Отчет о проделанной работе с использованием оборудования ИВЦ НГУ

Тема работы:

Сборка и анализ генома человека

Состав коллектива:

- Штокало Дмитрий Николаевич, ИСИ СО РАН, н.с., к.ф.-м.н., shtokalod@gmail.com
- Вяткин Юрий Викторович, ООО "АкадемДжин", биоинформатик, руководитель проектов, vyatkin@gmail.com
- Антонец Денис Викторович, ФГУН ГНЦ ВБ "Вектор" Роспотребнадзора, с.н.с., к.б.н., antonec@yandex.ru
- Кабилов Марсель Расимович, ИХБФМ СО РАН, руководитель ЦКП "Геномика" СО РАН, к.б.н., kabilov@niboch.nsc.ru

Научное содержание работы:

Постановка задачи

Проект направлен на изучение методов сборки генома человека *de novo* из данных секвенирования, полученных на приборах секвенирвания второго и третьего поколения. Получаемые сборки геномов сравниваются с существующими референсными геномами и между собой. Кроме того, задача предполагает получение из данных секвенирования дополнительных сведений, таких как состав однонуклеотидных и структурных (протяженных) вариаций по отношению к референсному геному и прочих.

Подробное описание работы, включая используемые алгоритмы

В работе предполагается (пере-)сборка геномов автохтонных народов Сибири по данным секвенирования ДНК с помощью существующих методов и алгоритмов, таких как процессирование графов де Брёйна, реализованных в ряде пакетов ПО, таких как SPAdes, Velvet, SOAPdenovo, ABySS. У коллектива имеются данных секвенирования 9-ти представителей бурятского этноса и 2-х представителей якутского этноса, полученные на секвенаторе SOLiD 5500xl, а также публичные данные секвенирования человека, такие как GIAB NA12878, полученные на различных платформах секвенирования. Полученные фрагменты геномов (контиги и скаффолды) планируется сравнивать с существующими рефернсными сборками геномов человека, такими как GRCh37 и GRCh38, путем картирования фрагментов на целый геном с помощью таких методов как LastZ и MultiZ.

Предполагается также разработка собственных методов сборки генома, основанных на использовании графов де Брёйна и реализующих данный подход библиотек, таких как GATB и bruno.

Также необходимо проведение филогенетического анализа полученных геномов и фрагментов геномов существующими методами, такими как Phylip. Филогенетические расчеты необходимо проводить как для целых хромосом, включая хромосому Y и митохондриальную хромосому, так и для отдельных интересных фрагментов с большим числом однонуклеотидных вариаций. Поиск вариации геномов для анализа также проводится существующими и оригинальными методами. Для этого существующие данные секвенирования картируются на референсный геном человека с помощью такого ПО как STAR, BWA и HISAT2, а затем производится поиск однонуклеотидных вариаций в выравнивания с помощью пакета GATK.

Полученные результаты

В результате работы были протестированы инструменты сборки геномов *de novo*, такие как SPAdes, ABySS и SOAPdenovo. Полученные сборки фрагментов геномов человека для некоторых представителей бурятского этноса были проанализированы с точки зрения качества и соответствия референсному геному человека. Качество полученных сборок не было признано удовлетворительным, что говорит о необходимости, во-первых, дальнейших настроек параметров используемого ПО, а во-вторых, о необходимости разработки иных подходов для работы с такими данными.

Эффект от использования кластера в достижении целей работы

Работа не может быть выполненна без привлечения вычислительных узлов с большим объемом оперативной памяти. По нашим оценкам, для сборки генома человека минимальный необходимый объем памяти составляет окого 800Гб. Данные ресурсы доступны в ИВЦ НГУ.

Перечень публикаций, содержащих результаты работы

- нет