Домашнее задание – Практикум 11

Общая задача: поиск и аннотация вариантов одного человека по данным экзомного секвенирования на примере одной хромосомы

Задача практикума: подготовить необходимые файлы (парно-концевые прочтения и последовательность референсного генома), изучить качество предложенных чтений и проиндексировать референс.

1. Подготовка референса

Для себя лично и возможности восстановить картину происходящего в будущем настоятельно рекомендую сохранить команды подготовки референса и проверки качества чтений. Включать этап подготовки файлов и проверки качества чтений в программный сценарий (см. ниже) не нужно!!!

1) Получение референса

Создайте директорию для последовательности генома и индекса к программе для картирования с помощью hisat2, скопируйте в нее файл с вашей хромосомой. Теперь это ваш референс.

Индексация для hisat2

Проиндексируйте референсный геном (это ваша хромосома).

Мануал к hisat2 - http://daehwankimlab.github.io/hisat2/manual/

Команда для индексирования референса для последующего картирования: **hisat2-build chrN.fa prefix**

При индексации референса можно использовать данные генной разметки, особенно при анализе RNA-seq, но сейчас мы этого делать не будем.

В данном случае prefix - префикс, с которого будут начинаться индексные файлы (должно получиться 8 файлов .ht2).

2) Индексация samtools

Многие программы перед работой с большими файлами требуют предварительной индексации согласно своим алгоритмам.

Одной из таких программ является **samtools**.

Индексирование: **samtools faidx chrN.fa**

На выходе должен получиться файл **chrN.fa.fai**

Описание можно прочитать тут: https://manpages.ubuntu.com/manpages/bionic/man5/faidx.5.html

Из полученного **chrN.fa.fai** узнайте точное имя своей хромосомы и длину вашей хромосомы в нуклеотидах.

(*) Объясните каждую цифру из файла, полученного после индексирования референса с помощью samtools. В описании обратите внимание на раздел с примером, там все подробно описано.

2. Чтения ДНК

1) Описание образца

Найдите ID вашего образца в базе NCBI (https://www.ncbi.nlm.nih.gov/) в разделе SRA.

Укажите:

a) SRR ID образца ДНК-чтений (см. ведомость)
b) ссылку на информацию об образце из NCBI
c) прибор для секвенирования
d) организм
e) стратегию секвенирования (полногеномное, экзомное, таргетная панель)
f) парноконцевые или одноконцевые чтения
g) сколько чтений ожидается (spots)

2) **Проверка качества исходных чтений**

Проанализируйте качество исходных чтений с помощью программы **fastqc**, исследуйте оба (!!!) файла (помните, что у вас парно-концевые чтения). Мануал: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Запустить **fastqc** можно с помощью команды **fastqc file.fastq.gz**

Программа умеет работать с архивированными файлами.

Не удаляйте получившиеся файлы (.html), они могут пригодиться на зачете. Положите их туда, где можно быстро взять.

Укажите:

a) какое количество пар чтений получилось
b) совпадает ли количество чтений у “прямых” чтений и “обратных” чтений
c) краткий комментарий качества пар чтений по результатам fastqc (картинка Per base sequence quality - 2 штуки(!!))
d) краткий комментарий о длине ваших чтений по результатам fastqc (картинка Sequence Length Distribution - 2 штуки(!!))
e) (*) краткий комментарий о любых других результатах fastqc. Помните, что на странице программы есть примеры для плохих и хороших чтений, а также подробно объяснено, что выдает программа

3) **Фильтрация чтений**

Вне зависимости от качества исходных чтений, проведите их фильтрацию с помощью **trimmomatic**.

Мануал: http://www.usadellab.org/cms/?page=trimmomatic

Обратите внимание, что на вход вы подаете 2 файла с чтениями, а на выходе получаете 4 файла (2 paired и 2 unpaired).

Программа запускается так: **TrimmomaticPE** ИЛИ **TrimmomaticSE**

Помните, что у вас парноконцевые чтения и -phred33.

Удалите с КОНЦА чтений нуклеотиды с качеством ниже 20 (без прохода окном!), оставьте только такие чтения, длина которых не ниже 50 нуклеотидов за одну команду. Параметры, которые не нужны для выполнения этих манипуляций, использовать не нужно.

Подумайте о том, почему после работы trimmomatic получается именно 4 файла и что содержится в каждом из них.

4) **Проверка качества триммированных чтений**

Проанализируйте качество чтений после обработки программой Trimmomatic (4 файла!!!) с помощью программы fastQC.

Укажите:

a) какое количество пар(!) чтений осталось (paired) в штуках
b) какой процент пар(!) чтений остался (paired) (процент от исходного количества пар чтений)
c) краткий комментарий о сравнении качества чтений после(!!) триммирования: paired vs unpaired
d) краткий комментарий о сравнении качества чтений до и после триммирования (только paired)
e) как изменилась длина чтений после триммирования?

5) **(*) Сводный отчет о качестве чтений**

Вы проанализировали качество чтений (исходных и после улучшения качества).

Для того, чтобы понять, что у вас все прошло хорошо, вам нужно было просмотреть несколько html файлов. И это только для одного образца.

Попробуйте разобраться с программой **multiqc**

https://multiqc.info/

С помощью этой программы можно собирать отчеты после работы разных программ, в том числе и fastqc.

Воспользуйтесь этой программой, чтобы собрать отчеты о качестве всех fastq файлах, которые были вами использованы при выполнении предыдущих пунктов задания.