Studying antibiotic resistance via alignment and variants detection

Kamilla Faizullina, Ignat Sonets

In this project, we analyze sequencing data from a strain of E. coli in order to find variants which could lead to the resistance to the antibiotic ampicillin. We use open-source tools to process data and detect mutations. We found the positions of variants in the genome and suggest the possible effects of each mutations.

1   Introduction

Antibiotics are the powerful drugs which used to treat or prevent some types of bacterial infection. However, the enormous and irresponsible use of the antibiotics has contributed significantly to the advent of the resistant strains [1], [2]. Development of new antibiotics can be helpful. To that end, scientists carry out researches related to mechanism of resistance.

In this work, we use the set of real sequencing data from a strain of E.coli resistant to the antibiotic ampicillin in order to analyze it and detect single-nucleotide polymorphism (SNP). The project aims to find the mutated genes to research the mechanism of resistance and attempt to make recommendations for alternative antibiotics.

2   Methods

2.1   Data

Sequencing data	Number of reads
Original	455876
Improved	446259

Table 1: Number of reads

First, We use the reference sequence of the Escherichia coli strain [3]. This is E.coli strain K-12 substrain MG1655, laboratory workhorse. Second, we use the sequencing data of an E. coli strain which is resistant to the antibiotic ampicillin [4].

We use Fastqc for the quality control [5] (See 5.1). With the aim to improve the overall quality of the sequencing data Trimmomatic tool is applied [6]. We cut off off the start and the end of a read if quality below 20. We also drop the reads if their lengths are below 20. The window size is 10 and the average quality within the window 20.

Both forward and reverse initial data consist of 455876 reads. The improved data contains 446259 reads.

2.2   Aligning sequences to reference

As we need to map the data from the resistant strain to the reference sequence, we use the aligner called BWA-MEM [7]. In order to prepare the data we use [8]. This utility allows to index and sort bam-format files.

VarScan is used to find SNP [9]. VarScan allows to change the threshold which regulates if the position is called mutation. We decided to use –min-var-frequency=0.95.

3   Results

First, we improve sequencing data quality of the resistant strain and index data. After, we align the sequence from resistant strain to the reference sequence to detect variants.

The filtered data is used to detect the variants. Five variants is found.

Position	Ref	Alt
93043	C	G
482698	T	A
852762	A	G
3535147	A	C
4390754	G	T

Table 2: Results

4   Discussion

Position 852762 does not codes protein according to NCNI [10].

The changing in position 852762 could affect the EG11704 (EcoCyc) which codes multidrug efflux pump RND permease AcrB. This is the inner membrane component of a tripartite.

The variants in position 3535147 could change the gene EG10269 (EcoCyc). This is the member of the two-component regulatory system EnvZ/OmpR.

Mutation in position 93043 affects the gene G7841 (EcoCyc) which codes small ribosomal subunit biogenesis GTPase RsgA. RsgA induces local conformational changes in the 30S subunit and may thereby destabilize kinetically trapped assembly intermediates

Mutation in position 93043 affects the gene EG10341 (EcoCyc) [11]. This part of the genome decodes FtsI (penicillin-binding protein 3, PBP3) [12]. Penicillin-binding protein is essential cell division protein that catalyzes cross-linking of the peptidoglycan cell wall at the division septum. The variants of this gene could make strain resistant to antibiotic.

5 Supplement

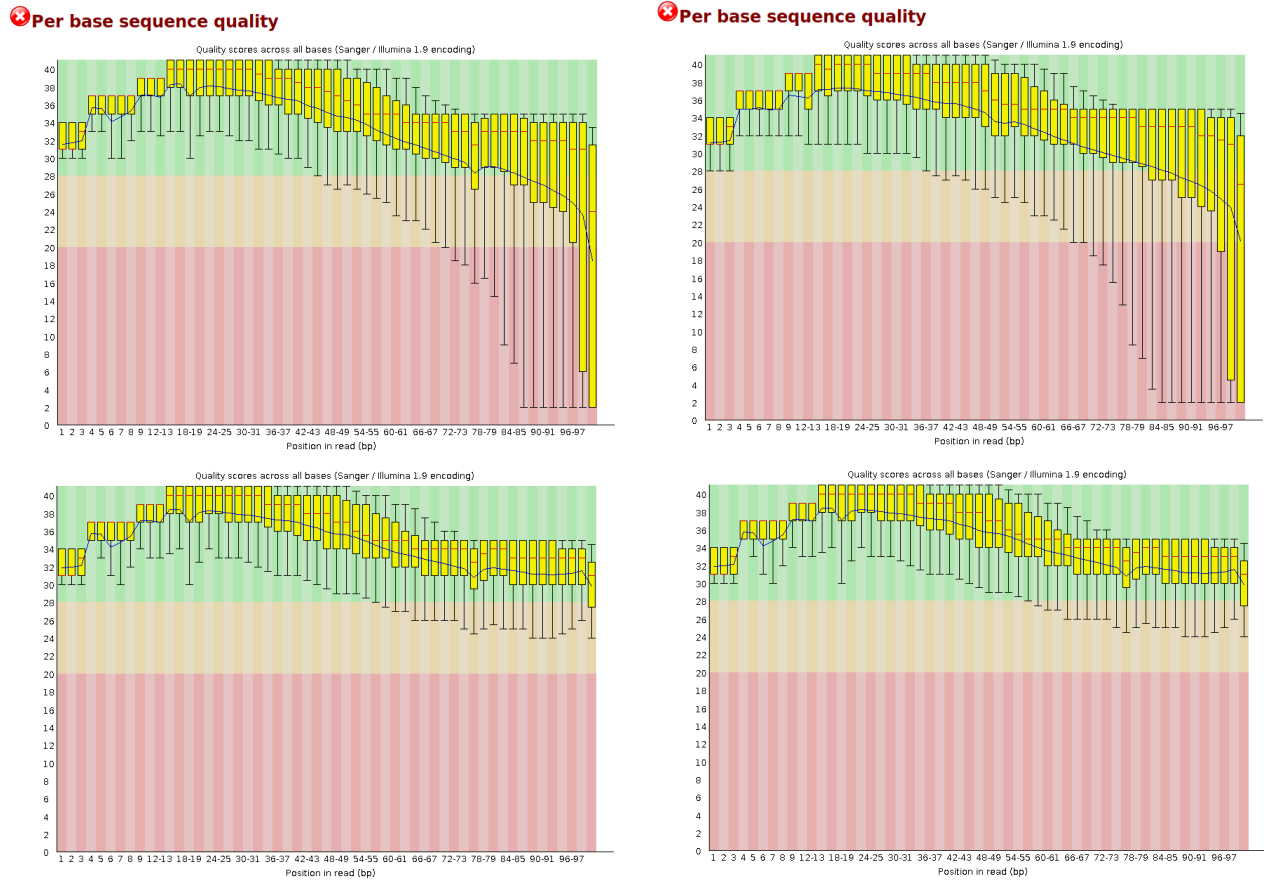## 5.1 Improvement of the overall quality of the initial data

Figure 1: The quality of forward (left) and reverse (right) sequence before and after applying Trimmomatic

References

[1] Zaman SB, Hussain MA, Nye R, Mehta V, Mamun KT, Hossain N. A Review on Antibiotic Resistance: Alarm Bells are Ringing. Cureus. 2017;9(6):e1403. Published 2017 Jun 28. doi:10.7759/cureus.1403

[2] Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. P T. 2015;40(4):277–283.

[3] The reference data, retrieved from NCBI FTP

[4] The data of resistant strains: first and second

[5] Fastqc

[6] Trimmomatic: A flexible read trimming tool for Illumina NGS data

[7] Burrows-Wheeler Aligner

[8] SAM (Sequence Alignment Map)

[9] VarScan

[10] NCBI

[11] EcoCyc

[12] UniProt