Bioinformatics Pathway Analysis Pipeline for NGS Transcriptome Profile Data on Nasopharyngeal Carcinoma

Digdo Sudigyo1,2, Gisti Rahmawati1, Dicka W. Setiasari2, Risky H. Poluan1, Tjeng Wawan Cenggoro1,2, Arif Budiarto1,2, Alam Ahmad Hidayat1, Sri R. Indrasari1, Afiahayati1, Sofia M. Haryana3, and Bens Pardamean1,6

1Bioinformatics & Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia 11480
2Study Program of Biotechnology, Gadjah Mada University, Yogyakarta, Indonesia 55281
3Faculty of Medicine, Public Health, and Nursing, Gadjah Mada University, Yogyakarta, Indonesia 55281
4Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science, Gadjah Mada University, Yogyakarta, Indonesia 55281
5Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
6Computer Science Department, BINUS Graduate Program - Master of Computer Science Program, Bina Nusantara University, Jakarta, Indonesia 11480

Correspondent author: digdo.sudigyo@binus.edu

Abstract. Next-Generation Sequencing (NGS)-based genomics data have a huge potential to be used in transcriptomic profiling of Nasopharyngeal Carcinoma (NPC) to study the biosynthesis mechanism behind it. The high dimensionality of NGS data is the main challenge in performing the data analysis to extract useful information. In this workflow pipeline, memory-efficient Linux-based software such as HISAT2 and HTSeq are utilized to process the raw NGS data. Furthermore, Differential Expression Gene (DEG) list can be obtained by performing advanced analysis to the aligned Ribonucleic Acid (RNA) sequence using the edgeR protocol. This DEG list is one of the main inputs of biological pathway analysis that can be done in DAVID and PANTHER web-based software. Both tools generate a different pathway result related to inflammation.

Keywords—Transcriptome Profiling, NGS, Whole Genome, Cancer, Bioinformatics
1. Introduction

Based on GLOBOCAN 2018 [1], Nasopharyngeal Carcinoma (NPC) had a dangerous malignancy tumor disease with a massive prevalence of approximately 48,401 about the last 5 years. Inflammation plays the role of one main door inducing the tumor environment and one of the Hallmark of Cancer in the carcinogenesis process [2]. NPC shown dysregulation of epigenetic, which causes a change in several expression genes in carcinogenesis [3]. Morbidity of gene expression in cancer can be arranged to profile the transcriptome of cancer looking for specific biomarkers [4,5]. Several recent studies have to use bioinformatic analysis and update technology for sequencing tools to look at specific epigenetic regulation of cancer. Next-Generation Sequencing (NGS) can read several expressions of a gene, and dysregulation epigenetic of whole-genome sequences [6]. NGS technology can characterize functional genomes and can see changes in genomic and gene expression that can be seen through bioinformatics analysis [7]. NGS has an enhanced understanding of oncology with an emphasis on molecular genetics. The massive large amount of data also becomes an obstacle in processing data from the whole-genome sequence. It takes an advanced device with a large storage memory capacity to process and accommodate NGS data in silico. Large Random-Access Memory (RAM) size and storage memory capacity of more than 1 TB (1000 GB) are factors that are considered for processing NGS data using supercomputers [8,9]. High throughput of NGS data especially for the human genome requires efficient storage memory, retrieving, and processing large amounts of data. E-waste from computer and laptop components is a toxic and hazardous waste due to metal content. Heavy metals such as mercury, chromium, and lead are very persistent and harmful to the environment [10]. On the other hand, the bioinformatics world is growing rapidly in the world of research in all fields [11–14], especially health, which often analyzes the human genome [4,5]. The use of redundant analytical methods will increase the high consumption of computer components, leading to the accumulation of electronic waste.

Therefore, a suitable and easy bioinformatics software from NGS data is used to overcome these limitations with pipeline bioinformatics for transcriptomic analysis of human whole-genome by creating a transcriptome profile on nasopharyngeal cancer in Indonesia. Also through this bioinformatic pipeline can be utilized as an option to decrease the accumulation of e-waste (electronic waste) which reduces the purchase of components, especially RAM to process the human genome. In Indonesia, bioinformatics research on the cancer genome is infrequently carried out, however, SNP (Single-Nucleotide Polymorphisms) analysis has begun to be researched. GWAS (Genome-Wide Association Study) bioinformatic research on cancer colorectal Indonesian genome using data variant calling format (vcf) focuses on SNP analysis [15–17].

2. Literature Review

Several studies on transcriptomic profiling through the bioinformatic analytical approach were substantially determined using the software. NGS is one of the current technologies in clinical research to determine the mechanism of a disease, one of them through Ribonucleic Acid (RNA)-seq data besides microarray. In data mining research for profiling autoimmune genome transcriptome inner ear diseases on endolymphatic sac related to inflammatory responses [18], data were taken from the European Bioinformatics Institute database (ebi.ac.uk). The analyzed data were raw data obtained from the microarray sequencing analysis. Expression level changes were analyzed using the Robust Multichip Average algorithm in R software and pathway analysis using DAVID web-based software. A study by Zhe Ji et al [18] used breast cancer cells sequenced with a microarray of associated inflammation-related cancer regulation compared with all cancers in the Cancer Cell Line Encyclopaedia (CCLE) database. Affymetrix Expression Console Software was used to analyze gene expression levels and DAVID for gene ontology analysis. Microarrays in RNA-seq sequencing has a low sensitivity and limited dynamic range in sequencing samples compared to NGS [19].

Transcriptome profiling of ovarian carcinoma in Sallinen et al [20] study used a bioinformatic pipeline from a Hiseq 2000 type NGS machine and Differential Expression Genes (DEG) edgeR analysis software. Focusing on gene ontology in discussing the role of metastases in ovarian carcinoma, this study used edgeR software for differential expression gene, DAVID web-based, and IPA for ontology analysis. DAVID bUSED software not only for ontology analysis based on KEGG pathway but also for biopathway analysis of functional genes. Functional genes were analyzed into the biopathway can
determine the role of these genes and their relationship with other genes related to inflammation. Some software [2,18,20] can be recommended for the analysis of biosynthesis on cancer-related to inflammation mechanisms for transcriptome profiling. Bioinformatic software with minimum memory utilization, updating in database updates, and being free to use is a priority for analyzing human genome data from NGS, which has quite expensive sequencing costs.

3. The pipeline of Profiling Transcriptomic Nasopharyngeal Carcinoma

This study of profiling transcriptome NPC used NGS to present the regulation of genes that plays a role in inflammation in the process of carcinogenesis in Indonesia by showing a unique transcriptome profile. **Figure 1** shows the wet lab process of synthesis cDNA libraries from the isolation of total RNA samples. The qPCR analyzes the concentration of the cDNA library. One pM is a standard minimum value of cDNA library concentration used for the normalization process. If the cDNA library concentration meets the criteria in the sample using an index with the same value, then the sample with the highest concentration is used for normalization and pooling. After pooling, selected samples sequence to NGS machine (**Figure 1**).

![Figure 1. The workflow of RNA Isolation to Sequencing of Raw Data NGS (Wet Lab).](image)

The NGS machine type that was used in this study, Illumina Nextseq 550, can produce high throughput data of the highest quality for the whole data sequence. The scope of strategies for cancer sequencing ranges from targeted gene panels spanning several thousand bases through Whole Exome Sequencing (WES) analysis of 22,000 human proteins encoding genes (40-50 million bases) in Whole Genome Sequencing (WGS) across 3.3 billion bases of the whole human genome. Total RNA that represents the human whole genome can be run with NGS [8,9].
Illumina Nextseq 550 is able to produce high throughput data of the highest quality. The dry lab process shown in Figure 2 is the analysis of the raw data to profile the transcriptome with bioinformatics tools. FastQC, the quality control software for quality of the results of the NGS sequencing with the Linux operating system shows the quality value of raw data resulting from sequencing. If the value of the poor-quality sequence is obtained, the sequence needs to be trimmed or cut to reduce the value of poor quality in the sample. Trimming aims to eliminate unnecessary sequences such as adapters or excessive polyA, which causes a decrease in sequence quality in the analyzed sample [21].

Raw data of NGS sequencing results from cDNA amplification on the forward and reverse strand separately, in which the reading results of one sample contain two paired-end data. Paired-end data must be an assembly for mapping sequences to the human genome using Geneious software [22,23]. The alignment and mapping process used Linux-based software HISAT2 mapped sequences with genome sequence references with the resulting output in the form of SAM (Sequence Alignment Map) format. HISAT2 for the whole genome highly recommended to align and mapping the human genome. This platform quickly performs alignment and mapping without going through the stages of indexing the human genome reference. In the indexing stage, it takes a long time for the human genome and on other platforms, it requires large RAMs. HISAT2 needs low RAM requirement and make it easier for researchers to analyze with limited tools. HISAT2 protocols can be used to process large reads sequencing raw data and analyze gene transcripts and levels of expression quickly [24,25]. Although HISAT2 is still new and needs updating of information and databases, this software is free to access, which makes it easy to process and analyze big data on the human genome.

Alignment sequence calculation using HTseq-count software easily to match the calculated sequence with the reference code on input data like SAM format. Data from samples were calculated and tabulated in one data file in CSV (Comma Separated Values) format, which is used as input data for the differential analysis of gene expression. In the sequence calculation results using HTseq-count, genes were obtained, which have sequences with unique and specific alignments compared to the RefSeq database. HTSeq can count matrix with default parameters, and rapid counting for high-throughput sequencing (HTS) like the human genome. HTSeq count software recommends for pre-processes data like RNA-seq to analyze differential expression genes by counting overlap reads based on the reference sequence database [26].

For the differential analysis of gene expression (DEGs), we used the criteria from RStudio software with the edgeR package. The package annotates the sequence code from HTSeq according to the genes from the reference code in the Ensembl symbol database. The database was accessed from the UCSC genome browser. Then the annotated genes were selected into genes from the filtering process to eliminate each gene that had a transcript number of less than 10 and normalized to determine the normal factor and library size to determine the p-value and False Discovery Rate (FDR) values for each differential gene expression value detected. The use of edgeR for the analysis of the value of differences in gene expression based on replicated count data, further explains the biological variation with the addition of the limma package for RNA-seq from the Whole Genome Human data [27,28]. NPC samples are difficult to collect and the criteria for NGS are tremendously strict for sequencing. Therefore, the method of edgeR is used to calculate until minimum replication level of at least one replication in an experiment. edgeR has a recommended command protocol for analyzing differential expression gene (DEG) in oncology studies by comparing cancer and healthy (control) samples.
Figure 2. The workflow of Bioinformatics Analysis for Profiling Transcriptomic (Dry Lab).

Significant DEG data (differential expression in each gene) was obtained to be analyzed using DAVID and PANTHER, pathway analysis web-based tools to determine the role of genes related to nasopharyngeal cancer inflammation. PANTHER analysis is based on the ontology pathway detected genes mapped on pathways on the database. Meanwhile, DAVID analysis is related to the homo sapiens database search results by looking at the functional annotation of its pathway from the KEGG (Kyoto Encyclopaedia of Genes and Genomes). PANTHER and DAVID have different user interfaces and pathway analysis systems. DAVID is more up to date than PANTHER for functional genes based on pathway analysis. The KEGG pathway that is commonly used for pathway analysis references can be found on DAVID. It can integrate and process high-throughput genomic experimental data such as the human genome as well as to analyze functional genes as well as perform grouping to visualize multiple genes to many terms' relationships of a large gene list based on biopaths [29,30]. However, to find functional genes based on biological processes is easier in PHANTER than DAVID. PHANTER classifies functional genes into the biological process related to inflammation first. Afterward, it classifies them based on the biological process directed to the associated pathway. The pathway detected in PHANTER only leads to the protein family and subfamily for the input gene list. If more than one functional gene in a protein class is detected, it is difficult to know the specific role and relationship...
between individual genes [31,32]. Because of these considerations, DAVID and PHANTER are used to see the role of differential gene expression to see the role of detected functional genes and a more comprehensive relationship between genes in the bio-pathway related to the inflammatory process of nasopharyngeal carcinoma.

Table 1. Results of the Transcriptomic Profile of Nasopharyngeal Cancer In Inflammation

| No. | Gen      | Fold Change | P-Value | Pathway                                      | Software |
|-----|----------|-------------|---------|----------------------------------------------|----------|
| 1   | MYH2     | 7,240639    | 0,00122 |                                              |          |
| 2   | ALOX15   | -7,16467    | 0,000629|                                              |          |
| 3   | COL20A1  | 9,403125    | 0,000043| Inflammation Mediated by Chemokine and Cytokine Signaling | DAVID   |
| 4   | ACTC1    | 9,041164    | 0,000070|                                              |          |
| 5   | MYLK2    | 7,822418    | 0,000426|                                              |          |
| 6   | CXCL10   | 5,834063    | 0,002965|                                              |          |
| 7   | CXCL9    | 5,8664      | 0,003107| Toll-Like Receptor Signalling Pathway         |          |
| 8   | CXCL11   | 6,0372      | 0,002302|                                              |          |
| 9   | TRPA1    | 5,9087      | 0,002992|                                              |          |
| 10  | HTR2A    | 7,6709      | 0,000448| Inflammatory Mediator Regulation of TRP Channels Pathway | PHANTER |
| 11  | NGF      | 8,2358      | 0,000222|                                              |          |
| 12  | PRKCG    | 5,8831      | 0,003569|                                              |          |
| 13  | ADCY8    | 9,2334      | 0,000054|                                              |          |

From this bioinformatic pipeline, 13 significant genes related to the inflammatory process on nasopharyngeal cancer were obtained in 3 different pathways based on DAVID and PHANTER web-based software. The summarized transcriptome profile in Table 1 shows the effect of differential expression genes that affect the inflammatory pathway associated with nasopharyngeal carcinoma in Indonesia. Eventually, this information is possible to diagnose cancer early without going through invasive sampling methods especially serum and exosomal fluid by observing the transcriptome profile. In the future, gene therapy can be realized with the considerable research on transcriptome profile.

4. Conclusion
We presented a transcriptome profiling analysis pipeline related to the inflammatory process from NGS whole-genome data on nasopharyngeal carcinoma cases. The analysis of human cancer genome data from NGS sequencing in this study used a bioinformatic pipeline with simple commands, fast running data processes, and free access. Making transcriptome profiles from the human genome, which requires large RAM utilization, can be minimized by this bioinformatic pipeline. Therefore, the limitations of analyzing the human genome of cancer can be done using this pipeline, one of them for transcriptomic analysis.
5. Acknowledgment

We would like to acknowledge Fundamental Research Grants, Ministry of Higher Education Indonesia (KEMENRISTEK DIKTI) for funding this research, Dr. Med. dr. Indwiani Astuti from Faculty of Public Health and Nursing (FKKMK) for leading this research, Dr. Didik Setyo Heriyanto, Sp.PA, Ph.D., Tirta Wardana S.SI, M.Biotech, and Salsabila Luthfi Sesotyosari for as a member of project research in assisting permissance and execution of the wet lab process, Department of Computer Science and Electronics Gadjah Mada University for providing comprehension and lend tools of dry lab process, staff of Poli-THT Sardjito Hospital in Yogyakarta and Dr. dr. Cita Herawati, Sp.THT-KL from Dharmais Cancer Hospital in Jakarta, Indonesia for collecting and providing information that supports the research sample information.

References

[1] GLOBOCAN 2018 Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2018. https://gco.iarc.fr › data › 360-indonesia-fact-sheets. Accessed 27 October 2019.

[2] Ji Z, He L, Regev A, Struhl K 2018 Regulatory network controlling tumor-promoting inflammation in human cancers bioRxiv 352062.

[3] Heyn H, Sayols S, Moutinho C, Vidal E, Sanchez-Mut J V, Stefansson OA, et al 2014 Linkage of DNA Methylation Quantitative Trait Loci to Human Cancer Risk Cell. Rep. 7 331–8.

[4] Subramanian A, Tamayo P, Mootha V K, Mukherjee S, Ebert B L, Gillette M A, et al 2005 Gene set enrichment analysis: a knowledge-based approach to interpreting genome-wide expression profiles Proc. Natl. Acad. Sci. U S A 102 (43) 15545-50.

[5] Wang Z, Gerstein M, Snyder M 2009 RNA-Seq: a revolutionary tool for transcriptomics Nat. Rev. Genet. 10 (1) 57-63.

[6] Ross D T, Scherf U, Eisen M B, Perou C M, Rees C, Spellman P, et al 2000 Systematic variation in gene expression patterns in human cancer cell lines Nat. Genet. 24 227–35.

[7] Spencer D H, Sehn J K, Abel H J, Watson M A, Pfeifer J D, Duncavage E J 2013 Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J. Mol. Diagn. 15 623 –33.

[8] Ng S B, Buckingham K J, Lee C, Bigham A W, Tabor H K, Dent K M, Huff C D, Shannon P T, Jabs E W, Nickerson D A, Shendure J, Bamshad M J 2010 Exome sequencing identifies the cause of a Mendelian disorder Nat. Genet. 42 30 – 35.

[9] Bamshad M J, Ng S B, Bigham A W, Tabor H K, Emond M J, Nickerson D A, Shendure J 2011 Exome sequencing as a tool for Mendelian disease gene discovery Nat. Rev. Genet. 12 745–755.

[10] Liu X, Tanaka M, Matsui Y 2006 Generation amount prediction and material flow analysis of electronic waste: a case study in Beijing, China Waste Manag. Res. 24 434–45.

[11] Baurley J W, McMahan C S, Ervin C M, Pardamean B, Bergen AW 2018 Biosignature discovery for substance use disorders using statistical learning Trends Mol. Med. 24 221–35.

[12] Budiarto A, Mahesworo B, Baurley J, Suparyanto T, Pardamean B 2019 Fast and Effective Clustering Method for Ancestry Estimation Procedia Comput. Sci. 157 306–12.

[13] Lumbanraja F R, Mahesworo B, Cenggoro T W, Budiarto A, Pardamean B 2019 An evaluation of deep neural network performance on limited protein phosphorylation site prediction data Procedia Comput. Sci. 157 25–30.

[14] McMahan C, Baurley J, Bridges W, Joyner C, Kacamarga M F, Lund R, et al 2017 A Bayesian hierarchical model for identifying significant polygenic effects while controlling for confounding and repeated measures Stat. Appl. Genet. Mol. Biol. 16 407–19.

[15] Pardamean B, Baurley J W, Pardamean C I, Figueiredo J C 2016 Changing colorectal cancer trends in Asians Int. J. Colorectal Dis. 31 1537–8.

[16] Joyner C, McMahan C, Baurley J, Pardamean B 2020 A two-phase Bayesian methodology for the analysis of binary phenotypes in genome-wide association studies Biometrical J. 62 191–201.

[17] Yusuf I, Miskad U A, Lusikooy R E, Arsyad A, Irwan A, Mathew G, et al 2019 Genetic risk factors for colorectal cancer in multiethnic indonesians BioRxiv 626739.

[18] Zhang J, Wang N, Xu A 2018 Screening of genes associated with inflammatory responses in the
endolymphatic sac reveals underlying mechanisms for autoimmune inner ear diseases *Exp. Ther. Med.* **16** (3) 2460–2470.

[19] Asmann Y W, Wallace M B, Thompson E A 2018 Transcriptome profiling using next-generation sequencing *Gastroenterology* **135** (5) 1466 – 1468.

[20] Sallinen H, Janhonen S, Pölönen P, Niskanen H, Liu O H, Kivelä A, et al 2019 Comparative transcriptome analysis of matched primary and distant metastatic ovarian carcinoma *BMC Cancer* **19** 1121.

[21] Andrews S 2010 *FastQC: a quality control tool for high throughput sequence data* http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Accessed 27 October 2019.

[22] Meyer M, Kircher M 2010 Illumina sequencing library preparation for highly multiplexed target capture and sequencing *Cold Spring Harb. Protoc.* (6) pdb-prot5448.

[23] Šarhanová P, Pfanzelt S, Brandt R, Himmelbach A, Blattner FR 2018 SSR-seq: Genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring *Ecol. Evol.* **8**(22) 10817-10833.

[24] Kim D, Langmead B, Salzberg SL 2015 HISAT: a fast spliced aligner with low memory requirements *Nat. Methods* **12** (4) 357-360.

[25] Pertea M, Kim D, Pertea G M, Leek J T, Salzberg S L 2016 Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown *Nat. Protoc.* **11** 1650–1667.

[26] Anders S, Phyl P T, Huber W 2015 HTSeq: a Python Framework to Work with High-Throughput Sequencing Data *Bioinformatics* **31** (2) 166 – 169.

[27] Marioni J C, Mason C E, Mane S M, Stephens M, Gilad Y 2008 RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays *Genome Res.* **18** (9) 1509-1517.

[28] Robinson M D, McCarthy D J, Smyth G K 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data *Bioinformatics* **26** (1) 139-140.

[29] Huang D W, Sherman B T, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler M W, Lane H C, Lempicki R A 2007 DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists *Nucleic Acids Res.* **35** W169--W175.

[30] Huang D W, Sherman B, Lempicki R 2008 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources *Nat. Protoc.* **4** 44-57. https://doi.org/10.1038/nprot.2008.211.

[31] Thomas P D, Campbell M J, Kejariwal A, Mi H, Karlan B, Daverman R, Diemer K, Muruganujan A, Narechania A 2003 PANTHER: a library of protein families and subfamilies indexed by function *Genome Res.* **13** (9) 2129–2141.

[32] Mi H, Muruganujan A, Thomas P D 2013 PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees *Nucleic Acids Res.* **41** D377–D386.