COMPARISON ANALYSIS OF SOFTWARE AND IDENTIFICATION OF AN EFFECTIVE ALGORITHM FOR WHOLE GENOME SEQUENCING OF BACTERIA

Shevtsov V.A.

Object of research:

Software:

Data sets obtained by sequencing bacterial genomes based on MiSeq and IonTorrent platform.

Novelty:

Will be carried out an analysis of genome-wide data of strains circulating in Kazakhstan.

Purpose and main objectives of the study:

The purpose of this research is to study the varieties of algorithms and to identify the most effective for processing of full-genome data of bacteria.

Expected results:

An efficient algorithm for processing of full-genome bacteria data will be identified.

Research problem:

- To review the literature describing methods of Assembly and analysis of genomic data of bacteria;

- To study the principle of operation of different types of software for assembling full-genome data;

- To obtain and evaluate the suitability of whole-genomic data of bacteria circulating in Kazakhstan for bioinformatic analysis;

- Develop algorithms for assembling of full genomes using various software;

- Conducting of simulation experiments on real and synthetic genomes to characterize the performance of algorithms for final processing of whole-genomic data.

Methods: Bioinformatics.

Practical and theoretical significance:

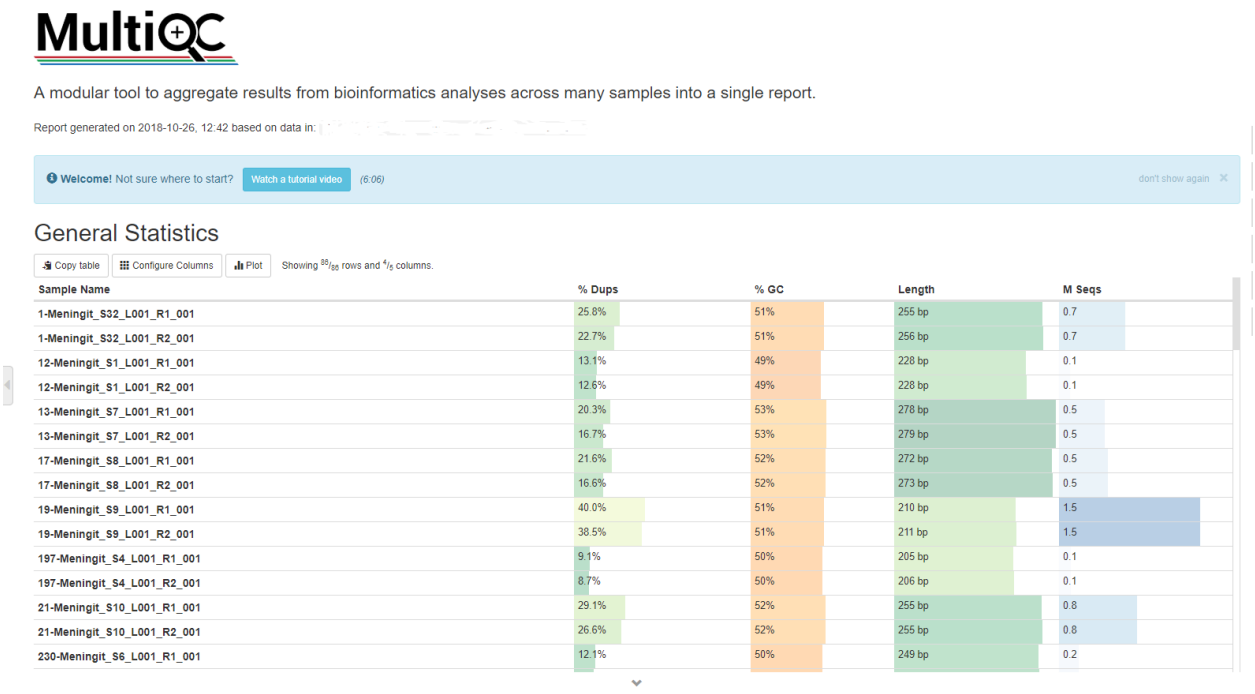
In Kazakhstan, there is a rapid introduction of technologies of full-genome sequencing of bacterial strains in clinical epidemiology, fundamental and applied biotechnology. Currently, full-genome data allow to establish the factor of pathogenicity or drug resistance of bacterial pathogens and to correct therapy in a timely manner. Also in biotechnology, full-genome data allow us to select high-value and promising strains. However, the bioinformatics school remains underdeveloped. As part of this work, the algorithm of processing of full-genome data of bacteria will be optimized, which can be used in practice in epidemiological and biotechnological areas.

The relevance of research:

Determination of nucleotide sequence of genomes is currently the main technology in biological researches. 20 years ago, the study of genomic data seemed expensive and difficult to solve in the field of nucleotide sequence determination. The primary cost of the human genome was estimated at $ 3 billion. Progress in the development of next-generation sequencers (NGS) instantly changed the situation, at the moment the human genome which costs about $ 1000 became a reality 4. Reducing the cost of genomic research has led to the rapid development of this area, and already implemented ambitious projects aimed at genomic sequencing of 5000 insects Arthropod Genomic Consortium, 2014, 5, ten thousand genomes of vertebrates Genome 10K Community of Scientists, 2009, millions of microorganisms, etc. However, the generation of data is only a part of the representation of full-genome data, the next no less time-consuming step is the Assembly of short sequences into a coherent informative "thread of life" with annotation and prediction of the open reading framework function. There are various technologies in obtaining full-genome data, but the most common are developed by Illumina and LifeTechnology. Which generates many short fragments 100-250 BP in length. Due to short reads, there are difficulties in the Assembly of the genome associated with tandem repeats, which exceeds the length of the reading, palindromes, etc.. That does not allow to correctly sew the resulting

contigs and many genome assemblies remain fragmented with a large number of errors in the arrangement of the fragments. Confirmation of this fact is described in the work of Salzberg and Yorke, which indicates hundreds of incorrect compounds in the genomes. Alkan and co-authors reported that the Assembly of the "de novo" human genome using short fragments is 16% shorter than the genome assembled using much more expensive approaches. The difficulty in processing short fragments in the Assembly of complete genomes provoked the development of a wide range of programs collectors. And yet, despite the number of tools available, there are no universal algorithms for processing genomes and data sets obtained from different platforms.

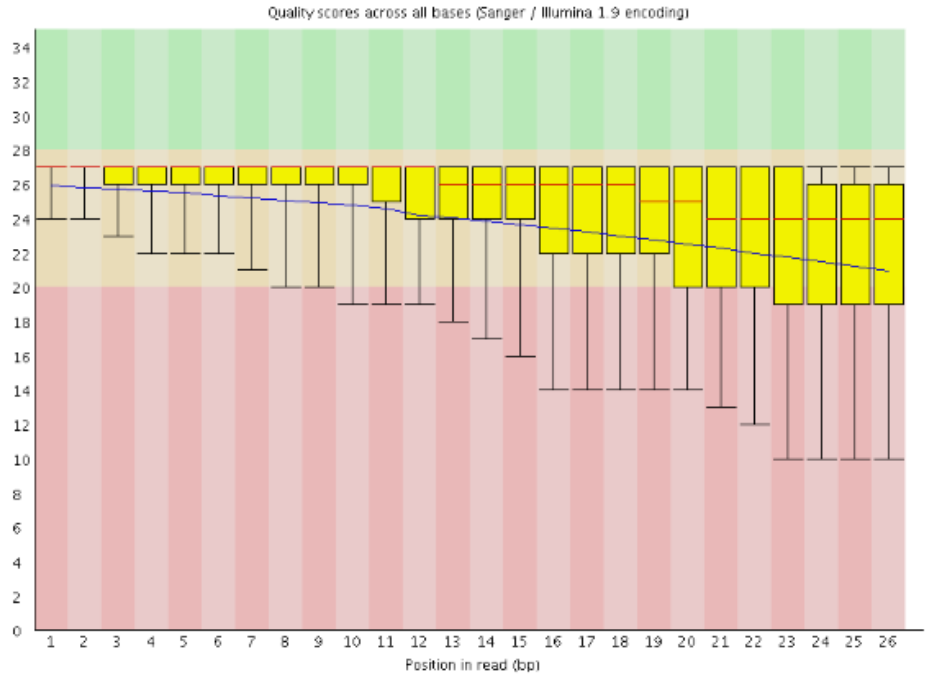
At this stage of the study was cloned and sequenced several samples of menigococcal, and was conducted a quality control of samples (Fig.1) by using of FastQC software. The main function of this software is to provide an easy way to check the quality of raw sequence data coming from high throughput sequencing pipelines. (Babraham Bioinformatics - FastQc)



Pic.1 Visual presentation of quality control of samples.

The quality control verification phase is necessary to verify and visualize the suitability or unsuitability of data for further processing. If the sequence is successful, then the absolute majority of positions in the reading will be in the green zone as shown in Fig. 2, if some of the readings are in the red zone, which means that these data cannot

be used in their process. Processing, in this case, software is usually used to separate useful data from unnecessary.



Pic.2 Data quality control chart

In a further study, will be used a software to trim the data from the data unsuitable for further manipulation over them. Such solutions as Trimmomatic, SeqPurge, AfterQC and identification of the best solution for cropping unsuitable data will be applied.

Conclusion

The genome Assembly and annotation of the genome is a field where there is no gold standard. Projects are often research projects, and knowing whether the results are good or bad is often difficult to determine. This is especially true when working with organisms that only vaguely resemble already sequencing and published organisms, with the result that there is little to compare them with.

With the development of new sequencing technologies, high-quality genome Assembly is becoming more feasible than ever, and a well-assembled and annotated genome will be a resource that can be used for many years.