Домашнее задание – Практикум 11

Часть 1

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, проиндексировать референс.

1. Подготовка референса

Для себя лично и возможности восстановить картину происходящего в будущем настоятельно рекомендую сохранить команды подготовки референса. Включать этап подготовки референса 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**.

Индексировать мы будем референсный геном.

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 40 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.

Теперь мы имеем все необходимые входные файлы и перейдем к второй части практикума.

Часть 2

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

Задача практикума: картировать чтения хорошего качества на референсный геном и отобрать только такие чтения, которые удалось картировать в корректных парах

1. Картирование чтений на референсный геном

Используйте чтения хорошего качества, полученные после триммирования, только парные. Выполните картирование чтений с помощью программы hisat2 (мануал), не разрешая картирование с разрывами. Помните, что у вас парные чтения, а в качестве выходного файла создайте файл с расширением **.sam**. Воспользуйтесь инструкцией, не забудьте сохранить логи.

Вам потребуются следующие параметры:

1) -x

2) -1

3) -2

4) -p

5) (!) параметр, запрещающий возможность сплайсинга - найдите его самостоятельно в инструкции

2. Конвертация sam в bam

1) Описание sam/bam файла

a) Сколько весит sam файл в Гб?

Sam файл очень тяжелый, переконвертируйте его в сортированный bam файл и удалите(!) sam, больше он не пригодится. Для конвертации и сортировки используйте команду:

samtools sort -o file.bam file.sam

b) Сколько весит bam файл в Гб?

2) **Проиндексируйте получившийся bam файл**

Используйте команду **samtools index file.bam**

3. **Анализ bam файла**

Заглянуть в bam файл просто так не получится, он бинарный.

Проанализируйте bam файл с помощью возможностей программы samtools.

samtools flagstat file.bam, результат запишите в текстовый файл.

Изучите полученный файл и приведите его полностью в отчете.

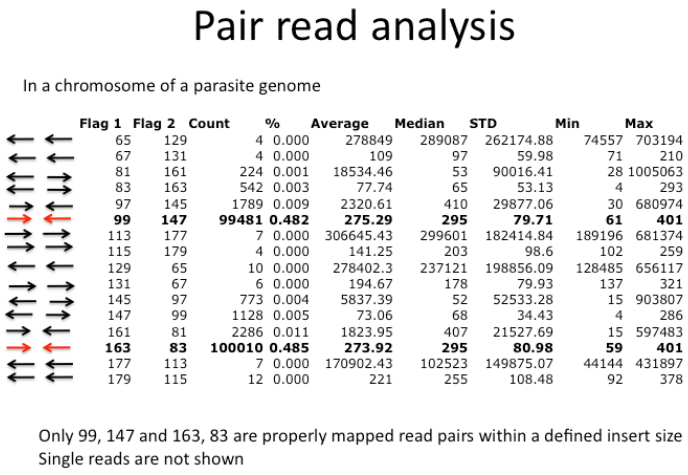
Мануал в помощь.

Ответьте на следующие вопросы:

1) Что значит число в поле «in total»?
2) Сколько чтений (не пар!) поступило на картирование?
3) Сколько чтений картировано на референс в корректных парах в штуках?
4) Сколько чтений картировано на референс в корректных парах в процентах относительно потупивших на картирование?

Не забудьте подробно описать, откуда вам удалось взять эти числа.

Помните, что у вас парные чтения, т.е. это сиквенс одного не очень большого фрагмента ДНК. Мы ожидаем, что чтения из одной пары должны картироваться недалеко друг от друга и быть направлены друг к другу. Пример возможных расположений чтений, красным отмечены варианты корректно картированных пар чтений:



Также чтения могут быть картированы на геном не один раз.

Еще несколько полезных ссылок: раз, два, три, четыре

Заглянуть в bam файл вы можете без конвертации его обратно в sam командой: **samtools view file.bam** (используйте | head или | less)

4. Получение чтений, картированных на вашу хромосому

Вам предоставлены чтения полного экзома, но картируем мы только на одну хромосому, т.е. ожидается много некартированных чтений.

Получим чтения, картированные только на вашу хромосому.

Используйте команду **samtools view** (полезная ссылка). Выходной файл должен быть в формате bam, samtools умеет принимать сразу несколько параметров. Посмотрите, как называется ваша хромосома. Узнать это можно, применив к файлу с хромосомой команду **samtools faidx** (уже делали в практикуме 11).

samtools view -h -bS file.bam chrName > file.chr.bam

Объясните каждый параметр.

Примените в полученному bam файлу уже знакомую нам команду samtools flagstat. Результат также приведите в явном виде в отчете.

Чем этот файл отличается от аналогичного файла из п.3?

5. Получение только правильно картированных пар чтений

Воспользуйтесь командой **samtools view -f 2 -bS file.bam**

Что указано в качестве значений для параметра -f?

К полученному bam файлу, содержащему только правильно картированные на вашу хромосому пары чтений, примените уже знакомую команду samtools flagstat, сохранив выход в отдельный файл. Результат приведите в явном виде в отчете.

Изучите полученный файл.

Чем этот файл отличается от аналогичного файла из п.4?

Проиндексируйте полученный bam файл, содержащий только правильно спаренные картированные чтения нужной вам хромосомы.

Далее работайте только с этим bam файлом и его индексами. Из него мы будем добывать варианты!!!

6. **Получение чтений, картированных только в границы экзома**

Возьмите bam файл, полученный в п.5., т.е. с чтениями, картированными на референс только на вашу хромосому и в правильных парах.

Оставьте только такие чтения, которые картировались в пределах экзома.

Файл с координатами экзома:

/mnt/scratch/NGS/DATA/genes/seqcap_hg38.bed

Воспользуйтесь средствами bedtools intersect. Обратите внимание, что в конце есть инструкция для пересечения bam файла с bed.

В итоге должен получиться bam файл.

Примените к нему samtools flagstat, приведите результат в явном виде и опишите.

7. **(*) Получение чтений, картированных в границы расширенного экзома**

   Повторите пункт 6, но в качестве разметки экзома возьмите расширенный файл:
   /mnt/scratch/NGS/DATA/genes/seqcap_hg38_50.bed

   Что изменилось?