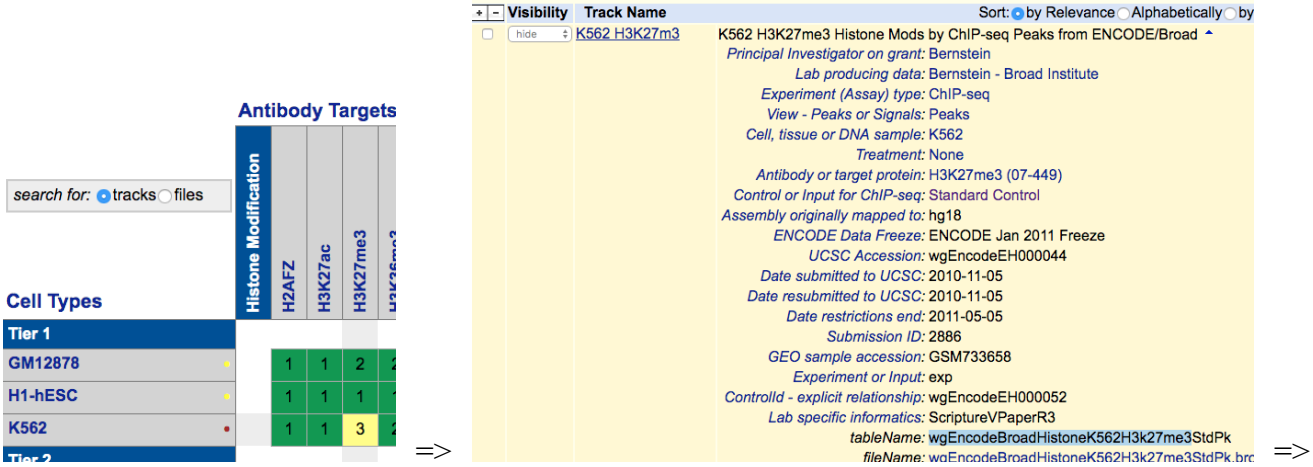
Домашнее задание №3 (ChromHMM)

Введение

Целью данного домашнего задания является разбивка (аннотация) генома человека на разные типы эпигенетических состояний. Это будет сделано на основании данных о наличии различных гистоновых модификаций (гистоновый код) в соответствующих участках генома. Для выполнения этого задания мы будем работать с чтениями, полученными в ChIP-seq экспериментах из проекта ENCODE (https://www.encodeproject.org/), которые были выровнены на геном человека (версия hg19) – т.е. у нас будут уже готовые bam-файлы. Автоматическая разбивка генома на эпигенетические типы будет осуществляться с помощью программы ChromHMM. В результате итеративной процедуры (алгоритм Баума-Велша) программа определит сочетание гистоновых меток, которые характерны для каждого из N разных эпигенетических типов (число N указывается пользователем при запуске программы). Наша задача будет заключаться в том, чтобы на основании косвенных независимых наблюдений вручную приписать каждому из N эпигенетических типов возможную биологическую функцию.
Данное задание было подготовлено с использованием руководства пользователя программы ChromHMM (http://compbio.mit.edu/ChromHMM/ChromHMM_manual.pdf), а также аналогичного задания с сайта ENCODE:
https://www.encodeproject.org/documents/d0a10470-b049-4da1-9de2-01449ddfa6a5/@@download/attachment/ChromHMM_tutorial.pdf

Обязательная часть задания (8 баллов)

- На сайте github.com создаем публичный репозиторий «hse_hw3_chromhmm» и приводим ссылку на этот репозиторий в общей гугл-таблице в лист HW3.
    - https://docs.google.com/spreadsheets/d/10r73G68KiXa-7Kf__u1Nir1cITd-3xy1NbBX7_kGk0A/edit#gid=0
    - Мы продолжаем работу данными из ENCODE желательно с той же клеточной линией, для которой вы делали ДЗ-2 (про ChIP-seq на гистоновую метку) -- повторите название этой клеточной линии и на вкладке HW3 и общей гугл таблице
- Для начала работы необходимо скачать набор bam-файлов для «своего» типа клеток с сайта UCSC Genome Browser:
    - Все доступные ChIP-seq эксперименты для всех типов клеток указаны в табличке: https://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html
    - Сами bam-файлы находятся тут: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/
    - Соответствие между меткой из матрицы экспериментов и bam-файлом можно узнать “вручную” следующим образом:



=> =>

=>

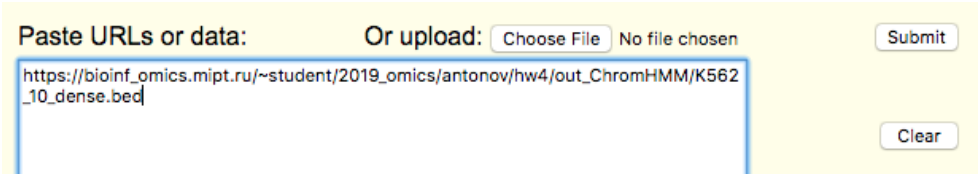
```
wgEncodeBroadHistoneK562H3k27acStdRawDataRep1.fastq.gz      26-Oct-2010 12:34  782M
wgEncodeBroadHistoneK562H3k27acStdRawDataRep2.fastq.gz      06-Jan-2009 15:27  496M
wgEncodeBroadHistoneK562H3k27acStdSig.bigWig                27-Jan-2011 21:18  292M
wgEncodeBroadHistoneK562H3k27me3StdAlnRep1.bam              29-Oct-2010 11:16  518M
wgEncodeBroadHistoneK562H3k27me3StdAlnRep1.bam.bai          05-Nov-2010 15:12  5.8M
wgEncodeBroadHistoneK562H3k27me3StdAlnRep2.bam              29-Oct-2010 12:46  473M
wgEncodeBroadHistoneK562H3k27me3StdAlnRep2.bam.bai          05-Nov-2010 15:14  5.7M
```

- Желательно скачать выравнивания для по крайней мере 10-ти разных модификаций гистонов. Одной реплики для каждой модификации будет достаточно.
- Также необходимо скачать файл с контрольным экспериментом для соответствующего типа клеток, например:

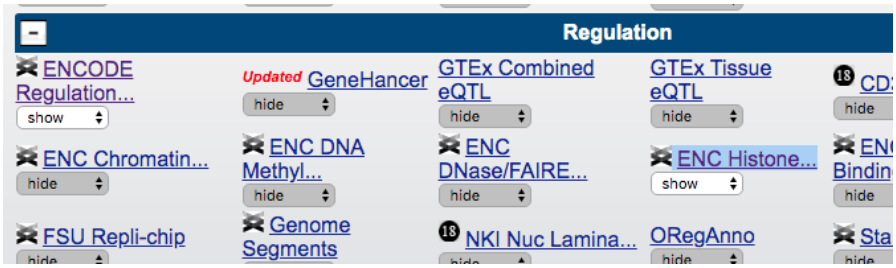
```
wgEncodeBroadHistoneK562Chd7a301223a1PK.broadPeak.gz        25-Jul-2012 20:02  488K
wgEncodeBroadHistoneK562Chd7a301223a1RawDataRep1.fastq.gz   18-Jul-2012 08:51  1.2G
wgEncodeBroadHistoneK562Chd7a301223a1Sig.bigWig             18-Jul-2012 02:20  318M
wgEncodeBroadHistoneK562ControlStdAlnRep1.bam               29-Oct-2010 13:55  759M
wgEncodeBroadHistoneK562ControlStdAlnRep1.bam.bai           05-Nov-2010 15:18  6.0M
wgEncodeBroadHistoneK562ControlStdRawDataRep1.fastq.gz      26-Oct-2010 12:50  801M
wgEncodeBroadHistoneK562ControlStdSig.bigWig                27-Jan-2011 21:00  394M
```

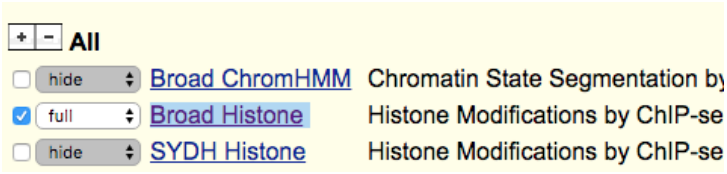
- Необходимо проанализировать все эти метки с помощью ChromHMM и получить разбивку генома по состояниям.
  - Образец Google Colab ноутбука с примерами запуска ChromHMM доступен по ссылке: https://colab.research.google.com/drive/1w7wjrlX3FM55tfiWkeTww6T5VEfFTcLG?usp=sharing

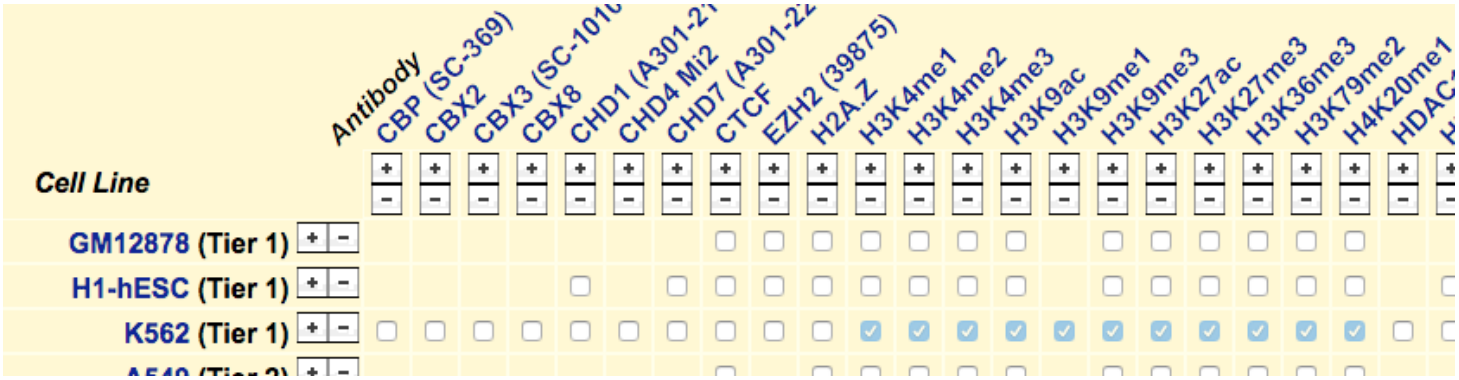
- Наша дальнейшая задача заключается в том, чтобы «вручную» присвоить каждому из полученных эпигенетеических состояний (на которые был разбит геном) имя, описывающее его возможную биологическую роль. Для этого необходимо визуализировать расположение на геноме полученных состояний:
  - Открываем UCSC GenomeBrowser (http://genome.ucsc.edu) и выбираем версию генома человека hg19. Переходим в раздел My Data => Custom Tracks и загружаем один из .bed файлов (*_dense.bed или *_expanded.bed) созданных программой ChromHMM.

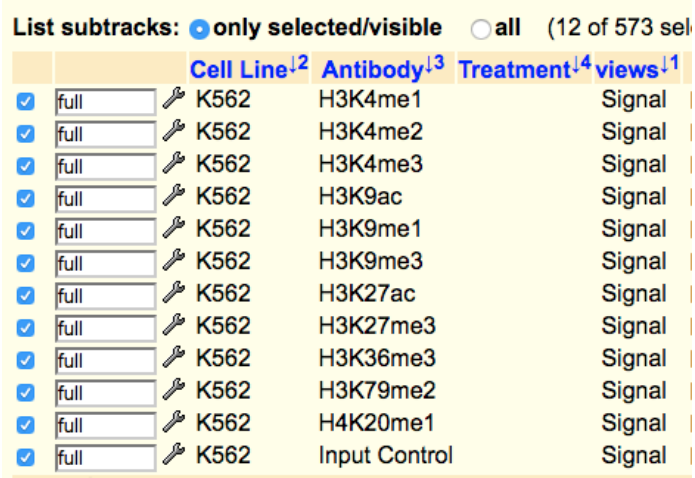


- Также можно добавить панели с профилями гистоновых модификаций, на основании которых и были определены эпигенетические пики. Не забывайте выбрать именно тот тип клеток, с которым вы работали:

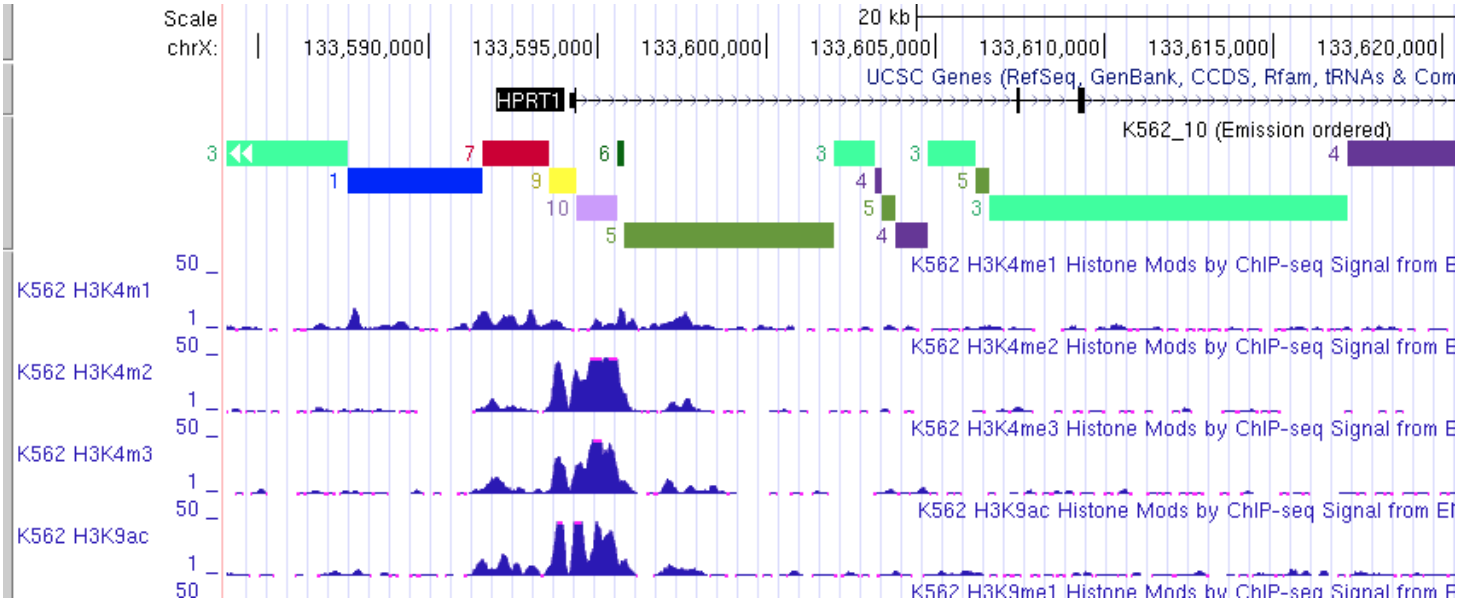




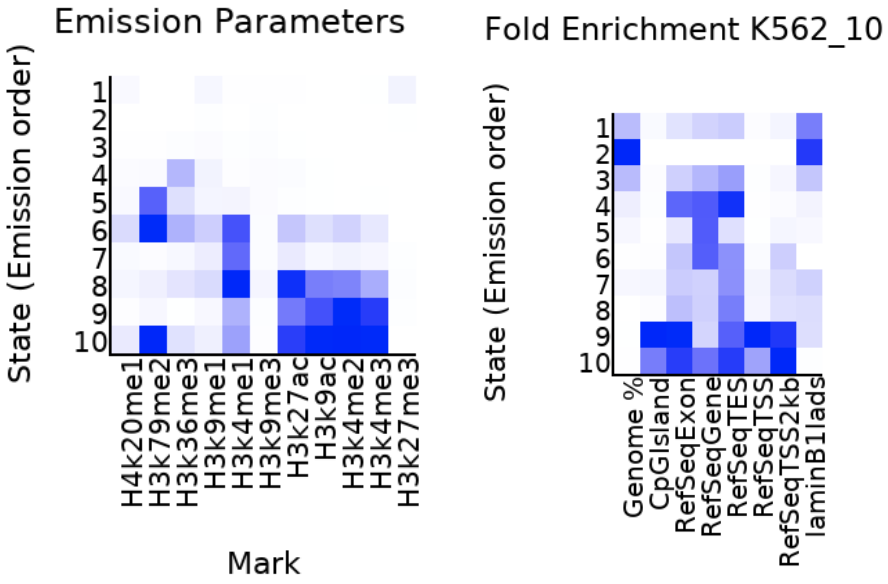




- Это позволит сразу увидеть и разбивку ChromHMM, и соответствующие профили гистоновых меток, например:



- Дальше смотрим, где на геноме (относительно аннотированных генов) располагаются разные состояния, а также смотрим на картинки из HTML-отчета ChromHMM о том, какие метки типичны для разных состояний:

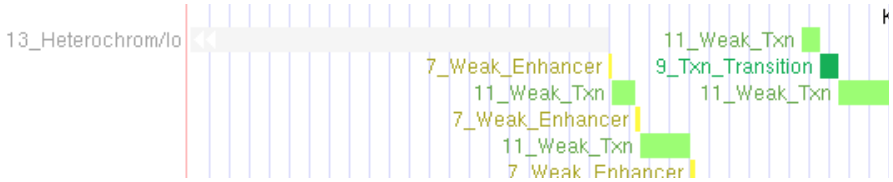


  - Для каждого из типов (state -- состояния) можно посмотреть характерные гистоновые модификации, а также его типичное расположение относительно CpG островков (часто соответствуют промоторам), аннотированных генов, доменов, ассоциированными с ядерной ламиной (lamina associated domains – обычно являются репрессированным гетерохроматином).
  - Примеры названий эпигенетических типов можно посмотреть в модели из 15-ти состояний, которая была использована в проекте ENCODE: http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=wgEncodeBroadHmm

- State 1 - Bright Red - Active Promoter
- State 2 - Light Red -Weak Promoter
- State 3 - Purple - Inactive/poised Promoter
- State 4 - Orange - Strong enhancer
- State 5 - Orange - Strong enhancer
- State 6 - Yellow - Weak/poised enhancer
- State 7 - Yellow - Weak/poised enhancer
- State 8 - Blue - Insulator
- State 9 - Dark Green - Transcriptional transition
- State 10 - Dark Green - Transcriptional elongation
- State 11 - Light Green - Weak transcribed
- State 12 - Gray - Polycomb-repressed
- State 13 - Light Gray - Heterochromatin; low signal
- State 14 - Light Gray - Repetitive/Copy Number Variation
- State 15 - Light Gray - Repetitive/Copy Number Variation

Бонусная часть задания (2 балла)

- В .bed файле, который был создан программой ChromHMM (*_dense.bed или *_expanded.bed), заменить номера эпигенетических типов (4-й столбец) на их соответствующие названия. Это позволит более удобно визуализировать полученную эпигенетическую аннотацию генома, что и нужно сделать:



Список файлов для сдачи

- В репозитории в файле README.md:
  - Список 10-ти гистоновых меток (и соотв имен файлов) , для которых был сделан анализ
  - Файл cellmarkfiletable.txt
  - Папку с выдачей ChromHMM
  - Картинки из выдачи ChromHMM
  - Табличка с номерами эпигенетических типов, их характерные эпигенетические метки и другие свойства, а также присвоенные им названия
  - Одну или несколько картинок из UCSC GenomeBrowser, показывающих различные участки генома и соответствующие эпигенетические типы (и, желательно, профили эпигенетических меток)
  - Список всех запущенных команд
  - Результат бонусного задания (если есть)

Форма отчетности

Github репозиторий, содержащий все полученные результаты.

Последний срок сдачи: 5 апреля до 23:59 (будет отслеживаться по последнему коммиту в репозиторий). Штраф -0.5 балла за каждый день просрочки.

В случае возникновения вопросов обращаться telegram: @PlainSight