**Software benchmarking of the assembly strategies of the PATRIC BRC using *Janthinobacterium sp.* data sets**

**Student**: Juan Pablo Rojas-Sossa

**Department:** Biosystems and Agricultural Department.

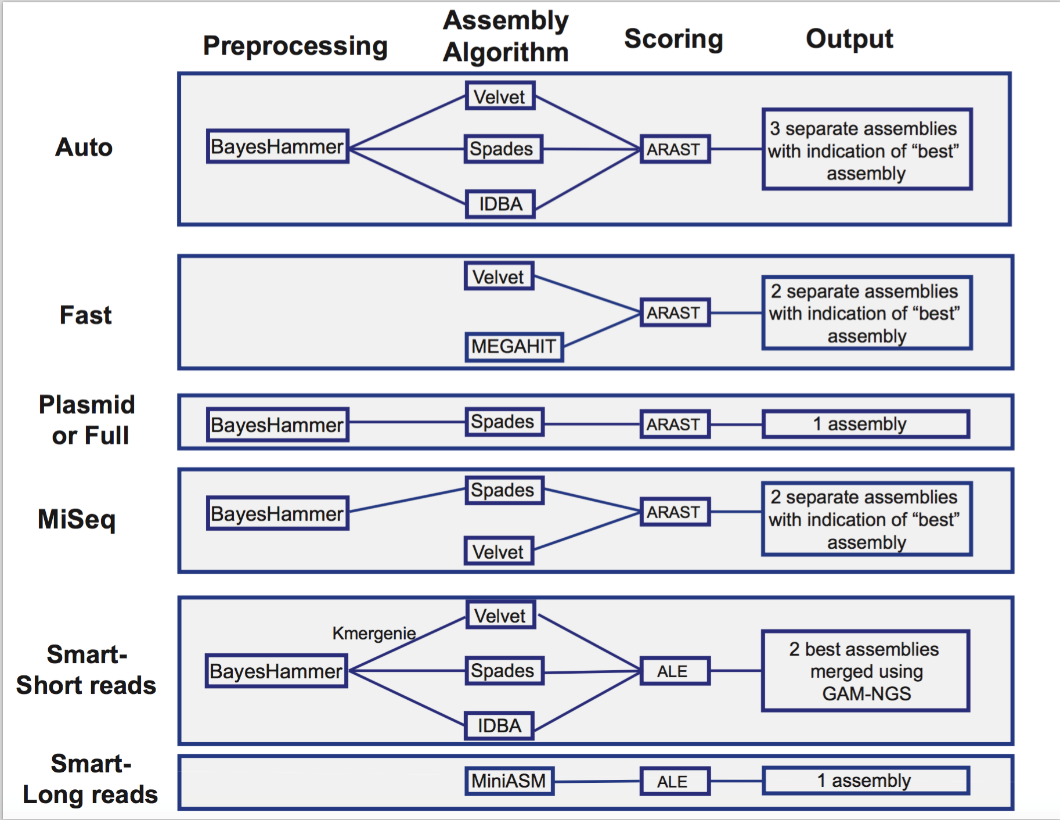
**Mentor**: Terence Marsh, PhD.

* **Problem**

Evaluate the genomic differences between six sets of *Janthinobacterium* strain that shown phenotypic differences. The goals be will include the assembly and analysis the *Janthionobacterium* sequences along with a comparative genomics analysis plus the addition of phenotypic analysis. Because this data is not ready jet to finish the entire project. All the computation will be made on PATRIC bnr.

* **Current approaches**

The actual approach used on PATRIC is a selection of one this method, and later obtain and output as explain the Figure 1. PATRIC allows the use of different strategies to assembly a set of reads. This project will be focus in five of the methods: auto, fast and full spades, Miseq, Smart-short reads.



**Figure 1.** Assembly’s strategies in PATRIC BRC.

In Auto and Full spades, there is an preprocessing step using the algorithm BayesHammer (Nikolenko, Alekseyev, Korobeynikov, & Alekseyev, 2013), this tool is an error correction on non-uniform coverage base in integration of Hamming graphs and Bayesian clustering also called machine learning classification. This step is not use in the fast strategy.

Later, the different strategies is used the assembly algorithm that produce the graphs. The auto strategy used three different assembly tools (Velvet, Spades, IDBA), this process will give as result three different assemblies and ARAST scoring software selects the best assembly. The fast strategies uses Velvet and MEGAHIT to assembly the raw reads and then use ARAST to score the assembly and output the best assembly. Finally, full strategy uses only Spades to make the assembly and output just one assembly.

* **New Approach**

The objective will be compare 5 *de novo* assembly strategies (Auto, Fast and Full spades) and later make an alignment with a finished genome. The final step will be compare the results of assemblies with a reference genome assembly and see the effect of the computing efficiency and scoring the real value. Using the different database like JGI database will be necessary to found one-finished genomes, and then make a reference alignment using this finished one and using SAM software to make a mapping alignment and later make comparison between methods.

* **Significance & Impact**

The main goal of the project is to deliver best data assembly result to an ongoing data sets provided, this a collaboration project is happening in project that required the data results as the same of being a good benchmarking effort of the PATRIC platform. This data later will be used to a comparative genomics analysis and the addition of phenotypic data that are getting collected, right now. My mentor is fully expect to publish these results in a proper publication.

This process will be a very important input to Janthinobacterium’s project first, later the results will allow to review the PATRIC strategies and will provides a performance comparison between them. We aim to highlight the experimental parameters and the computing preparation of each method and provide explicit guidance for practitioner.

* **Project Goals**

1. The first step will be make assembly of the five PATRIC’s methods.
2. Later will be necessary to make an assembly by reference or by mapping using SAM tools.
3. Finally, will be necessary analyze the data obtain and make plots and comparisons of the entire benchmarking exercise.

* **Division of Work/ Division of Work Project Goals**

Because for this project I do not have partners, I will be doing the project only by myself.

* **Milestones**

Last week the first results of the assembly were done all the calculation was made on the PATRIC’s cloud, the results of the assembly from three of the PATRIC’s methods Fast, Full\_Spades and Auto. The QUAST 4.6.3 parameters obtain are presented in Figure 2. The bigger number of contigs is the Fast in comparison with Spades\_Full and the Auto method. The higher N50 number obtain is the assembly made by the full SPADES, fallow by the Auto. The results obtain for the fast methods are significantly lower. N50 is the length for which the collection of all contigs of that length or longer covers at least half an assembly. Preliminary showing that the Full spades method is the best assembly obtain from the sequence data.

**Figure 2**. QUAST metrics reported in three of five PATRIC methods.

* **Datasets**

The data used for the analysis are pair end shotgun sequence reads. Six paired end files were obtain from the thirty cultures performed in the laboratory to make the analysis. Six pair end files of the culture of 30 samples. This QC values were obtain from the reports of FastQC and QUAST.

The description of the number of sequences. All this files has been using together to perform the assemblies pipelines before descripted. This number are presented in the Table 1.

**Table 1.** Number of sequences obtain in each paired end sequence file.

|  |  |
| --- | --- |
| **Filename** | **Total Sequences** |
| Janthinobacterium\_3-1\_S3\_L001\_R1\_001.fastq.gz | 870,217 |
| Janthinobacterium\_19-1\_S2\_L001\_R1\_001.fastq.gz | 990,766 |
| Janthinobacterium\_35-1\_S1\_L001\_R1\_001.fastq.gz | 964,218 |
| Janthinobacterium\_51-1\_S11\_L001\_R1\_001.fastq.gz | 824,914 |
| Janthinobacterium\_MN-18\_S4\_L001\_R1\_001.fastq.gz | 3,373,652 |
| Janthinobacterium\_MN-39\_S5\_L001\_R1\_001.fastq.gz | 2,769,829 |

* **Challenges**

As main challenge that will be, address correctly the mapping assembly evaluation. The computation time and the usage of the SAM tools, this software has been compiled in C, right now the knowledge in C is basic. So, the usage of this assembly methods implies a challenge in the future.

* **Computation Cost**

The computational cost was lower for the Auto and the Fast assembly, the smart, Miseq (Kiki) and Full required more time to obtain the data, the data is visible in Table 2. Also, Is important to say that this values depend of the disponible servers in PATRIC.

**Table 2.** Computation cost of assembly in PATRIC.

|  |  |  |  |
| --- | --- | --- | --- |
| **Strategies** | **Start** | **End** | **Time consume (h)** |
| Auto | 3/15/2018 17:27 | 3/15/2018 18:06 | 0:39 |
| Smart | 3/14/2018 8:02 | 3/15/2018 22:08 | 14:06 |
| Full | 2/26/2018 11:26 | 2/27/2018 1:30 | 14:04 |
| Fast | 2/26/2018 11:23 | 2/27/2018 21:05 | 9:42 |
| Miseq | 2/23/2018 1:16 | 2/23/2018 21:59 | 20:43 |

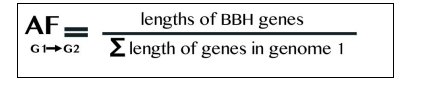
**Downstream Process**

After obtaining the contigs file several downstream assembly parameters were trying to study. Below some of these.

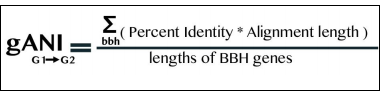
* *ANI Tools*

Stands for Average Nucleotide Identity (ANI). Is a measure of nucleotide-level genomic similarity between the coding regions of two genomes. This were calculated in the cloud using [IMG/JGI](https://img.jgi.doe.gov/) database but scaffold file was required

Alignment fraction:



ANI Formula:



* *COMPASS*

Is implemented in R, COMpare a DNA sequence ASSembly to a trusted reference sequence. Features are: (i)Increased sensitivity and selectivity of homology detection; (ii) longer, more complete alignments; and (iii) faster computational speed

* *SSPACE*

Stand-alone program for scaffolding pre-assembled contigs. Uses the distance information of paired-end to assess the order, distances and orientation of contigs. Implemented in linux. Installation process process is tricky. Finally, I install the software and yesterday I started to scaffold the first data set.

**References**

Nikolenko, S. I., Alekseyev, M. A., Korobeynikov, A. I., & Alekseyev, M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, *14*(1), S7. http://doi.org/10.1186/1471-2164-14-S1-S7