

Пайплайн для обработки данных метабаркодинга грибов природных сообществ

Ишманов Т.^{1*}, Звягина Е.^{1,2}, Рудыкина Е.¹, Филиппов И.¹, Бульонкова Т.¹, Добрынина А.¹, Филиппова Н.¹

¹ Югорский государственный университет, Ханты-Мансийск, Россия

² Московский государственный университет, Москва, Россия

* ishmanov2003tagir@mail.ru

Ключевые слова: метабаркодинг; eDNA; микобиота; DNA-derived-data; Мухрино; Западная Сибирь

Мотивация и цель: Оценка разнообразия микробиомов природных сообществ в последнее десятилетие претерпела существенный рывок с развитием методов метабаркодинга. На данный момент для обработки данных секвенирования eDNA критически необходима оптимизация набора и параметров работы отдельных модулей обработки под особенности полученных данных (наличие остатков адаптеров, праймеров и т. д.) и систематическую группу исследуемых организмов (растения, грибы, беспозвоночные, ...).

Для использования метабаркодинга в анализе структуры разнообразия грибных сообществ природных экосистем был собран и оптимизирован пайплайн, позволяющий эффективно обработать последовательности ITS2 грибов, полученные на платформе Illumina MiSeq.

Методы и алгоритмы: Пайплайн подготовлен на Python 3 в интерфейсе Artifact API для QIIME 2 [1], который позволяет модифицировать параметры в зависимости от целей исследования и качества сырых данных. Этапы пайплайна: 1) подача файлов fastq; 2) поиск и удаление адаптеров секвенирования функцией `trim_paired` (`adapter_f`, `adapter_r`) (последовательности адаптеров AmpliSeq были взяты из протокола 16S Metagenomic Sequencing Library Preparation); 3) объединение парных чтений при помощи `merge_pairs` (`minlen` = 150, `maxns` = 5, `maxee` = 2); 4) фильтрация по качеству с использованием функции `q_score` (`min_quality` = 20, `max_ambiguous` = 2); 5) дерепликация функцией `dereplicate_sequences`; 6) кластеризация до OTU с целью увеличения производительности программы (Operational Taxonomic Units) функцией `cluster_features_de_novo` (`perc_identity` = 0.99); 7) удаление химер в `uchime_ref`; 8) идентификация организмов, где каждая OTU классифицируется на таксономический уровень, была произведена при помощи `classify_sklearn(classifier)`, где `classifier` – это предварительно обученный на таксономической базе UNITE (version 9.0 2022.10.16) [2] (наивный байесовский классификатор `fit_classifier_naive_bayes` (`reference_reads`, `reference_taxonomy`)); 9) получение набора данных в виде двух архивов формата .qza (таблица частот и последовательности ДНК таксонов), пригодном для дальнейшего анализа разнообразия в QIIME 2.

Данный пайплайн был апробирован для анализа результатов секвенирования тотальной ДНК 210 образцов из проб торфа, растительного опада, древесины и микоризных корневых окончаний, отобранных в 2022 и 2023 годах на верховом болоте Мухрино в окрестностях города Ханты-Мансийск с целью оценки

разнообразия грибов верховых болот. ДНК выделяли набором SileksMagNA. Для метабаркодинга использовали регион ITS2 как наиболее информативный и традиционный для метабаркодинга грибов [3] и обеспеченный курируемыми референсными базами, такими как, например, UNITE. Секвенировали участок ITS2 в границах праймеров ITS7 и ITS4 [4, 5] по протоколу 16S Metagenomic Sequencing Library Preparation [6] на платформе Illumina MiSeq. Были получены сырые данные с разбросом от 1,5К до 250К прочтений на 1 пробу ДНК, из которых были отброшены пробы с числом прочтений < 10000. Сырые чтения (архивы FastQ и таблица метаданных) были загружены в NCBI Sequence Reads (номер биопроекта PRJNA1007262) и представлены в виде обработанной матрицы OTU на портале GBIF [7].

Результаты: Анализ последовательностей выявил 1609 операционных таксономических единиц (OTUs), классифицированных до 686 видов, 496 родов, 246 семейств, 89 отрядов, 33 классов, 9 типов и одного царства с уровнем сходства 99 %. Около 47 % таксонов идентифицированы на уровне вида, 22 % – на уровне рода, остальные – на более высоких таксономических уровнях. Основные порядки: Leotiomycetes, Dothideomycetes, Eurotiomycetes и др. (Ascomycota) и Agaricomycetes, Tremellomycetes, Mycobotriomycetes и Exobasidiomycetes, (Basidiomycota).

Анализ структуры сообществ, выполненный для всех проб методом ординации, выявил четыре выделяющихся облака точек, соответствующих основным типам субстратов: поверхностный торф, древесина, опад растений, и глубинный торф. В свою очередь, ординация проб из каждой субстратной группы показала дробление в подгруппы по другим параметрам среды (например, пробы из опада растений – по видам хозяев). Полученные результаты свидетельствуют о достоверности полученных данных о видовом разнообразии и структуре сообществ грибов и корректной работе пайплайна.

Для проверки качества работы пайплайна также была произведена ручная классификация последовательностей грибов-макромицетов. Данную группу грибов предварительно изучали на участке Мухрино в течение 9 лет на постоянных площадках; была создана база последовательностей ITS и опубликован список видов [8]. При помощи BLAST [9] метабаркодинговые последовательности сравнивали с базой ITS плодовых тел видов, собранных на данном участке, и последовательностями ITS типовых образцов, представленных в NCBI. Последовательности были сгруппированы и проанализированы по семействам Agaricaceae, Auriscalpiaceae, Boletaceae, Clavariaceae, Cortinariaceae, Crepidotaceae, Entolomataceae, Hygrophoraceae, Inocybaceae, Lycoperdaceae, Lyophyllaceae, Mycenaceae, Omphalotaceae, Paxillaceae, Physalacriaceae, Pluteaceae, Psathyrellaceae, Russulaceae, Strophariaceae, Suillaceae, Thelephoraceae, Tricholomataceae, принадлежащих к трем основным порядкам Agaricales, Boletales и Russulales.

Результаты ручной классификации полностью совпали с полученными посредством пайплайна на уровне класса, порядка, семейства и рода. Однако на видовом уровне 23 % (27 из 118 видов) получили другие названия. В результате изучения проб грунта и опада список макромицетов на модельном участке верхового болота после метабаркодинга составил 161 вид. Выявленное разнообразие увеличилось на 66 видов (40 % от общего числа видов).

Выводы: подготовленный пайплайн позволяет корректно проанализировать секвенированные данные и с уверенностью определить до рода большинство последовательностей ITS2 грибов, полученных из проб грунта и растительных остатков. Классификация до видового уровня может успешно осуществляться с использованием ручной проверки на небольших датасетах. Данный пайплайн является частью разрабатываемого программного обеспечения, которое позволит в перспективе анализировать метабаркодинговые данные до уровня структуры сообществ.

Финансирование: Субсидия из федерального бюджета на выполнение государственного задания «Молекулярно-генетические методы в изучении и оценке состояния биоразнообразия Северных регионов (FENG-2024-0003)» (№ 1023041300017-6-1.6.4 от 13.03.2024).

Pipeline for processing metabarcoding data of fungi from natural communities

Ishmanov T.^{1*}, Zvyagina E.^{1,2}, Rudykina E.¹, Filippov I.¹, Bulyonkova T.¹, Dobrynina A.¹, Filippova N.¹

¹ Yugra State University, Khanty-Mansiysk, Russia

² Moscow State University, Moscow, Russia

* ishmanov2003tagir@mail.ru

Key words: metabarcoding; eDNA; mycobiota; DNA-derived-data; Mukhrino; Western Siberia

Motivation and Aim: The assessment of microbiome diversity in natural communities has undergone a significant breakthrough in the last decade with the development of metabarcoding methods. However, at the present stage, an important challenge of eDNA sequencing data processing is the need to optimize the set and operating parameters of individual processing modules for the characteristics of the NGS-data obtained (presence of adapter residues, primers, etc.) and the systematic group of barcoded organisms (plants, fungi, invertebrates, etc.) To use the Illumina NGS metabarcoding in analyzing the diversity structure of fungal communities in natural ecosystems, a pipeline was assembled and optimized.

Methods and Algorithms: The pipeline was prepared in Python 3 in the Artifact API for QIIME 2 [1], which allows modifying the parameters depending on the goals of the study and the quality of the raw data. Pipeline stages: 1) submission of fastq files; 2) search and removal of sequencing adapters using the trim_paired function (adapter_f, adapter_r) (AmpliSeq adapter sequences were taken from the 16S Metagenomic Sequencing Library Preparation protocol); 3) combining paired reads using merge_pairs (minlen = 150, maxns = 5, maxee = 2); 4) filtering by quality using the q_score function (min_quality = 20, max_ambiguous = 2); 5) dereplication with the dereplicate_sequences function; 6) clustering to OTUs (Operational Taxonomic Units) in order to increase the performance of the program using the function cluster_features_de_novo (perc_identity = 0.99); 7) removal of chimeras in uchime_ref; 8) identification of organisms, where each OTU is classified to a certain taxonomic level, was carried out using classify_sklearn(classifier), where the classifier was pre-trained on the UNITE taxonomic database (version 9.0 2022.10.16) [2] (naive Bayes classifier

fit_classifier_naive_bayes (reference_reads, reference_taxonomy)); 9) obtaining a data set in the form of two archives in .qza format (table of frequencies and DNA sequences of taxa), suitable for further analysis of diversity in QIIME 2.

This pipeline was used to analyze the results of total DNA sequencing of 210 samples from peat, plant litter, wood and mycorrhizal root endings. The samples were collected in 2022 and 2023 in the Mukhrino raised bog near the city of Khanty-Mansiysk. DNA was isolated using the SileksMagNA kit. We used the ITS2 region for metabarcoding as it is traditionally considered to be the most informative for metabarcoding of fungi [3], and provided with curated reference databases, such as UNITE. The ITS2 region was sequenced within the boundaries of the ITS7 and ITS4 primers [4, 5] according to the 16S Metagenomic Sequencing Library Preparation protocol [6] on the Illumina MiSeq platform. Raw data was obtained ranging from 1.5K to 250K reads per 1 DNA sample, from which samples with the number of reads < 10,000 were discarded. Raw reads (FastQ archives and metadata table) were uploaded to NCBI Sequence Reads (bioproject number PRJNA1007262) and presented as a processed OTU matrix in the GBIF portal [7].

Results: Sequence analysis identified 1609 operational taxonomic units (OTUs), classified into 686 species, 496 genera, 246 families, 89 orders, 33 classes, 9 phyla and one kingdom with a similarity level of 99 %. About 47 % of taxa have been identified to the species level, 22 % to the genus level, and the rest to higher taxonomic levels. The main orders revealed were the Leotiomycetes, Dothideomycetes, Eurotiomycetes, etc. (Ascomycota) and Agaricomycetes, Tremellomycetes, Mycobotriomycetes and Exobasidiomycetes, (Basidiomycota). Analysis of the community structure, performed for all samples using the ordination method, revealed four prominent clouds of points corresponding to the main types of substrates: surface peat, wood, plant litter, and deep peat. In turn, samples from each group are combined into their own subgroups according to other environmental parameters (for example, samples from plant litter – by the host species). Our results indicate the reliability of the data obtained on the species diversity and structure of fungal communities and the correct operation of the pipeline. To check the quality of the pipeline, manual classification of sequences belonging to fleshy fungi was also performed. This group of fungi was previously studied at the Mukhrino site for 9 years on permanent sites, a reference database of ITS sequences from fungal basidiomata was created, and a list of species has been published [8]. We have compare the metabarcoding sequences with the reference database and Type specimens sequences using BLAST [9]. The main families Agaricaceae, Auriscalpiaceae, Boletaceae, Clavariaceae, Cortinariaceae, Crepidotaceae, Entolomataceae, Hygrophoraceae, Inocybaceae, Lycoperdaceae, Lyophyllaceae, Mycenaceae, Omphalotaceae, Paxillaceae, ae, Pluteaceae, Psathyrellaceae, Russulaceae, Strophariaceae, Suillaceae, Thelephoraceae, Tricholomataceae, belonging to the three main orders Agaricales, Boletales and Russulales were analyzed. The results of manual classification completely matched those obtained using the pipeline at the level of class, order, family and genus. However, at the species level, 23 % (27 of 118 species) received different names. As a result of studying soil and litter samples, the list of fleshy fungi in the model area of the raised bog was expanded by 66 species (40 % of the total list).

Conclusion: the prepared pipeline allows us to correctly analyze the sequenced data and confidently determine to the genus the majority of ITS2 sequences of fungi obtained from soil and plant residue samples. Species-level classification can be successfully accomplished using manual inspection on small datasets. This pipeline is a part of

currently developed software that will allow analyzing metabarcoding data down to the community structure level.

Funding: Subsidy from the federal budget for the implementation of the state task “Molecular genetic methods in the study and assessment of the state of biodiversity of the Northern regions (FENG-2024-0003)” (No. 1023041300017-6-1.6.4 dated 03/13/2024).

Список литературы/References

1. Bolyen E., Rideout J.R., Dillon M.R. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol.* 2019;37:852-857
2. Abarenkov K., Nilsson R.H., Larsson K.-H. et al. The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. *Nucleic Acids Res.* 2023;52(D1):D791-D797. doi 10.1093/nar/gkad1039
3. Tedersoo L., Bahram M., Zinger L., Nilsson R.H., Kennedy P.G., Yang T., Anslan S., Mikryukov V. Best practices in metabarcoding of fungi: From experimental design to results. *Mol Ecol.* 2022;31(10):2769-2795. doi 10.1111/mec.16460
4. White T., Bruns T., Lee S., Taylor J., Innis M., Gelfand D., Sninsky J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: *PCR protocols: A guide to methods and applications.* 1990;315-322. doi 10.1016/B978-0-12-372180-8.50042-1
5. Ihrmark K., Bödeker I.T., Cruz-Martinez K. et al. New primers to amplify the fungal ITS2 region--evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiol Ecol.* 2012;82(3):666-677. doi 10.1111/j.1574-6941.2012.01437.x
6. Sinclair L., Osman O.A., Bertilsson S., Eiler A. Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the illumina platform. *PLoS One.* 2015;10(2):e0116955. doi 10.1371/journal.pone.0116955
7. Filippova N., Zvyagina E., Rudykina E.A., Ishmanov T.F., Filippov I.V., Bulyonkova T.M., Dobrynina A.S. DNA-based occurrence dataset on peatland fungal communities studied by metabarcoding in north-western Siberia. *Biodivers Data J.* 2024;12:e119851. doi 10.3897/BDJ.12.e119851
8. Filippova N., Zvyagina E., Rudykina E., Dobrynina A., Bolshakov S. The diversity of macromycetes in peatlands: nine years of plot-based monitoring and barcoding in the raised bog "Mukhrino", West Siberia. *Biodivers Data J.* 2023;11:e105111. doi 10.3897/BDJ.11.e105111
9. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-410. doi 10.1016/S0022-2836(05)80360-2