doi 10.18699/bgrs2024-8.4-08

Оптимизация биоинформатической обработки данных 16S секвенирования микробиоты кишечника

Скоповец Е.Я.*, Вертелко В.Р.
РНПЦ детской онкологии, гематологии и иммунологии, Минск, Беларусь
** Skopovets@yandex.ru*

Ключевые слова: биоинформатика; 16S; ДНК; микробиота

Мотивация и цель: В последние десятилетия технологии секвенирования нового поколения существенно повлияли на исследования микробиома человека, позволяя лучше понять и охарактеризовать взаимодействия микробиома и хозяина. Поскольку большинство видов кишечных микробов трудно культивировать, для изучения кишечного микробиома широко применяются технологии секвенирования «нового» поколения, включая 16S рРНК, 18S рРНК, секвенирование внутреннего транскрибируемого спейсера (ITS), метагеномное секвенирование методом дробовика, метатранскриптомное секвенирование и виромное секвенирование. Целью исследования являлась оптимизация анализа данных 16S секвенирования микробиоты кишечника пациентов с онкогематологическими заболеваниями с помощью разработанного программного ресурса MicrobiotaAnalysisToolkit.
Методы и алгоритмы: Материалом для исследования послужили fastq файлы, полученные в результате парноконцевого NGS секвенирования микробиоты кишечника 24 здоровых детей и 43 пациентов с врожденными дефектами иммунной системы (ПИД) с помощью платформы Illumina Miseq (v3 600 циклов). Биоинформатическая обработка данных проводилась с использованием языка Python. В основе работы программного ресурса MicrobiotaAnalysisToolkit – конвейер DADA2. Назначение таксономии проводилось с помощью базы данных SILVA и Greengenes.
Результаты: Метод 16S секвенирования кишечной микробиоты включает множество этапов: забор материала у пациентов, выделение ДНК и подготовку библиотеки для секвенирования, непосредственно секвенирование, биоинформатическую и статистическую обработку [1].
Биоинформационная обработки данных метагеномного секвенирования включает следующие этапы (рис. 1):
1. Удаление адаптеров, фильтрация и демультиплексирование прочтений.
2. Формирование обучающей выборки и последующее обучение модели на основе алгоритма наивного Байеса для классификации таксономической принадлежности
3. Присвоение таксономии с помощью баз данных GreenGenes и SILVA.
4. Построение сводных графиков и таблиц на основе полученных результатов.

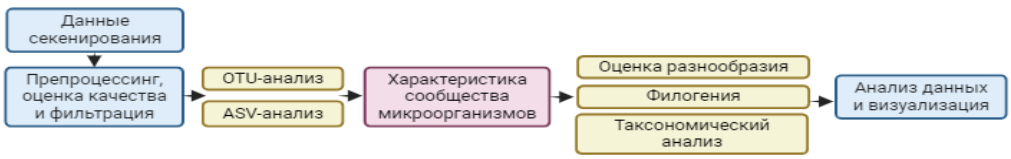

Рис. 1. Алгоритм анализа данных секвенирования 16S рРНК

Для осуществления этапов оценки качества и фильтрации данных секвенирования использован алгоритм DADA2, в частности плагин q2-dada2. Для работы с парноконцевыми данными секвенирования, в плагине q2-dada2 использован подмодуль denoise_paired. При данных вводных параметрах выполняли фильтрацию по качественному признаку, проверяли библиотеку на наличие химерных последовательностей и объединяли парноконцевые прочтения в единый пул данных.

Для работы подмодуля denoise_paired требовалась настройка параметров, которые подобрали эмпирически. Для используемого набора данных (более чем из 30 пациентов) удалось выделить ряд параметров, которые позволили улучшить качественную характеристику набора данных: p-trim-left-f и p-trim-left-r (параметры указывающие сколько нуклеотидов необходимо убрать с начала каждого фрагмента, для прямых и обратных прочтений соответственно), p-trunc-len-f и p-trunc-len-r (параметры указывающие сколько нуклеотидов необходимо оставить, для прямых и обратных прочтений соответственно), p-min-fold-parent-over-abundance (параметр указывающий минимальное количество перекрывающихся нуклеотидов, используется при объединении парноконцевых прочтений), p-max-ee-f и p-max-ee-r (параметры определяющие максимальное количество возможных несоответствий при объединении прочтений). В дальнейшем использовали обученную модель на основе алгоритма наивного Байеса для классификации таксономической принадлежности сформированных ранее последовательностей.

Завершающий этап присвоение таксономии проводили с использованием двух доступных баз данных: SILVA или Greengenes. Определение эталонной базы данных может существенно повлиять на все результаты анализа: процесс группировки/кластеризации бактерий в значительной степени привязан к выбранной эталонной базе данных, что напрямую влечет за собой риск получения противоречивых результатов при прямом сравнении (рис. 2) [2, 3].

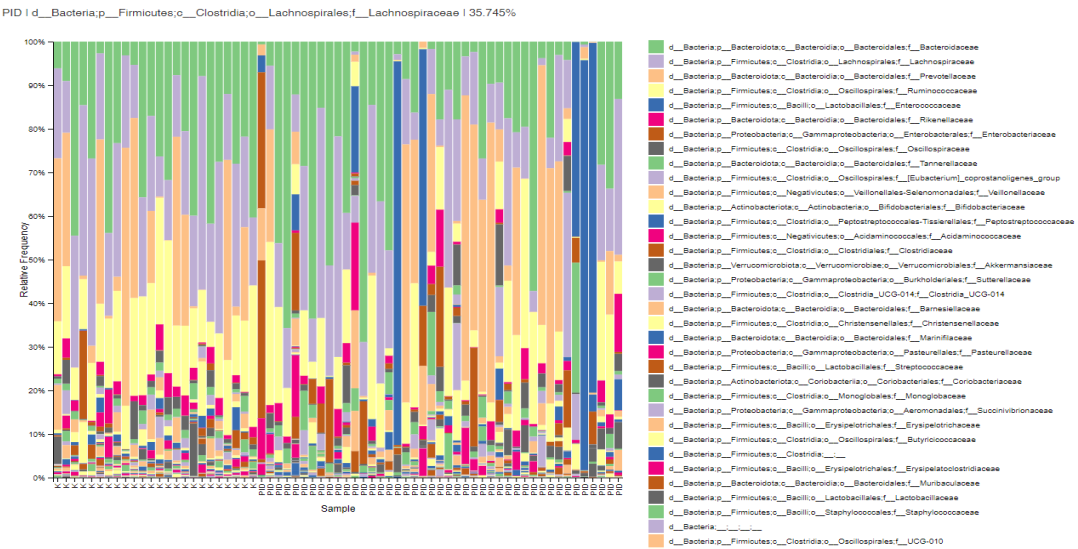

Рис. 2. Таксономический состав микробиоты кишечника на уровне семейства здоровых лиц и пациентов с врожденными дефектами иммунной системы

Процент идентичности группировки также играет важную роль, поскольку и Greengenes, и SILVA обеспечивают разные уровни идентичности, при этом 97 % является наиболее популярным и общепринятым: считается, что штаммы, принадлежащие к одному и тому же виду бактерий, имеют разницу < 3 % в их последовательности 16S рРНК.
Выбор базы данных влияет не только на таксономический состав (особенно на уровне рода), но и на все филогенетические оценки, как на метрики альфа-разнообразия, так и на расстояния бета-разнообразия, поскольку каждая база данных имеет свой определенный множественно-выровненный эталонный набор в качестве руководства по составлению филогенетического дерева.
Следующий шаг – это вычисление ряда общих показателей разнообразия в нашей функциональной таблице. Для этого использован плагин core-metrics-phylogenetic и пайплайн, представляющий собой набор различных команд для QIIME2.
Выводы: Для осуществления процедуры метагеномного анализа, необходимо обладать знаниями в области медицины и программирования, системного администрирования, а также ряд узкоспециализированных навыков из области биоинформатики, математического моделирования и статистики, что в значительной степени увеличивает необходимый уровень квалификации специалиста. В свою очередь, использование предлагаемого веб-ресурса, позволяет автоматизировать все описанные ранее этапы обработки – в значительной степени снизив уровень вхождения и, что немаловажно, повысив удобство и скорость обработки, так как конечному пользователю необходимо лишь загрузить данные через графический интерфейс, выбрать из встроенного перечня необходимые параметры и нажать кнопку запуска.

Optimization of bioinformatics analysis of 16S sequencing data of gut microbiota

Skopovets K.Ya.*, Vertelko V.R.
Belarusian Research Center for Pediatric Oncology, Hematology and Immunology, Minsk, Belarus
** Skopovets@yandex.ru*

Key words: bioinformatics; 16S; DNA; microbiota

Motivation and Aim: In recent decades, next-generation sequencing technologies have significantly impacted human microbiome research, allowing for a better understanding and characterization of microbiome-host interactions. Because most gut microbial species are difficult to culture, “next-generation” sequencing technologies, including 16S rRNA, 18S rRNA, internal transcribed spacer (ITS) sequencing, metagenomic shotgun sequencing, metatranscriptome sequencing, and virome sequencing, have been widely used to study the gut microbiome. The purpose of the study was to optimize the analysis of 16S sequencing data of the intestinal microbiota of patients with oncohematological diseases using the developed software resource MicrobiotaAnalysisToolkit.
Methods and Algorithms: The material for the study was fastq files obtained as a result of paired-end NGS sequencing of the intestinal microbiota of 24 healthy children and 43 patients with inborn errors of immunity (PID) using the Illumina Miseq platform (v3 600 cycles). Bioinformatics data processing was carried out using the Python language.

The operation of the MicrobiotaAnalysisToolkit software resource is based on the DADA2 pipeline. Taxonomy assignment was done using the SILVA and Greengenes database.

Results: The 16S intestinal microbiota sequencing method includes many stages: collecting material from patients, DNA extraction and library preparation for sequencing, sequencing itself, bioinformatics and statistical processing [1].

Bioinformatics processing of metagenomic sequencing data includes the following steps (Fig. 1):

1. Removing adapters, filtering and demultiplexing reads
2. Formation of a training sample and subsequent training of the model based on the Naive Bayes algorithm for taxonomic classification
3. Taxonomy assignment using GreenGenes and SILVA databases.
4. Construction of summary graphs and tables based on the results obtained.

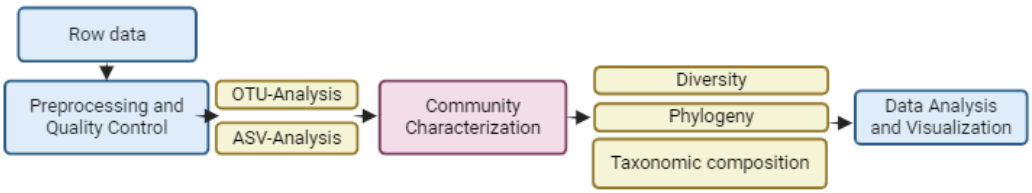

Fig. 1. Algorithm for analysis 16S rRNA sequencing data

To carry out the stages of quality assessment and filtering of sequencing data, the DADA2 algorithm, in particular the q2-dada2 plugin, was used. To work with paired-end sequencing data, the q2-dada2 plugin uses the denoise_paired submodule. With these input parameters, filtering was performed based on qualitative criteria, the library was checked for the presence of chimeric sequences, and paired-end reads were combined into a single data pool.

For the denoise paired submodule to work, it was necessary to configure parameters that were selected empirically. For the data set used (more than 30 patients), it was possible to identify a number of parameters that made it possible to improve the qualitative characteristics of the data set: p-trim-left-f and p-trim-left-r (parameters indicating how many nucleotides need to be removed from the beginning of each fragment, for forward and reverse reads, respectively), p-trunc-len-f and p-trunc-len-r (parameters indicating how many nucleotides need to be left, for forward and reverse reads, respectively), p-min-fold-parent-over- abundance (a parameter indicating the minimum number of overlapping nucleotides, used when merging paired-end reads), p-max-ee-f and p-max-ee-r (parameters defining the maximum number of possible mismatches when merging reads). Subsequently, a trained model based on the Naive Bayes algorithm was used to classify the taxonomic affiliation of the previously generated sequences.

The final stage of taxonomy assignment was carried out using two available databases: SILVA or Greengenes. The definition of a reference database can have a significant impact on the overall results of the analysis: the process of grouping/clustering bacteria is largely tied to the selected reference database, which directly entails the risk of producing inconsistent results in direct comparisons (Fig. 2) [2, 3].

Grouping percentage identity also plays an important role, as both Greengenes and SILVA provide different levels of identity, with 97 % being the most popular and generally accepted: strains belonging to the same bacterial species are considered to have < 3 % difference in their 16S rRNA sequences.

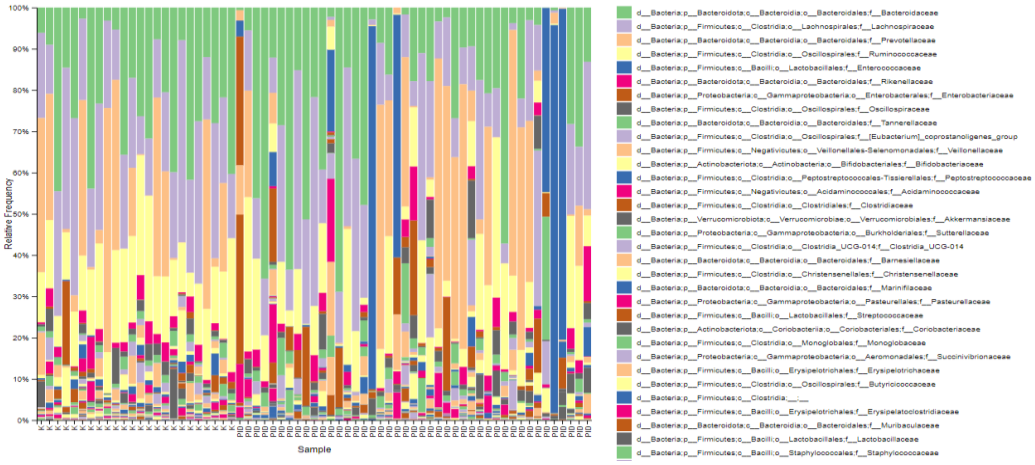

Fig. 2. Taxonomic composition of the intestinal microbiota at the family level of healthy individuals and patients with inborn errors of immunity

The next step is to calculate a number of common diversity indicators in our functional table. For this, the core-metrics-phylogenetic plugin and pipeline, which is a set of various commands for QIIME2, were used.

Conclusion: To carry out the metagenomic analysis procedure, it is necessary to have knowledge in the field of medicine and programming, system administration, as well as a number of highly specialized skills in the field of bioinformatics, mathematical modeling and statistics, which significantly increases the required level of specialist qualifications. In turn, the use of the proposed web resource allows you to automate all the previously described processing stages – significantly reducing the level of entry and, importantly, increasing the convenience and speed of processing, since the end user only needs to download data through a graphical interface, select from the built-in list the required parameters and press the start button.

Список литературы/References

1. Gao B., Chi L., Zhu Y., Shi X., Tu P., Li B., Yin J., Gao N., Shen W., Schnabl B. An introduction to next generation sequencing bioinformatic analysis in gut microbiome studies. *Biomolecules*. 2021;11(4):530. doi 10.3390/biom11040530
2. Yen S., Johnson J.S. Metagenomics: a path to understanding the gut microbiome. *Mamm Genome*. 2021;32(4):282-296. doi 10.1007/s00335-021-09889-x
3. Liu Y.X., Qin Y., Chen T., Lu M., Qian X., Guo X., Bai Y. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell*. 2021;12(5):315-330. doi 10.1007/s13238-020-00724-8