The landscape of non-coding RNA in triple negative breast cancer.

Shahan Mamoor
San Diego, CA 92101
shahanmamoor@gmail.com

Breast cancer is one of the leading causes of death from cancers for women in the United States (1). Triple negative breast cancer is a more aggressive and harder to treat form of the cancer (2). We mined a public dataset containing transcriptome data from tumors isolated from 266 women with triple negative breast cancer (3). Differential gene expression analysis of these tumors based on stratification dependent on survival revealed that eleven non-coding RNA were among the RNA messages most differentially expressed when comparing the tumors of patients that survived versus those who had expired, including five separate anti-sense transcripts. Understanding the transcriptional behavior of tumors in women that survive the disease versus women who expire from it may facilitate discovery of targeted therapeutics. This analysis demonstrates that multiple species of non-coding RNA constitute an important part of the transcriptional landscape of triple negative breast cancers in humans, a disease causing over 42,000 deaths each year in the United States.

Keyword: differential gene expression analysis, triple negative breast cancer, non-coding RNA, antisense RNA, microRNA, pseudogenes, systems oncology

Introduction

Breast cancer is the second leading cause of cancer death in the United States for women, costing the lives of 42,000 women each year (1). Breast cancer can be classified based on the type of receptors expressed by the tumor, such as ER+ for tumors expressing the estrogen receptor, or HER2+ for tumors over-expressing the epidermal growth factor receptor (2). Likewise, breast cancers respond to treatment based on these biological classifications, and ER+ breast cancers benefit from treatments targeting the estrogen receptor such as tamoxifen, and women diagnosed with HER2+ breast cancers benefit from treatment with monoclonal antibodies targeting the epidermal growth factor receptor such as trastuzumab (4, 5). Women diagnosed with triple negative breast cancer (TNBC) bear tumors that express neither the estrogen receptor nor the progesterone receptor, and their cancers are not HER2+, meaning they do not over-express the epidermal growth factor receptor (2). As such, chemotherapies such as tamoxifen and trastuzumab will be less beneficial, if at all, to women diagnosed with TNBC. A diagnosis of TNBC brings along with it a shorter median survival from time of diagnosis, higher likelihood of death from the disease, and shorter time to relapse with a metastasis - in general, a more aggressive disease course with less available treatment options (2).

Targeted therapeutics in cancer have dual benefits in that they specifically address an underlying mechanism that drives or sustains tumorigenesis or the tumorigenic state, and because they specifically address that mechanism, in general, they carry fewer and less severe “side effects” (toxicity) (4-8). Systems-level analysis of the RNA output of cancers, including triple negative breast cancer, can provide critical insights into the transcriptional behavior of these tumors, and how they differ from healthy breast tissue to tumor tissue, or how they differ in tumors of women who survive the disease versus those who expire from the disease (9-15). We performed differential gene expression analysis on 209 tumors from a cohort of 266 African American women diagnosed with triple negative breast cancer, stratified in two groups, based

on whether they had survived or expired at time of analysis (3). This analysis revealed that eleven of the 250 RNA transcripts that were most differentially expressed, or whose quantities changed the most between the 86 women who expired versus the 123 women who survived, were RNA messages of the non-coding type (ncRNA). This group of differentially expressed non-coding RNA included five antisense transcripts.

Methods

GEO2R was used for this analysis, along with dataset GSE142102 (3) using platform GPL17692. The authors of the study from which this dataset was used (Purrington, KS; Wayne State University; NCBI Geo DataSets; unpublished) utilized Affymetrix Human Gene 2.1 ST Array microarray technology to obtain total RNA expression data from fresh frozen paraffin embedded (FFPE) tumors from 226 African American women diagnosed with triple negative breast cancer, or TNBC. For this global differential gene expression analysis, the p -value adjustment was set to Benjamini and Hochberg; log transformation of data was set to “Auto-detect”, and the submitter supplied category of platform annotation was used. Statistical significance of differential expression at the global level was determined by GEO2R. Statistical significance comparing the mean value of RNA expression between dead and alive groups was performed using PRISM (PRISM 8.1.2) (227) to conduct a two-tailed, unpaired t-test with Welch’s correction.

Results

We performed global differential gene expression profiling on tumor transcriptome data generated by microarray analysis of 266 primary breast tumors from women diagnosed with triple negative breast cancer from a public dataset (3). Of these women, 86 were designated as having expired from the disease at time of analysis, and 123 as being alive. Their global gene expression profiles were separated based on survival, and then compared using GEO2R to determine what the most significant transcriptional differences were between these two

groups. We found that eleven non-coding RNA transcripts were represented among the 250 most differentially expressed genes in the triple negative breast cancer transcriptome when stratified based on survival. These included five antisense transcripts, three pseudogenes, two microRNA and one intronic transcript

Five antisense transcripts are encoded by the most differentially expressed genes in a cohort of patients with triple negative breast cancer when comparing patients based on survival

We discovered five separate antisense RNA that were among the transcripts most differentially expressed in the tumor transcriptome of this cohort of patients with triple negative breast cancer when stratifying patients based on survival (Table 1). Antisense transcripts are transcribed in the opposite direction as the sense strand of a genomic locus and can exert transcriptional regulation by a variety of means including hybridization to the sense strand (16). Out of 53617 total transcripts detected by microarray in this dataset, ATP1B3 antisense RNA 1, or ATP1B3-AS1 was ranked 11th most differentially expressed between the tumors of patients that survived and the tumors of patients that had expired ($p = 0.00021917$), RRM1 antisense RNA 1, or RRM-AS1 was ranked 59th ($p = 0.00146944$), WWC3 antisense RNA 1, or WWC3-AS1 was ranked 156th ($p = 0.003813142$), SHANK2 antisense RNA 2, or SHANK2-AS2 was ranked 186th ($p = 0.0046377$), and SAMD12 antisense RNA, or SAMD12-AS1 was ranked 211th ($p = 0.00695859$).

We next retrieved the individual expression levels of each antisense transcript, for each patient (209 patients in total). ATP1B3-AS1 was expressed at significantly higher levels in the breast tumors of patients who had survived the disease at time of analysis (Figure 1; $p = 0.0001$).

RRM-AS1 was expressed at significantly lower levels in the breast tumors of TNBC patients who had survived the disease (Figure 1, $p = 0.0011$), as was WW3-AS1 (Figure 1, $p = 0.0077$), SHANK2-AS2 (Figure 1, $p = 0.0072$), and SAMD12-AS1 (Figure 1, $p = 0.0074$).

In summary, five antisense transcripts were among the RNA messages most differentially expressed between the tumors of patients with TNBC that had expired versus the

tumors of patients with TNBC that had survived, and all antisense transcripts except for ATP1B3-AS1 were expressed at significantly higher levels in patients with TNBC who had expired from the disease (Figure 1).

Three pseudogenes are encoded by the most differentially expressed genes in a cohort of patients with triple negative breast cancer when comparing patients based on survival

Non-coding RNA can exist in several forms, one of which are pseudogenes.

Pseudogenes, in general, possess a reference gene in the genome from which they are derived, but do to duplication, retroviral insertions, base pair mutation, and other genetic processes, have become non-functional with respect to translation of the gene product (17). This analysis revealed that three pseudogenes were represented among the most differentially expressed genes when comparing the global gene expression profiles of tumors from patients with TNBC based on survival.

Out of 53617 total transcripts detected by microarray in this dataset, the U7 small nuclear 71 pseudogene, or RNU7-71P was ranked 76th most differentially expressed (Table 1; $p = 0.00189562$), the ribosomal protein L23a pseudogene 64, or RPL23AP64 was ranked 77th (Table 1; $p = 0.00190801$), and the U7 small nuclear 69 pseudogene, or RNU7-69P was ranked as 125th most differentially expressed (Table 1; $p = 0.00308564$).

Both U7 small nuclear pseudogene transcripts were expressed at significantly lower levels in the breast tumors of TNBC patients who had survived (Figure 2; RNU7-71P: $p = 0.0035$; RNU7-69P: $p = 0.0057$). While the expression of RNU7-71P and RNU7-69P were significantly lower in the tumors of TNBC patients who had survived the disease, Rpl23-AP64 expressed was significantly higher in the tumors of TNBC patients who had survived (Figure 2; $p = 0.0015$)

Two microRNA are among the most differentially expressed genes in the tumor transcriptome of a cohort of patients with triple negative breast cancer when comparing patients based on survival

MicroRNA (miRNA) regulate gene expression after transcription and before translation by hybridizing to a transcript at its 3' untranslated region (UTR) using a seed region that is 6-8 nucleotides in length, thus interfering with its translation (18). Two miRNAs were represented among the two hundred and fifty most differentially expressed genes when comparing the transcriptomes of breast tumors from patients with TNBC who had expired versus those who had survived: miR4664 and miR367 (Table 1). Out of 53617 transcripts detected by microarray in the dataset utilized for this analysis, miR4664 ranked 146 ($p = 0.00350973$) and miR367 ranked 239 ($p = 0.00695859$).

While miR4664 expression was significantly higher in the tumors of TNBC patients that had expired from the disease (Figure 3; $p = 0.0054$), expression of miR367 was significantly lower in the tumors of TNBC patients that had expired from the disease (Figure 3; $p = 0.0054$).

A non-coding transcript from an intronic region was among the most differentially expressed genes when comparing the global gene expression profiles of breast tumors from patients with TNBC, stratified based on survival outcomes

Finally, one intronic ncRNA transcript, or non-coding RNA message produced from an intronic region of the genome (19, 20) was among the most differentially expressed genes when comparing the transcriptomes of breast tumors from patients with TNBC that had survived versus those who had expired: MAGI2 intronic transcript 1, or MAGI2-IT1 (Table 1). MAGI2-IT1 was the 193 most differentially expressed gene out of 53617 genes detected by microarray and included in the dataset utilized (Table 1; $p = 0.00479855$). The expression of MAGI2-IT1 was significantly lower in the tumors of patients with TNBC who had survived the disease (Figure 4; $p = 0.0081$).

Discussion

Triple negative breast cancer is a type of breast cancer with inferior survival outcomes and no available targeted treatments. Systems analysis of differential gene expression makes no assumptions regarding importance of any gene or pathway and instead asks what the most significant changes are in the transcriptional output of any two given groups of tissues, cells, or tumors. Here, we performed global differential gene expression profiling by using a microarray dataset from a cohort of 209 women diagnosed with TNBC, comparing the transcriptomes of breast tumors from 123 patients with TNBC that had survived versus the transcriptomes of breast tumors from 86 patients with TNBC that had expired. Our analyses revealed that eleven individual non-coding RNA transcripts were among the most significantly differentially expressed genes out of 53617 total transcripts detected by microarray.

Interestingly, there were two pseudogenes from the U7 family of small nuclear RNAs, important for processing of the 3' end of histone pre-mRNAs (21), that we found were some of the most differentially expressed genes when comparing the total RNA produced by breast tumors in patients that had survived versus those who had expired: RNU7-69P and RNU7-71P.

Women with triple negative breast cancer face a shorter time to relapse, a shorter median survival time, a shorter 5-year survival rate, and higher likelihood of metastases, the leading cause of death in women with breast cancer (2). An understanding of the transcriptional behavior of tumors of the triple negative sub-type, and how they differ from healthy to tumor tissue, from tumor to metastasis, and from the tumors of survivors to the tumors of those who will die from the disease, is critical for the development of novel targeted therapeutics. Non-coding RNA can regulate gene expression programs through multiple mechanisms (16-20) and these results merit further investigation into the role of each of these transcripts in the transcriptional biology of triple negative breast cancers.

References

1. DeSantis, C.E., Fedewa, S.A., Goding Sauer, A., Kramer, J.L., Smith, R.A. and Jemal, A., 2016. Breast cancer statistics, 2015: Convergence of incidence rates between black and white women. *CA: A Cancer Journal for Clinicians*, 66(1), pp.31-42.
2. Dent, R., Trudeau, M., Pritchard, K.I., Hanna, W.M., Kahn, H.K., Sawka, C.A., Lickley, L.A., Rawlinson, E., Sun, P. and Narod, S.A., 2007. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical cancer research*, 13(15), pp.4429-4434.
3. Purrington, KS; Wayne State University; GEO Data Sets GSE142102.
4. Piccart-Gebhart, M.J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., Gianni, L., Baselga, J., Bell, R., Jackisch, C. and Cameron, D., 2005. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *New England Journal of Medicine*, 353(16), pp.1659-1672.
5. Romond, E.H., Perez, E.A., Bryant, J., Suman, V.J., Geyer Jr, C.E., Davidson, N.E., Tan-Chiu, E., Martino, S., Paik, S., Kaufman, P.A. and Swain, S.M., 2005. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *New England Journal of Medicine*, 353(16), pp.1673-1684.
6. Nahta, R., Yu, D., Hung, M.C., Hortobagyi, G.N. and Esteva, F.J., 2006. Mechanisms of disease: understanding resistance to HER2-targeted therapy in human breast cancer. *Nature Reviews Clinical Oncology*, 3(5), p.269.
7. DeVita, V.T. and Chu, E., 2008. A history of cancer chemotherapy. *Cancer research*, 68(21), pp.8643-8653.
8. Chabner, B.A. and Roberts Jr, T.G., 2005. Chemotherapy and the war on cancer. *Nature Reviews Cancer*, 5(1), p.65.
9. Chin, L., Andersen, J.N. and Futreal, P.A., 2011. Cancer genomics: from discovery science to personalized medicine. *Nature medicine*, 17(3), p.297.
10. Cieřlik, M. and Chinnaiyan, A.M., 2018. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics*, 19(2), pp.93-109.
11. Liotta, L. and Petricoin, E., 2000. Molecular profiling of human cancer. *Nature Reviews Genetics*, 1(1), p.48.
12. Ramaswamy, S. and Golub, T.R., 2002. DNA microarrays in clinical oncology. *Journal of clinical oncology*, 20(7), pp.1932-1941.
13. Liang, P. and Pardee, A.B., 2003. Analysing differential gene expression in cancer. *Nature Reviews Cancer*, 3(11), p.869.
14. Fitzgerald, J.B., Schoeberl, B., Nielsen, U.B. and Sorger, P.K., 2006. Systems biology and combination therapy in the quest for clinical efficacy. *Nature chemical biology*, 2(9), p.458.

15. Liu, Y.R., Jiang, Y.Z., Xu, X.E., Hu, X., Yu, K.D. and Shao, Z.M., 2016. Comprehensive transcriptome profiling reveals multigene signatures in triple-negative breast cancer. *Clinical cancer research*, 22(7), pp.1653-1662.
16. Brantl, S., 2002. Antisense-RNA regulation and RNA interference. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1575(1-3), pp.15-25.
17. Mighell, A.J., Smith, N.R., Robinson, P.A. and Markham, A.F., 2000. Vertebrate pseudogenes. *FEBS letters*, 468(2-3), pp.109-114.
18. Lewis, B.P., Burge, C.B. and Bartel, D.P., 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *cell*, 120(1), pp.15-20.
19. Heo, J.B. and Sung, S., 2011. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science*, 331(6013), pp.76-79.
20. Louro, R., Smirnova, A.S. and Verjovski-Almeida, S., 2009. Long intronic noncoding RNA transcription: expression noise or expression choice?. *Genomics*, 93(4), pp.291-298.
21. Ruepp, M.D., Vivarelli, S., Pillai, R.S., Kleinschmidt, N., Azzouz, T.N., Barabino, S.M. and Schümperli, D., 2010. The 68 kDa subunit of mammalian cleavage factor I interacts with the U7 small nuclear ribonucleoprotein and participates in 3'-end processing of animal histone mRNAs. *Nucleic acids research*, 38(21), pp.7637-7650.

Figure 1: Five antisense RNA are among the transcripts that are most differentially expressed in the tumor transcriptome of a cohort of patients with triple negative breast cancer when comparing patients based on survival

Left: The mean RNA expression value of each antisense transcript is shown with error bars designating standard error of the mean (SEM), and a p-value determined by an unpaired t-test with Welch's correction to compare the significance of the difference between the groups.

Middle: Same as in left, but with each patients expression data displayed as a dot, and the mean expression value displayed as a black bar.

Right: Same as in left and middle, but with each patients expression data spread across the X axis to display the range of expression values at the higher and lower bounds of expression on the Y axis.

Figure 2: Three pseudogenes are represented among the most differentially expressed genes in the tumor transcriptome of a cohort of patients with triple negative breast cancer when comparing patients based on survival

Left: The mean RNA expression value of each pseudogene transcript is shown with error bars designating standard error of the mean (SEM), and a p-value determined by an unpaired t-test with Welch's correction to compare the significance of the difference between the groups.

Middle: Same as in left, but with each patients expression data displayed as a dot, and the mean expression value displayed as a black bar.

Right: Same as in left and middle, but with each patients expression data spread across the X axis to display the range of expression values at the higher and lower bounds of expression on the Y axis.

Figure 3: Two microRNA are among the most differentially expressed genes in the tumor transcriptome of a cohort of patients with triple negative breast cancer when comparing patients based on survival

Left: The mean RNA expression value of each microRNA transcript is shown with error bars designating standard error of the mean (SEM), and a p-value determined by an unpaired t-test with Welch's correction to compare the significance of the difference between the groups.

Middle: Same as in left, but with each patients expression data displayed as a dot, and the mean expression value displayed as a black bar.

Right: Same as in left and middle, but with each patients expression data spread across the X axis to display the range of expression values at the higher and lower bounds of expression on the Y axis.

Figure 4: One intronic transcript is represented among the most differentially expressed genes in the tumor transcriptome of a cohort of patients with triple negative breast cancer when comparing patients based on survival

Left: The mean RNA expression value of the intronic non-coding transcript is shown with error bars designating standard error of the mean (SEM), and a p-value determined by an unpaired t-test with Welch's correction to compare the significance of the difference between the groups.

Middle: Same as in left, but with each patients expression data displayed as a dot, and the mean expression value displayed as a dot.

Right: Same as in left and middle, but with each patients expression data spread across the X axis to display the range of expression values at the higher and lower bounds of expression on the Y axis.

Differentially Expressed ncRNA in the Triple Negative Breast Cancer Transcriptome Based on Survival

Rank	NCBI ID	ncRNA type	p-value	t	B	Transcript	Ensembl/Refseq ID
11	16959868	antisense	0.00021917	3.6958318	-2.48	ATP1B3-AS1 // ATP1B3 antisense RNA 1	ENST00000492725
59	16734755	antisense	0.00146944	-3.1806521	-3.09	RRM1-AS1 // RRM1 antisense RNA 1	ENST00000529323
76	16828226	pseudogene	0.00189562	-3.1061174	-3.17	RNU7-71P // RNA, U7 small nuclear 71 pseudogene	ENST00000516059
77	16745144	pseudogene	0.00190801	3.1041906	-3.17	RPL23AP64 // ribosomal protein L23a pseudogene 64	NR_003040
125	17113919	pseudogene	0.00308564	-2.9590754	-3.32	RNU7-69P // RNA, U7 small nuclear 69 pseudogene	ENST00000516369
146	17082258	microRNA	0.00350973	-2.9191636	-3.36	MIR4664 // 100616318	NR_039810
156	17109037	antisense	0.00381342	-2.8931975	-3.38	WWC3-AS1 // WWC3 antisense RNA 1	ENST00000430057
186	16728429	antisense	0.0046377	-2.8311781	-3.44	SHANK2-AS2 // SHANK2 antisense RNA 2	ENST00000429268
193	17059082	intronic	0.00479855	-2.8202558	-3.45	MAGI2-IT1 // MAGI2 intronic transcript 1	ENST00000440555
211	17072104	antisense	0.00518751	-2.7951539	-3.48	SAMD12-AS1 // SAMD12 antisense RNA	NR_038210
239	16979108	microRNA	0.00695859	2.6988202	-3.57	MIR367 // microRNA 367	NR_029860

Table 1: Five antisense transcripts, three pseudogenes, two microRNA and one intronic transcript are encoded by the most differentially expressed genes in the tumor transcriptomes of a cohort of patients with triple negative breast cancer when comparing patients based on survival

Data shown includes the rank of differential expression compared to the total transcriptome, raw *p*-value for statistical significance of differential expression of the gene globally, the B-statistic, or log-odds that the gene is differentially expressed, and a moderated *t* statistic generated by GEO2R, calculating the significance of differential expression by comparison of the two groups (dead and alive).