Тема работы:

Технология структурирования и обработки транскриптомных данных на основе гибридного использования RDBMS и NoSQL подходов

Состав коллектива

Генаев М.А. н.с., к.б.н., ИЦиГ СОРАН, НГУ

Мухин А.М. м.н.с., ИЦиГ СОРАН

Афонников Д. А. в.н.с, к.б.н., ИЦиГ СОРАН, НГУ

Информация о гранте:

Работа была поддержана грантом РНФ №18-14-00293

Научное содержание работы:

Постановка задачи.

Эксперимент по секвенированию транскриптома (RNA-seq) у растений стал практически рутинной процедурой для изучения как модельных организмов, так и для сельскохозяйственных культур. В результате биоинформатической обработки таких экспериментов получаются объемные разнородные данные, представленные нуклеотидными последовательностями транскриптов, аминокислотными последовательностями и их структурно-функциональной аннотацией. Полученные данные важно представить широкому кругу исследователей в виде баз данных (БД).

В работе предложен гибридный подход к созданию молекулярно-генетических баз данных, которые содержат информацию о последовательностях транскриптов и их структурно-функциональной аннотации. Сущность подхода в одновременном хранении в БД информации как структурированного типа, так и слабо структурированных данных. Технология использована для реализации БД транскриптомов сельскохозяйственных растений. В работе рассматриваются особенности реализации такого подхода и примеры формирования как простых, так и сложных запросов к такой базе данных на языке SQL.

Современное состояние проблемы.

Изучение транскриптомов растений с помощью высокопроизводительного секвенирования (секвенирование РНК, RNA-seq) широко используется в настоящее время для решения таких задач как оценка экспрессии генов для разных генотипов и в разных условиях среды, идентификация последовательностей РНК (для не модельных организмов), поиск маркеров к функционально важным генам [1,2]. Эксперимент RNA-seq стал практически рутинной процедурой для изучения как модельных организмов (Arabidopsis thaliana) [3], так и для сельскохозяйственных культур (томат, кукуруза, ячмень, пшеница и др.) [4]. Результаты транскриптомного эксперимента представляют собой короткие фрагменты нуклеотидных последовательностей и лишь биоинформатическая обработка, включающая несколько стадий [5–8], позволяет получить на их основе последовательности транскриптов и их функциональную аннотацию. Именно результаты биоинформатической обработки представляют интерес для биолога и могут быть интерпретированы в терминах функций генов, их продуктов, уровней экспрессии, генетических вариаций и т.п. [9,10]. Необходимо отметить, что в результате транскриптомного эксперимента получается большое количество данных (десятки и сотни тысяч последовательностей), свободный и удобный доступ к которым важно предоставить широкому кругу биологов, далеких от рутинной биоинформатической обработки результатов секвенирования. Этой цели служат базы данных, имеющие удобный пользовательский интерфейс и организующие связи между биологическими последовательностями и их функциональной аннотацией. Среди таких баз можно указать

Expression Atlas Европейского Института Биоинформатики [11], EGENES, базу данных информации о метаболических путях генов, основанную на транскриптомных данных [12], базы данных по экспрессии генов для определенных видов организмов: TodoFirgene для пихты Abies sachalinensis [13]; атлас экспрессии генов для розы [14]; базу аннотированных транскриптомов приморской сосны EuroPineDB [15]; и др.

Результаты обработки RNA-seq экспериментов, представленные в таких базах данных, являются комплексными и включают: последовательности транскриптов, их локализацию в геноме, классификацию по типам генов (мРНК, днкРНК, миРНК, тРНК и пр.), функциональную аннотацию транскриптов, оценку уровней экспрессии, оценку вариантов изоформ транскриптов. Эти результаты представлены в виде бинарных и текстовых файлов в различных форматах. Это могут быть файлы последовательностей (форматы FASTA, FASTQ), выравниваний (форматы BAM, SAM, PSL и т.д.) или разметки (BED, GFF, GTF) [16–19]. Результаты анализа дифференциальной экспрессии генов обычно представляют в виде таблиц (форматы TSV, XLS) [20,21]. Подобные хорошо структурированные данные удобно описывать в виде классической реляционной модели отношений RDBMS (Реляционная система управления базами данных, англ. Relational Database Management System): например, у одного гена может быть несколько изоформ, в эксперименте рассматриваются несколько образцов ткани разных различных особей и т.д.

Отметим, однако, что наряду с хорошо структурированными данными, в результате анализа RNA-seq экспериментов генерируются слабо структурированные и неструктурированные данные, которые не могут быть описаны с помощью реляционной модели. Сложности могут возникать в силу разнородности данных, получаемых в процессе выполнения биоинформатических вычислительных конвейеров. Эта разнородность обусловлена разнообразием методов, которые вовлечены в конвейеры биоинформатической обработки транскриптомных данных [7,22]. Узлами в конвейерах являются различные программы, которые реализуют методы обработки данных. На практике обычно бывает так, что в существующий вычислительный конвейер для решения какой-то задачи в процессе работы вносятся изменения: например, происходит замена некоторых узлов на новые, которые реализуют более точные или более производительные методы обработки данных. Поэтому заранее невозможно полностью декларировать структуру данных, которая будет получена в результате их обработки. Описание таких слабоструктурированных данных удобнее делать с использованием технологий NoSQL (не только SQL, англ. not only SQL) [23,24].

Подробное описание работы, включая используемые алгоритмы

В настоящей работе мы предлагаем комплексный подход к описанию транскриптомных данных, который заключается в использовании элементов RDBMS при описании хорошо структурированных данных и NoSQL для описания слабо структурированных данных, полученных в результате широкомасштабного биоинформатического анализа транскриптомных экспериментов у 5 сельскохозяйственных растений (кукурузы, риса, ячменя, томата и картофеля). Анализ этих данных был направлен на идентификацию новых транскриптов, которые либо не выравниваются на референсный геном растения, либо выравниваются на его неаннотированные участки и, таким образом, представляют собой новую, ранее неисследованную часть транскриптома. На примере задачи массового анализа транскриптомов сельскохозяйственных растений мы предлагаем наборы реляционных отношений для описания основных сущностей: исследование, эксперимент, нуклеотидные и белковые последовательности. В то же время, для каждой их этих сущностей мы предлагаем вводить возможность для аннотации наборами слабоструктурированных данных, формат представления которых может быть заранее неизвестен. На основе предложенного гибридного подхода разработана база данных OORT (Out Of Reference Transcripts), которая позволяет пользователям, с помощью поисковых запросов, извлекать информацию о структуре и функциях ранее неаннотированной части транскриптомов сельскохозяйственных растений, в

частности: идентифицировать новые гены устойчивости к заболеваниям и абиотическому стрессу, длинные некодирующие РНК, последовательности мРНК, получать оценки уровня экспрессии этих транскриптов. База данных построена на основе анализа 1241 транскриптомных экспериментов и содержит информацию о 20440228 нуклеотидных и 4055996 аминокислотных аннотированных последовательностях. Функциональные возможности базы данных OORT демонстрируются на примере нескольких запросов.

В качестве исходных данных использованы архивы SRA (Sequence Read Archive), которые хранят «сырые» данные секвенирования транскриптомных библиотек. Архивы загружены с сайта ENA [25,26]. Каждый архив в базе данных ENA сопровождается метаинформацией, структура которой включает описание: идентификатор биологического проекта, в рамках которого проводилось секвенирование (BioProject) идентификатор исследование (study), к которому относится библиотека, образец, для которого получено секвенирование транскриптома. Метаданные для исследования содержат также его краткое описание, список публикаций в которых результаты транскриптомного эксперимента были опубликованы, данные об исследуемых образцах (BioSample, sample), например, вид, пол, ткань или орган, геометка и т.д. Из каждого образца может быть получено несколько препаратов для секвенирования (experiment), в метаданных эксперимента описан метод выделения РНК, экспериментальную платформу секвенирования. С экспериментом ассоциировано один или несколько SRA архивов, каждый из которых соответствует набору прочтений, полученных в результате секвенирования образца на конкретном секвенаторе. Каждый запуск секвенатора соответствует в терминах NCBI SRA одному файлу архива. Детальное описание структуры отношений между указанными уровнями описания данных по секвенированию РНК представлено на сайте NCBI [27].

При формировании базы данных OORT рассматривались эксперименты для пяти видов сельскохозяйственных растений: *Hordeum vulgare* (ячмень, taxonomy ID 4513); *Oryza sativa* (рис, taxonomy ID 4530); *Solanum lycopersicum* (томат, taxonomy ID 4081); Solanum tuberosum (картофель, taxonomy ID: 4113); *Zea mays* (кукуруза, taxonomy ID 4577). Для анализа были отобраны SRA архивы с библиотеками RNA-seq, которые соответствовали следующим критериям: платформа секвенирования Illumina HiSeq 2000 или Illumina HiSeq 2500; данные полученные методами PolyA, Inverse rRNA или cDNA; длина прочтений библиотеки не менее 75. В итоге, запрос для получения списка данных из базы SRA формулировался следующим образом:

"instrument_platform == 'ILLUMINA' & library_strategy == 'RNA-Seq' & library_source == 'TRANSCRIPTOMIC' & (instrument_model=='Illumina HiSeq 2000' | instrument_model=='Illumina HiSeq 2500') & sra_has_ftp == True & mean_read_len>=75 & (library_selection == 'cDNA' | library_selection == 'PolyA' | library_selection == 'Inverse rRNA')".

Данному запросу удовлетворяло 3883 SRA архива, 200 исследований, 3419 образцов и 3578 экспериментов. Из полученного списка файлов мы отобрали данные, на которые есть ссылки в публикациях. В результате мы получили список из 69 исследований. Эти данные включают 1395 SRA файла общим размером 695 ГБ.

Первый этап формирования базы данных OORT заключался в *de novo* сборке последовательностей транскриптов из сырых прочтений. Для этого был разработан конвейер, который состоит из четырех последовательных шагов: (1) извлечение прочтений из SRA файла с помощью пакета SRA Toolkit [28]; (2) подготовка данных (сырые данные были подвергнуты фильтрации при помощи программы fastp [29]); (3) сборка, с использованием программы Trinity-v2.6.6. [5]; (4) оценка уровня экспрессии транскриптов с помощью программы Kallisto [21]. В качестве количественной меры экспрессии транскриптов мы использовали Transcripts Per Million (TPM).

Всего нами было обработано 1298 SRA файлов - библиотек коротких прочтений. Для классификации транскриптов в собранных нами транскриптомах мы использовали программу rnaQUAST v. 1.5 [30]. С использованием выравнивания транскриптов на референсный геном и известную его аннотацию, программа rnaQUAST классифицирует все транскрипты на 5 групп: (1) Unaligned - невыравненные транскрипты на геном; (2) Multiply aligned - транскрипты, которые выравниваются на 2 и более участка референсного генома; (3) Misassembled - транскрипты, выравнивание, которых имеет разногласия с аннотацией; (4) Uniquely aligned - транскрипты, имеющие ровно 1 выравнивание на референсный геном, которые при этом не содержат разногласий с аннотацией; (5) Unannotated - транскрипты, которые выровнены на референсный геном на его неаннотированные участки. Для выравнивания транскриптов на референсный геном в программе rnaQUAST использовался пакет gmap v. 2018-07-04 [31]. Мы использовали последовательности и аннотации референсных геномов пяти растений, загруженные с сайта Ensembl Plants [32,33]. Нами были использованы следующие версии сборок и аннотаций: для ячменя v. 42 (Hordeum_vulgare.IBSC_v2.42), для риса v. 40 (Oryza_sativa.IRGSP-1.0.40), для томата v. 40 (Solanum_lycopersicum.SL2.50.40), для картофеля v. 40 (Solanum_tuberosum.SolTub_3.0.40), для кукурузы v. 40 (Zea_mays.AGPv4.40).

Для контигов, которые были классифицированы как Unaligned и Unannotated, с помощью программы TransDecoder v5.5.0. [5] были получены предполагаемые белок-кодирующие последовательности,. Для аннотации аминокислотных последовательностей использовалась программа InterproScan v.5.36-75 [34].

Общий объем полученных таким образом данных составил 20440228 нуклеотидных последовательностей транскриптов суммарной длиной 9659793403 нуклеотидов и 4055996 аминокислотных последовательностей суммарной длиной 877229075 аминокислот. В дальнейшем эти данные, а также данные об исследованиях, экспериментах и результаты аналитических программ были экспортированы в базу данных OORT

Содержание БД OORT включает:

- метаинформацию о библиотеках транскриптомных экспериментов;
- нуклеотидные последовательности транскриптов, полученные в результате *de novo* сборки;
- аминокислотные последовательности, полученные в результате трансляции нуклеотидных последовательностей транскриптов, кодирующих белки;
- аннотация нуклеотидных последовательностей (предсказание кодирующего потенциала, выравнивание на референсный геном, оценку уровня экспрессии);
- аннотация аминокислотной последовательности (предсказание функциональных доменов, ассоциированные с последовательностью термины онтологии генов).

При работе с базой данных для пользователя важно решать ряд поисковых задач, связанных с идентификацией последовательностей в базе данных по метаинформации об эксперименте, принадлежности к организму, гомологии, функциональным характеристикам, уровню экспрессии транскрипта. Подобный поиск может осуществляться как для нуклеотидных, так и для аминокислотных последовательностей. При этом следует принимать во внимание отношения между аминокислотной последовательностью и нуклеотидной последовательностью транскрипта, из которого она была получена.

В качестве СУБД при формировании базы данных OORT мы использовали PostgreSQL версии 12 [35,36]. С ее помощью была построена реляционная схема данных связывающая исследования, эксперименты, контиги и белки (см. раздел “Структура базы данных”). База данных позволяет определять реляционные отношения с помощью первичных и внешних ключей.

С помощью языка программирования Python данные из файлов с результатами биоинформатической обработки (см раздел “Содержание БД OORT) были преобразованы в структуру типа “словарь”, которую затем конвертировали в JSON формат с помощью библиотеки SQLAlchemy [37]. SQLAlchemy позволяет транслировать код на языке Python в команды на языке SQL и передавать их на выполнение СУБД, получая результаты запросов в виде объектов языка Python. Далее эти данные были преобразованы в структуры формата JSONB для дальнейшего хранения и доступа.

Создаваемая нами схема базы данных разрабатывалась для поиска информации, представленных как в типизированных (хорошо структурированных), так и в слабоструктурированных полях. Типизированные поля предназначены для описания сущностей в БД, типы которых известны и определены однозначно. Например, последовательность представлена в виде текстовой строки, длина последовательности описана полем типа int, название организма описано в виде текстового поля. Для поиска информации в типизированных данных (int, real, и text) проводилась индексация с помощью структуры B-tree, которая в дальнейшем позволяла выполнять в том числе и операции поиска с учетом отношений “равно”, “больше”, “меньше” [38]. Структура B-tree позволяет оптимизировать выполнение запросов с указанными операциями.

Помимо типизированных, хорошо структурированных данных, содержимое базы включало слабо структурированную информацию. К таким данным относились в том числе и аннотации аминокислотных последовательностей в терминах онтологии генов (GO) [39] и белковых доменов [34]; результаты выравнивая контигов на референсные геномы; результаты оценки уровня экспрессии транскриптов и т.д. Данные подобного сорта представлялись в виде текстовых строк в формате JSON: записи вида ключ-значение без определенных заранее ограничений. СУБД PostgreSQL 12 позволяет хранить информацию в формате JSON в структуре реляционной базы как поля специального типа, JSONB [40]. Это данные в формате JSON, представленные в бинарном виде для быстрого обращения к структуре таких полей без повторного синтаксического разбора, при этом ключи и значения в этом поле не описываются в формате языка SQL. Поверх ключей поля JSONB можно создавать индексы для быстрого поиска данных, также некоторые индексы позволяют использовать новые поисковые опции над значениями, например, поиск искомого объекта в массиве. Индексация полей формата JSON проводилась с помощью GIN индексов [40]. Такая индексация позволяет быстро производить поиск текстовых строк в указанном массиве терминов, например, терминов онтологии и белковых доменов.

Для индексации координат контигов в геноме использовалась структура “материализованное представление”, специальная техника, реализованная в СУБД PostgreQSL, которая позволяет сохранить в виде отдельной таблицы базы данных результат определенного запроса [41]. Это позволяет, после выполнения запроса в виде “материализованного представления”, обращаться к нему как к отдельной таблице реляционной БД. Такая технология позволяет существенно ускорить выполнение запросов (особенно сложных) к базе данных, поскольку в случае повторного запроса обращение производится только к этой таблице, а не к реальным полям БД.

Доступ к извлечению и модификации данных в БД OORT производился на языке запросов SQL. Типичный запрос на поиск данных включает, как правило, следующие секции этого языка:

- SELECT - секция выбора полей таблиц. Здесь через запятую нужно указать имена колонок, которые нужно отобразить, или * чтобы отобразить все поля.
- FROM - секция таблиц. В этой секции следует указать таблицы, из которых нужно получить результаты. Также в этой секции может использоваться оператор JOIN, с помощью которого выполняется объединение таблиц.

- WHERE - секция условия. В ней пользователь указывает условия поиска и объединяет их с помощью логических операторов AND, OR и NOT (условие не выполняется).

Следует отметить, что СУБД PostgreSQL версии 12 расширяет диалект языка SQL:2016 за счет дополнительных операторов, которые служат для извлечения данных из полей формата JSONB [40]. К этим операторам относятся оператор '->', который используется для извлечения вложенного JSON-значения по ключу, и оператор '->>', который используется для получения строковых значений по ключу. Использование этих операторов продемонстрировано ниже в разделе “Примеры поисковых запросов”.

Интерфейс SQL к разработанной БД OORT может быть реализован с помощью таких приложений, как DataGrip [42], pgAdmin [43] или консольного приложения psql СУБД PostgreSQL.

1. Martin L.B.B. и др. Catalyzing plant science research with RNA-seq // Frontiers in Plant Science. Frontiers Research Foundation, 2013. Т. 4, № APR. С. 66.
2. Usadel B., Fernie A.R. The plant transcriptome—from integrating observations to models // Front. Plant Sci. Frontiers Research Foundation, 2013. Т. 4, № MAR. С. 48.
3. Klepikova A. V. и др. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling // Plant J. Blackwell Publishing Ltd, 2016. Т. 88, № 6. С. 1058–1070.
4. Strickler S.R., Bombarely A., Mueller L.A. Designing a transcriptome next-generation sequencing project for a nonmodel plant species // American Journal of Botany. John Wiley & Sons, Ltd, 2012. Т. 99, № 2. С. 257–266.
5. Haas B.J. и др. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis // Nat. Protoc. Nature Publishing Group, 2013. Т. 8, № 8. С. 1494–1512.
6. Kim D., Langmead B., Salzberg S.L. HISAT: A fast spliced aligner with low memory requirements // Nat. Methods. Nature Publishing Group, 2015. Т. 12, № 4. С. 357–360.
7. Bryant D.M. и др. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors // Cell Rep. Elsevier B.V., 2017. Т. 18, № 3. С. 762–776.
8. Bolger M.E., Arsova B., Usadel B. Plant genome and transcriptome annotations: From misconceptions to simple solutions // Brief. Bioinform. Oxford University Press, 2018. Т. 19, № 3. С. 437–449.
9. Glagoleva A.Y. и др. Metabolic pathways and genes identified by RNA-seq analysis of barley near-isogenic lines differing by allelic state of the Black lemma and pericarp (Blp) gene // BMC Plant Biol. BioMed Central Ltd., 2017. Т. 17, № S1. С. 182.
10. Shmakov N.A. и др. Identification of nuclear genes controlling chlorophyll synthesis in barley by RNA-seq // BMC Plant Biol. BioMed Central Ltd., 2016. Т. 16, № 3. С. 119–138.
11. Papatheodorou I. и др. Expression Atlas update: From tissues to single cells // Nucleic Acids Res. Oxford University Press, 2020. Т. 48, № D1. С. D77–D83.
12. Masoudi-Nejad A. и др. EGENES: Transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG // Plant Physiol. American Society of Plant Biologists, 2007. Т. 144, № 2. С. 857–866.
13. Ueno S. и др. TodoFirGene: Developing transcriptome resources for genetic analysis of abies sachalinensis // Plant Cell Physiol. Oxford University Press, 2018. Т. 59, № 6. С. 1276–1284.

14. Dubois A. и др. Transcriptome database resource and gene expression atlas for the rose // BMC Genomics. BioMed Central, 2012. Т. 13, № 1. С. 638.

15. Fernández-Pozo N. и др. EuroPineDB: A high-coverage web database for maritime pine transcriptome // BMC Genomics. BioMed Central Ltd., 2011. Т. 12, № 1. С. 366.

16. Barnett D.W. и др. Bamtools: A C++ API and toolkit for analyzing and managing BAM files // Bioinformatics. Oxford Academic, 2011. Т. 27, № 12. С. 1691–1692.

17. Quinlan A.R., Hall I.M. BEDTools: A flexible suite of utilities for comparing genomic features // Bioinformatics. Oxford Academic, 2010. Т. 26, № 6. С. 841–842.

18. Li H. и др. The Sequence Alignment/Map format and SAMtools // Bioinformatics. Oxford Academic, 2009. Т. 25, № 16. С. 2078–2079.

19. Pertea G., Pertea M. GFF Utilities: GffRead and GffCompare // F1000Research. F1000 Research Ltd, 2020. Т. 9. С. 304.

20. Anders S., Huber W. Differential expression of RNA-Seq data at the gene level-the DESeq package.

21. Bray N.L. и др. Near-optimal probabilistic RNA-seq quantification // Nat. Biotechnol. Nature Publishing Group, 2016. Т. 34, № 5. С. 525–527.

22. Gunbin K. V. и др. Computer System for Analysis of Molecular Evolution Modes (SAMEM): Analysis of molecular evolution modes at deep inner branches of the phylogenetic tree // In Silico Biol. IOS Press, 2011. Т. 11, № 3. С. 109–123.

23. Han J. и др. Survey on NoSQL database // Proceedings - 2011 6th International Conference on Pervasive Computing and Applications, ICPCA 2011. 2011. С. 363–366.

24. Gabetta M. и др. BigQ: A NoSQL based framework to handle genomic variants in i2b2 // BMC Bioinformatics. BioMed Central Ltd., 2015. Т. 16, № 1. С. 415.

25. ENA Portal [Электронный ресурс]. URL: https://www.ebi.ac.uk/ena/portal/api/ (дата обращения: 23.10.2020).

26. Harrison P.W. и др. The European Nucleotide Archive in 2018 // Nucleic Acids Res. Oxford University Press, 2019. Т. 47, № D1. С. D84–D88.

27. Submit your project and biological samples [Электронный ресурс]. URL: https://www.ncbi.nlm.nih.gov/sra/docs/submitbio/ (дата обращения: 23.10.2020).

28. Staff S.R.A.S. Using the SRA Toolkit to convert .sra files into other formats. National Center for Biotechnology Information (US), 2011.

29. Chen S. и др. Fastp: An ultra-fast all-in-one FASTQ preprocessor // Bioinformatics. Oxford University Press, 2018. Т. 34, № 17. С. i884–i890.

30. Bushmanova E. и др. RnaQUAST: A quality assessment tool for de novo transcriptome assemblies // Bioinformatics. Oxford University Press, 2016. Т. 32, № 14. С. 2210–2212.

31. Wu T.D., Watanabe C.K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences // Bioinformatics. Oxford Academic, 2005. Т. 21, № 9. С. 1859–1875.

32. Ensembl Plants [Электронный ресурс]. URL: https://plants.ensembl.org/index.html (дата обращения: 23.10.2020).

33. Kersey P.J. и др. Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species // Nucleic Acids Res. Oxford University Press, 2018. Т. 46, № D1. С. D802–D808.

34. Jones P. и др. InterProScan 5: Genome-scale protein function classification // Bioinformatics. Oxford University Press, 2014. Т. 30, № 9. С. 1236–1240.

35. PostgreSQL: The world's most advanced open source database [Электронный ресурс]. URL: https://www.postgresql.org/ (дата обращения: 23.10.2020).

36. Schönig H.-J. Mastering PostgreSQL 11: Expert techniques to build scalable, reliable, and fault-tolerant database applications. Packt Publishing Ltd, 2018.

37. SQLAlchemy - The Database Toolkit for Python [Электронный ресурс]. URL: https://www.sqlalchemy.org/ (дата обращения: 23.10.2020).

38. PostgreSQL: Documentation: 12: 11.2. Index Types [Электронный ресурс]. URL: https://www.postgresql.org/docs/12/indexes-types.html (дата обращения: 23.10.2020).

39. Carbon S. и др. The Gene Ontology Resource: 20 years and still GOing strong // Nucleic Acids Res. Oxford University Press, 2019. Т. 47, № D1. С. D330–D338.

40. Petković D. JSON integration in relational database systems // Int J Comput Appl. 2017. Т. 168, № 5. С. 14–19.

41. Kaur M., Shaik B. PostgreSQL Development Essentials. Packt Publishing Ltd, 2016.

42. DataGrip: кросс-платформенная среда разработки для баз данных и SQL [Электронный ресурс]. URL: https://www.jetbrains.com/ru-ru/datagrip/ (дата обращения: 23.10.2020).

43. pgAdmin - PostgreSQL Tools [Электронный ресурс]. URL: https://www.pgadmin.org/ (дата обращения: 23.10.2020).

Полученные результаты.

Реляционная схема базы данных включает описание метаданных транскриптомного эксперимента и результатов сборки транскриптов *de novo* в виде следующих объектов: исследование (study), эксперимент (exp), нуклеотидная последовательность (contig) и аминокислотная последовательность (pep) (рис. 1). Как было указано в разделе "Формирование базы данных, индексация", при создании ссылок между таблицами БД мы использовали первичные и вторичные ключи, которые были созданы следующим образом:

- Каждая из таблиц БД хранит поле id, которое является первичным ключом
- Таблица exp содержит вторичный ключ study_id, который ссылается на таблицу study.
- Таблица contig содержит вторичный ключ exp_id, который ссылается на таблицу exp.
- Таблица pep содержит вторичные ключи contig_id и exp_id, которые ссылаются на таблицы contig и exp, соответственно.

В результате, между указанными таблицами были созданы реляционные связи, позволяющие эффективно осуществлять поиск информации по эксперименту, исследованию и полученных в результате эксперимента последовательностях. Структура отношений между таблицами БД OORT представлена на рисунке 1. Примеры запросов для решения различных биологических задач показаны на рисунке 2.

Созданная нами база данных на основе гибридного метода описания комплексных молекулярно-биологических данных в виде как хорошо структурированных (типизированных) данных, так и слабо структурированной информации, обладает рядом преимуществ. Прежде всего, такая организация хранения данных облегчает их сопровождение: при изменении набора

программ аннотации, их версий или набора выводимых ими параметров, вместо того, чтобы модифицировать реляционную схему базы и повторно загружать данные в базу, разработчик может составлять слабосвязанные структуры в формате JSON.

Предложенная архитектура БД сохраняет связность таблиц при помощи методов первичных и вторичных ключей. Следует отметить, что структура JSON-описания нетипизированных данных строго не описана и для разных записей даже одной таблицы может отличаться. В конечном итоге, эта структура зависит от того, какие данные были внесены в конкретное поле БД. Поддержание однородности данных в формате JSON при такой организации лежит на разработчике: контроль структуры полей JSON происходит либо в момент экспорта данных, и\или путем реализации функций-обработчиков, которые запускаются при каждом обновлении и добавлении данных.

Иллюстрации, визуализация результатов.

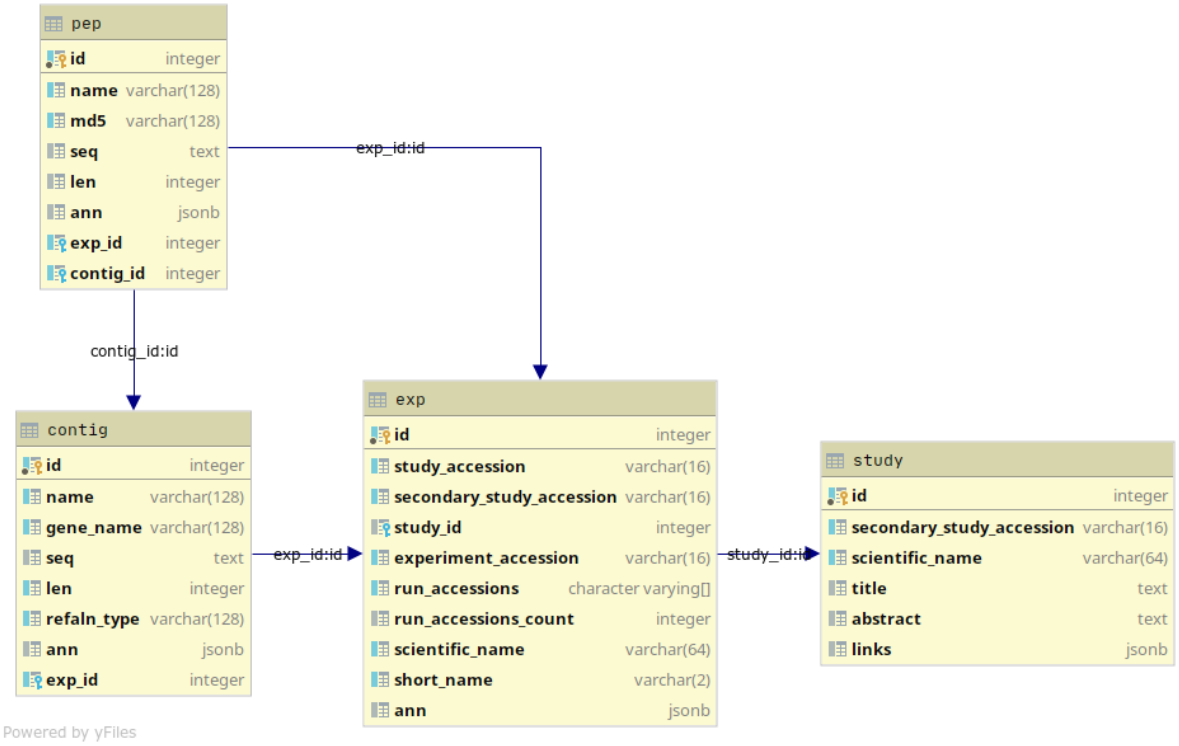

Рис. 1. Структура реляционной базы данных OORT. Показаны таблицы БД и отношения между ними.

```
# Запрос 1: Получить последовательности генов для вида 'Oryza sativa'
01 select contig.seq
02 from contig join exp on exp.id = contig.exp_id
03 where scientific_name = 'Oryza sativa'
```

```
# Запрос 2: Найти все гены, с уровнем экспрессии в пределах от 3.0 до 7.0 TPM
01 select *
02 from contig
03 where ((ann -> 'kallisto' -> 'g' ->> 'tpm')::real)
```

```
04 between 3.0 and 7.0
```

Запрос 3: Вывести список генов, кодирующих идентичные пептиды с разным уровнем экспрессии в разных экспериментах

```
01 select subseq.exp_id, subseq.study_id,
02 max(subseq.contig_g_tpm) as contig_g_tpm,
03 study.scientific_name, tissue_type
04 from (
05    select (contig.ann->'kallisto'->'g'->'tpm')::real
06    as contig_g_tpm, contig.id as contig_id, pep.id as
07    pep_id, contig.exp_id, e.study_id, pep.md5,
08    e.ann->'ebi'->'tissue_type' as tissue_type
09    from pep join contig on pep.contig_id = contig.id
10    join exp e on contig.exp_id = e.id
11    where pep.md5 = '857c1634328c5333ab73efcc11a45038'
12 )
13 subseq join study on subseq.study_id = study.id
14 group by exp_id, study_id, study.scientific_name,
15 tissue_type
```

Запрос 4: Найти все аминокислотные последовательности с определенным набором терминов Генной Онтологии

```
# создание индекса для выполнения запроса
01 create index ix_pep_ann_interpro_go_ann on pep using
02 gin (((ann -> 'interpro'::text) -> 'go_ann'::text))
```

поиск белков согласно аннотации генной онтологии на основе созданного индекса

```
03 select * from pep
04 where (ann -> 'interpro' -> 'go_ann')
05 ?& array['GO:0055085', 'GO:0006811']
```

Запрос 5: Поиск слова в описаниях исследования

```
# создание индекса для полнотекстового поиска
01 create index study_abstract_idx on study
02 (to_tsvector('english'::regconfig, abstract));
# выполнение полнотекстового поиска
03 select *
04 from study
05 where to_tsvector(study.abstract)
06 @@ plainto_tsquery('Categorizing')
```

```
# Запрос 6. Найти все белки, которые транслируются из контигов в геноме Zea mays во 2 хромосоме между 2000 и 5000 нуклеотидами.

# реализация материального представления contig_gmap_view_sqlalchemy

create materialized view
contig_gmap_view_sqlalchemy as
select db.id, jsonb_array_elements(
db.ann -> 'gmap'::text) AS gmap
from (
select contig.id, contig.ann
from contig ORDER BY contig.id
) db;

# создание индексов поверх колонок GMAP

create index gmap_start_end_chr_idx
on contig_gmap_view_sqlalchemy
(((gmap ->> 15)::integer), ((gmap ->> 16)::integer),
(gmap ->> 13));

# поиск белков, синтезирующихся со 2 хромосомы с 2000 по 5000

# позиции в геноме

select pep.id AS pep_id
from exp join contig ON exp.id = contig.exp_id
join contig_gmap_view_sqlalchemy
on contig.id = contig_gmap_view_sqlalchemy.id
join pep ON pep.contig_id = contig.id
where
cast(contig_gmap_view_sqlalchemy.gmap ->> 15
as integer) >=2000 AND
cast(contig_gmap_view_sqlalchemy.gmap ->> 16
as integer) <= 5000
and(contig_gmap_view_sqlalchemy.gmap ->> 13) = '2'
and exp.scientific_name like 'Zea mays'
```

Рис. 2. Реализация запросов 1 и 6 к базе данных OORT на языке SQL.

Эффект от использования кластера в достижении целей работы.

Denovo сборка последовательностей транскриптов, на основе которых создавалась база данных OORT выполнялась на кластере НГУ. Процессорное время, которое потребовалось для denovo сборки этих данных на сервере с двумя процессорами Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz (суммарно 28 ядер) и 128GB оперативной памяти составляет ~9011.7 часов (или ~375.4 дней).

Результаты работы данной работы опубликованы в статье:

Мухин А. М. и др. Технология структурирования и обработки транскриптомных данных на основе гибридного использования RDBMS и NoSQL подходов Математическая биология и биоинформатика. 2020; 15 (2): 455-470. doi: 10.17537/2020.15.455.