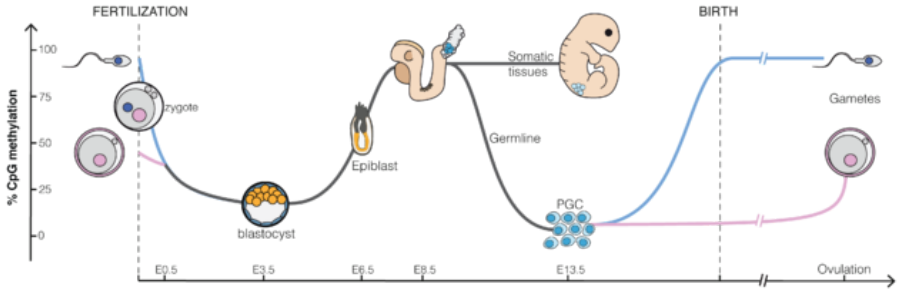
Домашнее задание №1

Введение

Целью данного домашнего задания является изучение глобального изменения уровня CpG метилирования ДНК при раннем эмбриональном развитии мыши. Считается, что при развитии эмбриона происходят так называемые волны деметилирования-метилирования, т.е. на ранних стадиях CpG метилирование уменьшается до некоторого минимума (около 25%), а затем по мере дифференцировки тканей, оно сильно увеличивается (около 90%) и остается таким на протяжении всей жизни организма:

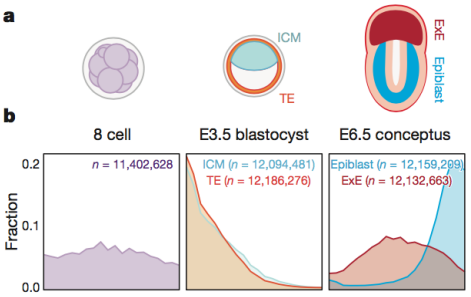


https://en.wikipedia.org/wiki/DNA_methylation#During_embryonic_development

Для выполнения данной задачи мы будем изучать следующие образцы WGBS (Whole genome bisulfite sequencing), соответствующие разным стадиям эмбрионального развития мыши:

- 8cell – 8-клеточный эмбрион, примерно 2.25 дня после оплодотворения яйцеклетки
- ICM – внутренняя клеточная масса бластоциста, примерно 3.5 дня после оплодотворения яйцеклетки
- Epiblast – стадия эпибласта, примерно 6.5 дней после оплодотворения яйцеклетки

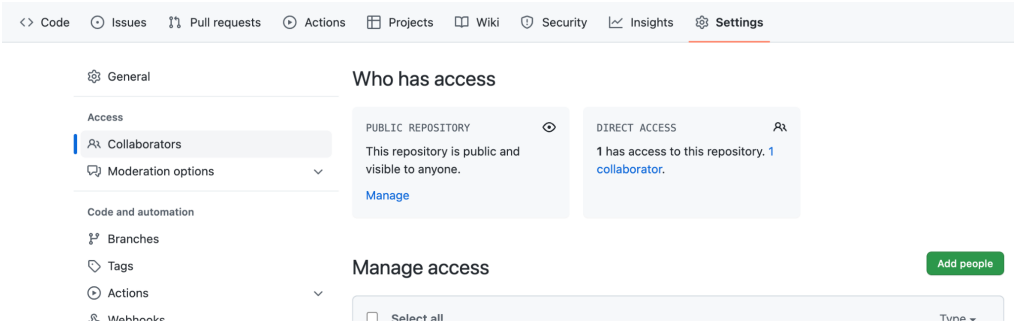
В соответствующей статье [PMID: 28959968] авторы приводят следующие распределения метилирования цитозинов:



В целях экономии времени мы будем анализировать только одну из двух реплик, а также выравнивать чтения только на одну из хромосом мыши.

Обязательная часть задания (8 баллов)

- На сайте github.com создаем **приватный** репозиторий и приводим ссылку на этот репозиторий в общей гугл-таблице (**вкладка HW1**) https://docs.google.com/spreadsheets/d/10r73G68KiXa-7Kf_u1Nir1cITd-3xy1NbBX7_kGk0A/edit#gid=0
    - Также необходимо дать доступ ассистенту (**efrsw**) к репозиторию для будущей проверки (Settings => Collaborators => Add people):



- Рекомендуется выполнять работу в Google Colab ноутбуках.
    - Если вы будете выполнять работу на сервере или на своем компьютере, необходимо будет также загрузить написанный код на Github
- В данном задании будут проанализированы следующие 3 BS-Seq образца, полученные на разных стадиях развития эмбриона мыши:
    - SRR5836473 - 8 Cell
    - SRR5836475 - ICM
    - SRR3824222 - Epiblast
- Скачайте любой из запусков и проведите анализ QC прочтений.Для скачивания можно воспользоваться SRAtoolkit или скачать с European Nucleotide Archive (просто вбиваете в поиске номер запуска или эксперимента или по FTP-ссылке скачиваете файлы с помощью wget). **Какие особенности можно наблюдать по сравнению с секвенированием ДНК или РНК? Загрузите отчет html в репозиторий.**
- В целях экономии времени и ресурсов, прочтения 3 образцов были выровнены на 11 хромосому мыши. Файлы с выравниваниями и отчеты о работе Bismark размещены здесь:
    - https://drive.google.com/drive/folders/1athn52a93obOwfuHynDCRPRhjAPB80aO?usp=sharing
- Пример Google Colab ноутбука с примерами запуска только для одного файла (образца). **Вам следует сделать это для 3 образцов**.
    - https://colab.research.google.com/drive/1xonncIMhr8llOhpJg5AlnpEM_-q7iilF?usp=sharing
- Работа с bam-files с выравниваниями BS-seq ридов на **11-ю хромосому** мыши (используем samtools view) :
    a. Результатом этой части задания будет сводная таблица, в которую необходимо записать число ридов, закартированных на участки *11347700-11367700; 40185800-40195800*. Занесите таблицу в read.me.
- Проведите дедупликацию файлов выравниваний:
    a. Сколько процентов прочтений дуплицированно в каждом из образцов? Занесите таблицу в read.me.
- Проведите коллинг метилирования цитозинов.

- Выведите отчет в формате html. Файл html загрузите в директорию github. Прикрепите скриншот M-bias plot и кратко опишите, что вы на нем видите (читайте manual) .
- С помощью файла bismark.cov или .bedgraph постройте гистограмму распределения метилирования цитозинов по хромосоме (отображение насколько часто метилируются цитозины в данном образце: по X процент метилированных цитозинов, по Y - частота). Сделайте выводы. Код построения гистограммы нужно прикрепить. Можно, в Python или R.

Бонусная часть задания (2 балла)

- Визуализируйте уровень метилирования и покрытия для каждого образца (для этого нужно составить файл с трэками, читайте мануал). С помощью pyGenomeTracks. Или другим любым способом (можно через UCSC GenomeBrowser).

Список файлов для сдачи

- В репозитории в файле *README*.md
    - Ссылки на google colab ноутбуки
    - Скриншоты/файлы html и ответы на вопросы
    - Таблицы/таблица со статистикой по каждому из 3 образцов:
        - Сколько ридов пришлось на целевые регионы
        - Сколько дуплицированных чтений в каждом образце
    - Гистограмма с общим уровнем метилирования для каждого из образца
    - Рисунки с уровнем метилирования и покрытием на любом регионе (лучше больше 10 000 нуклеотидов) для каждого образца
- В репозитории в папке src – любой другой код, который был использован для выполнения задания (например, для создания гистограмм)

Форма отчетности

Github репозиторий, содержащий все полученные результаты.

Последний срок сдачи: среда, 8 февраля до 23:59 (будет отслеживаться по последнему коммиту в репозиторий). Штраф -0.5 балла за каждый день просрочки.

В случае возникновения вопросов обращаться по каналам связи telegram ассистентов: @efrsw