Описание проекта “Методы визуализации агрегированных эпигенетических данных"

<https://internship.jetbrains.com/projects/1000/>

### **Описание**

Группа [BioLabs](https://research.jetbrains.org/ru-ru/groups/biolabs/) разрабатывает новые алгоритмы и методы анализа данных биологических экспериментов, и работает в тесном сотрудничестве с биологами над различными экспериментами по изучению старения.

Один из проектов группы – это масштабное лонгитюдное исследование связи изменений метилирования ДНК, генной экспрессии и метаболомики в процессе старения человека на выборке из более чем 100 людей. Для экспериментов с большим кол-вом образцов, задачи анализа и визуализации данных требуют дополнительных шагов по агрегации геномных данных.

В рамках стажировки вы будете работать над агрегацией и визуализацией экспериментальных данных, которые позволят выдвинут новые гипотезы о данных или наглядно представить результаты их анализа. Например, для различных типов данных нужно будет реализовать построение BigWig треков, отражающих:

* корреляцию сигнала в соседних геномных регионах;
* усредненный сигнал и его дисперсию по возрастным группам;

Построение геномных треков и таблиц должно быть оформлено в виде Snakemake pipeline.

По результатам работы над проектом возможно участие в публикации.

Перед собеседованием нужно выполнить тестовое задание, можно сделать его частично, но чем полнее, тем выше шансы на получение стажировки в этом проекте.

Вопросы можно задавать по почте: roman.chernyatchik@jetbrains.com

### **Требования**

* Понимание основ биоинформатики
* Базовые знания Python, умение работать с библиотеками Pandas/Numpy
* Понимание форматов файлов: Bed, BedGraph, Wig, BigWig
* Уверенное владение командной строкой Unix
* Приветствуется базовые знания статистики; опыт написания pipelines на Snakemake; владение [Plot.ly](http://plot.ly/), Matplotlib, Seaborn

Задание

* Скачайте таблицу с данными по methylation level цитозинов <https://artyomovlab.wustl.edu/publications/supp_materials/aging/rrbs/filtered_cytosines_freq.tsv.gz>. Данные приведены для сборки генома hg19. Первый столбец соответствует позиции в геноме, остальные уровню метилирования в соответствующем доноре.
* Требуется выполнить пункты ниже для размера бина 100bp, 1kbp, 10kbp, 100kpb, 1mb и в итоге получить 5 BigWig треков.
* Разбейте геном на непересекающиеся бины заданной длинны. Например, chr1:0-1000, chr1:1000-2000, chr1:2000-3000,.... для размера бина 1 kbp.
* Для каждого бина в каждом доноре вычислите средний уровень метилирования. За средний уровень метилирования в бине можно принять среднее арифметическое уровня метилирования по всем цитозинам, для которых есть данные.
* Таким образом, каждому бину можно сопоставить вектор среднего уровня метилирования по всем донорам. Посчитайте пирсоновскую корреляцию между данными в бине и в соседнем бине справа.
* Сгенерируйте BigWig трек, где в каждом бине будет отображаться значение посчитанной корреляции, и все бины будут сдвинуты на половину длина бина вправо (на границу между соседними бинами). Например, для бинов chr1:0-1000, chr1:1000-2000, chr1:2000-3000,.... нужно получить трек с значениями коэф корреляции в регионах: chr1:500-1500, chr1:1500-2500, chr1:2500-3500, .., где региону chr1:1500-2500 соответствует коэф. корреляции векторов chr1:1000-2000, chr1:2000-3000.
* Вычисления нужно оформить в виде python / bash / snakemake или тп скрипта
* В качестве ответа дайте ссылку на скачивание архива с скриптами + bigwig треками. Если Вы используете Google Drive, то не забудьте сделать ссылку публичной или предоставьте права на скачивания для пользователя с почтой roman.chernyatchik@jetbrains.com