Node and Edge Enrichment Analysis through Bipartite Networks: Application to Gene Mutations in Breast Cancer

Suresh K. Bhavnani PhD[1], Jeremy L. Warner MD, MS[3], Tianlong Chen PhD[1], Weibin Zhang PhD[1], Zhou Zhang PhD[5], Charles Balch MD[4], Sandra Hatch MD[2], Suzanne Klimberg MD PhD[2], Ning Liao MD PhD[6]

[1]Preventive Medicine and Population Health, [2]Cancer Center, University of Texas Medical Branch, Galveston, TX; [3]Div. of Hematology/Oncology and Dept of BMI, Vanderbilt University, Nashville, TN; [4]Div. of Surgery, MD Anderson Cancer Center, Houston, TX, USA; [5]Burning Rock Biotech, Shanghai, [6]Cancer Center, Guangdong General Hospital, Guangzhou, China

Introduction

A primary goal of precision medicine is to identify patient subgroups based on how they share key characteristics, and infer their underlying disease processes in order to design interventions that are targeted to those processes.[1] For example, breast cancer patients have been classified into five molecular subtypes (Luminal-A, Luminal-B, Triple-negative/basal like, HER2-enriched, and Normal-like).[2] However, despite the identification of these subtypes, little is understood about why subsets of patients have heterogeneous responses to current treatment paradigms.[3] For example, while most ER/PR+ HER2- patients (a clinical phenotype comprising the Luminal-A, Luminal-B, and Normal-like molecular subtypes) present at an early stage, and are cured with surgery and adjuvant therapy, a subset have poor response to current treatments, suggesting the existence of yet-to-be discovered molecular and clinical heterogeneities.

A common approach used to identify subtypes has been through the use of unipartite methods (e.g., k-means clustering, hierarchical clustering, and factor analysis), which identify *uniclusters* such as how patients cluster based on genes, or how genes co-occur across patients, with post-hoc approaches such as heatmaps to combine them. However, more recently the use of bipartite methods has shown improvements in the accuracy and interpretability of results by identifying *biclusters* such as simultaneously identifying patient subgroups and their frequently co-occurring gene mutations. Here we demonstrate how one such method called bipartite networks,[4] not only enabled the automatic identification and visualization of patient-gene biclusters, but also enabled identification of (a) which biclusters had a significantly higher proportion of patients with a specific outcome versus the rest of the data (**node enrichment**), and (b) which biclusters had genes with a significantly higher proportion of a specific mutational type within the cluster versus outside (**edge enrichment**). We demonstrate the efficacy and interpretability of this approach on a dataset of ER/PR+ HER2- Chinese breast cancer patients.

Method

***Data*.** We used a subset of data from a previous study[5] consisting of 217 Chinese breast cancer patients clinically phenotyped as ER/PR+ HER2-, and their mutational profile on 32 candidate genes. Based on their function, the mutations were categorized into four types: (1) missense variant, conservative in-frame deletion, conservative in-frame insertion, disruptive in-frame deletion, disruptive in-frame insertion, in-frame deletion, in-frame insertion, and fusion; (2) fusion and cm amp if they occur at the same time, cn amp); (3) frameshift variant, cn delete, splice acceptor variant, splice donor variant, splice region variant, splice variant, start lost, stop gained, stop lost; and (4) synonymous variant. As this subtype has low mortality rates, the outcome variables consisted of the fraction of Ki67 protein in tumor cells categorized into low (1-20), mid (21-50), and high (>50) levels, and the androgen receptor (AR) status in the tumor categorized as positive or negative. High levels of Ki67, and a negative status of AR are strong prognostic biomarkers for aggressive breast cancer in ER/PR+ HER2- patients.[6]

***Analysis*.** The analysis consisted of 4 steps: **(1) Bicluster Identification and Visualization.** (a) Represented the data as a bipartite network (Fig. 1), where nodes (circles and triangles) represented either patients or genes, the edges (lines) connecting the patient-gene pairs represented the presence of a gene mutation. Furthermore, the patient node color represented status on either Ki67 or AR (using separate networks), and the edge color represented one of four mutation types for specific patient-gene mutation pairs; (b) used bicluster modularity[4] to identify the number and boundaries of patient-symptom biclusters and the degree of biclustering (Q); (c) measured the significance of Q by comparing it to a distribution of Q generated from 1000 random permutations of the network; and (d) used the force-directed algorithm *Kamada-Kawai* to lay out the network, and *ExplodeLayout*[7] to separate the identified biclusters to improve their interpretability. **(2) Node Enrichment.** Used chi-squared with FDR correction, to measure the difference in proportion of high, mid, and low Ki67, and the proportion of positive and negative AR, in each bicluster compared to the rest of the data. **(3) Edge Enrichment.** Of those biclusters with significant node enrichment, we used chi-squared with FDR correction to measure the proportion of the mutation types for each gene-patient pair within that bicluster, versus outside that bicluster. **(4) Interpretation.** A team of oncologists interpreted the results based on: (a) the molecular mechanisms in specific biclusters that resulted in significantly high Ki67 and negative AR; and (b) potential targeted treatments.

Results

***Bicluster Identification and Visualization*.** As shown in Fig. 1, the bipartite network analysis identified 8 biclusters consisting of subgroups of breast cancer patients, and their most frequently co-occurring mutated genes, which had significant biclusteredness (Q=0.419, p<.001, z=6.0, two-tailed).

Node Enrichment**.** **Bicluster-A** had a significantly higher proportion $[\chi^2(1, N=217)=22.81, p<.001]$ of patients with high Ki67, and a significantly higher proportion $[\chi^2(1, N=217)=20.59, p<.001]$ of patients with negative AR, compared to the rest of the patients. **Bicluster-B** had a significantly higher proportion of patients with low Ki67 $[\chi^2(1, N=217)=11.06, p<.01]$, compared to the rest.

Edge Enrichment**.** **Bicluster-A**: (1) **TP53** had a significantly higher proportion of Type-1 and Type-3 mutations $[\chi^2(1, N=217)=69.49, p<.001]$; (2) **MYC** had a significantly higher proportion of Type-2 mutations $[\chi^2(1, N=217)=48.95, p<.001]$; and (3) **FGFR1** had a significantly higher proportion of Type-2 mutations $[\chi^2(1, N=217) =27.14, p<.001]$. **Bicluster-B**: (1) **GATA3** had a significantly higher proportion of Type-1 and Type-3 mutations $[\chi^2(1, N=217)=75.63, p<.001]$; (2) **AKT1** had a significantly higher proportion of Type-1 and Type-3 $[\chi^2(1, N=217)=30.67, p<.001]$.

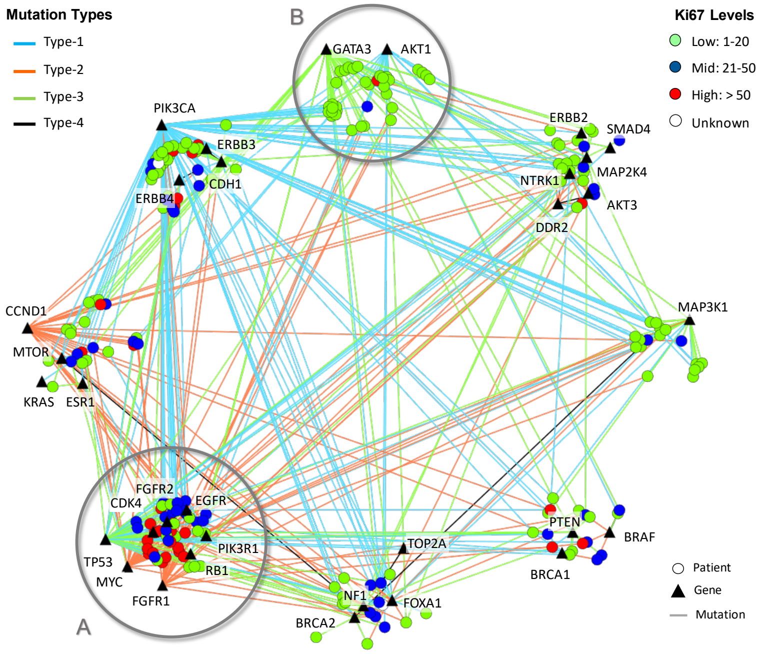

Fig. 1. 8 biclusters showing the co-occurrence of 32 genes mutated across 217 ER/PR+ HER2- Chinese breast cancer patients.

Interpretation**.** (1) **Bicluster-A**. This bicluster had a significantly higher proportion of patients with high Ki67, with Type-1 and Type-3 mutations more likely to be associated with loss of function, and Type-2 with gain of function. This is because the bicluster contained TP53 and RB1, two well-known tumor suppressor proteins, suggesting a loss of function as a driver event. Furthermore, the MYC and EGFR oncogenes could be additional drivers explaining the relative enrichment for high Ki67 in this cluster. Finally, the co-occurrence of FGFR1/2 and CDK4 suggests potential efficacy for CDK4/6 inhibitors (e.g., palbociclib), and for FGFR inhibitors (e.g., erdafitinib), requiring prospective validation. (2) **Bicluster-B.** This bicluster had a significantly higher proportion of patients with low Ki67, and contained AKT1 and GATA3, which might define a distinct subtype of patients with differential response to treatment.[8] Although treatment data is not yet available, the overall results suggest that treatment-emergent resistance could explain the biclustering and node/edge enrichment. For example, the bicluster containing ESR1 might represent patients with acquired endocrine therapy resistance[9]; the bicluster containing ERBB2 may represent acquired HER2 mutations[10], and the biclusters containing BRCA1/2 are likely enriched for germline variants of these mutations, which are mechanistically distinct from other etiologies of breast cancer.

Conclusions and Future Research.

The results suggest that node and edge enrichment analysis used in combination with bipartite network analysis and visualization enabled a multi-channel interpretation of the data. Our future research will integrate data related to treatment responses with the goal of generating testable hypotheses for interventions targeted to specific patient subgroups.

Acknowledgements. Funded in part by UTMB Cancer Center, UTMB CTSA (UL1 TR001439), and NIH U01 (CA231840).

References

1. Collins FS, Varmus H. A new initiative on precision medicine. *The New England journal of medicine.* 2015;372(9):793-795.
2. Dai X, Li T, Bai Z, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research.* 2015;5(10):2929-2943.
3. Yang L, Ye F, Bao L, et al. Somatic alterations of TP53, ERBB2, PIK3CA and CCND1 are associated with chemosensitivity for breast cancers. *Cancer science.* 2019;110(4):1389-1400.
4. Newman MEJ. *Networks: An Introduction.* Oxford, United Kingdom: Oxford University Press; 2010.
5. Chen B, Zhang G, Wei G, et al. Heterogeneity of genomic profile in patients with HER2-positive breast cancer. *Endocrine-related cancer.* 2020;27(3):153-162.
6. Vera-Badillo FE, Chang M, Kuruzar G, et al. Association between androgen receptor (AR) expression, Ki-67, and the 21-gene recurrence score in early breast cancer. *Journal of Clinical Oncology.* 2014;32(15_suppl):547-547.
7. Bhavnani SK, Chen T, Ayyaswamy A, et al. Enabling Comprehension of Patient Subgroups and Characteristics in Large Bipartite Networks: Implications for Precision Medicine. *Proceedings of AMIA Joint Summits on Translational Science.* 2017:21-29.
8. Smyth LM, Zhou Q, Nguyen B, et al. Characteristics and outcome of AKT1 E17K-mutant breast cancer defined through AACR GENIE, a clinicogenomic registry. *Cancer discovery.* 2020:CD-19-1209.
9. Jeselsohn R, Buchwalter G, De Angelis C, Brown M, Schiff R. ESR1 mutations—a mechanism for acquired endocrine resistance in breast cancer. *Nature Reviews Clinical Oncology.* 2015;12(10):573-583.
10. Nayar U, Cohen O, Kapstad C, et al. Acquired HER2 mutations in ER+ metastatic breast cancer confer resistance to estrogen receptor–directed therapies. *Nature genetics.* 2019;51(2):207-216.