Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(СПбГУ)

Индекс УДК 577.2.08
Рег. № НИОКТР АААА-А16-116101210020-7

УТВЕРЖДАЮ
Проректор по научной работе
СПбГУ
_____ С.В. Микушев
«19» мая 2022 г.

ОТЧЁТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

ЛАБОРАТОРИЯ «ЦЕНТР БИОИНФОРМАТИКИ И АЛГОРИТМИЧЕСКОЙ БИОТЕХНОЛОГИИ» В СПБГУ
(промежуточный, этап 8)

Руководитель НИР,
профессор,
кандидат биологических наук _____ А.Л. Лапидус

подпись, дата 24.11.2022

Санкт-Петербург 2022

СПИСОК ИСПОЛНИТЕЛЕЙ

Отв. исполнитель,
профессор, канд. биол. наук  А.Л.Лапидус
(введение, раздел 2.2, 2.6, заключение)

_____20.11.2022_
подпись, дата

Исполнители:

Доцент, канд. физ.-мат. наук  А.И.Коробейников
(разделы 2.2, 2.3)

_____20.11.2022_
подпись, дата

Ст. науч. сотр., канд. биол. наук  И.А.Александров
(раздел 2.4)

_____20.11.2022_
подпись, дата

Ст. науч. сотр., канд. биол. наук  М.П.Райко
(раздел 2.2)

_____20.11.2022_
подпись, дата

Ст. науч. сотр., канд. физ.-мат. наук  А.Д.Пржибельский
(введение, раздел 2.1, заключение)

_____20.11.2022_
подпись, дата

Ст. науч. сотр., канд. физ.-мат. наук  А.А.Михеенко
(раздел 2.1, 2.4)

_____20.11.2022_
подпись, дата

Науч. сотр.  Д.Ю.Антипов
(раздел 2.2)

_____20.11.2022_
подпись, дата

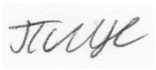
Мл. науч. сотр. О.А.Кунявская
(раздел 2.4)

_____20.11.2022_
подпись, дата

Мл. науч. сотр. Д.А.Мелешко
(раздел 2.2, 2.3)

_____20.11.2022_
подпись, дата

Мл. науч. сотр. И.Н.Толстоганов
(раздел 2.3)



_____ 20.11.2022_
подпись, дата

Мл. науч. сотр. С.Д. Очкалова
(раздел 2.3)

_____ 20.11.2022_
подпись, дата

Программист Т.Е.Дворкина
(раздел 2.4)

_____20.11.2022_
подпись, дата

Специалист Д.Шафранская
(раздел 2.1)

_____20.11.2022_
подпись, дата

Нормоконтролер Н.В.Квадрициус

_____20.11.2022_
подпись, дата

РЕФЕРАТ

Отчет 44 с., 9 рис., 3 табл., 48 источн.

Ключевые слова: БИОИНФОРМАТИКА, СБОРКА ГЕНОМА, МЕТАГЕНОМИКА, ТРАНСКРИПТОМИКА, АНАЛИЗ ДАННЫХ СЕКВЕНИРОВАНИЯ, БИОЛОГИЧЕСКИ АКТИВНЫЕ СОЕДИНЕНИЯ, МАСС-СПЕКТРОМЕТРИЯ

Объектом исследования лаборатории являются геномные, транскриптомные и метаболомные данные, полученные при исследовании вирусов, микро- и макроорганизмов и микробных сообществ. Цель работы – создание новых алгоритмических подходов и программных инструментов для анализа и обработки больших объемов биологических данных.

Результатами данного этапа стало создание новых вычислительных методов и программных продуктов, а также существенная переработка и расширение функциональности созданных ранее инструментов. В 2022 году были созданы программы VerityMap и HORmon (для работы со сборками генома человека) и программа NPOmix (для работы с геномными и метаболомными данными). Также в результате активной разработки были выпущены новые версии широко используемых в научном сообществе программных пакетов SPAdes, IsoQuant, Patheacer и Nerpa. Значимость выполненных работ подтверждается публикацией результатов исследований в высокорейтинговых международных журналах и высоким уровнем цитирования уже в первый год выхода статей. Создаваемые в лаборатории программы используются в широком спектре генетических исследований в биологии, сельском хозяйстве, медицине и других областях.

Доклады о созданных программных продуктах были представлены на ведущих международных конференциях по биоинформатике.

В 2022 году сотрудники лаборатории продолжили совместную работу в рамках международных консорциумов: «The Telomere-to-Telomere consortium», «The Long-read RNA-seq Genome Annotation Assessment Project», «Serratus», «SEQC2 consortium», «Critical Assessment of Metagenome Interpretation (CAMI) initiative», а так приняли участие в научно-исследовательских проектах совместно с учеными ведущих университетов мира. В

результате их сотрудничеств был опубликован целый ряд статей в ведущих мировых журналах.

В отчетном году расширилась вовлеченность в образовательные процессы СПбГУ, включающая участие сотрудников лаборатории в преподавании в рамках магистерской программы «Биоинформатика», межуниверситетского курса «Bioinformatics Algorithms», в образовательной программе профессиональной переподготовки «Биоинформатика».

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящем отчете о НИР применяют следующие термины и сокращения с соответствующими определениями:

- БАС — биологически активные соединения природного происхождения;
- БГК — биосинтетический генный кластер;
- Геномные карты — способ получения геномной информации, когда для молекулы ДНК или ее фрагментов определяется расстояние между позициями, в которых встречается, например, некоторый заданный мотив. Геномные карты разделяют по способу получения на физические (рестрикционные), оптические и электронные;
- Граф де Брюйна — структура данных, используемая для хранения информации о прочтениях, и являющаяся базовой структурой данных для многих геномных сборщиков;
- Граф сборки — граф, используемый алгоритмом в процессе сборки. Основной особенностью такого графа является то, что его ребрам соответствуют нуклеиновые последовательности. В настоящем отчете, как правило, под графом сборки подразумевается упрощенный граф де Брюйна;
- Граф персон — граф, в котором каждая вершина исходного графа представлена серией реплик, называемых персонами. Каждая из персон представляет собой экземпляр вершины в локальном сообществе, к которому она принадлежит;
- ДНК — дезоксирибонуклеиновая кислота;
- Изоформа — вариант сплайсинга гена у эукариот;
- кб — килобаза (тысяча парных оснований);
- Контиг — последовательность, выдаваемая сборщиком. Как правило, соответствует фрагменту или целой последовательности исходной молекулы ДНК (РНК);
- Масс-спектр — зависимость интенсивности ионного тока (количества вещества) от отношения массы к заряду (природы вещества);
- Мономер (в контексте тандемных повторов) — последовательность нуклеотидов, представляющая собой единичный блок для построения повторов более высокого порядка. В центромерных регионах мономерами являются альфа-сателлиты;
- Мономер (в контексте БАС) — единичный строительный блок БАС, например, аминокислота для БАС класса НРП; при биосинтезе БАС, каждый модуль геномного

кластера отвечает за добавление одного мономера в итоговую структуру соединения;

- н.к. — нуклеотид, нуклеотидов;
- НРП — нерибосомный пептид;
- п.н. — пара нуклеотидов. Используется как единица измерения длины нуклеотидной последовательности;
- ПЦР — полимеразная цепная реакция;
- РНК — рибонуклеиновая кислота;
- рРНК — рибосомальная РНК;
- Скаффолдинг — объединение контигов в более длинные фрагменты;
- СММ — скрытые марковские модели;
- Тандемный повтор — последовательности почти идентичных повторяющихся фрагментов ДНК, следующие непосредственно друг за другом в геноме;
- Центромера — часть хромосомы, которая связывает сестринские хроматиды и обеспечивает правильное движение хромосом во время деления клетки;
- *de novo* — заново, с нуля (лат.)
- e-Value — значение статистической значимости прикладывания профильной скрытой марковской модели на последовательность. Чем выше e-Value, тем меньше вероятность, что прикладывание произошло случайно;
- F-мера, F1 — критерий, объединяющий в себе точность и полноту, среднее гармоническое этих двух метрик;
- HOR — повтор высокого порядка (англ. Higher-Order Repeat);
- Illumina — технология секвенирования второго поколения;
- ONT — Oxford Nanopore, технология секвенирования третьего поколения;
- PacBio — Pacific Biosciences, технология секвенирования третьего поколения.

СПИСОК РИСУНКОВ

СПИСОК ТАБЛИЦ

1 ВВЕДЕНИЕ

Основными направлениями деятельности лаборатории «Центр биоинформатики и алгоритмической биотехнологии» СПбГУ (ЦБАБ) являются исследования в области геномики, метагеномики, транскриптомики и метаболомики. Значительная часть проектов лаборатории направлена на разработку алгоритмов и создание компьютерных программ для анализа данных секвенирования различной природы, а также данных масс-спектрометрии.

Результатами научной деятельности лаборатории являются новые вычислительные методы и программные продукты, позволяющие осуществлять различные виды анализа биологических данных. Разработанное программное обеспечение используется биологами и биоинформатиками всего мира для анализа геномов и транскриптомов отдельных организмов и микробных сообществ, а также упрощает задачу поиска новых антибиотиков и производящих их организмов. Сотрудники лаборатории также принимают непосредственное участие в конкретных биологических исследованиях, в которых требуется реализация новых или адаптация уже имеющихся вычислительных методик анализа.

Одним из важнейших факторов, позволяющих регулярно создавать качественное и востребованное программное обеспечение, является налаженное научное сотрудничество с различными лабораториями и институтами — мировыми лидерами в области биоинформатики. Такие контакты дают возможность лучше получать доступ к самым современным данным, решать действительно актуальные на сегодняшний день задачи и применять разработанные алгоритмы в передовых исследованиях. В 2022 году сотрудники лаборатории продолжили свое участие в таких крупных международных консорциумах, как The Telomere-to-Telomere consortium, The Long-read RNA-seq Genome Annotation Assessment Project, SEQC2, Serratus,.

В 2022 году ключевыми темами НИР в лаборатории являлись разработка новых алгоритмических подходов к анализу данных секвенирования ДНК и РНК, а также в других смежных направлениях, включая поиск антибиотиков и других биологически активных соединений.

Были решены задачи по следующим научным направлениям:

- анализ транскриптомных данных третьего поколения;
- метатранскриптомика;
- анализ метагеномных данных;

- поиск антибиотиков и других биологически активных соединений природного происхождения;
- анализа повторных регионов в новых сборках геноме человека;
- разработка алгоритмов для анализа графов сборки;
- развитие и поддержка инструментов на базе геномного сборщика SPAdes.

Проводимые научные исследования тесно связаны с образовательной деятельностью и организацией мероприятий. Так, в 2022 году состоялся второй выпуск магистров программы Биоинформатика, прочитан межуниверситетский курс «Bioinformatics Algorithms, создана программа профессиональной переподготовки «Биоинформатика».

2 ОСНОВНАЯ ЧАСТЬ ОТЧЕТА О НИР

2.1 Анализ транскриптомных и метатранскриптомных данных

В рамках транскриптомного направления был разработан вычислительный конвейер для совместного анализа метагеномных и метатранскриптомных данных секвенирования второго поколения. Данный инструмент основан на программах rnaSPAdes и metaSPAdes, которые были созданы в лаборатории на предыдущих этапах. Основная идея разработанного конвейера — использование метагеномной сборки для восстановления транскриптомных последовательностей из того же микробного сообщества. По результатам тестирования на симулированных и реальных данных метагеномного и метатранскриптомного секверннирования, было показано, что разработанный инструмент позволяет значительно увеличить количество полностью собранных транскриптов (до 2.5 раз), а также практически полностью избавиться от ошибочных последовательностей, присутствующих в изначальной сборки. Разработанный конвейер выложен в публичный доступ.

Также были проведены работы по усовершенствованию сборщика rnaSPAdes, который позволяет осуществлять сборку эукариотических транскриптомов по данным Illumina. Основной идеей данного подпроекта стала кластеризация ребер графа сборки при помощи дополнительных данных, таких как, например, данные секвенирования третьего поколения или последовательности референсных транскриптов для родственных организмов. Был проведен обзор существующих методов поиска пересекающихся сообществ в графе. Для реализации была выбрана методика построения графа персон, в котором ребру из графа

сборки сопоставляется одна или несколько вершин. Таким образом, при кластеризации одно и то же ребром может попасть в несколько кластеров. Полученные кластера используются в модифицированном алгоритме продления путей. Для кластеризации используются связи, которые могут быть получены путем выравнивания дополнительных данных на граф сборки. Веса связей зависят от точности каждого определенного типа данных. Именно в комбинировании различных типов данных на ребрах одного графа и заключается универсальность разработанного подхода. В данный момент ведется подготовка публикации, описывающая разработанные алгоритмы. Ввиду того, что вслед за геномным секвенированием, в транскриптомике на смену технологии Illumina приходят технологии третьего поколения, в 2022 году фокус данного направления был смещен в сторону разработки методов для анализа данных, полученных на платформах Pacific Biosciences и Oxford Nanopore.

В рамках участия в международных консорциумах SEQC и LRGASP была доработана программа IsoQuant, которая позволяет определять новые гены и изоформы в эукариотических геномах по данным секвенирования РНК третьего поколения (Oxford Nanopore и Pacific Biosciences). За отчетный период в IsoQuant был реализован новый режим, позволяющий находить новые гены без существующей геномной аннотации. Основной идеей алгоритма для восстановления транскриптома является построение графа интронов по выровненным ридам и его последующая фильтрация. В предыдущей версии программы фильтрация осуществлялась при помощи набора известных генов. За отчетный период алгоритм был значительно переработан таким образом, что фильтрация осуществляется непосредственно по данным секвенирования и не требует дополнительной информации. По сравнению с аналогичными методами, IsoQuant показал высокую полноту и точность в определении новых изоформ как при наличии геномной аннотации (Таблица 2.1) так и без нее (Таблица 2.2). В результате данной работы в публичный доступ была выложена версия IsoQuant 3.0 (https://github.com/ablab/IsoQuant), а соответствующая статья была принята в печать в международный рецензируемый журнал Nature Biotechnology, IF = 68.19 [12].

В рамках сотрудничества с Weill Cornell Medicine (Нью-Йорк, США) при помощи IsoQuant были исследованы транскриптомные данные, полученные из РНК в ядрах человеческих клеток новой методикой. Статья, описывающая эту методику подготовки библиотек и последующий анализ данных, опубликована в международном журнале Nature Biotechnology, IF = 68.19 [13].

	Программа	TALON	FLAIR	Bambu	StringTie	**IsoQuant**
Все транскрипты	Полнота, %	59.0	70.6	80.0	81.1	**85.0**
	Точность, %	12.7	51.6	67.8	75.2	**95.6**
	F1	0.21	0.60	0.73	0.78	**0.90**
Новые транскрипты	Полнота, %	44.0	38.5	1.0	51.2	**62.6**
	Точность, %	1.6	8.6	69.9	29.6	**86.3**
	F1	0.03	0.14	0.02	0.38	**0.73**

Таблица 2.1.1. Сравнение разработанного алгоритма (IsoQuant) с существующими аналогами (TALON [10], FLAIR [2], Bambu [11], StringTie2 [1]) на симулированных данных Oxford Nanopore (3% ошибок) при наличии неполной базы референсных транскриптов. Для симуляции был выбран транскриптом мыши (Gencode M26).

Данные	PacBio		ONT R10.4		ONT R9.4	
Программа	StringTie	IsoQuant	StringTie	IsoQuant	StringTie	IsoQuant
Транскриптов	30,318	29,103	36,275	24,287	41,791	23,397
Полнота, %	**80.1**	79.3	**65.1**	58.7	**62.2**	57.5
Точность, %	93.8	**96.7**	63.6	**85.7**	52.8	**87.3**
F1	0.86	**0.87**	0.64	**0.70**	0.57	**0.69**

Таблица 2.1.2. Сравнение разработанного алгоритма (IsoQuant) с аналогом StringTie2 [1]) на симулированных данных PacBio (1.5% ошибок) и двумя типами данных Oxford Nanopore (3% и 15% ошибок соответственно) без базы референсных транскриптов. Для симуляции был выбран транскриптом мыши (Gencode M26).

В процессе разработки IsoQuant был также реализован ряд методов, позволяющих осуществлять детальное сравнение методов секвенирования третьего поколения. В 2022 году проект был завершен, разработанные методы были выложены в публичный доступ (https://github.com/ablab/platform_comparison), а статья описывающая особенности

секвенирования РНК на платформах Pacific Biosciences и Oxford Nanopores опубликована в международном рецензируемом журнале Genome Research, IF = 9.034 [14].

В рамках консорциума SEQC при помощи программы IsoQuant был обработан большой объем транскриптомных данных секвенирования третьего поколения, полученный из различных образцов человеческих тканей. Основной целью консорциума является поиск новых изоформ и генов, а также определение оптимальных методик подготовки библиотек секвенирования. В данный момент ведется подготовка публикации.

Консорциум LRGASP главным образом организован для определения лучших методик продготовки библиотек, секвенирования и обработки транскриптомных данных третьего поколения. Одним из способов сравнения вычислительных методов является симуляция данных. Для этой цели был создан симулятор транскриптомных данных, в основу которого легли существующие симуляторы Trans-NanoSim [15], IsoSeqSim (https://github.com/yunhaowang/IsoSeqSim) и RSEM [16]. Разработанный метод позволяет добавлять искусственные неизвестные изоформы при помощи утилиты SQANTI [17], генерировать профиль экспрессии генов по реальным данным и производить симулированные данные транскриптомного секвенирования, имитирующие платформы Illumina, Oxford Nanopore и Pacific Biosciences. При помощи данной программы было создано два набора симулированных данных (мышь и человек), которые в дальнейшем были использованы консорциумом при оценке разли качества различных вычислительных методов. Разработанный симулятор выложен в публичный доступ (https://github.com/andrewprzh/lrgasp-simulation), симулированные наборы данных опубликованы в базе Synapse (https://www.synapse.org/#!Synapse:syn25683370). Статья, описывающая методики работы консорциума, принята в печать в международном журнале Nature Methods, IF = 47.99 [18].

В целом по направлению за 2022 год:

- Разработан и выложен в публичный доступ вычислительный конвейер для совместного анализа метагеномных и метатранскриптомных данных секвенирования второго поколения;
- Проведено усовершенствование сборщика rnaSPAdes при помощи алгоритмов кластеризации на графе сборки;

- Разработаны алгоритмы для восстановления новых изоформ по данным транскриптомного секвенирования третьего поколения без геномной аннотации. В публичный доступ выложена новая версия IsoQuant 3.0, соответствующая статья принята в печать в журнал Nature Biotechnology, IF = 68.19;
- Статья, посвященная методике анализа ядерных транскриптомов опубликована в международном журнале Nature Biotechnology, IF = 68.19 [13];
- Статья, описывающая особенности секвенирования РНК на платформах Pacific Biosciences и Oxford Nanopores опубликована в журнале Genome Research, IF = 9.034 [14];
- В рамках консорциумов SEQC и LRGASP при помощи IsoQuant был произведен анализ большого массива транскриптомных данных;
- В рамках консорциума LRGASP разарботан универсальный симулятор транскриптомных данных. Статья, описывающая методики работы консорциума, принята в печать в журнале Nature Methods, IF = 47.99.

2.2 Анализ метагеномных данных

2.2.1 Исследование метагеномов почвенных сообществ

В рамках направления, посвященного анализу метагеномных данных различной природы, получаемых с помощью высокопроизводительного секвенирования генов 16S/18S рРНК и ITS и полногеномного метагеномного секвенирования, в 2022 году решалось несколько задач.

Так, созданные ранее в лаборатории программные продукты и вычислительные цепочки были использованы при описании микробиоты почв бореальных лесов Западной Сибири. В рамках этого проекта были получены данные об особенностях и специфике состава микробных сообществ, потенциально приводящих к феномену высокотравья и необычайно высокой скорости разложения растительных остатков (в первую очередь листового опада) в отдельных регионах Западной Сибири - т.н. «черневой тайге».

Анализ данных продемонстрировал, что в ризосфере растений в исследуемых почвах значительно возрастает доля представителей типов Bacilli, Verrucomicrobia, alpha-Proteobacteria и Actinobacteria. Этот вывод также получил подтверждение в

лабораторном эксперименте с выделением бактериальных изолятов с высокой фитостимулирующей активностью, относящихся к pp. *Paenibacillus, Streptomyces, Azospirillum, Pseudomonas, Methylobacterium.*

Кроме того, эксперимент с выращиванием культур редиса (*Raphanus raphanistrum subsp. Sativus*, cultivar «Red light») и пшеницы (*Triticum aestivum* L., cultivar «Lada») в лабораторных контролируемых условиях, проведенный коллаборatorами из Института микробиологии им. С.Н. Виноградского, подтвердил, что микробные сообщества ризосферы отличаются по составу в инокулированной и контрольной почвах и в случае лабораторного эксперимента. В то же время, значительную часть в составе ризосферного микробиома занимают нитрифицирующие археи *Nitrosospaera* и Cand. Nitrosocosmius, которые ранее не обнаруживались в ризосфере (Рис. 2.2.1.).

Проведенный анализ также позволил впервые обнаружить такие некультивируемые бактерии как *Cand. Saccharimonadia* (Patescibacteria) и *Candidatus Udaeobacter* (Verrucomicrobia), а также целый ряд некультивируемых представителей Chloroflexi и Elusimicrobia. Полученные результаты опубликованы в журнале Microorganisms [48].

Анализ полногеномных сиквенсов, полученных в ходе этого эксперимента позволил собрать полный геном потенциально нового штамма азотфиксирующей бактерии *Azospirillum doebereinerae*. Геном собран в семь кольцевых контигов и в настоящий момент проводится их анализ (ни один из них не определяется однозначно как хромосома или плазмида). Для многих видов *Azospirillum* характерно наличие нескольких мегарепликонов, число и размер которых отличается даже у близких видов, поэтому полученные данные представляют интерес для изучения эволюции и закономерностей строения бактериальных геномов.

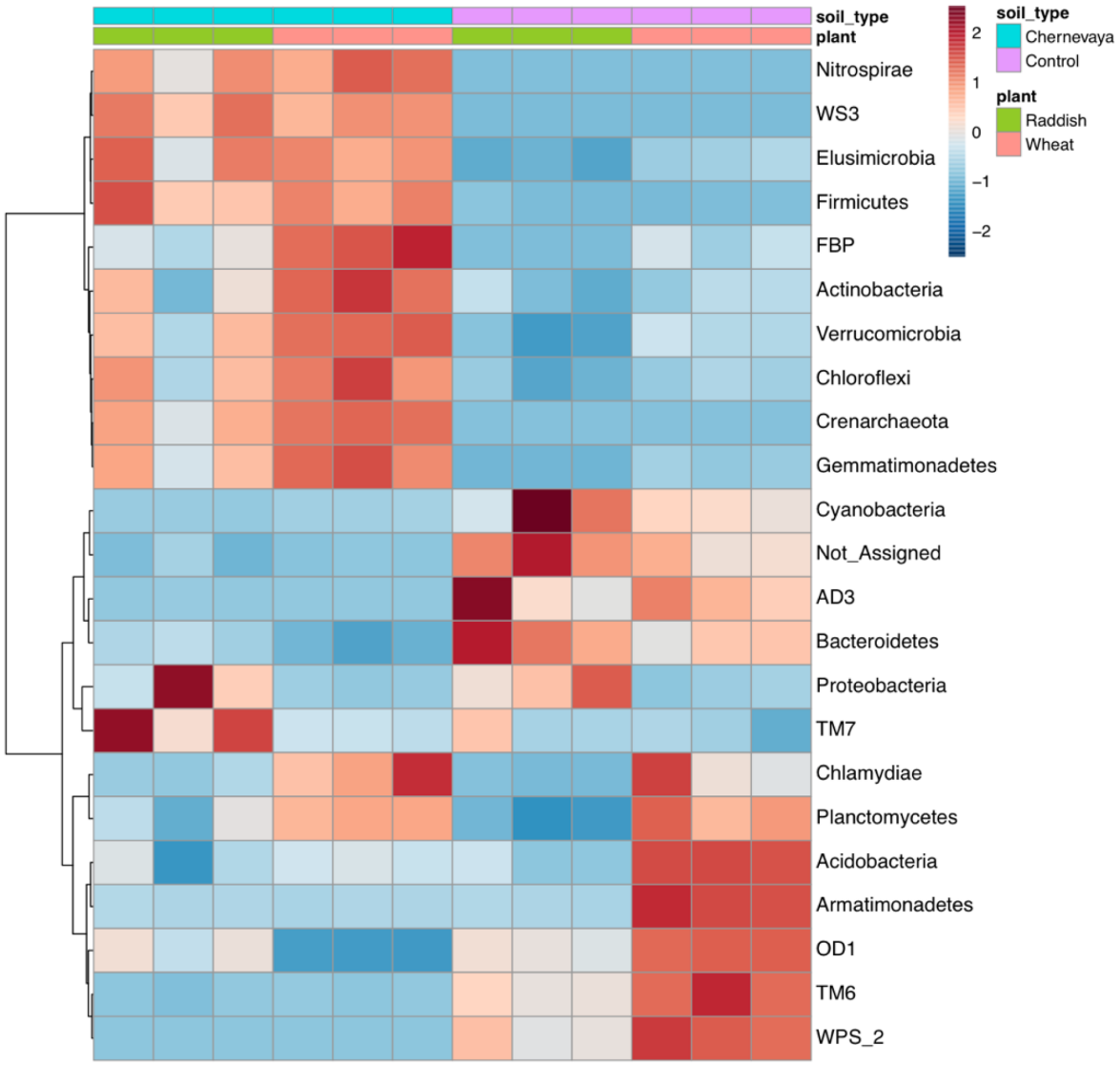


Рис. 2.2.1. Кластеризация ризосферных сообществ почвенных проб на уровне типов. Для кластеризации используется метод Уорда и евклидово расстояние.

Помимо анализа непосредственно данных полногеномного секвенирования в рамках данного направления проводились работы по оптимизации потребления памяти и ускорению геномного сборщика metaSPAdes. Характерной особенностью данных секвенирования метагеномов почв оказалось огромное видовое разнообразие, при этом относительная представленность (покрытие) геномов отдельных представителей оказалось достаточно низким: наибольшее покрытие в подавляющих случаях не превышало 15x.

Как следствие этого, получаемые метагеномные сборки оказывались крайне фрагментированными, что в свою очередь приводило к очень большому потреблению памяти во время работы алгоритмов разрешения геномных повторов в сборщике metaSPAdes (пилотные сборки требовали свыше 400 Гб оперативной памяти).

В 2022 была проведена оптимизация потребления памяти алгоритмов разрешения повторов, внедрены более эффективные структуры данных. Это позволило уменьшить потребление памяти на этапе разрешения повторов более чем в два раза, сократив, например, 400 Гб до 170 Гб. Разработанные изменения были включены в релизы SPAdes 3.15.x.

2.2.2 Поиск новых вирусов в метагеномах кишечника человека

В 2022 году была продолжена ранее проводимая в лаборатории [29,30] работа по анализу вирусов в метагеноме человеческого кишечника. В рамках этой темы был проведен поиск и анализ новых вирусов реалма Varidnaviria, отличительной особенностью которых является особая икосаэдрическая структура капсидного белка, т. н. «double jelly roll» (DJR). В лаборатории были проанализированы более 33 тысяч метагеномных образцов человеческого кишечника, и затем с помощью специфических HMM-профилей были отобраны 218 вирусных геномов DJR-содержащих вирусов.

Затем для найденных бактериофагов был выполнен поиск возможных бактерий-хозяев с помощью поиска соответствующих CRISPR-спейсеров. Наши коллаборатоы (группа Евгения Кунина из NIH) провели дальнейший анализ обнаруженных вирусных геномов, и показали наличие как минимум трёх новых вирусных семейств в пределах порядка Vinavirales (одного из порядков в Varidnaviria) на основании филогенетического дерева по DJR MCP белкам. Полученные результаты опубликованы в журнале Viruses [31].

2.2.3. Метагеномный анализ систем паразит-хозяин

В сотрудничестве с коллегами из Института Цитологии РАН и кафедры зоологии беспозвоночных СПбГУ был проведен анализ нескольких сложных датасетов - геномов простейших, их внутриклеточных эукариотических и бактериальных паразитов и симбионтов. Секвенирование выполнялось методом «single cell» (MDA), полученные данные представляли собой метагеномы с крайне неравномерным покрытием. Сборщик SPAdes

изначально был разработан для данных такого типа, и нам удалось собрать метагеномы, выполнить биннинг и для каждого исследуемого организма получить однокопийные ортологи для последующего филогеномного анализа.

В ходе анализа полученных результатов был собран геном нового вида свободноживущей амёбы *Balamuthia spinosa*, и было показано, что она является новым вектором для человеческого патогена *Legionella pneumophila* в окружающей среде, что может иметь значение для медицины и здравоохранения. Полученные результаты опубликованы в журнале Parasitology Research [45].

2.2.4. Разработка методов оценки качества обработки метагеномных данных

Совместно с международным консорциумом Critical Assessment of Metagenome Interpretation (CAMI) были разработаны и описаны методы оценки качества интерпретации метагеномных данных разной сложности, полученных из коротких и длинных прочтений. Были оценены 76 программных инструментов для разных этапов обработки данных. Результаты анализа и рекомендации по использованию вычислительный инструментов для сборки, биннинга и профайлинга метагеномных данных были опубликованы в журнале Nature Methods, IF=47.99 [35].

В целом по направлению за 2022 год:

- Проведено секвенирование и анализ метагеномов почвенных проб отдельных регионов бореальных лесов России.
- Обнаружены характерные виды бактерий и архей для экосистемы черневой тайги Новосибирской области, потенциально связанные с высокой продуктивностью почвы. Описаны новые некультивируемые организмы
- Проведена оптимизация потребления памяти и времени работы геномного сборщика metaSPAdes.
- Обнаружено три новых семейства DJR-вирусов (Varidnaviria) в пределах порядка Vinavirales
- Проведена сборка, биннинг и анализ геномов простейших (амёб и микроспоридий) из ряда образцов, разделены геномы паразитов и хозяина, обнаружен новый природный резервуар патогенной бактерии *Legionella pneumophil*a.

- В рамках сотрудничества с консорциумом Critical Assessment of Metagenome Interpretation (CAMI) подготовлена и опубликована в журнале Nature Methods (IF=47.99, Q1) статья об оценке качества методов обработки метагеномных данных [35].

2.3 Разработка методов и алгоритмов анализа графа сборки

Для графов сборок метагеномов характерен достаточно большой размер, обусловленный достаточно большим разнообразием геномов, составляющих этот метагеном (так, например, сборка типичного почвенного метагенома может легко включать более тысячи бактериальных и эукариотических геномов и иметь суммарную длину свыше нескольких миллиардов пар оснований).

Однако, часто для дальнейшего анализа требуется сборка не всего метагенома, а лишь отдельных, часто достаточно небольших частей: например только вирусных геномов, плазмид и прочих экстрахромосомных элементов. В отдельных случаях требуется лишь собрать один или несколько интересующих генов.

Традиционный подход к решению вышеобозначенной задачи подразумевает полную сборку метагенома с дальнейшей фильтрацией полученных результатов. Даже с использованием специализированных инструментов, таких как metaSPAdes [20], metaplasmidSPAdes [21] и metaviralSPAdes [22], сборка больших метагеномов может потребовать достаточно больших вычислительных и временных ресурсов, тем самым делая задачу труднорешаемой. Альтернативный подход, предусматривающий фильтрацию отдельных прочтений, как правило, не в состоянии выделить прочтения, покрывающие весь интересующий геном, итоговые сборки отобранных прочтений оказываются неполными и фрагментированными.

Для решения этой задачи были разработаны методы локальной таргетированной пересборки, не требующие с одной стороны сборки полного метагенома, но позволяющие выделить интересующие фрагменты графа сборки.

Подход основан на построении вероятностного графа де Брюйна [25] из исходных прочтений. Предусмотрено два варианта структур данных для хранение графа: считающий фильтр Блума [28] и считающий фильтр остатков [27]. Обе эти структуры данных достаточно хорошо зарекомендовали себя для хранения спектра k-меров максимально эффективным

образом. При этом, фильтр Блума обеспечивает достаточно малое время запроса присутствия отдельного k-мера, но при этом потребляет больше памяти по сравнению со считающим фильтром остатков, имеющим противоположные характеристики. Для оценки числа уникальных k-меров используется вероятностный алгоритм HyperLogLog [26]. Одно из основных отличий построенного вероятностного графа де Брюйна от ранних алгоритмов заключается в хранении также покрытия каждого k-мера.

Построенный граф далее используется для выделения компоненты, содержащей k-меры из последовательности-затравки (например, генома близкого вида или же интересующей последовательности гена). От стартовых k-меров затравки осуществляется обход графа с ограничением по глубине, а также с фильтрацией по покрытию (таким образом осуществляется эффективное отсечение ложных переходов). Наконец, выделяются прочтения, содержащиеся в полученной компоненте, а также строится обычный граф де Брюйна по k-мерам из выделенной компоненты.

Реализованный подход NFilter отличается предсказуемым потреблением памяти (8-16 бит на каждый уникальный k-мер), реализован на основе кодовой базы сборщика SPAdes с переиспользованием эффективных алгоритмов и структур данных. Алгоритмы распараллелены с использованием технологии OpenMP.

NFilter был апробирован на выделении и сборке DJR (double jelly roll) вирусных геномов из данных SRA. Исходные затравочные последовательности были взяты из [24]. Коллаборация Serratus из ~5.5 млн библиотек секвенирования, депонированных в SRA, отобрала 1235, содержащие последовательности схожие с DJR MCP (major capsid protein). Далее, на вычислительном кластере СПбГУ была осуществлена таргетированная сборка полученных библиотек. Из полученных сборок были отобраны 6751 последовательностей, длиннее 1500 пар оснований, имеющих схожесть с MCP, и откластеризованы по уровню схожести 70% посредством mmseq linclust [23]. В результате этой процедуры было получено 3968 уникальных последовательностей-представителей каждого кластера.

Далее, полученные последовательности были попарно выровнены друг на друга и сгруппированы по отсечке e-value в 10^{-3}. Полученная корреляционная картина показана на рис. 2.2. При этом цветом выделены последовательности уже известных групп DJR фагов, а розовым – новые. Как видно, анализ данных позволил увеличить число известных последовательностей DJR фагов в ~20 раз. Также, возможно, была обнаружена новая группа DJR фагов, неизвестная ранее.

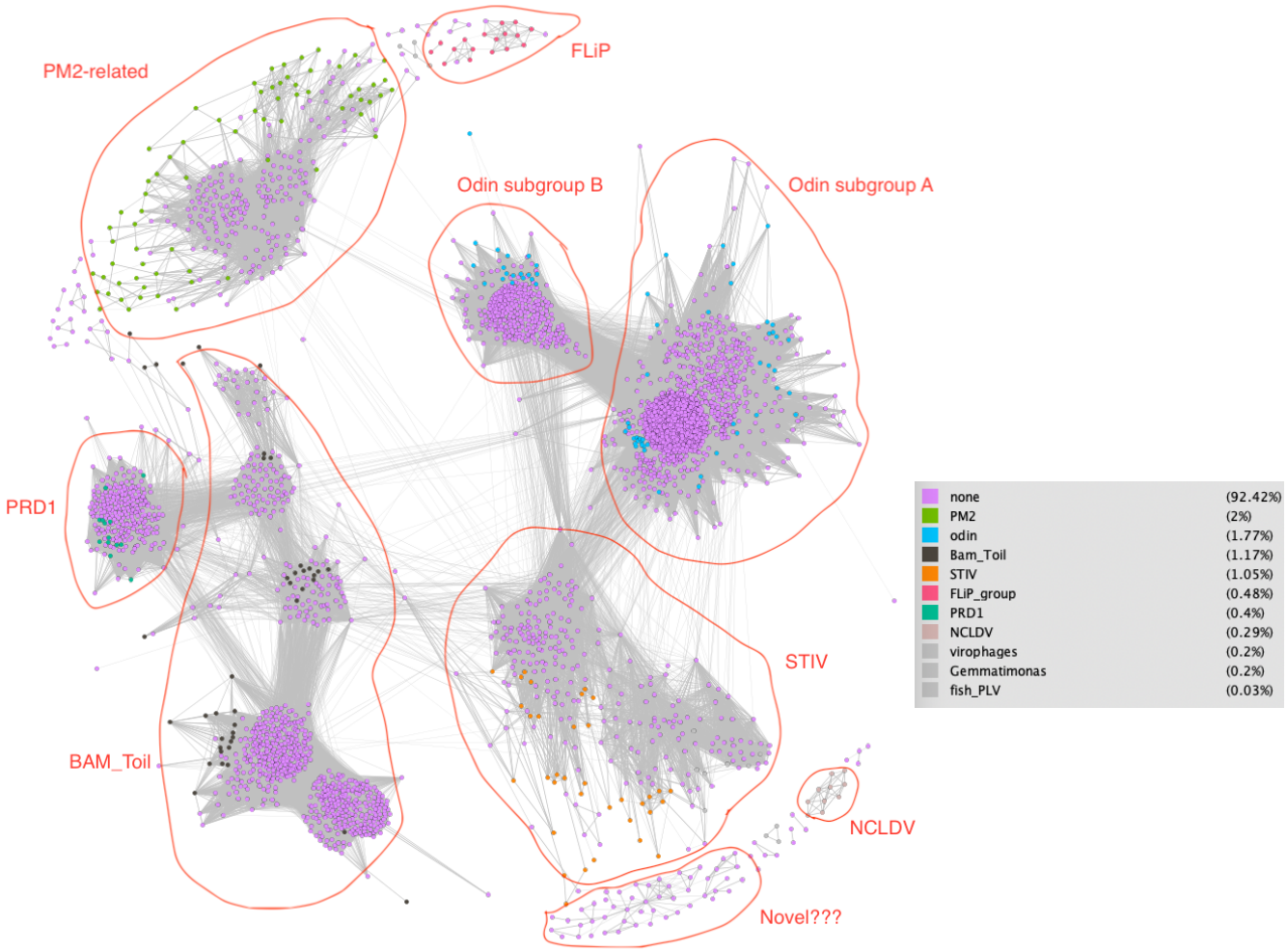


Рис. 2.3.1. Визуализация корреляционных связей последовательностей MCP. Цветом, отличным от фиолетового показаны известные последовательности семейств DJR MCP.

Ранее в лаборатории был разработан инструмент PathRacer, позволяющий прикладывать профильные скрытые марковские модели к графам сборки. PathRacer послужил основой для нескольких вычислительных цепочек, таких как Orfograph и GraphAMR, однако поддерживал только графы сборки, полученные из сборщика SPAdes (или MEGAHIT): предполагалось, что граф сборки представляет собой полноценный граф де Брюйна с одинаковым перекрытием последовательностей ребер в вершинах.

Данное предположение очевидным образом не выполнялось для графов сборки, получаемых из других сборщиков, в частности, сборщиков длинных прочтений Flye и minigraph.

В отчетном периоде в инструменте PathRacer была реализована прослойка т.н. «курсора», позволяющая абстрагироваться от вида конкретного графа, и делающая алгоритм PathRacer универсальным: теперь поддерживаются как графы сборки типа графов де Брюйна, так и другие графы сборки с (возможно) произвольным, или даже отсутствующим перекрытием в вершинах. Также при помощи реализованной абстракции курсора удалось исправить некоторые проблемы с точностью фрагментированных прикладываний, что улучшило результаты выделения генов, обуславливающих антибиотикорезистентность, в вычислительной цепочке GraphAMR, разработанной ранее. Обновленная версия инструмента PathRacer доступна в репозитории Conda.

В целом по направлению:

- Разработаны алгоритмы и подходы для локальной таргетированной пересборки последовательностей генов, профагов и т.п. в сложных метагеномах без сборки всего метагенома
- Утилита PathRacer доработана на предмет использования графов сборки, полученных сборщиками Flye и minigraph
- Улучшено восстановление генов обуславливающих антибиотикорезистентность из графов сборки метагеномов

2.4 Разработка алгоритмов для анализа повторных регионов в геноме человека

2.4.1 Алгоритмы для выравнивания длинных прочтений в повторных регионах

В рамках сотрудничества с консорциумами Telomere-to-Telomere (T2T) и Human Pangenome Reference Consortium (HPRC) была разработана программа VerityMap для высокоточного выравнивания длинных прочтений на сборку, поиска ошибок и гетерозиготных вариантов в сборке на основе полученных выравниваний. Программа была использована для анализа новых сборок генома человека и подтвердила свою высокую эффективность и точность прикладывания в повторных регионах. Также, в отличие от предыдущей программы TandemMapper, разработанной прежде всего для анализа повторных регионов, VerityMap позволяет осуществить прикладывание длинных прочтений и

проанализировать полный геном человека менее чем за сутки. Это стало возможным благодаря разработке новых высокопроизводительных алгоритмов и оптимизации программы для многопоточной работы. Значительно был переработан первый этап алгоритма: создание индекса k-меров для сборки и прочтений. В VerityMap для этой цели используется высокоэффективный фильтр Блума, позволяющий сократить потребление памяти для хранения индекса. Выбранная структура данных позволяет быстро оценить частоту встречаемости каждого k-мера в сборке и прочтениях и создать массив «надежных» k-меров, которые будут использованы затем в процессе прикладывания.

Для каждого отдельного прочтения VerityMap хранит набор «надежных» k-меров и на их основе строит граф совместимости, где набор вершин состоит из k-меров, общих для прочтения и сборки. Результатом прикладывания является самый длинный путь в графе. Веса ребер графа определяются в зависимости от частоты k-меров, а также разницы в расстояниях между этими k-мерами в прочтении и в сборке. Данная структура позволяет находить ошибки сборки на основе анализа разницы в расстояниях между надежными k-мерами, даже в высокоповторных регионах.

Было произведено сравнение VerityMap с наиболее популярными алгоритмами прикладывания minimap2 [3] и Winnowmap2 [5] на новых сборках генома человека, полученных в рамках консорциума T2T [7], а также созданных с помощью новых алгоритмов сборки, таких как программа LJA [34]. Результаты показали, что minimap2 и Winnowmap2 значительно уступают в точности в прикладывании прочтений к повторным регионам, особенно в присутствии ошибок сборки.

Алгоритм VerityMap также был значительно усовершенствован для учета особенностей диплоидных сборок и позволяет эффективно находить в них ошибки, в том числе ошибки при неправильном разрешении гаплотипов, и гетерозиготные варианты (Рис. 2.4.1). Существующие алгоритмы прикладывания не позволяли находить ошибки в сложных или повторных регионах сборок диплоидных геномов. Это особенно важно с учетом роста объема соответствующих данных и создания новых алгоритмов для сборок диплоидных геномов.

Рис. 2.4.1. Визуализация прикладываний VerityMap, minimap2 и Winnowmap2 в геномном браузере IGV [6]. С помощью данных программ было выполнено прикладывание длинных прочтений на симулированную диплоидную сборку центромеры X хромосомы человека. Как видно, VerityMap, в отличие от двух других программ, показывает ошибку в месте неправильного разрешения гаплотипов. В то же время, minimap2 и Winnowmap2 прикладывают прочтения в неправильные места генома, что подтверждается большим количеством замен по сравнению с референсным геномом.

Статья, описывающая алгоритм и результаты работы программы VerityMap, принята к публикации в журнале Genome Research (IF = 9.043).

2.4.2 Алгоритмы для аннотации центромерных регионов

Помимо разработки инструментов для анализа произвольных повторных регионов, была продолжена работа над проектом по усовершенствованию программных продуктов для автоматического анализа центромерных участков в геноме человека. Центромерные участки генома человека представляют собой повторы высокого порядка (higher-order repeats, HORs). На Рисунке 2.4.2 представлена структура центромеры хромосомы X: её последовательность состоит из повторов высокого порядка H, в свою очередь каждый такой повтор состоит из блоков меньшего размера, называемых альфа-сателлитами или мономерами (длина ~171 п.н.).

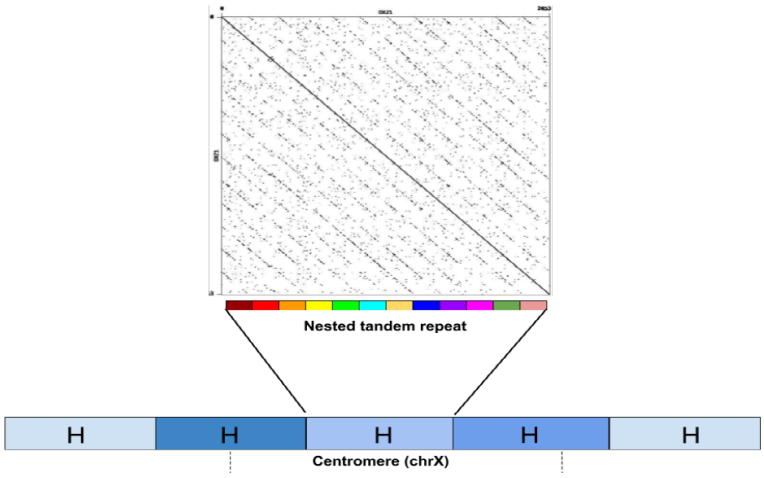


Рис. 2.4.2 Структура центромеры хромосомы X. Центромера хромосомы X состоит из повторяющихся последовательностей H длины ~2 килобазы, каждая из которых состоит из 12 мономеров длины 171 п.н. Сходство между двумя последовательностями H порядка 98-100%, в то время как сходство между двумя мономерами порядка 80-90%.

До недавних пор полные последовательности центромер были неизвестны. В рамках работы Telomere-to-Telomere (T2T) Консорциума были получены первые сборки последовательностей центромер человека [7], что позволило приступить к их детальному анализу. Большое количество данных требует разработки автоматических решений для анализа. В нашей лаборатории в рамках сотрудничества с T2T Консорциумом был разработан новый программный продукт HORmon [8], который принимает на вход последовательности центромерных участков (целые сборки или набор прочтений) и производит полную аннотацию данных участков. HORmon использует часть решений, предложенных в инструменте CentromereArchitect [9], такие как генерация первичных последовательностей мономеров, поиск гибридных мономеров и идентификацию немономерных участков. В свою очередь модуль для генерации повторов высокого порядка был полностью переработан. В него были внесены изменения в соответствии с текущими представлениями биологического сообщества об устройстве и эволюции центромер.

На основе современных представлений об эволюции центромер в рамках работы над HORmon, был разработан Постулат Эволюции Центромер (Centromere Evolution Postulate). Также была разработана новая концепция представления центромеры в виде графа (Рис. 2.4.3), которая позволила представить центромерные последовательности в более компактном

виде. Предложена новая концепция разбиения центромерной последовательности на повторы высокого порядка.

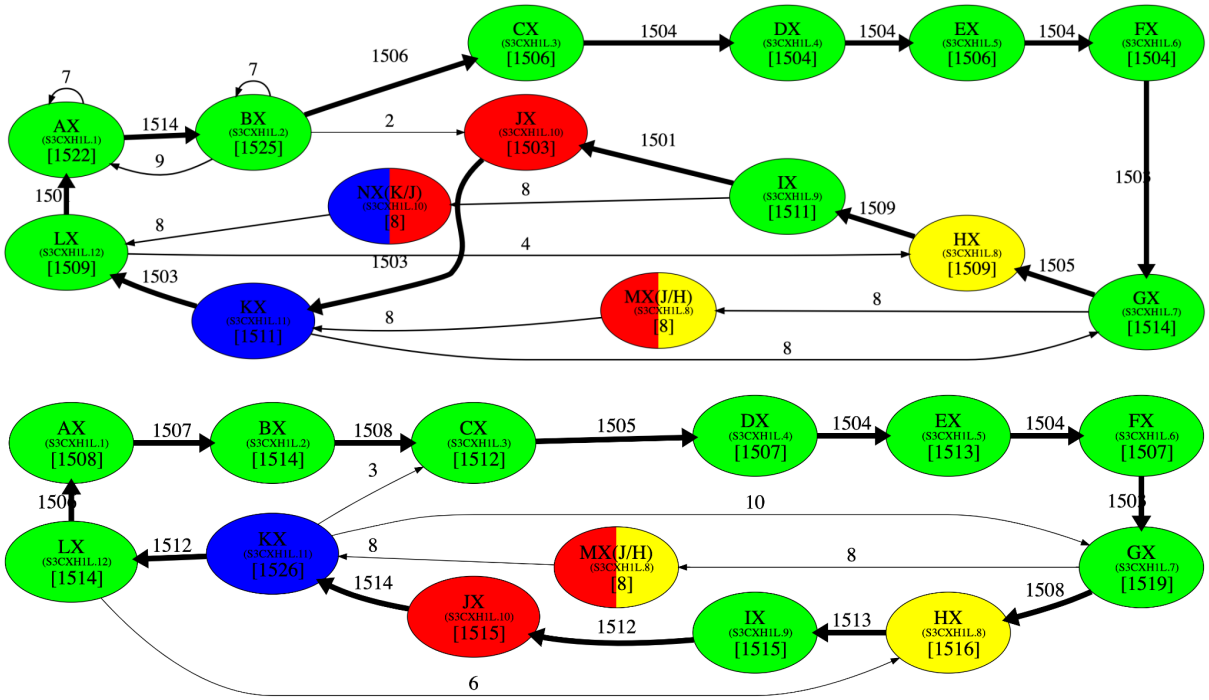


Рис. 2.4.3. Представление последовательности центромеры в виде графа для геномов CHM13 (вверху) и HG002 (внизу), полученных в рамках T2T Консорциума. Каждая вершина представляет собой мономер, вес ребра соответствует количеству раз, которое в последовательности центромеры два мономера идут друг за другом. Гибридные мономеры (мономеры, образованные слиянием двух других) показаны двумя цветами.

Инструмент HORmon использовался для разработки первой полуручной аннотации центромерных сборок в проекте по анализу центромер с коллегами из T2T Консорциума, что позволило провалидировать автоматические аннотации, полученные с помощью данного инструмента. На данный момент HORmon является единственным программным продуктом, решающим данную задачу и отвечающим требованиям биологического сообщества.

2.4.3 Алгоритмы для аннотации и сравнения центромерных регионов без получения полных сборок центромер

В рамках работы в сотрудничестве с T2T Консорциумом и HPRC Консорциумом над построением наиболее полного описания центромерных участков хромосом [32], стало понятно, что при автоматической обработке HiFi прочтений центромер различных индивидуумов, можно получить важные результаты о различиях в структуре этих центромер и об их мономерном составе (Рис. 2.4.4). В ходе этой работы HORmon был улучшен для корректной обработки длинных прочтений с низким количеством ошибок, а именно подобраны новые настройки для выделения мономеров и повторов высокого порядка и ускорен алгоритм построения консенсуса повторов высокого порядка.

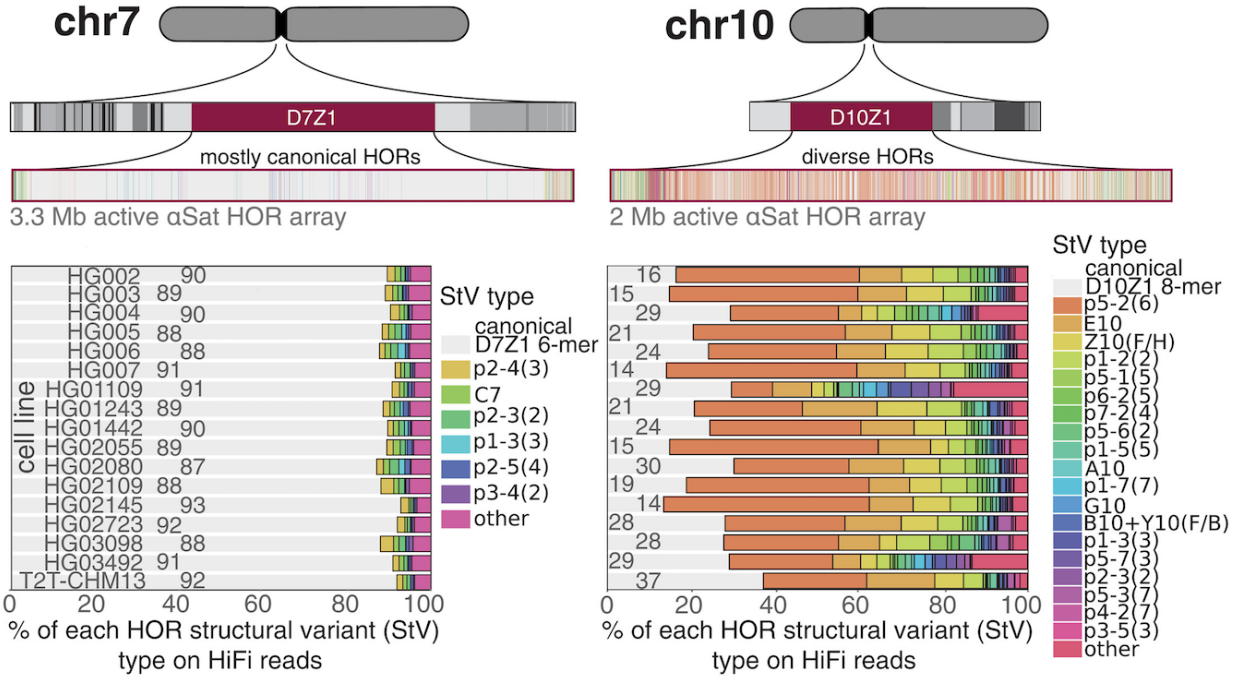


Рис. 2.4.4. Сравнение структур центромер хромосом 7 и 10 различных индивидов с помощью результатов HORmon на длинных прочтениях [32]. Верхние изображения показывают разницу в составе уже собранных центромер 7 и 10 для модельной клеточной линии CHM13, в то время как нижние гистограммы отражают разницу в структуре центромер на основе HiFi прочтений из данных различных индивидов. Каждый столбик соответствует одному индивиду, а количество повторов высокого порядка отражено длиной соответствующей линии в столбике.

Далее в проекте по обработке центромерных участков планировалось разработать алгоритм для выравнивания центромерных сборок, главной задачей которого являлось бы сравнение сборок центромер из геномов различных индивидуумов. Благодаря получившимся с помощью длинных прочтений результатам, было принято решение продолжить работу над улучшением алгоритма HORmon, в частности для длинных прочтений:

- Был полностью изменен модуль построения консенсусов мономеров. Если раньше консенсус каждого мономера считался отдельно по всем прикладываниям данного мономера к прочтениям, то теперь было решено строить консенсус пар подряд идущих мономеров, так нам удалось учесть нуклеотиды, которые часто пропадают на концах мономеров. Благодаря этому мономеры, сгенерированные HORmon оказались более точными, чем изначальные мономеры, сгенерированные в полуручном режиме. Данные результаты были отражены в последних версиях статей [8] и [35].
- Одной из серьезных проблем для обработки длинных прочтений HORmon, как и CentromereArchitect, является долгое построение множества изначальных мономеров. Алгоритм кластеризации мономеров в CentromereArchitect представляет собой итеративную процедуру и на каждой итерации CentromereArchitect запускает программу StringDecomposer [36] для оптимального разбиения центромеры на текущее множество мономеров. Был ускорен алгоритм StringDecomposer для работы с большим количеством мономеров. Были проведены эксперименты по замене текущего алгоритма кластеризации на алгоритм HDBSCAN из стандартных библиотек (Таблица 2.4.1), которое показало сравнимое качество кластеризации за гораздо более короткое время работы. Данные изменения будут отражены в ближайшем релизе HORmon.

	HORmon Clustering	HDBSCAN
Время (16 потоков)	00:14:57	**00:01:12**
Память	2862.5 Мб	**176.1** Мб
Количество кластеров	12	12
Silhouette score	0.848	**0.909**

Таблица 2.4.1. Сравнение кластеризации HORmon и алгоритма HDBSCAN на последовательности центромеры X модельной клеточной линии CHM13. Как видно количество кластеров в обоих случаях было определено правильно (12), в то же время по затраченному времени и используемой памяти HDBSCAN выигрывает у кластеризации, реализованной в HORmon. По метрике Silhouette score, оценивающий качество кластеров, алгоритмы показали похожий результат.

В целом по направлению за 2022 год:

- В рамках сотрудничества с консорциумом T2T опубликованы статьи в журналах Science (IF=63.7) [7,32] и Nature Methods (IF=47.99) [33], описывающие создание новой референсной сборки человека.
- Программный продукт VerityMap для высокоточного прикладывания длинных прочтений на сборку реализован и публично доступен на Github (https://github.com/ablab/VerityMap).
- Алгоритм VerityMap адаптирован для корректной работы с диплоидными сборками.
- VerityMap использован для анализа новых сборок генома человека, как полученных в рамках консорциума T2T, так и созданных с помощью новых алгоритмов сборки.
- Статья, описывающая алгоритм VerityMap и результаты работы на сборках генома человека, принята к публикации в журнале Genome Research (IF=9.043).
- Программный продукт HORmon для аннотации человеческих центромер реализован и публично доступен на Github (https://github.com/ablab/HORmon).

- Произведена проверка автоматической аннотации с помощью HORmon на соответствие предыдущим ручным аннотациям и постулату центромерной эволюции.
- По результатам работы над инструментом HORmon опубликована статья «Automated annotation of human centromeres with HORmon» в журнале Genome Research (IF=9.043) [8].
- Последовательности всех центромер представлены в новом виде более удобном для дальнейшего ручного анализа и опубликованы в дополнительных материалах к статье «Automated annotation of human centromeres with HORmon».
- Консенсусы всех повторов высокого порядка и их мономеров извлечены и опубликованы в дополнительных материалах к статье «Automated annotation of human centromeres with HORmon».
- Усовершенствован алгоритм HORmon для анализа архитектуры центромер по длинным прочтениям с низким количеством ошибок, предложен более эффективный и точный алгоритм кластеризации изначальных мономеров и построения их консенсусов
- Построенные с помощью новой версии HORmon консенсусы были использованы в статье [32] в качестве эталонных мономеров центромер человека.

2.5 Разработка методов и программ для поиска антибиотиков и других биологически активных соединений природного происхождения

Биологически активные соединения природного происхождения (БАС), прежде всего бактериального, являются ценным источником промышленно и фармакологически ценных веществ, таких как антибиотики, иммуносупрессоры и токсины [37]. Современные технологии секвенирования и масс-спектрометрии позволяют быстро и относительно дешево получать большие объемы данных о БАС и синтезирующих их организмах. Однако высокопроизводительная обработка этих данных требует специализированных методов и программ. В 2022 году была как усовершенствована ранее созданная программа идентификации БАС класса нерибосомные пептиды (Nerpa), так и разработан новый вычислительный метод сопоставления геномных и метаболомных данных БАС, работающий с различными химическими классами этих соединений (NPOmix). Ниже описаны выполненные работы по каждому из этих направлений.

2.5.1. Улучшение скоринг-функции программы Nerpa

Ранее в лаборатории была создана программа Nerpa для поиска биосинтетических генных кластеров известных БАС класса нерибосомные пептиды (НРП) [38]. Программа принимает на вход геном организма-продуцента БАС и базу химических структур известных соединений класса НРП. Nerpa использует antiSMASH [39] для идентификации в геноме вероятного биосинтетического генного кластера, ответственного за синтез БАС, и rBAN [40] для декомпозиции химических структур НРП на отдельные строительные блоки-мономеры, чаще всего аминокислоты. Ключевой модуль Nerpa сопоставляет мономеры, предсказанные antiSMASH по генному кластеру, с декомпозированными структурами из базы известных НРП и ранжирует сопоставления для нахождения наиболее вероятной пары НРП-кластер.

В оригинальной публикации для ранжирования пар НРП-кластер использовалась эвристическая скоринг-функция, построенная на основе известных биосинтетических генных кластеров из базы MIBiG [41]. В прошлом отчетном периоде был разработан более универсальный подход к оцениванию пар на основе генеративной скрытой марковской модели (СММ) с тремя состояниями – «сопоставление двух мономеров» (совпадения/замены, match/mismatch), «вставка мономера из геномного предсказания» и «вставка мономера из структуры» (инсерция/делеция, insertion/deletion). Параметры обоих подходов были обучены на одинаковом тренировочном наборе данных из базы MIBiG и показали схожие по точности результаты. Однако у обоих скоринг-функций оставались два существенных недостатка: (1) итоговый скор, числовую оценку соответствия пары НРП-кластер, крайне трудно трактовать – у него нет теоретического минимума и максимума и для отсечения «надежных» пар от «ненадежных» приходится использовать эвристическую отсечку, полученную на тестовом наборе данных (скор равный 6.0 [38]), (2) скоры заведомо смещены в сторону более длинных НРП и биосинтетических генных кластеров.

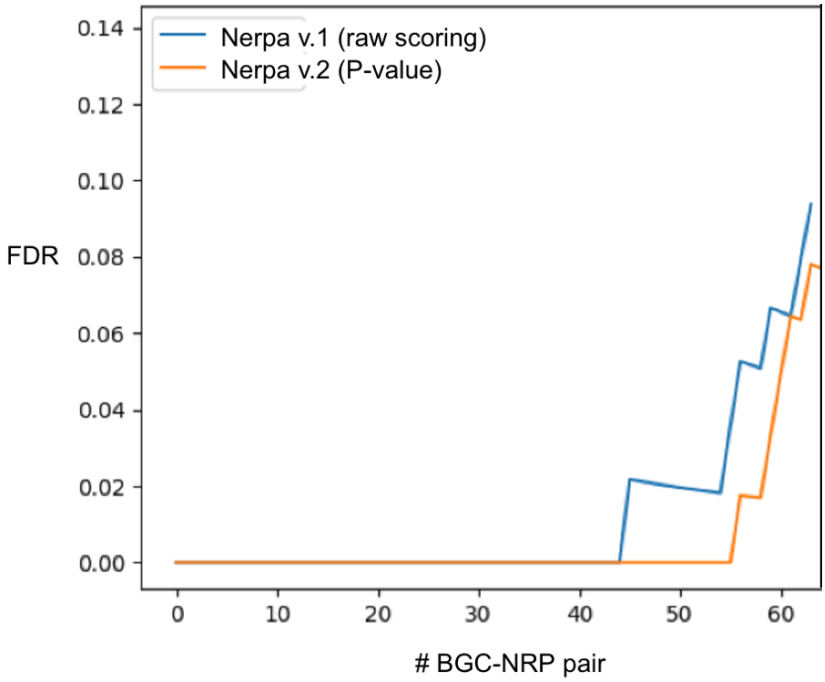


Рис. 2.5.1 Точность работы Nerpa с различными скоринг-функциями на 64 парах НРП-кластер из базы MIBiG. По оси X номера пар НРП-кластер, отсортированных согласно оригинальной (синий, абсолютное значение скора) и новой (оранжевый, P-значение скора) скоринг-функцией. По оси Y доля ложноположительных пар (false discovery rate, FDR).

В отчетном периоде был разработан и апробирован новый подход к оцениванию соответствия пар НРП-кластер, который лишен обозначенных недостатков. Новая скоринг-функция опирается на генеративную СММ, разработанную на прошлом этапе, но использует ее не только для вычисления абсолютного значения скора, но и для оценки статистической значимости (P-значение) данного скора для данной пары НРП-кластер. Полученное P-значение выдается в качестве итого уровня соответствия пары, по которому происходит ранжирование и выбор наиболее вероятного НРП, синтезируемого данным генным кластером. P-значения вычисляется методом Монте-Карло и Выборкой по значимости (importance sampling) по выборке значений скоров для случайно сгенерированных пар НРП-кластер. При этом размеры генерируемых НРП и кластеров соответствуют размеру оцениваемой в данный момент реальной пары НРП-кластер, а скоры вычисляется с помощью генеративной СММ. Таким образом, была убрана явная зависимость скора от размера оцениваемых НРП и кластера, осуществлен переход от абсолютного значения к P-значениям в шкале от 0.0 (лучшее) до 1.0 (худшее) и пользователям предоставлена возможность отсеивать

«ненадежные» пары по желаемому уровню значимости, например, 0.05. На Рис. 2.5.1 показано преимущество нового подхода по сравнению с опубликованным ранее в версии Nerpa v.1. При нулевом уровне ложноположительных идентификаций новая-скоринг функция позволяет найти на 25% больше пар НРП-кластер (55 против 44).

Кроме того в отчетном периоде начата подготовка расширенного тренировочного набора известных пар НРП-кластер из опубликованной в 2022 базы MIBiG v3 [42], содержащей более 800 записей о биосинтетических генных кластерах НРП и НРП-подобных соединений. Планируется использовать этот набор для переобучения параметров генеративной СММ, что позволит улучшить точность работы как непосредственно этого метода, так и построенной на его основе в отчетном периоде скоринг-функции. Улучшения 2022 года в Nerpa публично выпущены на платформе GitHub: https://github.com/ablab/nerpa, там же планируется выпустить версию Nerpa с новой скоринг-функцией после ее апробации на больших объемах данных и переобучения с расширенным тестовым набором известных пар НРП-кластер.

2.5.2. Разработка вычислительного метод сопоставления геномных и метаболомных данных БАС

В отчетном периоде в коллаборации с учеными Калифорнийского университета в Сан-Диего и Университета Сан-Паулу был разработан новый вычислительный инструмент для сопоставления биосинтетических генных кластеров с продуцируемыми ими БАС. Инструмент получил название NPOmix и в его основе лежит метод k-ближайших соседей (k-nearest neighbor algorithm). В отличии от Nerpa, обрабатывающей только БАС класса нерибосомные пептиды, NPOmix может анализировать и другие основные классы БАС, включая поликетиды, терпены и сахариды. Схема вычислительного конвейера NPOmix показана на Рис. 2.5.2.

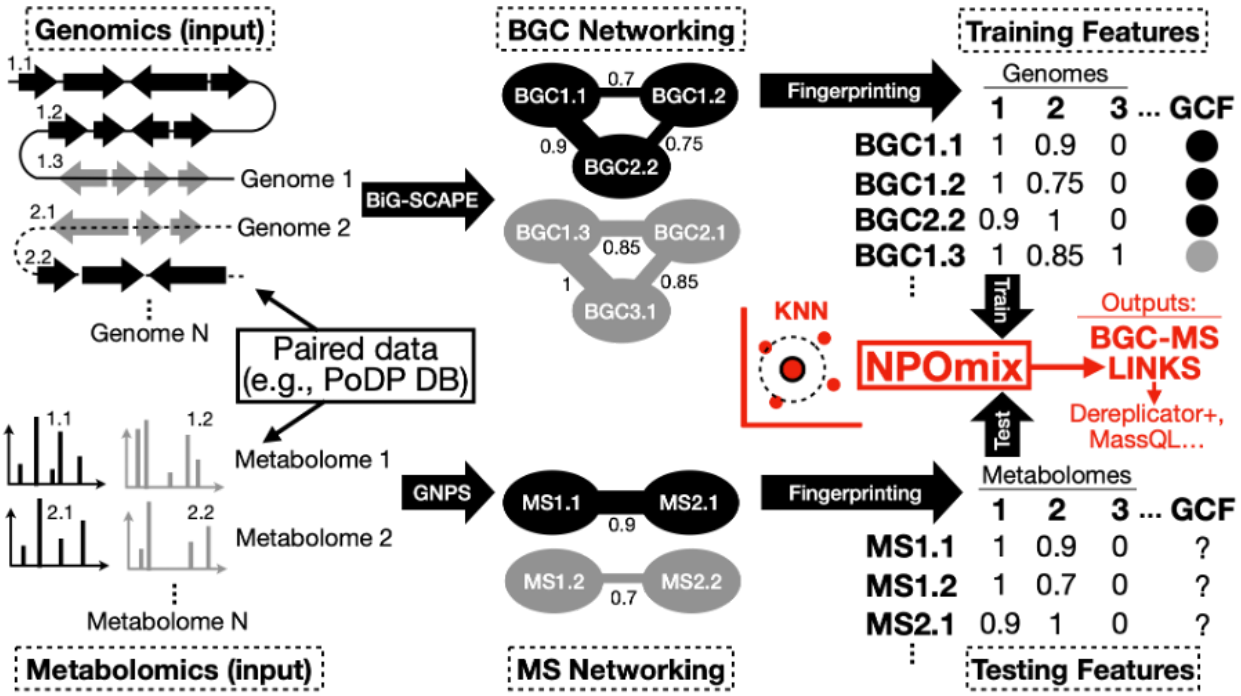


Рис. 2.5.2. Вычислительный конвейер NPOmix (адаптировано из [43]). Входными данными для конвейера являются геномы организмов-продуцентов БАС (левый верхний угол) и масс-спектрометрические замеры метаболома соответствующих образцов (левый нижний угол). В геномах предсказываются биосинтетические генные кластеры (biosynthetic gene clusters, BGC), которые далее объединяются в семейства похожих кластеров (центр верх). Схожие масс-спектры группируются в молекулярные сети (центр низ). В обоих типах групп (геномных и метаболомных) выделяются характерные признаки, которые подаются в NPOmix в качестве тренировочных и тестовых данных (правая часть рисунка). Сопоставляя наборы признаков по методу k-ближайших соседей, NPOmix находит наиболее вероятные пары биосинтетический генный кластер-метаболит (масс-спектр). Ранжированный список пар является выходом NPOmix, который затем может быть проанализирован с использованием специализированных программ для идентификации метаболита по масс-спектру или вероятного продукта синтеза по генному кластеру.

Вычислительный конвейер NPOmix опирается на ряд сторонних инструментов: antiSMASH [39] для предсказания биосинтетических генных кластеров (БГК) во входных геномных данных, BiG-SCAPE [44] для группировки БГК в семейства, молекулярные сети

(GNPS [46]) для кластеризации схожих масс-спектров. Для обучения классификатора использовались известные пары БГК-масс-спектр с платформы PODB [47].

Было проведено сравнение NPOmix с основными программами-конкурентами на 22 валидированных БАС различных классов с доступными геномными, масс-спектрометрическими и структурными данными (Рис. 2.5.3). NPOmix оказался единственным инструментом, который смог обработать все входные данные. NPOmix уступил Nerpa в точности (77.3% против 100%), но значительно опередил ее по полноте результатов (100.0% против 20.0%), остальные программы показали худшие результаты, чем NPOmix по обоим параметрам (лучшая точность 75%, лучшая полнота 25%).

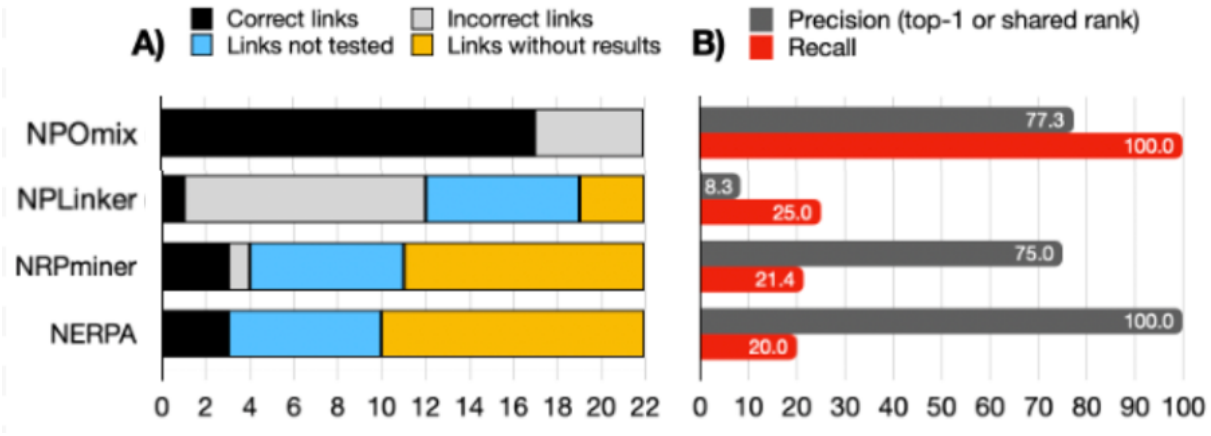


Рис. 2.5.3. Сравнение точности и полноты результатов NPOmix с программами-аналогами на 22 известных парах БГК-БАС (адаптировано из [43]). (A) Количество правильных (черный), неправильных (серый) и неидентифицированных (желтый) сопоставлений БГК-БАС. Программы-аналоги работают только с определенными классами БАС, поэтому часть данных ими не обрабатывалась (голубой). (B) Точность (серый) и полнота (красный) результатов.

Программа NPOmix выпущена в открытом доступе как инструмент командной строки по адресу https://github.com/tiagolbiotech/NPOmix_python. Статья о NPOmix в октябре 2022 принята к печати в международный рецензируемый журнал PNAS Nexus [43] (журнал запущен в 2022 поэтому еще не имеет собственного импакт-фактора, IF головного журнала PNAS = 12.779).

В целом по направлению за 2022 год:

- Разработана новая скоринг-функция для программы Nerpa, на тестовом наборе данных из 64 известных пар нерибосомный пептид-геномный кластер из базы MIBiG показано улучшение результата на 25% при нулевом уровне ложноположительных идентификаций;
- Разработан и выпущен в публичный доступ инструмент NPOmix для сопоставления геномных и метаболомных данных БАС на основе метода k-ближайших соседей;
- Статья об NPOmix принята в печать в рецензируемый международный журнал PNAS Nexus.

2.6 Организация и проведение образовательных и научных мероприятий

В отчетном году лаборатория продолжила проводимую уже много лет образовательную деятельность. Так в 2022 году 145 слушателей первого курса магистратуры прошли курс «Биоинформатика». В рамках межуниверситетского курса «Bioinformatics Algorithms» (https://www.multi-bioalgorithms.org/) продолжилось чтение курса «Алгоритмы в биоинформатике» на факультете Математических и компьютерных наук СПбГУ.

3 ЗАКЛЮЧЕНИЕ

В 2022 году научно-исследовательская работа лаборатории «Центр биоинформатики и алгоритмической биотехнологии» велась по 5 направлениям в области вычислительной биологии: транскриптомика, метагеномика, разработка инструментов для анализа графа сборки, анализ длинных тандемных повторов в новых сборках генома человека, а также поиск антибиотиков и других БАС природного происхождения. В рамках каждого направления основная часть проектов была направлена на создание новых алгоритмов и их реализации. Также сотрудники лаборатории принимали участие в проектах, направленных на решение прикладных биоинформатических задач.

В отчетном периоде было разработано несколько новых программных продуктов, таких как NPOmix, VerityMap, HORmon. Целый ряд вычислительных методов был усовершенствован за счет создания новых алгоритмов, или же адаптирован для поддержки новых типов данных: IsoQuant, rnaSPAdes, metaSPAdes, PathRacer, StringDecomposer, Nerpa и др. Разрабатываемые и поддерживаемые программы используются исследователями во всем мире, включая дружественные научные коллективы, что позволяет получать своевременную обратную связь, а также развивать и совершенствовать вычислительные методы в соответствии с наиболее актуальными в области задачами. Среди научных сотрудничеств лаборатории стоит упомянуть такие передовые исследовательские центры и сообщества как UC Santa Cruz, UC San Diego, Weill Cornell Medicine, NIH, NCBI (все США), а также несколько международных консорциумов: The Telomere-to-Telomere consortium (T2T), The Long-read RNA-seq Genome Annotation Assessment Project (LRGASP), SEQC2, Critical Assessment of Metagenome Interpretation (CAMI), Serratus.

Результаты исследований вошли в 16 статей, опубликованных в рецензируемых международных журналах Scopus/WoS: из них 12 — в журналах первого квартиля и 10 в ведущих специализированных журналах с IF больше 8.0, в том числе Science (IF 63.714), Nature Methods (IF 47.99) и Nature Biotechnology (IF 68.16). По результатам НИР было представлено 2 устных доклада на национальной и международной конференциях. Востребованность создаваемых в лаборатории программных методов подтверждается крайне высоким уровнем цитирования соответствующих публикаций. Статьи, вышедшие в 2022 году, уже были процитированы более 500 раз (по данным Google Scholar), то есть уже активно используются научным сообществом.

Кроме того, большое внимание уделялось вопросу подготовки кадров по направлению «Биоинформатика». Центр организовывал и проводил образовательные и научные мероприятия участвуя в руководстве и проведении магистерской программы «Биоинформатика» на Биологическом факультете СПбГУ и межуниверситетского курса «Bioinformatics Algorithms на Факультете математики и компьютерных наук, программы профессиональной переподготовки «Биоинформатика».

4 СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Kovaka S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2 // Genome biology. – 2019. – Т. 20. – №. 1. – С. 1-13.
2. Tang A. D. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns //Nature communications. – 2020. – Т. 11. – №. 1. – С. 1-12.
3. Li H. Minimap2: pairwise alignment for nucleotide sequences // Bioinformatics. 2018. Т. 34. № 18. С. 3094–3100.
4. Epasto A., Lattanzi S., Paes Leme R. Ego-splitting framework: From non-overlapping to overlapping clusters //Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2017. – С. 145-154.
5. Jain, C. et al. Long-read mapping to repetitive reference sequences using Winnowmap2 // Nature Methods. – 2022. – Т. 19. – С. 705–710.
6. Robinson, T. et al. Integrative genomics viewer // Nature biotechnology. – 2011–Т.29. № 1. С. 24-26.
7. Nurk, S. et al. The complete sequence of a human genome // Science – 2022. – Т. 376. – №. 6588. – С. 44-53.
8. Kunyavskaya, O., et al. Automated annotation of human centromeres with HORmon // Genome Research – 2022.
9. Dvorkina, T., et al. CentromereArchitect: inference and analysis of the architecture of centromeres // Bioinformatics – 2011. – Т. 37. – №. Supplement_1. – С. 196-204.
10. Wyman, D., et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. // Biorxiv. – 2020 – 672931.
11. Chen Y. et al. Context-Aware Transcript Quantification from Long Read RNA-Seq data with Bambu //bioRxiv. – 2022.
12. Prjibelski, A., et al. IsoQuant: a tool for accurate novel isoform discovery with long reads. // Researchsquare. – 2022 – doi:10.21203/rs.3.rs-1571850/v1.
13. Hardwick, S.A., et al. Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. // Nature Biotechnology. – 2022 – С. 1-11.

14. Mikheenko, A., et al. Sequencing of individual barcoded cDNAs using Pacific Biosciences and Oxford Nanopore Technologies reveals platform-specific error patterns. // Genome Research. – 2022 – C. 726-737.
15. Hafezqorani, S., et al. Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. // GigaScience – 2020 – giaa061.
16. Li, Bo, and Colin N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. // BMC bioinformatics – 2011 – C. 1-16.
17. Tardaguila, M., et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. // Genome research. – 2018 – C. 396-411.
18. Pardo-Palacios, F., et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. // Researchsquare. – 2021 – doi:10.21203/rs.3.rs-1571850/v1.
19. Kolmogorov M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs // Nature Methods. – 2020. – Т. 17. – №. 11. – С. 1103-1110.
20. Nurk S. и др. MetaSPAdes: A new versatile metagenomic assembler // Genome Res. 2017. Т. 27. № 5. С. 824–834.
21. Antipov D. et al. Plasmid detection and assembly in genomic and metagenomic data sets // Genome research. – 2019. – Т. 29. – №. 6. – С. 961-968.
22. Antipov D. et al. Metaviral SPAdes: assembly of viruses from metagenomic data // Bioinformatics. – 2020. – Т. 36. – №. 14. – С. 4126-4129.
23. Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets // Nat. Biotechnol. 2017. Т. 35. № 11. С. 1026–1028.
24. Yutin N. et al. Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis // Virology journal. – 2018. – Т. 15. – №. 1. – С. 1-17.
25. Pell J. et al. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs // Proceedings of the National Academy of Sciences. – 2012. – Т. 109. – №. 33. – С. 13272-13277.
26. Flajolet P. et al. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm //Discrete Mathematics and Theoretical Computer Science. – Discrete Mathematics and Theoretical Computer Science, 2007. – С. 137-156.

27. Pandey P. et al. A general-purpose counting filter: Making every bit count //Proceedings of the 2017 ACM international conference on Management of Data. – 2017. – С. 775-787.
28. Guo D. et al. False negative problem of counting bloom filter //IEEE transactions on knowledge and data engineering. – 2010. – Т. 22. – №. 5. – С. 651-664.
29. Yutin N. et al. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features // Nature communications. – 2021. – Т. 12. – №. 1. – С. 1-11.
30. Benler S. et al. Thousands of previously unknown phages discovered in whole-community human gut metagenomes // Microbiome. – 2021. – Т. 9. – №. 1. – С. 1-17.
31. Yutin N. et al. Varidnaviruses in the human gut: a major expansion of the order Vinavirales // Viruses. – 2022. – Т. 14. – №. 9. – С. 1842.
32. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. // Science. – 2022. – Т. 376. – №. 6588.
33. McCartney, M. et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. // Nature Methods. – 2022. Т.19. – С.687–695
34. Bankevich, A. et al. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. // Nature Biotechnology. – 2022. Т.40. – С.1075–1081
35. Meyer F. et al. Critical Assessment of Metagenome Interpretation: the second round of challenges // Nature Methods. – 2022. – Т. 19. – №. 4. – С. 429–440
36. Dvorkina T. et al. The string decomposition problem and its applications to centromere analysis and assembly. // Bioinformatics. – 2020. – Т. 36(Suppl_1) – С. 93-101
37. Hug J. J., Krug D., Müller R. Bacteria as genetically programmable producers of bioactive natural products // Nature Reviews Chemistry. – 2020. – Т. 4. – №. 4. – С. 172-193.
38. Kunyavskaya O. et al. Nerpa: A Tool for Discovering Biosynthetic Gene Clusters of Bacterial Nonribosomal Peptides // Metabolites. – 2021. – Т. 11. – №. 10. – С. 693.
39. Blin K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities // Nucleic Acids Research. – 2021.
40. Ricart E. et al. rBAN: retro-biosynthetic analysis of nonribosomal peptides //Journal of cheminformatics. – 2019. – Т. 11. – №. 11. – С. 13.
41. Kautsar S. A. et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function //Nucleic acids research. – 2020. – Т. 48. – №. D1. – С. D454-D458.

42. Terlouw B. et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters // Nucleic Acids Research. – 2022.
43. Leão T.F. et al. NPOmix: a machine learning classifier to connect mass spectrometry fragmentation data to biosynthetic gene clusters // PNAS Nexus. В печати.
44. Lotonin K. et al. Balamuthia spinosa n. sp.(Amoebozoa, Discosea) from the brackish-water sediments of Nivå Bay (Baltic Sea, The Sound)—a novel potential vector of Legionella pneumophila in the environment //Parasitology Research. – 2022. – Т. 121. – №. 2. – С. 713-724.
45. Navarro-Muñoz J. et al. A computational framework to explore large-scale biosynthetic diversity // Nature Chemical Biology. – 2020. – Т. 16. – №. 1. – С. 60–68
46. Wang M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking // Nature Biotechnology. – 2016. – Т. 34. – №. 8. – С. 828–837
47. Schorn M.A. et al. A community resource for paired genomic and metabolomic data mining // Nature Chemical Biology. – 2021. – Т. 17. – №. 4. – С. 363–368
48. Kravchenko I. et al. Agricultural Crops Grown in Laboratory Conditions on Chernevaya Taiga Soil Demonstrate Unique Composition of the Rhizosphere Microbiota //Microorganisms. – 2022. – Т. 10. – №. 11. – С. 2171.