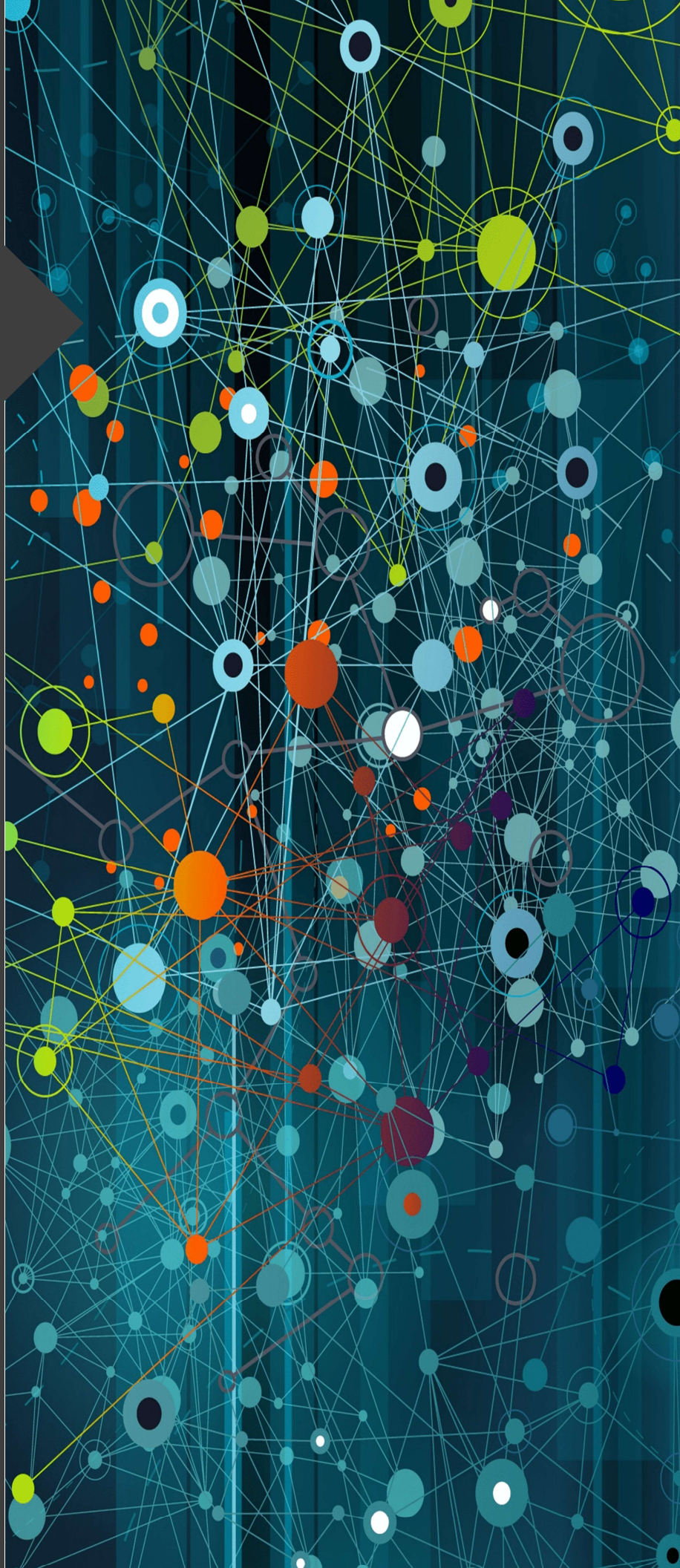
PATTERN COMPUTER®

ON THE ROAD TO PERSONALIZED MEDICINE: DISCOVERY OF PROGNOSTIC COMBINATORIAL HIGH-ORDER INTERACTIONS IN BREAST CANCER

Nidhi Singh,[1,4] Meenakshi Venkatasubramanian,[1,4] Irshad Mohammed,[1] Michael Dushkoff,[1] Ben Brown[2-4]

[1]Pattern Computer Inc., 38 Yew Lane, Friday Harbor, WA 98250.
[2]Statistics Department, University of California, Berkeley, CA 94720.
[3]Centre for Computational Biology, School of Biosciences, University of Birmingham, Edgbaston B15 2TT, United Kingdom.
[4]Molecular Ecosystems Biology Department, Biosciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.



Pattern Computer Inc.
38 Yew Lane, Friday Harbor, WA 98250.
USA

Introduction

Decades of research has demonstrated that breast cancer is a heterogenous complex of diseases with distinct biological features and clinical outcomes. Genome-wide association studies (GWAS) have successfully identified variants associated with disease [1] but of the 46 known drug targets, only one has been discovered through GWAS. Indeed, GWAS genes rarely constitute actionable intelligence. This is because such studies provide only a parts list – they don't indicate how genes work together to effect outcomes.

Disruptive advances in machine learning and computing enable fundamentally new types of genetic and genomic studies – where we search for important aspects of genomic architecture; for pathways, or relationships between pathways, rather than individual genes. We move beyond lists of parts, we learn how the parts assemble into the machine – form and function.

Previously, such studies have been frustrated by the "**curse of dimensionality**" – the fact that searching for collections of variants or genes that exhibit signatures of interactions requires the exploration of an intractably large space. Current methods using statistics to assess the effects of pairs of variants requires conducting $2x10^{13}$ tests. With triplets that's up to 10^{19}, and quadruplets would require over 300M hours on largest supercomputers in North America.

With new tools, we can search for interactions of any form or order at the same computational cost as individual variants. We can map response surfaces, and use these to understand relationships between, for instance, the expression levels of collections of genes and clinical outcomes. We are working to improve diagnosis and prognosis to develop individualized therapy recommendation systems and to identify new actionable therapeutic targets. Further, in our learning framework, these goals are all interlinked: our learning machines are transparent – **prognostic panels are not black boxes** – users can explore the joint effects of genetic variants or changes in gene expression. Viewing cancer through the lens of genomic landscapes, rather than individual genes, variants, or quantitative trait loci (QTLs) may help us better understand cancer biology and to develop new, more personalized therapeutic strategies.

Objective

Our goal is to identify novel genes and gene interactions specific to individual breast cancer subtypes that can serve as potential target(s) for developing more effective, personalized treatment options for combating breast cancer. The extent to which genetic background and genomic context is important to oncogenesis has remained opaque. We provide a new view of the genomic landscape of cancer, and conclude that modeling interactions between genes is a valuable step toward accurate prognostics and the rational development of therapeutic strategies.

Using publicly available gene expression datasets and our cutting-edge machine learning tools, we generated: (1) novel gene panels that are capable of accurate prognosis and subtype identification, and (2) a "hypothesis generator" for the identification of higher-order gene-gene interactions within subtypes. We illustrate the power of these approaches in a few case-studies. Follow-on studies will focus on the validation of our findings in pre-clinical models.

"We have demonstrated the capacity of our algorithms to learn 6^{th} order interactions in a search space larger than 10^{22} at the same computational cost as the identification of individual genes."

Prognostic Gene Panels: Subtype & Risk Classification

The first step to accomplishing our goal was the development of better, more accurate and robust multivariate prediction models for the identification of biomarkers. Our aim is to simultaneously classify tumors by their molecular subtypes and also to provide accurate identification of patients with low-risk versus high-risk disease-states to inform treatment decisions. Figure 1 outlines our workflow to design and develop predictive classifiers.

Using our feature-selection engine, high-dimensional genomic datasets were reduced from around 20,000 features (genes) to the order of 10s of genes. Multiple gene panels were derived using our proprietary machine learning tools, which enabled the identification of the top-weighted genes that, together, reproducibly identify subtype and survival. This was followed by retraining the calibration engine with gene panels with varying numbers of genes to enhance predictive power. The overall accuracy for the calibrated model (Pattern BC38) was then evaluated at approximately 90%, Fig. 2. We predict that accuracy will be further improved by repeated testing of tumor sub-samples – under a Bayesian model, 99% accuracy is obtainable after testing in only biological triplicate.

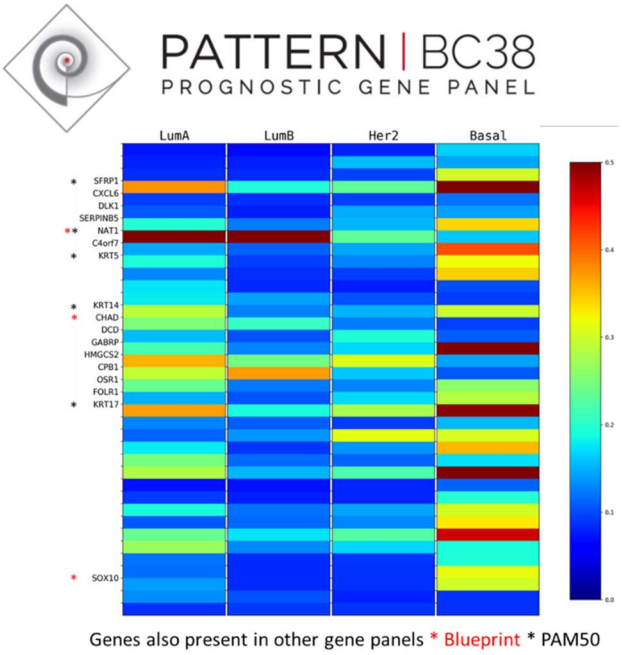


Figure 2a. The Pattern BC38 gene panel for breast cancer subtype and survival classification. The bar next to it shows expression levels from low (blue) to high (red). Redacted gene references represents proprietary PCI content.

The top 6 genes account for 95.5% of the variability of the Pattern BC38, prompting us to study a reduced six-gene panel, Pattern BC06 shown in Figs. 2 and 3. This panel provides adequate classification for both subtype and survival with fewer genes in a robust, and cost-efficient manner.

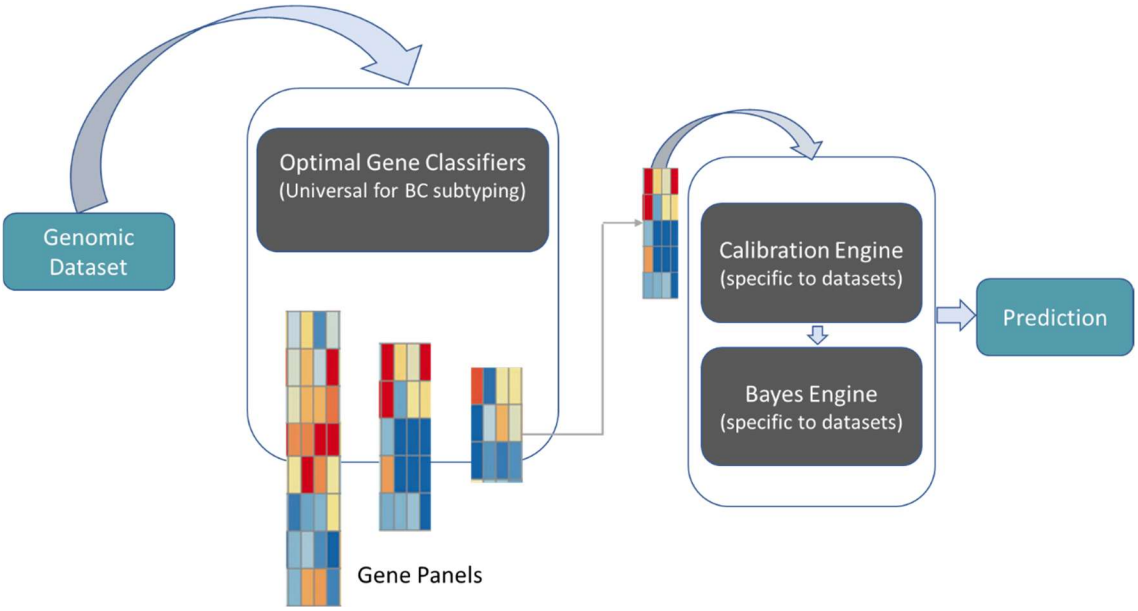


Figure 1. An outline of the approach to design classifying gene panels using biomarker classifier.

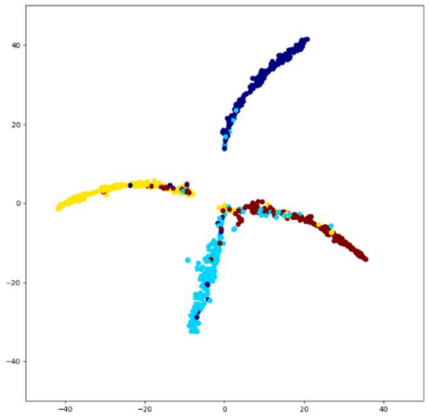


Figure 2b. A 2D representation of breast cancer subtypes generated using t-SNE dimensional reduction technique.

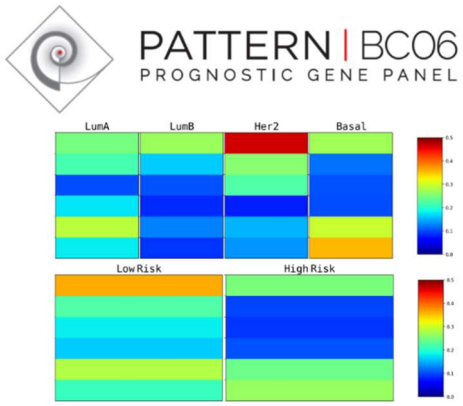


Figure 3. The Pattern BC06 gene panel for breast cancer subtype and survival classification. Redacted gene references represent proprietary PCI content.

Finally, the performance of our panel to assign the same tumor to the same subtype was assessed on external, independent breast cancer datasets.

It was found that the simplified gene panel had an overall **prediction accuracy of ~86%** for test samples, which we project will obtain >99% accuracy after testing in biological quadruplicate.

High-Order Interaction Detection

Using our proprietary algorithms built into our "Pattern Discovery Engine™", our next step was to attempt to map the gene expression architecture that underlies disease risk in human-navigable representations. Fig. 4 provides an outline of how the Pattern Discovery Engine™ works.

Briefly, large genomic datasets are ingested by the dimensionality reduction engine that reduces its size to the order of 10s of genes. This is followed by feature discovery, selection and consolidation to learn high-order interactions that correspond to testable hypotheses at the basis of disease progression. Finally, based on their respective statistical scores and generated probability cubes, a handful of interactions are selected for further biological investigation.

Methods exist for identifying two-way relationships or predefined (hypothesis-based) high-order interactions, and many "black box" machine learning architectures take advantage of complex interactions but extracting them for human exploration and hypothesis generation

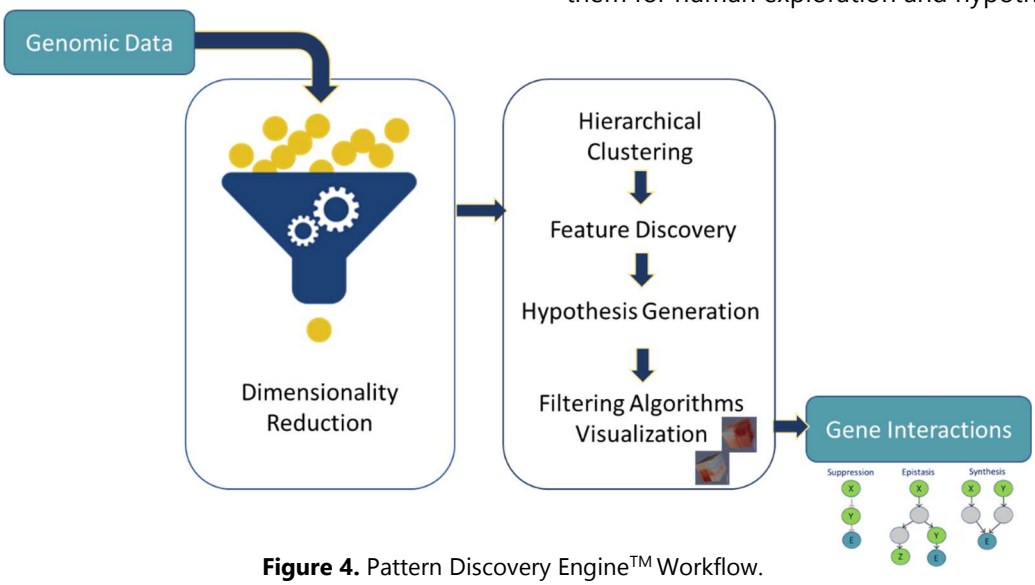


Figure 4. Pattern Discovery Engine™ Workflow.

remains a foundational challenge for the field. "Open box" procedures, like forward regression, become computationally prohibitive for even relatively small datasets. This is where our system shines:

- We have demonstrated the capacity of our algorithms to learn 6th order interactions in a search space larger than 10^{22} at the same computational cost as the identification of individual genes.

This presents a substantial advantage over existing approaches and uniquely places our technologies for the discovery of complex, nonlinear interactions permitting inquiry into the high-order mechanisms underlying functional regulation.

To explore the utility of our engine for pattern discovery, we present a three-way gene interaction between BUB1, FOXM1 and CHEK1 identified from among the high-risk group within the basal subtype of breast cancer. We present the architecture of the association between these three genes and disease prognosis as a "probability cube" for visualization. The probability cube describing this gene-gene interaction represents the relationship of the expression levels of these genes to survival. Here we see that high expression of all three genes is indicative of poor prognosis (high risk, Fig. 5).

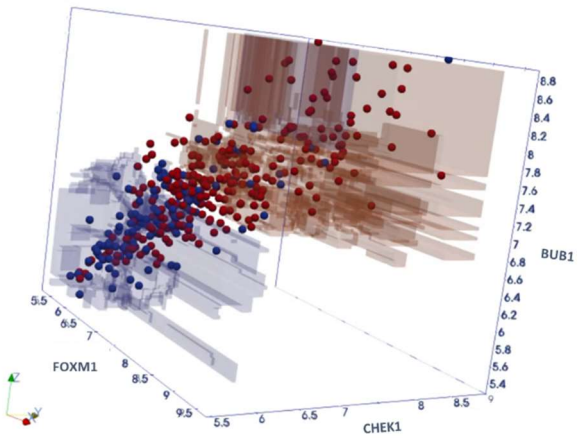


Figure 5. The probability cubes showing relationship between expression levels of FOXM1, BUB1 and CHEK1 with respect to survival. The blue and red colored areas represent regions of low risk (low mortality) and high risk (high mortality) for breast cancer of the basal subtype.

This is further evidenced by the Kaplan-Meier curve that shows the collective ability of the three genes to predict overall survival with high statistical significance (Cox p = 0.0022, log rank test; Fig. 6a). We further plotted the correlation of BUB1 and CHEK1 as a function of the expression levels of FOXM1. Based on Fig. 6b, we hypothesized that FOXM1 may act as a regulator of CHEK1 and BUB1.

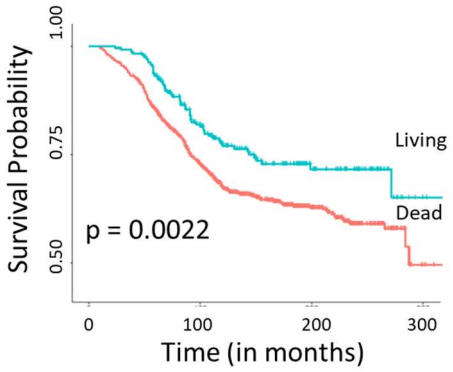


Figure 6a. The Kaplan-Meier curve demonstrating the ability of FOXM1-BUB1-CHEK1 to predict overall survival.

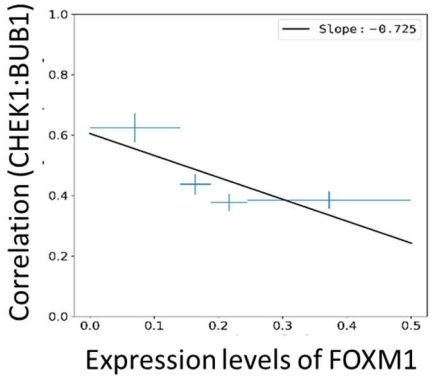


Figure 6b. A plot of correlation between BUB1 and CHEK1 as a function of expression levels of FOXM1 – exogenous (high) levels of FOXM1 expression are associated with the discoordination of CHEK1 and BUB1 expression, which, under nominal conditions, are tightly correlated.

To validate our computationally-derived hypothesis, we looked into published literature to understand the functional relationship between FOXM1, BUB1, and CHEK1. The protein-protein interaction network generated for the aforementioned genes using the online database resource - Search Tool for the Retrieval of Interacting Genes (StringDB) [2] - indicates functional associations (Fig. 7). Prior literature reveals the involvement of FOXM1 in the regulation of the

transcription of cell cycle progression genes, that CHEK1 as an important regulator in the DNA damage response pathway [3], and that BUB1 as mitotic checkpoint protein with an important role in chromosome segregation [4]. In fact, FOXM1 is a direct transcription regulator of both CHEK1 [5] and BUB1 [6].

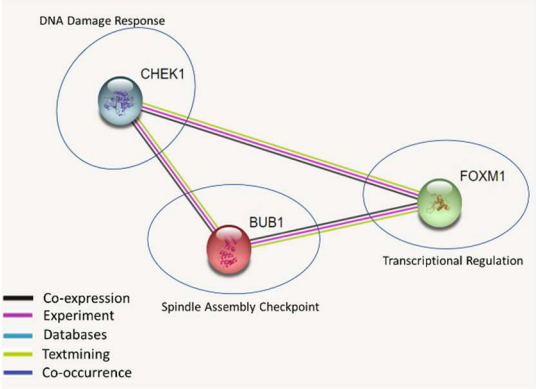


Figure 7. Protein interaction network showing putative interactions between FOXM1, BUB1 and CHEK1 generated by StringDB.

Conclusions

Due to heterogeneity in breast cancer, identifying subtype-specific gene interactions associated with survival will be useful in providing guidance for improved meta-dimensional prognostic biomarkers and tailoring newer therapeutic strategies. Further, as we learn to explore the space of high-dimensional interactions, we may learn that numerous distinct subtypes exist within current classifications, based on linear and low-dimensional models.

In summary, we developed a systematic workflow that incorporates biomarker classifier and our Pattern Discovery Engine for accurate biomarker prediction and for the discovery of novel gene interactions in search for personalized strategies for combating breast cancer. Higher-order interactions were identified and validated based on published literature. Our methods provide novel insights into gene interaction patterns in breast cancer and deliver candidates for further study. The proposed workflow can be broadly applied to other forms of cancers, and provides a unique view of the genomic landscape of disease states.

References

1. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington Z, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, and Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017; 45 (Database issue):D896.
2. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45(D1):D362.
3. Bryant C, Rawlinson R, Massey AJ. Chk1 inhibition as a novel therapeutic strategy for treating triple-negative breast and ovarian cancers. BMC Cancer. 2014;14:570.
4. Han JY, Han YK, Park GY, Kim SD, Lee CG. Bub1 is required for maintaining cancer stem cells in breast cancer cell lines. Sci Rep. 2015;5:15993.
5. Tan Y, Chen Y, Yu L, Zhu H, Meng X, Huang X, Meng L, Ding M, Wang Z, Shan L. Two-fold elevation of expression of FoxM1 transcription factor in mouse embryonic fibroblasts enhances cell cycle checkpoint activity by stimulating p21 and Chk1 transcription. Cell Prolif. 2010;43(5):494.
6. Wan X, Yeung C, Kim SY, Dolan JG, Ngo VN, Burkett S, Khan J, Staudt LM, Helman LJ. Cancer Res. Identification of FoxM1/Bub1b signaling pathway as a required component for growth and survival of rhabdomyosarcoma. 2012;72(22):5889.