PHENOTYPIC CHARACTERIZATION OF BREAST INVASIVE CARCINOMA VIA TRANSFERABLE TISSUE MORPHOMETRIC PATTERNS LEARNED FROM GLIOBLASTOMA MULTIFORME

Ju Han[1] *Gerald V. Fontenay*[2] *Yunfu Wang*[2,3] *Jian-Hua Mao*[2] *Hang Chang*[1,2]

[1] Department of Electrical and Biomedical Engineering, University of Nevada, Reno, Nevada, USA
[2] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[3]Department of Neurology, Taihe Hospital, Hubei University of Medicine, Shiyan, Hubei, China

ABSTRACT

Quantitative analysis of whole slide images (WSIs) in a large cohort may provide predictive models of clinical outcome. However, the performance of the existing techniques is hindered as a result of large technical variations (e.g., fixation, staining) and biological heterogeneities (e.g., cell type, cell state) that are always present in a large cohort. Although unsupervised feature learning provides a promising way in learning pertinent features without human intervention, its capability can be greatly limited due to the lack of well-curated examples. In this paper, we explored the transferability of knowledge acquired from a well-curated Glioblastoma Multiforme (GBM) dataset through its application to the representation and characterization of tissue histology from the Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma (BRCA) cohort. Our experimental results reveals two major phenotypic subtypes with statistically significantly different survival curves. Further differential expression analysis of these two subtypes indicates enrichment of genes regulated by NF-kB in response to TNF and genes upregulated in response to IFNG.

Index Terms— Breast invasive carcinoma, unsupervised feature learning, knowledge sharing, predictive sparse decomposition, consensus clustering, survival analysis, enrichment analysis

1. INTRODUCTION

Tumor histology provides a detailed insight into cellular morphology, organization, and heterogeneity. For example, tumor histological sections can be used to identify mitotic cells, cellular aneuploidy, and autoimmune responses. More importantly, if tumor morphology and architecture can be quantified on large histological datasets, then it will pave the way for constructing histological databases that are prognostic, the same way that genome analysis techniques have identified molecular subtypes and predictive markers. Genome-wide analysis techniques (e.g., microarray analysis and next generation sequencing (NGS)) have the advantages of standardized tools for data analysis and pathway enrichment, which enables hypothesis generation for the underlying mechanism. On the other hand, histological signatures are hard to compute because of the technical variations and biological heterogeneities in the stained histological sections; however, they offer insights into tissue composition as well as heterogeneity (e.g., mixed populations) and rare events.

Histological sections are often stained with hematoxylin and eosin stains (H&E).Traditional histological analysis is performed by a trained pathologist through the characterization of phenotypic content, such as various cell types, cellular organization, cell state and health, and cellular secretion. One of the main technical barriers for processing a large collection of histological data is that the color composition is subject to technical variations (e.g., fixation, staining) and biological heterogeneities (e.g., cell type, cell state) across histological tissue sections, especially when these tissue sections are processed and scanned at different laboratories. Here, a histological tissue section refers to an image of a thin slice of tissue applied to a microscopic slide and scanned from a light microscope. From an image analysis perspective, color variations can occur both within and across tissue sections. For example, within a tissue section, some nuclei may have low chromatin content (e.g., light blue signals), while others may have higher signals (e.g., dark blue); nuclear intensity in one tissue section may be very close to the background intensity (e.g., cytoplasmic, macromolecular components) in another tissue section.

In this paper, we aim to explore the transferability of knowledge acquired from a well-curated GBM dataset through its application to the phenotypic characterization of breast invasive carcinoma. We suggest that, unsupervised feature learning is capable of generating transferable knowledge in tissue histology that can potentially be shared across cohorts with different tumor types, which provides an effective solution when well-curated examples are not available.

Organization of this paper is as follows: Section 2 reviews related works. Section 3 describes the details of our proposed approach. Section 4 elaborates the details of our experimental setup, followed by detailed discussion on the experimental results. Lastly, section 5 concludes the paper.

2. RELATED WORK

Several outstanding reviews for the analysis of histology sections can be found in [1, 2]. From our perspective, four distinct works have defined the trends in histology image analysis: (i) one group of researchers proposed nuclear segmentation and organization for tumor grading and/or the prediction of tumor recurrence [3, 4, 5, 6]. (ii) A second group of researchers focused on patch level analysis (e.g., small regions) [7, 8, 9, 10], using color and texture features, for tumor representation. (iii) A third group focused on block-level analysis to distinguish different states of tissue development using cell-graph representation [11, 12]. (iv) Finally, a fourth group has suggested detection and representation of the auto-immune response as a prognostic tool for cancer [13].

The major challenge for computational histopathology is the large amounts of technical variations and biological heterogeneities

This work was supported by NIH R01 CA184476 carried out at Lawrence Berkeley National Laboratory.

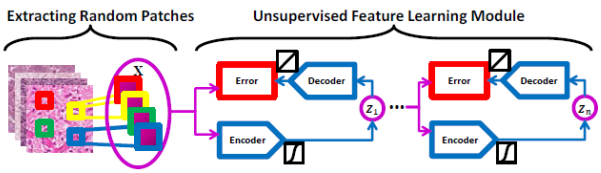

Fig. 1. Computational workflow of unsupervised feature learning with predictive sparse decomposition (PSD)

.

in the data [14], which typically results in techniques that are tumor type specific. To overcome this problem, recent studies have focused on either fine tuning human engineered features [7, 8, 14, 15], or applying automatic feature learning [16, 17, 18], for robust representation.

3. APPROACH

The proposed approach for tissue phenotypic characterization and integrated analysis includes: transferable knowledge learning through predictive sparse decomposition (PSD), tissue phenotypic representation and subtyping via consensus clustering, survival analysis and genomic association.

3.1. Transferable Knowledge Learning

Given many of the shared visual concepts among different tumor types (e.g., cell), we employed predictive sparse decomposition (PSD) [19] to learn transferable knowledge (i.e., sparse tissue morphometric patterns) from a well-curated GBM dataset [15, 17, 18], as shown in Figure 1. Unlike many other unsupervised feature learning algorithms [20, 21, 22, 23], the feed-forward feature inference of PSD is very efficient, as it involves only element-wise nonlinearity and matrix multiplication, which is crucial to the characterization and representation of large cohort of WSIs.

Given $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{m \times N}$ as a set of vectorized image patches, we formulate the PSD optimization problem as:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{Z},\mathbf{G},\mathbf{W}} \quad & \|\mathbf{X} - \mathbf{B}\mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}\|_1 + \|\mathbf{Z} - \mathbf{G}\sigma(\mathbf{W}\mathbf{X})\|_F^2 \\ \text{s.t.} \quad & \|\mathbf{b}_i\|_2^2 = 1, \forall i = 1, \ldots, h \end{aligned} \tag{1}$$

where $\mathbf{B} \in \mathbb{R}^{m \times h}$ is a set of the basis functions; $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_N] \in \mathbb{R}^{h \times N}$ is the sparse feature matrix; $\mathbf{W} \in \mathbb{R}^{h \times m}$ is the auto-encoder; $\mathbf{G} = \text{diag}(g_1, \ldots, g_h) \in \mathbb{R}^{h \times h}$ is a scaling matrix with diag being an operator aligning vector $[g_1, \ldots, g_h]$ along the diagonal, $\sigma(\cdot)$ is the element-wise sigmoid function and λ is a regularization constant. The goal of jointly minimizing Eq. (1) with respect to the quadruple $\langle \mathbf{B}, \mathbf{Z}, \mathbf{G}, \mathbf{W} \rangle$ is to enforce the inference of the nonlinear regressor $\mathbf{G}\sigma(\mathbf{W}\mathbf{X})$ to be resemble to the optimal sparse codes $\mathbf{Z}$ that can reconstruct $\mathbf{X}$ over $\mathbf{B}$ [19]. An iterative process is employed for optimizing Eq. (1), and the details can be found in our previous publication [17].

3.2. Tissue Phenotypic Representation and Subtyping

With the transferable knowledge derived from the GBM dataset through the unsupervised feature learning procedure, as shown in Figure 1, each image patch (i.e., 20×20 sub-image) in the WSI can be represented by a sparse feature vector $\mathbf{z}$. Each WSI is then represented by summarizing the feature vectors of all non-overlap, non-background and non-border patches within the WSI (e.g., moments of each individual feature dimension, etc.).

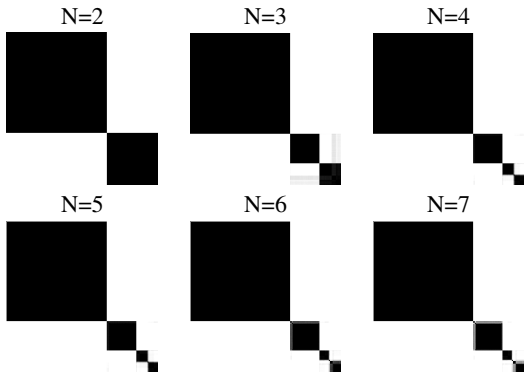

Fig. 2. Consensus clustering matrices of 273 TCGA patients with BRCA for cluster number of $N = 2$ to $N = 7$ based on tissue morphometric features.

Consensus clustering [24] is performed for identifying subtypes/clusters across tissue sections of TCGA BRCA cohort. The input of consensus clustering are the summarized features from all tissue sections. Consensus clustering aggregates consensus across multiple runs for a base clustering algorithm. Moreover, it provides a visualization tool to explore the number of clusters in the data, as well as assessing the stability of the discovered clusters.

3.3. Integrated Analysis with Genomic Signatures and Clinical Outcomes

Tissue phenotypic subtypes derived from consensus clustering of learned features are then associated with clinical outcomes, genomic/methylation subtypes and molecular data for integrated analysis. The Kaplan-Meier estimator, a non-parametric statistic, is used to estimate the survival function from clinical outcomes. Log-rank test, a nonparametric test designed for data of right skewed and censored, is used to compare the survival distributions of two subtypes. Fisher's exact test is used for the enrichment analysis between tissue phenotypic subtypes and genomic/methylation subtypes. Linear models are used for assessing differential expression of genes between tissue phenotypic subtypes.

4. RESULTS AND DISCUSSION

The proposed approach has been applied on the TCGA BRCA cohort, including 273 tissue sections from 273 patients each of which has the labels of both the 50-gene PAM50 subtypes and methylation subtypes [25]. For the quality control purpose, background and border portions of each whole slide image were detected and removed from the analysis.

4.1. Consensus clustering

The representation for each tissue section consists of a 1024-dimensional feature (average sparse feature over all non-overlap, non-background and non-border patches in the tissue section). Hierarchical clustering was chosen as the cluster algorithm for consensus clustering, where the distance function is Pearson correlation. The procedure was run for 500 iterations with a sampling rate of 0.8 on 273 tissue sections. Consensus clustering is implemented through

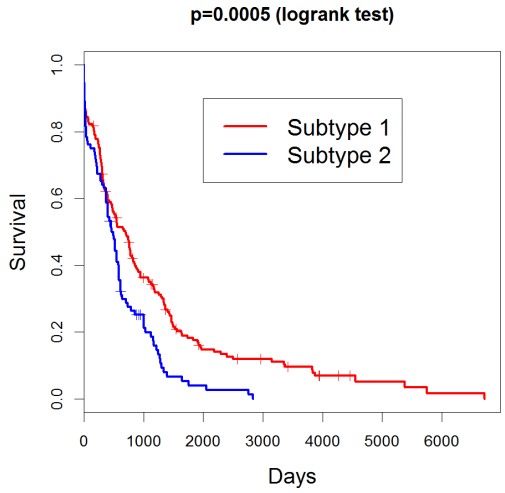

Fig. 3. Kaplan-Meier plot for the two subtypes associated with patient survival from the two-cluster consensus clustering result (181 patients in Subtype 1 and 92 patients in Subtype 2).

the R Bioconductor *ConsensusClusterPlus* package. Consensus clustering matrices with 2 to 7 clusters are shown in Figure 2, where the matrices with 2 to 4 clusters reveal different levels of similarity among tissue sections and matrices with 5 to 7 clusters provide little further details. Interestingly, the top left cluster in the matrices with 2 to 4 clusters remains the same, while the bottom right samples are further divided into sub-clusters.

4.2. Survival analysis and genomic association

Survival analysis is implemented through the R *survival* package. Figure 3 shows the Kaplan-Meier plot for two subtypes associated with patient survival from the two-cluster consensus clustering result. The log-rank test p-value of 0.0005 indicates that the difference between survival times of these two subtypes is statistically significant. Due to the short median overall follow up and the small number of overall survival events, survival analysis was not performed on the three-cluster and four-cluster consensus clustering results.

Figure 4 and 5 show tissue phenotypic subtypes and corresponding 50-gene PAM50 subtypes / methylation subtypes [25] for each tissue section in the consensus clustering results of two, three and four clusters, respectively. Fisher's exact test reveals no enrichment between tissue phenotypic subtypes and 50-gene PAM50 subtypes / methylation subtypes.

Fifty-six differential expressed genes between the two subtypes, as shown in Figure 4d, indicates enrichment of genes regulated by NF-kB in response to TNF and genes up-regulated in response to IFNG (via MSigDB [26]). TNF refers to a group of cytokines that induce proliferation, and inflammation and apoptosis depending upon the adaptor proteins. It is shown that TNF-α acting on TNFR1 promotes breast cancer growth via p42/P44 MAPK, JNK, Akt and NF-kB-dependent pathways [27]. IFNG is an inflammatory cytokine that induces the expression and function of IRF1, a tumor suppressor gene that can increase antiestrogen responsiveness. Observations also support the exploration of clinical trials combining antiestrogens and compounds that can induce IRF1 for the treatment of some ER-positive breast cancers [28].

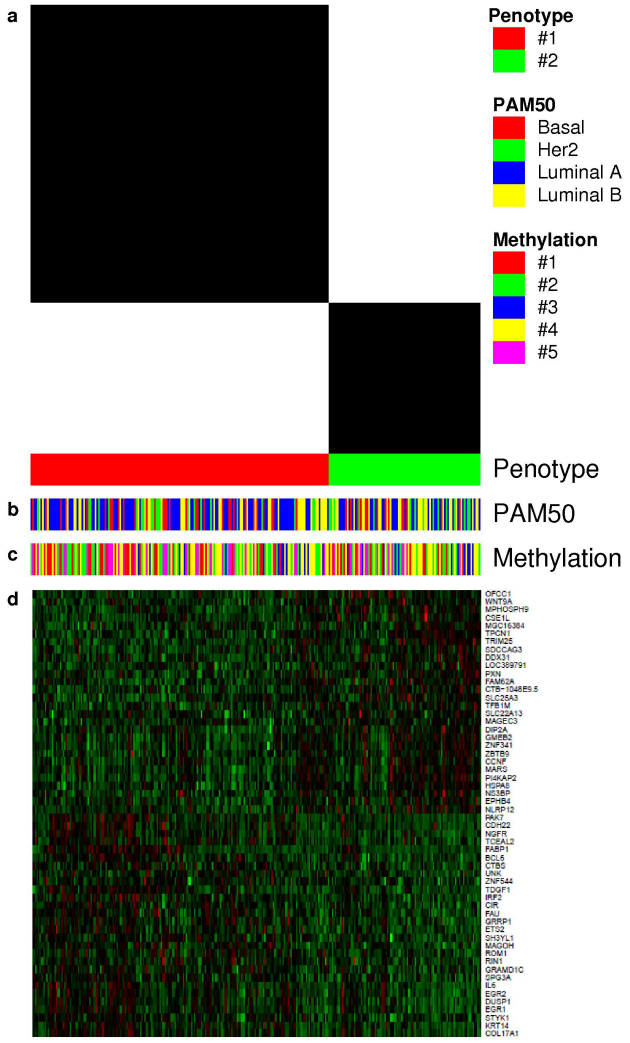

Fig. 4. Coordinated analysis for two-cluster consensus clustering result: a. Phenotypic subtypes; b. 50-gene PAM50 subtypes [25]; c. Methylation subtypes [25]; d. Genes that are differently expressed between the two phenotypic subtypes (FDR-adjusted p-value<0.05).

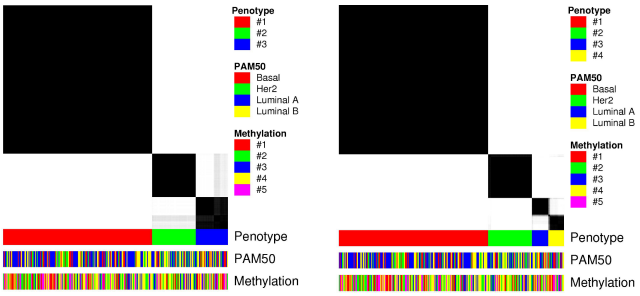

Fig. 5. Coordinated analysis for three-cluster and four-cluster consensus clustering results.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a knowledge sharing approach based on unsupervised feature learning (i.e., predictive sparse decomposition) for tissue phenotypic characterization, followed by phenotypic subtyping and genomic and clinical association. The knowledge (i.e., sparse tissue morphometric patterns) was initially learned from a well-curated GBM dataset, and then transferred to the TCGA BRCA cohort. Experimental results indicate no enrichment between the tissue phenotypic subtypes and 50-gene PAM50 subtypes / methylation subtypes. Instead, they reveal two major phenotypic subtypes with statistically significantly different survival curves. Further differential expression analysis of these two subtypes indicates enrichment of genes regulated by NF-kB in response to TNF and genes up-regulated in response to IFNG.

We suggest that such an approach can be potentially applied for many other biomedical applications when well-curated examples are not easily available. And our future work will focus on (i) validating our findings on breast invasive carcinoma with an independent cohort; and (ii) validating our approach as well as the transferability of the pre-built knowledge (i.e., sparse tissue morphometric patterns) with other biomedical applications, such as the study of biological responses to environmental challenges.

6. REFERENCES

[1] C. Demir and B. Yener, "Automated cancer diagnosis based on histopathological images: A systematic survey," *Technical Report, Rensselaer Polytechnic Institute, Department of Computer Science.*, 2009.

[2] M. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, NM Rajpoot, and Y. Bulent, "Histopathological image analysis: a review," *IEEE Transactions on Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[3] D. Axelrod, N. Miller, H. Lickley, J. Qian, W. Christens-Barry, Y. Yuan, Y. Fu, and J. Chapman, "Effect of quantitative nuclear features on recurrence of ductal carcinoma in situ (DCIS) of breast," *Cancer Informatics*, vol. 4, pp. 99–109, 2008.

[4] M. Datar, D. Padfield, and H. Cline, "Color and texture based segmentation of molecular pathology images using HSOMs," in *ISBI*, 2008, pp. 292–295.

[5] A. Basavanhally, J. Xu, A. Madabhushu, and S. Ganesan, "Computer-aided prognosis of ER+ breast cancer histopathology and correlating survival outcome with oncotype DX assay," in *ISBI*, 2009, pp. 851–854.

[6] S. Doyle, M. Feldman, J. Tomaszewski, N. Shih, and A. Madabhushu, "Cascaded multi-class pairwise classifier (CASCAMPA) for normal, cancerous, and cancer confounder classes in prostate histology," in *ISBI*, 2011, pp. 715–718.

[7] R. Bhagavatula, M. Fickus, W. Kelly, C. Guo, J. Ozolek, C. Castro, and J. Kovacevic, "Automatic identification and delineation of germ layer components in $h\&e$ stained images of teratomas derived from human and nonhuman primate embryonic stem cells," in *ISBI*, 2010, pp. 1041–1044.

[8] J. Kong, L. Cooper, A. Sharma, T. Kurk, D. Brat, and J. Saltz, "Texture based image recognition in microscopy images of diffuse gliomas with multi-class gentle boosting mechanism," in *ICASSAP*, 2010, pp. 457–460.

[9] J. Han, H. Chang, L. Loss, K. Zhang, FL Baehner, JW Gray, PT Spellman, and Bahram Parvin, "Comparison of sparse coding and kernel methods for histopathological classification of glioblastoma multiforme," in *ISBI*, 2011, pp. 711–714.

[10] A. Mujahid Khan, K. Sirinukunwattana, and N.M. Rajpoot, "A global covariance descriptor for nuclear atypia scoring in breast histopathology images," *IEEE J. Biomedical and Health Informatics*, vol. 19, no. 5, pp. 1637–1647, 2015.

[11] E. Acar, G.E. Plopper, and B. Yener, "Coupled analysis of in vitro and histology samples to quantify structure-function relationships," *PLoS One*, vol. 7, no. 3, pp. e32227, 2012.

[12] C.C. Bilgin, S. Ray, B. Baydil, W.P. Daley, M. Larsen, and B. Yener, "Multiscale feature analysis of salivary gland branching morphogenesis," *PLoS One*, vol. 7, no. 3, pp. e32906, 2012.

[13] H. Fatakdawala, J. Xu, A. Basavanhally, G. Bhanot, S. Ganesan, F. Feldman, J. Tomaszewski, and A. Madabhushi, "Expectation-maximization-driven geodesic active contours with overlap resolution (EMaGACOR): Application to lymphocyte segmentation on breast cancer histopathology," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1676–1690, 2010.

[14] S. Kothari, J.H. Phan, A.O. Osunkoya, and M.D. Wang, "Biological interpretation of morphological patterns in histopathological whole slide images," in *ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 2012.

[15] H. Chang, A. Borowsky, P.T. Spellman, and B. Parvin, "Classification of tumor histology via morphometric context," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2013.

[16] C.H. Huang, A. Veillard, N. Lomeine, D. Racoceanu, and L. Roux, "Time efficient sparse analysis of histopathological whole slide images," *Computerized medical imaging and graphics*, vol. 35, no. 7-8, pp. 579–591, 2011.

[17] H. Chang, Y. Zhou, A. Borowsky, K.E. Barner, P.T. Spellman, and B. Parvin, "Stacked predictive sparse decomposition for classification of histology sections," *IJCV*, vol. 113, no. 1, pp. 3–18, 2015.

[18] Y. Zhou, H. Chang, K.E. Barner, P.T. Spellman, and B. Parvin, "Classification of histology sections via multispectral convolutional sparse coding," in *CVPR*, 2014, pp. 3081–3088.

[19] K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "Fast inference in sparse coding algorithms with applications to object recognition," Tech. Rep. CBLL-TR-2008-12-01, Computational and Biological Learning Lab, Courant Institute, NYU, 2008.

[20] H. Lee, A. Battle, R. Raina, and A.Y. Ng, "Efficient sparse coding algorithms," in *NIPS*, 2007, pp. 801–808.

[21] H. Lee, C. Ekanadham, and A.Y. Ng, "Sparse deep belief net model for visual area v2," in *Advances in Neural Information Processing Systems 20*. 2008, MIT Press.

[22] C. Poultney, S. Chopra, and Y. Lecun, "Efficient learning of sparse representations with an energy-based model," in *Advances in Neural Information Processing Systems (NIPS 2006*. 2006, MIT Press.

[23] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., 2009, pp. 2223–2231.

[24] S. Monti, P. Tamayo, J. Mesirov, and T.R. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Mach Learn*, vol. 52, pp. 91–118, 2003.

[25] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumors," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[26] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich A, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *PNAS*, vol. 102, no. 43, pp. 15545–50, 2005.

[27] M.A. Rivas, R.P. Carnevale, C.J. Proietti, C. Rosemblit, W. Beguelin, M. Salatino, E.H. Charreau, I. Frahm, S. Sapia, P. Brouckaert, P.V. Elizalde, and R. Schillaci, "Tnf alpha acting on tnfr1 promotes breast cancer growth via p42/p44 mapk, jnk, akt and nf-kappa b-dependent pathways," *Exp Cell Res*, vol. 314, no. 3, pp. 509–29, 2008.

[28] Y. Ning, R.B. Riggins, J.E. Mulla, H. Chung, A. Zwart, and R. Clarke, "Ifngamma restores breast cancer sensitivity to fulvestrant by regulating stat1, ifn regulatory factor 1, nf-kappab, bcl2 family members, and signaling to caspase-dependent apoptosis," *Mol Cancer Ther*, vol. 9, no. 5, pp. 1274–85, 2010.