Identifying Cyanobacteria through Next-Generation Sequencing Technology for Modern Agriculture

Joko Pebrianto Trinugroho1,*, Faisal Asadi1, and Bens Pardamean2

1Bioinformatics and Data Science Research Center, 11480 Bina Nusantara University, Indonesia
2Computer Science Department BINUS Graduate Program, Master of Computer Science, 11480 Bina Nusantara University, Indonesia

Abstract. As the global demand for food continue to increase, it is important to find a way to meet the demand without creating any problems to the environment. Cyanobacteria have a prospective to be utilised for the modern agriculture, as they contribute to the improvement of the soil fertility, the crop yield, and they also do not harm the environment. Therefore, it is crucial to understand the species of cyanobacteria or the characteristics that could be used for modern agriculture. The development of Next-Generation Sequencing (NGS) technologies enables us to study the genome of cyanobacteria. Thus, we can study their characteristics by analysing the NGS data. This paper aims to elaborate a pipeline for genomic analysis on cyanobacteria from NGS data. We used a free Linux-based software tool, namely Bresseq to process the NGS sequencing raw data. This tool predicts mutations that occur in the genome of the sample, including single-nucleotide variation, insertions, and deletions which could be beneficial for the identification of a new species or a mutant of cyanobacteria which has the right characteristics for modern agriculture utilisation.

1 Introduction

Cyanobacteria are organisms with a microscopic size and the ability to perform photosynthesis, which can be found in different habitats, including marine, lakes, ponds, rocks, and soils [1-3]. It is believed that cyanobacteria have been living in the Earth for more than 2 billion years, which is much earlier than the other organisms such as plants and animals [4]. Cyanobacteria are also known to play a critical role in the agriculture as they can improve the soil fertility and the crop yield [5]. This makes cyanobacteria have a huge potential to be explored for the modern sustainable agriculture. Although there are a lot of species of cyanobacteria, not all of them could be utilised for agriculture. Therefore, effort is needed in identifying and discovering the right species and characteristics of cyanobacteria for modern agricultural utilisation.

One of the promising approaches to identify cyanobacteria is by studying their genome, which can be achieved using next-generation sequencing (NGS) technology. Unlike the standard sequencing method, the NGS method can easily capture the information of the entire genome of an organism. The development of NGS technologies has been extremely rapid for the last 20 years, which enables us to obtain the genomic or transcriptomic sequencing data quickly in a much more affordable price [6-15]. Currently, there are several NGS technologies that are widely known, including Illumina, Roche 454, IonTorrent, Pacific Biosciences (PacBio), and Oxford Nanopore [16, 17].

Analysing NGS data is one of the most important steps in studying the genome of cyanobacteria. Most often, we compare the genome of a newly identified cyanobacterium or a mutant to the reference organism, a cyanobacterium which its genome has been studied before [18]. In order to compare the difference between the genome of a new cyanobacterium with the reference sequence, it is crucial to have a parameter. There are several parameters that is commonly used for genomic analysis: single-nucleotide variation (SNV) or single-nucleotide polymorphism (SNP), short insertion or deletion (indel), and structural variation (SV) which comprises of long deletions or insertions, and rearrangements [19-22]. To perform genomic analysis, many software tools have been created and developed. However, most of the software tools for genomic analysis are created to analyse a large genome, such as human genome, so some reads pointing to the repetitive sequences are overlooked to speed up the process [23, 24]. Therefore, finding the right software tool to perform genomic analysis on cyanobacteria is important so that we can study the genome of cyanobacteria well.

Here, we describe a pipeline that is designated for performing identifying cyanobacteria through their genome. The pipeline consists of wet lab and dry lab, which are discussed in this paper. The key steps of the pipeline, which leads to the results, are also demonstrated.

* Corresponding author: joko.trinugroho@binus.edu, faisal.asadi@binus.edu, bpardamean@binus.edu

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
2 Literature review

2.1 Overview of NGS genome analysis of cyanobacteria

Overall, the cyanobacterial genome analysis pipeline consists of two parts, the wet lab and dry lab parts. The wet lab focuses on performing Whole Genome Sequencing (WGS) using Next Generation Sequencing technique. The workflow of the wet lab part is shown in Figure 1. The most important step of the Whole Genome Sequencing is to obtain the correct cyanobacterial sample. It is crucial to make sure that we have a sufficient number of cyanobacterial cells for genomic DNA extraction. After the genomic DNA of the sample has been successfully extracted, it is subsequently subjected to DNA purity measurement and DNA concentration measurement. This process is performed to ensure that the quality and quantity of genomic DNA meet the standard [25, 26]. If both the concentration and purity of the genomic DNA meet the standard, the next step is library preparation. This step is performed to make the genomic DNA sample compatible with the NGS sequencing technology/system that is used [27]. After the library preparation was successfully performed, the sample was then subjected to NGS sequencing with Ilumina Nextseq550. Ilumina Nextseq550 is a powerful system with a high accuracy and robust performance, which has been vastly used to study cyanobacterial genome [28-30]. The system produces sequencing raw data that needs to be further analysed.

3 Methodology

3.1 Data acquisition

All genome sequences were obtained from an online database. The reference sequence of cyanobacterium Synechocystis sp. PCC 6803 (BA000022.2) was retrieved from https://www.ncbi.nlm.nih.gov, while the sample of genomic data of Synechocystis sp. PCC 6803 GT-1 strain was retrieved from https://www.ebi.ac.uk. Both reference sequence and sample were then subjected to data analysis.

3.2 NGS sequencing data analysis

Analysis of the NGS sequencing data from cyanobacterial sample was performed using bresseq (v.0.35.5) software tool [31, 32]. This tool is free-access, based on Linux, and has been extensively used to analyse genome’s sequencing data from prokaryotes, including cyanobacteria ([18], [33-36]). To prepare the input for bresseq, we first extract fastq.gz files containing raw sample genomic data into fastq files. These files were then used for the analysis. In addition, we used the genebank (.gbk) extension file as the reference sequence, since it has already been annotated. We run bresseq in a default mode without any modifications. The detailed steps are explained in the results section.

4 Results and discussion

The dry lab part of the cyanobacterial genome analysis pipeline focuses on analysing NGS sequencing raw data using bioinformatic approach. The workflow of the dry lab part is shown in Figure 2. As mentioned in the methodology, we used bresseq software tool for the genomic analysis. Both sequencing raw data (FASTQ) and reference sequences are needed for the input. The first step is mapping reads using bowtie2 [37], which is already integrated in bresseq software. This step generates read alignments (SAM/BAM extension file). From the best-read alignments, the software then creates several evidences, such as New Junction Evidence, Missing Coverage Evidence, and Read Alignment Evidence, that are used to predict mutations [31]. The predicted mutations include large deletions or insertions, short deletions or insertions, and substitutions. The software then creates a list of mutations and evidence in three different formats, genome diff, Variant Call Format (VCF) [38], and Genome Variant Format (GVF) [39], which can be visualised interactively using Integrative Genomics Viewer [40]. Finally, the output of the data analysis is created in Hypertext Markup Language (HTML) extension file, which contains annotated mutation and evidence presented in a table [31].
The result of genomic analysis from cyanobacterial sample (Synechocystis sp. PCC 6803 GT-1 strain) is displayed in Table I. The table informed important information from the analysis, including evidence, position, type of mutation, annotation, gene, and description. The evidence tells the type of evidence that are used to predict the mutations as mentioned previously. The position reflects the location of the mutation within the genome. The mutation gives information about the type of mutation, while the annotation tells the change in coding sequence/amino acid of the genomic DNA. If the reference sequence has been well curated and annotated, the detailed information about the name of the gene and its description are also shown. This helps us to further analyse the data.

The genomic analysis predicted 27 mutations from the sample (Table I). This includes substitutions, short insertions or deletions, and large deletions. Single-nucleotide variations (SNV) and short insertions or deletions (indels) were the most predicted mutations from the analysis. SNV and Indels commonly occurred in the genome, so they are used as genetic markers [41-43]. Among all predicted mutations, the mutation in the psaA and sps genes may be beneficial for the growth of cyanobacteria, as both genes are involved in the metabolism [44]. psaA gene encodes photosystem I protein subunit, which is crucial for photosynthesis, while sps gene functions in sucrose synthesis [44, 45]. Variation in these genes could lead to the improvement of cells’ biomass, which could be further utilised for agriculture [45]. Overall, our genomic analysis has successfully predicted mutations from NGS sequencing data of the cyanobacterial genomic DNA.

### Table 1. List of predicted mutations.

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Position</th>
<th>Mutation</th>
<th>Annotation</th>
<th>Gene</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C=7</td>
<td>P=104</td>
<td>C(CT→CT)</td>
<td>ORF</td>
<td>P=104 -&gt;</td>
</tr>
<tr>
<td>2</td>
<td>C=7</td>
<td>R=140</td>
<td>(CGC→AGC)</td>
<td>ORF</td>
<td>R=140 -&gt;</td>
</tr>
<tr>
<td>3</td>
<td>C=7</td>
<td>S=294</td>
<td>(GAG→AAG)</td>
<td>ORF</td>
<td>S=294 -&gt;</td>
</tr>
<tr>
<td>4</td>
<td>G=6</td>
<td>V=104</td>
<td>(GTC→ATC)</td>
<td>ORF</td>
<td>V=104 -&gt;</td>
</tr>
<tr>
<td>5</td>
<td>G=6</td>
<td>I=185</td>
<td>(CTG→CTG)</td>
<td>ORF</td>
<td>I=185 -&gt;</td>
</tr>
<tr>
<td>6</td>
<td>G=6</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>7</td>
<td>A=7</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>8</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>9</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>10</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>11</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>12</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>13</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>14</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>15</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>16</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>17</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>18</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>19</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>20</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>21</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>22</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>23</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>24</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>25</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>26</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
<tr>
<td>27</td>
<td>T=4</td>
<td>L=220</td>
<td>(CTT→TCA)</td>
<td>ORF</td>
<td>L=220 -&gt;</td>
</tr>
</tbody>
</table>

### 5 Conclusion

In this paper, we presented a genomic analysis pipeline from NGS data of cyanobacteria, which was obtained
from an online database. The genomic analysis in this study used a free-access software tool with simple commands. Our results have predicted different mutations, including SNV and indels. Hence, this pipeline will be useful to identify and discover a new species or a mutant of cyanobacteria which has the right characteristics for modern agriculture utilisation.

References


15. A. Budiarto, B. Pardamean, Explainable supervised method for genetics ancestry estimation, in 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI) (2021)


17. V. Tripathi, P. Kumar, P. Tripathi, A. Kishore, M. Kamle, Next-generation sequencing (NGS) platforms: an exciting era of genome sequence analysis, in Microbial Genomics in Sustainable Agroecosystems, pp. 89–110 (Springer, 2019)


21. B. Mahesworo, A. Budiarto, B. Pardamean, Systematic evaluation of cross population polygenic risk score on colorectal cancer, Procedia Computer Science, pp. 1–8 (2020)


30. A. V. Bryanskaya, A. A. Shipova, A. S. Rozanov, O. A. Volkova, E. V. Lazareva, Y. E. Uvarova, T. N. Goryachkovskaya, S. E. Peltek, Metagenomics dataset used to characterize microbiome in water and sediments of the lake solenoe (novosibirsk region, Russia), Data in Brief 34, 106709 (2021).


