Домашнее задание – Практикум 12

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:

Pair read analysis

In a chromosome of a parasite genome

	Flag 1	Flag 2	Count	%	Average	Median	STD	Min	Max
← ←	65	129	4	0.000	278849	289087	262174.88	74557	703194
← ←	67	131	4	0.000	109	97	59.98	71	210
← →	81	161	224	0.001	18534.46	53	90016.41	28	1005063
← →	83	163	542	0.003	77.74	65	53.13	4	293
→ ←	97	145	1789	0.009	2320.61	410	29877.06	30	680974
→ ←	**99**	**147**	**99481**	**0.482**	**275.29**	**295**	**79.71**	**61**	**401**
→ →	113	177	7	0.000	306645.43	299601	182414.84	189196	681374
→ →	115	179	4	0.000	141.25	203	98.6	102	259
← ←	129	65	10	0.000	278402.3	237121	198856.09	128485	656117
→ →	131	67	6	0.000	194.67	178	79.93	137	321
← →	145	97	773	0.004	5837.39	52	52533.28	15	903807
← →	147	99	1128	0.005	73.06	68	34.43	4	286
→ ←	161	81	2286	0.011	1823.95	407	21527.69	15	597483
→ ←	**163**	**83**	**100010**	**0.485**	**273.92**	**295**	**80.98**	**59**	**401**
← ←	177	113	7	0.000	170902.43	102523	149875.07	44144	431897
← ←	179	115	12	0.000	221	255	108.48	92	378

Only 99, 147 and 163, 83 are properly mapped read pairs within a defined insert size
Single reads are not shown

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?