Файлы .fastq с ридами Illumina, полученные в результате ChIP-seq эксперимента, разделены на отдельные файлы, соответствующие участкам хромосом. Они лежат в директории /srv/databases/ngs/chipseq\_y14 на диске P. Соответствие между студентами и файлами в приложенной таблице.

Задание (Автор — Данила Бредихин)

Сделайте контроль качества прочтений с помощью программы FastQC. С ней вы знакомились в прошлом семестре: https://kodomo.fbb.msu.ru/wiki/2015/3/task13. Опишите получаемые на этом шаге результаты.

1. Если это необходимо (обоснуйте), отфильтруйте данные. Для этого можно использовать программу Trimmomatic. Вам предложены одиночные прочтения. Используйте соответствующие параметры для программы Trimmomatic.
2. Картирование прочтений на геном человека hg19 выполните с помощью программы BWA.   
   Вы можете воспользоваться проиндексированным геномом человека (его можно получить самостоятельно, выполнив команду bwa index для последовательности генома человека): он доступен по адресу /srv/databases/ngs/hg19.

$ bwa mem path/to/hg19/GRCh37.p13.genome.fa chipseq\_chunkX.fastq > chipseq\_chunkX.sam

1. Опишите полученные результаты, в т. ч. укажите сколько прочтений откартировалось на геном и сколько прочтений было изначально. Предположите, прочтения с какой хромосомы были предложены вам для анализа.   
   При ответе на вопросы выше могут пригодиться следующие команды:  
   $ samtools view -bSo chipseq\_chunkX.bam chipseq\_chunkX.sam  
   $ samtools sort chipseq\_chunkX.bam -T chip\_temp -o chipseq\_chunkX.sorted.bam  
   $ samtools index chipseq\_chunkX.sorted.bam  
   $ samtools idxstats chipseq\_chunkX.sorted.bam > chipseq\_chunkX.idxstats  
   $ samtools view -c chipseq\_chunkX.sorted.bam  
   При исполнении команд командной строки, пожалуйста, указывайте в протоколе, что именно они выполняют.
2. Для поиска пиков (обычно это называют *peak calling*) воспользуйтесь программой MACS (она установлена на kodomo): <http://liulab.dfci.harvard.edu/MACS/>.  
   В простом случае её запуск выглядит так:  
   macs2 callpeak -t chipseq\_chunkX.sorted.bam  
   Если пиков слишком мало, то программа сообщит вам об этом. Тогда можно запустить ее с другими параметрами:  
   macs2 callpeak -t chipseq\_chunkX.sorted.bam --nomodel  
   Самостоятельно (macs2 callpeak -h) научитесь использовать параметр для изменения названия эксперимента, в соответствии с которым называются выходные файлы.  
   Итогом работы программы являются три выходных файла:
   * \*\_peaks.narrowPeak
   * \*\_peaks.xls
   * \*\_summits.bed

Все файлы содержат информацию о найденных пиках и в зависимости от того или иного формата более удобны для последующего анализа.  
Наиболее полная информация представлена в файле \*\_peaks.xls.

1. Опишите полученные результаты, в т. ч. укажите, сколько пиков найдено программой. Визуализируйте информацию (например, из файла \*\_peaks.narrowPeak) с помощью геномного браузера.  
   Если вы используете [UCSC Genome Browser](https://genome.ucsc.edu/), то загрузить трек можно в разделе меню My Data > Custom Tracks.  
   При этом в начало файла \*\_peaks.narrowPeak следует поместить дополнительную информацию так, как это описано по ссылке:  
   [http://genome.ucsc.edu/FAQ/FAQformat.html#format12](http://genome.ucsc.edu/FAQ/FAQformat.html" \l "format12).   
   Пример:  
   track type=narrowPeak visibility=3 db=hg19 name="my\_peaks" description="Peaks from chunk X"  
   browser position chr1:1000000-1010000  
   Можно также загрузить файл \*\_summits.bed, в котором, однако, будут отображены только «вершины» пиков.  
   Приведите соответствующий скриншот.
2. Изучите несколько пиков в более крупном масштабе в геномном браузере. (Если среди найденных пиков есть наиболее отчетливые, то рекомендуется начать рассмотрение именно с них.)  
   Ответы на следующие вопросы могут помочь при описании результатов:
   * Какова ширина пика?
   * Насколько достоверны пики? (Об этом можно судить, например, по *p-value* или *q-value*, указанным в файле \*\_peaks.narrowPeak. В восьмой и девятой колонках этого файла приведены соответственно -log10pvalue и -log10qvalue. Большее или меньшее значение этих величин соответствует более достоверным пикам?)
   * Как расположена вершина пика относительно его крайних точек? (Расстояние вершины от старта пика указана в десятой колонке файла \*\_peaks.narrowPeak .)
   * Пересекаются ли пики с функциональными элементами генома? (Если пик расположен перед геном (учитывайте направление гена!), то укажите Ensembl ID ENSG\* этого гена. Можно также указать информацию о продукте этого гена: например, *является ли ген белок-кодирующим*; если да, то *какую функцию выполняет этот белок*.)

Дополнительно

1. Конвертируйте файл \*\_\_peaks.narrowPeak в формат fasta.  
   Например:  
   /srv/databases/ngs/tools/seqtk/seqtk subseq path/to/hg19/GRCh37.p13.genome.fa my\_peaks.narrowPeak > my\_peaks.fa
2. Попытайтесь построить выравнивание последовательностей.  
   Удается ли идентифицировать консенсусную последовательность (~ 4 - 8 nts)? Почему? Почему в полноценных экспериментальных данных это (зачастую) легче сделать?   
   Вы также можете использовать сервис [MEME](http://meme-suite.org/tools/meme) (или [MEME-ChIP](http://meme-suite.org/tools/meme-chip)).