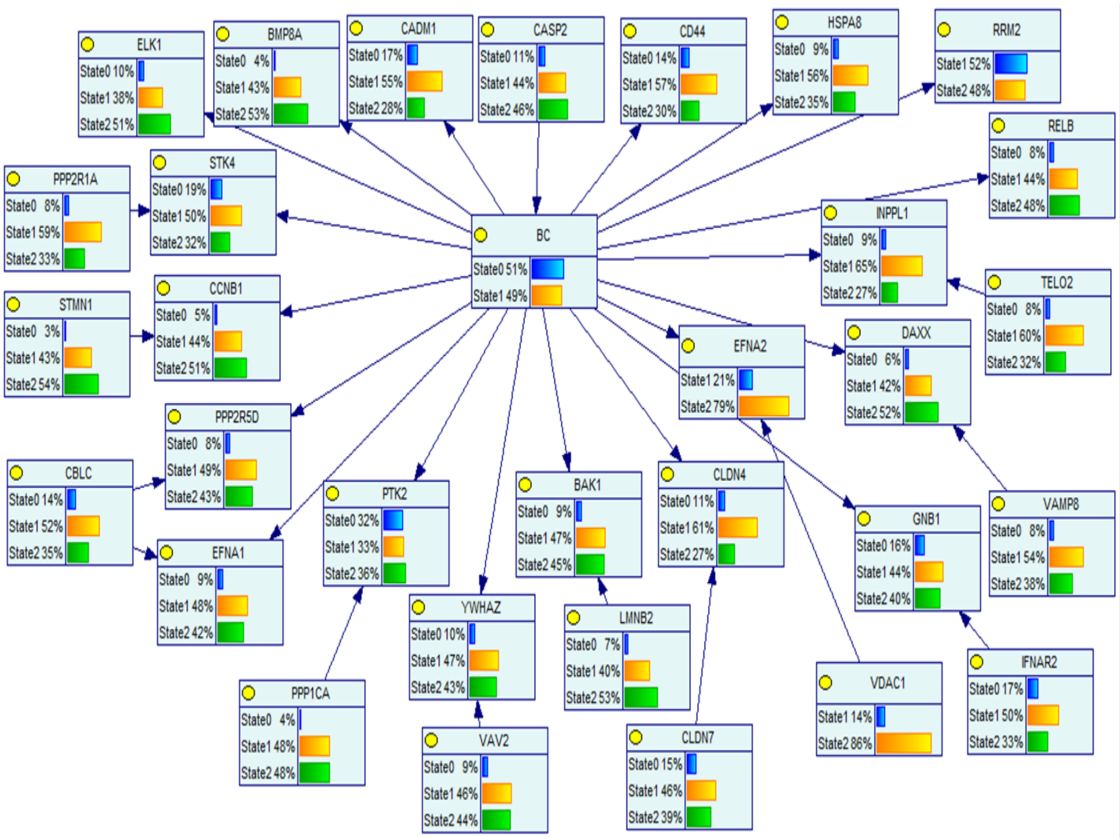


Figure 7. Bayesian network (BN) of causal hypothesis genes learned from the dataset of Non-Hispanic White Triple-Negative breast cancer (TNBC) cluster (*n* = 74) and normal samples (*n* = 79). Gene expression was categorized as follows: 0 = down, 1 = Normal, and 2 = upregulated. BC was categorized 1 for breast cancer samples and 0 for normal. This figure shows how that the joint probability of TNBC for this cluster was 49%.

CASP2 (caspase 2) is part of the group of genes involved in the apoptosis signaling pathway which contributes to the hallmarks of cancer Evading Growth Suppressors and Resisting cell death (Table 3). Apoptosis is known to play a role in tumorigenesis and also contributes to the development of resistance to cancer therapies. CASP2 produces several alternative splicing isoforms that play antagonistic roles, while Casp-2L promotes apoptosis; Casp-2S protects cells against apoptosis (Fushimi et al., 2008). Sensitivity analysis in this model suggests that CASP2 overexpression exerts a protective role for cancer cells in TNBC.

For ELK1, it was reported that higher mRNA expression was associated with worse recurrence-free survival in TNBC patients (Liu et al., 2017). Table 11 also shows that the biggest impact on increasing breast cancer risk was obtained with the upregulation of HSPA8. Previous studies have found HSPA8 upregulated in the early stages of breast cancer (Hou et al., 2016).

Triple-Negative Breast Cancer in Non-Hispanic Black

Table 12 shows that ATP6V1C2 had the highest individual impact on breast cancer risk, increasing the probability 2.85 times. ATP6V1C2 (ATPase, H+ transporting, lysosomal 42kD, V1 subunit C isoform 2) is one of the proteins called V-ATPases reported in the literature as playing a role in breast cancer growth and metastasis (McConnell et al., 2017). Table 12 also shows that switching CCNB1 (parent of BC in the network) to 100% normal presented the highest effect on lowering breast cancer risk for this cluster (RR = 0.15). As mentioned earlier, overexpression of CCNB1 was found associated with poor prognosis for distant metastasis-free survival and overall survival in breast cancer patients with ER + breast cancer (Ding et al., 2014).

Table 12

Results of Sensitivity Analysis for Triple-Negative BN Model in Non-Hispanic Black Cluster

TNBC Non-Hispanic Black			
Candidate Gene	**Simulated change in gene expression**	**Probability of breast Cancer (%)**	**Relative Risk (RR)compared to Initial Pr=26 %**
ATP6V1C2	UP	74	2.85
ATG3	UP	73	2.81
ATG4D	UP	73	2.81
CCNB1	UP	53	2.04
BIRC5	Normal	48	1.85
CCNB1	Down	25	0.96
ATP6V1C2	Down	21	0.81
ATG4D	Down	15	0.58
ATG3	Down	14	0.54
ATG4D	Normal	11	0.42
ATG3	Normal	10	0.38
BIRC5	UP	10	0.38
ATP6V1C2	Normal	7	0.27
CCNB1	Normal	4	0.15

Note. Notice how Relative Risk figures indicate that the biggest impact in increasing Pr. (BC) from 26% to 74% (RR = 2.85) was achieved by switching LLLATP6V1C2 to 100% probability of upregulation (initial probability of upregulation was 26%). CCNB1 has the highest effect in the opposite direction when switched to 100% normal expression.

Unlike the TNBC model for Non-Hispanic Whites, in the BN model for TNBC in Non-Hispanic Blacks, we did not find genes common to all three local BNs generated by Banjo. However, we found six genes common to two of them: ATG3, ATG4D, ATP6V1C2, BIRC5, ECSIT, and LMNB2. Sensitivity analysis was performed by simulating changes in expression levels of these genes to observe the impact on the joint probability of TNBC that was initially estimated at 26% for this cluster (Figure 8).

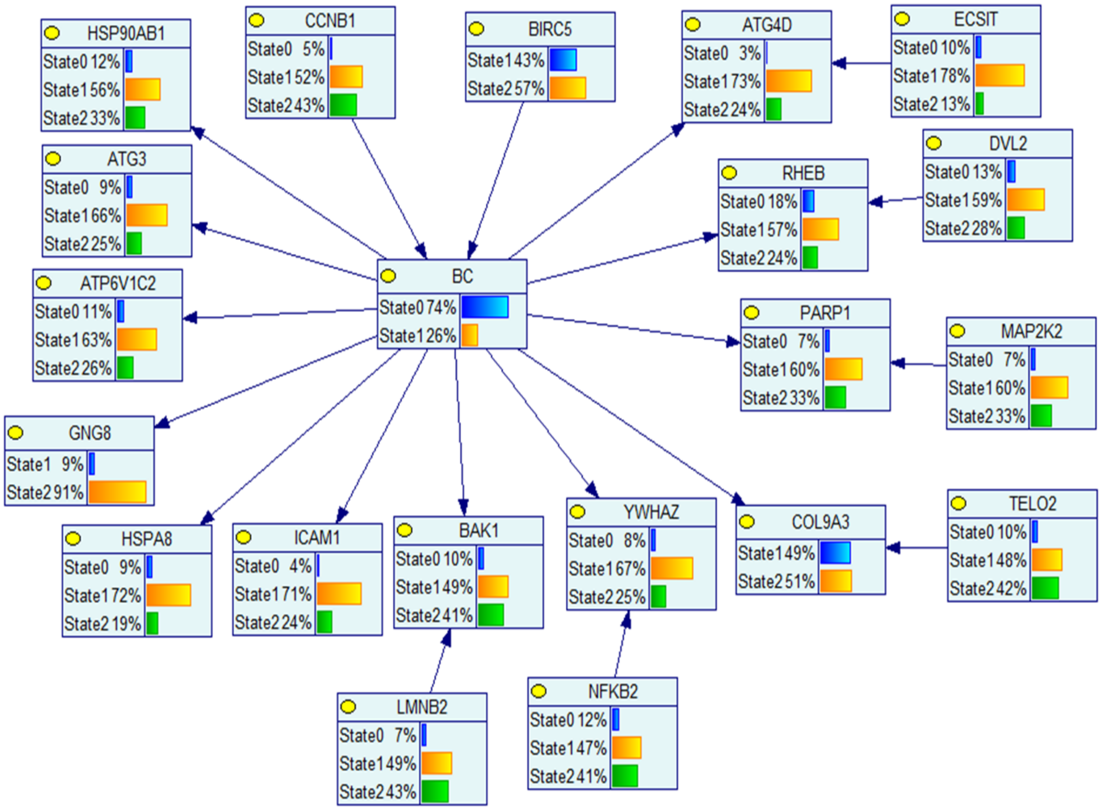


Figure 8. Bayesian network (BN) of causal hypothesis genes learned from the dataset of Non-Hispanic BlackTriple-negative breast cancer (TNBC) cluster (*n* = 48) and normal samples (*n* = 79). Gene expression was categorized as follows: 0 = down, 1 = normal, and 2=upregulated. BC was categorized 1 for breast cancer samples and 0 for normal. This figure shows that the joint probability of TNBC for this cluster was 26%.

Individual and simultaneous changes were simulated, with the results that when ATG3, ATG4D, ATP6V1C2, and BIRC5 were set up to 100% normal expression, the probability of TNBC decreased to 2%. When individual changes to 100% upregulation were simulated, all of them increased the probability of TNBC except BIRC5, which lowered it. This was a surprise because Surviving (also known as baculoviral inhibitor of apoptosis repeat-containing 5), the protein encoded by this gene that belongs to the inhibitor of the apoptosis (IAP) family, is very well known for its dual role as an inhibitor of apoptosis and regulator of cell division. These are both involved in tumorigenesis (Vequaud, Desplanques, Jezequel, Juin, & Barille-Nion, 2016).

Surviving has been found upregulated in breast cancer and is a poor prognostic marker associated with low overall survival (Brennan et al., 2008). ATG3, on the other hand, is part of the Autophagy-related family of proteins (ATG) that regulates autophagy. These proteins can be either protumorigenic or antitumorigenic (Shen et al., 2015).

HER2 Enriched Breast Cancer in Non-Hispanic Whites

Table 13 shows the results of individual gene sensitivity analysis for the HER2 breast cancer model among Non-Hispanic Whites, the only cluster found with enough samples in the TCGA dataset for HER2 enriched breast cancer. GTSE1 ranked number one in affecting breast cancer risk in both directions when switching between the only two states in the sample (state 2 and state 1). This gene is part of the p53 signaling pathway which plays a role in two hallmarks of cancer: Resisting cell death and Enabling replicative immortality.

Table 13

Results of Sensitivity Analysis for HER2 Enriched Model in Non-Hispanic White Cluster

HER2 enriched Non-Hispanic White			
Candidate Gene	**Simulated change in gene expression**	**Probability of breast Cancer (%)**	**Relative Risk (RR)compared to Initial Pr=18 %**
GTSE1	Normal	80	4.44
VDAC1	Normal	66	3.67
ATP6V1A	UP	53	2.94
BMP8A	UP	44	2.44
MAPK13	UP	44	2.44
CDK1	Normal	37	2.06
BMP8A	Down	19	1.06
ATP6V1A	Down	13	0.72
MAPK13	Down	11	0.61
CDK1	UP	10	0.56
VDAC1	UP	8	0.44
ATP6V1A	Normal	5	0.28
BMP8A	Normal	4	0.22
MAPK13	Normal	4	0.22
GTSE1	UP	1	0.06

Note. Notice how Relative Risk figures indicate that the biggest impact in increasing Pr. (BC) from 18% to 80% (RR = 4.44) was achieved by switching GTSE1 to 100% probability of normal expression (initial probability of normal expression was 22%). This gene was primarily overexpressed in this sample—778%). The same gene has the highest effect in the opposite direction of lowering breast cancer risk when switched to 100% upregulation.

GTSE1 (G2 and S-phase expressed 1) has been reported overexpressed in patients with poor outcomes (Canevari et al., 2016) and as a cell migration promoter whose expression is correlated with invasive potential, tumor stage, and distant metastasis in breast tumors (Scolz et al., 2012). It is worth mentioning that GTSE1 is not shown as a candidate gene in any of other seven clusters studied.

Four genes were common to all the three local BNs (ATP6V1A, GTSE1, MAP4K2, and VDAC1). Sensitivity analysis showed that after changing to 100% normal expression the genes ATP6V1A, MAP4K2 and BMP8A, the probability of HER2+ breast cancer decreased to 0% from the initial joint probability of 18%.

Furthermore, we also simulated the overexpression of the same three genes, and the probability of the disease increased to 93%. These findings suggest that these three genes may be implicated in breast cancer development and progression. ATP6V1A is part of the mTOR Pathway and MAP4K2 is involved in the MAP-kinase Pathway (Table 3). Both signaling pathways are associated to sustaining proliferative signaling, one of the hallmarks of cancer. BMP8A is part of the TGF-beta Signaling pathway linked to Evading Growth Suppressors as we had mentioned it before. Figure 9 shows the Bayesian network (BN) of causal hypothesis genes learned from the dataset of Non-Hispanic White HER2 enriched breast cancer cluster and normal samples.

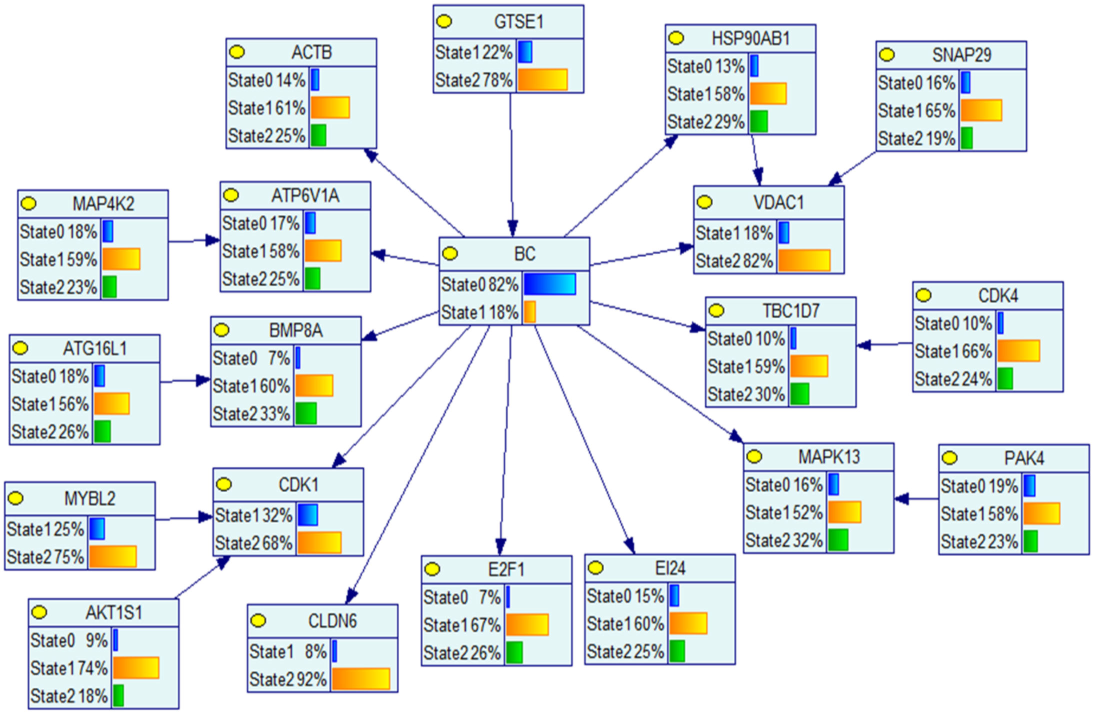


Figure 9. Bayesian network (BN) of causal hypothesis genes learned from the dataset of Non-Hispanic White HER2 enriched breast cancer cluster (*n* = 17) and normal samples (*n* = 79). Gene expression was categorized as follows: 0 = down, 1 = normal, and 2 = upregulated. BC was categorized 1 for breast cancer samples and 0 for normal. This figure shows how that the joint probability of HER2+ breast cancer for this cluster was 18%.

Luminal B in Non-Hispanic Whites

The results of sensitivity analysis are presented in Table 14. Joint probability of breast cancer for this cohort was initially 51% (Figure 10). Only one gene, PARP1, was common to all three local Bayesian Networks generated by Banjo and was the gene with the highest impact on elevating and reducing breast cancer risk for this cluster (RR s = 1.84 and 0.06). PARP1 [Poly (ADP-ribose) polymerase 1] is very well known for its role in DNA repair and is found commonly upregulated in cancer (Ko & Ren, 2012; Rouleau, Patel, Hendzel, Kaufmann, & Poirier, 2010).

Table 14

Results of Sensitivity Analysis for Luminal B Model in Non-Hispanic White Cluster

Luminal B Non-Hispanic White			
Candidate Gene	**Simulated change in gene expression**	**Probability of breast Cancer (%)**	**Relative Risk (RR)compared to Initial Pr=51 %**
PARP1	UP	94	1.84
RRM2	Normal	93	1.82
SLC7A5	UP	88	1.73
TUBA1C	UP	88	1.73
SHISA5	UP	86	1.69
SNAP29	UP	86	1.69
BMP8A	UP	85	1.67
CACNA1E	UP	81	1.59
BAK1	UP	78	1.53
EIF4E	UP	75	1.47
EIF4E	Down	55	1.08
SNAP29	Down	54	1.06
EIF4E	Normal	38	0.75
CACNA1E	Down	37	0.73
BAK1	Down	31	0.61
CACNA1E	Normal	29	0.57
SHISA5	Normal	28	0.55
SLC7A5	Normal	28	0.55
BAK1	Normal	26	0.51
SLC7A5	Down	22	0.43
BMP8A	Down	20	0.39
SNAP29	Normal	19	0.37
TUBA1C	Down	13	0.25
SHISA5	Down	12	0.24
TUBA1C	Normal	11	0.22
PARP1	Down	10	0.20
RRM2	UP	5	0.10
BMP8A	Normal	4	0.08
PARP1	Normal	3	0.06

Note. Notice how Relative Risk figures indicate that the biggest impact in increasing Pr. (BC) from 51% to 94% (RR = 4.44) was achieved by switching PARP1 to 100% probability of upregulation (initial probability of upregulation was 53%). PARP1 also has the highest effect in the opposite direction of lowering breast cancer risk when switched to 100% normal expression (RR = 0.06). This BN model also shows that RRM2 has a significant effect on modulating BC risk in both directions (RR = 1.82 and 0.10).

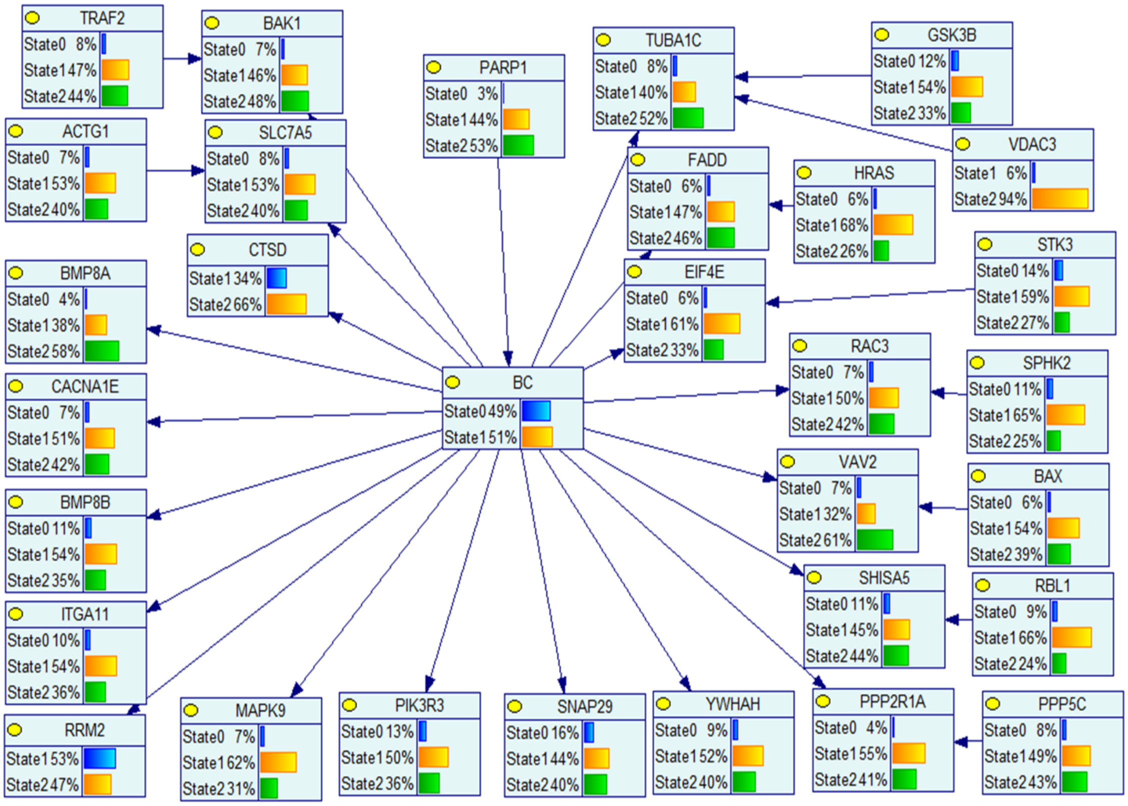


Figure 10. Bayesian network (BN) of causal hypothesis genes learned from the dataset of Non-Hispanic White-Luminal B enriched breast cancer cluster (*n* = 83) and normal samples (*n* = 79). Gene expression was categorized as follows: 0 = down, 1 = normal, and 2 = upregulated. BC was categorized 1 for breast cancer samples and 0 for normal. This figure shows that the joint probability of Luminal B breast cancer for this cluster was 51%.

Rojo et al. (2012) found PARP1 overexpressed in 31.2% of breast cancer samples (*n* = 330), especially in triple-negative breast cancer (51%). PARP1 overexpression was also found associated with a poor prognosis for disease-free and overall survival among all patients. PRP1 inhibitors are actually approved by the FDA under specific parameters for some cases of ovarian, fallopian tube, peritoneal, and breast cancer (Pettitt & Lord, 2018). The Bayesian network model developed by us is in agreement with these reported findings. Simulating overexpression of PARP1 (state 2) from an initial probability of 53% upregulation (Figure 10) to 100 % upregulation resulted in increase of breast cancer probability from 51% to 94% (Relative Risk = 1.84). Conversely, its downregulation, which would be equivalent to treatment with PRP1 inhibitors, lowered the probability of BC to 10% (RR = 0.196). Furthermore, simulating its expression into normal expression levels reduced BC probability to 3% (RR = 0.06).

Sensitivity analysis was also performed on the genes that appeared two out of three times in the BNs of this cluster: BAK1, BMP8A, CACNA1E, EIF4E, RRM2, SHISA5, SLC7A5, SNAP29, TRAF2, and TUBA1C. All had the effect of increasing the probability of breast cancer (RR ranging from 1.47 to 1.73) except for TRAF2. This did not have any effect on the Pr (BC) and RRM2, with the opposite effect of reducing it to 5% (RR = 0.10). Interestingly, RRM2 has been reported to be downregulated in breast cancer metastasis compared to primary breast tumor (Bell, Barraclough & Vasieva, 2017).

Discussion

Dysregulation of transcription factors is a key aspect of cancer development, progression, and therapy resistance (Bhagwat & Vakoc, 2015). Transcription Factor activity profiles between clusters of cancer subtypes and ethnicity may help to elucidate the outcome disparities. Although multiple analysis and comparisons can be conducted from the results of our work, our main goal was focused on triple-negative breast cancer (ER-/ PR- / HER2-) the most aggressive subtype that is also present in higher proportion in Non-Hispanic Blacks. Results of Transcription Factor Target Enrichment Analysis (TFTEA) showed that upregulation of NRF1 activity occurs in all eight cancer subtypes grouped by race and ethnicity. However, the strength of that upregulation as well as the number of signature (differential expressed) genes varies. Based on *p* values (Table 2), the difference in increase of NRF1 activity is more significant in HER2 enriched breast tumors of Non-Hispanic White patients than in all other clusters. HER2 enriched is one of the two more aggressive breast cancers, with triple-negative breast cancer (TNBC), which has a higher incidence rate among African Americans.

The TNBC proportion among all breast cancer cases in the United States general population is between 15% and 20% but in African Americans approximately 30%. TNBC affects more young premenopausal women, and African Americans also present higher mortality rate (Hicks et al., 2013). A survival rate of 5 or 10 years in African American women is significantly worse than in to Non-Hispanic Whites (Doepker, Holt, Durkin, Chu, & Nottingham, 2018). A comparison of TFTEA results between the two TNBC clusters of AA and and Non-Hispanic Whites (Table 2) shows that the number of DE genes is very similar (270), with a higher significance of NRF1 activity increase in the Non-Hispanic White group.

A more interesting comparison is found in analysis of the causal hypothesis genes resulting from the Bayesian Network Analysis listed in Table 4. We found six common causal genes for both clusters (BAK1, CCNB1, HSPA8, LMNB2, TELO2, and YWHAZ); 14 unique in the African American cluster (ATG3 , ATG4D , ATP6V1C2 , BIRC5 , COL9A3 , DVL2 , ECSIT , GNG8 , HSP90AB1 , ICAM1 , MAP2K2 , NFKB2 , PARP1, and RHEB); and 25 unique in the Non-Hispanic White cluster (BMP8A , CADM1 , CASP2 , CBLC , CD44 , CLDN4 , CLDN7 , DAXX , EFNA1 , EFNA2 , ELK1 , GNB1 , IFNAR2 , INPPL1 , PPP1CA , PPP2R1A , PPP2R5D , PTK2 , RELB , RRM2 , STK4 , STMN1 , VAMP8 , VAV2 , and VDAC1) (Figure 11).

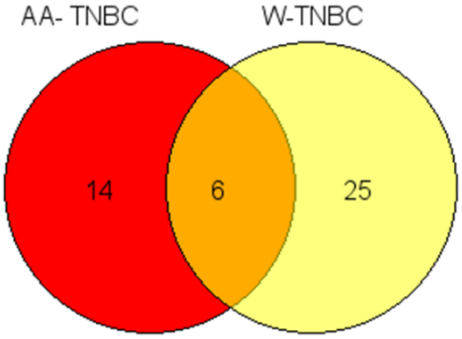


Figure 11. Venn diagram showing overlap of differentially expressed causal hypothesis genes in TNBC between African American (AA) and Non-Hispanic White (W) tumor samples. (Diagram constructed using Core Graphic Module by Vijayaraj Nagarajan, and Web implementation by Mehdi Pirooznia. October 2006, usm.edu)

Can the NRF1 regulated causal hypothesis genes that are unique for the TNBC African American cluster explain the increase mortality rate compared to Whites? We learned that three genes--DVL2, MAP2K2, and NFKB2—were part of the KEGG breast cancer pathway, and two of them have already been linked to breast cancer. DVL2 (disheveled segment polarity protein 2) is involved in promoting migration of breast cancer cells via Wnt signaling, which has been found dysregulated in TNBC and associated with metastasis (Dey et al., 2013; Pohl et al., 2017; Zhu et al., 2012).

NFKB2 (nuclear factor kappa B subunit 2) encodes a subunit of the transcription factor complex nuclear factor-kappa B (NF-κB) (Stelzer et al., 2017). Nuclear factor-kappa B (NF-κB) signaling has been reported involved in the regulation of breast cancer stem cells properties (Yeo, French, Spada, & Clarkson, 2017). Among the other unique causal hypothesis genes in the African American-TNBC cluster, ATP6V1C2 (ATPase, H+ transporting, lysosomal 42kD, V1 subunit C isoform 2) t showed the highest individual impact on breast cancer risk in our sensitivity analysis (RR = 2.85) and had been reported associated to breast cancer growth and metastasis (McConnell et al., 2017).

For a graphical view of the NRF1 activity profile showing causal hypothesis genes for the eight clusters, we developed a heat map (Figure 12). Notice how TNBC in the Non-Hispanic Black cluster has the profile with the highest number of upregulated genes, followed by TNBC in the Non-Hispanic White cluster. Again, the heat map also shows that the three genes mentioned before have the highest expression level in the TNBC / Non-Hispanic Black cluster (TNB-BLACK).

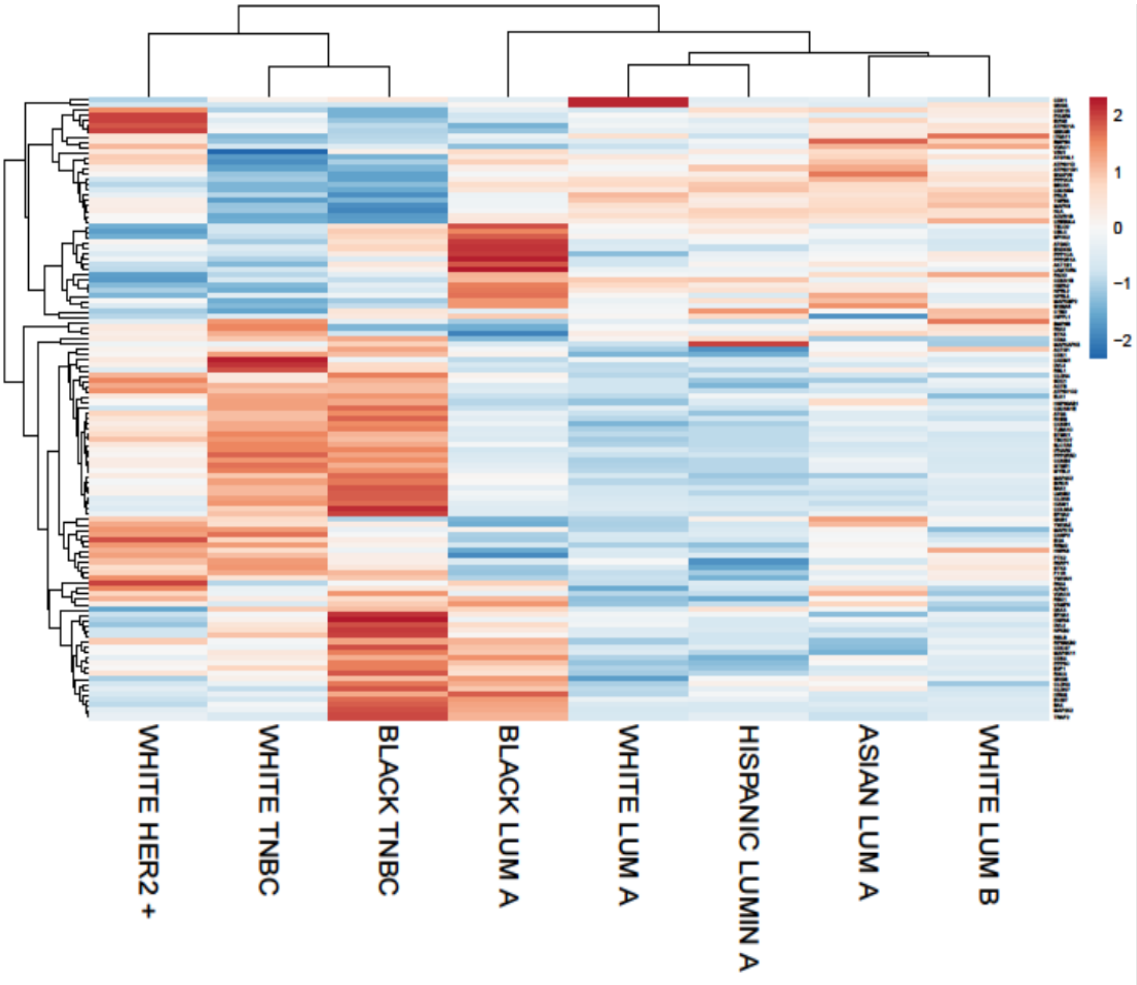


Figure 12. Heat map of causal hypothesis genes generated from Bayesian network analysis for all eight clusters. Notice how TNBC clusters for Black and White patients show the profiles with the highest number of upregulated genes. TNBC is the most aggressive subtype of breast cancer and is present in higher proportion in African Americans. Heat map constructed using Clustvis (Metsalu & Vilo. 2015).

Conclusion

Breast cancer incidence, death rates, and overall survival vary depending on molecular subtypes, race, and ethnicity. Age of diagnosis, proportion of more aggressive tumors, and survival rates are worse among Non-Hispanic Black (African American) compared to Non-Hispanic White women. Biological and nonbiological factors may explain these disparities. Several studies have been conducted addressing the nonbiological factors, such as access to health care, cultural issues, and comorbidities, unlike biological factors that still lack understanding.

Triple-negative breast cancer is the most aggressive subtype, which is also present in higher proportion in African Americans. Our results show how NRF1 sensitivity, including comparison of NRF1 activity profile of causal hypothesis genes in TNBC samples from African Americans versus TNBC samples from Non-Hispanic Whites, may explain the disparities in outcomes such as lower overall survival. Fourteen genes were found to be in the list of causal hypothesis genes that are unique to the TNBC- African American cluster These genes included DVL2 (disheveled segment polarity protein 2) previously reported to be associated with promoting migration of breast cancer cells, NFKB2 (nuclear factor kappa B subunit 2) involved in regulation of breast cancer stem cells properties, and ATP6V1C2 (ATPase, H+ transporting, lysosomal 42kD, V1 subunit C isoform 2) involved in breast cancer growth and metastasis.

The heat map (Figure 12) provides important information of NRF1 activity profiles for all eight clusters. This map can lead to new analysis involving breast cancer subtypes other than TNBC, our main focus. Our findings help to elucidate the role of NRF1 sensitivity in the development of TNBC in different racial/ethnic groups of breast cancer patients. Our findings may help in the future development of novel therapies.

Methods

RNA-Seq and Clinical Data

RNA-Seq gene expression of 20,502 genes and clinical data corresponding to 1,212 breast cancer and normal tissues samples were downloaded from TCGA with Broad Institute's Firehose tool (Version: std. data 2016-01-28). RNA-Seq data collected were level 3, specifying data that had already been normalized and assembled in counts per gene. Clinical data of patients were deidentified and included several characteristics, such as race, ethnicity, receptor status, cancer stage, and age at diagnosis.

Receptor status (ER / PR / HER2) information was used to classify samples into different clusters based on molecular subtypes: luminal A (ER+ and/or PR+, HER2-), luminal B (ER+ and/or PR+, HER2+), triple-negative (ER-, PR-, HER2-), and HER2 (ER-, PR-, HER2+) (Stewart, Luks, Roycik, Sang, & Zhang, 2013). Race and ethnicity were also recorded (Table 1). The number of samples was small in some of the groups; therefore, only clusters with enough samples (eight in total) were selected for the study.

Differential Expression Analysis in Breast Tumor Compared to Normal Tissue

Differential expression of all 20,502 genes in breast tumor compared to normal samples was estimated with the limma package (Ritchie et al., 2015). This package uses several statistical principles that makes it effective for gene expression studies involving large number of genes. Limma works with a matrix of gene expression values, in which genes are listed in rows and samples in columns.

First, we used the voom function in limma to transform the normalized RNA-Seq counts into log2 counts per million (log CPM) in order to continue with the downstream analysis of differential expression. The principal tool of the limma package is to develop a linear model for each gene (row), including calculation of regression coefficients and standard errors. This method allows the design of different experiments more complex that the common comparison of two phenotypes. Linear modeling was performed using the lmFit function. Subsequently, the contrast matrix was developed to compute the log2 fold expression changes and *t* statistics (breast cancer vs. normal tissue). Finally, the empirical Bayes approach with the eBayes function was used to estimate differential expression. The moderated *t* statistics were used for significance testing.

The results provided by limma included logFC (log2 fold change), average expression, and adjusted *p* value (Table 15). Table 15 lists the top 10 differentially expressed genes in Luminal A breast cancer samples / Non-Hispanic White group and is a partial view of the complete list, presented here for illustrative purposes, given the extension of the file (more than 20,000 genes). The genes were ranked by adjusted *p* value to prepare the matrix for input into LRpath, the software used for Transcription Factor Target Enrichment Analysis (TFTEA). Differential expression using limma was performed for each of the eight clusters.

Table 15

Top 10 Differentially Expressed Genes Ranked by Adjusted p Value in Luminal A Non-Hispanic White Breast Cancer Samples (N = 373) Compared to Normal Samples (N = 79)

Gene	logFC	AveExpr	t	P.Value	adj.P.Val	B
VEGFD	-6.1834741	-0.0306428	-36.924037	1.23E-138	2.52E-134	306.117173
HIF3A	-5.2653731	-0.5516659	-32.940671	2.83E-122	2.90E-118	268.528229
LYVE1	-5.206841	1.97975346	-32.743262	1.93E-121	1.32E-117	266.799124
DMD	-3.7915362	3.87204437	-31.999619	2.76E-118	1.15E-114	259.60918
CD300LG	-6.1113788	1.13376895	-31.998388	2.80E-118	1.15E-114	259.523984
PAMR1	-3.7200305	2.94456505	-31.655643	8.15E-117	2.78E-113	256.213842
SCARA5	-6.1865318	1.09163597	-30.156982	2.42E-110	7.10E-107	241.318589
RYR3	-3.7268943	-0.3157998	-29.737855	1.64E-108	4.20E-105	236.828528
BTNL9	-4.0324113	3.03237913	-29.698696	2.43E-108	5.54E-105	236.760205
MYOM1	-3.8374644	1.5576487	-29.541303	1.19E-107	2.44E-104	235.121503

Note. Columns are log 2 fold change (logFC), average gene expression, adjusted *p* value and log-odds that the gene is differentially expressed (B).

Transcription Factor Target Enrichment Analysis (TFTEA) to Estimate Changes in NRF1 Activity

TFTEA was the method used to measure NRF1 activity based on the combined changes in activity of its target genes (Falco, Bleda, Carbonell-Caballero, & Dopazo, 2016). TFTEA is a Gene Set Enrichment (GSE) analysis that aims to detect asymmetrical distribution of the transcription factor target genes in the top (or the bottom) of the list of differentially expressed genes ranked by the adjusted *p* value (Falco et al., 2016). NRF1 target genes were selected from the results we reported previously using NRF1 ChIP sequence data from the HCC1954 breast cancer cell line, a good model of HER2 enriched breast cancer (Ramos et al., 2018).

Several studies have been conducted to determine which transcription factor binding's sites are functional. The results have shown that the most prevalent transcriptionally functional mechanisms involve binding next to the TSS (Tabach et al., 2007). Consequently, we selected 8,443 genes with NRF1 peaks located in the promoter region. For this study, the promoter region was defined as -5,000 to +1,000 bp from the Transcription Start Site (TSS), as proposed by Falco et al. (2016). However, the downstream coordinate was revised to +1,000 bp (proposed by cited reference = up to first exon). We considered + 1,000 bp from TSS to be a good approximation because the average length of partially coding first exon in the human genome is 348 bp and the average 5' UTR is 210 bp (Davuluri, Grosse, & Zhang, 2001; Mignone, Gissi, Liuni, & Pesole, 2002).

After we obtained the lists of differentially expressed genes from limma and the list of NRF1 target genes, in the first step we used these two files as input into the web-based application LRpath (http://lrpath.ncibi.org/) to perform the TFTEA (Kim et al., 2012; Lee, Patil, & Sartor, 2016; Sartor et al., 2009). Initially, the input file (limma output) containing the list of all 20,502 genes in rows and three columns displaying log fold change (logFC) and adjusted *p* value and average gene expression was uploaded into the LRpath web server. The second step was to upload the dataset to search against the list of NRF1 target genes. The final step before execution of the LRpath search was to set up the searching parameters, which included the selection of directional test option to detect up or down regulation of the gene set under investigation.

This process was repeated with all eight clusters. The LRpath method consists of using linear regression to find the functional relationship between the odds of a gene to be part of a gene set (NRF1 target gene in this case) with statistical significance of its differential expression (adjusted *p* value). To measure the statistical significance, LRpath computes the *p* value adjusted for multiple testing using the Benjamini-Hochberg procedure to reduce the false discovery rate (FDR). By our setting up LRpath to perform a directional test, the software was able to determine whether the NRF1 target set was enriched with genes up or down regulated.

Based on these results, it could be inferred whether NRF1 activity was up or down regulated. LRpath output included the number and symbols of NRF1 target genes with DE $p < 0.05$ (signature genes), the direction of the enrichment (upregulation or downregulation) and the *p* value. The selection of genes for Bayesian network modeling was based on the list of signature genes generated by LRpath.

Selection of Genes for Bayesian Network Modeling

Differential expression analysis and TFTEA, as mentioned, allowed us to obtain the list of NRF1 target genes with significant differential expression in breast cancer compared to normal for each cluster (signature genes). The strategy chosen to identify the drivers of the disease was to construct a Bayesian network model using the software Banjo. Since the number of signature genes for each cluster was too large (ranging from 3,103 Luminal A to 2,252 HER2+ in Non-Hispanic Whites) to perform the Bayesian network modeling, we lowered the number of genes by selecting only those involved in the processes of cells acquiring the hallmarks of cancer (Hanahan & Weinberg, 2011).

For that purpose, we searched against the list of 902 NRF1 target genes involved in hallmarks of cancer we had previously discovered using ChIP-Seq data of HCC1954 (breast cancer cells) and HMEC (normal human mammary epithelial cells) cell lines (Table 3) (Ramos et al., 2018). Table 3 shows genes classified by signaling pathways (PI3K-Akt Signaling, MAP-kinase Pathway, mTOR Pathway, Cellular Senescence, p53 Signaling, Apoptosis, TGF-beta Signaling, Autophagy, VEGF Signaling, ECM-receptor interaction, Cell adhesion molecules (CAMs), T cell receptor signaling pathway, and B cell receptor signaling pathway) and mapped to the hallmarks of cancer (Sustaining proliferative signaling, Evading growth suppressors, Resisting cell death, Enabling replicative immortality, Inducing Angiogenesis, Activating invasion and metastasis, and Evading immune destruction). Use of this list as a filter not only contributed to a lower number of candidate genes but also to incorporate biological knowledge to the model. Table 16 shows the number of genes that finally were used for input into Banjo for Bayesian network learning.

Table 16

Number of NRF1 Target Genes Selected for Bayesian Network Modeling in Each Cluster After Selecting Among Differentially Expressed (Signature Genes) Those Involved in Signaling Pathways Linked to the Hallmarks of Cancer

	Breast Cancer samples			**Normal samples**			**Bayesian Network Modeling**
MOL SUBT	**ETHNICITY AND RACE**		**# of samples**	**ETHNICITY AND RACE**		**# of samples**	**Number of NRF1 target genes with DE p<0.05 and also involved in hallmarks of cancer signaling pathways**
HER2 Enriched	Non-Hispanic	White	17	Non-Hispanic	White	79	138
Luminal A	Non-Hispanic	Asian	22	Non-Hispanic	White	79	163
Luminal A	Non-Hispanic	White	373	Non-Hispanic	White	79	181
Luminal B	Non-Hispanic	White	83	Non-Hispanic	White	79	161
Luminal A	Hispanic	White	21	Non-Hispanic	White	79	158
Triple-negative	Non-Hispanic	White	74	Non-Hispanic	White	79	181
Luminal A	Non-Hispanic	Black	69	Non-Hispanic	White	79	163
Triple-negative	Non-Hispanic	Black	48	Non-Hispanic	White	79	162

Bayesian Network Modeling: Structural and Parameter Learning

Bayesian networks (BNs) are graphical representation of joint probability distributions. A BN consists of a number of variables represented by nodes which are connected by edges representing causal probabilistic relationship between the variables. To develop a BN structure, two aspects need to be learned from the dataset: the structure and the parameters (Fuster-Parra et al., 2016). Banjo free software developed under the direction of Alexander J. Hartemink in the Department of Computer Science at Duke University was used to obtain the structures (https://users.cs.duke.edu/~amink/software/banjo/) (Hartemink, 2010).

The data matrix for input into Banjo consists of the list of variables as row names (selected genes, age of patients at diagnosis, and breast cancer status) and sample IDs as column names. The matrix is completed with the corresponding values. Banjo required the data to be categorized. For gene expression we used three tiers with cutoff points equal to the mean plus or minus one standard deviations of the particular gene expression in the group of normal tissue samples.

Any value between the two cutoff points was considered normal with categorical value equal to one (1). Values greater than the mean plus one standard deviation were considered upregulated with an assigned value of two (2), and values below the mean minus one standard deviation were considered downregulated with a value of zero (0). For age we used three tiers: less than 50 years of age at the time of diagnosis was categorized as equal to 0, between 50 and 60 years was categorized as equal to 1, and more than 60 years old was categorized as equal to 2. Disease status was categorized 0 for normal tissue samples and 1 for breast cancer samples.

Structural learning of BNs from data is considered an NP-hard problem, and the number of possible networks increases exponentially with the increase in the number of variables (Adabor, Acquaah-Mensah, & Oduro, 2015). Banjo's approach to structural learning is based on searching and scoring for structure inference (Hartemink, 2010). The metric used for scoring is the Bayesian Dirichlet Equivalence (BDe), which is proportional to the posterior probability of the network given the data. The single highest scoring network is selected after searching millions of structures. Given the extent of the search, Banjo provides the best local network found after the time limit has been reached. We set up 8 hours as the time limit for the search and ran Banjo three times for each cluster. Each run provided the best local network, after which we selected the one with the best BDe score as the closest approximation to the global network.

Selected networks for each cluster were used to identify the Markov blanket genes (also called causal hypothesis genes) of the breast cancer node (BC), our variable of interest. Markov blanket genes of a node are its parents, children, and other children's parents. This variable (node) is conditionally independent of all the other variables (nodes), and therefore Markov blanket genes are the only ones we need to incorporate in our final BN model, whose goal is to identify the drivers of the disease (Figure 2).

After the network and Markov blanket genes have been selected, the final step for Bayesian network modeling is parameter learning. During parameter learning, a software program is used to estimate the conditional probabilities given the structure (Fuster-Parra et al., 2016) and the data matrix. This data matrix is basically the same matrix used for network learning except that it includes omly the Markov blanket genes. We used GeNIe Modeler, software developed at the

Decision Systems Laboratory, University of Pittsburgh (currently licensed to the company BayesFusion) for parameter learning. GeNIe can be downloaded free for academics using the software for teaching and research at https://download.bayesfusion.com/files.html?category=Academia.

Validation of Proposed Bayesian Networks (Bns)

The main objective of the constructed BN model is to predict the probability of breast cancer based on the values of the variables (gene expression). GeNIe can use different methods for validation. We used the most powerful crossvalidation method, known as K-fold crossvalidation. In this method, the data are divided into K equally sized groups. The model is trained with K-1 and is validated using the Kth group. This process is repeated with different parts of the dataset. For our validation process we selected K = 10. Validation results from GeNIe include a file with predictive and real values of breast cancer probability and the values of sensitivity, accuracy, and area under the Receiver Operating Characteristic (ROC) curve. Table 6 shows the results of crossvalidation for all eight clusters. After proving with validation that these were good models for breast cancer prediction, we proceeded to carry out sensitivity analysis to identify those genes that have the greatest impact on breast cancer.

Sensitivity Analysis of Bayesian Networks (Bns)

In statistical terms, GeNIe allows us to automatically estimate the posterior probability distribution after observing evidence. This observing evidence may be changes in levels of gene expression. Since we wanted to identify those genes with the highest impact on the probability of breast cancer, we developed a strategy of systematically changing the levels of gene expression in the causal hypothesis genes to detect what type of changes and what genes affected the most the probability of development of breast cancer according to the model.

Sensitivity analysis was concentrated on the parents of the BC node and the genes that appeared in the Markov blanket genes in at least two of the three best local networks. However, we did not discard any of the discovered Markov blanket genes as potential drivers of the disease for each cluster studied here. Our search was not restricted to individual gene changes but included simultaneous changes in several genes. We considered that sensitivity analysis using GeNIe is easier to explain with the use of practical examples and thus decided to incorporate most of the details of the methods used in this part of our research into the results section. With this information, readers may become informed about the results and methods at the same time.

REFERENCES

Adabor, E. S., Acquaah-Mensah, G. K., & Oduro, F. T. (2015). SAGA: A hybrid search algorithm for Bayesian network structure learning of transcriptional regulatory networks. *Journal of Biomedical Informatics, 53*, 27-35.

Alarmo, E. L., & Kallioniemi, A. (2010). Bone morphogenetic proteins in breast cancer: Dual role in tumourigenesis? *Endocrine-Related Cancer, 17*(2), R123-R139.

Bell, R., Barraclough, R., & Vasieva, O. (2017). Gene expression meta-analysis of potential metastatic breast cancer markers. *Current Molecular Medicine, 17*(3), 200-210.

Bhagwat, A. S., & Vakoc, C. R. (2015). Targeting transcription factors in cancer. *Trends in Cancer, 1*(1), 53-65.

Brennan, D. J., Rexhepaj, E., O'Brien, S. L., McSherry, E., O'Connor, D. P., Fagan, A., . . . Landberg, G. (2008). Altered cytoplasmic-to-nuclear ratio of Survivin is a prognostic indicator in breast cancer. *Clinical Cancer Research, 14*(9), 2681-2689.

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., & Mountain, J. L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *American Journal of Human Genetics, 96*(1), 37-53.

Canevari, R. A., Marchi, F. A., Domingues, M. A., de Andrade, V. P., Caldeira, J. R., Verjovski-Almeida, S., . . . Reis, E. M. (2016). Identification of novel biomarkers associated with poor patient outcomes in invasive breast carcinoma. *Tumour Biology, 37*(10), 13855-13870.

Chen, D., Li, Y., Wang, L., & Jiao, K. (2015). SEMA6D expression and patient survival in breast invasive carcinoma. *International Journal of Breast Cancer, 2015,* 1-10.

Davuluri, R. V., Grosse, I., & Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. *Nature Genetics, 29*(4), 412-417.

Dey, N., Barwick, B. G., Moreno, C. S., Ordanic-Kodani, M., Chen, Z., Oprea-Ilies, G., . . . Ambramovitz, M. (2013). Wnt signaling in triple-negative breast cancer is associated with metastasis. *BMC Cancer, 13,* 1-15.

Ding, K., Li, W., Zou, Z., Zou, X., & Wang, C. (2014). CCNB1 is a prognostic biomarker for ER+ breast cancer. *Medical Hypotheses, 83*(3), 359-364.

Doepker, M. P., Holt, S. D., Durkin, M. W., Chu, C. H., & Nottingham, J. M. (2018). Triple-negative breast cancer: A comparison of race and survival. *American Surgeon, 84*(6), 881-888.

Falco, M. M., Bleda, M., Carbonell-Caballero, J., & Dopazo, J. (2016). The pan-cancer pathological regulatory landscape. *Scientific Reports, 6,* 1-13.

Fushimi, K., Ray, P., Kar, A., Wang, L., Sutherland, L. C., & Wu, J. Y. (2008). Up-regulation of the proapoptotic caspase 2 splicing isoform by a candidate tumor suppressor, RBM5. *Proceedings of the National Academy of Sciences of the United States of America, 105*(41), 15708-15713.

Fuster-Parra, P., Tauler, P., Bennasar-Veny, M., Ligeza, A., Lopez-Gonzalez, A. A., & Aguilo, A. (2016). Bayesian network modeling: A case study of an epidemiologic system analysis of cardiovascular risk. *Computer Methods and Programs in Biomedicine, 126*, 128-142.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646-674.

Hartemink, A. J. (2010). *Banjo.* Retrieved from 06/11/2018http://www.cs.duke.edu/~amink/software/banjo/

Hicks, C., Kumar, R., Pannuti, A., Backus, K., Brown, A., Monico, J., & Miele, L. (2013). An integrative genomics approach for associating GWAS information with triple-negative breast cancer. *Cancer Informatics, 12,* 1-20.

Hou, L., Chen, M., Wang, M., Cui, X., Gao, Y., Xing, T., . . . Jiang, J. (2016). Systematic analyses of key genes and pathways in the development of invasive breast cancer. *Gene, 593*(1), 1-12.

Kim, J. H., Karnovsky, A., Mahavisno, V., Weymouth, T., Pande, M., Dolinoy, D. C., . . . Sartor, M. A. (2012). LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types. *BMC Genomics, 13*(526), 1-16.

Ko, H. L., & Ren, E. C. (2012). Functional aspects of PARP1 in DNA repair and transcription. *Biomolecules, 2*(4), 524-548.

Lee, C., Patil, S., & Sartor, M. A. (2016). RNA-enrich: A cut-off free functional enrichment testing method for RNA-seq with improved detection power. *Bioinformatics, 32*(7), 1100-1102.

Liu, C. Y., Huang, T. T., Huang, C. T., Hu, M. H., Wang, D. S., Wang, W. L., . . . Chen, M. H. (2017). EGFR-independent Elk1/CIP2A signalling mediates apoptotic effect of an erlotinib derivative TD52 in triple-negative breast cancer cells. *European Journal of Cancer, 72,* 112-123.

McConnell, M., Feng, S., Chen, W., Zhu, G., Shen, D., Ponnazhagan, S., . . . Li, Y. P. (2017). Osteoclast proton pump regulator Atp6v1c1 enhances breast cancer growth by activating the mTORC1 pathway and bone metastasis by increasing V-ATPase activity. *Oncotarget, 8*(29), 47675-47690.

Metsalu, T., & Vilo, J. (2015). ClustVis: A web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Research, 43*(W1), W566-W570.

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biology, 3*(3), 1-10.

Pettitt, S. J., & Lord, C. J. (2018). PARP inhibitors and breast cancer: Highlights and hang-ups. *Expert Review of Precision Medicine and Drug Development, 3*(2), 83-94.

Pohl, S. G., Brook, N., Agostino, M., Arfuso, F., Kumar, A. P., & Dharmarajan, A. (2017). Wnt signaling in triple-negative breast cancer. *Oncogenesis, 6*(4), e310.

Ramos, J., Das, J., Felty, Q., Yoo, C., Poppiti, R., Murrell, D., . . . Roy, D. (2018). NRF1 motif sequence-enriched genes involved in ER/PR -ve HER2 +ve breast cancer signaling pathways. *Breast Cancer Research and Treatment*, *8,* 1-17.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7), 1-26.

Roberts, P. J., Bisi, J. E., Strum, J. C., Combest, A. J., Darr, D. B., Usary, J. E., . . . Sharpless, N. E. (2012). Multiple roles of cyclin-dependent kinase 4/6 inhibitors in cancer therapy. *Journal of the National Cancer Institute, 104*(6), 476-487.

Rojo, F., Garcia-Parra, J., Zazo, S., Tusquets, I., Ferrer-Lozano, J., Menendez, S., . . . Lobo, F. (2012). Nuclear PARP-1 protein overexpression is associated with poor overall survival in early breast cancer. *Annals of Oncology*, *23*(5), 1156-1164.

Rouleau, M., Patel, A., Hendzel, M. J., Kaufmann, S. H., & Poirier, G. G. (2010). PARP inhibition: PARP1 and beyond. *Nature Reviews Cancer, 10*(4), 293-301.

Sartor, M. A., Leikauf, G. D., & Medvedovic, M. (2009). LRpath: A logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics, 25*(2), 211-217.

Scolz, M., Widlund, P. O., Piazza, S., Bublik, D. R., Reber, S., Peche, L. Y., . . . Ellenberg, J. (2012). GTSE1 is a microtubule plus-end tracking protein that regulates EB1-dependent cell migration. *PloS One, 7*(12), 1-17.

Shen, M., Duan, W. M., Wu, M. Y., Wang, W. J., Liu, L., Xu, M. D., . . . Gong, F. R. (2015). Participation of autophagy in the cytotoxicity against breast cancer cells by cisplatin. *Oncology Reports, 34*(1), 359-367.

Shubbar, E., Kovacs, A., Hajizadeh, S., Parris, T. Z., Nemes, S., Gunnarsdottir, K., . . . Helou, K. (2013). Elevated cyclin B2 expression in invasive breast carcinoma is associated with unfavorable clinical outcome. *BMC Cancer, 13*, 1-10.

Song, Y., Zhao, C., Dong, L., Fu, M., Xue, L., Huang, Z., . . . Lu, N. (2008). Overexpression of cyclin B1 in human esophageal squamous cell carcinoma cells induces tumor cell invasive growth and metastasis. *Carcinogenesis, 29*(2), 307-315.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., . . . Kaplan, S. (2016). The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics, 54*, 1-33.

Stewart, P. A., Luks, J., Roycik, M. D., Sang, Q. X., & Zhang, J. (2013). Differentially expressed transcripts and dysregulated signaling pathways and networks in African American breast cancer. *PloS One, 8*(12), 1-13.

Tabach, Y., Brosh, R., Buganim, Y., Reiner, A., Zuk, O., Yitzhaky, A., . . . Domany, E. (2007). Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PloS One, 2*(8), 1-14.

Vassilev, L. T., Tovar, C., Chen, S., Knezevic, D., Zhao, X., Sun, H., . . . Chen, L. (2006). Selective small-molecule inhibitor reveals critical mitotic functions of human CDK1. *Proceedings of the National Academy of Sciences of the United States of America, 103*(28), 10660-10665.

Vequaud, E., Desplanques, G., Jezequel, P., Juin, P., & Barille-Nion, S. (2016). Survivin contributes to DNA repair by homologous recombination in breast cancer cells. *Breast Cancer Research and Treatment, 155*(1), 53-63.

Yeo, S. K., French, R., Spada, F., & Clarkson, R. (2017). Opposing roles of Nfkb2 gene products p100 and p52 in the regulation of breast cancer stem cells. *Breast Cancer Research and Treatment, 162*(3), 465-477.

Zabkiewicz, C., Resaul, J., Hargest, R., Jiang, W. G., & Ye, L. (2017). Bone morphogenetic proteins, breast cancer, and bone metastases: Striking the right balance. *Endocrine-Related Cancer, 24*(10), R349-R366.

Zhu, Y., Tian, Y., Du, J., Hu, Z., Yang, L., Liu, J., & Gu, L. (2012). Dvl2-dependent activation of Daam1 and RhoA regulates Wnt5a-induced breast cancer cell migration. *PloS One, 7*(5), 1-12.

CHAPTER VI

OVERALL CONCLUSIONS

The primary goal of this dissertation was to decipher mechanisms by which nuclear respiratory factor 1 (NRF1) coordinates changed in the transcriptional and chromatin landscape, affecting development and progression of invasive breast cancer. This study was undertaken to contribute to clarification of the molecular basis underlying the aggressiveness of some breast cancer subtypes and disparities associated with race and ethnicity. Based on previous research conducted by our laboratory and the current literature review demonstrating the involvement of the transcription factor NRF1 in the control of breast cancer cells cycle progression, we hypothesized that NRF1 reprogramming of the transcription of tumor initiating gene(s) and tumor suppressor gene(s) contribute to the development and progression of invasive breast cancer.

Three specific goals were established to test our hypothesis: (a) Decipher regulatory landscape of NRF1 networks in breast cancer. (b) Determine the role of NRF1 gene networks in different subtypes of breast cancer. (c) Determine differential NRF1 gene network sensitivity contributing to breast cancer disparities.

To accomplish the first goal, we used published NRF1 ChIP-Seq data from different breast cancer cells (MCF7, T47D, and HCC1954) and normal human mammary epithelial cells (HMEC) to identify approximately 10,000 potential NRF1 target genes with NRF1 binding sites next to the Transcription Start Site (TSS) and enhancer regions located hundreds of thousands of cells from the promoter region. We also found that NRF1 regulatory network was cell context dependent. Using

Gene Ontology and Pathway Analysis, we confirmed the participation of NRF1 regulated genes in signaling pathways and biological processes important in breast cancer development and progression.

To determine the role of NRF1 gene networks in different subtypes of breast cancer, we used a large set of RNA-Seq gene expression (dataset 20,502 genes) corresponding to 1,212 samples from the Cancer Genome Atlas (TCGA). A systematic integration of ChIP DNA-seq, RNA-Seq data combined with NRF1 protein-DNA motif binding, signal pathway analysis, and Bayesian machine learning were used to identify differentially regulated NRF1 target genes involved in ER/PR - Her2 + (HER2 enriched) breast cancer. Contribution to the susceptibility of the disease may be via perturbation of regulation of diverse growth factor receptors, PI3K-Akt-mTOR, MAPK, E2Fs, and Wnt pathways. We also observed new roles for NRF1 in the acquisition of breast tumor initiating cells, regulation of epithelial to mesenchymal transition (EMT), and invasiveness of breast cancer stem cells. The NRF1 motif was one of the principal regulatory motifs significantly associated with worsening histological grades and poor breast cancer prognosis.

Finally, using differentially expressed genes, transcription factor target enrichment analysis (TFTEA) and Bayesian network analysis to investigate breast cancer disparities, we discovered 14 causal hypothesis genes that may explain the outcome disparities in TNBC when we compared African American with Non-Hispanic White patients. Our findings were aligned with previous studies reporting that the genes DVL2, NFKB2, and ATP6V1C2 were linked to growth, migration, and metastasis of breast cancer cells.

Clinical confirmation of our study will have a significant impact on the understanding of the role of NRF1 as a valuable additional biomarker for assessing resistance to therapeutic response in HER2+ and TNBC, the two most aggressive breast cancer subtypes.

Limitations

Methods used in the search of NRF1 motif sequence-enriched genes involved in er-pr-her2+ breast cancer signaling pathways (Chapter IV) and breast cancer disparities associated with aggressive subtypes HER2+ and triple- negative breast cancer—TNBC (Chapter V) involved the use of Bayesian network analysis. We used the Bayesian score, which is the posterior probability of the network given the data P (G/D), to select the structure with the best score and to make inferences assuming this was the true model.

Even though this approach is widely used, it lacks consideration of the uncertainty of the model. This uncertainty is particularly risky when there are a large number of structures with highest scores that are very close to each other. Several methods have been proposed to account for model uncertainty, referred to as Bayesian Model Averaging (BMA). Thus, it would be advisable to use BMA to confirm our results. Nevertheless, the candidate hypothesis genes we discovered were confirmed with the mathematical validation of the model to predict breast cancer status. We established with biological knowledge that some cellular processes and signaling pathways known to play important roles in cancer development and progression were enriched with these NRF1 targets.

Future Studies

The heterogeneity of breast cancer is widely known. Scientists have made great progress in finding common ground by categorizing breast tumors into five to 10 molecular subtypes. Here we have shown that the genetic profile of breast cancer can be different from one individual to another based not only on molecular subtype but also on race and ethnicity. The methods we used in this research and the results may be employed in the future towards a more personalized approach aimed at identifying patient-specific genetic profiles of tumors to identify gene drivers specific to patients. Corresponding personalized treatment and therapy could be aimed at increasing the overall survival of breast cancer patients, especially patients affected with the more aggressive subtypes of breast cancer.

VITA

JAIRO RAMOS

1976	B.S. Mechanical Engineering Universidad del Norte Barranquilla, Colombia
1987	M.B.A. Rutgers University Newark, New Jersey
2013	M.P.H. Florida International University Miami, Florida
2018	Ph.D. in Public Health candidate Florida International University Miami, Florida
July-December 2015	Intern Florida Department of Health Miami Dade County Division of Environmental Health and Engineering Miami, Florida
May-November 2013	Research Assistant Intern University of Miami Miller School of Medicine Department of Neurology Brain Endowment Bank Miami, Florida
September 2010- December 2013	Adjunct Faculty Miami Dade College School of Business Miami, Florida

PUBLICATIONS AND PRESENTATIONS

Ramos, J., Felty, Q., Yoo. C., & Roy, D. (In preparation). Working title: Differential NRF1 gene network sensitivity contributing to breast cancer disparities.

Ramos, J., & Roy, D. (In preparation). Working title: Integrated chip-seq and rna-seq data analysis to investigate regulatory mechanisms of NRF1 transcription factor on target genes.

Ramos, J., Das, J., Felty, Q., Yoo, C., Poppit, R., Murrell, D., Foster. P. J., & Roy, D. (2018). NRF1 motif sequence-enriched genes involved in ER-PR-HER2+ breast cancer signaling pathways. *Breast Cancer Research and Treatment*, *8,* 1-17.

Ramos, J., Das, J., & Roy, D. (2018, April) *Identifying the regulatory network structure of the genes involved in signaling pathways underlying ER-PR-HER2+ breast cancer using Bayesian Modelling of genome-wide NRF1 DNA motif sequence-enriched genes*. Poster presented at the 2018 Annual Meeting of the American Association for Cancer Research (AACR), Chicago, IL.

Ramos, J., Das, J., & Roy, D. (2018, March). *Pharmacological estrogen-17α-ethinyl estradiol (EE)-responsive NRF1 gene networks in human breast cancer cells: Its involvement in the carcinogenic effect of EE*. Poster presented at the 2018 Annual Meeting of the Society of Toxicology (SOT), San Antonio, TX.