Домашнее задание №1

Введение

Целью данного задания является сборка генома бактерии, выделенной из воды с нефтью, на основании парно-концевых (paired-end, PE) и чтений типа mate-pairs (MP).

Обязательная часть задания (8 баллов)

- На сайте github.com создаем публичный репозиторий «hse22_hw1» и приводим ссылку на этот репозиторий в общей гугл-таблице: https://docs.google.com/spreadsheets/d/1bdKBlDMFvhX8xStZUL2YHRTGLpmdiLNU5JcVmy8K-Yw/edit?usp=sharing
- Исходные файлы в формате .fastq лежат на сервере в папке /usr/share/data-minor-bioinf/assembly/ и на гугл диске https://drive.google.com/drive/folders/0B1L-4UZI5kYmTjF4S1lfQ2dlQzQ?resourcekey=0-dBYLwDGRcSleV5Ic2sqeNQ&usp=sharing
    - Чтения типа paired-end: oil_R1.fastq, oil_R2.fastq
    - Чтений типа mate-pairs: oilMP_S4_L001_R1_001.fastq, oilMP_S4_L001_R2_001.fastq
- Работать и создавать файлы следует только внутри своей домашней папки:
    - Обратите внимание, что копировать файлы с сырыми данными себе в папку необязательно (чтобы не занимать лишнее место на сервере). Вместо этого у себя в папке можно создать символическую ссылку на каждый из файлов, например:

```
ln -s /usr/share/data-minor-bioinf/assembly/oil_R1.fastq
```

- Запускать долго-работающие программы можно с помощью команды screen или tmux:
    - Все программы запускаем на одном процессоре
    - Перед запуском проверяем, что на данный момент на сервере есть свободные процессоры (например, командой htop)
- С помощью команды seqtk выбираем случайно 5 миллионов чтений типа paired-end и 1.5 миллиона чтений типа mate-pairs (чтобы у каждого получился свой уникальный результат)
    - Для параметра -s (random seed) важно использовать одно и то же число (например, месяц и дата вашего рождения):

```
seqtk sample -s722 read1.fq 10000 > sub1.fq
seqtk sample -s722 read2.fq 10000 > sub2.fq
```

- С помощью программы fastQC и multiQC оценить качество исходных чтений и получить по ним общую статистику
    - Привести эту информацию и картинки в отчете на github-е
- С помощью программ platanus_trim и platanus_internal_trim подрезать чтения по качеству и удалить адаптеры
    - Рекомендации по использованию программ пакета Platanus можно найти по ссылке http://platanus.bio.titech.ac.jp/platanus-assembler/sample-page
- **ВАЖНО – после подрезания чтений удалите исходные .fastq файлы, полученные с помощью программы seqtk (они больше не будут нужны)**

- С помощью программы fastQC и multiQC оценить качество подрезанных чтений и получить по ним общую статистику
    - Привести эту информацию и картинки в отчете на github-е
- С помощью программы “platanus assemble” собрать контиги из **подрезанных** чтений
- Написать код (в Jupiter ноутбуке или в Google Colab) для анализа полученных контигов (общее кол-во контигов, их общая длина, длина самого длинного контига, N50)
    - Загрузить Jupiter ноутбук с кодом на github в папку src или, если писали в Google Colab, то укажите ссылку на колаб-ноутбук
- С помощью программы “ platanus scaffold” собрать скаффолды из **контигов**, а также из **подрезанных** чтений
- Написать код (в Jupiter ноутбуке или в Google Colab) для анализа полученных скаффолдов (общее кол-во скаффолдов, их общая длина, длина самого длинного скаффолда, N50)
    - Загрузить Jupiter ноутбук с кодом на github в папку src или, если писали в Google Colab, то укажите ссылку на колаб-ноутбук
- Для самого длинного скаффолда посчитать количество гэпов (участков, состоящих из букв NNNN) и их общую длину
- С помощью программы “ platanus gap_close” уменьшить кол-во гэпов с помощью **подрезанных** чтений
- **ВАЖНО – удалите .fastq файлы, содержащие подрезанные чтения (они больше не будут нужны)**
- Для самого длинного скаффолда посчитать количество гэпов (участков, состоящих из букв NNNN) и их общую длину

Бонусная часть задания (2 балла)

- Собрать геном из меньшего кол-ва чтений и посмотреть как это влияет на качество сборки, которое можно оценить по следующим параметрам:
    - кол-во контигов и скаффолдов
    - значение N50 для контигов и скаффолдов
    - Длина самого длинного контига и скаффолда
    - Кол-во гэпов в самом длинном контиге и скаффолде
    - и т.д.

Список файлов для сдачи

- В репозитории в файле *README*.md
    - Список всех команд, которые были выполнены на сервере
    - Скриншоты и статистика из файлов multiQC
    - Ссылки на google colab или jupiter ноутбуки (если имеются)
    - Результаты бонусной части задания (если имеются)
- В репозитории в папке data
    - Файл contigs.fasta – все полученные контиги
    - Файл scaffolds.fasta – все полученные скаффолды (после команды gap_close)
    - Файл longest.fasta – последовательность самого длинного скаффолда
- В репозитории в папке src – любой код, который был использован для выполнения задания

Форма отчетности

Github репозиторий, содержащий все полученные результаты (в README.md), файлы, собранный геном, а также команды и программный код, которые были использованы для работы (например, в виде ссылки на ноутбук Google colab).

Последний срок сдачи: среда, 5 октября 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: @PlainSight
ВК: https://vk.com/plainsight