Домашнее задание №3

**Введение**

В данном практическом задании вы научитесь определять сайт связывания транскрипционного фактора на основании ChIP-Seq данных.

Для этого мы будем реконструировать результаты, полученные в следующей статье: [www.ncbi.nlm.nih.gov/pubmed/23932716](http://www.ncbi.nlm.nih.gov/pubmed/23932716).

Известно, что многие длинные некодирующие РНК (lncRNAs) специфически экспрессируются в головном мозге. В рассматриваемом исследовании идентифицировали группу lncRNAs, которые необходимы для нейрогенеза. В том числе транскрипт, ассоциированный с рабдомиосаркомой (RMST). RMST участвует в модулировании нейрогенеза, – экспрессия RMST специфична для головного мозга, регулируется репрессором транскрипции REST и увеличивается во время дифференцировки нейронов, что указывает на роль в нейрогенезе.

RMST взаимодействует с транскрипционным фактором SOX2, и они совместно регулируют группу нижестоящих генов, участвующих в нейрогенезе. Было обнаружено, что RMST необходим для связывания SOX2 с промоторными областями нейрогенных факторов транскрипции. Это указывает на участие RMST в качестве ко-регулятора транскрипции SOX2 и на его ключевую роль в регуляции дифференцировки нейральных стволовых клеток.

Каждому из вас нужно будет анализировать сайты связывания транскрипционного фактора SOX2 в пределах одной определенной хромосомы. Ниже приведен пример запуска программ для X-хромосомы – не забывайте во всех командах поменять chrX на номер вашей хромосомы.

**Обязательная часть задания (15 баллов)**

Скачиваем программы bowtie2, MACS2, Homer. Скачиваем данные из SRA.

Cоздаем директорию с названием своей хромосомы. Внутри нее следует проделывать всю дальнейшую работу:

* mkdir chrX

cd chrX

Делаем контроль качества ридов, если все хорошо, то переходим на следующий шаг, если что-то не нравится — делаем обрезку по качеству.

Делаем выравнивания и преобразуем его в .bed

Подробнее почитать про разные форматы можно на сайте GenomeBrowser: <https://genome.ucsc.edu/FAQ/FAQformat.html>

Берем из этих файлов только данные соответствующие вашей хромосоме, например:

grep '^chrX\s' ~/your\_folder/input.bed > input.chrX.bed

С помощью программы MACS2 определяем координаты мест связывания SOX2 (пики) на нашей хромосоме (с нормировкой на контрольный эксперимент). Координаты пиков будут находиться в файле SOX2.chrX.macs2\_peaks.narrowPeak:

macs2 callpeak -t SOX2.chrX.bed -c input.chrX.bed -g hs -n SOX2.chrX.macs2

Используя программу Homer и полученные пики, ищем мотив, с которым связывается SOX2 (лучше запускать внутри screen-а):

findMotifsGenome.pl SOX2.chrX.macs2\_peaks.narrowPeak ~/data/hg38\_ucsc/hg38.fa homer\_chrX

Скачиваем папку выдачи Homer-а с сервера, открываем homerResults.html и смотрим, если известный мотив был найден.

Лого мотива для SOX2 из статьи приведено ниже. Имейте в виду, что reverse complement этой последовательности также является верным мотивом (т.к. ДНК двуцепочечная):



Копируем скриншот первых 5-ти мотивов, найденных Homer-ом в файл отчета и пишем вывод (получился ли правильный мотив на вашей хромосоме или нет).

Далее необходимо проделать вышеописанную процедуру для ChIP-Seq эксперимента, выполненного на клетках, в которых была подавлена экспрессия некодирующей РНК RMST

* При поиске пиков не забывайте также нормировать на контроль

В статье авторы показали, что некодирующая РНК RMST связывается с SOX2 и каким-то образом определяет специфичность его связывания. Это утверждение доказывалось тем, что в отсутствии RMST мотив связывания белка SOX2 кардинально менялся.

Добавить в отчет топ-5 получившихся мотивов, а также написать в отчете подтверждается ли вывод авторов, если мы рассматриваем не весь геном, а только одну хромосому.

**Необязательная часть задания (10 баллов)**

Изучение дифференциального связывания.

Предлагается провести downstream analysis, который делится на два основных этапа:

* Сопоставление из набора данных ChIP определенным генам. Например, можно воспользоваться функцией makeChIPGeneDB из пакета [TFEA.Chip](https://www.bioconductor.org/packages/devel/bioc/vignettes/TFEA.ChIP/inst/doc/TFEA.ChIP.html) (эта функция ставит в соответствие пик, перекрывающий определенную регуляторную область, с определенной меткой - геном, ассоциированным, с данным регионом).
* GSEA анализ обогащенности полученного набора генов. Необходимо построить картинки, отражающие результаты анализа GSEA.

**Список файлов для сдачи**

Отчет в формате .pdf или .docx.

На сервере в папках с названиями хромосом должны быть папки с выдачей программы Homer.

**Форма отчетности**

Прислать на электронную почту файл (.pdf или .docx) с полученными результатами и выводом. Имя файла должно состоять из номера хромосомы, нижнего подчеркивания и фамилии студента. Например, chrX\_Ivanov.task3.docx

**Дедлайн по сдаче задания: 8 мая, 23:59.**