

# Инструкции для выполнения проекта (ВШЭ, майнор, 2023)

## Введение

Целью работы над проектом является изучение ранней эволюции белков, выполняющих различные эпигенетические модификации в клетках человека, методами сравнительной геномики.

## Подготовка

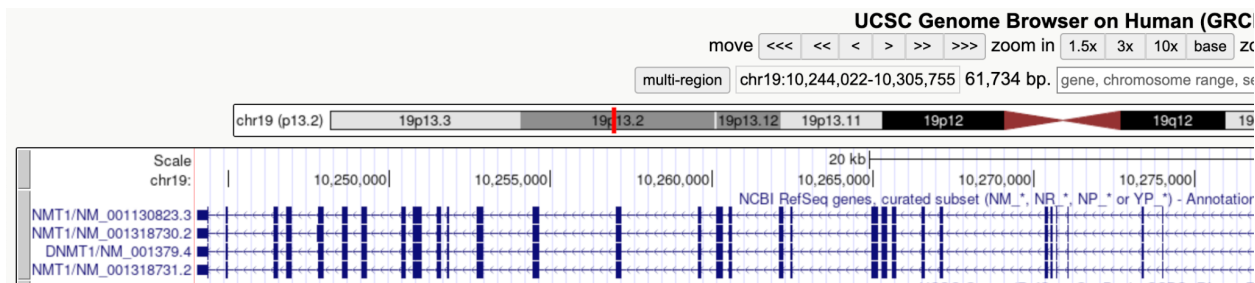
Из [таблицы](#) выбрать белок (-и), который делает какую-либо эпигенетическую модификацию в клетках человека, прописать название белка в [таблицу](#) (вкладка Project). При выборе эпигенетического белка имеет смысл ориентироваться на кол-во статей (неск сотен, а лучше больше 1000) по этому белку (или по комплексу, в состав которого данный белок входит) -- это можно проверить в БД NCBI (Pubmed) или Google Scholar. Сами статьи (если они в закрытом доступе) можно получить через ресурс <https://sci-hub.ru/>

Каждая группа готовит предварительную презентацию по введению в свою эпигенетическую модификацию (на основании литературного анализа):

- Таблица со списком участников группы (+выбранные белки и их функции)
- С какой модификацией связана функция выбранных белков
- Для каждого белка привести ссылку на статью, где сказано, что он действительно связан с этой модификацией (у всех белков группы должна быть одна и та же эпигенетическая модификация)
- В какие комплексы входят выбранные белки
- В каких тканях человека экспрессируются данные гены
  - См например БД NCBI Gene - <https://www.ncbi.nlm.nih.gov/gene/>
  - БД GTEx <https://gtexportal.org/home/gene/KDM6B>
- Доменная структура выбранных белков
  - БД Pfam
  - БД NCBI conserved domains - <https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>

## Обязательная часть задания

- Скачиваем посл-ть белка из человека. Если у выбранного гена несколько изоформ, выбираем самый длинный белок, например для гена DNMT1:



### RefSeq Gene DNMT1

RefSeq: [NM\\_001318731.2](#) Status: Reviewed  
Description: DNA methyltransferase 1, transcript variant 4  
Molecule type: mRNA  
Source: BestRefSeq  
Biotype: protein\_coding  
Synonyms: ADCADN,AIM,CXXC9,DNMT,HSN1E,m.Hsal,MCMT  
Other notes: isoform d is encoded by transcript variant 4  
OMIM: [126375](#)  
Protein: [NP\\_001305660.1](#)  
HGNC: [2976](#)

### DNA (cytosine-5)-methyltransferase 1 isoform a [Homo sapiens]

NCBI Reference Sequence: NP\_001124295.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to: ☐

LOCUS NP\_001124295 1632 aa linear PRI 29-MAY-2023  
DEFINITION DNA (cytosine-5)-methyltransferase 1 isoform a [Homo sapiens].  
ACCESSION NP\_001124295  
VERSION NP\_001124295.1  
DBSOURCE REFSEQ: accession [NM\\_001130823.3](#)  
KEYWORDS RefSeq; MANE Select.

- Запускаем поиск BLASTp вашего белка против 11-ти протеомов
  - Протеомы на сервере находятся тут  
`/mnt/storage/project_2023/teomes`
  - если делаете это через питоновский ноутбук, необходимо ставить ! в начале строки:

`blastp -query YOUR_PROTEIN.fasta -db`

`/mnt/storage/project_2023/teomes/drosophila.faa -out drosophila.blast -outfmt 7`

- В итоге получается 11 файлов с выдачей BLAST по каждому протеому

Выписать лучше (самые маленькие, первые по списку) E-value из каждого поиска в единую табличку -- желательно с помощью кода на Питоне (например в ноутбуке) . Если поиск не дал результатов, считаем, что E-value = 1 (те его log будет равен 0).

	human	mouse	..	Dropohilla	Yeast	Am	Inf
H2a	0	0		2.00E-69			
H2b							
H3							
H4							
YOUR_PROTEIN							

- Установка пакетов
  - При необходимости можно поставить модули питона командой:
    - `pip3 install pandas`
  - Либо создать окружение и ставить программы и модули туда:
    - `conda install -c bioconda muscle`
    -
- Конвертируем в  $-\log(\text{Evalue})$  -- Если  $\text{evalue} < 1\text{e-}300$ , то считать его  $1\text{e-}300$

	human	mouse	..	Dropohilla	Yeast	Am	Inf
H2a	300	300		69			
H2b							
H3							
H4							
YOUR_PROTEIN							

#### Анализ самих гистонов

- Каждый гистон может кодироваться несколькими генами (см <https://en.wikipedia.org/wiki/Histone>). Необходимо проверить, насколько они правда похожи друг на друга.
  - **Белковые посл-ти гистонов на сервере находятся тут:**
  - **`/mnt/storage/project_2023/histones`**
- Для этого предлагается скачать аминокислотные последовательности гистонов H2A, H2B, H3, H4, которые можно найти [тут](#) или на сервере в папке `/mnt/storage/project_2023/histones`. В файлах содержатся все аминокислотные последовательности, которые относятся к каждому гистону (включая разные изоформы генов). После скачивания необходимо провести выравнивание белковых последовательностей (например, с помощью программы MUSCLE, CLUSTAL, MEGAX), проанализировать получившиеся результаты, сделать выводы о том, являются ли эти гены копиями, и объяснить из-за чего могут быть различия. Результаты выравнивания и выводу нужно привести в репозитории в файле README.

- Для вызова программы CLUSTAL на кластере используйте команду "clustalw2"
- Далее нужно выбрать по одной аминокислотной последовательности для каждого гистона и сопоставить их с модельными организмами. Для этого нужно выбрать одну белковую последовательность из соответствующего файла и сохранить ее в отдельный файл в формате fasta. Далее нужно запустить blastp для выбранного белка и выбранного организма (пример запуска blastp [тут](#)). Это необходимо проделать для каждого гистона и каждого организма.
- Получившиеся результаты нужно добавить на тепловую карту.
  - a. В строках гистоны
  - b. В столбцах -- протеомы (записывается название организма)
  - c. Цвет ячейки --  $-\log(\text{BLASTp evaluate})$
  - d. **Порядок столбцов:**
    - i. **многокл позвоночные -- human, mouse, zebrafish**
    - ii. **многокл беспозвоночные -- drosophila, c.elegans**
    - iii. **однокл эукартиоты -- ciliate, yeast**
    - iv. **Археи -- methanocaldococcus, thermococcus**
    - v. **бактерии -- e.coli, tuberculosis**
  - e. Пример создания тепловой карты с помощью R-библиотеки ComplexHeatmap:  
[https://github.com/vanya-antonov/article-chelatase/blob/master/images\\_R/heatmap\\_full.R](https://github.com/vanya-antonov/article-chelatase/blob/master/images_R/heatmap_full.R)
  - f. Пример создания тепловой карты с помощью Python:  
[https://colab.research.google.com/drive/1ec\\_IBTF6EIUfQzIBudq8R1tc-aYLOCMW?usp=sharing](https://colab.research.google.com/drive/1ec_IBTF6EIUfQzIBudq8R1tc-aYLOCMW?usp=sharing)

## Требования к отчету по индивидуальному заданию

1. Отчет оформлен на github в файле README.md
2. Краткое описание выбранного гена (эпигенетическая функция соотв белка, ссылки на статьи (по крайней мере 2), где указана его связь с эпигенетической меткой, где экспрессируется и тп)
3. Множественное белковое выравнивание каждого из 4-х гистонов (или скриншоты из программы просмотра этого выравнивания, например [AliView](#))
4. Таблички с E-value и  $-\log(\text{Evalue})$  для 4х гистонов и вашего белка (и 11 протеомов в столбцах)
5. Тепловая карта созданная по табличке  $-\log(\text{E-value})$

## 6. Вывод о том, насколько рано в эволюции появился Ваш выбранный белок

# Требования к финальной групповой презентации

- Краткое описание того, что было сделано
  - Табличка с информацией по последовательностям белков человека (queries), которые были использованы для поиска по протеомам. В эту табличку включить следующую информацию:
    - название гена (например, AIRE)
    - количество изоформ у данного гена (по аннотации RefSeq или UCSC)
    - идентификатор белка, выбранного для поиска (например NP\_000374.1)
    - Длина выбранного (например 545 аминокислотных остатков)
    - Домены, которые присутствуют в данном белке
  - Как именно был проведен поиск BLASTp по протеомам, какие пороги использовались, сколько хитов (результатов поиска) было найдено в каждом поиске, уменьшалось ли кол-во хитов при удалении от человека по эволюционной лестнице.
- Объединить результаты по всем белкам группы и показать общую тепловую карту.
- Для каждого из 10-ти белков группы выбрать хит BLASTp (найденный белок) в протеоме наиболее удаленного от человека организме. Важно, чтобы этот дальний гомолог имел достаточно "сильное" E-value (например  $< 1e-10$ ) и его длина была сопоставима с длиной исходного человеческого белка (разница не более 100 или 200 а.о.).
- Подготовить табличку с информацией о отобранных далеких гомологах -- название гена, идентификатор белка и описание его функции длина бе
- Для каждого из выбранных далеких гомологов привести следующую информацию:
  - Название гена (если есть) и идентификатор белка-гомолога
  - описание функции данного белка-гомолога
  - набор доменов в белке-голомоле
- Посмотреть статьи о наличии данной модификации у наиболее древних организмах
  - в частности поискать информацию о том, действительно ли функция найденных далеких белков-ортологов связана с выбранной эпигенетической метков
- В конце презентации слайд Выводы
  - Ваше мнение о том, в каком из наиболее ранних организмах появилась выбранная эпигенетическая метка (только у позвоночных, или беспозвоночные тоже имели данную модификацию или возможно даже у одноклеточных эукариот и тп)

# Форма отчетности

Индивидуальный Github репозиторий, содержащий все полученные результаты и код.

Групповой Github репозиторий, содержащий файлы с 2-мя презентациями (вводной и финальной) + любой код, который Вы использовали (например, для создания общей тепловой карты).

Последний срок сдачи: **воскресенье, 11 июня до 23:59** (будет отслеживаться по последнему коммиту в репозиторий).