Домашнее задание 3
Построение пайплайна получения генетических вариантов

Юрий Викторович Вяткин

E-mail: vyatkin@gmail.com

Факультет информационных технологий
Новосибирский государственный университет
Весенний семестр 2023

Домашнее задание 3

- Решение необходимо оформить в репозитории на github.com и выслать преподавателю на почту vyatkin@gmail.com, с обязательным указанием ФИО слушателя, ссылку на репозиторий

- Срок выполнения задания – 3 недели (25.05.23)

- Наличие выполненного задания и срок сдачи влияет на допуск к экзамену и оценку

Домашнее задание 3

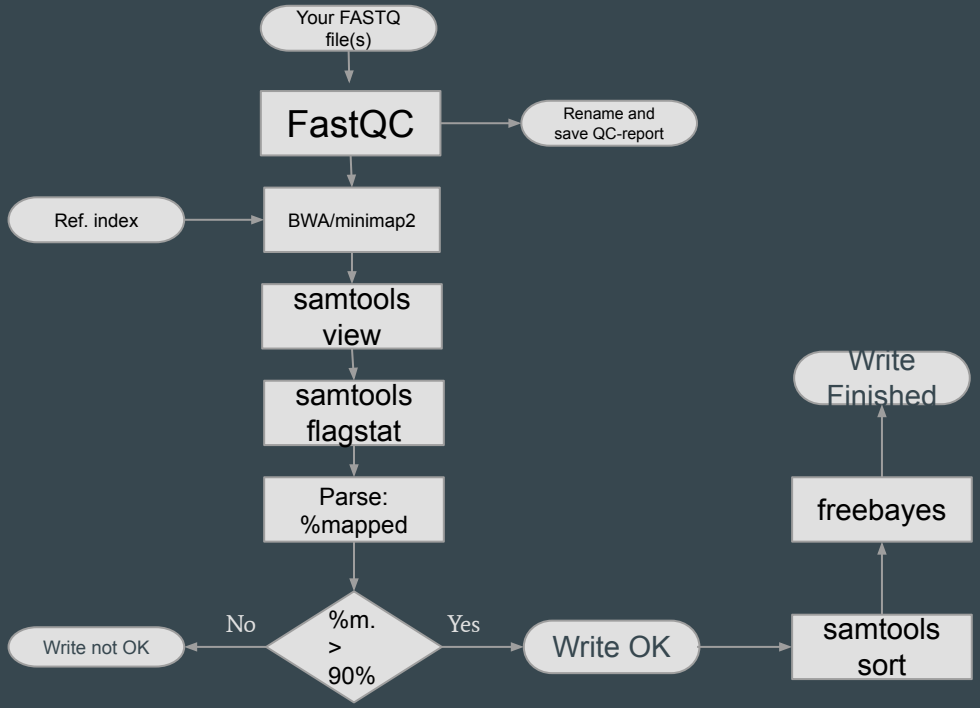
1. Найти Linux, вспомнить bash, завести репозиторий на github
2. Найти на NCBI SRA и скачать результат секвенирования (набор ридов) *Escherichia coli* (*e.coli*) ИЛИ *Homo sapiens (WES/WXS - whole exome sequencing (2-20Gb), WGS - whole genome sequencing (большой файл!))*
3. Скачать референсный геном *e.coli* https://www.ncbi.nlm.nih.gov/assembly/GCF_000005845.2/ или Homo sapiens https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz
4. Скачать и установить (скомпилировать или бинарный файл) консольные версии программ: FastQC, bwa/minimap2, samtools
5. Изучить простой запуск этих программ (см. Getting started, Quick start и тд.)
6. Индексировать референсный геном соответствующим инструментом
7. Написать скрипт (bash/Python) разбора результатов samtools flagstat для получения % картированных ридов
8. Реализовать “алгоритм оценки качества картирования” на bash со всеми элементами (см. ниже), в том числе вывод сообщения вида “OK/not OK”
9. Найти, скачать и установить (развернуть) фреймворк создания пайплайнов
10. Написать короткую инструкцию по скачиванию и установке фреймворка
11. Изучить базовые возможности фреймворка (см. Tutorials, youtube и тд.), написать тест “Hello world”
12. Реализовать пайплайн оценки качества картирования на фреймворке
13. Визуализировать полученный пайплайн автоматическими инструментами фреймворка
14. Описать использованный способ визуализации и отличия полученного DAG от блок-схемы алгоритма

Домашнее задание 3

Необходимо получить и выложить в репозиторий результаты:

1. Ссылку на загруженные прочтения из NCBI SRA
2. Скрипт на bash с реализованным алгоритмом
3. Результат команды samtools flagstat
4. Скрипт разбора файлов с этими результатами
5. *Опционально файлы FASTQ, SAM/BAM, VCF в архивах (большой размер!)
6. Инструкцию по развертыванию и установке фреймворка
7. Код любого тестового пайплайна (“Hello world”) на фреймворке
8. Результаты работы пайплайна на фреймворке и лог-файлы
9. *Опционально описание использованных инструментов для визуального создания пайплайнов (скриншоты)
10. Код пайплайна “оценки качества картирования” на фреймворке
11. Выведенные результаты работы пайплайна на загруженных данных в отдельном файле
12. Лог-файлы работы пайплайна на загруженных данных
13. Визуализацию пайплайна в виде графического файла
14. Описание использованного способа визуализации и отличия полученной визуализации от блок-схемы алгоритма в свободной форме

Алгоритм получения генетических вариантов



Burrows-Wheeler Aligner

BWA (bwa mem)
https://github.com/lh3/bwa

Minimap2

https://github.com/lh3/minimap2

Индексирование референса

Вход: hg38.fa

Выход: hg38.fa.*
(hg38.mmi)

Картирование

Вход: hg38.fa.*,
sample_1.fastq(.gz),
sample_2.fastq(.gz)

Выход: sample.sam

Конвертация форматов SAM/BAM

Samtools

https://github.com/samtools/samtools

samtools view

SAM->BAM

Вход: sample.sam

Выход: sample.bam, sample.bai

BAM->SAM

Вход: sample.bam

Выход: sample.sam

Оценка SAM/BAM

Samtools

https://github.com/samtools/samtools

samtools flagstat

Вход: sample.bam

Выход: sample.txt

Оценка SAM/BAM

samtools flagstat:

```
1099585 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
159 + 0 supplementary
183658 + 0 duplicates
1097662 + 0 mapped (99.83% : N/A)
1099426 + 0 paired in sequencing
549713 + 0 read1
549713 + 0 read2
1091988 + 0 properly paired (99.32% : N/A)
1095974 + 0 with itself and mate mapped
1529 + 0 singletons (0.14% : N/A)
3566 + 0 with mate mapped to a different chr
2892 + 0 with mate mapped to a different chr (mapQ>=5)
```

Оценка SAM/BAM

```
1097662 + 0 mapped (99.83% : N/A)
```

> 90%

OK!

< 90%

not OK...

Сортировка BAM

Samtools

https://github.com/samtools/samtools

samtools sort

Вход: sample.bam

Выход: sample.sorted.bam

Коллинг генетических вариантов FreeBayes

FreeBayes
https://github.com/freebayes/freebayes

Вход: hg38.fa,
sample.sorted.bam

Выход: sample.vcf