


Multiple Omics Data Integration to Identify Long Noncoding RNA Responsible for Breast Cancer–Related Mortality

Tapasree Roy Sarkar^{1,2} , Arnab Kumar Maity³, Yabo Niu² and Bani K Mallick²

¹Department of Biology, Texas A&M University, College Station, TX, USA. ²Department of Statistics, Texas A&M University, College Station, TX, USA. ³Early Clinical Development Oncology Statistics, Pfizer Inc, San Diego, CA, USA.

Cancer Informatics
Volume 18: 1–3
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176935119871933



ABSTRACT: Long non-coding RNAs (lncRNAs) are a large and diverse class of transcribed RNAs, which have been shown to play a significant role in developing cancer. In this study, we apply integrative modeling framework to integrate the DNA copy number variation (CNV), lncRNA expression, and downstream target protein expression to predict patient survival in breast cancer. We develop a 3-stage model combining a mechanical model (lncRNA regressed on CNV and target proteins regressed on lncRNA) and a clinical model (survival regressed on estimated effects from the mechanical models). Using lncRNAs (such as *HOTAIR* and *MALAT1*) along with their CNV, target protein expressions, and survival outcomes from The Cancer Genome Atlas (TCGA) database, we show that predicted mean square error and integrated Brier score (IBS) are both lower for the proposed 3-step integrated model than that of 2-step model. Therefore, the integrative model has better predictive ability than the 2-step model not considering target protein information.

KEYWORDS: Long noncoding RNA, breast cancer, integrative modeling, survival model, TCGA

RECEIVED: July 16, 2019. **ACCEPTED:** July 21, 2019.

TYPE: Short Report

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: T.R.S. was supported through the NIH T32 Training grant (PI: Dr Raymond J Carroll); A.K.M. and B.K.M. were supported through NIH R01CA194391 (PI: Dr B.K.M.).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Tapasree Roy Sarkar, Department of Biology, Texas A&M University, College Station, TX 77843, USA. Email: tsarkar@bio.tamu.edu

Introduction

Several evidences highlight the emerging impact of long non-coding RNAs (lncRNAs) in cancer progression.^{1–4} The aim of this study is to identify the predictive capability of some oncogenic lncRNAs in tumor progression and prognosis of breast cancer.

Breast cancer is the most common malignancy and the leading cause of cancer death in women. By focusing on a single type of genetic alteration such as copy number variation (CNV), scientists have identified significant genes that may contribute to cancer progression.^{5–8} Due to its complexity, the study of cancer should focus on incorporating data from multiple platforms ranging from genes, transcripts, and proteins found in cancer cells,⁹ to whole biological systems, represented by molecular pathways and cell populations.¹⁰ The integration, where multiple levels of omics data (ie, CNV, methylation, and gene expression) are gathered from the same subjects and analyzed, is known as vertical integration.^{10–12}

In this study, we introduce an easy and simplified way to integrate multiple omics data to show that the survival prediction due to the presence of lncRNAs increases significantly in breast cancer. We consider the genomic platform such as CNV, mRNA expression, proteomic platform such as protein expression, and the phenotype such as the survival of the patients. This study focuses only on the lncRNA expressions from The Cancer Genome Atlas (TCGA) breast cancer data. We consider the target protein expressions as proteomics data.

An Integrative Model

We consider a 3-stage model here. Suppose that n is the number of patients, p is the number of lncRNAs, and L is the number of CNV expressions.

The mechanistic model for each lncRNA can be expressed as

$$lncRNA_k = \sum_{l=1}^L c_{kl} \alpha_l + O_k, \quad k = 1, \dots, p \quad (1)$$

where $lncRNA_k$ is the level of gene expression for gene k , $k = 1, \dots, p$, and is of dimension $n \times 1$; c_{kl} is part of the $lncRNA_k$ expression that is attributed to the l th CNV; O_k is the other (remaining) part of the gene expression which is not regulated by CNV and is of dimension $n \times 1$; and α is the regression coefficient vector.

Next, the downstream target protein of each specific lncRNA was identified from PubMed articles, TCGA RNA-Seq database, and other extensive analyses such as differential expression analysis. The mechanistic model for each protein (for every lncRNA) can be expressed as

$$Protein = C\gamma_1 + O\gamma_2 + O^* \quad (2)$$

where $C = (c_{kl})_{p \times L}$ and O^* represents the “other” part of the protein expression that is not regulated by lncRNA. γ_1 and γ_2 are the regression coefficients corresponding to the CNV expressions and the error part from equation (1), respectively.



The clinical component part models the effect of the mechanistic parts of the genes on a clinical outcome of interest and can be written as

$$\log t = \text{IncRNA}_1 \beta_1 + O_2^* \beta_2 + \epsilon \quad (3)$$

where t is the survival outcome, ϵ is the error term, and β_1 and β_2 are the usual regression coefficients corresponding to lncRNA and the estimated error part from equation (2), respectively.

The variable **Protein** represents the vectorized downstream gene effects attributed to protein expressions and is estimated from the second-stage mechanistic model. Therefore, the clinical component additively models the effects of all the gene expressions and their components—derived from different sources (gene expression, CNV) in a unified manner.

Assumptions such as $O \sim N(0, \sigma_1^2 I_n)$ and $O^* \sim N(0, \sigma_2^2 I_n)$ give rise to the usual linear model, whereas we obtain the log-normal accelerated failure time (AFT) model when we assume $\epsilon \sim N(0, \sigma_3^2 I_n)$.

In the presence of right censoring, we observe the tuple (t_i^*, δ_i) , $i = 1, \dots, n$, where $\delta_i = 1$ if the event is observed (death in this case), and 0 otherwise; $t_i^* = \min(t_i, c_i)$ with c_i being the censoring time. A standard statistical software can be used to fit a log-normal AFT model and the other linear regression models.

To quantify the prediction accuracy, we consider a standard comparative predictive approach Brier score (BS)¹³ which uses the predicted survival times

$$\text{BS}(t) = n^{-1} \sum_{i=1}^n \frac{\hat{S}(t | x_i)^2 I(t_i \leq t \wedge \delta_i = 1)}{\hat{K}(t_i)} + \frac{(1 - \hat{S}(1 | x_i))^2 I(t_i > t)}{\hat{K}(t)}$$

where $\hat{K}(\cdot)$ denotes the Kaplan-Meier estimate of the censoring distribution which is based on the observations $(t_i, 1 - \delta_i)$, and $\hat{S}(\cdot)$ stands for the estimated survival function. As the mathematical form suggests, BS provides a numerical comparison between the observed and estimated survival functions. Brier score is defined for each time point t and hence can be added for the entire time range to obtain IBS, $\text{IBS} = \max(t_i)^{-1} \int_0^{\max(t_i)} \text{BS}(t) dt$. We can see that models with smaller scores are preferred. We compute integrated Brier score (IBS) using *ipred* package.¹⁴

Nevertheless, we also compute the prediction square error by comparing the observed data and their posterior predicted values.

From TCGA database, we consider the information of 222 breast tumor samples with their survival data. We observe that at least 82% data are right censored.

Along with the clinical observations, we also collected measurements of 12 lncRNA expressions (Table 1). Among those, we found the CNV information available for 9 genes (or

Table 1. The lncRNA considered for our experiment.

GENE	FUNCTION
<i>BCAR4</i> ^a	Oncogenic, promotes invasion and metastasis ¹⁵
<i>BCYRN1</i>	Oncogenic, promotes tumor progression ¹⁶
<i>GAS5</i> ^a	Tumor suppressor ¹⁷
<i>H19</i> ^a	Oncogenic, promotes proliferation and metastasis ¹⁸
<i>HOTAIR</i> ^a	Oncogenic, promotes EMT, proliferation, and metastasis ¹⁹
<i>MALAT1</i> ^a	Oncogenic, promotes proliferation, invasion, and migration ²⁰
<i>MEG3</i> ^a	Tumor suppressor, induces accumulation of p53 ²¹
<i>PVT1</i> ^a	Oncogenic, promotes tumor progression ²²
<i>SOX2OT</i>	Oncogenic, promotes tumor growth and metastasis ²³
<i>SRA1</i> ^a	Oncogenic ²⁴
<i>UCA1</i> ^a	Oncogenic, promotes cell growth, suppresses the tumor suppressor p27 ²⁵
<i>XIST</i>	Tumor suppressor ²⁶

Abbreviation: EMT, epithelial-mesenchymal transition.

^aThe copy number variation available (among those lncRNAs, *SRA1* transcribes both long noncoding and protein-coding RNAs which are produced by alternative splicing).

Table 2. MSPE and IBS for fitted models in TCGA breast cancer data.

MODELS	MSPE	IBS
2-stage	1.903	0.488
3-stage	1.196	0.395

Abbreviations: IBS, integrated Brier score; MSPE, mean squared prediction error; TCGA, The Cancer Genome Atlas.

lncRNAs). We also consider 64 target protein expressions for these genes.

We apply the integrative modeling in these data and obtain the results shown in Table 2. We notice that the mean squared prediction error and IBS are both lower for the proposed model than for the 2-stage model after omitting the protein expressions from the analysis.

In this article, we have shown that when the contribution of lncRNA's target protein expression measurement is not ignored, then the survival prediction has improved dramatically. Toward this, we have developed a simple yet integrative modeling strategy which borrows strengths from all 3 platforms such as DNA CNV, mRNA expressions for the long noncoding genes, and their target protein expressions to predict the survival of the subjects. We have shown that this integrated model outperforms its closest competitor.

Acknowledgements

The authors thank the editor and the reviewers for their helpful suggestions which substantially improved this paper.

Author Contributions

TRS, AKM, and BKM designed the study. AKM and YN collected and analyzed the data. TRS and AKM wrote the manuscript.

ORCID iD

Tapasree Roy Sarkar  <https://orcid.org/0000-0001-8022-6760>

REFERENCES

- Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. *Cancer Discov.* 2011;1:391-407.
- Zhang S, Wang J, Ghoshal T, et al. lncRNA gene signatures for prediction of breast cancer intrinsic subtypes and prognosis. *Genes.* 2018;9:65.
- Sun M, Wu D, Zhou K, et al. An eight-lncRNA signature predicts survival of breast cancer patients: a comprehensive study based on weighted gene co-expression network analysis and competing endogenous RNA network. *Breast Cancer Res Treat.* 2019;175:59-75.
- The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490:61-70.
- Fan B, Dachrut S, Coral H, et al. Integration of DNA copy number alterations and transcriptional expression analysis in human gastric cancer. *PLoS One.* 2012;7:e29824.
- Bass AJ, Watanabe H, Mermel CH, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nat Genet.* 2009;41:1238-1242.
- Nanjundan M, Nakayama Y, Cheng KW, et al. Amplification of MDS1/EVI1 and EVI1, located in the 3q26.2 amplicon, is associated with favorable patient prognosis in ovarian cancer. *Cancer Res.* 2007;67:3074-3084.
- Scott KL, Kabbarah O, Liang MC, et al. GOLPH3 modulates mTOR signaling and rapamycin sensitivity in cancer. *Nature.* 2009;459:1085-1090.
- Wang W, Baladandayuthapani V, Morris JS, et al. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics.* 2012;29:149-159.
- Chu SH, Huang YT. Integrated genomic analysis of biological gene sets with applications in lung cancer prognosis. *BMC Bioinformatics.* 2017;18:336.
- Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010;86:929-942.
- Liu L, Lei J, Sanders SJ, et al. DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Mol Autism.* 2014;5:22.
- Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med.* 1999;18:2529-2545.
- Peters A, Hothorn T. Ipred: improved predictors (R package version 0.9-6). <https://CRAN.R-project.org/package=ipred>. Updated 2017.
- Xing Z, Park PK, Lin C, Yang L. lncRNA BCAR4 wires up signaling transduction in breast cancer. *RNA Biol.* 2015;12:681-689.
- Ren H, Yang X, Yang Y, et al. Upregulation of lncRNA BCYRN1 promotes tumor progression and enhances EpCAM expression in gastric carcinoma. *Oncotarget.* 2018;9:4851-4861.
- Li S, Zhou J, Wang Z, Wang P, Gao X, Wang Y. Long noncoding RNA GAS5 suppresses triple negative breast cancer progression through inhibition of proliferation and invasion by competitively binding miR-196a-5p. *Biomed Pharmacother.* 2018;104:451-457.
- Barsyte-Lovejoy D, Lau SK, Boutros PC, et al. The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res.* 2006;66:5330-5337.
- Zhang H, Cai K, Wang J, et al. MiR-7, inhibited indirectly by lincRNA HOTAIR, directly inhibits SETDB1 and reverses the EMT of breast cancer stem cells by downregulating the STAT3 pathway. *Stem Cells.* 2014;32:2858-2868.
- Xu S, Sui S, Zhang J, et al. Downregulation of long noncoding RNA MALAT1 induces epithelial-to-mesenchymal transition via the PI3K-AKT pathway in breast cancer. *Int J Clin Exp Pathol.* 2015;8:4881-4891.
- Sun L, Li Y, Yang B. Downregulated long non-coding RNA MEG3 in breast cancer regulates proliferation, migration and invasion by depending on p53's transcriptional activity. *Biochem Biophys Res Commun.* 2016;478:323-329.
- Tang Y, He Y, Zhang P, et al. lncRNAs regulate the cytoskeleton and related Rho/ROCK signaling in cancer metastasis. *Mol Cancer.* 2018;17:77.
- Shi XM, Teng F. Up-regulation of long non-coding RNA Sox2ot promotes hepatocellular carcinoma cell metastasis and correlates with poor prognosis. *Int J Clin Exp Pathol.* 2015;8:4008-4014.
- Leygue E, Dotzlaw H, Watson PH, et al. Expression of the steroid receptor RNA activator in human breast tumors. *Cancer Research.* 1999;59:4190-4193.
- Huang J, Zhou N, Watabe K, et al. Long non-coding RNA UCA1 promotes breast tumor growth by suppression of p27 (Kip1). *Cell Death Dis.* 2015;5:e1008.
- Huang YS, Chang CC, Lee SS, Jou YS, Shih HM. Xist reduction in breast cancer upregulates AKT phosphorylation via HDAC3-mediated repression of PHLPP1 expression. *Oncotarget.* 2016;7:43256-43266.