

Выявление участков пластовов, потенциально пригодных для метабаркодинга видов растений сем. Роасеае

Криницына А.^{1, 2*}, Никитина О.^{1, 2}, Широбоков В.¹, Никитин П.¹, Северова Е.¹

¹ Московский государственный университет им. М.В. Ломоносова, Москва, Россия

² ООО «Эвоген», Москва, Россия

* ankrina@gmail.com

Ключевые слова: Роасеае; идентификация пыльцы; метабаркодинг; пластовы

Мотивация и цель: Аллергены, содержащиеся в переносимой по воздуху пыльце, вызывают астму и аллергический ринит, от которых страдают около 300 миллионов человек по всему миру. Одним из наиболее распространенных аллергенов является пыльца злаковых (Роасеае). Это семейство включает более 11 000 видов [1], большинство из которых опыляются ветром. Пыльца злаков присутствует в воздухе в течение очень длительного периода, поэтому пациенты с аллергией на пыльцу злаковых могут страдать от симптомов сезонного аллергического ринита долгое время. Пыльца всех видов злаков относится к единому палиноморфологическому типу и при проведении аэриобиологического мониторинга стандартными методами идентифицируется только до уровня семейства. В то же время аллергенность пыльцы разных видов злаков различается, поэтому возникает необходимость детализации кривых пыления. Идентификация биоразнообразия с помощью метабаркодирования представляет собой еще один вариант идентификации состава сложных смесей, в том числе для идентификации и количественного определения пыльцы и образцов из окружающей среды. Наиболее популярными генными маркерами, используемыми для филогенетической классификации, являются внутренние транскрибируемые спейсеры (ITS1 и ITS2), а также участки пластовой ДНК – часть последовательностей генов *rbcL* и *matK*. Технология секвенирования длинных участков позволяет расширить спектр маркерных последовательностей, перспективных для определения систематической принадлежности растений в сложных растительных смесях. Цель нашей работы – выявить участки пластовой ДНК, пригодные для определения видовой принадлежности пыльцы злаков в аэриобиологических образцах.

Методы и алгоритмы: На основании последовательностей полных пластовов 52 видов, относящихся к сем. Роасеае, был проведен поиск наиболее вариативных участков с дальнейшей теоретической оценкой возможности их использования для метабаркодинга. Множественное выравнивание последовательностей пластовов проводили с помощью программы MAFFT v7.511. Поиск наиболее полиморфных участков выполняли следующим образом: с помощью Jupyter notebook строили гистограмму плотности полиморфизмов по сделанному выравниванию с использованием python 3.9.12 и библиотек AlignIO и Seaborn. Алгоритмом разметки регионов по плотности гэпов в выравнивании с шагом в 5000 и 1000 для лучшего масштабирования. С использованием выявленных участков проводили кластеризацию видов в программе MEGA v 11.0.13 (maximum likelihood (ML), bootstrap = 1000). Аннотацию выбранных участков выполняли с использованием MPI-MP CHLOROBX – GeSeq (GeSeq – annotation of Organellar Genomes). Подбор праймеров на выявленные участки пластовов осуществляли в программе Primer-Blast. Получение продуктов амплификации проводили с

использованием в качестве матрицы ДНК, полученной из искусственных смесей пыльцы разных видов злаков, протокол ПЦР описан в работе [2].

Результаты: В результате проведенного анализа нами были отобраны семь наиболее вариативных фрагментов участков пластомной ДНК размером 800–1000 п. н. Из них только четыре позволили провести кластеризацию одноименных родов в общие группы. Филогенетические отношения видов отражают деревья, полученные с использованием участков r2_cpDNA_3 (фрагмент гена *rps16*) и r2_cpDNA_4 (фрагмент пластома, содержащий участок генов *rps16* и *trnQ-UUG*). Участки пластома, содержащие ген *rps16*, а также межгенные спейсеры с соседними генами, показали существенную значимость в филогенетических исследованиях разных подсемейств злаковых (например, [3, 4]). Менее достоверный результат был получен с использованием участков r7_cpDNA_1-2 (гены *trnG-UCC*, *trnT-GGU* и *trnE-UUC*). Вариабельные участки r22_cpDNA_1 (*ndhF*) и r22_cpDNA_2 (*rpl32*) оказались непригодными для кластеризации одноименных родов в общие группы. Вероятно, участок пластомного генома, содержащий гены *ndh*, отражает экологическую адаптацию, позволяющую покрытосеменным выживать во многих стрессовых условиях, поддерживая эффективный фотосинтез (Sabater, 2021). Для выбранных четырех вариантов последовательностей удалось получить продукт амплификации указанного размера с использованием ДНК из искусственных пыльцевых смесей.

Выводы: Для определения систематической принадлежности возможно использовать фрагменты пластома, содержащие фрагмент генов *rps16* и *trnQ-UUG*. Дополнительно можно использовать последовательности генов и межгенных спейсеров *trnG-UCC*, *trnT-GGU* и *trnE-UUC*.

Финансирование: Исследование выполнено в рамках госзадания Московского государственного университета им. М.В. Ломоносова № 121032500082-2.

Identification of regions of chloroplast DNA potentially useful for metabarcoding of plant species of the fam. Poaceae

Krinitina A.^{1,2*}, Nikitina O.^{1,2}, Shirobokov V.¹, Nikitin P.¹, Severova E.¹

¹ Lomonosov Moscow State University, Moscow, Russia

² LCT Evogen, Moscow, Russia

* ankrina@gmail.com

Key words: Poaceae; pollen identification; metabarcoding; plastome

Motivation and Aim: Allergens in airborne pollen cause asthma and allergic rhinitis, which affect about 300 million people worldwide. One of the most common allergens is pollen from cereals (Poaceae). This family includes more than 11,000 species [1], most of which are wind pollinated. Grass pollen is present in the air for a very long period, so patients allergic to grass pollen may suffer from symptoms of seasonal allergic rhinitis for a long period of time. The pollen of all Poaceae belongs to a single palynomorphologic type and is identified only to the family level during aerobiologic monitoring by standard methods. At the same time, allergenicity of pollen of different species differs, therefore, there is a need to detail pollen curves. Biodiversity identification using metabarcoding is another option for identifying the composition of complex mixes, including for the identification and quantification of pollen and environmental samples. The most popular gene markers used for molecular phylogeny are internally transcribed spacers (ITS 1 and ITS2), as well as regions of chloroplast DNA – part of

the *rbcL* and *matK* gene. The technology of sequencing long DNA regions allows us to expand the range of marker sequences promising for determining the identity of plants in complex mixtures. The aim of our work was to identify regions of chloroplast DNA useful for the determination the species identity of cereal pollen in aerobic samples.

Methods and Algorithms: Based on the sequences of complete cpDNA of 52 species of Poaceae, the most variable sites were searched and further theoretically evaluated for their use for metabarcoding. Multiple sequence alignment of plastome sequences was performed using the MAFFT v7.511. The search for the most polymorphic regions was performed as follows: using Jupyter notebook we constructed a histogram of polymorphism density based on the made alignment using python 3.9.12 and AlignIO and Seaborn libraries. Algorithm partitioned regions by density gaps in the alignment with steps of 5000 and 1000 for better scaling. Using the identified regions, species clustering was performed using MEGA v 11.0.13 (maximum likelihood (ML), bootstrap = 1000). Annotation of selected regions was performed using MPI-MP CHLOROBOX – GeSeq (GeSeq – annotation of Organellar Genomes). Primers for the identified plastome sites was performed using Primer-Blast. Amplification products were obtained using DNA obtained from artificial mixes of pollen of different cereal species as a matrix; the PCR protocol is described in [2].

Results: As a result of our analysis, we selected 7 most variable fragments of 800–1000 bp cpDNA regions. Only 4 of them allowed clustering of the same-named genera into common groups. Phylogenetic relationships of species are reflected by phylogenetic trees obtained using the r2_cpDNA_3 (part of *rps16* gene) and r2_cpDNA_4 (cpDNA regions containing the part of *rps16* and *trnQ*-UUG genes). cpDNA regions containing the *rps16* gene as well as intergenic spacers with neighboring genes have shown significant significance in phylogenetic studies of different subfamilies (e.g. [3, 4]). Less reliable results were obtained using the r7_cpDNA_1-2 (*trnG*-UCC, *trnT*-GGU and *trnE*-UUC genes). Variable sites of r22_cpDNA_1 (*ndhF*) and r22_cpDNA_2 (*rpl32*) were not suitable for clustering the same-named genera into common groups. It is likely that the region of the cpDNA containing the *ndh* genes reflects an ecological adaptation that allows covered seeds to survive under many stress conditions by maintaining efficient photosynthesis (Sabater, 2021). For the selected 4 sequence variants, we were able to obtain an amplification product of the indicated size using DNA from artificial pollen mixtures.

Conclusion: Use fragments of cpDNA containing a part of the *rps16* and *trnQ*-UUG genes is probably a good fit for identification of the pollen species. It is also possible to additionally use sequences of the genes and intergenic spacers *trnG*-UCC, *trnT*-GGU, and *trnE*-UUC.

Funding: The study is supported to the scientific programs of M.V. Lomonosov Moscow State University No. 121032500082-2.

Список литературы/References

1. Christenhusz M.J.M., Byng J.W. The number of known plants species in the world and its annual increase. *Phytotaxa*. 2016;261(3):201-217
2. Krinitsina A.A. et al. Aerobiological monitoring and metabarcoding of grass pollen. *Plants*. 2023;12(12):2351-2023
3. Guo-Ye G. et al. Phylogenetic relationships among *Leymus* and related diploid genera (Triticeae: Poaceae) based on chloroplast *trnQ*–*rps16* sequences. *Nor J Bot*. 2014;32(5):658-666
4. Krawczyk K. et al. Plastid super-barcodes as a tool for species discrimination in feather grasses (Poaceae: *Stipa*). *Sci Rep*. 2018;8:1924