Data in Brief 37 (2021) 107241

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Datasets exploring putative lncRNA-miRNA-mRNA axes in breast cancer cell lines

Marie-Claire D Wasson[a,1], Justin M Brown[a,1], Jaganathan Venkatesh[a], Wasundara Fernando[a], Paola Marcato[a,b,*]

[a] Department of Pathology, Dalhousie University, Halifax, NS, B3H 4R2, Canada
[b] Department of Microbiology & Immunology, Dalhousie University, Halifax, NS, B3H 4R2, Canada

ARTICLE INFO

Article history:
Received 26 May 2021
Revised 9 June 2021
Accepted 21 June 2021
Available online 24 June 2021

Keywords:
Long non-coding RNA
microRNA
mRNA
Breast cancer
Regulation
Oncogenesis

ABSTRACT

Long non-coding RNA (lncRNA)/microRNA (miRNA)/messenger RNA (mRNA) interactions regulate oncogenesis and tumour suppression in breast cancer. Oncogenic lncRNA/miRNA/mRNA axes may offer novel therapeutic targets; therefore, identifying such axes is a clinically relevant undertaking. To explore miRNAs regulated by oncogenic lncRNAs, we queried the NCBI Gene Expression Omnibus (GEO) database to find datasets that profiled gene expression changes upon lncRNA knockdown in breast cancer. We identified four microarray datasets that permitted our interrogation of genes regulated by lncRNAs LincK, LincIN, SPRY4-IT1 and AC009283.1. We specifically analysed changes in miRNA transcripts within these datasets to study miRNAs regulated by each of the four lncRNAs. We subsequently identified the predicted mRNA targets for these miRNAs to uncover possible lncRNA/miRNA/mRNAs axes in breast cancer. These axes may be candidates for future investigation of gene regulation in breast cancer.

DOI of original article: 10.1016/j.canlet.2021.04.002
* Corresponding authors.
E-mail address: paola.marcato@dal.ca (P. Marcato).
Social media: (P. Marcato)
[1] These authors contributed equality to this work and are co-first authors

https://doi.org/10.1016/j.dib.2021.107241

Specifications Table

Subject	Cancer research
Specific subject area	Long non-coding RNA and microRNA interactions in breast cancer
Type of data	Figure Table
How data were acquired	The gene expression profiling by array datasets were collected from NCBI's Gene Expression Omnibus (GEO) database (GSE79214, GSE62507, GSE134254, GSE109007). R packages biomaRt and hoardeR were used to identify lncRNA/miRNA/mRNA axes from these datasets.
Data format	Analyzed raw .CEL files from previously conducted array profiling studies.
Parameters for data collection	Gene expression profiling by array studies in which a single long non-coding RNA was knocked down in a breast cancer cell line model were considered.
Description of data collection	GSE79214: Transcriptome analysis was performed using the GeneChip® Human Gene 2.0 ST Array. GSE62507: Agilent-026652 Whole Human Genome Microarray 4 × 44K v2 (Probe Name version) GSE134254: The Affymetrix Human Transcriptome Array 2.0 [transcript (gene) version] GSE109007: Transcriptome analysis was performed using the GeneChip® Human Gene 2.0 ST Array platform
Data source location	GSE79214: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79214 GSE62507: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62507 GSE134254: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134254 GSE109007: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109007
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/wstfmr4z57.1 Direct URL to data: https://data.mendeley.com/datasets/wstfmr4z57/draft?a=c5e63c44-a3a2-4486-9050-c0dc3b68a3aa
Related research article	Venkatesh J, Wasson MD, Brown JM, Fernando W, Marcato P. LncRNA-miRNA axes in breast cancer: Novel points of interaction for strategic attack. Cancer Lett. 2021 Jul 1;509:81-88. doi:10.1016/j.canlet.2021.04.002.

Value of the Data

- This data is of value for cancer researchers studying lncRNA/miRNA/mRNA interactions in breast cancer. These axes may represent novel therapeutic targets.
- This data may be used to identify specific miRNAs regulated by lncRNAs LincIN, SPRY4-IT1, AC009283.1 and LincK in the associated breast cancer cell lines and their downstream mRNA targets. This may inform lncRNA function.
- These datasets provide the predicted mRNA targets for the majority of miRNAs in the Affymetrix Human Gene 2.0 ST Array [transcript (gene) version], Aligent-026652 Whole Human Genome Microarray 4 × 44k v2 (Probe Name Version) and Affymetrix Human Transcriptome Array 2.0 [transcript (gene) version] arrays.
- Our study provides a foundation for similar analyses using primary or secondary microarray datasets. Our methodology for predicting lncRNA/miRNA/mRNA axes can be applied to explore the regulatory roles of additional lncRNAs of interest.

1. Data Description

Long non-coding RNAs (lncRNAs), microRNAs (miRNAs) and messenger RNAs (mRNAs) form specific interaction networks to regulate gene expression in cancer [1,2]. Here, we have

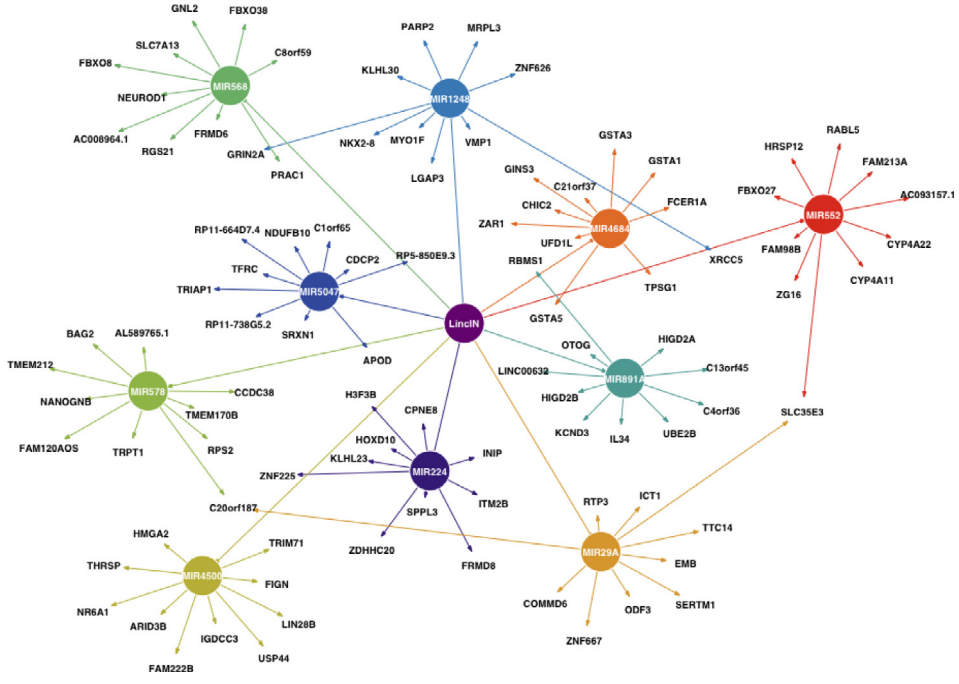

Fig. 1. Predicted LincIN/miRNA/mRNA axes in MDA-MB-231 breast cancer cells. The ten miRNAs with the highest absolute $\log_2$ fold expression change and p-value < 0.05 upon LincIN knockdown by shRNA #2 in MDA-MB-231 cells (obtained from the NCBI GEO GSE79214 dataset) with mRNA targets (predicted via TargetScan) are shown. Up to ten of the strongest mRNA targets for each of the miRNAs (predicted via TargetScan) are plotted. Together, these interactions reveal one hundred potential miRNA/mRNA interactions stemming from LincIN in MDA-MB-231 cells. The miRNAs, their fold expression change induced by LincIN knockdown, and their predicted mRNA targets shown in this plot are detailed in Dataset 1A and Dataset 1B. This graph was generated using the igraph package in Rv4.0.4 [8].

conducted a network analysis to identify potential lncRNA/miRNA/mRNA axes in breast cancer for oncogenic lncRNAs LincK, LincIN, SPRY4-IT1 and AC009283.1. We provide data illustrating changes in miRNA expression following knockdown of each of the four lncRNAs in breast cancer cell lines to identify potential lncRNA/miRNA interactions. We then predicted the mRNA targets of the lncRNA-regulated miRNAs. Together, these two sets of results reveal potential lncRNA/miRNA/mRNA axes in breast cancer. These data offer a more complete look into the regulatory functions of lncRNAs LincK, LincIN, SPRY4-IT1 and AC009283.1 and may inform the identification of novel targetable axes.

Datasets 1 – 4 contain the probe IDs, HGNC gene symbol, $\log_2$ fold change and p-value for each miRNA identified in the primary microarray. They also contain predicted mRNA targets for miRNAs in the TargetScan database [3].

Dataset 1A. Potential LincIN/miRNA/mRNA axes in MDA-MB-231 breast cancer cells. The $\log_2$ fold change in expression of 1963 miRNAs following LincIN knockdown by shRNA #2 in MDA-MB-231 cells (GSE79214 [4]). This dataset can be used to predict mRNA targets for miRNAs in the Affymetrix Human Gene 2.0 ST Array [transcript (gene) version].

Dataset 1B. Potential LincIN/miRNA/mRNA axes in MDA-MB-231 breast cancer cells (miRNAs that meet threshold criteria). The $\log_2$ fold change in expression of 225 miRNAs following LincIN knockdown by shRNA #2 that meet the following criteria: $\log_2$ fold expression change $\geq$0.5 or $\leq$–0.5 and p-value < 0.05. The top hundred LincIN/miRNA/mRNA axes, predicted *in silico*, are depicted in Fig. 1.

Dataset 2A. Potential SPRY4-IT1/miRNA/mRNA axis in MDA-MB-231 breast cancer cells. The $\log_2$ fold change in expression of 213 miRNAs following SPRY4-IT1 knockdown in

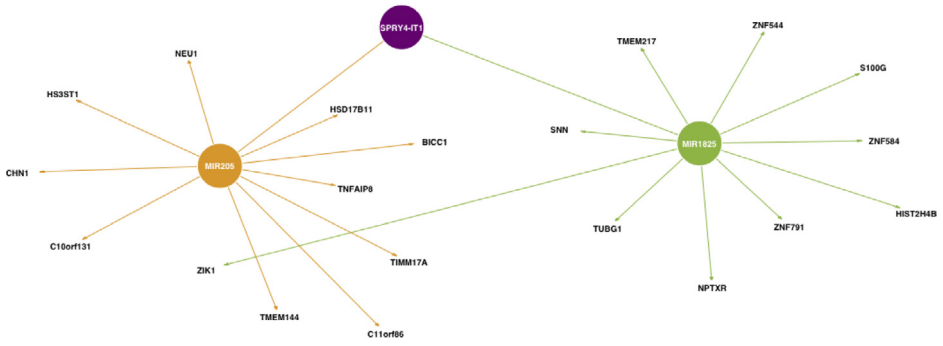

Fig. 2. Predicted SPRY4-IT1/miRNA/mRNA axes in MDA-MB-231 breast cancer cells. Two miRNAs with significant fold expression changes ($p < 0.05$) upon SPRY4-IT1 shRNA knockdown in MDA-MB-231 cells (obtained from the NCBI GEO GSE62507 dataset) with mRNA targets (predicted via TargetScan) are shown. Up to ten of the strongest mRNA targets for each of the miRNAs (predicted via TargetScan) are plotted. Together, these interactions reveal twenty potential miRNA/mRNA interactions stemming from SPRY4-IT1 in MDA-MB-231 cells. The miRNAs, their fold expression change induced by SPRY4-IT1 knockdown, and their predicted mRNA targets shown in this plot are detailed in Dataset 2A and Dataset 2B. This graph was generated using the igraph package in Rv4.0.4 [8].

MDA-MB-231 cells (GSE62507 [5]). This dataset can be used to predict mRNA targets for miRNAs in the Agilent-026652 Whole Human Genome Microarray 4 × 44K v2 (Probe Name version) array.

Dataset 2B. Potential SPRY4-IT1/miRNA/mRNA axis in MDA-MB-231 breast cancer cells (miRNAs that meet threshold criteria). The $\log_2$ fold change in expression of 2 miRNAs following SPRY4-IT1 knockdown that meet the following criteria: $\log_2$ fold expression change $\geq$0.5 or $\leq$−0.5 and p-value < 0.05. The top twenty SPRY4-IT1/miRNA/mRNA axes, predicted *in silico*, are depicted in Fig. 2.

Dataset 3A. Potential AC009283.1/miRNA/mRNA axis in SKBR3 breast cancer cells. The $\log_2$ fold change in expression of 1874 miRNAs following SPRY4-IT1 knockdown in SKBR3 cells (GSE134254 [6]). This dataset can be used to predict mRNA targets for miRNAs in the Affymetrix Human Transcriptome Array 2.0 [transcript (gene) version].

Dataset 3B. Potential AC009283.1/miRNA/mRNA axis in SKBR3 breast cancer cells (miRNAs that meet threshold criteria). The $\log_2$ fold change in expression of eight miRNAs following AC009283.1 knockdown that meet the following criteria: $\log_2$ fold expression change $\geq$0.5 or $\leq$−0.5 and p-value < 0.05. The top hundred AC009283.1/miRNA/mRNA axes, predicted *in silico*, are depicted in Fig. 3.

Dataset 4A. Potential LincK/miRNA/mRNA axis in MCF7 breast cancer cells. The $\log_2$ fold change in expression of 1963 miRNAs following LincK knockdown by shRNA #1 in MCF7 cells (GSE109007 [7]). This dataset can be used to predict mRNA targets for mRNAs in the Affymetrix Human Gene 2.0 ST Array [transcript (gene) version].

Dataset 4B. Potential LincK/miRNA/mRNA axis in MCF7 breast cancer cells (miRNAs that meet threshold criteria). The $\log_2$ fold change in expression of 45 miRNAs following LincK knockdown by shRNA #1 that meet the following criteria: $\log_2$ fold expression change $\geq$0.5 or $\leq$−0.5 and p-value < 0.05. The top hundred LincK/miRNA/mRNA axes, predicted *in silico*, are depicted in Fig. 4.

2. Experimental Design, Materials and Methods

The data analysis pipeline used to profile putative lncRNA/miRNA/mRNA axes is summarized in Fig. 5. To identify potential lncRNA/miRNA/mRNA axes in breast cancer, we searched for studies that profiled changes in gene expression in the context of lncRNA knockdown in a

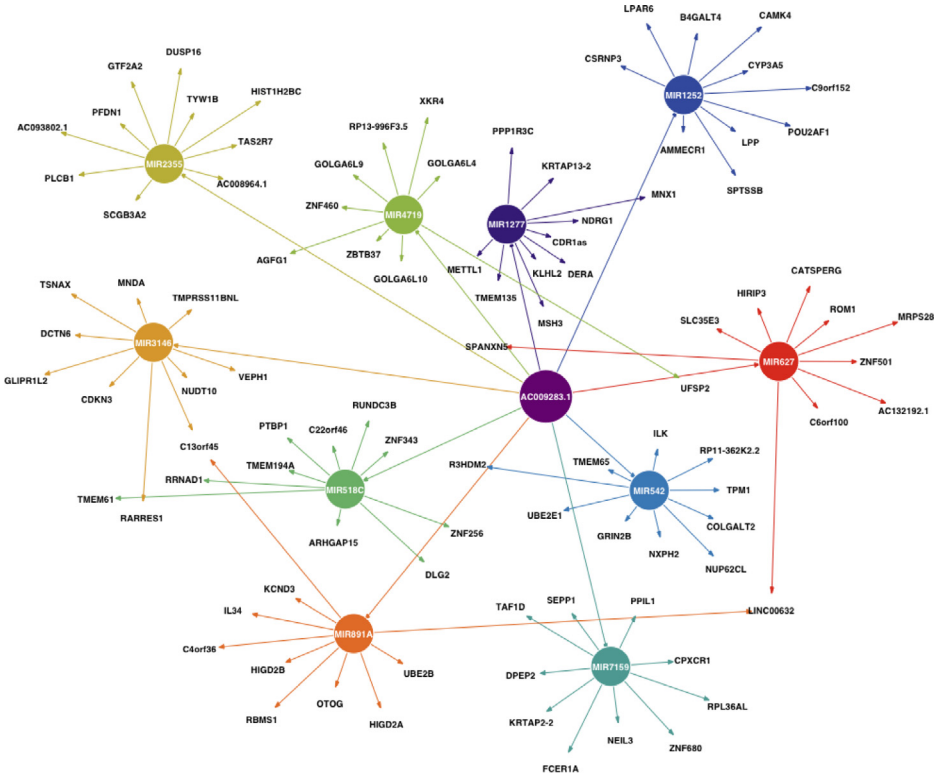

Fig. 3. Predicted AC009283.1/miRNA/mRNA axes in SKBR3 breast cancer cells. The ten miRNAs with the highest absolute $\log_2$ fold expression change and p-value < 0.05 upon lncRNA AC009283.1 knockdown by shRNA in SKBR3 cells (obtained from the NCBI GEO GSE134254 dataset) with mRNA targets (predicted via TargetScan) are shown. Up to ten of the strongest mRNA targets for each of the miRNAs (predicted via TargetScan) are plotted. Together, these interactions reveal one hundred potential miRNA/mRNA interactions stemming from AC009283.1 in SKBR3 cells. The miRNAs, their fold expression change induced by SPRY4-IT1 knockdown, and their predicted mRNA targets shown in this plot are detailed in Dataset 3A and Dataset 3B. This graph was generated using the igraph package in Rv4.0.4 [8].

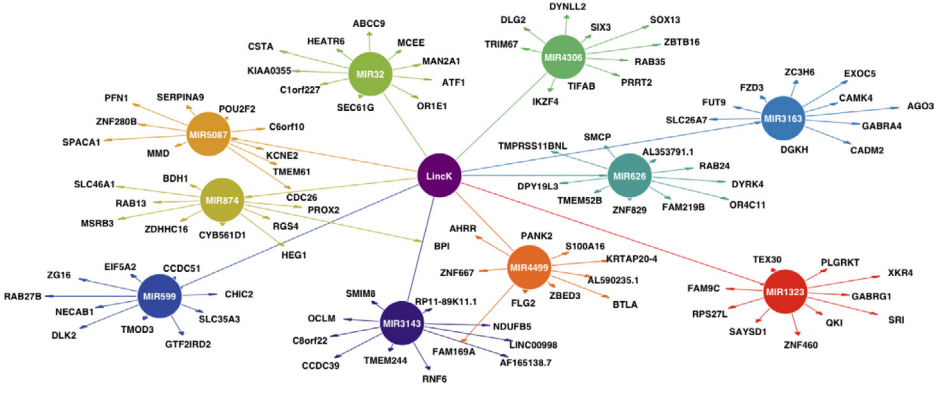

Fig. 4. Predicted LincK/miRNA/mRNA axes in MCF7 breast cancer cells. The ten miRNAs with the highest absolute $\log_2$ fold expression change and p-value < 0.05 upon LincK knockdown by shRNA #1 in MCF7 cells (obtained from the NCBI GEO GSE109007 dataset) with mRNA targets (predicted via TargetScan) are shown. Up to ten of the strongest mRNA targets for each of the miRNAs (predicted via TargetScan) are plotted. Together, these interactions reveal one hundred potential miRNA/mRNA interactions stemming from LincK in MCF-7 cells. The miRNAs, their fold expression change induced by LincK knockdown, and their predicted mRNA targets shown in this plot are detailed in Dataset 4A and Dataset 4B. This graph was generated using the igraph package in Rv4.0.4 [8].

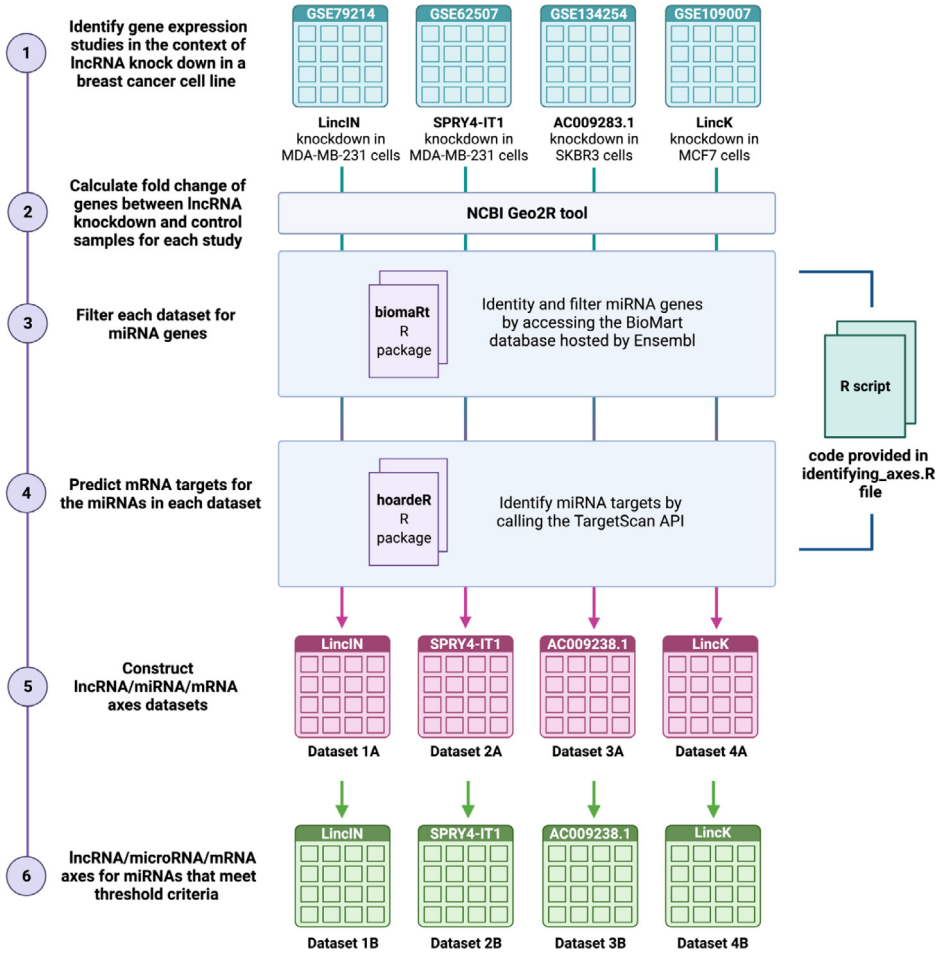

Fig. 5. Overview of methodology used to acquire, analyze and filter the four selected NCBI GEO datasets to identity putative lncRNA/miRNA/mRNA axes in breast cancer. The numbers 1, 2, 3, 4, 5, and 6 describe the sequential steps in this workflow.

breast cancer cell line on the GEO database. We identified four studies that met these criteria: GSE79214 [4], GSE62507 [5], GSE134254 [6], and GSE109007 [7].

The GSE79214 study used two shRNAs to knockdown LincIN in triple negative breast cancer MDA-MB-231 cells[4]. Our analysis compares the changes in gene expression between the control and shRNA #2-treated samples. The authors of this study performed transcriptome analysis using the GeneChip® Human Gene 2.0 ST Array.

The authors of the GSE62507 study knocked down lncRNA SPRY4-IT1 using short interfering RNAs (siRNAs) in MDA-MB-231 cells[5]. Changes in gene expression levels between the siRNA and control samples were performed using the Agilent-026652 Whole Human Genome Microarray 4 × 44K v2 (Probe Name version).

LncRNA AC009283.1 was knocked down using short hairpin RNAs (shRNAs) in the HER2-negative breast cancer cell line SKBR3 in the GSE134254 study [6]. Changes in gene expression between the treatment and control samples were obtained through the Affymetrix Human Transcriptome Array 2.0 [transcript (gene) version].

The GSE109007 study used two shRNAs to knockdown LincK in ER+ MCF-7 breast cancer cells [7]. Our analysis tracks the changes in gene expression between the control and shRNA #1-treated samples obtained using the GeneChip® Human Gene 2.0 ST Array platform.

The $\log_2$ fold change ($\log_2$FC) expression of the genes between the lncRNA knockdown and control samples were computed using the NCBI GEO2R tool (https://www.ncbi.nlm.nih.gov/geo/geo2r/). The summary data was downloaded and imported into RStudio (running R v4.0.2) [9,10]. Array probe IDs for GSE79214 (AFFY HuGene 2 0 st v1 probe set), GSE10907 (AFFY HuGene 2 0 st v1 probe set), GSE62507 (AGILENT WholeGenome 4 × 44k v2 probe set) and GSE134254 (AFFY HTA 2 0 probe set) were converted to HNGC gene symbols using the "biomaRt" R package (from the Bioconductor project) [11–13] (the code is provided in file identifying_axes.R).

The transcript type (i.e. protein-coding, miRNA, long non-coding RNA, etc.) for each probe was determined using the "biomaRt" R package (using the "gene_biotype" attribute). Probes with a gene_biotype value of "miRNA" were extracted and any duplicate entries in the datasets were removed. The fold change of the miRNA genes upon lncRNA knockdown were visualised using the "EnhancedVolcano" R package (from the Bioconductor project) [13,14] (the code is provided in file identifying_axes.R).

The predicted mRNA targets for each miRNA were acquired using the "hoardeR" R package, which calls the TargetScan [3] API via the targetScan() function (the code is provided in file identifying_axes.R). This data was merged with the miRNA fold change data to create the Datasets 1A, 2A, 3A and 4A. Each dataset was filtered to only include miRNAs with a $\log_2$ fold expression change $\geq$0.5 or $\leq$–0.5 and p-value $<$ 0.05. These miRNAs represent the strongest lncRNA-interacting candidates. This data is located in Datasets 1B, 2B, 3B, 4B.

Ethics Statement

NA

CRediT Author Statement

MCW: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing; **JMB:** Conceptualization, Writing – Original Draft, Writing – Review & Editing; **JV:** Writing – Review & Editing; **WF:** Writing – Review & Editing; **PM:** Conceptualization, Methodology, Validation, Writing – Review & Editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

MCDW and **JMB** are funded by the Dalhousie Medical Research Foundation (DMRF) Genomics in Medicine scholarships, Beatrice Hunter Cancer Research Institute (BHCRI) Cancer Research Training Program (CRTP) awards and Research Nova Scotia (RNS) Scotia Scholars awards. **MCDW** is also supported by Dalhousie University's Faculty of Medicine graduate scholarship. **JMB** is also supported by the Terry Fox Research Institute. **JV** and **WF** are funded by an operating grant to **PM** from the Canadian Institutes of Health Research (CIHR, PJT 162313). **JV** is also supported by a DMRF PhD Fellowship for Breast Cancer Research. Fig. 5 was created with BioRender.com. This study was funded by CIHR funding to **PM** (PJT 162313).

References

[1] L. Salmena, L. Poliseno, Y. Tay, L. Kats, P.P. Pandolfi, Leading Edge Essay A ceRNA Hypothesis: the Rosetta Stone of a Hidden RNA Language? Cell 146 (2011) 353–358, doi:10.1016/j.cell.2011.07.014.

[2] J. Venkatesh, M.-C.D. Wasson, J.M. Brown, W. Fernando, P. Marcato, LncRNA-miRNA axes in breast cancer: novel points of interaction for strategic attack, Cancer Lett. (2021).

[3] V. Agarwal, G.W. Bell, J.W. Nam, D.P. Bartel, Predicting effective microRNA target sites in mammalian mRNAs, ELife 4 (2015), doi:10.7554/eLife.05005.

[4] Z. Jiang, C.M. Slater, Y. Zhou, K. Devarajan, K.J. Ruth, Y. Li, K.Q. Cai, M. Daly, X. Chen, LincIN, a novel NF90-binding long non-coding RNA, is overexpressed in advanced breast tumors and involved in metastasis, Breast Cancer Res. 19 (2017) 62, doi:10.1186/s13058-017-0853-2.

[5] Y. Shi, J. Li, Y. Liu, J. Ding, Y. Fan, Y. Tian, L. Wang, Y. Lian, K. Wang, Y. Shu, The long noncoding RNA SPRY4-IT1 increases the proliferation of human breast cancer cells by upregulating ZNF703 expression, Mol. Cancer 14 (2015) 51, doi:10.1186/s12943-015-0318-0.

[6] A. Cedro-Tanda, M. Ríos-Romero, S. Romero-Córdoba, M. Cisneros-Villanueva, R.G. Rebollar-Vega, L.A. Alfaro-Ruiz, S. Jiménez-Morales, C. Domínguez-Reyes, F. Villegas-Carlos, A. Tenorio-Torres, V. Bautista-Piña, F.O. Beltrán-Anaya, A. Hidalgo-Miranda, A lncRNA landscape in breast cancer reveals a potential role for AC009283.1 in proliferation and apoptosis in HER2-enriched subtype, Sci. Rep. 10 (2020) 1–19, doi:10.1038/s41598-020-69905-z.

[7] J. Li, Y. Hao, W. Mao, X. Xue, P. Xu, L. Liu, J. Yuan, D. Zhang, N. Li, H. Chen, L. Zhao, Z. Sun, J. Luo, R. Chen, R.C. Zhao, LincK contributes to breast tumorigenesis by promoting proliferation and epithelial-to-mesenchymal transition, J. Hematol. Oncol. 12 (2019), doi:10.1186/s13045-019-0707-8.

[8] G. Csardi, T. Nepusz, The igraph software package for complex network research, InterJournal. Complex Syst. (2006) 1695–1695.

[9] R Core Team, R: The R Project for Statistical Computing, (n.d.). https://www.r-project.org/ (accessed March 29, 2021).

[10] RStudio TeamRStudio: Integrated Development for R, 2020 Http://Www.Rstudio.Com/.

[11] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, W. Huber, BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis, Bioinformatics 21 (2005) 3439–3440, doi:10.1093/bioinformatics/bti525.

[12] S. Durinck, P.T. Spellman, E. Birney, W. Huber, Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt, Nat. Protoc. 4 (2009) 1184–1191, doi:10.1038/nprot.2009.97.

[13] W. Huber, V.J. Carey, R. Gentleman, S. Anders, M. Carlson, B.S. Carvalho, H.C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K.D. Hansen, R.A. Irizarry, M. Lawrence, M.I. Love, J. MaCdonald, V. Obenchain, A.K. Oles, H. Pagès, A. Reyes, P. Shannon, G.K. Smyth, D. Tenenbaum, L. Waldron, M. Morgan, Orchestrating high-throughput genomic analysis with Bioconductor, Nat. Methods 12 (2015) 115–121, doi:10.1038/nmeth.3252.

[14] K. Blighe, S. Rana, M. Lewis, GitHub - kevinblighe/EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling, (n.d.). https://github.com/kevinblighe/EnhancedVolcano (accessed March 25, 2021).