Identification of microRNAs targeting NAT1 and NAT2 gene transcripts in prostate cancer patients observed in different races

M Zainul Arifin N1, David Agustriawan*1, Arli Aditya Parikesit1, Rizky Nurdiansyah1, Kevin Nathanael Ramanto1

1Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta

*Corresponding email: david.agustriawan@i3l.ac.id

Abstract
Prostate cancer has the second highest death rate second only to lung cancer. Mutation in a single gene does not cause prostate cancer. Instead, many different genes are responsible, including NAT1 and NAT2 gene. By finding microRNAs that can suppress NAT1 and NAT2 gene, a novel prostate cancer treatment can be developed. Prostate cancer is also more commonly found in African-American than White-American. Therefore, this in silico study aimed to find several microRNAs targeting NAT1 and NAT2 gene observed in black, white, and other racial groups. 100 white patients, 100 races not reported patients, 11 black patients, and 638 combined races patient’s expression data were collected with TCGA-Assembler in R from The Cancer Genomic Atlas (TCGA). Next, Spearman correlation analysis was performed in R to find microRNAs that are negatively correlated with NAT1 and NAT2 gene. MicroRNAs were validated with miRTarBase and RNAhybrid. Only microRNAs which are located in combined patients and found in at least another race group is considered, i.e., hsa-mir-103a-1, hsa-mir-183, hsa-mir-32, and hsa-mir-96. Race-specific microRNA cannot be determined due to small sample size in black race group. Further study is needed to confirm the interaction between listed microRNAs with NAT1 and NAT2 gene expressions.

1. Introduction
Prostate cancer is a type of cancer characterized by the development of cancer in the prostate gland in the male reproductive system [1]. According to American cancer society, prostate cancer is diagnosed in 1 out 9 men during his lifetime. Just behind lung cancer, prostate cancer has the second highest death rate; about 1 out of 41 man died due to prostate cancer [2]. Prostate cancer is also more commonly found in African-American relative to White American [3].

Prostate cancer is not caused by mutation in a single gene. Instead, many different genes are responsible for prostate cancer [4]. Two genome-wide association studies in 2008, links several single nucleotide polymorphisms (SNPs) to prostate cancer [5]. One of their findings is that people with TT allele at SNP rs10993994 were reported to be 1.6 times likelier to develop prostate cancer than the one with CC allele pairs. The C allele is less prevalent in African-American compared to White-American, which explain the higher occurrence of prostate cancer in African-American descendants. This SNP is located in the promoter region of MSMB gene, thus it affects the amount of MSMB protein synthesized and secreted by epithelial cells of the prostate [6].
Biomarker is measurable substance that can be used to infer the organism biological state or condition [7]. In many diseases including prostate cancer, biomarker differentiates normal and prostate cancer patient. A good biomarker is disease specific and mutation in such biomarker tend to produce the disease [8]. Various type of biomarkers exists, such as the expression of gene and microRNA [9]. For many years, the screening of prostate-specific antigen (PSA) remains as the most common method of prostate cancer detection [10]. However, whether PSA is the main factor for prostate cancer development or not is still unclear [11].

Many genes are responsible for prostate cancer. Arylamine N-acetyltransferase (NAT) 1 and 2 are genes found in human genome which likely to play a pivotal role in prostate cancer. These genes, arylamine N-acetyltransferase enzyme that, catalyzes the N- or O- acetylation of various arylamine and heterocyclic amine substrate, a well known carcinogens [12]. In order for arylamine to be carcinogenic, it needs to be converted into its electrophilic intermediate [13]. Hein et.al (2002) has investigated bio-transformation of arylamine by NAT1 and NAT2. The high frequency of the NAT1 and NAT2 gene polymorphisms and the exposure to arylamine or heterocyclic amine in human population contributes to higher incidence of various cancers, but has never been tested in prostate cancer tissue.

MicroRNA are small endogenous noncoding RNA that negatively regulate expression of protein-coding genes post transcription [14]. Some evidences, such as the abnormal expression of microRNA, suggest their involvement in cancer development. A study by Porrka et al in 2007 successfully differentiates normal with prostate cancer patients, indicating microRNA as viable biomarker in prostate cancer [15]. As microRNA play a role in gene silencing, it is possible to down-regulate the expression of NAT1 and NAT2 gene through microRNA. If a negative correlation is obtained from the correlation analysis between NAT1 and NAT2 gene expression with microRNAs expression, indication of NAT1 and NAT2 expressions down-regulation through microRNA is present.

Based on these arguments. This in silico study is aimed to identify the relationship between NAT1 and NAT2 gene transcripts with its corresponding microRNA among prostate cancer patients based on TCGA dataset. Since prostate cancer is more commonly found in African-American, this study is also interested in identifying race specific microRNAs targeting NAT1 and NAT2 gene transcripts. This study hopes to provide list of microRNAs that has potential to target NAT1 and NAT2 gene transcripts for future studies.

2. Methods

2.1 microRNA and mRNA data acquisition

All dataset was obtained from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/). For this study, the data used are gene transcripts (mRNA) and microRNA data of prostate cancer patients, divided further into four portions based on the race, white, black, not reported, and all race.

These keywords were used in the repository section of the web-page to obtain the mRNA data: white, black, or not reported (race), prostate gland (primary site), TCGA-PRAD (project id), HTSeq – FPKM-UQ (Workflow type), transcriptome profiling (data category), and gene expression quantification (data type). These keywords were used for microRNA data: white, black, not reported, (race), prostate gland (primary site), TCGA – PRAD (project id), transcriptome profiling (data category), and microRNA expression quantification (data type). To get all prostate cancer patient’s regardless of the race, the race parameter is left unchecked. The metadata of these files were downloaded in json file format. Then, the 8 json files were converted into CSV file format using an online converter (https://konklone.io/json/).

A custom python script was used for pattern matching between microRNA and mRNA data in order to find patient’s ID which has both microRNA and mRNA data belong to the same race group. The result of the script is the list of patient’s IDs and their amount.

TCGA-Assembler version 2.0.6 were used to get the mRNA and microRNA data from patient’s ID [16,17]. For this study, 100 white patients, 100 not reported patients, 11 black patients, and 638
combined race patient’s expression ID were used. The first twelve characters of each patient’s ID was used as input into the R script provided by TCGA-Assembler. The result of this step is a text file of microRNA and mRNA data matrix which then converted into xlsx file for spearman correlation analysis.

2.2 Spearman correlation of microRNA and mRNA data
The microRNA and mRNA files were subjected into spearman correlation analysis in Matlab. It is performed to find out which microRNA has moderate to high negative correlation with NAT1 and NAT2 gene expression.

Only results with spearman correlation value (rho) and p-value less than -0.3 and 0.05, respectively, were retrieved for further validation. The cutoff for rho was set below -0.3 to get microRNA with moderate negative correlation value at minimum. The p-value cutoff was set below 0.05 to filter insignificant results.

2.3 Result validation
The chosen microRNAs were validated with miRTarBase (http://mirtarbase.mbc.nctu.edu.tw); a website that records all validated microRNA mRNA interaction [18]. Furthermore, microRNA and mRNA possible interaction was also simulated using RNAhybrid (https://bibiserv2.cebitec.uni-bielefeld.de/); a website that can predict the minimum free energy (mfe) required for microRNA and mRNA to interact [19]. To calculate the mfe, microRNA sequence and mRNA, taken from mirBase and NCBI, respectively, were inserted into the submission page. FireBrowse (http://firebrowse.org/) gene expression viewer is also used to further support the result [20].

3. Results and discussion
In TCGA-PRAD, there are 638 patient’s expression data for both gene expressions (mRNA) and microRNA expressions. the first 100 IDs were taken for white patients and not reported patients. Only 11 expressions data is available in TCGA for black patients, even though prostate cancer is more prevalent in African-American population [3].

![Figure 1](image.png)

**Figure 1.** NAT1 and NAT2 gene expression in RPM from 638 prostate cancer patients combined race group.
Figure 2. The expression of NAT1 and NAT2 in different cancers; (A) NAT1 and (B) NAT2. Sorted from the highest from left to right. NAT1 and NAT2 gene expressions are indicated by blue arrow. NAT1 ranks 18th and NAT2 ranks 23 across 37 type of cancer, respectively. The legends are shown at the top-right of the graph. Boxplots were taken from FireBrowse.

Figure 1 shows the gene expression in RPM for combined group with 648 data for both NAT1 and NAT2. Gene expression median of NAT1 and NAT2 is relatively similar (figure 1), but multiple outliers exists in NAT1 gene expression. This pattern is also observed from NAT1 FireBrowse boxplot (Figure 2A). For prostate cancer (PRAD), the median of tumor sample expressions are lower relative to normal for both NAT1 and NAT2. This may suggest both NAT1 and NAT2 down-regulation contributes to prostate cancer development. However, high variation can also be seen from figure 2. PRAD tumor gene expression first and third quartile are far from the median even surpassing normal sample first and third quartile which turns the previous statement inconclusive without further laboratory study.

We found 39 microRNAs that are negatively correlated with NAT1 gene expression and 10 microRNAs that are negatively correlated with NAT2 gene expression in white patient’s data. In black
patients, we found 103 microRNAs that are negatively correlated with NAT1 gene expression and 54 microRNAs that are negatively correlated with NAT2 gene expression. We found 14 microRNAs that are negatively correlated with NAT1 gene expression and 14 microRNAs that are negatively correlated with NAT2 gene expression in patient with unreported race. Additionally, we found 4 microRNAs that are negatively correlated with NAT1 gene expression and 21 microRNAs that are negatively correlated with NAT2 gene expression in combined patient data. Surprisingly, we found no microRNA that has negative correlation observed in all race group (white, black, unreported, and combined). Furthermore, the 28 microRNAs, which have at least moderate negative correlation, in unreported race group were not found in at least 2 other groups. There is also no microRNA with at least moderate negative correlation found in all race groups. Therefore, we only listed microRNAs that can be found in the combined patient’s group and found in at least two other race groups.

Table 1. List of microRNAs which are found in combined patient’s data and found in at least one or more race group that are at least moderately negatively correlation.

| microRNA       | Targeted gene transcript | Spearman correlation value (rho) | P-value |
|----------------|--------------------------|----------------------------------|---------|
|                |                          | Combined | White | Black  |        |
| hsa-mir-103a-1 | NAT2                     | -0.301   | -0.6545 | -0.789 | < 0.01 |
| hsa-mir-141    | NAT2                     | -0.3302  | -   | -0.623 | < 0.01 |
| hsa-mir-17     | NAT2                     | -0.354   | -   | -0.7789 | < 0.01 |
| hsa-mir-182    | NAT1                     | -0.3217  | -   | -0.763 | < 0.01 |
| hsa-mir-182    | NAT2                     | -0.364   | -   | -0.724 | < 0.01 |
| hsa-mir-183    | NAT1                     | -0.331   | -0.609 | -0.836 | < 0.01 |
| hsa-mir-183    | NAT2                     | -0.362   | -   | -0.697 | < 0.01 |
| hsa-mir-200c   | NAT2                     | -0.349   | -   | -0.779 | < 0.01 |
| hsa-mir-20a    | NAT2                     | -0.351   | -   | -0.816 | < 0.01 |
| hsa-mir-25     | NAT2                     | -0.389   | -   | -0.605 | < 0.01 |
| hsa-mir-3074   | NAT2                     | -0.3138  | -   | -0.779 | < 0.01 |
| hsa-mir-30d    | NAT2                     | -0.3139  | -   | -0.844 | < 0.01 |
| hsa-mir-32     | NAT1                     | -0.314   | -0.818 | -0.8 | < 0.01 |
| hsa-mir-375    | NAT2                     | -0.34004 | -0.682 | -0.614 | < 0.01 |
| hsa-mir-5694   | NAT1                     | -0.3209  | -0.7304 | - | < 0.01 |
| hsa-mir-708    | NAT2                     | -0.3098  | -   | -0.605 | < 0.01 |
| hsa-mir-92a-1  | NAT2                     | -0.3552  | -   | -0.816 | < 0.01 |
| hsa-mir-92a-2  | NAT2                     | -0.368   | -   | -0.7706 | < 0.01 |
| hsa-mir-93     | NAT2                     | -0.3551  | -   | -0.779 | < 0.01 |
| hsa-mir-96     | NAT1                     | -0.3217  | -0.672 | -0.836 | < 0.01 |

Because genetic variation exists between races, microRNAs might be race dependent. Listed microRNAs might be unsuccessful in degrading the mRNA for some race groups. Therefore, we listed three microRNAs with their potential target for known race groups (table 2).
Table 2. MicroRNAs that are found only in one race group (race specific) which have the highest correlation value for each group.

| microRNA     | Race | Targeted gene transcript | Spearman correlation value | P-value |
|--------------|------|--------------------------|----------------------------|---------|
| hsa-mir-5704 | White| NAT1                     | -0.8318                    | < 0.01  |
| hsa-mir-5683 | White| NAT1                     | -0.8091                    | < 0.01  |
| hsa-mir-6894 | White| NAT1                     | -0.8002                    | < 0.01  |
| hsa-mir-200c | Black| NAT1                     | -0.9272                    | < 0.05  |
| hsa-mir-4791 | Black| NAT1                     | -0.9174                    | < 0.05  |
| hsa-mir-141  | Black| NAT1                     | -0.9091                    | < 0.05  |

The correlation value gets closer to zero as the sample size increase. All rho value in either white or black has high correlation, whereas all the rho value found in combined patients are only moderately correlated. It is also important to note, that all microRNAs that have negative correlation found in combined patient cannot be found in patients with their race not reported (table 1).

MicroRNAs that can be found in combined patients and in at least one or more race group are chosen. Those microRNAs can be a potential marker or even as a molecule that can target the expressions of NAT1 and/or NAT2 gene in more than one race group. Even though the correlation results in all group are significant (p-value < 0.05), the rho value found in combined patients is more significant relative to the other group. Higher significance results are obtained due to the sample size even though their correlation values are moderate. Therefore microRNAs chosen for further validations are: **hsa-mir-103a-1, hsa-mir-183, hsa-mir-32, hsa-mir-375 and hsa-mir-96**.

For microRNA mRNA interaction validation, unfortunately, there is minimum experiment that confirm the interaction between microRNAs and NAT1 and NAT2 in miRTarBase. In fact, there is only 2 validated microRNA for NAT1 and there is no record for NAT2.

Therefore, we validate our result through possible interaction between selected microRNA and NAT1 and NAT2 transcripts or mRNA. The NAT1 mRNA sequence used was NAT1*10 allele. NAT*10 allele, with 2 single base mutation (T1088A and C1095A), is shown to be associated with several cancers, including prostate cancer [21]. For NAT2 mRNA, it is shown that fast acetylator NAT2 allele is associated with the increase of colorectal cancer risk, one of which is NAT2*14 which is used as a target sequence [22].

The lowest mfe for NAT1 mRNA is with the interaction with hsa-mir-183 microRNA with the mfe of -27.2 kcal/mol. The lowest mfe for NAT2 mRNA is with the interaction with hsa-mir-103a-1 with the mfe of -27.5 kcal/mol. Unfortunately, even though interaction is theoretically possible because the minimum energy requirement is less than 0, indicating spontaneous reaction, the significance value are all undefined. Further wet lab study can be performed to test NAT1 and NAT2 transcripts interaction with specified microRNAs.

MicroRNA can silence a gene by making a complementary base pairing with a gene mRNA product so they can never be translated to protein [14]. In prostate cancer, more specifically in NAT1 and NAT2 gene regulation, it is still unclear whether microRNA functions as tumor suppressor or tumor promoters. Further study is needed on NAT1 and NAT2 gene to know the genes involvement in prostate cancer specifically. Once they are known, the role of microRNAs can be determined as they function as NAT1 and NAT2 gene silencer.

4. Conclusion
How race affect the correlation between microRNA and mRNA cannot be determined from this study. The discrepancy of number of samples used promotes bias even though the result is statistically significant. Only microRNAs which are found in combined patients and found in at least
another race group is considered, those are: hsa-mir-103a-1, hsa-mir-183, hsa-mir-32, and hsa-mir-96.

Theoretically it is possible for the microRNAs listed to interact with the gene transcript based on the simulation result. However, the significance of this interaction is undefined. Racial implication also cannot be determined due to the insufficient sample available for African-American race group. Further study is needed to confirm stated microRNA-mRNA interactions and NAT1 and NAT2 involvement in prostate cancer to confirm the role of microRNAs.

Acknowledgment
This research was supported by Indonesia International Institute for Life Sciences (I3L). We thank David Agustriawan PhD, Dr.rer.nat Arli Parikesit, and Rizky Nurdiansyah M.Si for comments that greatly improved the manuscript. We would also like to thank Kevin Nathanael for assisting the research.

References
[1] National Cancer Institute. NCI Dictionary of Cancer Terms. National Cancer Institute. 2013.
[2] Society AC. Cancer Facts and Figures 2018. American Cancer Society. 2018.
[3] Brambilla E, Travis WD. World Cancer Report 2014. Geneva: WHO. 2014.
[4] Giles GG. Prostate Cancer. In: International Encyclopedia of Public Health. 2016.
[5] Eeles R a, Kote-jarai Z, Giles GG, Amin A, Olama AA Al, Guy M, et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet. 2008;
[6] Whitaker HC, Kote-Jarai Z, Ross-Adams H, Warren AY, Burge J, George A, et al. The rs10993994 risk allele for prostate cancer results in clinically relevant changes in microseminoprotein-beta expression in tissue and urine. PLoS One. 2010;
[7] Atkinson AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clinical Pharmacology and Therapeutics. 2001.
[8] Wang Q, Chaerkady R, Wu J, Hwang HJ, Papadopoulos N, Kopelovich L, et al. Mutant proteins as cancer-specific biomarkers. Proc Natl Acad Sci. 2011;
[9] Henry NL, Hayes DF. Cancer biomarkers. Molecular Oncology. 2012.
[10] Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, et al. Delineation of prognostic biomarkers in prostate cancer. Nature. 2001;
[11] Prensner JR, Rubin MA, Wei JT, Chinnaiyan AM. Beyond PSA: The next generation of prostate cancer biomarkers. Science Translational Medicine. 2012.
[12] Bateman A, Martin MJ, O’Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 2017;
[13] Hein DW. Molecular genetics and function of NAT1 and NAT2: Role in aromatic amine metabolism and carcinogenesis. Mutat Res - Fundam Mol Mech Mutagen. 2002;
[14] Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell. 2004.
[15] Porkka KP, Pfeiffer MJ, Waltering KK, Vessella RL, Tam melanchorpi T. MicroRNA expression profiling in prostate cancer. Cancer Res. 2007;
[16] Zhu Y, Qiu P, Ji Y. TCGA-Assembler: Pipeline for TCGA Data Downloading, Assembling, and Processing. HealthBsdUchicagoEdu [Internet]. 2014;1–8. Available from: http://health.bsd.uchicago.edu/yji/TCGA-Assembler-files/TCGA-Assembler.pdf%5Cnfile:///Users/rvanjaarsveld/Documents/Papers2/Articles/Unknown/Zhu/ Zhu.pdf%5Cnpapers2://publication/uuid/AADBE192-C6C4-40DC-A681-399931FEC8A5
[17] Wei L, Jin Z, Yang S, Xu Y, Zhu Y, Ji Y. TCGA-assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data. Bioinformatics. 2018;
[18] Chou CH, Chang NW, Shrestha S, Hsu S Da, Lin YL, Lee WH, et al. miRTarBase 2016: Updates to the experimentally validated microRNA-target interactions database. Nucleic Acids Res. 2016;44(D1):D239–47.
[19] Krüger J, Rehmsmeier M. RNAhybrid: MicroRNA target prediction easy, fast and flexible. Nucleic Acids Res. 2006;
[20] Deng M, Brägelmann J, Kryukov I, Saraiva-Agostinho N, Perner S. FirebrowseR: An R client to the Broad Institute’s Firehose Pipeline. Database. 2017;
[21] Bouchardy C, Mitrunen K, Wikman H, Husgafvel-Pursiainen K, Dayer P, Benhamou S, et al. N-acetyltransferase NAT1 and NAT2 genotypes and lung cancer risk. Pharmacogenetics. 1998;
[22] Da Silva TD, Felipe AV, de Lima JM, Oshima CTF, Forones NM. N-acetyltransferase 2 genetic polymorphisms and risk of colorectal cancer. World J Gastroenterol. 2011;