Transcriptome Analysis Identifies GATA3-AS1 as a Long Noncoding RNA Associated with Resistance to Neoadjuvant Chemotherapy in Locally Advanced Breast Cancer Patients

Contreras-Espinosa, Laura; Alcaraz, Nicolás; De La Rosa-Velázquez, Inti A.; Díaz-Chávez, José; Cabrera-Galeana, Paula; Rebollar-Vega, Rosa; Reynoso-Noverón, Nancy; Maldonado-Martínez, Héctor A.; González-Barrios, Rodrigo; Montiel-Manríquez, Rogelio; Bautista-Sánchez, Diana; Castro-Hernández, Clementina; Alvarez-Gomez, Rosa M.; Jiménez-Trejo, Francisco; Tapia-Rodríguez, Miguel; García-Gordillo, José A.; Pérez-Rosas, Augusto; Bargallo-Rocha, Enrique; Arriaga-Canon, Cristian; Herrera, Luis A.

Published in:
Journal of Molecular Diagnostics

DOI:
10.1016/j.jmoldx.2021.07.014

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Transcriptome Analysis Identifies GATA3-AS1 as a Long Noncoding RNA Associated with Resistance to Neoadjuvant Chemotherapy in Locally Advanced Breast Cancer Patients


From the Unidad de Investigación Biomédica en Cáncer,* Instituto Nacional de Cancerología—Instituto de Investigaciones Biomédicas, National Autonomous University of Mexico (UNAM), Tlalpan, Mexico City, México; The Bioinformatics Centre,† Department of Biology, University of Copenhagen, Copenhagen, Denmark; the Genomics Laboratory,‡ Red de Apoyo a la Investigación, Universidad Nacional Autónoma de México, Tlalpan, Mexico City, México; the Departamento de Tumores Mamarios,§ Instituto Nacional de Cancerología, Tlalpan, Mexico City, México; the Centro de Investigación en Prevención,§ the Patología Molecular e Inmunología,∥ the Clinica de Cáncer Hereditario,∥∥ Instituto Nacional de Cancerología, Tlalpan, Mexico City, México; the Instituto Nacional de Pediatría,‡‡ Coyoacán, Mexico City, México; the Instituto de Investigaciones Biomédicas,‡‡‡ Universidad Nacional Autónoma de México, Tercer Circuito Exterior S/N, Ciudad Universitaria, Mexico City, México; and the Instituto Nacional de Medicina Genómica,¶¶ Tlalpan, Mexico City, México

Breast cancer is one of the leading causes of mortality in women worldwide, and neoadjuvant chemotherapy has emerged as an option for the management of locally advanced breast cancer. Extensive efforts have been made to identify new molecular markers to predict the response to neoadjuvant chemotherapy. Transcripts that do not encode proteins, termed long noncoding RNAs (lncRNAs), have been shown to display abnormal expression profiles in different types of cancer, but their role as biomarkers in response to neoadjuvant chemotherapy has not been extensively studied. Herein, lncRNA expression was profiled using RNA sequencing in biopsies from patients who subsequently showed either response or no response to treatment. GATA3-AS1 was overexpressed in the nonresponder group and was the most stable feature when performing selection in multiple random forest models. GATA3-AS1 was experimentally validated by quantitative RT-PCR in an extended group of 68 patients. Expression analysis confirmed that GATA3-AS1 is overexpressed primarily in patients who were nonresponsive to neoadjuvant chemotherapy, with a sensitivity of 92.9% and a specificity of 75.0%. The statistical model was based on luminal B-like patients and adjusted by menopausal status and phenotype (odds ratio, 37.49; 95% CI, 6.74–208.42; \( P = 0.001 \)); GATA3-AS1 was established as an independent predictor of response. Thus, lncRNA GATA3-AS1 is proposed as a potential predictive biomarker of nonresponse to neoadjuvant chemotherapy. (J Mol Diagn 2021, 23: 1306–1323; https://doi.org/10.1016/j.jmoldx.2021.07.014)

Supported by the National Cancer Institute of Mexico and Consejo Nacional de Ciencia y Tecnología (CONACYT) grant A3-S-46689.

L.C.-E. is a master student in the Programa de Posgrado en Ciencias Biológicas, UNAM, and received a fellowship from CONACYT with Curriculum Vitae Único (CVU)-1003211.

L.C.-E. and N.A. contributed equally to this work.

Disclosures: The authors have submitted a patent at the Mexican Institute of Industrial Property; it is in process with registration key IMPI-00-009, registration number Mx/a/2018/015065, and file Mx/a/2018/015065.

Current address of I.A.D.L.R.-V., Next Generation Sequencing Core Facility, Helmholtz Zentrum Muenchen, Ingolstaedter Landstr, Neuherberg, Germany.

Copyright © 2021 Association for Molecular Pathology and American Society for Investigative Pathology. Published by Elsevier Inc.

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0).

https://doi.org/10.1016/j.jmoldx.2021.07.014
Breast cancer (BC) is one of the main causes of death in women worldwide, with >600,000 deaths annually (https://gco.iarc.fr/today/home, last accessed February 25, 2021). Furthermore, as the leading cause of cancer in women, it constitutes a public health burden. In particular, patients with locally advanced breast cancer (LABC) represent a heterogeneous group with variable local recurrence and global survival. These patients have a significant risk for local recurrence and metastatic progression, in addition to presenting low rates of global survival compared with patients with BC at early stages (stage I to IIA). Hence, a comprehensive approach to LABC patients to achieve local and distant control of disease has become a challenge, as well as monitoring disease progression and treatment efficacy. On the other hand, pathologic complete response (pCR) is one of the most important parameters to consider in patient prognosis. However, different studies have shown that <50% of patients with LABC achieve pCR after neoadjuvant chemotherapy (NAC). NAC was initially used in the context of LABC because it has several advantages, such as making inoperable tumors surgically resectable (stage T4, N2, or N3) and increasing the rates of breast-conserving surgery. Although some biomarkers have proved useful for improving treatment efficiency, most are still in the clinical testing stage and have not yet been approved for standardized use. Expression status of estrogen receptor (ER), progesterone receptor, and ERBB2 [human epidermal growth factor receptor 2 (HER2)] by immunohistochemical evaluation is currently the gold standard for determining management and response to treatment. In addition, gene expression panels, such as Oncotype Dx and Mammaprint, are examples of sets of biomarkers used to provide further clinical support by predicting response to chemotherapy. Nevertheless, these types of tests have mostly been restricted to adjuvant chemotherapy, whereas only a few biomarkers, such as Ki-67 have been proposed for response to NAC.

Recently, in addition to gene expression of coding genes, it has been proposed that noncoding transcripts, such as long noncoding RNAs (lncRNAs), may also serve as molecular markers for BC diagnosis and prognosis. These transcripts are defined as having >200 bases and lacking open reading frames, making them unable to be translated into proteins.

Therefore, transcriptome analysis by ab initio assembly established the potential importance of lncRNAs in cancer, suggesting that this kind of noncoding transcript could be useful in cancer pathogenesis and biomarker development. The lncRNA HOTAI has been proposed as a potential prognostic biomarker, and its overexpression was associated with metastasis-free survival and overall survival (OS), suggesting that this noncoding transcript is a powerful predictor of metastasis and death. Another example is the ER-regulated lncRNA DSCAM-AS1, which is overexpressed in ER-positive tumors and was shown to be of clinical relevance as a good predictor of tumor progression and tamoxifen resistance. In another comprehensive analysis of RNA sequencing (RNA-Seq) data from The Cancer Genome Atlas, Berger et al. constructed gene correlation networks and detected significant gene-lncRNA interactions in breast cancer between coding genes (ESR1 and DKCI) and lncRNAs (NEAT1, TUG1, and TERC). Despite their potential, to date, few studies have investigated or reported lncRNAs as biomarkers of response to systemic therapies in BC and specifically to NAC in BC. Therefore, focusing on lncRNAs may aid in identifying novel and more accurate biomarkers for predicting the response to systemic neoadjuvant therapy in BC.

In this study, using RNA-Seq profiling and machine learning, first, a group of lncRNAs was identified; these lncRNAs are differentially expressed in NAC-resistant LABC Mexican patients compared with NAC-sensitive patients (nonresponders and responders, respectively) and were also predictive and stable features in random forest model. In particular, GATA3-AS1 was identified as a divergent lncRNA that acts as a predictive biomarker of the response to NAC. Expression profiling by RT-qPCR on a larger cohort confirmed that GATA3-AS1 is overexpressed only in nonresponder patients (n = 68). In addition, univariate and multivariate analyses established that GATA3-AS1 distinguishes between nonresponder and responder patients with a sensitivity of 92.9%, a specificity of 75.0%, and an area under the curve (AUC) of approximately 0.90. Finally, GATA3-AS1 is suggested as a novel biomarker for predicting NAC response in patients with LABC and provides the first evidence of this lncRNA as a prediction biomarker for NAC response in breast cancer patients with positive hormonal receptors that correspond to luminal B-like HER2-positive and HER2-negative phenotypes.

Materials and Methods

Breast Sample Collection

Eleven RNA samples, obtained from biopsies of female patients diagnosed with locally advanced mammary adenocarcinoma (stage IIB to IIIIC) belonging to the Mexican National Cancer Institute population who were candidates for the administration of neoadjuvant chemotherapy, were sequenced by RNA-Seq. A validation cohort was collected from another 68 biopsies diagnosed with primary breast tumors, all of whom were patients at the Breast Tumor Division in the Mexican National Cancer Institute between January 2012 and December 2015. Samples were collected from tissue of the initial biopsy (taken with a thick needle) before treatment began (Figure 1). All patients were previously confirmed with primary breast carcinoma and locally advanced disease (clinical stage II to IIIIC) by histologic studies. Patient selection was performed with the support of oncologists and pathologists.

The Journal of Molecular Diagnostics  jmdjournal.org  1307
All patients included in this study received systemic NAC according to the recommendations of the National Comprehensive Cancer Network guidelines. Chemotherapy regimens were based on anthracyclines and taxanes in sequential scheduling, as described below: four taxane cycles every 21 days (paclitaxel, 80 mg/m², on days 1, 8, and 14; or docetaxel, 100 mg/m², on day 1) followed by four doses of fluorouracil-Adriamycin-cyclophosphamide every 21 days (fluorouracil, 500 mg/m², Adriamycin, 50 mg/m², and cyclophosphamide, 500 mg/m², on day 1). Subsequently, all patients underwent local control by mastectomy, breast conservation surgery mastectomy, or breast conservation surgery. At the end of the neoadjuvant regimen, treatment response was evaluated in surgical specimens obtained by the oncopathologist. A pCR was classified according to the most accepted definition: the absence of infiltrating components in the breast and lymph nodes (ypT0/is or ypN0). Luminal B-like phenotype was assigned according to the expression of estrogen and progesterone receptors (>1%), in absence or presence of HER2 overexpression. If HER2 was not overexpressed, it was considered for subtyping the expression of Ki-67 (>20%). In this study, patients were classified as responders if they presented pCR and nonresponders if they did not show pCR. Informed consent was obtained, and this study was approved by the ethical and research committee of the Mexican National Cancer Institute (018/055/DII CEI/1302/18).

External RNA-Seq Data Collection of Breast Cancer Cell Lines

RNA-Seq results of the breast cancer cell lines were obtained from the data set contained in the Breast Cancer Profiling Project, Gene Expression 1: baseline mRNA sequencing on 35 cell lines, which forms part of the Library of Integrated Network-Based Cellular Signatures, which includes 33 breast cancer cell lines and 2 transformed noncancerous breast cell lines (http://lincs.hms.harvard.edu/db/datasets, last accessed May 14, 2020).

RNA-Seq Data Analysis

The quality of the sequencing files was determined with reports generated using FastQC version 0.11.9. Filtering of low-quality reads and adapter removal were performed with trimmomatic version 0.39. Reads were mapped to the human reference genome assembly version hg38 using
STAR aligner version 2.7.1a. Gene expression quantification was performed with featureCounts from the rsubread package version 1.34.7 on the STAR bam files. Gene expression quantification was also performed by aligning to the transcriptome with Salmon version 0.14.1. All gene quantification was performed on gencode version 31 annotations. Tumor purity was inferred using the ESTIMATE algorithm version 2.0.

Differential expression analysis was performed with DESeq2 version 1.24 and was subsequently filtered to analyze only the subset of lncRNAs. The cutoff points were a log2 fold change (FC) value of 1.5 for overexpressed lncRNAs and −1.5 for underexpressed lncRNAs (false discovery rate < 0.05). To corroborate grouping of the patient tumor samples, principal component analysis was performed on the basis of the expression profile of the whole transcriptome, where a batch effect for ER status was detected and added as a covariate in DEseq2 design (~ ER_status + response).

Random forest models were trained on the log-transformed normalized counts (log = FALSE parameter) to predict patient response, using the package randomForest version 4.6.14 in a leave-two-out cross-validation scheme, where two samples (one responsive and one nonresponsive) were always left out. The mean decrease in accuracy was stored for each gene and each fold. To assess the stability of the feature importance, the quartile coefficient of dispersion of the mean decrease in accuracy over all folds was computed for each gene: $(Q_3 - Q_1)/(Q_3 + Q_1)$, where $Q_3$ and $Q_1$ are the third and first quartiles, respectively, of their mean decrease in accuracy.

ConsensusPathDB was performed using clusterProfiler. The Bioconductor package on all genes identified differential expression (DE) in at least one of the two quantification methods using all expressed genes as background. Gene set enrichment analysis was performed on all expressed genes ordered by their DE, $-\log_{10}$ FDR value from DESeq2. Pathway enrichment of lncRNAs was performed using the LncPath (LncRNAs2Pathways) package version 1.1.

RNA Isolation and Quantitative Real-Time PCR Assays

MCF-10A, MCF-7, BT474, and MDA-MB-231 cell cultures were performed following respective ATCC (Manassas, VA) culture protocols. Total RNA was isolated from cultured cells using TRIzol (Thermo Fisher Scientific, Waltham, MA), according to the manufacturer’s instructions. For patient samples, RNA was isolated using the AllPrep kit (Qiagen, Germantown, MD; number 80204); RNA concentration and quality analysis (RNA integrity number value) were performed by a Tape Station 2200 bioanalyzer (Agilent Technologies, Santa Clara, CA). Then, 1 μg RNA was treated with DNase I, RNase free (molecular biology; Thermo Fisher Scientific; reference EN0525). cDNA was synthesized from 1 μg total RNA using a High Capacity cDNA Reverse Transcription (Applied Biosystems, Thermo Fisher Scientific; reference 4368814).

Finally, real-time PCR was performed with SYBR Green/ROX qPCR Master Mix (molecular biology; Thermo Fisher Scientific; reference K0223) on a QuantStudio 3 Real-Time PCR System (Applied Biosystems, Thermo Fisher Scientific). Relative gene expression was determined by fold change calculation ($\Delta$ΔCt) for cell lines and ΔCt for patient samples, normalized to RPS28 as a housekeeping gene. The primers used are listed in Table 1.

RNA-Seq

To ensure good quality of the samples for sequencing, Nanodrop, Qubit 2.0 (Life Technologies, Carlsbad, CA), and the Agilent 2100 Bioanalyzer (Agilent Technologies) were used to detect the purity, concentration, and integrity of RNA samples, respectively. All samples had RNA integrity numbers $>8.0$. A total of 1 μg RNA was used to generate sequencing libraries using the TruSeq Stranded mRNA library prep kit from Illumina, Inc. (San Diego, CA), according to the manufacturer’s instructions. After construction of the libraries, their concentrations and insert sizes (approximately 260 bp) were detected using Qubit 2.0 and the Agilent 2100 Bioanalyzer, respectively. The library was then sequenced using an Illumina HiSeq 2500 sequencer with paired-end 2 × 125 cycles using Illumina TruSeq version 4 sequencing by synthesis (SBS) chemistry and following the manufacturer’s instructions. The depth of sequencing was $>25$ million reads using a HiSeq2500 sequencer by Illumina. RNA-Seq data are available from National Center for Biotechnology Information Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo, accession number GSE159448).

Immunohistochemistry

Primary antibody for GATA3 protein was obtained from BioSB (mouse monoclonal; clone EP368; catalog number BSB 3333; BioSB, Santa Barbara, CA). The reaction was performed on a Benchmark Ultra automated immunohistochemistry device (Ventana Medical Systems, Tucson, AZ). Briefly, the slides were deparaffinized at 72°C for 4 minutes on xylene-free dewaxing reagent (EZprep; Ventana Medical Systems). Antigen retrieval followed, for 8 minutes at 95°C in alkaline solution (CC1; Ventana Medical Systems). Primary antibody was incubated for 16 minutes at 36°C, diluted 1:50 (DaVinci Green diluent; catalog number PD900; Biocare Medical, Pacheco, CA). Finally, enhanced polymer-based detection with diaminobenzidine chromogen was employed (Optview DAB Kit; Ventana Medical Systems). In between all steps, thorough washings were made (Reaction Buffer; Ventana Medical Systems). The slides were then manually counterstained with Harris hematoxylin and mounted with nonaqueous medium.

Statistical Analysis

Descriptive statistics of the main demographic and clinical variables were performed, presenting the median and
Table 1  Primers Used to Detect GATA3-AS1 Expression by RT-qPCR

<table>
<thead>
<tr>
<th>Sequence 5′-3′</th>
<th>Bases, n</th>
<th>Amplicon, bp</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>5′-CCGACACGGAGAGCAGG-3′</td>
<td>20</td>
<td>133</td>
<td>GATA3-AS1 F</td>
</tr>
<tr>
<td>5′-GCTGGGAATGGGGAGGCACT-3′</td>
<td>20</td>
<td>133</td>
<td>GATA3-AS1 R</td>
</tr>
<tr>
<td>5′-CGATCCATCGCCGCAATG-3′</td>
<td>19</td>
<td>101</td>
<td>RPS28 F</td>
</tr>
<tr>
<td>5′-AGCCAGCTCGCCGACAC-3′</td>
<td>18</td>
<td></td>
<td>RPS28 R</td>
</tr>
</tbody>
</table>

F, forward; GATA3-AS1, GATA3 antisense RNA 1; RT-qPCR, quantitative RT-PCR; R, reverse; RPS28, ribosomal protein S28.

interquartile range of continuous variables and the proportion of qualitative variables. Differences in the variables collected in the study groups (response versus no response) were identified according to the type of variable with the U-test or the χ² test. Identification of the effect of overexpression or down-regulation of GATA3-AS1 on clinical response was adjusted to the main clinical variables and was performed by statistical analysis of 66% with the sample available. A P < 0.05 was considered statistically significant. Analysis of variance, followed by the Tukey test, was performed to determine significant differences in GATA3-AS1 expression among different cell lines, in RNA-Seq results analysis, and in quantitative real-time PCR relative expression analysis. For clinical data, GATA3-AS1 expression differences between responder and nonresponder groups were determined by t-test (two tailed, nonpaired, 95% confidence value), assuming variance homogeneity (determined by the Levene test, P = 0.06). This was verified by the Fisher test and χ² test (P < 0.05). In addition, because of sample size, normal distribution was determined by Lilliefors, Kolmogorov-Smirnov, and Shapiro-Wilk tests, indicating that the sample did not show normal distribution (P > 0.05). For that reason, a nonparametric U-test/Wilcoxon test was implemented, showing differences between medians (P < 0.05).

For The Atlas of Noncoding RNAs in Cancer (TANRIC, https://www.tanric.org, last accessed February 25, 2021) survival plot analysis, Cox regression was implemented to determine survival time, and the log-rank test was used to compare survival distribution. Significance was defined as P < 0.05. To determine the specificity and sensitivity of GATA3-AS1 overexpression in neoadjuvant resistance prediction, receiver operating characteristic curve analysis was performed. For this analysis, the STATA software version 14 (StataCorp, College Station, TX) was used.

Results

Clinicopathologic Characteristics of Locally Advanced Breast Cancer Patients

This study was primarily focused on identifying biomarkers to predict the response to NAC in patients with LABC within luminal B-like phenotype. The main features that define each patient are the molecular subtype, age, clinical stage, and response to treatment, among others (Table 2). Expression of hormonal receptors corresponds to histologic classification of the molecular luminal B-like HER2-positive and HER2-negative phenotype (Materials and Methods).

Transcriptome Profiling of lncRNAs in Breast Cancer Patients Is Associated with Nonresponder Patients

With the objective of detecting lncRNAs that could potentially serve as molecular biomarkers of predicted response to NAC, transcriptome profiling was performed in a poly A−enriched population of RNAs using RNA-Seq. The discovery phase included snap-frozen pretreatment core needle biopsies from primary breast cancer patients who responded to NAC (n = 4) and from those who did not respond to the system treatment (n = 7), referred to as responder and nonresponder patients, respectively. Principal component analysis considering the complete expression profile, including mRNAs and lncRNAs, showed no distinction between groups; however, grouping based on ER status was observed and adjusted for these in subsequent analyses (Supplemental Figure S1). Next, DE analysis was performed comparing nonresponder patients with responder patients and identified 69 lncRNAs that were underexpressed and 10 lncRNAs that were overexpressed, according to the established cutoff point (|log2FC| > 1.5 and false discovery rate ≤ 0.05) (Figure 2A). Unsupervised hierarchical clustering of patients shows that these DE lncRNAs place responders and nonresponders into two well-defined groups (Figure 2B).

Given that most of the lncRNAs identified in this study are synthesized from the antisense strand with limited information about their biological functions or molecular characterization, to investigate their association with different cell types, FARN server was used (https://www.cbr.kaust.edu.sa/farna, last accessed February 25, 2021), which infers the function of noncoding RNAs based on the function of their coexpressed genes in multiple data sets (Supplemental Table S1). Results demonstrated that most lncRNAs overexpressed in the nonresponder group participate in processes, such as chromatin remodeling, miRNA interactions, and cancer progression. In contrast, most underexpressed lncRNAs were related to apoptosis and interacting competing endogenous RNA networks. Derived from the fact that most lncRNAs found in this study had not been previously reported or functionally characterized, the study was focused on GATA3-AS1 (FC, 3.02), first because this lncRNA is the only lncRNA that was overexpressed in all nonresponder
patients (Figure 2B), and second, it was the only lncRNA that had been previously well characterized in lymphocytes. In addition, this lncRNA is near GATA3, an important gene in breast cancer. Interestingly, in RNA-Seq differential expression analyses between responder and nonresponder patients, GATA3 is not significantly overexpressed genes in nonresponder patients (Supplemental Table S2).

Overexpressed and Underexpressed LncRNAs, as well as Machine Learning Analysis, Define Patients Who Are Nonresponders to NAC

RNA sequencing is a robust tool for measuring all transcripts, especially when used for identifying differentially expressed genes or noncoding RNA genes, such as lncRNAs, between sample groups. Depending on the pipelines used in RNA-Seq analysis, an incorrect estimate of transcript abundance can be obtained, indicating that differences between pipelines contribute to overall uncertainty in estimates of transcript abundance. Given this premise, the decision to use two different pipelines to analyze RNA sequencing data was taken. Salmon + DESeq2 and STAR + FeaturesCounts + DESeq2 were used to identify differential expression of lncRNAs. With the Salmon pipeline, 70 underexpressed lncRNAs were obtained compared with 64 lncRNAs with the STAR pipeline, and 40 lncRNAs appeared in both methods (Figure 3A and Supplemental Table S3). For overexpressed lncRNAs, when the Salmon pipeline was applied, 10 lncRNAs were found compared with 6 when the STAR pipeline was used, with an overlap of 4 lncRNAs between the two methods (Figure 3B and Supplemental Table S4). Interestingly, one of the overexpressed lncRNAs that coincided in the two pipelines was GATA3-AS1, the same lncRNA overexpressed in all nonresponder patients.

Furthermore, to identify lncRNAs associated with response to NAC, random forest models were built on subsets of patient samples by leaving two patients out (Material and Methods), and the mean decrease in accuracy of each variable was recorded for each model (Figure 3C). The top 50 genes sorted by median mean decrease in accuracy were selected, where only 6 of these were also found to be differentially expressed. This highlights the importance of considering nonlinear predictive models for biomarker discovery in complex and heterogeneous diseases, such as cancer. To assess the stability of the feature importance in the random forest models, the quartile coefficient of dispersion was calculated (Figure 3C). It was observed that GATA3-AS1 exhibited the most stable importance across the different models. Given its high importance scores and stability across the random forest models and its detection as DE, GATA3-AS1 was selected as a top candidate for further validation.
GATA3-AS1 Is Overexpressed in Breast Cancer Cell Lines and in Patients Resistant to NAC

To corroborate the expression profile of GATA3-AS1, RNA-Seq data from breast cancer patients and breast cancer cell lines were used. For this purpose, expression levels of GATA3-AS1 were evaluated from RNA-Seq data in patients derived from the discovery phase. It was found that in samples from nonresponder patients, GATA3-AS1 was overexpressed compared with responder patients, showing basal expression.

**Figure 2** Transcriptome profiling of long non-coding RNAs (lncRNAs) in breast cancer (BC) patients is associated with nonresponder patients. A: Pie charts showing the proportion of underexpressed (left panel) and overexpressed (right panel) RNA biotypes in the nonresponder group. mRNA, lncRNA, and other RNA biotypes are indicated in blue, yellow, and green, respectively. Below is a volcano plot of the identified differentially expressed lncRNAs in nonresponder patients. Blue, yellow, and green dots correspond to mRNA, lncRNA, and other RNAs, respectively. False discovery rate (FDR) < 0.05, log2 fold change > 1.5 for up-regulation and < −1.5 for down-regulation. Red dots correspond to RNA with no significant changes in the nonresponder group. B: Heat map of hierarchical clustering analysis of the top 44 differentially expressed lncRNAs between responder and nonresponder patients. Rows and columns represent differentially expressed lncRNAs and tissue samples, respectively. The color scale represents expression levels. Red and green colors represent up-regulated and down-regulated lncRNAs, respectively (FDR < 0.05). n = 11 (B). FC, fold change.
when visualized by histograms (Figure 4A). Even in the expression analysis from RNA-Seq data quantified by transcript per million, GATA3-AS1 was overexpressed only in nonresponder patients (Figure 4B), suggesting its importance as a prediction biomarker of response to NAC in LABC patients with a luminal B-like phenotype.

In addition, to evaluate oncogenic potential and tissue specificity of GATA3-AS1, expression validation in breast cancer cell lines was needed. RNA-Seq data from the Library of Integrated Network-Based Cellular Signatures (http://lincs.hms.harvard.edu/db/datasets, last accessed May 14, 2020) were used; this library includes 33 breast cancer cell lines and 2 transformed noncancerous breast cell lines (see Materials and Methods). RNA-Seq histograms established that several breast cancer cell lines overexpressing GATA3-AS1 and MCF10A, a nontumorigenic epithelial cell line, exhibited basal expression levels of GATA3-AS1 (Figure 4C). Interestingly, in bar plots where GATA3-AS1 expression levels were measured by reads per kilobase of transcript per million reads mapped (RPKM) (Figure 4D), neoplastic cell lines, such as MDA-MB-157, MDA-MB-436, HCC1395, and CAL51, among others, showed basal expression levels, such as the noncancerous breast cell lines MCF10A and HME1. However, in the other 27 breast cancer cell lines analyzed, GATA3-AS1 was overexpressed; in particular, MCF7 and T47D breast cancer cell lines showed the highest expression levels of this lncRNA, suggesting that GATA3-AS1 is a tissue- and stage-specific overexpressed lncRNA. Regarding GATA3, the adjacent coding gene, its expression was similar between responder and nonresponder patients (Supplemental Figure S2, A and B). Also, GATA3 was widely expressed among the analyzed breast cancer cell lines (Supplemental Figure S2, C and D). Similarly, GATA3 expression was widespread in the different cancer types analyzed (Supplemental Figure S3). On the contrary, analysis of RNA-Seq data from MiTranscriptome (http://mitranscriptome.org, last accessed February 25, 2021) demonstrated that GATA3-AS1 is only overexpressed in breast and bladder cancer tissues, establishing a highly cancer-specific expression pattern and highlighting its importance in breast cancer (Supplemental Figure S4). In addition, this lncRNA is overexpressed in positive hormonal receptor phenotypes...
from RNA-Seq data obtained from TANRIC (Supplemental Figure S5), suggesting a cancer-specific expression pattern in positive hormone receptor phenotypes. Interestingly, GATA3 was overexpressed in hormone-positive subtypes (Supplemental Figure S6) in a similar manner to the expression pattern seen for GATA3-AS1.

Higher Relative Expression Levels of GATA3-AS1 Are Associated with Response to NAC

Once overexpression of GATA3-AS1 was determined in breast cancer cell lines and nonresponder patients through RNA-Seq analysis, expression levels of GATA3-AS1 were examined by RT-qPCR in human breast cancer cell lines and in samples from an independent cohort of patients. For breast cancer cell lines, results showed that the MCF-7 cell line exhibited increased expression levels of GATA3-AS1 by >300-fold, followed by 100-fold overexpression in the MDA-MB-231 and BT474 cell lines, which presented 40-fold lower expression values of GATA3-AS1 than the normal MCF10A cell line (Figure 5A). Hence, results revealed that all breast cancer cell lines analyzed in this study displayed higher expression of GATA3-AS1 when their levels of expression were compared with nonneoplastic cell lines, such as MFC-10A.

Once expression levels of GATA3-AS1 in human breast cancer cell lines were validated, its expression levels were analyzed by RT-qPCR in a validation cohort of 68 patients that included nonresponders and responders to NAC treatment within the luminal B-like phenotype. It was found that
nonresponder patients overexpressed GATA3-AS1, whereas responder patients exhibited underexpression of GATA3-AS1 (Figure 5B). Subsequently, specificity and sensitivity evaluations were performed, constructing an adjusted receiver operating characteristic curve. Figure 5C shows receiver operating characteristic curve analysis to evaluate the predictive capacity of GATA3-AS1 expression by RT-qPCR between nonresponder and responder patients. The analysis demonstrates that GATA3-AS1 has a sensitivity of 92.9% and a specificity of 75% with an AUC of approximately 0.90, which indicates that the use of GATA3-AS1 in the prediction of response to NAC distinguishes between patients who will not respond to NAC from those who will show response to system treatment (nonresponders versus responders, respectively). To determine whether there is a relationship between GATA3-AS1 expression and tumor purity, RNA-Seq data analysis was used. As shown in Figure 5D, there was no relation between GATA3-AS1 expression and tumor purity, suggesting that overexpression of GATA3-AS1 and tumor purity is not associated with NAC response.

In addition, it was determined that GATA3-AS1 and its adjacent coding gene GATA3 were co-expressed in the cohort.
of Mexican patients (Spearman = 0.63) (Supplemental Figure S7A) as well as in the public database TANRIC (Spearman = 0.80) (Supplemental Figure S7B). Interestingly, GATA3 showed overexpression in breast cancer cell lines (Supplemental Figure S8A), but this coding gene was not significantly overexpressed in nonresponder patients compared with responder ones (Supplemental Figure S8B). Furthermore, GATA3 showed low sensitivity of 52.4%, with a specificity of 73.9% and an AUC of 0.60 (Supplemental Figure S8C). Besides, no relationship was found between GATA3 expression and tumor purity (Supplemental Figure S8D). Finally, GATA3 protein expression by immunohistochemistry was not shown to distinguish responders from nonresponders (Supplemental Figure S9), suggesting that in LABC within the luminal B-like phenotypes GATA3 is not a good molecular marker for predicting the response to NAC contrary to GATA3-AS1. Taken together, these results suggest that GATA3-AS1 is overexpressed only in nonresponder patients, indicating that GATA3-AS1 represents a potential biomarker of NAC resistance in patients with LABC within the luminal B-like phenotypes who do not respond to systemic treatment.

**GATA3-AS1 Is a Predictive Molecular Biomarker of Response to NAC in Breast Cancer Patients**

To determine the role of GATA3-AS1 in prognosis, a Kaplan-Meier curve was generated. No relationship was detected for GATA3-AS1 overexpression in OS when the cohort of 68 Mexican patients was used, suggesting that this lncRNA is not a prognostic factor in LABC patients within the luminal B-like phenotype (Supplemental Figure S10). Even when the relationship of GATA3-AS1 with OS from TANRIC RNA-Seq luminal B phenotypes (Supplemental Figure S11A) and in all molecular subtypes (Supplemental Figure S11B) was validated, no relationship with OS was observed, confirming that GATA3-AS1 is not related to prognosis. Therefore, these results suggest that GATA3-AS1 is not a prognostic biomarker in the cohorts analyzed.

Furthermore, multivariate logistic regression showed that GATA3-AS1 overexpression was an independent predictor of response adjusted by menopausal status and phenotype, proving to be statistically significant in this model with 37.49-fold (95% CI, 6.74–208.42) more probability in nonresponders with GATA3-AS1 compared with responder patients who did not express GATA3-AS1 (Table 3). Finally, from all analyses and results obtained in this study, GATA3-AS1 is proposed as a novel divergent lncRNA that may serve as a potential predictive molecular biomarker of response to NAC, which could be included in clinical practice to manage Mexican LABC patients with a luminal B-like phenotype.

### Functional Analysis of Long Noncoding RNAs in Nonresponder Patients to NAC

To identify potential affected pathways in nonresponder patients, IncRNA Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis was performed with LncRNAs2Pathways using the top 79 differentially expressed IncRNAs between nonresponder and responder groups from RNA-Seq data analyzed. It was found that the term taste transduction was significantly enriched for underexpressed IncRNAs (P < 0.05), whereas olfactory transduction and renin-angiotensin system were significantly enriched for overexpressed IncRNAs (P < 0.05) (Figure 6) and found that the most significantly enriched terms were taste transduction, Parkinson disease, Alzheimer disease, oxidative phosphorylation, and regulation of autophagy (Figure 6). These results indicate that IncRNAs may influence these pathways and biological processes associated with NAC response; however, more analysis is needed to confirm this.
Functional and Pathway Analysis of mRNAs in Patients Nonresponsive to NAC

Pathway enrichment analyses were performed using clusterProfiler on a collection of pathways from multiple sources and were collected by ConsensusPathDB to determine functional enrichment analysis of the differentially expressed mRNAs in RNA-Seq data. The top 20 functionally enriched biological processes obtained from clusterProfiler analysis under Gene Ontology terms are indicated in a bar chart of the 214 differentially expressed mRNAs. The most significantly enriched biological processes were mesodermal commitment pathway, transmission across chemical synapsis, and neuronal system (P < 0.05) (Figure 7A). The interaction networks between enriched biological processes were analyzed, yielding an interaction network among the biological processes related to mesodermal commitment pathway and cell differentiation, as well as neuronal system and transmission across chemical synapsis, among others (Figure 7B). To identify genes involved in each functionally enriched biological process, a heat map of enriched mRNAs was constructed (Figure 7C). Finally, the network of gene pathways with significantly altered expression (FC > 1.5 and P < 0.05) was delineated using clusterProfiler (Figure 7D). These results suggest that mRNAs may influence these pathways and biological processes associated with NAC response.

Discussion

In general, lncRNAs are differentially expressed among human tissues. These expression profiles are related to the different roles of lncRNAs in the cellular physiology of tissues, and alterations in the transcriptional rate of lncRNAs could lead to the development of pathologies, such as cancer. Transcriptomic analysis of lncRNAs in different human neoplastic tissues demonstrated that these transcripts are differentially expressed among human cancers. In particular, RNA-Seq and microarray studies have shown that breast cancer is characterized by specific lncRNA expression profiles among molecular subtypes.
Figure 7  Functional and pathway analysis of mRNAs in nonresponder patients to neoadjuvant chemotherapy. A: Top 20 enriched biological processes of differentially expressed mRNAs in the nonresponder group. The significance of the $P$ value is indicated by the intensity of the red color. B: Interaction networks of the top 20 enriched biological processes. Nodes are indicated with dots, and relationships between nodes are indicated by gray lines. C: Enrichment map for differentially expressed mRNAs in the nonresponder group. Columns show the mRNA symbol, and rows show the enriched term for biological processes. The enrichment fold change is indicated by the intensity of the color scales. D: Network of gene pathways showing differentially expressed mRNAs in the nonresponder group. Gene names are indicated in red and green dots, whereas gene category is indicated in gray dots. Nodes are indicated with dots, and relationships between nodes are indicated by gray lines. The number of genes of every node is indicated by the size of the dot, and the color of dots indicates the fold change (maximum fold change in red, and minimum fold change in green).
(luminal, HER2 enrichment, and basal)\textsuperscript{15} that have been related to prognostic variables, such as OS, recurrence, progression, metastasis risk, treatment efficacy, and resistance to treatment.\textsuperscript{44,45} Because of their association with treatment resistance, several lncRNAs have been identified for potential predictive use in endocrine therapy, such as \textit{HOTAIR},\textsuperscript{26} \textit{lncRNA-ATB} for antibody administration,\textsuperscript{47} and \textit{LINC00472} for adjuvant chemotherapy.\textsuperscript{38} However, little is known about the association of lncRNAs and response to neoadjuvant chemotherapy, which is the standard treatment for breast cancer patients with locally advanced disease.\textsuperscript{89,90} Several studies based on microarray assays have shown that lncRNAs are related to pathologic complete response in breast cancer patients in all molecular subtypes.\textsuperscript{24,51} Currently, a few studies on transcriptome analysis by RNA-Seq in NAC resistance have been performed\textsuperscript{52–54} to better understand the molecular biology of resistance to NAC in LABC patients.

In this study, a subset of long noncoding genes differentially expressed in LABC within luminal-like B phenotype patients who did not respond to NAC treatment were identified. From this subset, only the most characterized lncRNAs were selected, and identified the divergent lncRNA \textit{GATA3-AS1}, which had been previously reported to be overexpressed in breast cancer patients.\textsuperscript{51} In addition, \textit{GATA3-AS1}, \textit{RP11-279F6}, and \textit{AC017048} showed specific and high expression levels in ER-positive (ER\textsuperscript{+}) compared with ER-negative (ER\textsuperscript{-}) cancers and normal breast tissue samples.\textsuperscript{91} \textit{GATA3-AS1} belongs to a class of divergent lncRNAs; it is located on chromosome 10 and is approximately 2 Kb. The transcription start site of this noncoding gene is approximately 1 Kb from the first exon of the adjacent \textit{GATA3} gene and is synthesized from the antisense strand and consists of two exons.\textsuperscript{92} Experimental validation by RT-qPCR of \textit{GATA3-AS1} expression in breast cancer cell lines MCF-7, MDA-MB-231, and BT474 demonstrated an association between \textit{GATA3-AS1} overexpression and neoplastic disease in mammary cells. This was further corroborated in samples from LABC patients in a discovery phase and an independent breast cancer patient cohort (validation phase), where it was observed that overexpression of \textit{GATA3-AS1} is associated with patients who do not respond to systemic treatment, which is indicative of the relationship between \textit{GATA3-AS1} overexpression and resistance to NAC. Moreover, multivariate logistic regression analysis was performed to demonstrate this association, showing that \textit{GATA3-AS1} overexpression is an independent predictor of response and proving that \textit{GATA3-AS1} is statistically significant in this model, with 37.49\% (95\% CI, 6.74\%–208.42\%). Moreover, high sensitivity and specificity (92.9\% and 75\%, respectively) with an AUC of approximately 0.90 suggested that \textit{GATA3-AS1} is a potential biomarker for predicting NAC response in clinical practice to improve therapy for LABC within luminal B-like phenotype patients. Several preclinical studies have identified lncRNAs that proposed as response biomarkers despite the reduced use of patient samples, such as \textit{UCAI},\textsuperscript{23} \textit{TP53COR1} (or lncRNA-p21),\textsuperscript{36} \textit{GAS5},\textsuperscript{21} and \textit{HOTAIR},\textsuperscript{36} but clinical application of each lncRNA analyzed is relevant to each type of treatment in the cancer patients in whom they were evaluated. The lncRNA \textit{GATA3-AS1}, despite having been validated in a small number of patients, might be an important biomarker of response to chemotherapy similar to the previously mentioned lncRNAs. Further studies will be needed to accurately evaluate the potential use of this lncRNA as a clinical predictive biomarker in the luminal B-like phenotype in a large cohort of patients.

Furthermore, patients in the luminal B group are of clinical interest because luminal B is a specific subtype that contains clinically aggressive ER-positive breast cancers in which patients present an intermediate prognosis with a high variety of responses to different treatments.\textsuperscript{58} In particular, it was determined that \textit{GATA3-AS1} is overexpressed in luminal B-like patients who do not respond to neoadjuvant chemotherapy. A clear association was observed between \textit{GATA3-AS1} overexpression and nonresponders; however, no association was found between patient outcomes when OS was assessed against \textit{GATA3-AS1} overexpression. Interestingly, the expression profile of \textit{GATA3-AS1} is similar to that of lncRNAs, such as \textit{DSCAM-AS1}, where its expression is not directly associated with prognosis but rather with response to treatment or progression of the disease.\textsuperscript{93} Moreover, both lncRNAs exhibited higher expression in luminal B patients analyzed. It is also necessary to include patients diagnosed with luminal A breast cancer subtypes to elucidate whether it is possible to extend the predictive value of \textit{GATA3-AS1} in this group and to generalize its use as a molecular biomarker. LncRNAs are known useful factors in treatment selection, independently of molecular cancer subtype;\textsuperscript{15} however, this will likely require additional large cohorts to be analyzed that include both hormone-positive subtypes.

Conversely, the coding gene \textit{GATA3} has a wide expression pattern in patients and breast cancer cell lines, as observed by RNA-Seq analysis. As previous studies have demonstrated, mRNAs tend to have less tissue- and stage-specific expression\textsuperscript{92,95,60} in contrast to lncRNAs, which tend to have more tissue-specific\textsuperscript{19,91,92} and stage-specific expression in disease, which is one of the main reasons lncRNAs have been proposed as molecular biomarkers in cancer.\textsuperscript{16,63,64} Alternatively, results by receiver operating characteristic curve showed that \textit{GATA3} had sensitivity of 52.4\% and specificity of 73.9\% with an AUC of 0.60 in RT-qPCR expression analyses; the results suggest that \textit{GATA3} mRNA expression has low ability to distinguish responders from nonresponders to NAC due to its low sensitivity. Currently, \textit{GATA3} is a molecular marker in breast cancer that has had a controversial clinical role.\textsuperscript{96} \textit{GATA3} protein expression has been associated with a favorable prognosis and increased survival in patients with invasive breast carcinoma; however, \textit{GATA3} has not been shown to be a
reliable prognostic factor regardless of ER status.65–67 Likewise, clinical data regarding the role of GATA3 in treatment response prediction have also been controversial.68,69 There is evidence suggesting that GATA3 could be used as a biomarker for predicting response to NAC. It was observed that absence of GATA3 is an independent pathologic complete response predictor to neoadjuvant chemotherapy through multivariate analysis, suggesting that GATA3 might be clinically useful as a predictor of poor response to chemotherapy.70 However, these results did not evaluate the sensitivity and specificity of GATA3; thus, more studies are necessary to determine the implications of GATA3 as a predictor biomarker to NAC response. To date, there is controversy over the use of GATA3 as a predictive biomarker for NAC69–71 and, because of the low sensitivity found in the present study, the coding gene GATA3 cannot be proposed as a biomarker for NAC resistance, even though it is co-expressed with GATA3-AS1. The results obtained in this study indicate that there is a moderate correlation between GATA3-AS1 and GATA3 mRNA expression. However, in the literature, it has been suggested that in breast cancer, GATA3-AS1 does not regulate GATA3 mRNA expression; instead, GATA3-AS1 regulates GATA3 protein accumulation level through a degradation mechanism.72 Therefore, GATA3 protein levels, measured by immunohistochemistry, cannot distinguish patients who will be resistant to neoadjuvant chemotherapy, but the possibility that GATA3 protein accumulation levels could be regulated by GATA3-AS1, as has been demonstrated in triple-negative breast cancer cell lines, cannot be ruled out.73 However, this is beyond the scope of this study, and further research is needed. Thus, in locally advanced breast cancer luminal B-like phenotype, regulation of GATA3 mRNA could not depend on GATA3-AS1 despite there being a positive correlation in the expression levels of both transcripts.

On the other hand, there is scientific evidence that altered gene expression in breast cancer tumors is involved in neuronal-related pathways75 and processes.74–76 These results showed that tumors of nonresponders to NAC overexpressed IncRNA genes that are related to Kyoto Encyclopedia of Genes and Genomes pathways, such as taste transduction and Parkinson disease. This is in accordance with other reports in which IncRNA expression in breast cancer cells is related to the dysregulation of neuronal-related pathways, such as IncRNA IRAIN and its targets involved in cholinergic synapses77 and the IncRNA-mRNA coexpression network, which is related to taste transduction in docetaxel-resistant breast cancer cell lines.78 In addition, this study demonstrated that the expression profile of mRNA in breast cancer tumors is involved in the neuronal system signaling, as has been described for other processes, such as cranial nerve and neural crest development in breast cancer patients.76,79 Together, these data indicate that genes involved in neuronal processes are regulated by IncRNAs in breast cells and, in cancer conditions, might acquire oncogenic potential, contributing to breast cancer progression, resistance to treatment, or metastasis development, as shown for IncRNA BORG, which is associated with brain metastasis.80

Finally, the divergent IncRNA GATA3-AS1 is a non-coding transcript described in T lymphocytes with an important role in T-cell differentiation,78 and it has been associated with respiratory pathologies, such as asthma and rhinitis.81 In this study, it was found that GATA3-AS1 is also expressed in mammary cells, but until now, the function of GATA3-AS1 in breast tissue has not been described, as seen in hepatocellular carcinoma.82 Further research is needed to establish the function and molecular mechanisms of GATA3-AS1 in mammary tissue and how it associates with breast neoplastic disease and NAC resistance.

In conclusion, the presence of an IncRNA profile capable of defining patients nonresponsive to NAC in LABC luminal B-like patients by RNA-Seq analysis was demonstrated. Particularly, the divergent IncRNA GATA3-AS1 showed high specificity and sensitivity associated with its predictive value in nonresponders to NAC treatment, making it the first molecular biomarker with a potential use in clinical practice in the prediction of NAC treatment response in breast cancer. Further investigation is needed to discover whether its predictive value is applicable to other molecular subtypes and to uncover the molecular mechanisms of this IncRNA in NAC resistance.

Acknowledgment

We thank the National Cancer Institute of Mexico for support.

Supplemental Data

Supplemental material for this article can be found at http://doi.org/10.1016/j.jmoldx.2021.07.014.

References


Contreras-Espinosa et al.


67. Afzalajaran F, Sadir AS, Savas S, Pasdar A: GATA3 somatic mutations are associated with clinicopathological features and