Наименование проекта: ИРН AP19678041 «Разработка программного обеспечения для идентификации тандемных повторов при полногеномном секвенировании»

Актуальность:

За последние десятилетия кардинально изменилось представление роли повторяющихся последовательности в геноме, и из разряда «мусорной ДНК», повторяющие элементы оказывают большое влияние на функционирование и эволюцию геномов их хозяев, способствуя генетическому разнообразию и появлению новых регуляторных элементов. Дальнейшее развитие технологии секвенирования и в особенности секвенирование третьего поколение значительно способствует в изучении тандемных повторов, что привело к появлению новых данных для детального изучения. Установлено, что на долю коротких тандемных повторов приходится около 7% генома человека. Широкая представленность в геномах эукариот и прокариот, и их высокая скорость изменчивости, как один из ключевых факторов эволюции генома и генетической диверсификации, повторы будут систематически оцениваться на предмет их роли. Поскольку известно, что многие из этих элементов активируются при заболеваниях, возникает потенциал для персонализированной медицины и диагностики болезней относительно генетических изменений и ожидаемых последствий при анализе тандемных повторов и выявление ассоциаций биомаркеров и регулированием биологических процессов в организме. В связи с этим разработка продвинутых и простых в использовании биоинформатических инструментов для идентификации разнообразных форм тандемных повторов, является **актуальной задачей**.

Цель предлагаемого проекта заключается в разработке биоинформационного приложения с открытым доступом для идентификации и анализа вариабельности тандемных повторов, в том числе, в исходных данных при полногеномном секвенировании третьего поколения.

Ожидаемые и достигнутые результаты:

В рамках проекта будут применятся биоинформатические алгоритмы для идентификации родственных последовательностей с разным уровнем дивергенции, а также языки и средства программирования при разработке структуры и интерфейса приложения. Подтверждение достоверности результатов, получаемых в разрабатываемом программном обеспечении, будет проводится стандартными лабораторными молекулярно-генетическими методами, включая полногеномное секвенирование геномов прокариот и эукариот, и методы для дифференциации тандемных повторов с помощью капиллярного электрофореза. **Основным результатом** реализуемого проекта будет программное обеспечение с открытым доступом и пользовательским интерфейсом для идентификации тандемных повторов при полногеномном секвенировании. Программное обеспечение позволит идентифицировать

разнообразие целевых локусов с тандемными повторами, в том числе, в исходных данных полногеномного секвенирования и проводить статистический анализ идентифицированных вариантов.

В ходе реализации проекта, будут опубликованы не менее 2 (двух) статей и (или) обзоров в рецензируемых научных изданиях, входящих в 1 (первый) и (или) 2 (второй) квартиль по импакт-фактору в базе Web of Science и (или) имеющих процентиль по CiteScore в базе Scopus не менее 65 (шестидесяти пяти); либо не менее 1 (одной) статьи или обзора в рецензируемом научном издании, входящем в 1 (первый) квартиль в базе Web of Science или процентиль по CiteScore в базе Scopus не менее 95 (девяносто пяти). Все биоинформационные коды, скрипты будут размещены в постоянных, открытых репозиториях, а также размещены на Github со свободным доступом.

Члены исследовательской группы:

руководитель проекта – Исмаилова Айсулу Абжаппаровна, ГНС, PhD, ассоциированный профессор

ORCID: 0000-0002-8958-1846

Scopus/WoS (Hirsch Index = 3): Scopus Author ID: 56145830200

исследовательская группа:

1) **Календарь Руслан Николаевич**, ГНС, к.б.н., биолог-генетик, Профессор (Биология), Доцент генетики (университет Хельсинки)

ORCID: 0000-0003-3986-2460

Scopus/WoS (Hirsch Index = 34): ResearcherID: D-9751-2012

Scopus ID: 6602789279

2) **Бельдеубаева Жанар Толеубаевна**, ВНС, PhD

ORCID: 0000-0003-4056-6220

Scopus/WoS (Hirsch Index =2): Scopus Author ID: 56951278600

3) Сатыбалдиева (Сатекбаева) Айжан Жанабековна, ВНС, PhD

ORCID: 0000-0001-5740-7934

Scopus/WoS (Hirsch Index =2): Scopus Author ID: 56145597900

4) **Шевцов Владислав Александрович**, СНС, магистр технических наук докторант кафедры «Информационные системы» НАО «Казахский агротехнический университет им. С. Сейфуллина»

ORCID: 0000-0001-6202-2123

Scopus/WoS (Hirsch Index =3): Scopus Author ID: 57216896596

5) **Голенко Екатерина Сергеевна**, СНС, магистр технических наук

ORCID: 0000-0002-4643-4571

6) **Вакансия**, ВНС, IT архитектор, программист

7) **Вакансия**, НС, докторант

Информация для потенциальных пользователей:

Область применения разработанного программного обеспечения: биоинформатика, медицинская и сельскохозяйственная генетика, генетика

микроорганизмов. **Результаты этого проекта** имеют большое значение, в том числе, для фундаментальных наук. Программное обеспечение позволит эффективно идентифицировать тандемных повторов и установлению ассоциаций между разнообразием тандемных повторов с генетическими заболеваниями человека и с генетическим разнообразием микроорганизмов и их патогенностью. Реализация проекта позволит усилить направление биоинформатики в ведущем вузе страны и создаст платформу для специализации и профориентации обучающихся.

Полученные результаты по проекту за 2023 год.

1) Разработана библиотека классов, которая позволит идентифицировать последовательности, содержащие целевые локусы с тандемными повторами с известной природой, а также предсказывать тандемные повторы для участков со скрытой сигнатурой и неизвестной природой. Идентифицированные тандемы классифицированы в соответствии с их сигнатурой, природой повтора и гетерогенности тандемных блоков, для дальнейшего анализа и идентификации аллельных вариантов у сравниваемых генотипов. Был разработан алгоритм и программный код для выявления любых типов повторов в геномных последовательностях. Кроме того, анализ повторов проводится на полногеномных последовательностях из генбанка NCBI. Повторяющиеся последовательности являются функционально вездесущими структурными единицами, которые встречаются во всех геномах. Однако, из-за разнообразия повторов, каждый из которых имеет уникальную сигнатуру и структуру, затрудняет их классификацию. Чтобы преодолеть эту проблему, мы разработали инструмент для выявления любых типов повторов в геномных последовательностях. Java инструмент для идентификации повторов и результаты геномного анализа у различных таксономических видов, включающих геномы эукариот, грибов, микроорганизмов и гигантских вирусов доступен по адресу https://zenodo.org/records/8424601, а исходный код свободно доступен на GitHub по адресу https://github.com/rkalendar/Repeater.

2) Была проведена проверка эффективности разработанного кода идентификации последовательностей с тандемными повторами с использованием некоторых алгоритмов, включая линейные модели ближайших соседей, алгоритм Кнута-Морриса-Пратта, алгоритм Бойера-Мура, алгоритм Рабина-Карпа и алгоритм суффиксных деревьев. В качестве оценки эффективности при тестировании были выбраны следующие параметры: скорость работы каждого алгоритма, количество найденных тандемных повторов. В процессе проведения проверки эффективности кода для идентификации последовательностей тандемных повторов были выполнены следующие шаги:

- Подготовлены искусственные данные биологических последовательностей, содержащие различные паттерны тандемных повторов, которые требовалось идентифицировать.

- Проведены серии тестов, где каждый алгоритм был применен к подготовленным данным. При сравнении алгоритм с использованием суффиксных деревьев был определен как наиболее эффективный вариант алгоритма для идентификации тандемных повторов.

- Определен наилучший алгоритм по критериям эффективности нахождения паттернов тандемных повторов и быстродействии.

- Разработан алгоритм для выполнения поиска тандемных повторов, включающий методы, использующие суффиксные деревья.

В результате, суффиксные деревья, обобщенные суффиксные деревья, многострочный вариант суффиксных деревьев, можно использовать для решения задач, связанных с вычислительной биологией, в оптимальном пространстве и времени.

За текущий период реализации проекта были подготовлены 3 научные статьи:

1) **Kalendar R**, Karlov GI 2023. Editorial: Mobile Elements and Plant Genome Evolution, Comparative Analyses and Computational Tools, Volume II. ***Frontiers in Plant Science***, 14: 1308536. DOI: 10.3389/fpls.2023.1308536.

https://www.frontiersin.org/articles/10.3389/fpls.2023.1308536/full

WoS IF_{2022}=6.627 Q1;

Scopus 88th percentile

https://www.scopus.com/sourceid/21100313905

2) Belyayev A, **Kalendar R**, Josefiová J, Paštová L, Habibi F, Mahelka V, Mandák B, Krak K 2023. Telomere sequence variability in genotypes from natural plant populations: unusual block-organized double-monomer terminal telomeric arrays. BMC Genomics 24, 572 (2023).

https://doi.org/10.1186/s12864-023-09657-y

WoS IF_{2022}=4.4 Q1;

Scopus 76th percentile

https://www.scopus.com/sourceid/21727

3) **Shevtsov V., Ismailova A., Beldeubayeva Zh., Satybaldiyeva A.,** Nurpeisova A. MLVA as a method of genotyping and algorithms for its implementation using genome-wide data. News of the National academy of sciences of the Republic of Kazakhstan. Physico-mathematical series. Volume 4. № 348 (2023). P. 300-312 https://doi.org/10.32014/2023.2518-1726.235