Секвенирование с референсным геномом – домашнее задание.

В папке http://makarich.fbb.msu.ru/enabieva/ngs2017_2019/ лежат парные чтения в файлах reads_1.fastq.gz и reads_2.fastq.gz, а также референсный геном 13-й хромосомы (chr13.fa) из сборки референсного генома hg19 и сделанные из него файлы для индекса bwa и “sequence dictionary”.

Ваша задача:

- Выровнять («закартировать») эти чтения на референсный геном 13-й хромосомы.
    - отсортировать (`samtools sort`) и затем проиндексировать (`samtools index`) бам-файл с выравниванием
- дедуплицировать полученный файл (`picard MarkDuplicates`)
    - снова проиндексировать результат
- найти в дедуплицированном файле варианты, используя `GATK HaplotypeCaller` и выписывая файл, полученный в результате пересборки гаплотипов.
- Проаннотировать полученные варианты с помощью программы `snpEff`
- Установить у себя IGV и посмотреть на нем на некоторые из найденных вариантов

Подробности:

Картирование:

Для совместимости с последующими программами bwa нужно дать флажок `-M` и добавить read groups (`-R …`). Также имеет смысл сразу перенаправить результат на `samtools view -bS`, чтобы не писать sam-файл на диск, а сразу перевести его в bam, например:

```
bwa mem -M -R \
"@RG\tID:ERR232255_chr3\tLIB:ERR232255\tPL:ILLUMINA\tSM:ERR232255" chr13.fa \
reads_1.fastq.gz reads_2.fastq.gz  | samtools view -bS - > bwamem.chr13.bam
```

При variant calling ограничьтесь геном BRCA2, расположенном в интервале chr13:32889617-32973809. Для этого используйте опцию `-L chr13:32889617-32973809`. Также используйте опцию `-bamout reassembledfilename.bam`, чтобы записать результат пересборки HaplotypeCaller-ом в этот файл. Можно также использовать `-stand_call_conf 30` (поднимает порог для вариантов).

В IGV откройте полученный файл .vcf, а также bam-файл, используемый при variant calling (результат дедупликации), и файл с результатом пересборки HaplotypeCaller-ом (заданный опцией -bamout).

Проаннотируйте полученные варианты с помощью snpEff (или анаолгичной программой). Найдите потенциально функциональные варианты (аннотированные как “HIGH” или “MODERATE”. Используйте grep.

Посмотрите на эти инделы в IGV. Видите ли вы различия между двумя bam-файлами? (нужно смотреть одновременно на оба). Также посмотрите на несинонимичные однонуклеотидные замены. Для каждого изученного вами варианта делайте скриншоты IGV.

Нужные программы уже стоят на кластере.

Программы на java:

picard: (для MarkDuplicates)
`/usr/lib/jvm/java-8-openjdk-amd64/bin/java -jar /mnt/local/vse/shared/picard.jar`
GATK: `/usr/lib/jvm/java-8-openjdk-amd64/bin/java -jar /mnt/local/vse/shared/GenomeAnalysisTK-3.8-1/GenomeAnalysisTK.jar`
snpEff: `/usr/lib/jvm/java-8-openjdk-amd64/bin/java –jar /mnt/local/vse/shared/snpEff.jar`

Документация picard:

https://broadinstitute.github.io/picard/command-line-overview.html

Документация HaplotypeCaller (для GATK 3):

https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php

Пример запуска snpEff:

```
/usr/lib/jvm/java-8-openjdk-amd64/bin/java -jar /mnt/local/vse/shared/snpEff/snpEff.jar ann -c /mnt/local/vse/shared/snpEff/snpEff.config  hg19 my.vcf > my.snpeff.vcf
```

(одной строкой)

Что сдавать:

1. % закартированных чтений и % дуплицированных чтений.
2. Список мутаций в гене BRCA2, аннотированных snpEff как имеющих эффект “HIGH” или “MODERATE", и скриншоты IGV для них.

По желанию и на дополнительные баллы можно добавить шаги по «жесткой» фильтрации вариантов (см. https://software.broadinstitute.org/gatk/documentation/article.php?id=2806), а также перевыравниванию вокруг инделов или рекалибрации качества оснований и посмотреть, как это влияет на результат. Для последних двух шагов вам понадобятся популяционные данные – напишите мне, и я их пришлю (или скажу, как получить). Также можно посмотреть, что получится, если использовать альтернативные программы на разных этапах.

По умолчанию ответы присылайте на enabieva@gmail.com.