УДК 575.112

РАЗРАБОТКА МНОГОПОЛЬЗОВАТЕЛЬСКОЙ ПЛАТФОРМЫ ИНТЕРАКТИВНОЙ ВИЗУАЛИЗАЦИИ КАРТ Hi-C КОНТАКТОВ ДЛЯ СКАФФОЛДИНГА, ВАЛИДАЦИИ И СРАВНИТЕЛЬНЫХ ИССЛЕДОВАНИЙ ГЕНОМНЫХ СБОРОК

Синицын А.А. (ИТМО)
Научный руководитель – аспирант ФИТИП Замятин А.А.
(ИТМО)

Введение. Скаффолдинг является последним этапом процесса геномной сборки и заключается в упорядочивании и ориентировании контигов – однозначно определённых последовательностей ДНК, полученных на выходе автоматического сборщика, в последовательности большего размера – скаффолды, которые должны соответствовать истинной последовательности нуклеотидов в молекуле ДНК.

С появлением технологий секвенирования, таких как Oxford Nanopore и PacBio HiFi, позволяющих получать протяжённые области ДНК в одном прочтении, можно говорить о наступлении новой эры геномных сборок на уровне хромосомных скаффолдов, где один скаффолд почти полностью отражает последовательность всей молекулы ДНК (хромосомы). Крупные международные консорциумы, такие как The Human Pangenome Reference Consortium (HPRC) и The Vertebrate Genomes Project (VGP) на сегодняшний день заинтересованы в получении большого числа качественных сборок хромосомного уровня. Применение современных алгоритмов сборки для длинных прочтений позволяет получать протяженные контиги, каждый из которых определён однозначно, но неизвестен порядок их следования и ориентация. Для достижения хромосомного уровня сборки необходим этап скаффолдинга.

Скаффолдинг по данным Hi-C на данный момент является одним из основных методов, финализирующих процесс геномной сборки. Hi-C – это метод молекулярной биологии, позволяющий получить информацию о взаимном расположении участков ДНК в трёхмерном пространстве. Эти данные можно использовать в процессе скаффолдинга для того, чтобы правильно упорядочить и ориентировать контиги.

Существует несколько автоматических скаффолдов, таких как 3D-DNA, SALSA и YAHS, способных провести значительную часть скаффолдинга в полностью автоматическом режиме, однако для получения сборки наивысшего качества последний этап скаффолдинга на настоящий момент производится и проверяется человеком, что значительно затрудняет общую автоматизацию процесса геномной сборки хромосомного уровня. Таким образом, актуальной является разработка ПО для интерактивного скаффолдинга геномных сборок.

Основная часть. На данный момент существуют несколько программных инструментов, решающих эту задачу. JBAT, разработанный лабораторией Айдена, много лет являлся единственным инструментом, который используется локально для работы с файлами .hic, позволяет производить все необходимые операции, но имеет недостатки, такие как медленная работа с сильно фрагментированными геномами большого размера и большие требования к оперативной памяти. В международной лаборатории КТ Университета ИТМО был разработан инструмент HiCT решающий те же задачи, но при этом устранивший прежние недостатки. Засчет использования продвинутой модели данных, получилось ускорить процесс работы с картами контактов и уменьшить нагрузку на оперативную память. Следующим шагом должна стать разработка многопользовательской платформы для университета ИТМО.

Инструмент HiCT предоставил совершенно новый взгляд на интерактивное взаимодействие с картами Hi-C контактов. В свою очередь многопользовательская платформа позволит пользователям загружать различные данные для создания

глобального хранилища. Структурированное хранилище повысит эффективность использования инструмента.
Платформа позволяет легко масштабировать проекты и добавлять новых пользователей, не теряя при этом производительности. Доступ можно сделать ограниченным либо же полностью публичным. Платформа предоставляет интерфейс управления конфиденциальностью информации, а так же может модульно расширить функционал исходного приложения.
Многопользовательская платформа разработана как микросервисная архитектура, для большей масштабируемости и балансировки нагрузки. Был выделен отдельный сервис под управление всеми инстансами приложения HiCT, которые разворачиваются независимо для каждого пользователя для хранения и модификации его текущего состояния системы. Асинхронное общение микросерсисов, которое реализовано на основе брокера сообщений, повышает отзывчивость системы.

Выводы. В рамках этой работы, была реализована многопользовательская платформа, на которую исследователи могут загружать свои данные Hi-C и получать возможность интерактивно взаимодействовать с картой контактов в Web-интерфейсе, получать доступ к сформированной базе публичных данных для проведения сравнительных исследований.
В ближайшее время планируется внедрение разработанной платформы и ее развертывание на вычислительном кластере факультета, как основная работа в рамках выполнения гранта НИРМА номер 623082.

Список использованных источников:

1. Ghurye, Jay, et al. "Integrating Hi-C links with assembly graphs for chromosome-scale assembly." PLoS computational biology 15.8 (2019): e1007273.
2. J.A. Abraham, "Load Balancing in Distributed Systems" [Электронный ресурс] – https://ieeexplore.ieee.org/abstract/document/1702962.
3. Adam Phillippy, "The (near) complete sequence of a human genome", [Электронный ресурс] – https://genomeinformatics.github.io/CHM13v1/.
4. Dudchenko, Olga, et al. "De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds." Science 356.6333 (2017): 92-95.
5. Paulsen, Victor Kai Oscar, "Implementation of a component to manage authorization for a web application" [Электронный ресурс] – https://lup.lub.lu.se/student-papers/search/publication/9069204
6. Rhie, Arang, et al. "Towards complete and error-free genome assemblies of all vertebrate species." bioRxiv (2020).