BostonGene

ТЕСТОВОЕ ЗАДАНИЕ

Это задание нужно для проверки технических навыков и навыков поиска и анализа информации; бОльшая часть оценки будет основываться на выбранных и примененных решениях а не на сравнении итогового результата с неким “эталоном”;

Глобальная задача – найти CNA варианты в образце опухоли.
Входные данные:

- pileup файл, сгенерированный на основе normal.bam и tumor.bam
- vcf файл с соматическими мутациями
- Список генов, которые нужно учитывать в анализе*

(В качестве дополнительного референса можно передать bed файл с координатами генов)

Для уменьшения ресурсоемкости расчетов решение можно ограничить одной или несколькими хромосомами, если это необходимо, но крайне желательно чтобы в итоговом результате были разные события (делеции, нормы, амплификации).

Решение должно быть оформлено в виде докер образа. В качестве ENTRYPOINT необходимо выставить Python скрипт, обеспечивающий логику проводимого анализа данных. Все необходимые входные файлы должны передаваться аргументами командной строки (модуль argparse). Выполнение программы должно сопровождаться подробными логами (модуль logging), а вспомогательные сторонние программы (прим.: скрипты на R или другие CLI) должны вызываться внутри скрипта (модуль subprocess).

Ожидаемый формат выполненного задания – ссылка на github репозиторий со всем необходимым кодом, Dockerfile и подробной инструкцией в README.

Вещи которые могут помочь в выполнении задачи:

- Перед началом работы полезно ознакомиться с работой нескольких существующих тулов, выделить ключевые вещи с биологической и биоинформатической точки зрения и вещи которые могут влиять на результат (ключевые слова: purity, ploidy, depth ratio, BAF, normalisation etc.)
- Для работы с vcf файлами в питоне рекомендуется использовать библиотеку pysam