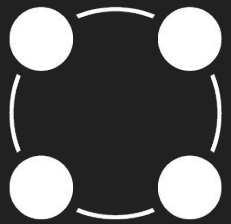
Геномная структура штаммов *Mycobacterium tuberculosis*, распространенных в различных регионах мира

Климов Владимир
Молчанов Владимир

Руководитель: Черняева Екатерина
СПбГУ, Центр геномной биоинформатики
им. Ф.Г. Добржанского

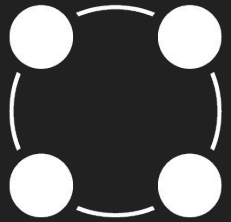


Напомним о проблеме

Проект направлен на анализ и систематизацию данных полногеномного секвенирования штаммов *Mycobacterium tuberculosis* (возбудитель туберкулеза), распространенных в различных регионах мира.

Почему?

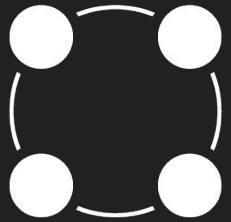
- Высокая эпидемиологическая значимость
- Большой круг заинтересованных лиц
- Постоянный прирост новых геномных данных



Цель проекта: Совершенствование базы данных геномных вариаций *M.tuberculosis* GMTV, Центра геномной биоинформатики им. Ф.Г. Добржанского СПбГУ

Задачи проекта:

- Осуществить поиск данных полногеномного секвенирования в нуклеотидных архивах
- Систематизировать информацию о секвенированных штаммах из литературных источников
- Найти нуклеотидные вариации (SNPs и Indels)
- Создание каталога известных мутаций ассоциированных с резистентностью к антибиотикам
- Провести филогенетический анализ и выявить закономерности распространения штаммов в различных географических регионах



Ход работы


Данные секвенирования
Фильтрация, Оценка
Поиск мутаций
Систематизирование информации
Филогенетический анализ
База данных

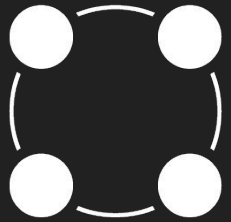
Климов Владимир Часть 1

Знакомство с литературой (Проблематика)

Знакомство с программами и существующими pipelines (Выбор “оптимального”)

Разработка скрипта (BWA-mem -- GATK) и его запуск (6 штаммов)

Проверка результатов работы скрипта (поиск известных мутаций в .vcf)

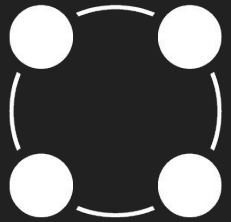


Климов Владимир Часть 2

Разработка скрипта для работы с большими объемами данных

Поиск мутаций в 999 штаммах (Республика Малави, проект: https://www.ebi.ac.uk/ena/data/view/PRJEB2794)

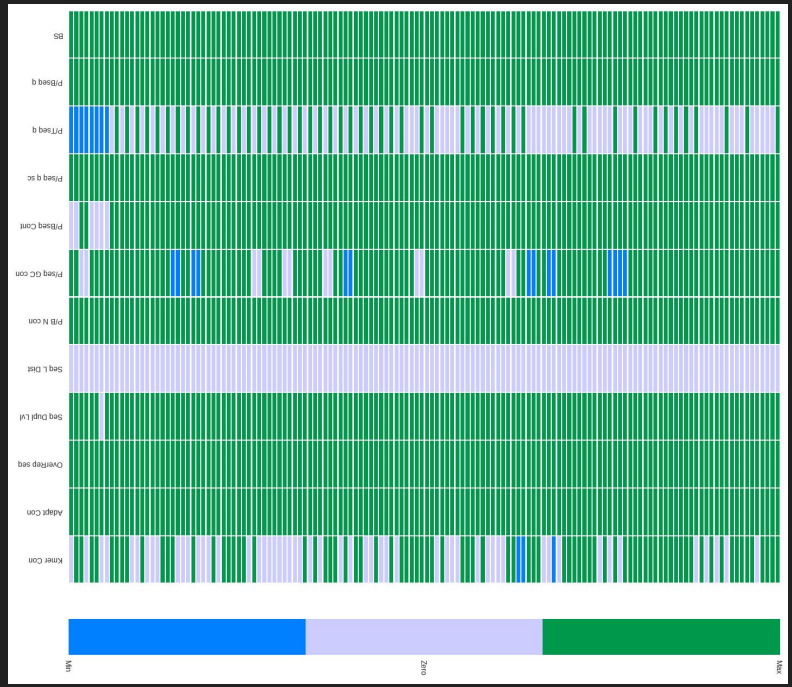
Филогенетический анализ полученных геномных вариаций



Климов Владимир Часть 2

Heatmap для оценки качества после тримминга

- fastQC Тест пройден - зеленый
- fastQC предупреждение - серый
- fastQC Тест не пройден - синий

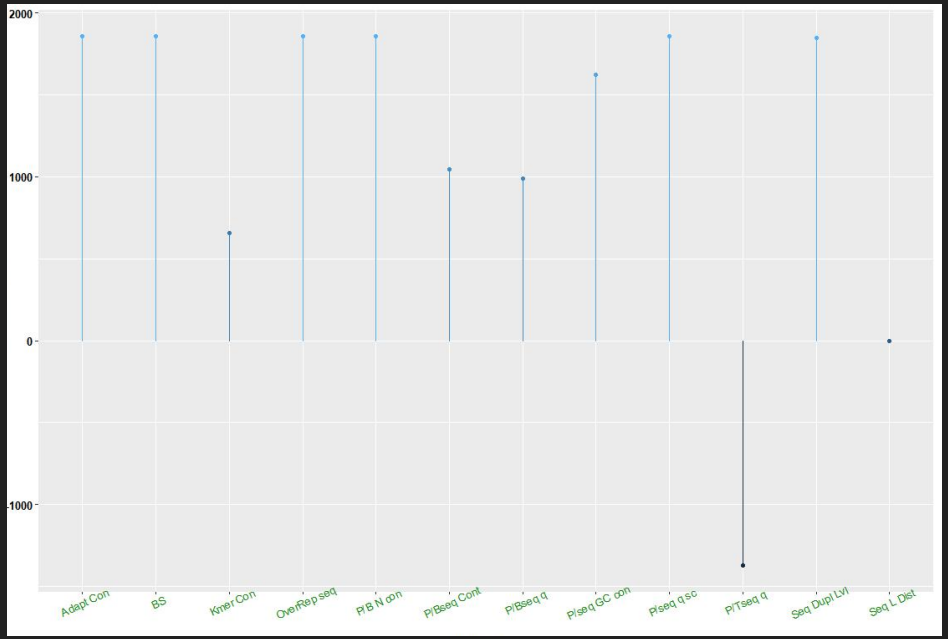


Климов Владимир Часть 2

geom_point() для оценки качества после тримминга

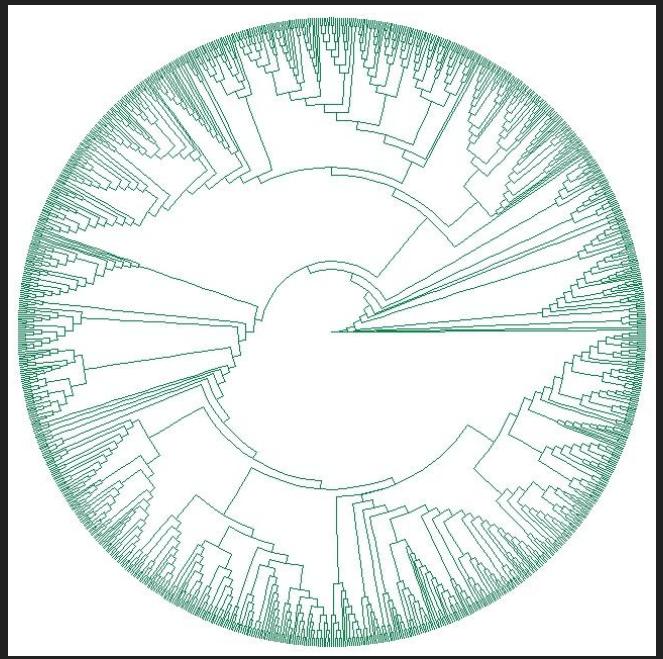
По оси x - fastQC тесты

По оси y - общая сумма оценок по тестам



Климов Владимир Часть 2

- 999 штаммов из Республики Малави Mycobacterium canettii - аутгруппа
- Парсинг .newick, выбор размера клад
- Пересечение .vcf клады - совокупность мутаций, которые определяют кладу?
- “Аннотация” геномных вариаций специфичных для клады
- Выбор мутаций ассоциированных с резистентностью



Молчанов Владимир (Часть 1)

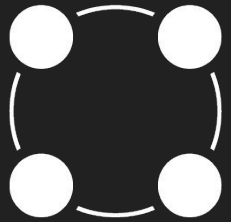
1. Сравнительный анализ программ и их комбинаций

2. Выбор альтернативного пайплайна

3. Создание скрипта пайплайна для поиска SNP с оценкой качества ридов и тримминга на основе результатов анализа

4. FastQC -> Summary -> Trimmomatic-0.36 -> **Bowtie2** -> **Samtools** -> .vcf

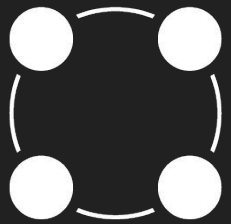
5. Тест пайплайна и сравнение результатов с альтернативным пайплайном

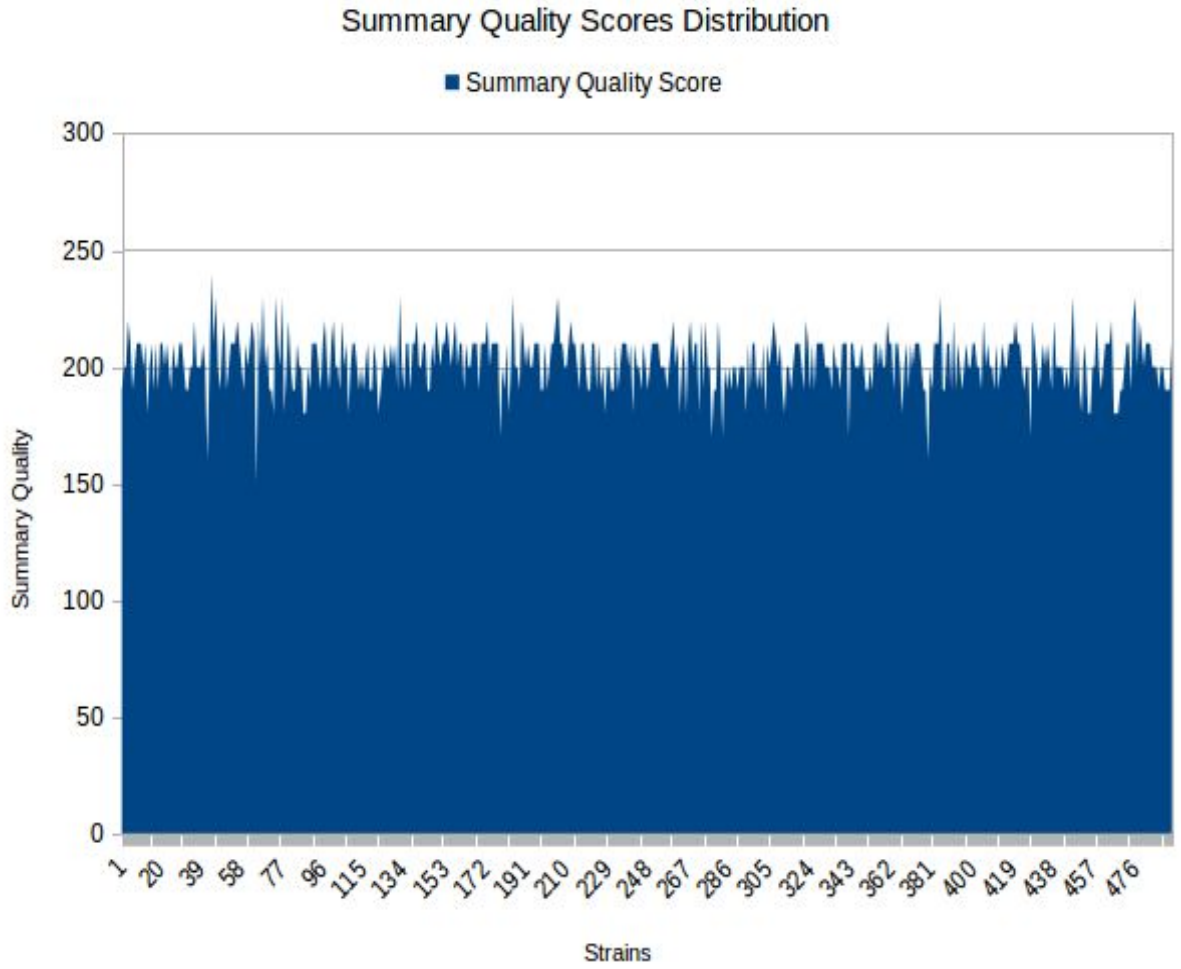


Молчанов Владимир (Часть 2)

1. Загрузка и обработка fastq файлов штаммов *M. tuberculosis* (Оксфордшир, Британия), Bioproject - PRJEB 5162

2. Оценка качества обработанных с помощью trimmomatic ридов > ***выполнен отчет по результатам оценки качества***

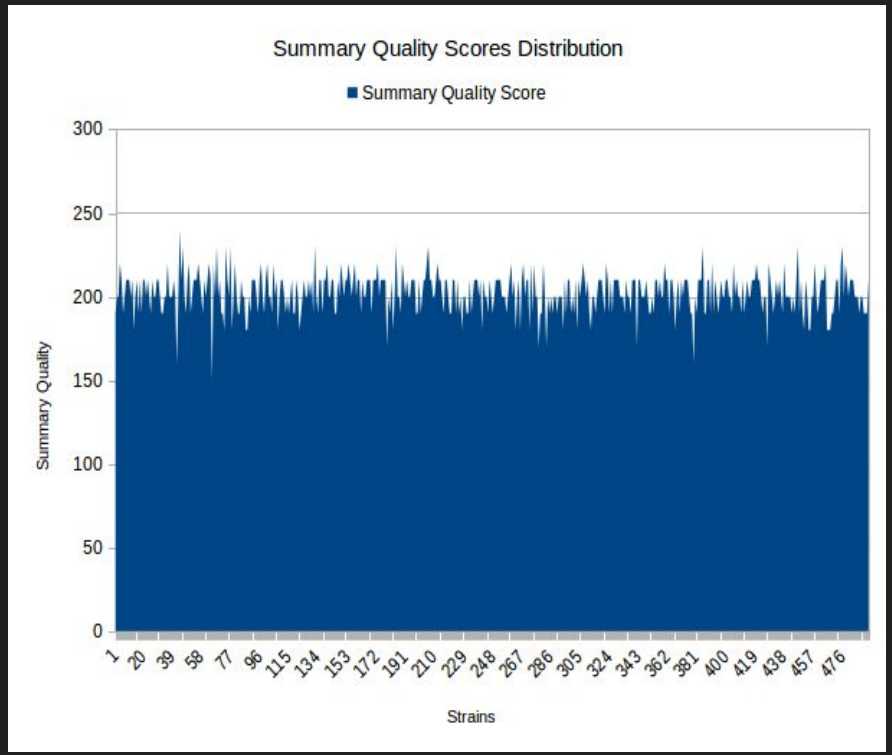




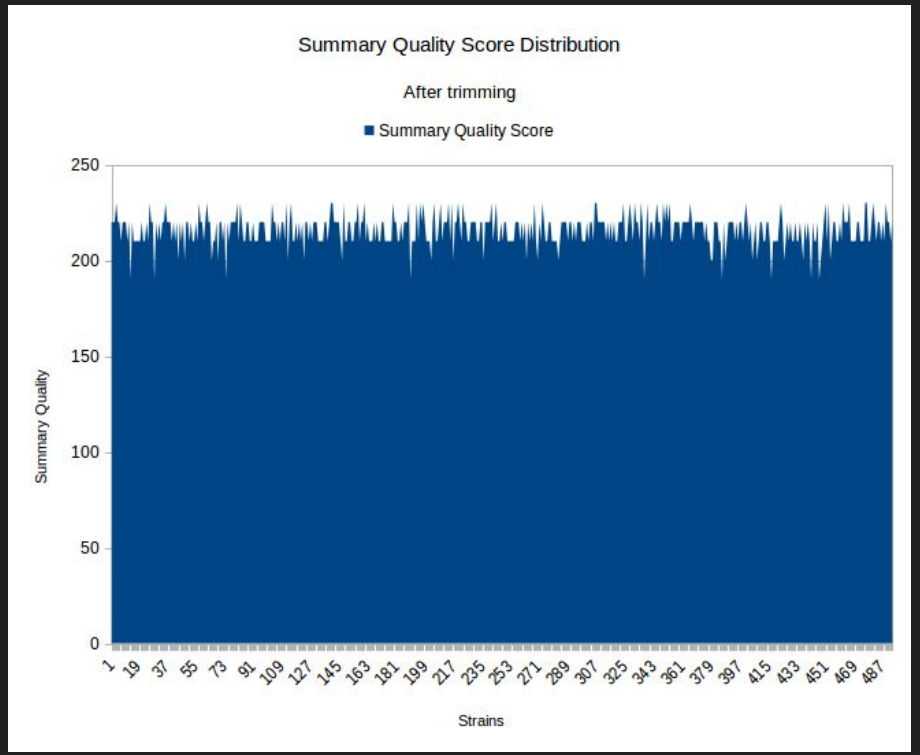
Визуализация результатов оценки качества ридов до обработки с помощью trimmomatic-0.36

** Качество рида определяется из суммы всех **pass (20)**, **warn(10)** и **fail(0)**, всего 12 параметров(**максимальное** качество - **240**)*

Визуализация результатов обработки ридов



До обработки

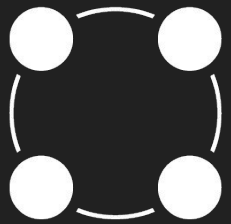


После обработки

Молчанов Владимир (Часть 2)

3. Поиск мутаций (BWA - GATK) > ***получены vcf файлы***

4. “Аннотация мутаций” vcf файлов (фильтр по качеству, пересечение с ref-gff) > ***получены vcf с информацией о генах в которых были обнаружены мутации***

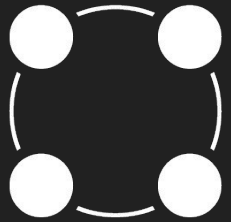


Молчанов Владимир (Часть 2)

5. Анализ литературных данных о связи мутаций в определенных генах с антибиотикорезистентностью > создан каталог мутаций, ассоциированных с резистентностью (лекарство - ген - мутация)

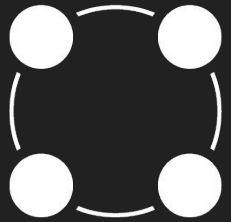
6. Разработка скрипта для поиска мутаций, связанных с антибиотикорезистентностью, в vcf файлах

7. Поиск и аннотирование мутаций (прогнозирование лекарственной устойчивости)



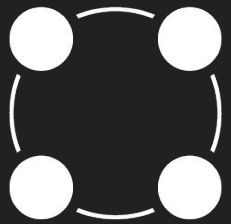
Что планировалось сделать:

- Пополнение базы данных
- Каталог известных мутаций, ассоциированных с резистентностью
- Прогноз для штаммов без данных по устойчивости
- Филогенетический анализ
- Выявление кладо-специфичных мутаций на основе NGS данных



Что сделано:

- Пополнение базы данных
- Каталог известных мутаций, ассоциированных с резистентностью
- Филогенетический анализ (999 штаммов)
- Прогноз для штаммов без данных по устойчивости
- Выявление кладо-специфичных мутаций на основе NGS данных



Литература

1. Hwang, Sohyun, et al. "Systematic comparison of variant calling pipelines using gold standard personal exome variants." Scientific reports 5 (2015).
2. Cornish, Adam, and Chittibabu Guda. "A comparison of variant calling pipelines using genome in a bottle as a reference." BioMed research international 2015 (2015).
3. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics 30.15 (2014): 2114-2120.
4. Glynn, Judith R., et al. "Whole genome sequencing shows a low proportion of tuberculosis disease is attributable to known close contacts in rural Malawi." PLoS One 10.7 (2015): e0132840.
5. Comas, Iñaki, et al. "Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans." Nature genetics 45.10 (2013): 1176-1182.
6. Cole, STea, et al. "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence." Nature 393.6685 (1998): 537-544.
7. Casali, Nicola, et al. "Microevolution of extensively drug-resistant tuberculosis in Russia." Genome research 22.4 (2012): 735-745.
8. Gagneux, Sebastien, et al. "Variable host–pathogen compatibility in Mycobacterium tuberculosis." Proceedings of the National academy of Sciences of the United States of America 103.8 (2006): 2869-2873.
9. Deniz Yorukoglu, Yun William Yu, Jian Peng & Bonnie Berger“, Compressive mapping for next-generation sequencing”, Nature Biotechnology(2016)
10. Adam Cornish, Chittibabu Guda, “A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference(2015)