Разработка инфраструктуры хранения данных геномного секвенировании и результатов их анализа

В. В. Скакун, Н. Н. Яцков, В. В. Гринев

Белорусский государственный университет, Минск; e-mail: skakun@bsu.by

Секвенирование генома позволяет идентифицировать множество сайтов генетического полиморфизма с целью диагностики генетических заболеваний человека. Автоматизация процесса анализа отсутствует, параметры методов и моделей анализа данных задаются пользователем вручную. Несистематизированное хранение множества данных и результатов их обработки сильно затрудняет их последующее использование. Все это требует разработки единого и высокопроизводительного программного комплекса. В статье описан процесс разработки базового элемента комплекса – инфраструктуры хранения данных генетического полиморфизма и результатов их анализа.

Ключевые слова: геномное секвенирование, генетический полиморфизм, базы данных, инфологическое моделирование

Введение

Полное секвенирование генома или секвенирование только функционально значимых регионов генома человека позволяет одновременно идентифицировать множество сайтов генетического полиморфизма, имеющих диагностическую или прогностическую значимость в отношении многих заболеваний человека [1, 2]. В настоящее время на первое место выходит не проблема получения первичных данных геномного секвенирования, необходимых для идентификации сайтов генетического полиморфизма, а трудности хранения, предобработки и анализа таких данных, которые эффективно решаются только через тесную кооперацию специалистов из молекулярной биологии, математики и информатики. Стандартным современным подходом в решении означенной проблемы является комбинирование доступных on-line/off-line компьютерных программ, каждая из которых позволяет осуществить тот или иной этап по идентификации сайтов генетического полиморфизма [3]. Однако такой подход имеет ряд недостатков. В частности, он времязатратен, особенно при использовании on-line сервисов, так как требует многократных повторных загрузок больших массивов данных. Кроме того, при таком подходе конечный пользователь может воспользоваться только теми возможностями, которые заложены в программу разработчиком, что накладывает ограничения по контролю всех этапов анализа, их точности и надежности, а также гибкости проведения анализа и его адаптации под нужды научно-исследовательских и диагностических лабораторий Республики Беларусь. В связи с этим критически важной и актуальной задачей является создание единой универсальной среды или программной инфраструктуры для хранения и обработки экспериментальных данных в ходе решения задачи идентификации сайтов генетического полиморфизма.

Анализ геномных последовательностей с целью поиска вариаций нуклеотидов в генах живых организмов выполняется с помощью R-пакетов и библиотек. Достоинствами пакетов является общедоступность и открытость распространения. Недостатки – характерны для некоммерческого программного обеспечения. Автоматизация программных средств отсутствует, параметры методов и моделей анализа данных задаются пользователем вручную. Промежуточные и конечные результаты обработки записываются в виде несистематизированных наборов файлов, а их местоположение документируется отдельно, часто в блокноте или лабораторном журнале. Для восстановления логической информации о выполненном анализе требуется упорядочивание множества взаимосвязанной информации о данных, включая сведения о версиях программ, референсных базах данных, протоколах подготовки

потоков данных и прочее. Для надежной воспроизводимости результатов анализа необходимо хранение гораздо большего объема данных, чем непосредственно самих экспериментальных данных и системных параметров настройки алгоритмов. Следовательно, требуется разработка специализированной инфраструктуры для хранения и обработки наборов больших данных геномного секвенирования.

Для структурированного и целостного хранения набора взаимосвязанных данных наилучшим решением является применение технологий баз данных (БД), основанных на реляционных моделях данных [4]. База данных выступает центральным ядром системы обработки и анализа геномных данных. Она позволит обеспечить также хранение актуальных версий программ анализа данных и автоматизировать весь процесс анализа. Хранение в БД множества взаимосвязанных данных предоставляет возможность их дополнительного статистического анализа, что позволяет повысить детализацию результатов анализа и физическую интерпретацию данных.

Цель работы – разработка базы данных хранения, предобработки и анализа данных геномного секвенирования для идентификации сайтов генетического полиморфизма.

Разработка базы данных

Файловая организация потоков данных, генерируемых в ходе решения задачи поиска вариаций нуклеотидов, показана на рис. 1.

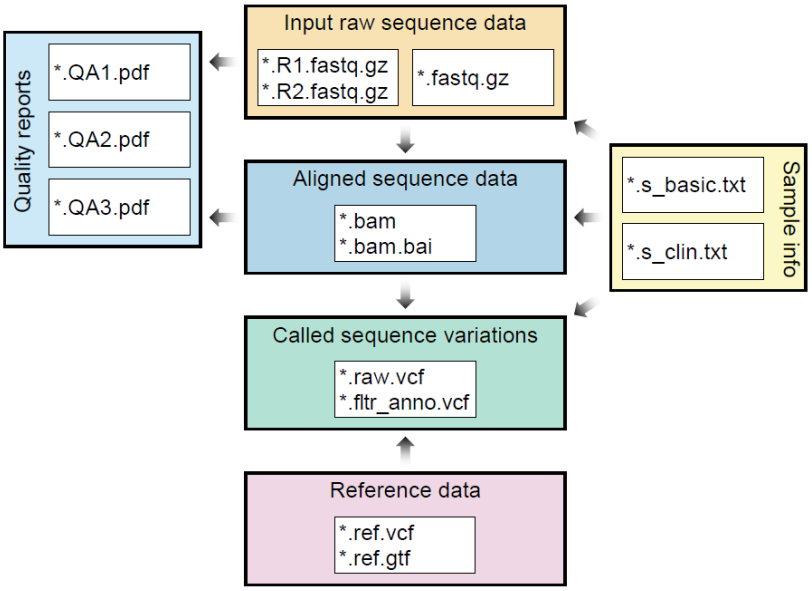

Рис. 1. – Схема потоков файлов геномных данных.

Прочитанные нуклеотидные последовательности в формата FASTQ сжатые архиватором Gzip, являются исходными данными (блок Input raw sequence data). Одно прочтение представлено двумя файлами, в названии которых включены суффиксы R1 и R2. Файлы могут иметь размер до десятков гигабайт. Файлы сопровождаются дополнительным описанием (блок Sample info) условий и параметров измерений, вносимых инженером-оператором, такими как пол и возраст пациента, параметрами оборудования, протокол измерения. Анализ включает несколько этапов, направленных на оценку качества данных (блок Qality reports), картирование последовательностей относительно референсного генома (блок Aligned sequence data), поиск и аннотирование нуклеотидных вариаций (блок Called sequence variations). Картирование прочтений и аннотирование идентифицированных сайтов полиморфизма требует референсных

данных о нуклеотидной последовательности целевого генома и известных полиморфных локусах (блок Reference data).

Разработка БД включает этап инфологического моделирования, заключающийся в анализе предметной области и создании концептуальной модели данных, цель которой – максимально отразить семантику предметной области в терминах модели данных. Общепризнанным «золотым» стандартом является технология IDEF1X.

При анализе предметной области выделены следующие высокоуровневые сущности: Samples (Образцы), RawData (Измеренные данные), Analyses (Анализы), Results (Результаты), ReferenceData (Референсные данные) и QualityReports (Отчеты по качеству). Для хранения детализированной информации о биологическом образце введены элементы Patient (Пациент) и Equipment (Оборудование). Между сущностями Analyses и RawData установлено отношение «многие-ко-многим» с помощью связующей сущности AnalysisConfiguration (Конфигурация анализа). Для повышения универсальности принято решение не выделять отдельные сущности под конкретные этапы анализа, а объединить их в одну – Results (Результаты), с указанием типа анализа. Сущность Results может хранить результаты, представленные от нескольких байт до десятков мегабайт данных. Для хранения параметров анализа введена сущность Settings (Настройки). Задание параметров в виде пары {Название параметра, Значение параметра} позволяет инициализировать произвольное количество всевозможных параметров. Введены сущности-справочники Targets (Назначения), HGSV_types (Типы HGSV), ResultTypes (Типы результатов), позволяющие добавлять новые виды и типы данных и результатов анализа. Перечисленные свойства обеспечивают модели данных универсальность и инвариантность к различным видам анализа.

Моделирование схемы БД реализовано с помощью системы CASE (Computer-aided Software Engineering) DBShema (https://dbschema.com). Разработанная схема БД представлена на рис 2. Схема соответствует требованиям стандарта IDEF1X с отображением связей по нотации «Crow’s feet» («воронья ножка»).

Концептуальная схема БД содержит 16 сущностей и может быть транслирована в реляционную модель, предоставляя уровень нормализации не ниже 3 нормальной формы. В схеме учтены необходимые ограничения на диапазон значений (поля Sex и Age) и ограничения ссылочной целостности. По внешним ключам, осуществляющим связь с сущностями, для которых предполагается много экземпляров, созданы индексы, позволяющие повысить скорость выполнения многотабличных запросов, поиска и фильтрации данных. С помощью специальных индексов реализовано требование уникальности значений полей FileName, id_Analysis и id_RawData. Прочие поля поиска, такие как ID, SubmissionID, FileName, сущности RawData, также проиндексированы.

DBShema содержит встроенные средства прямого и реверсного инжиниринга, позволяющего непосредственно создавать (не)реляционные БД на основе разработанной модели и обновлять модель в результате изменений, внесенных средствами СУБД. Разработанная модель данных транслирована в физическую модель (учтены типы данных, размеры полей и ограничения) и развернута на сервере БГУ под управлением СУБД SQL Server.

Заключение

Разработана инфраструктура хранения данных геномного секвенировании и результатов их анализа для идентификации сайтов генетического полиморфизма. Концептуальная схема БД характеризуется надежностью и гибкостью, обеспечивает согласованное и целостное хранение экспериментальных данных и результатов их анализа. Модель данных обладает достаточной универсальностью и позволяет хранить результаты разнообразных видов анализа.

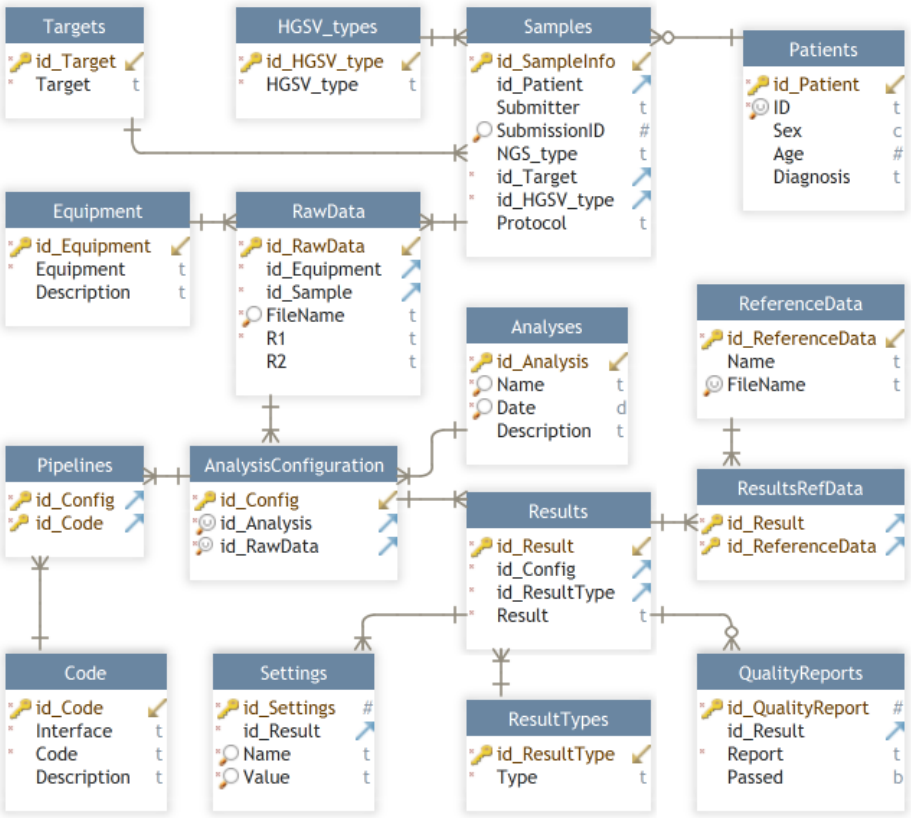

Рис. 2. – IDEF1X схема данных результатов анализа геномных последовательностей. Обозначения на схеме: пиктограмма ключа – первичный ключ; точки слева от названия поля – обязательные к заполнению поля; стрелки справа от названия поля – задание ограничений ссылочной целостности; символы c, d, t, # – типы данных, пиктограмма лупы –индексация поля.

Литература

1. Petersen B. S. et al. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genetics. 2017. Vol. 18(1):14.
2. Rabbani B. et al. Next generation sequencing: implications in personalized medicine and pharmacogenomics. Molecular Biosystems. 2016. Vol. 12, No. 6. P. 1818–1830.
3. De Los Campos G. et al. Complex-Trait Prediction in the Era of Big Data. Trends in Genetics. 2018. Vol. 34, No. 10). P. 746–754.
4. Скакун В. В. Системы управления базами данных: пособие. Минск: БГУ, 2020. 159 с.

Development of infrastructure for storing genomic sequencing data and the results of their analysis

V.V. Skakun, M.M. Yatskou, V.V. Grinev

Belarusian State University, Minsk, e-mail: skakun@bsu.by

Genome sequencing makes it possible to identify the sites of genetic polymorphism in order to diagnose human genetic diseases. There is no automation of the analysis process, the parameters of methods and models of data analysis are set manually by the user. Unsystematic storage of a lot of data and the results of their processing greatly complicates their subsequent use. All this requires the development of a unified and high-performance software package. The article describes the development process of the basic element of the package – the infrastructure for storing genetic polymorphism data and the results of their analysis.

Keywords: genome sequencing, genetic polymorphism, databases, infological modelling.