

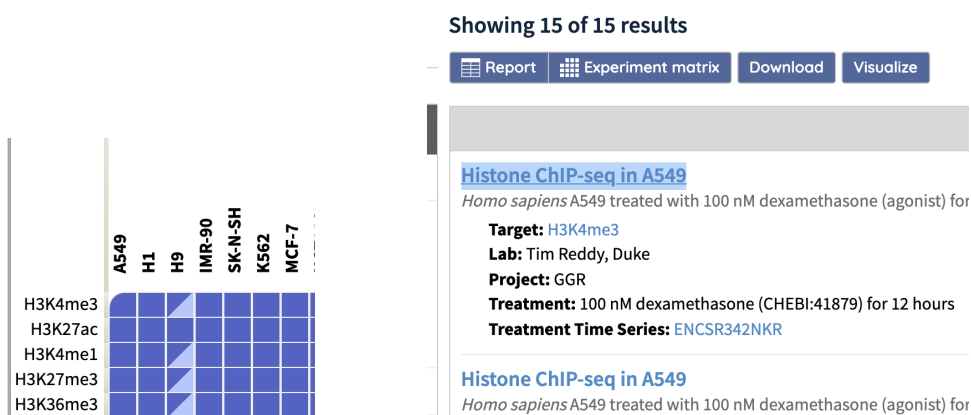
Домашнее задание №2

Введение

В данном практическом задании вы научитесь определять участки генома, где присутствует определенная гистоновая модификация в конкретном типе клеток с помощью анализа ChIP-Seq данных.

Обязательная часть задания (8 баллов)

1. На сайте github.com создаем публичный репозиторий «hse_hw2_chip» и приводим ссылку на этот репозиторий в общей гугл-таблице в лист HW2. https://docs.google.com/spreadsheets/d/10r73G68KiXa-7Kf_u1Nir1cITd-3xy1NbBX7_kGk0A/edit#gid=0
2. Для начала работы необходимо выбрать клеточную линию, гистоновую метку и файл контроля (ChIP-seq input):
 - a. Берем один из экспериментов ENCODE для одной клеточной линии человека (*homo sapiens*) и определенной гистоновой метки. Вписываем это в таблицу (см. пункт 1 -- столбцы "Клеточная линия" и "Гистоновая Метка")
https://www.encodeproject.org/chip-seq-matrix/?type=Experiment&replicates.library.biosample.donor.organism.scientific_name=Homo%20sapiens&assay_title=Histone%20ChIP-seq&assay_title=Mint-ChIP-seq&status=released



- b. В выбранном эксперименте должно быть по крайней мере 2 ChIP-seq реплики. Также копируем ID fastq файлов, соответствующих этим репликам в табличку (Реплика 1 и Реплика 2). **ВАЖНО - ID fastq файлов не должны пересекаться между студентами!**

Genome browser

Association graph

File details

GRCh38

Displaying 43 of 43 files

Raw sequencing data (3 Files)

Isogenic replicate	Library	Accession	File type	Run type	Read	Output type
1	ENCLB875ZJH	ENCFF891GXF	fastq	SE51nt		reads
2	ENCLB805CCP	ENCFF170YTR	fastq	SE51nt		reads
3	ENCLB827QLU	ENCFF020SER	fastq	SE51nt		reads

- с. Также выбираем хотя бы один fastq файл с соответствующим контролем (ID контрольного эксперимента указан в разделе Summary)

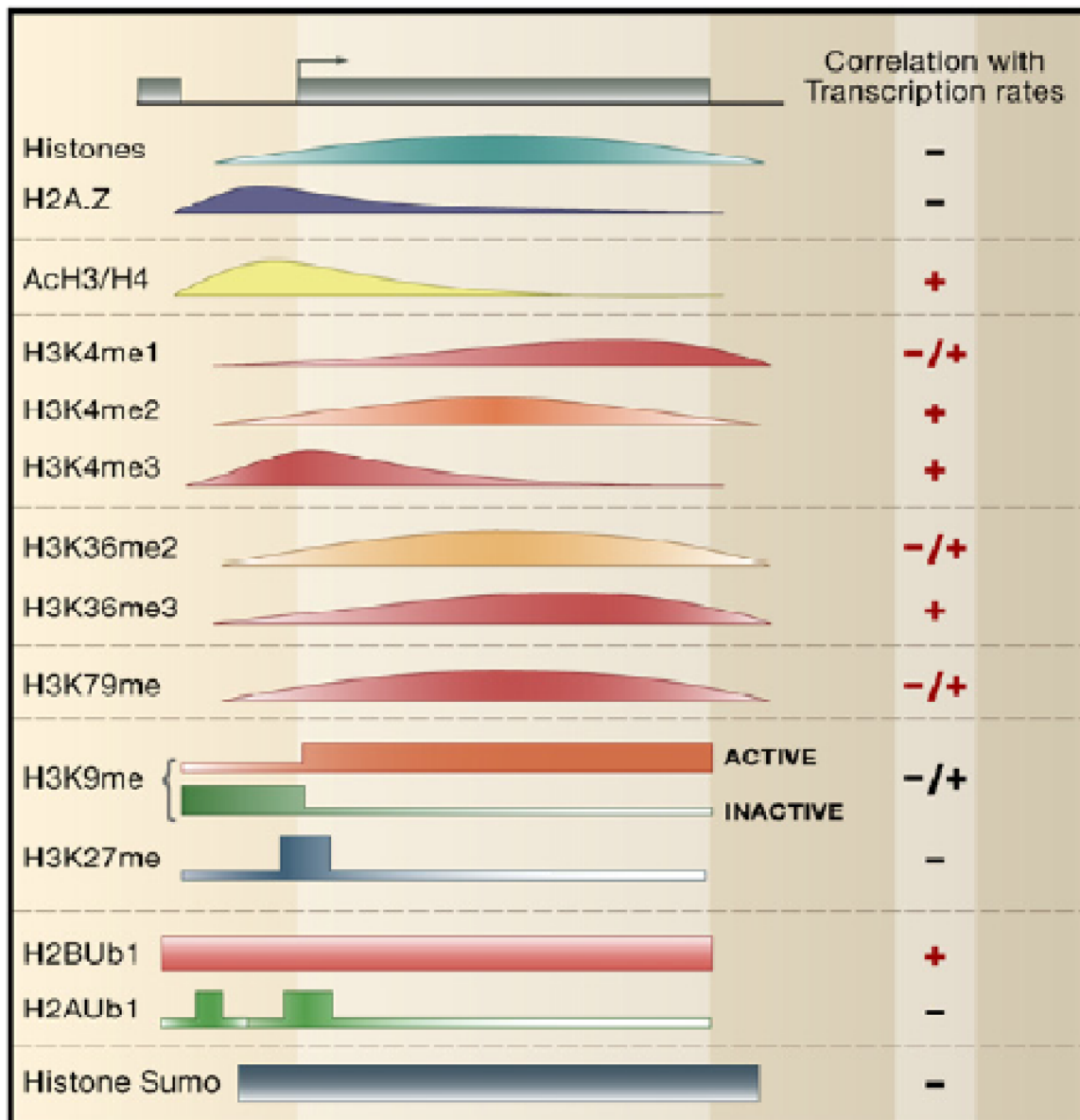
		Genome browser	Association graph	File details
		Displaying 1		
		Raw sequencing data (3 Files)		
Isogenic replicate	Library	Accession	File type	Run type
1	ENCLB309WZG	ENCFF524OKJ	fastq	SE51n
2	ENCLB984YWZ	ENCFF612KDN	fastq	SE51n
3	ENCLB702KNK	ENCFF232NPZ	fastq	SE51n

3. Образец Google Colab ноутбука с примерами анализа для fastq файла только одной из реплик. Вам следует сделать это для 2х реплик + контроль.

https://colab.research.google.com/drive/1cJ4E-iLq-x8rFxBeoCsbex_qs6JFsIFk?usp=sharing

Бонусная часть задания (2 балла)









Есть общая информация о типичном расположении гистоновой метки относительно генов (участков транскрипции). Задача посмотреть согласуется ли данные ChIP-seq эксперимента из ENCODE для выбранной гистоновой метки с картинкой ниже. Распределение сигнала метки из ENCODE эксперимента относительно генов можно получить с помощью программы ngs.plot (т.н. meta-gene plot) Это можно сделать с помощью ngs.plot.



Li e. al. (2007) Cell

Важно, чтобы версия генома по аннулированным генам в ngs.plot и версия генома, для которой был получен .bam/.bed файл совпадали.

Для этой задачи скачиваем 2 .bam файла с выравниваниями чтений на ВСЕ хромосомы

<div> <div>–</div> <div>Lab custom GRCh38 (ENCAN020ACA) processed data (12 Files)</div> </div>						
Accession		Default	File type	Output type	Isogenic replicate	Mapped read length
ENCFF662KPN	 		bed narrowPeak	peaks	1	
ENCFF211ZYL	 		bigBed narrowPeak	peaks	1	
ENCFF121HBN	 		bam	alignments	3	51

Для каждого .bam файла строим свой ngs.plot — и приводим оба графика в отчете Пробуем deepTools.

Список файлов для сдачи

- В репозитории в файле *README.md*
 - Ссылка на google colab ноутбук.
 - Выдача FastQC (или multiQC) для всех трех fastq файлов и анализ со скриншотами важных элементов в *README.md*. Указать, если была необходима фильтрация или подрезание чтений (и если это было сделано и как).
 - Таблицы/таблица со статистикой по каждому из 3 образцов:
 - Сколько ридов было в файле
 - Сколько ридов выравнилось уникально
 - Сколько ридов выравнилось НЕ-уникально
 - Сколько ридов НЕ выравнилось
 - Картинку с диаграммой Венна о пересечении наших MACS2 пиков и ENCODE пиков для двух реплик (можно оставить pdf в репозитории).
 - Ответы на все вопросы из колаба
 - Результаты выполнения бонусного задания

Форма отчетности

Github репозиторий, содержащий все полученные результаты.

Последний срок сдачи: среда, 1 марта до 23:59 (будет отслеживаться по последнему коммиту в репозиторий). Штраф -0.5 балла за каждый день просрочки.

В случае возникновения вопросов обращаться по telegram: @iv_sk