Bioinformatics, 35(22), 2019, 4830–4833
doi: 10.1093/bioinformatics/btz485
Advance Access Publication Date: 14 June 2019
Applications Note

OXFORD

Gene expression

SIGN: similarity identification in gene expression

Seyed Ali Madani Tonekaboni[1,2], Venkata Satya Kumar Manem[1,2,3], Nehme El-Hachem[4,5] and Benjamin Haibe-Kains[1,2,6,7,8,*]

[1]Princess Margaret Cancer Centre, [2]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada, [3]Institut Universitaire de Cardiologie et de Pneumologie de Québec, Université Laval, QC G1V 4G5, Canada, [4]Integrative Systems Biology, Institut de Recherches Cliniques de Montréal, [5]Department of Medicine, University of Montreal, Montréal, QC, Canada, [6]Department of Computer Science, University of Toronto, Toronto, ON M5T 3A1, Canada, [7]Ontario Institute of Cancer Research and [8]Vector Institute, Toronto, ON M5G 1L7, Canada

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on December 20, 2018; revised on April 12, 2019; editorial decision on May 25, 2019; accepted on June 10, 2019

Abstract

Motivation: High-throughput molecular profiles of human cells have been used in predictive computational approaches for stratification of healthy and malignant phenotypes and identification of their biological states. In this regard, pathway activities have been used as biological features in unsupervised and supervised learning schemes.

Results: We developed SIGN (Similarity Identification in Gene expressioN), a flexible open-source R package facilitating the use of pathway activities and their expression patterns to identify similarities between biological samples. We defined a new measure, the transcriptional similarity coefficient, which captures similarity of gene expression patterns, instead of quantifying overall activity, in biological pathways between the samples. To demonstrate the utility of SIGN in biomedical research, we establish that SIGN discriminates subtypes of breast tumors and patients with good or poor overall survival. SIGN outperforms the best models in DREAM challenge in predicting survival of breast cancer patients using the data from the Molecular Taxonomy of Breast Cancer International Consortium. In summary, SIGN can be used as a new tool for interrogating pathway activity and gene expression patterns in unsupervised and supervised learning schemes to improve prognostic risk estimation for cancer patients by the biomedical research community.

Availability and implementation: An open-source R package is available (https://cran.r-project.org/web/packages/SIGN/).

Contact: bhaibeka@uhnresearch.ca

1 Introduction

Messenger RNA (mRNA) expression is an important feature representative of the biological state of a cell or cell population. Activity of tissue-specific genes, master regulatory factors, tumor suppressor and oncogenes can play important roles in variety of healthy and disease phenotypes (Campbell and Marlow, 2013; Chatterjee and Vinson, 2012; Spitz and Furlong, 2012). External stress, such as drug treatment, hypoxia or other microenvironmental conditions of a tissue affect the mRNA transcription in cancer cells (Bindra *et al.*, 2005; Razorenova and Giaccia, 2010).

Enrichment of pathways and their activity have been used as features in machine learning frameworks to predict identity of cells, mechanism of action of drugs or mechanism of resistance of cancer cells (Karr *et al.*, 2012; Michelson and Young, 2011; Silberberg *et al.*, 2012). To facilitate this process, we developed SIGN (Similarity Identification in Gene expressioN) as an open-source R package. The expression

profiles of biological samples can be transformed to features at the pathway level using pathway enrichment scoring approaches like single-sample Gene Set Enrichment Analysis (GSEA) (Barbie *et al.*, 2009; Subramanian *et al.*, 2005) and Gene Set Variation Analysis (GSVA) (Hänzelmann *et al.*, 2013). These features are then used in a centroid classification scheme for supervised learning tasks. We further introduce the transcriptional similarity coefficient (TSC), an estimator of gene expression pattern similarity of pathway activity between biological samples based on the RV statistic (Smilde *et al.*, 2009). We used SIGN to classify breast tumors into subtypes and predict survival of patients in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort (Curtis *et al.*, 2012). We show that SIGN yields higher performance with respect to the top performing approaches from the community DREAM challenge for breast cancer survival prediction (Margolin *et al.*, 2013).

2 Materials and methods

We applied SIGN for subtype clustering and patient survival prediction using the transcriptomic profiles of patient tumors. The underlying centroid classification within SIGN is used to stratify patients with good survival from poor survival patients relying on gene expression profiles of their corresponding tumors. We further hypothesized that gene expression patterns, biological pathways as the relative expression of genes within the pathway, are determinants of cell identity. We defined TSC as a new way for comparing patterns of expression of biological pathways between tumors. Widely used pathway databases, such as KEGG and Reactome as well as Gene Ontology (GO) terms can be used as set of pathways in SIGN. GO terms in level C5 with 10–30 number of genes are used in this study to identify the similarity between samples based on their gene expression pattern (Liberzon *et al.*, 2011). We limited the number of genes in GO terms to exclude large GO terms (at the top of GO term hierarchy) that are parents of the GO terms in our study (at the bottom of the GO hierarchy).

2.1 Transcriptional similarity coefficient

Let P be the matrix of expression of genes within a pathway for a set of biological samples where rows are genes and columns are samples. SIGN defines the TSC between the two matrices using modified RV coefficient (Smilde *et al.*, 2009) as follows

$$TSC(P_1, P_2) = \frac{\sum_i (P_{10} \times P_{20})_{ii}}{\sqrt{\sqrt{\sum_{ij} (P_{10})_{ij}^2} \times \sqrt{\sum_{ij} (P_{20})_{ij}^2}}}$$

where P_1 and P_2 represent the matrix of gene expressions of a given pathway in two set of samples (population 1 and 2), i is row index (i.e. gene index) within each matrix, j is column index (i.e. sample index) within each matrix and P_{m0}

$$P_{m0} = P_m \times P'_m - \text{Diagonal}(P_m \times P'_m)$$

where the Diagonal function sets to zero the elements of the matrix that are not in the diagonal. The range of the TSC score lies between [−1, 1]. Higher scores indicates higher similarity of gene expression pattern of a given pathway between two populations. Therefore, TSC represents the similarity of a given pathway activity between two samples and/or sample sets. We identify similarity of pathways between two sample sets relying on distribution of TSCs between the populations.

2.2 Breast cancer subtype similarity

We used TSC as the measure of similarity of ESR1 and ERBB2 gene module activities (Haibe-Kains *et al.*, 2012) between breast tumor samples in the discovery cohort of METABRIC. The similarities between the samples were used to identify total Euclidean distance between the samples and cluster them accordingly.

2.2 Survival analysis

We assessed performance of SIGN to predict the overall survival of patients as an endpoint. We split the discovery cohort in METABRIC (Curtis *et al.*, 2012), for patients under the same treatment category (Hormonal therapy, Chemo+hormonal therapy, chemotherapy or under no drug treatment), into three groups of poor (10%), intermediate (80%) and good (10%) survival. We defined

$$\begin{aligned}\text{Delta} = &(\text{similarity to good survival cohort}) \\ &- (\text{similarity to poor survival cohort})\end{aligned}$$

and used it to identify difference of similarity of each patient tumor sample to good and poor survival populations. To assess the significance of the predictions using log-rank test, we binarized Delta as being Delta+ or Delta−. Log-rank function implemented in the survcomp R package (1.32.0) (Schroder *et al.*, 2011).

2.3 Research reproducibility

SIGN is publicly available as an open-source R package (https://cran.r-project.org/web/packages/SIGN/) and the results of this article can be reproduced using the cloud-based computational reproducibility platform CodeOcean (http://bit.ly/2PMwegY).

3 Results

We leveraged the METABRIC dataset of breast cancer patients to test whether TSC can be used to recapitulate subtyping of breast cancer patients. We used TSC comparing expression patterns within signatures of luminal, basal and HER2+ breast cancer subtypes (Desmedt *et al.*, 2008; Gendoo *et al.*, 2016; Haibe-Kains *et al.*, 2012) between tumor samples in discovery cohort of METABRIC. Clustering of the tumor samples relying on the identified TSCs agreed with different breast cancer subtyping methods, SCMOD2 (Haibe-Kains *et al.*, 2012) and PAM50 (Parker *et al.*, 2009), as well as histopathological status of ER and HER2 (Fig. 1A).

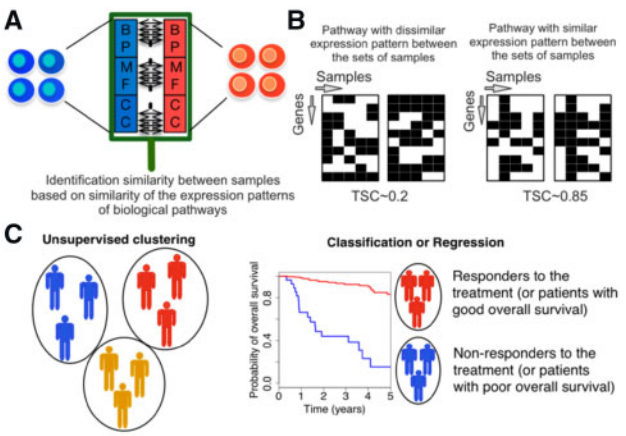

Fig. 1. Schematic representation of Similarity Identification in Gene expressioN (SIGN) and transcriptional similarity coefficient (TSC). (**A**) Gene Ontology (GO) terms (BP: biological processes; MF: molecular functions; CC: cellular components), or other pathway datasets, are used to identify similarity between samples. (**B**) TSC is used to identify similarity of expression pattern of the genes within a given pathway between two sets of samples. (**C**) Collection of TSCs pathways between two sets of samples are used in unsupervised or supervised learning schemes

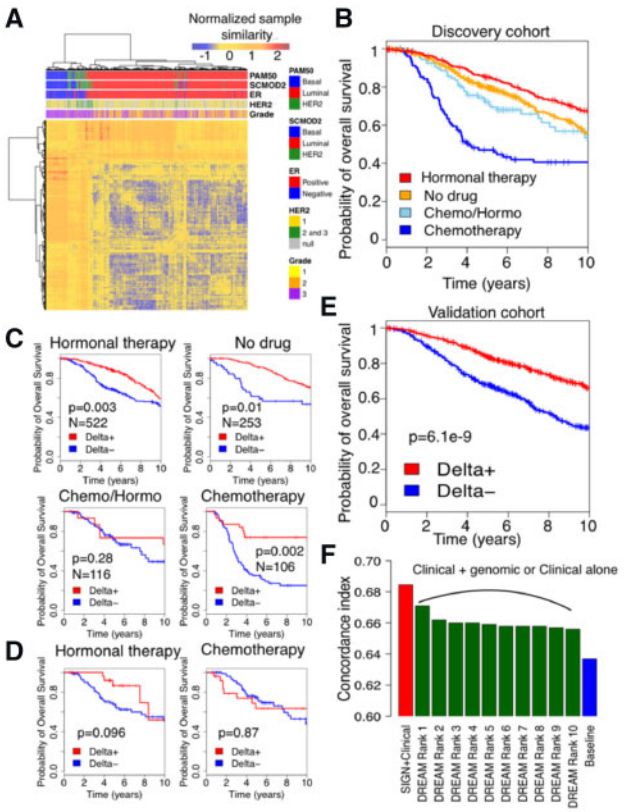

Fig. 2. Patient stratification performance for discovery and validation cohort in Molecular Taxonomy of Breast Cancer International Consortium (METABRIC). (**A**) Clustering of tumor samples in discovery cohort of METABRIC using ESR1 and ERBB2 gene signatures (Gendoo *et al.*, 2016) and TSC as the measure of distance between samples. (**B**) Survival of patients under different treatment regimens in discovery cohort of METABRIC. (**C**) Performance of SIGN on METABRIC discovery cohort stratified by treatments (log-rank test). (**D**) Performance of SIGN on predicting survival of patients under Chemotherapy+Hormonal therapy (Chemo/Hormo) in discovery cohort of METABRIC. (**E**) Performance of SIGN on METABRIC validation cohort. (**F**) Performance of SIGN with respect to the top 10 ranked methods in DREAM challenge (Bilal *et al.*, 2013) in predicting patient survival in validation cohort in METABRIC using Discovery cohort. Baseline is cox model of clinical features

Comparing three subtypes of breast cancer including using TSC illustrated ability of TSC for identifying true similarities between breast tumor samples (Fig. 2A).

We further used overall survival as an endpoint to examine ability of SIGN, as a classification framework, using TSC as a measure of similarity in stratifying patients based on their risk. Breast tumors are traditionally categorized to different subtypes in clinical setting using their histopathological or genomic information (Curtis *et al.*, 2012; Sotiriou *et al.*, 2006). The patients in each subtype are then assigned to different therapeutics available and approved for that subtype (Table 1) (Goldhirsch *et al.*, 2011). Hence, patients under different therapeutic regimen have different underlying biology and will have different survival rate (Fig. 2B). For example, patients with luminal breast cancer receive hormonal therapy and have higher survival rate with respect to patients with basal-like tumor type that receive chemotherapy (Table 1; Fig. 2B). To account for treatment differences, we applied SIGN, using TSC as measure of gene expression pattern similarity between patient gene expression profiles under similar therapeutic regimen in the discovery cohort of METABRIC.

We defined poor and good survival groups as the patients who die the earliest after diagnosis and the patients who survived the longest (10% of the cohort in each category). We then used the poor and good survival patient cohorts to train the model and check validity of the model in stratifying patients. We considered the rest of the population, patients with intermediate survival, as test set to assess the performance of SIGN in patient stratification. For each patient, the difference between the similarity with the poor and good groups is computed so that patients with positive Delta are predicted to have higher survival than the patients with negative Delta. We assessed the prognostic value of Delta using the log-rank test for each treatment group separately (Fig. 2C). SIGN yielded significant prognostic value for all the treatment groups of patients, under different therapeutic regimen, except the patients who received both hormonal therapy and chemotherapy (referred to as Chemo/Hormo; Fig. 2C). The patients under Chemo/Hormo are patients with aggressive luminal breast tumor who received chemotherapy upon showing low response to hormonal therapy. Tumors in this cohort of patients have high heterogeneity, and different underlying biology is potentially responsible for low response rate or even recurrence after hormonal therapy (Shipitsin *et al.*, 2007). We examined use of patient tumor profiles under only hormonal therapy or chemotherapy for predicting survival of patients under Chemo/Hormo. Tumor profile of patients under hormonal therapy were more informative, with respect to patients under chemotherapy, for predicting survival of patients under chemo/hormonotherapy (Fig. 2D).

We validated performance of SIGN trying to predict survival of patients in the validation cohort of METABRIC (Curtis *et al.*, 2012). SIGN could significantly predict survival of breast cancer patients in the validation cohort (Fig. 2E). We further compared the performance of SIGN with the best 10 predictive models from the DREAM challenge (Bilal *et al.*, 2013), to predict breast cancer patient survival using gene expression profiles of tumors, that showcased SIGN as the best model (Fig. 2F). We further showed SIGN outperformed than the best 10 predictive models from the DREAM challenge to predict breast cancer patient survival. The baseline model is the cox regression model trained and tested using clinical features, such as ER status, HER2 status, tumor size, age, grade, Lymph node status and assigned treatment (Bilal *et al.*, 2013). Moreover, PAM50 subtypes of breast tumor samples in discovery cohort of METABRIC were not significant predictors of survival if added to the baseline clinical model (P-value > 0.4).

In conclusion, SIGN is a classification tool that can be applied to predict cell identity and stratify patients based on their survival. With the increasing amount of gene expression and transcriptomic data, SIGN can be used in other applications, such as patient stratification across other cancer types, identification of different cell phenotypes and identification of mechanism of action of drugs using their genomic perturbation data (El-Hachem *et al.*, 2017).

Table 1. Number of patients in each breast subtyping group, identified using SCMOD2 gene signatures (Haibe-Kains *et al.*, 2012), in discovery cohort of METABRIC

Subtype	Hormonal therapy	No drug	Chemo/Hormo	Chemotherapy
Luminal	430	191	68	7
HER2	40	16	13	29
Basal	26	33	22	64

Acknowledgements

The authors thank the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) for sharing their valuable data with the scientific community.

Funding

This study was conducted with the support of the Cancer Research Society and the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. S.A.M.T was supported by Connaught International Scholarships for Doctoral Students, Genome Canada and the Ontario Research Funds. V.S.K.M. was supported by the Cancer Research Society. B.H.K. was supported by the Gattuso-Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre, the Natural Sciences and Engineering Research Council and the Canadian Institutes of Health Research.

Conflict of Interest: none declared.

References

Barbie,D.A. *et al*. (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.

Bilal,E. *et al*. (2013) Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS Comput. Biol*., **9**, e1003047.

Bindra,R.S. *et al*. (2005) Alterations in DNA repair gene expression under hypoxia: elucidating the mechanisms of hypoxia-induced genetic instability. *Ann. N.Y. Acad. Sci*., **1059**, 184–195.

Campbell,P.D. and Marlow,F.L. (2013) Temporal and tissue specific gene expression patterns of the zebrafish kinesin-1 heavy chain family, kif5s, during development. *Gene Expr. Patterns*, **13**, 271–279.

Chatterjee,R. and Vinson,C. (2012) CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. *Biochim. Biophys. Acta*, **1819**, 763–770.

Curtis,C. *et al*. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.

Desmedt,C. *et al*. (2008) Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res*., **14**, 5158–5165.

El-Hachem,N. *et al*. (2017) Integrative cancer pharmacogenomics to infer large-scale drug taxonomy. *Cancer Res*., **77**, 3057–3069.

Gendoo,D.M.A. *et al*. (2016) Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*, **32**, 1097–1099.

Goldhirsch,A. *et al*. (2011) Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann. Oncol*., **22**, 1736–1747.

Haibe-Kains,B. *et al*. (2012) A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst*., **104**, 311–325.

Hänzelmann,S. *et al*. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.

Karr,J.R. *et al*. (2012) A whole-cell computational model predicts phenotype from genotype. *Cell*, **150**, 389–401.

Liberzon,A. *et al*. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Margolin,A.A. *et al*. (2013) Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med*., **5**, 181re1.

Michelson,S. and Young,D.L. (2011) Introduction to systems biology in drug discovery and development. In: Young, D.L. and Michelson, S. (eds) *Systems Biology in Drug Discovery and Development*, pp. 1–5.

Parker,J.S. *et al*. (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol*., **27**, 1160–1167.

Razorenova,O.V. and Giaccia,A.J. (2010) Hypoxia, gene expression, and metastasis. In: *The Tumor Microenvironment*. Springer, NY, pp. 43–58.

Schroder,M.S. *et al*. (2011) survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, **27**, 3206–3208.

Shipitsin,M. *et al*. (2007) Molecular definition of breast tumor heterogeneity. *Cancer Cell*, **11**, 259–273.

Silberberg,Y. *et al*. (2012) Large-scale elucidation of drug response pathways in humans. *J. Comput. Biol*., **19**, 163–174.

Smilde,A.K. *et al*. (2009) Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics*, **25**, 401–405.

Sotiriou,C. *et al*. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst*., **98**, 262–272.

Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet*., **13**, 613–626.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.