Identification of Epigenetic Regulation on The Expression of The Aberrant Gene of Kidney Renal Clear Cell Carcinoma Patients Observed in a Specific Race

David Agustriawan1,*, Hardi Mulyono2, Arli Aditya Parikesit1, Rizky Nurdiansyah1
1Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Indonesia
2Department of Management, STIE MURA, Indonesia

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

Abstract. Aberrant expression of genes in cancer is mainly caused by a mutation where there is a change in DNA sequences. However, the aberrant expression was also found without a change in the DNA sequences where epigenetic modification such as DNA methylation, histone modifications and microRNA become the main regulator in another layer of the cancer mechanism. The nature of epigenetic is heritable and reversible. It is important to search for the epigenetic mechanism in disease development in order to design epigenetic therapy and drugs. Methylation inhibitors and HDAC inhibitors drug already yielded seven FDA approved epigenetic drugs for myelodysplastic syndrome, cutaneous T-cell lymphoma, Multiple myelomas, peripheral T-cell lymphoma. This study search for the epigenetic mechanism in kidney renal cell carcinoma patients in a white race. The dataset of transcriptome profiling and epigenetic was downloaded from The Cancer Genome Atlas (TCGA) database. Some programming languages such as R, Python, Matlab and MySQL database were used to pre-processing the datasets and correlation computation part. This study found 14 aberrant genes which significantly correlated with 19 aberrant methylation probes with the correlation score less than equal to -0.7 and p-value < 0.01. Some of those down-regulated genes such as ZNF542, ZFP28, TMEM25, STK33 are correlated with hypermethylation in more than one sites. It is suggested that those methylation sites can affect the down-regulation of their expression in cancer formation. Further study is needed to validate the results through wet lab analysis.

1. Introduction
Epigenetic is the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in the sequence of DNA [1]. DNA methylation, microRNA (miRNA) and histone modifications are the main epigenetic modification that plays a role in the development of cancer. The sequencing (NGS) technology like next-generation sequencing can produce a huge amount transcriptome profiling dataset such as coding and non-coding gene, DNA methylation expression dataset that can be downloaded from GDG data portal TCGA websites. Moreover, NGS was successfully able to sequence all the human genome. The role of epigenetics has been studied both in coding and non-coding genes such as miRNA and long non-coding RNA [2,3]. This new layer
of cancer formation is important to be identified since it can suggest the new drugs and therapy for the precision medicine and also personal medicine since epigenetic modifications are affected by environmental factors [4]. It is suggested that between the races or even between the people is possibly need different treatment.

There are already some FDA approved epigenetic drugs and also some candidate epigenetic drugs and their potential use in controlling or treating cancer formation [5]. The clinical implication of epigenetic drugs such as methylation inhibitors (DNMTi) can reverse the roles of methylation role in inhibition the gene expression. As the example, azacytidine (vidaza), and decitabine (dacogen) are able to reverse the methylation pattern that inhibits the gene regulation in myelodysplastic syndrome, it removes the inhibitory effect. As a result, the silenced gene will be active. Therefore, it is important to discover the methylation sites that negatively interact with the gene expression in cancer formation.

High-throughput sequencing technology has been generated a massive dataset on human cancer for many molecular biology data types which can be accessed in TCGA websites through this link https://portal.gdc.cancer.gov/. The current updated version of TCGA websites has released the updated dataset. TCGA data portal is open access. As a result, the researcher can download and analyze datasets. In order to study epigenetic, TCGA provided gene expression datasets, miRNA expression datasets and DNA methylation expression datasets both in cancer and normal patients. However, for some cancer types, the number of normal patients is limited. Expression dataset which provided by TCGA enabling the researcher to identify the role of epigenetic both in coding and non-coding genes through in silico study. Currently, many tools can be used to download the TCGA datasets, such as TCGA assembler. This tool can be used through the R programming language. As a result, it will return a table matrix for further analysis [6].

Many research has been studied the role of epigenetic regulation through in silico or in a wet lab experiment. From in silico study, many researchers utilized TCGA as the resource for the datasets. Since the TCGA has currently updated the format of the datasets, therefore mostly the previous study used TCGA old version of datasets where there is no a menu to filter the datasets yet for example filter based on races [2, 7-12]. Another study which does not use the TCGA or not in the in silico study has also been studied the effect of epigenetics [13-15]. There is still a lack of studies to use the current version of the TCGA datasets format. e.g., lack of research to identify the role of epigenetics in the specific race in silico analysis. Since epigenetic regulation can be affected by an environmental factor, it can suggest that there will be also the possibilities the epigenetic role in cancer is different in between the races. Therefore, this study conducted research to identify the epigenetic role in a specific race. In this study, kidney renal clear cell carcinoma (KIRC) in a white race was chosen for an illustrative purpose.

In this study, we investigate the aberrant expression and the association between a gene and DNA methylation expression in KIRC of a white race. It was found that aberrant DNA methylation-regulated gene might cause down-regulated of Tumor suppressor gene (TSGs). This pathway can provide useful knowledge for the researchers. It can serve as a helpful suggestion for further research to develop a personalized treatment program for KIRC.

2. Materials and methods

2.1. Datasets and pre-processing

The dataset of gene and DNA methylation expression was downloaded in January 2019 at GDC data portal of TCGA websites. First, this study filtered the datasets in TCGA websites menu. For the gene datasets, this research used RNA-seq dataset for the kidney of the TCGA project on a white rate. We selected a white race because it returned the more samples compared with other races such as black or African American, Asian, and some not reported races. As a result, we collected 450 TCGA ID of each case. For the DNA methylation, we selected the dataset from Illumina Human methylation 450 platforms with the same race on gene dataset selection. It returned 266 TCGA ID for each case. Second, based on the TCGA barcode, this study separated cancer and normal patients using python
programming. However, the number of normal samples is quite small compared with the number of cancer samples. Third, in order to conduct a correlation study, this epigenetic study matched the TCGA ID between gene and DNA methylation dataset. This generated 264 matching ID between gene and DNA methylation datasets. This research randomly selected 50 matched ID for further investigation considering the computation time and space of the processing the datasets. Then, this study downloaded all the 50 matching ID for both DNA methylation and the gene for normal and cancer datasets using R programming software package. This study used TCGA assembler to collect all the expression datasets. It yields a table matrix which consists of a set of matching ID patients. For each patient, it has a measurement of 20,531 genes expression and 485,577 methylation probes. It suggests that analyzing the whole datasets is becoming a problem in terms of time and memory. This problem is a reason why this study only randomly selected 50 matched ID for further in silico analysis. Moreover, this study filtered the datasets by deleted the unknown gene ID, and also all the gene ID or methylation probes which has a null value for the expression in more than 10% of the total samples. Therefore, for each patient, it yields 309,827 methylation probes and 20,506 genes ID.

2.2. Compute differentially expressed genes and DNA methylations genes

This study imported gene and DNA methylation table matrix both for cancer and normal patients into MySQL software. This study calculated the average score for all the patients in normal for each gene and methylation probe. Then, in normal samples, after calculated the average expression, this study calculated the standard error for each gene and probe. From the average expression and the standard deviation, we can infer the average expression and the aberrant expression. Since the goal of this study is to identify the possible epigenetic effect on gene regulation. This study set a threshold for the aberrant expression of genes or probes in cancer patients if the average expression is greater than or equal to the normal average score plus two times of normal standard deviation. It suggests that the up-regulation of the gene or the hyper-methylation probes. In the other hand, if the cancer average expression is lower than or equal to the normal average score minus two times of normal standard deviation. It yields the down-regulated genes expression or hypo-methylated probes. It yields 3,871 up-regulated genes, 603 down-regulated genes, 1,335 hyper-methylated probes, and 7,295 hypo-methylated probes. Since this study will only focus on the silencing effect of DNA methylation toward gene expression, therefore, down-regulated gene and hyper-methylated probes dataset will be used for further analysis.

2.3. Correlation analysis of hypermethylated DNA and down-regulated gene expression

After we obtained the hyper-methylated probes and down-regulated genes in the same set of samples. This study utilized MATLAB software to compute correlation by using Spearman correlation analysis. We computed 1,335 hyper-methylated probes with 603 down-regulated genes. It yields the table matrix which provided the correlation score and also the p-value of the correlation of each gene with all the methylation probes. Then, we matched the probes-related genes by using the information provided by the TCGA ID. In order to select a very strong and significant negative correlation, this study only considers the correlation score greater than or equal to -0.7 and a p-value less than 0.01. It suggests a linear negative correlation between the probes and also the gene expression. After we completed all the analysis (Figure 1) we need to investigate whether the resulted genes are a TSG or play a role in cancer formation through a literature review or a wet lab identification is required.
3. Results and discussion

This study found that there is a strong negative significant correlation between 14 down-regulated genes and 19 hyper-methylation probes that located in their genome location particularly in CpG island location. It has been known that methylated of CpG island in a cancer formation will inhibit the gene transcription or reduce the gene expression activity. As shown in table 1, this study found that a gene is negatively correlated with more than one methylation sites such as in STK33, TMEM25, ZFP28 and ZNF542. It is suggested that the regulation of multiple hyper-methylation can strongly repress the transcription of the genes. The effect of multiple methylations toward the regulation of gene expression can give the inhibition effect higher than a single hyper-methylation regulation in a gene since the negative repressor is more than one. The significantly inverse negative correlation where in this study we set the threshold with the correlation score is lower than or equal to -0.7 with the p-value lower than or equal to < 0.001. This threshold implicates the linear relationship between methylation sites and gene expression. It suggests that when the methylation-expression is higher than the gene regulation is lower and vice versa. It describes in figure 2.

The resulted genes in table 1 can potentially function as TSGs. Consequently, when the expression is inhibited by hyper-methylated, it can cause cancer development. Through a literature review, we found that ZNF582 is involved in DNA damage response, proliferation, cell cycle control, and neoplastic transformation. It suggests that ZNF582 role is a TSGs [16]. ZNF582 hypermethylation also has been reported in acute myeloid leukaemia and various invasive cancers [17]. Therefore, it suggests that our result can support the previous research which stated that ZNF582 function as TSGs in cancers. Furthermore, in order to find the function of the resulted gene and validate the result, further study in wet lab analysis is needed.
Table 1. The potential hyper-methylation regulated down-regulation of genes in KIRC with a statistically significant inverse correlation

| Gene ID   | Gene regulation | Methylation probe | Methylation regulation | SRCC | p-value |
|-----------|-----------------|-------------------|------------------------|------|---------|
| C17orf51  | down            | cg23291136        | up                     | -0.7 | < 0.001 |
| C1orf210  | down            | cg01413566        | up                     | -0.7 | < 0.001 |
| CGN       | down            | cg21157873        | up                     | -0.7 | < 0.001 |
| CLMN      | down            | cg04847275        | up                     | -0.7 | < 0.001 |
| FBXO34    | down            | cg12786198        | up                     | -0.7 | < 0.001 |
| GSTO2     | down            | cg23659134        | up                     | -0.8 | < 0.001 |
| IRF6      | down            | cg25192855        | up                     | -0.7 | < 0.001 |
| RIC3      | down            | cg17603132        | up                     | -0.7 | < 0.001 |
| SMTNL2    | down            | cg11486359        | up                     | -0.7 | < 0.001 |
| STK33     | down            | cg00450824        | up                     | -0.7 | < 0.001 |
| STK33     | down            | cg20771178        | up                     | -0.7 | < 0.001 |
| TMEM25    | down            | cg15694715        | up                     | -0.8 | < 0.001 |
| TMEM25    | down            | cg19715094        | up                     | -0.7 | < 0.001 |
| TMEM25    | down            | cg20001829        | up                     | -0.7 | < 0.001 |
| ZFP28     | down            | cg22054362        | up                     | -0.7 | < 0.001 |
| ZFP28     | down            | cg24152605        | up                     | -0.7 | < 0.001 |
| ZNF542    | down            | cg15708153        | up                     | -0.7 | < 0.001 |
| ZNF542    | down            | cg27062795        | up                     | -0.7 | < 0.001 |
| ZNF582    | down            | cg09568464        | up                     | -0.7 | < 0.001 |

Abbreviations: SRCC denotes the Spearman rank correlation coefficient, down denotes the decreasing expression of a gene in a cancer sample compared to the normal samples, up denoted the increasing methylation expression in cancer samples compared to the normal samples.

Figure 2. The plot of TMEM25 gene expression vs cg15694715 and cg19715094 methylation probes
4. Conclusion
This study is a preliminary study to provide the in silico prediction of the epigenetically regulated expression genes in a specific race. In order to obtain a better result, population sample should be considered and also another race should also be included in this study, therefore, we can study the environmental effect on the epigenetic modification. Furthermore, the computational time and space in using illumine Human Methylation 450 platform should be well estimated. High-performance computer cluster might help to process the big dataset from the Illumina human methylation 450. This study is important because the prediction results can be used for further analysis, e.g. validation in wet lab analysis. The silico study can accelerate the research time in finding future treatment for cancer patients. The epigenetic analysis in cancer formation can offer new hope for all of the researcher to fight against cancer development. The epigenetic regulation in cancer is interesting to be identified because the nature of the epigenetic is traversable and inheritance. Therefore, it is very possible to change its oncogenic roles and can be a very powerful treatment.

Acknowledgement
The authors would like to thank to Indonesia International Institute for Life Sciences (i3L), LPPM of i3L that funding this project.

References
[1] Armstrong, L. (2013). Epigenetics. Garland Science.
[2] Agustriawan, D., Huang, C. H., Sheu, J. J. C., Lee, S. C., Tsai, J. J., Kurubanjerdjit, N., & Ng, K. L. (2016). DNA methylation-regulated microRNA pathways in ovarian serous cystadenocarcinoma: A meta-analysis. Computational biology and chemistry, 65, 154-164.
[3] Wu, Y. S., Lee, Z. Y., Chuah, L. H., Mai, C. W., & Ngai, S. C. (2019). Epigenetics in metastatic breast cancer: its regulation and implications in diagnosis, prognosis and therapeutics. Current cancer drug targets, 19(2), 82-100
[4] Pace, R., & Blundell, R. (2018). Epigenetics and Depression: A Rabbit Hole of Discovery. Open Journal of Genetics, 8(03), 67.
[5] Verma, M., & Kumar, V. (2018). Epigenetic Drugs for Cancer and Precision Medicine. In Epigenetics of Aging and Longevity(pp. 439-451). Academic Press.
[6] Zhu, Y., Qiu, P., & Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. Nature methods, 11(6), 599.
[7] Wijaya, E. B., Lim, E., Agustriawan, D., Huang, C. H., Tsai, J. J., & Ng, K. L. (2018, June). Utilize Imputation Method and Meta-analysis to Identify DNA-Methylation-Mediated microRNAs in Ovarian Cancer. In International Conference on Algorithms for Computational Biology (pp. 140-153). Springer, Cham.
[8] Dessie, E. Y., Wijaya, E. B., Huang, C. H., Agustriawan, D., Tsai, J. J., & Ng, K. L. (2018, October). [Regular Paper] Identification of Several Core Overexpressed MicroRNAs that Could Predict Survival in Patients with Ovarian Cancer. In 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 263-268). IEEE.
[9] Hu, W., Ivan, C., Bottsford-Miller, J., Zand, B., Dalton, H., Nick, A., Lopez-Berestein, G., Coleman, R., Baggerly, K. and Sood, A., 2013. Comprehensive epigenetic analysis of Notch pathway in high-grade serous ovarian cancer. Gynecologic Oncology, 130(1), p.e115.
[10] Antony, P., Rose, M., Heidenreich, A., Knüchel, R., Gaisa, N. T., & Dahl, E. (2014). Epigenetic inactivation of ST6GAL1 in human bladder cancer. BMC cancer, 14(1), 901.
[11] Kloten, V., Rose, M., Kaspar, S., von Stillfried, S., Knüchel, R., & Dahl, E. (2014). Epigenetic inactivation of the novel candidate tumor suppressor gene ITIH5 in colon cancer predicts unfavorable overall survival in the CpG island methylator phenotype. Epigenetics, 9(9), 1290-1301.
[12] Huang, W. Y., Hsu, S. D., Huang, H. Y., Sun, Y. M., Chou, C. H., Weng, S. L., & Huang, H. D. (2014). MethHC: a database of DNA methylation and gene expression in human
cancer. Nucleic acids research, 43(D1), D856-D861.

[13] Egger, G., Liang, G., Aparicio, A., & Jones, P. A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. Nature, 429(6990), 457.

[14] Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nature reviews genetics, 8(4), 286.

[15] Jones, P. A., & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. Nature reviews genetics, 3(6), 415.

[16] Huang, R.L., Chang, C.C., Su, P.H., Chen, Y.C., Liao, Y.P., Wang, H.C., Yo, Y.T., Chao, T.K., Huang, H.C., Lin, C.Y. and Chu, T.Y., 2012. Methylomic analysis identifies frequent DNA methylation of zinc finger protein 582 (ZNF582) in cervical neoplasms. PLoS One, 7(7), p.e41060.

[17] Liou, Y.L., Zhang, Y., Liu, Y., Cao, L., Qin, C.Z., Zhang, T.L., Chang, C.F., Wang, H.J., Lin, S.Y., Chu, T.Y. and Zhang, Y., 2015. Comparison of HPV genotyping and methylated ZNF582 as triage for women with equivocal liquid-based cytology results. Clinical epigenetics, 7(1), p.50.