Jonathan Taylor

09/23/2020

Biol 493L

De novo Genome Assembly of Pelagibacter ubique

Introduction

P. ubique is a specific α-proteobacteria that is found in the SAR11 clade, which is found in the ocean. the SAR11 clade is one of the most successful microbial groups on Earth, with an estimated 2.4 × 1028 cells in the oceans (Morris et al. 2002). The relevance of P. ubique is that it has the smallest genome in the entire SAR11 clade, yet it has complete biosynthetic pathways for all 20 amino acids. There is also a very small percentage of nonfunctional and redundant DNA in this genome. The question is why has this genome become such a dominant α-proteobacteria in the SAR11 clade even though it has the smallest genome? It seems that evolutionary pressure has pushed P. ubique to expel everything from its genome except the most essential process for survival. This goes along with the genome streamlining theory. According to the genome streamlining theory, when effective population sizes are large, extreme selection for the efficient use of resources in nutrient-poor environments can result in genome reduction (Dufresne et al., 2003; Lynch and Conery, 2003; Giovannoni et al., 2005). In the strain of P. ubique there are 9,986,668 Million, 3G bp paired-end Illumina reads. This data was collected using Illumina HiSeq 2000. The goal is to assemble the genome of P. ubique and look at specific contigs assembled. By assembling a de novo genome of P. ubique it could potentially give crucial information as to why this specific α-proteobacteria has evolved into the dominant microorganism that it is.

Methods

The goal is to assemble a de novo genome of P. ubique. I am not very skilled in computer coding, so I will be doing a somewhat simpler approach to a genome assembly. I have experience in molelcular and genetic biology, this will help with understanding specifics about the genome sequence. To start, I need to retrieve the raw read sequence data. This is done by searching for it on the SRA database. After finding the raw read sequence data, the next step is to inspect the data for quality control. This is going to be done using the ‘fastqc’ software in git spruce repository. This will give important information about the raw data collected. After collecting this data, the next step is to trim/clean the reads. The software that will be used for this is ‘Trimmomatic’. This software is good for paired end reads. Within this software there is multiple functions that will need to be done in order to get the best possible output. Some of the functions may not need to be used, depending on the results from the fastqc report. The functions are as follows: Adapter trimming, Sliding window trimming, Trailing bases quality trimming, Leading bases quality trimming, and Minimum read length. After completing the functions necessary, it is important to look at the number of reads orphaned by the trimming/cleaning process, and the number of pairs lost. These outputs are what will be used for assembly. The next step is to do the assembly. The software that is going to be used in the assembly process is ‘velvet’. the de novo short-read assembler Velvet (Zerbino, 2010) is used to generate multiple assemblies by varying the k-mer size between 66 % and 90 % of the read length using VelvetOptimiser (Gladman & Seemann, 2008). It can form long contigs by manipulating de Bruijn graphs. A de Bruijn graph turns the formidable challenge of assembling a contiguous genome from billions of short sequencing reads into a tractable computational problem. (Compeau, P., et al. 2011). Velvet automatically searches for the optimal assembly to use. The important thing to note for velvet is choosing the most suitable read files and k-mer size search range. This can be done by inspecting the fastqc output data. The next step is to make sure the assembly quality is good and examine the draft contigs. You can check this in the velvet log file created from the velvet software. Now that the assembly is complete, it is important to look at the contigs that have been produce. Once everything looks good in the contigs, the genome assembly is finished.

Expected Outcomes

The overall assembly of the de novo genome should output significant and useful data to observe and give information about the functions of P. ubique. Although, it does seem that there may be some gaps in the genome assemblies contigs. This could be a problem with retrieve completely accurate information about P. ubique. There are certain spruce tools that could be useful in filling these gaps. There could be better software that could be used in the assembly of the genome. Other problems that could be encountered are that the quality checks may not be completed correctly and could skew data causing mistakes in the assembly, so it is essential to do multiple quality checks and make sure everything is correct. With this being said, after further investigation, there should sufficient data in the completed assembly to assume that evolution has depleted P. ubique of all non-essential cellular functions so that it replicates as efficiently as possible making it the oceans most dominant clade.

References

Compeau, P., Pevzner, P. & Tesler, G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 29, 987–991 (2011).

Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann I, Barbe V et al (2003). Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci USA 100: 10020–10025.

Gladman S., Seemann T.(2008). Velvet optimiser.

Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ. SAR11 clade dominates ocean surface bacterioplankton communities, Nature, 2002, vol. 420 (pg. 806-810)

Zerbino D. R.(2010). Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics 3111.5.1–11.511.