Personalized Annotation-based Networks (PAN) for the Prediction of Breast Cancer Relapse

Thin Nguyen[1]*, Samuel C. Lee[1], Thomas P. Quinn[1], Buu Truong[2], Xiaomei Li[2], Truyen Tran[1], Svetha Venkatesh[1], Thuc Duy Le[2]*

[1]Applied Artificial Intelligence Institute, Deakin University, Australia
[2]School of Information Technology and Mathematical Sciences, University of South Australia, Australia
*thin.nguyen@deakin.edu.au, thuc.le@unisa.edu.au

Abstract

The classification of clinical samples based on gene expression data is an important part of precision medicine. However, it is not a trivial task and it is difficult to accurately predict survival outcomes and treatment responses despite advancements in the field. In this manuscript, we show how transforming gene expression data into a set of personalized (sample-specific) networks can allow us to harness existing graph-based methods to improve classifier performance. Existing approaches to personalized gene networks, based on protein-protein interactions (PPI) or population-level models, all have the limitation that they depend on other samples in the data and must get re-computed whenever a new sample is introduced. Here, we propose a novel method, called Personalized Annotation-based Networks (PAN), that avoids this limitation by using curated annotation databases to transform gene expression data into a graph. These databases organize genes into overlapping gene sets, called annotations, that we use to build a network where nodes represent functional terms and edges represent the similarity between them. Unlike competing methods, PANs are calculated for each sample independent of the population, making it a more general solution to the single-sample network problem. Using two breast cancer datasets as a case study (METABRIC and a super-set of GEO studies), we show that PAN classifiers not only predict cancer relapse better than gene features alone, but also outperform PPI and population-level graph-based classifiers. This work demonstrates the practical advantages of graph-based classification for high-dimensional genomic data, while offering a new approach to making sample-specific networks.

1 Introduction

Breast cancer is one of the leading causes of death for women worldwide, with the incidence and mortality increasing globally [6].

Yet, it is not a single disease, but rather a collection of multiple biological entities, each with their own molecular signature and clinical implications [12]. Since prognosis and treatment response differs between and within cancer sub-types [16], there exists a strong motivation to develop methods that can accurately predict key patient outcomes, such as relapse after treatment, and so use them to tailor treatments to their patients [7]. However, despite advancements in the field, there remains much unexplained inter-tumour heterogeneity that drives substantial differences in survival outcomes and treatment responses [11]. Gene expression signatures, easily measured from a tissue sample using high-throughput assays, have been used as a means of stratifying breast cancer samples. This has resulted in computational methods that identify personalized "driver mutation" genes [18], differentially expressed genes and pathways [33], and individualized gene networks [34]. Although genes have been used successfully as biomarkers for cancer prediction tasks [17], it is not clear that gene biomarkers are the most appropriate substrate for classification. Rather, it may be more meaningful to describe diseases in terms of the dysfunction of specific systems, rather than the dysfunction of individual molecules [22]. This perspective is achieved by gene regulatory networks.

Gene regulatory networks represent genes as nodes and the interactions between them as edges. The interactions between genes can be inferred in three ways: from knowledge-driven methods, data-driven methods, or hybrid methods. Knowledge-driven methods use public databases which catalog experimentally confirmed (or predicted) information relating protein-protein interactions (**PPI**) or functionally associated gene sets (called *annotations*). Examples of these databases include the Kyoto Encyclopedia of Genes and Genomes (**KEGG**) [19], Human Phenotype Ontology (**HPO**) [29], Disease Ontology (**DO**) [2], HIPPIE v2.0 [1], among others [9, 31]. Data-driven methods are usually based on gene-gene correlation coefficients [13] or causal relationships [27], as inferred directly from gene expression data. Hybrid methods construct gene regulatory networks by combining gene expression data with the prior knowledge found in annotation databases [23]. Regardless of the method used, most studies compute gene networks at the population- or group-level instead of building personalized (sample-specific) gene regulatory networks. In principle, this results in a single model for all samples in a population, taking a "one-size-fits-all" approach that ignores the inter-tumour heterogeneity of breast cancer. Although population-level networks can help researchers understand a disease in the general sense, personalized gene regulatory networks could pave the way toward accurate and individualized disease prediction.

It is challenging to create sample-specific gene regulatory networks because individuals rarely have the multiple gene expression profiles necessary to compute intra-sample correlations. Borgwardt et al. [3] proposed an approach that uses a common PPI reference network to serve as a template from which edges are trimmed based on the gene co-expression status for *that individual* (relative to the population as a whole). Since this method trims edges by comparing a single sample with the population distribution, these PPI-based networks must be re-computed whenever a new sample is introduced. More recently, Kuijjer et al. [20] proposed a more formal method called LIONESS that builds a sample-specific network by estimating the contribution that each sample makes toward the population-level gene regulatory network. As such, LIONESS produces a unique graph for each sample without having to integrate external information. However, LIONESS has an important limitation: whenever a new sample is introduced into the dataset, all sample-specific networks must get re-computed. Liu et al. [22] have proposed another method, similar to LIONESS, but this has the same re-computation issue. All of these methods use population-level information to build template networks, making them impractical for machine learning applications where populations change, notably when data are streaming.

In response to these problems, we introduce a novel method for constructing sample-specific networks, called Personalized Annotation-based Networks (**PAN**), that uses curated annotation databases to transform gene expression data into a graph. Using gene set annotations, available through these databases, we build sample-specific networks where nodes represent functional terms and edges represent the similarity between them. Once a network is built for each sample, we can then use the graph properties (e.g., Closeness Centrality, Betweenness Centrality, or PageRank) as features for classification. Unlike the PPI-based and LIONESS sample-specific graphs, PAN is calculated for each sample agnostic to the population, making it a more general solution to the single-sample network problem. Using two large breast cancer datasets, we show that the graph properties of PAN not only predict breast cancer relapse data better than gene features alone, but also outperform the PPI-based and LIONESS graphs. In evaluating our method, we trial PAN with three separate annotation databases, and find that Disease Ontology networks consistently perform best. This work demonstrates the practical advantages of graph-based classification for high-dimensional genomic data, while offering a new approach to making sample-specific networks.

2 Methods

2.1 Overall objective

The ultimate goal of this work is to predict breast cancer relapse by using personalized (sample-specific) gene regulatory networks for classification. For this, we propose a new method for building sample-specific networks, called Personalized Annotation-based Networks (**PAN**), that we benchmark against established methods. Since PAN depends on the annotation database chosen, we test three databases: Kyoto Encyclopedia of Genes and Genomes (**KEGG**) [19], Human Phenotype Ontology (**HPO**) [29], and Disease Ontology (**DO**). To evaluate the performance of PAN, we use publicly available data from the Gene Expression Omnibus (**GEO**) and METABRIC. Our proposed PAN method is summarized in Figure 1.

2.2 Data acquisition

The data used for the prediction of events were obtained from two main sources, the Gene Expression Omnibus (**GEO**) and METABRIC. The first dataset, named GEO-5, is a super-set of five GEO collections (including GSE12276 [4], GSE20711 [14], GSE19615 [21], GSE21653 [30], and GSE9195 [24]). These datasets were selected because they all use the same micro-array plat-

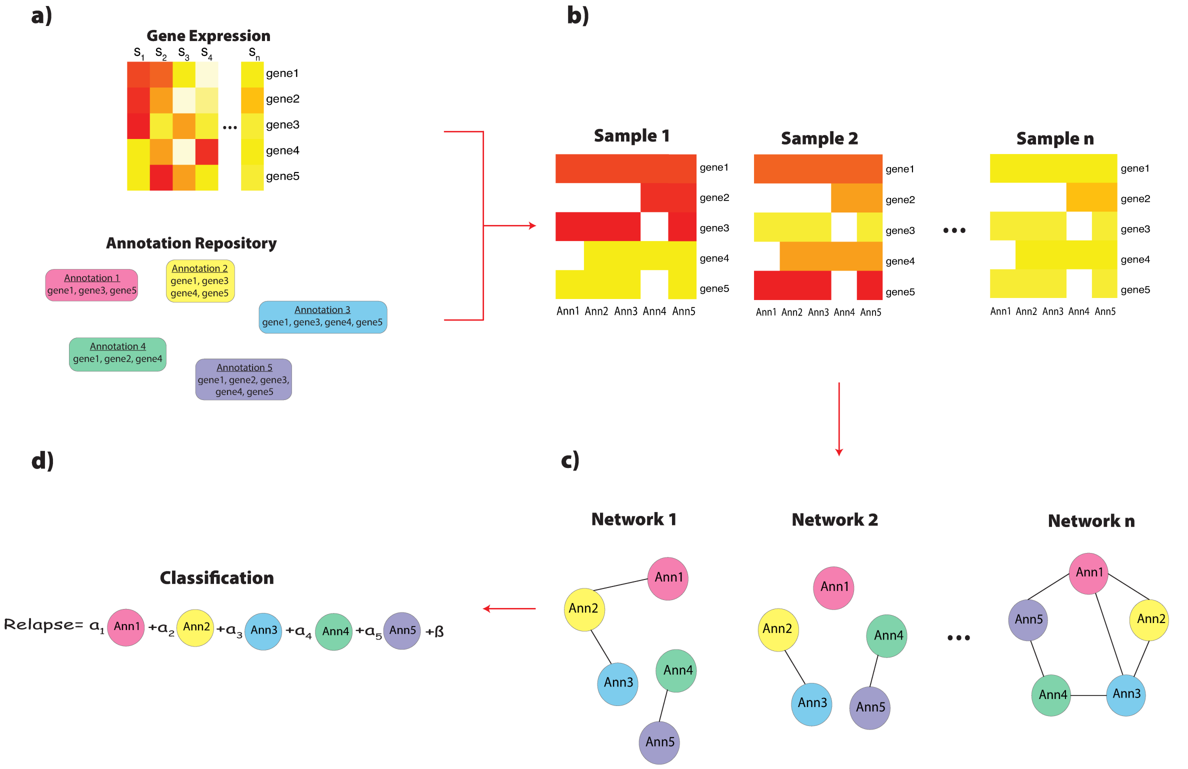
a)
Gene Expression
S1 S2 S3 S4 Sn
gene1
gene2
gene3
gene4
gene5
Annotation Repository
Annotation 1 gene1, gene3, gene5
Annotation 2 gene1, gene3 gene4, gene5
Annotation 3 gene1, gene3, gene4, gene5
Annotation 4 gene1, gene2, gene4
Annotation 5 gene1, gene2, gene3, gene4, gene5
b)
Sample 1
Sample 2
Sample n
Ann1 Ann2 Ann3 Ann4 Ann5
c)
Network 1
Network 2
Network n
Ann1
Ann2
Ann3
Ann4
Ann5
d)
Classification
Relapse= a1 Ann1 +a2 Ann2 +a3 Ann3 +a4 Ann4 +a5 Ann5 +ß

Figure 1: This figure illustrates how we create a Personalized Annotation-based Network (**PAN**) graph. **Panel A:** We obtain publicly available Gene Expression datasets (GEO-5 or METABRIC) and Gene Annotations (KEGG, HPO, or DO). **Panel B:** For each sample, we build an intermediate matrix by matching the Gene Expression data to the Gene Annotations: if there is an annotation for that gene, the value of the intermediate matrix becomes the expression of that gene. Otherwise, the value becomes zero. **Panel C:** From the intermediate matrix, we build a network where the nodes are the annotations and the edges are the Euclidean distance between them. To turn the annotation-annotation association weights into a discretized network, we create an adjacency matrix from the top 10% of edges as ranked by association strength. **Panel D:** After building the network, we use its properties (Betweenness Centrality, Closeness Centrality, or PageRank) as feature input for a classification algorithm (linear regression or support vector machine) to predict breast cancer relapse. Acronyms: S (Sample); Ann (Annotation).

form (Affymetrix Human Genome U133 Plus 2.0 Array) and include clinical information about relapse. The GEO-5 dataset contains 736 samples (349 relapse; 387 no relapse).

The second dataset was retrieved from METABRIC, a popular breast cancer dataset from the European Genome-Phenome Archive [10]. We obtained the gene expression data already pre-processed by limma [28] and fRMA [25] libraries. Next, we adjusted the data for batch effects using the ComBat algorithm from the sva library (with default parameters and covariates as tumor versus normal). The repository contains 2000 breast tumor samples. To be included in our study, we required that the sample had clinical information about relapse. Filtering samples with missing clinical data reduced the dataset to 1283 samples (422 relapse; 861 no relapse).

2.3 Defining sample-specific networks

2.3.1 Feature selection

To make the calculation of single-sample graphs tractable for high-dimensional gene expression data, we performed feature selection on the gene expression data. For each dataset, we included the top 100 genes ranked by total variance. Using the same list of 100 genes, we built sample-specific networks as described below. Note that the same 100 genes were also used to train the non-graph classifiers.

2.3.2 PPI-based networks

Although a personalized network is not immediately available from the data, we can build one from a protein-protein interaction (**PPI**) network [3]. In this method, the PPI serves as reference network that is subsequently pruned to establish sample-specific networks. For each sample, we keep any PPI edge where its constituent genes both have high (or low) relative expression. Specifically, we include all edges where both genes are in either the top or bottom quintiles, as compared with the other samples. For this application, we use the HIPPIE database of experimentally observed PPIs [1].

2.3.3 LIONESS networks

We applied LIONESS to the gene expression data using PyPanda [32], run with default parameters. To turn the gene-gene association weights into a discretized network, we create an adjacency matrix from the top 10% of edges as ranked by association strength.

2.3.4 Personalized Annotation-based Networks (PAN)

In this work, we propose a novel method, called Personalized Annotation-based Networks (**PAN**), to build personalized networks for individual samples based on gene expression data. It begins by converting gene expression data (using one of the KEGG, HPO, or DO databases) into annotation-based measures. Formally, each sample j is represented by the matrix of $M^j_{p,q}$, where p is a gene and q is an annotation associated with some genes. The value for each element $M^j_{p,q}$ equals the expression of gene p if it relates to the annotation q. Otherwise, it equals zero. Then, M^j is turned into a symmetric association matrix, where the value for each element A^j_{q,q^*} describes the Euclidean distance between an annotation q and another annotation q^*. For PAN, the nodes are annotations and the edges are the Euclidean distance between them. To turn the annotation-annotation association weights into a discretized network, we create an adjacency matrix from the top 10% of edges as ranked by association strength. Since a network is created for each sample independently, PAN does not have the limitation that the PPI-based and LIONESS approaches have. Our proposed PAN method is summarized in Figure 1.

2.4 Graph-based Representation

Once the discretized networks are constructed for each sample, their graph properties can be extracted and used as features for classification. Given a sample S, a graph $G^S\left(V^S, E^S\right)$ is constructed from its gene expression using the PPI-based method [3], the LIONESS method [20], or the proposed PAN method. The graph-based representation of S is denoted as $\mathbf{h}^S$ and defined based on G^S. In the following sub-sections, we present different ways to define $\mathbf{h}^S$. By using graph properties, $\mathbf{h}^S$ is represented as a vector of V^Sdimensions, i.e., $\mathbf{h}^S = \left[h_1^S, ..., h_{|V^S|}^S\right]$ where h_j^S is computed from properties of the vertex v_j^S.

2.4.1 Closeness Centrality

The Closeness Centrality of a node $v_j^S \in V^S$ is defined as the reciprocal of the sum of the shortest path distances from v_j^S to all other nodes [15],

$$h_j^S = CC\left(v_j^S\right) = \frac{|V^S| - 1}{\sum_{v_k^S \in V^S - \left\{v_j^S\right\}} l\left(v_j^S, v_k^S\right)}$$

where $l\left(v_j^S, v_k^S\right)$ is the length of the shortest-path from v_j^S to v_k^S.

2.4.2 Betweenness Centrality

The Betweenness Centrality of a node $v_j^S \in V^S$ is defined as the sum of the fraction of all-pairs shortest paths that pass through v_j^S [5],

$$h_j^S = BC\left(v_j^S\right) = \sum_{v_k^S, v_l^S \in V^S} \frac{\beta\left(v_k^S, v_l^S \middle| v_j^S\right)}{\beta\left(v_k^S, v_l^S\right)}$$

where $\beta\left(v_k^S, v_l^S\right)$ is the number of shortest $\left(v_k^S, v_l^S\right)$-paths and $\beta\left(v_k^S, v_l^S \middle| v_j^S\right)$ is the number of those paths passing through v_j^S other than v_k^S and v_l^S. Note, if $v_k^S = v_l^S$, then $\beta\left(v_k^S, v_l^S\right) = \beta\left(v_k^S, v_k^S\right) = 1$, and if $v_j^S \in \left\{v_k^S, v_l^S\right\}$, then $\beta\left(v_k^S, v_l^S \middle| v_j^S\right) = 0$.

2.4.3 PageRank

PageRank [26] was developed for measuring the importance of websites on the Internet. This method makes use of an underlying assumption that more important websites are likely to receive more links from others. In our case, we define the PageRank property of a node $v_j^S \in V^S$ as:

$$h_j^S = PR\left(v_j^S\right) = \sum_{v_k^S \in \mathcal{N}\left(v_j^S\right)} \frac{PR\left(v_k^S\right)}{L\left(v_k^S\right)}$$

where $\mathcal{N}(v_j^S)$ is the set of all nodes linking to node v_j^S and $L(v_k^S)$ is the number of links from v_k^S.

2.5 Classifier choice and performance

We trained linear regression (**LR**) and linear kernel support vector machine (**SVM**)

[8] classifiers using (a) the gene expression data (**non-graph**) and (b) the properties of the PPI-based, LIONESS, and PAN graphs. Since PAN depends on the database used, we used three databases to build three separate PAN models: **PAN_KEGG**, **PAN_HPO**, and **PAN_DO**. For each classifier, we report the accuracy, area under the receiver operating curve (**AUC**), and F1-score, as averaged across a 10-fold cross-validation of the dataset. Although we tested several graph properties, we only report the best performing classifier for each model (based on which graph property yielded the highest AUC). In most cases, Betweenness Centrality yielded the highest AUC.

3 Results and Discussion

3.1 Graph properties outperform gene expression

Once a single-sample network is created, its graph properties can be calculated and used as feature input for classification. Figures 2 & 3 show the average 10-fold cross-validation accuracies for graph and non-graph classifiers in the prediction of breast cancer relapse. These figures compare PAN classifier performance (built using three separate databases: KEGG, HPO, and DO) with the PPI, LIONESS, and non-graph classifiers. For the GEO-5 dataset, we see that all graph classifiers perform better than the commonly used non-graph (gene expression) classifier. Considering that the GEO-5 dataset is aggregated from multiple sources, this might suggest that single-sample networks capture a signal that is more robust to inter-batch differences. The PPI-based and LIONESS graphs did not perform as well for

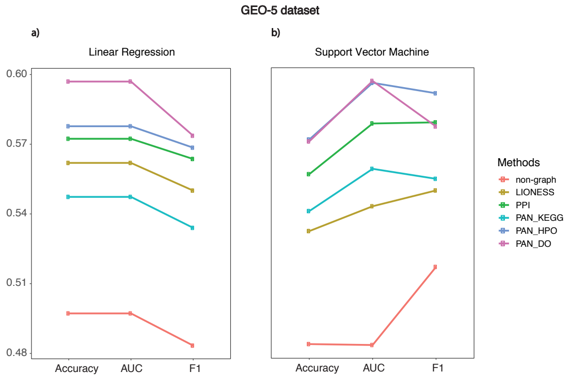

Figure 2: This figure shows the performance for six classifiers in the prediction of breast cancer relapse using the GEO-5 data. The left panel shows logistic regression performance, while the right panel shows support vector machine performance. The PPI, LIONESS, and PAN methods refer to graph-based methods, where graph properties are used as feature input. The non-graph method refers to using gene expression data as feature input. Note that we evaluated three separate PAN models, each built using a separate annotation database (KEGG, HPO, and DO). All performance metrics are averaged across 10-folds of cross-validation.

the METABRIC data as they did for the GEO-5 data. However, all PAN classifiers still outperform the non-graph (gene expression) classifier.

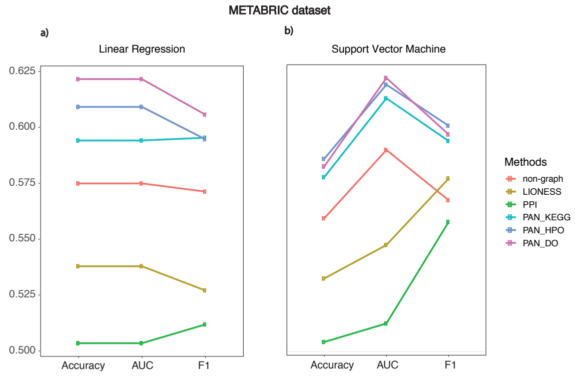

Figure 3: This figure shows the performance for six classifiers in the prediction of breast cancer relapse using the METABRIC data. The left panel shows logistic regression performance, while the right panel shows support vector machine performance. The PPI, LIONESS, and PAN methods refer to graph-based methods, where graph properties are used as feature input. The non-graph method refers to using gene expression data as feature input. Note that we evaluated three separate PAN models, each built using a separate annotation database (KEGG, HPO, and DO). All performance metrics are averaged across 10-folds of cross-validation.

3.2 PAN has consistently superior performance

The goodness of any classifier might depend largely on the dataset under study. Therefore, it is important to evaluate a classifier's performance across multiple datasets. When comparing the rank-order of the six classifiers between the GEO-5 and METABRIC data, we see that the PPI classifier has an inconsistent performance. Although the PPI classifier is among the best for GEO-5, it is among the worst for METABRIC. On the other hand, the LIONESS and non-graph (gene expression) classifiers have consistently poor performance, while the PAN_DO and PAN_HPO classifiers outperform all competing methods for both datasets. Likewise, PAN_KEGG always outperforms the non-graph method. Although the margin is small, PAN_DO usually performs better than PAN_HPO. As such, we see a stable rank-order among the PAN methods: PAN_DO > PAN_HPO > PAN_KEGG. It is interesting to note that the better performing annotation databases have fewer total annotations. This could suggest that PAN's success may have something to do with its ability to condense the high-dimensional gene expression data into a lower-dimensional space. Whatever the reason, PAN classifiers appear to perform reliably well for breast cancer relapse prediction.

4 Summary

In this paper, we have proposed a novel method for constructing sample-specific networks, called Personalized Annotation-based Networks (**PAN**), which use curated annotation databases to transform gene expression data into a graph. Using the properties of these graphs as feature input for classification, we show that PAN can not only predict breast cancer relapse better than non-graph (gene expression) classifiers, but also outperform competing sample-specific graph methods. Although PAN graphs depend on the annotation database used, we show that PAN classifiers perform consistently well across three separate databases (KEGG, HPO, and DO), with PAN_DO and PAN_HPO having superior performance in all tests. Our results support two principal conclusions. First, they suggest that applying graph-based models to the classification of gene expression data improves performance considerably. Second, they suggest that integrating annotation databases into classification pipelines is appropriate for clinically relevant classification problems, such as the prediction of breast cancer relapse. Although we showcase the PAN method on gene expression data, our method can be generalized to any classification problem where a relevant annotation database exists (e.g., for protein expression or methylation data).

References

[1] Gregorio Alanis-Lobato, Miguel A Andrade-Navarro, and Martin H Schaefer. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 45(D1):D408–D414, 2016.

[2] Susan M Bello, Mary Shimoyama, Elvira Mitraka, Stanley JF Laulederkind, Cynthia L Smith, Janan T Eppig, and Lynn M Schriml. Disease Ontology: improving and unifying disease annotations across species. *Disease Models & Mechanisms*, 11(3):dmm032839, 2018.

[3] Karsten M Borgwardt, Hans-Peter Kriegel, SVN Vishwanathan, and Nicol N Schraudolph. Graph kernels for disease outcome prediction from protein-protein interaction networks. In *Biocomputing*, pages 4–15. World Scientific, 2007.

[4] Paula D Bos, Xiang H-F Zhang, Cristina Nadal, Weiping Shu, Roger R Gomis, Don X Nguyen, Andy J Minn, Marc J van de Vijver, William L Gerald, John A Foekens, and Joan Massagué. Genes that mediate breast cancer metastasis to the brain. *Nature*, 459(7249):1005, 2009.

[5] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.

[6] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018.

[7] Sang-Hoon Cho, Jongsu Jeon, and Seung Il Kim. Personalized medicine

in breast cancer: A systematic review. *Journal of Breast Cancer*, 15(3):265, 2012.

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[9] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung, Ewan Birney, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database):D691–D697, 2010.

[10] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346, 2012.

[11] Ibiayi Dagogo-Jack and Alice T Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15:81, 2017.

[12] Xiaofeng Dai, Ting Li, Zhonghu Bai, Yankun Yang, Xiuxia Liu, Jinling Zhan, and Bozhi Shi. Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research*, 5(10):2929–2943, 2015.

[13] Maria Angels de Luis Balaguer, Adam P Fisher, Natalie M Clark, Maria Guadalupe Fernandez-Espinosa, Barbara K Möller, Dolf Weijers, Jan U Lohmann, Cranos Williams, Oscar Lorenzo, and Rosangela Sozzani. Predicting gene regulatory networks by combining spatial and temporal gene expression data in arabidopsis root stem cells. *Proceedings of the National Academy of Sciences*, 114(36):E7632–E7640, 2017.

[14] Sarah Dedeurwaerder, Christine Desmedt, Emilie Calonne, Sandeep K Singhal, Benjamin Haibe-Kains, Matthieu Defrance, Stefan Michiels, Michael Volkmar, Rachel Deplus, Judith Luciani, Françoise Lallemand, Denis Larsimont, Jérôme Toussaint, Sandy Haussy, Françoise Rothé, Ghizlane Rouas, Otto Metzger, Samira Majjaj, Kamal Saini, Pascale Putmans, Gérald Hames, Nicolas van Baren, Pierre G Coulie, Martine Piccart, Christos Sotiriou, and François Fuks. DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Molecular Medicine*, 3(12):726–741, 2011.

[15] Linton C Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1:215–239, 1979.

[16] Deena MA Gendoo, Natchar Ratanasirigulchai, Markus S Schröder, Laia Paré, Joel S Parker, Aleix Prat, and Benjamin Haibe-Kains. Genefu: an R/Bioconductor package for computation of gene expression-based signatures

in breast cancer. *Bioinformatics*, 32(7):1097–1099, 2015.

[17] Michael Gnant, Nadia Harbeck, and Christoph Thomssen. St. Gallen/Vienna 2017: A brief summary of the consensus discussion about escalation and de-escalation of primary breast cancer treatment. *Breast Care*, 12(2):102–107, 2017.

[18] Wei-Feng Guo, Shao-Wu Zhang, Li-Li Liu, Fei Liu, Qian-Qian Shi, Lei Zhang, Ying Tang, Tao Zeng, and Luonan Chen. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics*, 34(11):1893–1903, 2018.

[19] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[20] Marieke Lydia Kuijjer, Matthew Tung, GuoCheng Yuan, John Quackenbush, and Kimberly Glass. Estimating sample-specific regulatory networks. *arXiv*, 1505.06440, 2018.

[21] Yang Li, Lihua Zou, Qiyuan Li, Benjamin Haibe-Kains, Ruiyang Tian, Yan Li, Christine Desmedt, Christos Sotiriou, Zoltan Szallasi, J Dirk Iglehart, Andrea L Richardson, and Zhigang Charles Wang. Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. *Nature Medicine*, 16:214–218, 2010.

[22] Xiaoping Liu, Yuetong Wang, Hongbin Ji, Kazuyuki Aihara, and Luonan Chen. Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Research*, 44(22):e164–e164, 2016.

[23] Zhi-Ping Liu, Hulin Wu, Jian Zhu, and Hongyu Miao. Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza a virus infection. *BMC Bioinformatics*, 15(1):336, 2014.

[24] Sherene Loi, Benjamin Haibe-Kains, Christine Desmedt, Pratyaksha Wirapati, Françoise Lallemand, Andrew M Tutt, Cheryl Gillet, Paul Ellis, Kenneth Ryder, James F Reid, Maria G Daidone, Marco A Pierotti, Els MJJ Berns, Maurice PHM Jansen, John A Foekens, Mauro Delorenzi, Gianluca Bontempi, Martine J Piccart, and Christos Sotiriou. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, 9(1):239, 2008.

[25] Matthew N McCall, Benjamin M Bolstad, and Rafael A Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–253, 2010.

[26] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[27] Xiaojie Qiu, Arman Rahimzamani, Li Wang, Qi Mao, Timothy Durham, Jose L McFaline-Figueroa, Lauren Saunders, Cole Trapnell, and Sreeram

Kannan. Towards inferring causal gene regulatory networks from single cell expression measurements. *bioRxiv*, 426981, 2018.

[28] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[29] Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008.

[30] Renaud Sabatier, Pascal Finetti, Nathalie Cervera, Eric Lambaudie, Benjamin Esterni, Emilie Mamessier, Agnès Tallet, Christian Chabannon, Jean-Marc Extra, Jocelyne Jacquemier, Patrice Viens, Daniel Birnbaum, and François Bertucci. A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Research and Treatment*, 126(2):407–420, 2010.

[31] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, Michael Kuhn, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2014.

[32] David GP van IJzendoorn, Kimberly Glass, John Quackenbush, and Marieke L Kuijjer. PyPanda: a Python package for gene regulatory network reconstruction. *Bioinformatics*, 32(21):3363–3365, 2016.

[33] Hongwei Wang, Qiang Sun, Wenyuan Zhao, Lishuang Qi, Yunyan Gu, Pengfei Li, Mengmeng Zhang, Yang Li, Shu-Lin Liu, and Zheng Guo. Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics*, 31(1):62–68, 2014.

[34] Xiangtian Yu, Jingsong Zhang, Shaoyan Sun, Xin Zhou, Tao Zeng, and Luonan Chen. Individual-specific edge-network analysis for disease prediction. *Nucleic Acids Research*, 45(20):e170–e170, 2017.