

A framework for predicting lncRNAs expression in human dendritic cells in response to *M. tuberculosis* infection

Faizah Aplop¹, Saharuddin Mohamad²

¹Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia

²Institute of Biological Sciences, Universiti of Malaya, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Jun 20, 2022

Revised Aug 13, 2022

Accepted Aug 28, 2022

Keywords:

Convolutional neural networks

Dendritic cells

lncRNAs

RNA-seq expression analysis

Tuberculosis

ABSTRACT

Tuberculosis (TB) is an air-borne infectious diseases caused by *M. tuberculosis* bacteria that primarily affects human lungs. Existing vaccine does not work well due to the evolution and latent movement of this bacteria. Developing an effective vaccine to combat Tuberculosis is very difficult as the interaction between the bacteria and human immune system is not fully understood. With recent advancement of transcriptome profiling analysis, long noncoding ribonucleic acids (lncRNAs) are found to be widely expressed in immune system. However, the role of lncRNAs is still not been widely explored in understanding human immune response to TB infection. In this paper, we propose a general framework for predicting lncRNAs being expressed in human dendritic cells. By incorporating deep learning method with RNA-seq data analysis, we intend to identify and characterize the lncRNAs found in dendritic cells from two groups of TB resistant patients through their RNA-seq expression data.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Faizah Aplop

Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu

Unnamed Road, 21300 Kuala Terengganu, Terengganu, Malaysia

Email: faizah_aplop@umt.edu.my

1. INTRODUCTION

Tuberculosis (TB) is one of the most deadly contagious disease in the world that caused by antibiotic resistant bacteria known as *Mycobacterium tuberculosis* (*Mtb*). It primarily affects human lung and could travel through the bloodstream to infect other part of human body. Even worse, the bacteria could travel to the meninges, which are the membranes surrounding the brain and spinal cord. Since the function of meninges is to protect human central nervous system, the infected meninges causes a rise in pressure within the skull, resulting in nerve and brain tissue damage, which is often severe. This life-threatening condition is known as meningeal tuberculosis (TB meningitis). It has been reported that 10-20% of the TB meningitis patients will suffer long-term after-effect such as severe brain damage, epilepsy, paralysis, hearing loss, and blindness [1].

The human immune system ables to protect human body against diseases by identifying, attacking and destroying threats from viruses, bacteria and parasites when function properly. As for TB, the immune response of a patient is critical whereby it could either help the body to fight the progression of the disease or it could exacerbate the bacteria infection when there is involvement of certain key molecules. Even worse, tuberculosis is said to co-evolve with human and the ability of *Mtb* to manipulate human immune system to destruct lung tissue has made it an ultimate pathogen [2]. *M. bovis* *Bacillus* Calmette-Guérin (BCG), the vaccine that was introduced and being widely used since 1921 does not work well in protecting human against TB due to

evolvement and latent movement of Mtb. Developing an effective vaccine has been very challenging and difficult since the bacteria is able to evade immune system attack by co-opting the mechanisms of the immune system itself to its own advantage [3], [4]. The immune response against Mtb involves a network of innate and adaptive immune responses, where dendritic cells (DCs) are the key cells that bridge them. DCs is one of the main types of professional antigen-presenting cells (APCs) of the immune system [5]. Though the Mtb primarily resides in DCs and is able to interfere with their functions as it has the ability to impair host innate and adaptive immune responses, their interactions are less well understood [6]. Furthermore, existing research findings show that the interaction of DCs with Mtb is contradictory [7].

Long non-coding ribonucleic acids (lncRNAs) are RNA molecules with length exceeding more than 200 nucleotides, which do not code for proteins. Although lncRNAs are less understood, they do have crucial roles in diverse biological, pathological processes, and could cause prominent implications to various human diseases. They are expressed in a tissue-specific context and responsible for regulating transcriptional control. Recent studies show that they are functionally associated with various cancers [8]–[10] and immune-mediated diseases [11], [12]. They are the regulators of various immune function, where they have large effects on adaptive and innate immune system [6], [13], [14]. Microarray technology [15] and traditional wet-lab experiments [16] had been applied for many years to uncover lncRNAs differential expression in patients infected with tuberculosis. However, RNA-seq is proven to provide better estimates of transcript expressions [17]. As shown in Figure 1, we propose a general framework for predicting lncRNAs being expressed in human DCs. We intend to conduct RNA-seq expression analysis with convolutional neural networks (CNNs), a well-known deep learning (DL) technique to reveal the identification and characterization of lncRNAs found in DCs associated to TB infection for TB resistant patients who were identified to have non-infected and infected states as discussed in [18].

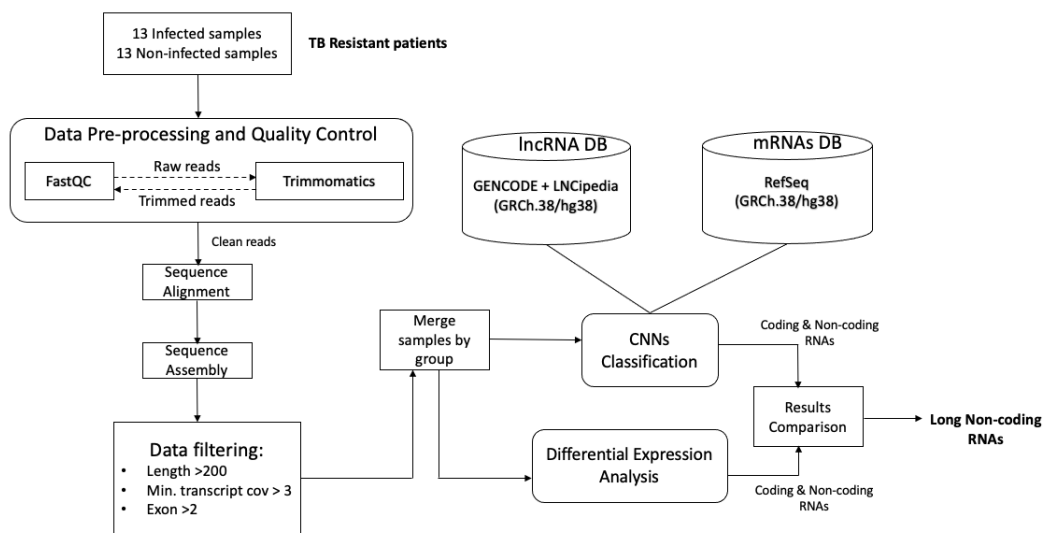


Figure 1. Schematic illustration framework of discovering lncRNAs being expressed in human DCs from RNA-seq data

2. METHOD

2.1. RNA-seq data

The RNA-seq datasets of resistant patients are obtained in a form of FASTQ file format from sequence read archive (SRA) database. They are single-end reads of next generation sequencing (NGS) library, using Illumina HiSeq 4,000 instrument, which sequencing was performed by Gilad Lab, University of Chicago [18]. The experiment of transcriptomic response of DCs towards Mtb infection will take into consideration the infected and non-infected group of patients. There are 26 samples altogether whereby 13 samples from infected patients, and another 13 samples are from non-infected patients. Information on raw RNA-Seq datasets used in this study are shown in Tables 1 and 2 respectively.

Table 1. Infected group

Sample size (bp)	SRA No
1.92G	SRR5206792
8.50G	SRR5206794
1.49G	SRR5206796
1.36G	SRR5206804
1.29G	SRR5206806
1.82G	SRR5206808
4.84G	SRR5206810
2.17G	SRR5206812
1.28G	SRR5206814
1.68G	SRR5206816
10.39G	SRR5206818
1.75G	SRR5206820
1.5G	SRR5206826

Table 2. Non-infected group

Sample size (bp)	SRA No
1.42G	SRR5206795
2.10G	SRR5206803
2.50G	SRR5206805
1.60G	SRR5206807
1.67G	SRR5206809
11.45G	SRR5206811
2.63G	SRR5206813
2.66G	SRR5206815
2.00G	SRR5206817
2.30G	SRR5206819
1.94G	SRR5206821
1.43G	SRR5206823
1.41G	SRR5206827

2.2. Data preprocessing and quality control

Initial quality control (QC) assessment will be performed on all data samples as an effort of checking whether these data has any problems before continuing further analysis. This process will be done using FASTQC (version 0.11.9) software, a simple graphical user interface (GUI) tool developed in Java by Babraham Bioinformatics group at the Babraham Institute. In a form of graph and tables (QC report), this tool provides a quick access and overview on problematic areas of raw sequence data coming from high throughput pipelines [19]. It has 12 analysis modules, which meet the quality control standard for raw reads.

FastQC program generates a QC report, which lead to the choice of preprocessing steps that need to be undertaken in order to fix the identified issues. A read trimming and filtering tool optimized for Illumina NGS data known as Trimmomatic [20] will be utilized to discard low-quality reads, trim adapter sequences and eliminate poor quality bases. In this work, we will use Trimmomatic to trim effected data samples that contains error in "per base sequence content" module as reported by QC report. After the effected reads were trimmed, FastQC will be run again to assess the quality of the trimmed reads. The QC report will be checked again to make sure that all the reads that previously have issue with per base sequence content are error free. Once the reads satisfied all the quality requirements, we will consider that these reads are good to go for the next step of RNA-seq data processes and analysis.

2.3. Reference-based RNA-seq read alignment

To infer which transcripts are expressed, identify genomic positions or estimate where the reads originated from, the sample RNA-Seq reads need to be aligned against reference genome. In this work, hierarchical indexing for spliced alignment of transcripts 2 (HISAT2) [21], a fast and sensitive graph-based alignment program, which is developed to map NGS reads to a single or population of human genomes will be used to align these reads against the selected human reference genome. HISAT2 ables to produce higher accuracy of sequencing reads alignment compared to original HISAT [22] system as it incorporates algorithmic improvements, where a hierarchical graph FM index (HGFM) is applied [23].

The strain of human reference genome used in this project is genome research consortium human build 38 (GRCh38). In order to run HISAT2, the reference genome need to be built in a form of indexes. Since the GRCh38 indexes are already available in HISAT2 website, we downloaded the HGFM index for reference plus transcripts directly from their website. These indexes use ensembl gene annotations, where it has many more transcripts compared to reference sequence (RefSeq) annotations [23]. The output of HISAT2 aligner is in sequence alignment map (SAM) format. Using Samtools, the SAM files will be sorted and converted into binary alignment map (BAM), which stores the same data but in a compressed binary representation for improved performance [24]. These BAM files will then be used as the input files in transcriptome assembly.

2.4. Transcriptome assembly

RNA-seq reads need to be reconstructed into a full length transcripts to allow for gene expression studies. For this purpose, we will be using StringTie, a transcript assembler and quantification tool for RNA-Seq, which was also developed by central for computational Biology (CCB) of John Hopkins University. This tool uses genome-guided transcriptome assembly together with de novo genome assembly approaches to improve transcript assembly. It applies network flow algorithm to estimate expression level for each transcripts

[25]. StringTie is claimed to be better than other leading assembler such as cufflinks as it could produce more complete and accurate reconstructions of genes and better expression level estimation [26], [27].

Each alignment files produced by HISAT2, which then will be converted into BAM format are assembled using StringTie2. StringTie2 is the latest release of StringTie program, which include a new method in better high rate error handling and has the ability to work with full length super-reads that brings to better quality of short-reads assemblies [27]. The assembly process requires for reference genome annotation in .gtf format. Therefore, we downloaded human GRCh38 (version 21) comprehensive gene annotation file from genome Encyclopedia Of DNA elements (GENCODE) [28] online database for this purpose.

2.5. Data filtering

Using in-house scripts, assembled transcripts will need to be checked and filtered to meet only criteria of a lncRNA. The selection criteria includes to select only transcripts with nucleotide sequence more than 200 base pairs, minimum coverage is at least 3x and the number of exon should be more than 2 as suggested in [29], [30]. All transcripts that do not meet these criteria will be discarded. Then, the lncRNA transcripts from the 26 samples will be merged into two different files according to patients group.

2.6. Differential expression analysis

One of the key steps in analyzing RNA-seq data is to perform the genes differential expression analysis between different biological conditions, where the summarized data will be assessed by statistical models [17]. To identify and quantify the changes in expression levels between the two groups of TB resistant patients, DESeq2 bioconductor software package will be used as the statistical tool to perform differential analysis of count data. As an improved package of DESeq [31], DESeq2 estimates and perform statistical inference on differential data based on negative binomial distribution. Using shrinkage estimators for the dispersion and fold changes in differential expression analysis allows DESeq2 to offer a sound, consistent performance and statistically well-founded solution to the wide dynamic range of RNA-seq experiments [32].

2.7. Deep learning method

In predicting non coding RNAs, it had been proven that the prediction performance of tools that apply deep learning methods exceeded other traditional machine learning methods in terms of identification reliability, ease of use and ability to utilize features not incorporated in the current knowledge [33], [34].

2.7.1. Convolutional neural networks

The considerable success of CNNs in various visions and imaging tasks has given significant impact to nearly all scientific fields including bioinformatics. Alzubaidi *et al.* [35], CNNs is the most utilized deep learning network type that could automatically identifies relevant features without any human intervention and considered to be more powerful than recurrent neural networks (RNNs), another well-known deep learning algorithm capable of processing sequential data. Therefore, as shown in Figure 1, CNNs technique is proposed to be implemented in identifying and classifying between lncRNAs and mRNAs being expressed in human DCs. We intend to explore the capability of CNNs to extract information from one-dimensional biological sequences data as discussed by [36], [37].

We will enhance the lncRNAs prediction by comparing, mapping and consolidating the results yielded from CNNs technique with those discovered by differential expression analysis method using DESeq2. These processes will be done by running our own in-house scripts. We expect that the annotated and putative lncRNAs could be identified and any possible predicted mRNAs will be discarded.

2.7.2. Training datasets

Datasets of human lncRNAs from two well-known public databases, which are GENCODE and LNCipedia, and human mRNAs datasets from the RefSeq, an NCBI RefSeq database will be used as the training datasets to discover and learn RNAs data patterns. As reported by [38], GENCODE had suggested that there are more than 16,000 lncRNAs in human genome. It has comprehensive gene annotation of lncRNA genes on the reference chromosomes. While LNCipedia currently contains 127,802 transcripts with 107,039 are considered as high-confidence set of lncRNAs [39]. As for RefSeq, it contains curated, non-redundant collection of sequences representing genomes, transcripts and proteins [40].

3. CONCLUSION

One of the important research applications of RNA-seq these days is to discover the lncRNAs expression profiles using computational tools and pipelines. There are studies to uncover lncRNAs differential expressions in patients with tuberculosis infections using microarray technology and traditional wet-lab experiments. However, RNA-seq is proven to come up with better estimates of transcript expressions. The sophisticated high-throughput RNA sequencing technology allows researchers to characterize and quantify differential expressions with higher sensitivity, higher speed and higher dynamic range. To further enhance and improve the prediction results, we propose a framework to discover lncRNA transcripts being expressed in human DCs of two TB resistant patient groups by incorporating CNNs classification technique with existing RNA-Seq expression analysis.





REFERENCES

- [1] K. J. Olbrich *et al.* "Systematic review of invasive meningococcal disease: Sequelae and quality of life impact on patients and their caregivers," *Infectious Diseases and Therapy*, vol. 7, pp. 421–438, 2018, doi:10.1007/s40121-018-0213-2.
- [2] P. T. Brace *et al.*, "Mycobacterium tuberculosis subverts negative regulatory pathways in human macrophages to drive immunopathology," *PLOS Pathogens*, vol. 13, no. 6, p. e1006367, 2017, doi: 10.1371/journal.ppat.1006367.
- [3] W. Zhai, F. Wu, Y. Zhang, Y. Fu, and Z. Liu, "A The immune escape mechanisms of *mycobacterium tuberculosis*," *International Journal of Molecular Sciences*, vol. 20, no. 2, p. 340, 2019, doi: 10.3390/ijms20020340.
- [4] J. D. Ernst, "Mechanisms of M. tuberculosis immune evasion as challenges to TB vaccine design," *Cell Host and Microbe*, vol. 24, no. 1, pp. 34–42, 2018, doi: 10.1016/j.chom.2018.06.004.
- [5] M. de Martino, L. Lodi, L. Galli, and E. Chiappini, "Immune response to mycobacterium tuberculosis: A narrative review," *Frontiers in Pediatrics*, vol. 7, no. 350, 2019, doi: 10.3389/fped.2019.00350.
- [6] R. Madan-Lala *et al.*, "Mycobacterium tuberculosis impairs dendritic cell functions through the serine hydrolase hip1," *The Journal of Immunology*, vol. 192, no. 9, pp. 4263–4272, 2014, doi: 10.4049/jimmunol.1303185.
- [7] A. Mihret, "The role of dendritic cells in mycobacterium tuberculosis infection," *Virulence*, vol. 3, 2012, doi: 10.4161/viru.22586.
- [8] Y. Qian, L. Shi, and Z. Luo, "Long non-coding RNAs in cancer: Implications for diagnosis, prognosis, and therapy," *Frontiers in Medicine*, vol. 7, p. 902, 2020, doi: 10.3389/fmed.2020.612393.
- [9] N. Gao *et al.*, "Long non-coding RNAs: The regulatory mechanisms, research strategies, and future directions in cancers," *Frontiers in Oncology*, vol. 10, p. 598817, 2020, doi: 10.3389/fonc.2020.598817.
- [10] Y. Fang and M. J. Fullwood, "Roles, functions, and mechanisms of long non-coding RNAs in cancer," *Genomics, Proteomics and Bioinformatics*, vol. 14, no. 1, pp. 42–54, 2016, doi: 10.1016/j.gpb.2015.09.006.
- [11] K. Hur, S.-H. Kim, and J.-M. Kim, "Potential implications of long noncoding RNAs in autoimmune diseases," *Immune Network*, vol. 19, no. 1, p. e4, 2019, doi: 10.4110/in.2019.19.e4.
- [12] J. Wang, F. Wei, and H. Zhou, "Advances of lncRNA in autoimmune diseases," *Frontiers in Laboratory Medicine*, vol. 2, no. 2, pp. 79–82, 2018, doi: 10.1016/j.flm.2018.07.004.
- [13] W. Ahmed and Z.-F. Liu, "Long non-coding RNAs: Novel players in regulation of immune response upon herpesvirus infection," *Frontiers in Immunology*, vol. 9, p. 761, 2018, doi: 10.3389/fimmu.2018.00761.
- [14] K. D. Mayer-Barber and D. L. Barber, "Innate and adaptive cellular immune responses to mycobacterium tuberculosis infection," *Cold Spring Harbor Perspectives in Medicine*, vol. 5, no. 12, p. a0184242015, 2015, doi: 10.1101/cshperspect.a018424.
- [15] J. He *et al.*, "Differential expression of long non-coding RNAs in patients with tuberculosis infection," *Tuberculosis*, vol. 107, pp. 73–79, 2017, doi: 10.1016/j.tube.2017.08.007.
- [16] S. Huang, Z. Huang, Q. Luo, and C. Qing, "The expression of lncRNA NEAT1 in human tuberculosis and its antituberculosis effect," *BioMed Research International*, pp. 1–8, 2018, doi: 10.1155/2018/9529072.
- [17] S. Zhao *et al.*, *Bioinformatics for RNA-Seq Data Analysis*, United Kingdom: IntechOpen, 2016, doi: 10.5772/61421.
- [18] J. D. Blischak *et al.*, "Predicting susceptibility to tuberculosis based on gene expression profiling in dendritic cells," *Scientific Reports*, vol. 7, no. 1, pp. 1–11, 2017, doi: 10.1038/s41598-017-05878-w.
- [19] S. Andrews, "FastQC: a quality control tool for high throughput sequence data," 2010.
- [20] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014, doi: 10.1093/bioinformatics/btu170.
- [21] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype," *Nature Biotechnology*, vol. 37, no. 8, pp. 907–915, 2019, doi: 10.1038/s41587-019-0201-4.
- [22] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements," *Nature Methods*, vol. 12, no. 4, pp. 357–360, 2015, doi: 10.1038/nmeth.3317.
- [23] D. Kim *et al.*, "HISAT2: graph-based alignment of next generation sequencing reads to a population of genomes," *Nature Biotechnology*, vol. 37, pp. 907–915, 2019, doi: 10.1038/s41587-019-0201-4.
- [24] H. Li *et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009, doi: 10.1093/bioinformatics/btp352.
- [25] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with hisat, stringtie and ballgown," *Nature Protocols*, vol. 11, pp. 1650–1667, 2016, doi: 10.1038/nprot.2016.095.
- [26] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, "Stringtie enables improved reconstruction of a transcriptome from RNA-seq reads," *Nature Biotechnology*, vol. 33, no. 3, pp. 290–295, 2015, doi: 10.1038/nbt.3122.
- [27] S. Kovaka, A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, and M. Pertea, "Transcriptome assembly from long-read rna-seq alignments with stringtie2," *Genome Biology*, vol. 20, no. 1, pp. 1–13, 2019, doi: 10.1186/s13059-019-1910-1.
- [28] A. Frankish *et al.*, "GENCODE reference annotation for the human and mouse genomes," *Nucleic acids research*, vol. 47, no. D1,





- pp. D766–D773, 2019, doi:10.1093/nar/gky955
- [29] H. Mirsafian *et al.*, “Long non-coding RNA expression in primary human monocytes,” *Genomics*, vol. 108, no. 1, pp. 37–45, Jul. 2016, doi: 10.4111/in.2019.19.e4.
- [30] S. Mathew *et al.*, “Methods to Study Long Noncoding RNA Expression and Dynamics in Zebrafish Using RNA Sequencing,” *Comp. Biology of Non-Coding RNA. Methods in Mol. Biology*, 2019, pp. 77–110, doi: 10.1007/978-1-4939-8982-9_4.
- [31] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biology*, vol. 11, no. R106, 2010, doi: 10.1186/gb-2010-11-10-r106.
- [32] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology*, vol. 15, no. 550, 2014, doi: 10.1186/s13059-014-0550-8.
- [33] T. Ammuneit, N. Wang, S. Khan, and L. L. Elo, “Deep learning tools are top performers in long non-coding RNA prediction,” *Briefings in Functional Genomics*, vol. 21, no. 3, pp. 230–241, 2022, doi: 10.1093/bfpg/elab045.
- [34] L. K. Xin and A. Abdullah, “Deep learning in non coding variant (a brief overview),” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 3, pp. 1432–1438, Jun. 2020 doi: 10.11591/ijeecs.v18.i3.pp1432-1438.
- [35] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, vol. 8, no. 1, pp. 1–74, 2021, doi: 10.1186/s40537-021-00444-8.
- [36] J. M. Vaz and S. Balaji, “Convolutional neural networks (CNNs): concepts and applications in pharmacogenomics,” *Molecular Diversity*, vol. 25, pp. 1569–1584, May 2021, doi: 10.1007/s11030-021-10225-3.
- [37] X. Tang and Y. Sun, “Fast and accurate microrna search using CNN,” *BMC Bioinformatics*, vol. 20, no. 23, pp. 1–14, 2019, doi: 10.1186/s12859-019-3279-2.
- [38] L. Statello, C.-J. Guo, L.-L. Chen, and M. Huarte, “Gene regulation by long non-coding RNAs and its biological functions,” *Nature Reviews Molecular Cell Biology*, vol. 22, pp. 96–118, 2020, doi: 10.1038/s41580-020-00315-9.
- [39] P.-J. Volders *et al.*, “LNCipedia 5: towards a reference set of human long non-coding RNAs,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D135–D139, 2019, doi: 10.1093/nar/gky1031.
- [40] K. D. Pruitt, T. Tatusova, and D. R. Maglott, “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins,” *Nucleic Acids Research*, pp. D61 – D65, 2007, doi: 10.1093/nar/gkl842.

BIOGRAPHIES OF AUTHORS



Faizah Aplop     received the Honours degree in Information System Management from Faculty of Information Studies, Universiti Teknologi MARA (UiTM) in 2001. She received the Master degree in Information Technology from Faculty of Information Technology and Quantative Sciences at the same university in 2007. She then completed her Ph.D in Computer Science (Bioinformatics) from Concordia University, Montreal, Canada in 2016. Currently, she is the member of Faculty of Ocean Engineering Technology and Informatics-Computer Science program, Universiti Malaysia Terengganu (UMT), Terengganu, Malaysia. Her research interests include bioinformatics, transcriptomics, metabolic network reconstructions, and machine learning. She can be contacted at email: faizah_aplop@umt.edu.my.



Saharuddin Mohamad     is an Associate Professor at the Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia. He received his Bachelor of Engineering (Bioengineering) in 1998 and Master of Engineering (Bioengineering) in 2000 from the University of Tokushima, Japan. He successfully completed his Doctor of Engineering degree in functional system engineering in 2003 at the same university. Then, he took up a post-doctoral position in bioinformatics at Nara Institute of Science and Technology, Japan until 2004. He was appointed as lecturer in Bioinformatics Program, Institute of Biological Sciences, Faculty of Science, University of Malaya ever since. He served as Programme Coordinator for Bachelor of Science (Bioinformatics) Degree Programme from 2013 to 2014 and re-appointed as the coordinator from 2019 to 2020. He was appointed as Advisory Board Member of MyBioInfoNet (Malaysia Bioinformatics Network) for 2019-2021 session. He was elected as vice president of the Malaysian Society of Bioinformatics and Computational Biology for 2018-2020 and 2020-2022 session. He was appointed as the Head of Institute of Biological Sciences, Faculty of Science from January 2020 to January 2022. Currently, he serves as the Head of Centre of Research in Systems Biology, Structural, Bioinformatics and Human Digital Imaging (CRYSTAL), University of Malaya since 2017. His current research interest focuses on structural bioinformatics, with special interest in computer-aided drug design and protein engineering. He is also engaged in molecular bioinformatics research projects related to analysis and data mining of the NGS data in genomics, metagenomics, whole-exomic, and transcriptomics. He is also involve in translational bioinformatics research with his collaborators from Institute of Medical Research, Malaysia. He can be contacted at email: saharuddin@um.edu.my.