Bioinformatics analyses to identify molecular gene signatures associated with breast cancer phenotypes

Federica Conte[1], Giulia Fiscon[1,2] and Paola Paci[1,2,3]
[1] *Institute for Systems Analysis and Computer Science "A. Ruberti" (IASI) National Research Council (CNR) of Rome, Rome, Italy*
[2] *Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy*
[3] *Karolinska Institutet, 17177 Stockholm, Sweden*

***Abstract*— Breast cancer is a heterogeneous and complex disease as witnessed by the existence of different subtypes with distinct morphologies and clinical implications. Despite the remarkable advances in understanding the mechanisms underlying breast cancer, this disease is still a major public health problem worldwide and poses significant open challenges. Here, we show how a multi-omics data integration analysis may provide useful insights in the identification of promising molecular signatures associated with the different breast cancer subtypes.**

***Keywords*— breast cancer subtypes, bioinformatics, computational medicine, gene signature**

I. Introduction

BREAST CANCER (BC) is the most common female cancer and, despite important advances in early detection and research development, it continues to be the second leading cause of death in women worldwide [1]. BC is a heterogeneous pathology as witnessed by the existence of different subtypes with distinct morphologies and clinical implications [2]. These subtypes are usually defined by using the immunohistochemical (IHC) classification, which is based on immunoprofile (i.e., Estrogen Receptor-ER, Progesterone Receptor-PR and Human Epidermal growth factor Receptor2-HER2 status) and Ki67 index [3]; or by using the genetic (PAM50) classification, based on the expression of a 50-gene signature [4], [5]. According to the IHC classification, breast cancers are stratified into four subtypes, i.e. luminal A, luminal B, Her2 positive and triple negative. According to the PAM50 classification, they are classified into four molecular intrinsic subtypes that are luminal A, luminal B, Her2 positive and basal-like. The most aggressive BC pathophenotypes are the triple-negative and the basal-like for the two classifications, respectively. Although triple-negative and basal-like are not the same by definition, these two terms are often used interchangeably, since most basal-like breast cancers also have a triple-negative phenotype (i.e., ER-negative, PR-negative, HER2-negative) [6] and since triple-negative breast cancers overlap with the molecular entity of basal-like [7], [8]. The identification of BC-associated biomarkers and the development of effective therapeutic strategies for the different subtypes are ongoing challenges to be faced.

Here, we present a bioinformatics approach for the integration of multi-omics data (e.g., transcriptomics, genomics, epigenomics, clinical) that can offer promising insights for understanding BC-related molecular mechanisms.

II. Materials and Methods

A. Data retrieval

Transcriptomic, clinicopathological, Copy Number Variations (CNVs) and DNA methylation data of patients affected by breast cancer were retrieved from The Cancer Genome Atlas (TCGA) repository [9]. Male samples, as well as samples undergoing a neoadjuvant treatment, were removed from the TCGA-BC cohort. The Human Protein Atlas (HPA) website [10] was leveraged to retrieve immunohistochemistry images and to evaluate changes in the proteins expression patterns.

B. SWIM algorithm

SWIM (SWItch Miner) is a recently developed tool able to unveil key (switch) genes within gene co-expression networks strongly associated with drastic changes in cell/tissue phenotype [11], [12]. SWIM algorithm has been implemented both in Matlab and in R language and encompasses a series of steps described in details in [11], [12].

C. In vitro and ex vivo experiments

To validate the common gene signature found to be altered in all BC subtypes, in vitro and ex vivo experiments were performed by using BC model cell lines and tissue specimens. The detailed description of these experiments was provided in [13].

D. Survival analysis

To analyze the correlation between the expression level of the basal-like specific switch genes and patient overall survival and therefore to evaluate their prognostic value, we used the RNA-sequencing data from TCGA to split the entire cohort of BC patients (1049 samples) into two groups (called low-expression and high-expression group) according to the upper and lower expression quartile. Low- and high-expression groups refer to patients with expression levels of the given switch gene lower and greater than the 50th percentile (i.e., median), respectively. For each patient cohort, the cumulative survival rates were computed for each switch gene according to the Kaplan-Meier (KM) method [14] on the clinical metadata provided by TCGA. For each switch gene, the survival outcomes of the two patients' groups were compared by the log-rank test. Switch genes with log-rank p-values less than 0.05 were suggested as candidate prognostic biomarkers.

E. Gene regulatory network analysis

To investigate putative transcription factor activities on the regulation of the basal-like specific switch genes, we built a gene regulatory network by integrating information from Pscan web tool [15], TRRUST database [16] and the human interactome assembled by Cheng and coauthors [17]. In particular, we firstly exploited Pscan web tool to predict TFs putatively able to bind the promoter regions of the selected switch genes. Then, we filtered the Pscan predictions keeping only the TFs known to physically interact with at least one switch genes in the human interactome. These TF-target relationships were finally complemented with those experimentally validated from TRRUST database.

III. Results and Discussion

A. Identification of a common gene signature altered in all BC subtypes

In our recent study [13], we applied SWIM methodology on the transcriptomic data of TCGA-BC patients stratified according to two different subtype classification (i.e., PAM50 and IHC classification) in order to identify switch genes both shared among all subtypes and specific for each subtype (Fig. 1). We focused on switch genes common to all subtypes of both classifications and we found a common gene signature composed of 11 genes.

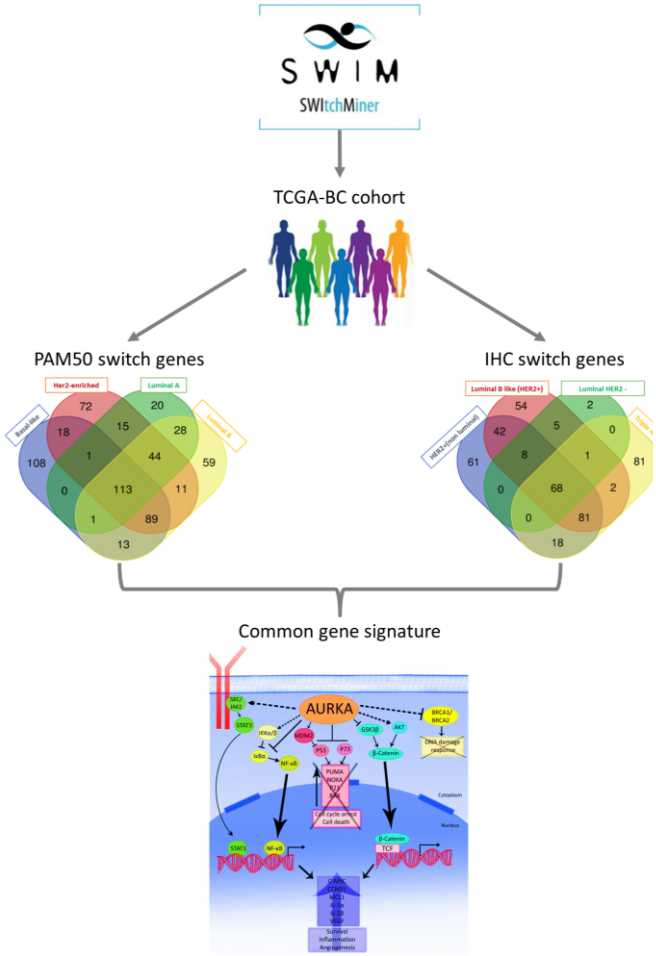

Fig. 1: Workflow for the identification of a common gene signature altered in all BC subtypes

Among them, Aurora Kinase A (AURKA) appears to be very promising since: i) it is a kinase known to have a key role in cell division and cell-cycle progression; ii) it plays a critical role in regulation of mitotic events like spindle assembly and chromosomal segregation; iii) it is deregulated in many human cancers; iv) it collaborates with several tumor suppressors like P53, BRCA1 and BRCA2; v) it is suggested as a pharmaceutical target for the treatment of various cancers [18].

The key role of AURKA in the context of BC subtypes was experimentally validated by showing that the encoded protein is always over-expressed both in BC cell lines (Fig.2A) and tissues (Fig.2B).

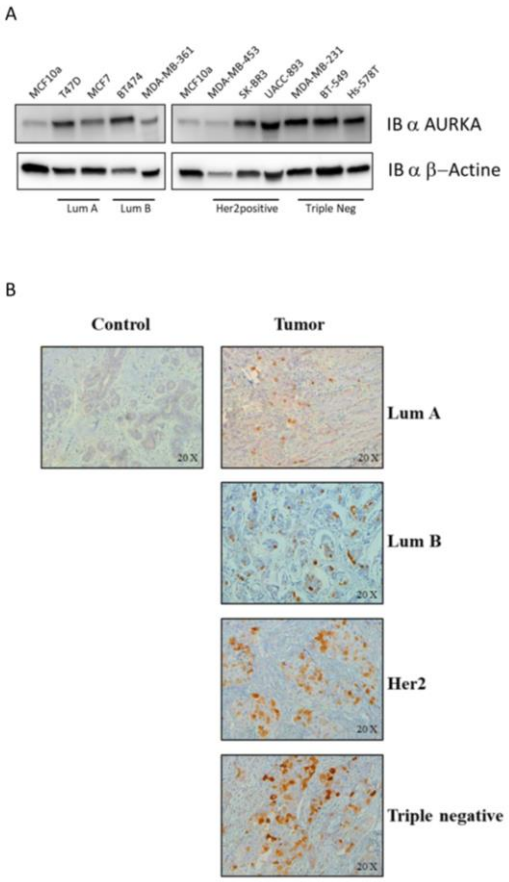

Fig. 2: AURKA protein expression examined in vitro by immunoblotting on BC subtype cell lines (A) and ex vivo by immunohistochemical analysis on surgical BC tissue specimens (B).

Moreover, we demonstrated that the AURKA inhibition by using alisertib drug led to a reduction in the cell growth of all BC subtype cell lines (Fig.3A) and to arrest their cell cycle (Fig.3B) up to 72 h after the treatment.

Taken together, all these findings supported the hypothesis that AURKA pathway could be a common mechanism univocally altered in all BC subtypes.

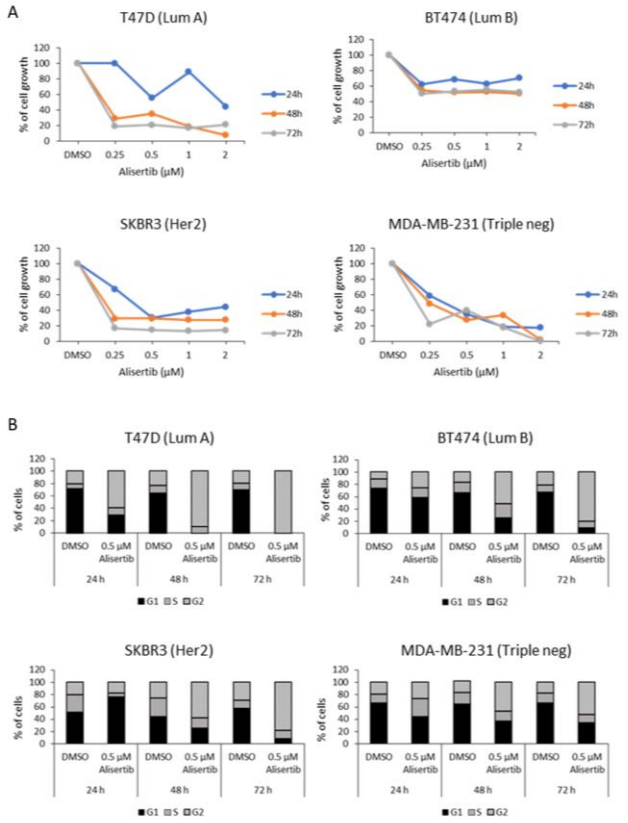

Fig. 3: Effect of AURKA's inhibitor (alisertib) assessed on cell growth (A) and cell cycle (B) of BC subtype cell lines.

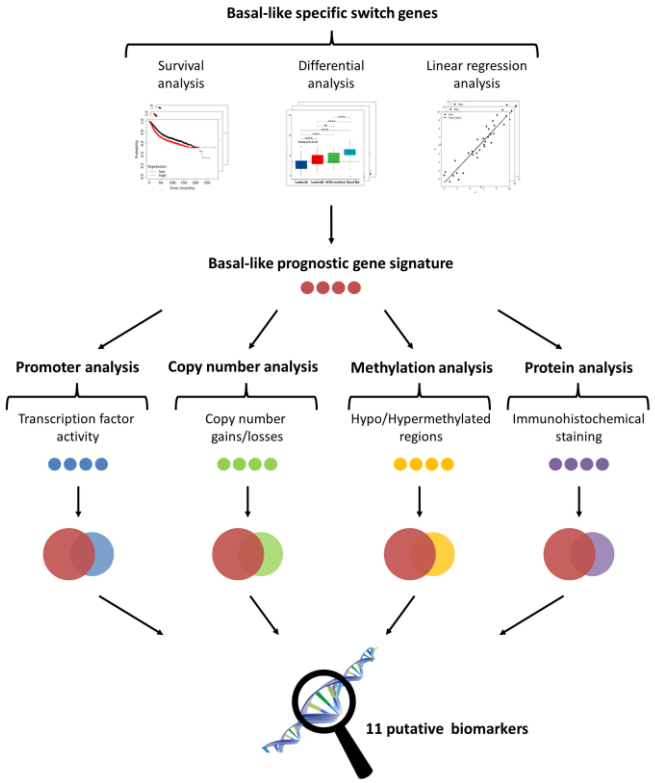

Fig. 4: Workflow for the identification of a prognostic gene signature for basal-like subtype.

B. Identification of a prognostic gene signature for basal-like subtype

Among the various recognized BC subtypes, basal-like is the most aggressive with a poor prognosis, a high risk of relapse and a high resistance to pharmacological treatments [19]. Currently, there are not available widely accepted prognostic biomarkers to predict the outcomes of basal-like patients. Therefore, in our recent study [20], we carried out a bioinformatics pipeline integrating transcriptomic, genomic, epigenomic, and clinical data in order to identify a prognostic gene signature for the basal-like subtype (Fig.4).

Our pipeline started from the 108 basal-like specific switch genes identified in [13] and performed a Kaplan-Meier (KM) survival analysis in order to explore their clinical relevance with respect to BC patients' overall survival (OS). This first step allowed us to identify 11 basal-like specific switch genes, i.e., CENPN, LRP8, DSCC1, CTPS, RCOR2, GINS4, TUBA1C, PRAME, SLC7A11, CDCA7, GSDMC, acting as unfavourable prognostic biomarkers (i.e., their higher expression was significantly associated with poorer BC patients' OS). The clinical relevance of these 11 switch genes was also confirmed using other larger BC datasets collected in the KM plotter website [21].

Next, we evaluated the gene expression of these 11 putative basal-like prognostic biomarkers demonstrating that they always reached the highest level in the basal-like condition.

To statistically quantify the increasing trend of their expression values as the phenotype varies from physiological to pathological condition passing across the different BC subtypes, we performed a linear regression analysis, where the index R-squared estimates the goodness-of-fit. We found that all of them showed a very strong straight-line relationship (R-squared >= 0.7) with respect to the tumor aggressiveness. These results were mostly confirmed also by performing the same analysis with respect to the pathological staging of the BC patients.

Moreover, we explored the expression patterns of the proteins encoded by the 11 prognostic switch genes through the HPA. We found that six of these proteins were overexpressed in BC tissues compared to normal breast tissues. For the remaining ones, there are pending cancer and normal tissue analysis on the HPA and the immunohistochemistry images are not currently available.

Eventually, we investigated if the overexpression of the 11 basal-like prognostic biomarkers may depend on the action of important transcription factors (TFs) as well as basal-like specific genomic alterations (CNVs) and/or epigenomic alteration (DNA methylation changes). In particular, to provide some hints on which TFs could regulate the expression of the 11 basal-like switch genes, we built a gene regulatory network by combining information on both computationally predicted and experimentally validated TF-target relationships (see Materials and Methods). The final gene regulatory network was composed of seven switch genes and twelve TFs, including well-known TFs that, if deregulated, contribute to neoplastic transformation as MYC, TP53 and NFKB. By performing hierarchical clustering analyses, we highlighted different CNVs (copy number gain/loss) and DNA methylation (hypo/hypermethylated

regions) status of the 11 putative prognostic biomarkers in basal-like subtype with respect to the less aggressive BC subtype, i.e., luminal A [20].

Taken together, all these findings prompted us to propose the 11 basal-like specific switch genes as a specific gene signature to evaluate the prognosis of basal-like BC patients.

IV. Conclusion

In the present work, we showed how bioinformatics may be exploited to provide a contribution in the identification of putative gene signatures associated with one of the most heterogeneous and widespread disease, i.e. breast cancer, thus helping in the discovery of effective therapies as well as prognostic biomarkers.

Acknowledgement

This work has been partially funded by BiBiNet project (grant number: H35F21000430002) within the POR-Lazio FESR 2014-2020 and by Progetto di Ricerca di Ateneo 2021 (grant n: RM12117A34663A2C).

References

[1] J. Ferlay *et al.*, "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods," *Int. J. Cancer*, vol. 144, no. 8, pp. 1941–1953, 2019, doi: 10.1002/ijc.31937.

[2] Z. D, I. A, S. C, and P. M, "Clinical management of breast cancer heterogeneity," *Nat. Rev. Clin. Oncol.*, vol. 12, no. 7, Jul. 2015, doi: 10.1038/nrclinonc.2015.73.

[3] A. Goldhirsch *et al.*, "Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013," *Ann. Oncol.*, vol. 24, no. 9, pp. 2206–2223, Sep. 2013, doi: 10.1093/annonc/mdt303.

[4] D. C. Koboldt *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, Art. no. 7418, Oct. 2012, doi: 10.1038/nature11412.

[5] J. S. Parker *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 27, no. 8, Art. no. 8, Mar. 2009, doi: 10.1200/JCO.2008.18.1370.

[6] A. C. Garrido-Castro, N. U. Lin, and K. Polyak, "Insights into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for Treatment," *Cancer Discov.*, vol. 9, no. 2, pp. 176–198, Feb. 2019, doi: 10.1158/2159-8290.CD-18-1177.

[7] P. Alluri and L. A. Newman, "Basal-like and triple-negative breast cancers: searching for positives among many negatives," *Surg. Oncol. Clin. N. Am.*, vol. 23, no. 3, pp. 567–577, Jul. 2014, doi: 10.1016/j.soc.2014.03.003.

[8] P. Gazinska *et al.*, "Comparison of basal-like triple-negative breast cancer defined by morphology, immunohistochemistry and transcriptional profiles," *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc*, vol. 26, no. 7, pp. 955–966, Jul. 2013, doi: 10.1038/modpathol.2012.244.

[9] K. Tomczak, P. Czerwinska, M. Wiznerowicz, and others, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemp Oncol Pozn*, vol. 19, no. 1A, pp. A68–A77, 2015.

[10] A. Asplund, P.-H. D. Edqvist, J. M. Schwenk, and F. Pontén, "Antibodies for profiling the human proteome-The Human Protein Atlas as a resource for cancer research," *Proteomics*, vol. 12, no. 13, pp. 2067–2077, Jul. 2012, doi: 10.1002/pmic.201100504.

[11] P. Paci, T. Colombo, G. Fiscon, A. Gurtner, G. Pavesi, and L. Farina, "SWIM: a computational tool to unveiling crucial nodes in complex biological networks," *Sci. Rep.*, vol. 7, p. srep44797, Mar. 2017, doi: 10.1038/srep44797.

[12] P. Paci and G. Fiscon, "SWIMmeR: an R-based software to unveiling crucial nodes in complex biological networks," *Bioinformatics*, vol. 38, no. 2, pp. 586–588, Jan. 2022, doi: 10.1093/bioinformatics/btab657.

[13] A. M. Grimaldi *et al.*, "The New Paradigm of Network Medicine to Analyze Breast Cancer Phenotypes," *Int. J. Mol. Sci.*, vol. 21, no. 18, Art. no. 18, Jan. 2020, doi: 10.3390/ijms21186690.

[14] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. J. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding Kaplan-Meier curves," *Otolaryngol. Neck Surg.*, vol. 143, no. 3, pp. 331–336, Sep. 2010, doi: 10.1016/j.otohns.2010.05.007.

[15] F. Zambelli, G. Pesole, and G. Pavesi, "Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes," *Nucleic Acids Res.*, vol. 37, no. suppl_2, pp. W247–W252, Jul. 2009, doi: 10.1093/nar/gkp464.

[16] H. Han *et al.*, "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D380–D386, Jan. 2018, doi: 10.1093/nar/gkx1013.

[17] F. Cheng *et al.*, "Network-based approach to prediction and population-based validation of in silico drug repurposing," *Nat. Commun.*, vol. 9, no. 1, Art. no. 1, Jul. 2018, doi: 10.1038/s41467-018-05116-5.

[18] R. Du, C. Huang, K. Liu, X. Li, and Z. Dong, "Targeting AURKA in Cancer: molecular mechanisms and opportunities for Cancer therapy," *Mol. Cancer*, vol. 20, no. 1, p. 15, Jan. 2021, doi: 10.1186/s12943-020-01305-3.

[19] N. U. Lin *et al.*, "Clinicopathologic features, patterns of recurrence, and survival among women with triple-negative breast cancer in the National Comprehensive Cancer Network," *Cancer*, vol. 118, no. 22, pp. 5463–5472, 2012, doi: 10.1002/cncr.27581.

[20] F. Conte, P. Sibilio, A. M. Grimaldi, M. Salvatore, P. Paci, and M. Incoronato, "In silico recognition of a prognostic signature in basal-like breast cancer patients," *PLOS ONE*, vol. 17, no. 2, p. e0264024, Feb. 2022, doi: 10.1371/journal.pone.0264024.

[21] B. Győrffy *et al.*, "An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients," *Breast Cancer Res. Treat.*, vol. 123, no. 3, pp. 725–731, Oct. 2010, doi: 10.1007/s10549-009-0674-9.