

Веб-приложение speCOIdent для анализа результатов генотипирования криптических видов байкальских амфипод рода *Eulimnogammarus*

Мутин А.Д.*, Саранчина А.Е., Ржечицкий Я.А., Помазкин В.К., Тимофеев М.А., Дроздова П.Б.

Научно-исследовательский институт биологии, Иркутский государственный университет, Иркутск, Россия

* andreimutin97@gmail.com

Ключевые слова: Байкал; амфиподы; молекулярная филогенетика; speCOIdent; цитохром-с-оксидаза (субъединица I)

Мотивация и цель: Озеро Байкал – одно из древнейших озер на планете, примечательной особенностью которого является обширное биоразнообразие фауны, насчитывающей более 2500 видов [1]. Наиболее значимым и многочисленным отрядом байкальских беспозвоночных являются амфиподы. В Байкале амфиподы представлены более чем 350 морфологическими видами и подвидами, среди которых отмечен 99 % эндемизм, однако степень криптической изменчивости остается неясной. Недавние исследования продемонстрировали, что одни из наиболее распространенных обитателей литоральной зоны озера, *Eulimnogammarus verrucosus* (Gerstfeldt, 1858) и *Eulimnogammarus vitatus* (Dybowsky, 1874), включают в себя разные генетические линии [2]. Помимо этого, у *E. verrucosus* было обнаружен репродуктивный барьер между двумя генетическими линиями данного вида, и подобные исследования продолжаются [3]. Наличие криптической изменчивости показывает, что для эффективного изучения видового разнообразия амфипод, помимо анализа морфологических различий, требуется исследование различий на генетическом уровне – генотипирование, которое обычно проводят с помощью секвенирования по Сэнгеру. Популярным инструментом для интерпретации результатов генотипирования является поиск в базе данных NCBI GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) и Barcode of Life Database (BOLD; <https://boldsystems.org>). Однако использовать эти сервисы для идентификации генетической группы не всегда эффективно. Пользователи NCBI BLAST часто вынуждены читать соответствующую публикацию для уточнения генетической линии, поскольку в базе Genbank могут быть указаны только виды. BOLD, в свою очередь, позволяет определять генетические группы, однако идентифицирует их как BIN (Barcode Index Number), что также требует от пользователя ознакомиться с соответствующей публикацией для того чтобы соотнести какой генетической группе соответствует конкретный BIN. Помимо этого, NCBI позволяет проводить идентификацию только на основе уже опубликованных последовательностей, что усложняет рутинную работу с новыми, неопубликованными последовательностями.

Таким образом, в связи с вышеперечисленным, целью данной работы является создание удобного, веб-приложения для эффективного определения видовой

принадлежности и/или генетической линии амфипод озера Байкал путем анализа результатов генотипирования.

Методы и алгоритмы: Используемая в приложении база данных была создана с помощью языка программирования Bash на основе стандартных филогенетических маркеров (цитохром-с-оксидаза I (COI) и 18S рРНК) амфипод Байкальского региона. База данных была сформирована из новых последовательностей, полученных с помощью секвенирования по Сэнгеру и дополнена последовательностями доступными в Genbank. Веб-приложение было создано на основе языка программирования R и с помощью ключевых пакетов Shiny [4], rBLAST [5], ggplot2 [6], mapview [7], leaflet [8].

Результаты: Было проведено генотипирование некоторых видов рода *Eulimnogammarus* из озера Байкал и реки Ангара и проанализирована на наличие различных генетических групп внутри каждого из видов. На основе этих данных была создана база данных, используемая в приложении speCOIdent.

Созданное веб-приложение speCOIdent размещено на одной странице и имеет три интерактивные панели для взаимодействия с пользователем: 1) Панель для ввода последовательности и настройки параметров идентификации. Пользователь имеет возможность скопировать последовательность в окно для ввода как в формате fasta, так и в формате нуклеотидной последовательности без заголовка. Помимо этого, в приложении присутствует возможность изменять параметры идентификации, такие как «e-value» (мера вероятности того, что найденное совпадение случайно) и «% ID» (процент идентичности последовательностей). 2) Результатом идентификации является всплывающее интерактивное окно, в котором содержится информация о типе гена (COI или 18S рРНК), видовой принадлежности и генетической группе. Помимо этого, на втором окне отображается динамически масштабируемая, интерактивная карта ареала вида или генетической группы, позволяющая с высокой точностью определить известные места их обитания. 3) Также дополнительным результатом идентификации является таблица из пяти лучших совпадений, содержащая название вида, название его генетической группы, название гена, а также процент идентичности. Данная таблица является интерактивной и позволяет вывести на экран полный список параметров идентификации и выравнивание на последовательность интереса. Дополнительными функциями speCOIdent является встроенная инструкция по использованию и возможность скачать полную базу данных в формате fasta.

Помимо этого, с помощью веб-приложения speCOIdent были проанализированы последовательности COI нескольких популяций амфипод вида *E. verrucosus* из озера Байкал и реки Ангара и определены их генетические линии.

Выводы: Было создано веб-приложение speCOIdent с базой данных из последовательностей стандартных филогенетических маркеров COI и 18S рРНК для определения видовой принадлежности, генетической группы и ареала амфипод Байкальского региона путем анализа результатов генотипирования. Приложение speCOIdent сочетает в себе преимущества как NCBI BLAST, так и BOLD, позволяя быстро и с высокой точностью идентифицировать не только видовую принадлежность, но и генетическую группу, параллельно отражая масштабируемую карту ареала. Помимо этого, speCOIdent является гибкой и легко адаптируемой программой, которую можно быстро настроить на использование различных баз данных с любыми маркерными последовательностями. Данное

приложение было сделано общедоступным и размещено на сервере научно-исследовательского института биологии ИГУ по адресу (<http://bioinformatics.isu.ru:3838/speCOIdent/>). Помимо этого, исходный код speCOIdent был опубликован на веб-сервисе GitHub по адресу (<https://github.com/MutinAndrei/Shiny-web-application-for-analyzing-genotyping-results>).

Финансирование: Исследование было проведено при поддержке гранта министерства науки и высшего образования России № FZZE-2024-0008.

Web application speCOIdent for the analysis of genotyping results for cryptic species of Baikal amphipods belonging to the genus *Eulimnogammarus*

Mutin A.D.*, Saranchina A.E., Rzhechitskiy Y.A., Pomazkin V.K., Timofeyev M.A., Drozdova P.B.

Institute of Biology at Irkutsk State University, Irkutsk, Russia

* andreimutin97@gmail.com

Key words: Baikal; amphipods; molecular phylogenetics; speCOIdent; cytochrome c-oxidase (subunit I)

Motivation and Aim: Lake Baikal is one of the oldest lakes in the world, a remarkable feature of which is the rich biodiversity of its fauna, including over 2500 species [1]. Amphipods are one of the most species-rich and important groups of Baikal invertebrates. In Baikal, amphipods are represented by more than 350 morphological species and subspecies, over 99 % of which are endemic, but the degree of cryptic diversity remains unclear. Recent studies have shown that some of the most common inhabitants of the littoral zone of the lake, *Eulimnogammarus verrucosus* (Gerstfeldt, 1858) and *Eulimnogammarus vitatus* (Dybowsky, 1874), comprise different genetic lineages [2]. In addition, a reproductive barrier between two genetic lineages of *E. verrucosus* has been found, and there is more to learn [3]. The presence of cryptic diversity shows that to effectively study amphipod species diversity, it is necessary, in addition to analyzing morphological differences, to study differences at the genetic level by genotyping, which is usually performed using Sanger sequencing. Popular tools for interpreting genotyping results are search engines of the NCBI GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) and the Barcode of Life (BOLD; <https://boldsystems.org>) databases. However, using these services to identify a genetic group is not always efficient. NCBI BLAST users often need to read the relevant publication to clarify the genetic lineage, as the Genbank database may only list species. BOLD, on the other hand, provides the identification of genetic groups, but identifies them as BINs (Barcode Index Numbers), which also requires the user to consult the relevant publication to determine which genetic group a particular BIN corresponds to. In addition, NCBI only allows identification based on previously published sequences, which makes routine work with new, unpublished sequences difficult. In view of the aforementioned considerations, the objective of this study is to create a user-friendly, web-based application for the efficient determination of the species and/or

genetic lineage of amphipods of Lake Baikal by means of the analysis of genotyping results.

Methods and Algorithms: The database utilized in this application was constructed using the Bash programming language, based on standard phylogenetic markers (cytochrome c-oxidase I (COI) and 18S rRNA) of amphipods from the Baikal region. The database was constructed from newly generated sequences obtained through Sanger sequencing and augmented with sequences available in GenBank. The web application was developed using the R programming language and several key packages, including Shiny [4], rBLAST [5], ggplot2 [6], mapview [7], and leaflet [8].

Results: Genotyping of some species of the genus *Eulimnogammarus* from Lake Baikal and the Angara River was conducted and analyzed to identify the presence of different genetic groups within each species. The resulting data were used to create a database, which was subsequently incorporated into the speCOIdent application.

The speCOIdent web application is comprised of a single page and features three interactive panels for user interaction: 1) a panel for entering the sequence and setting the identification parameters. The user may input the sequence in either FASTA format or nucleotide sequence format, without a header. Furthermore, the application enables the user to modify identification parameters, such as the e-value (a measure of the probability that the match is random) and the % ID (percentage of sequence identity). 2) The identification result is presented in an interactive pop-up window, which provides information about the type of the genetic marker (COI or 18S rRNA), species affiliation, and genetic group. Additionally, a second window displays a scalable, interactive map of the known records for the species or genetic group, allowing for precise identification of its range. Additionally, the identification process yields a table of the top five matches, which includes the species name, the name of the genetic group, gene name, and identity percentage. This table is interactive, allowing the user to view the full list of identification parameters and alignment to the sequence of interest. SpeCOIdent also features a built-in instruction manual and the possibility to download the full database in FASTA format.

Using the speCOIdent web application, the COI sequences of several populations of amphipod species *E. verrucosus* from Lake Baikal and the Angara River were analyzed and their genetic lineages were determined.

Conclusion: A web application, speCOIdent, was created with a database of sequences of standard phylogenetic markers, COI and 18S rRNA. The application was used to determine the species, genetic group, and range of amphipods in the Baikal region by analyzing genotyping results. The speCOIdent application integrates the strengths of both NCBI BLAST and BOLD, enabling rapid and precise identification of species affiliation and genetic group, while reflecting a user-friendly range map. Moreover, speCOIdent is a versatile and adaptable program that can be rapidly configured to utilize disparate databases with any marker sequences. The application is now publicly available and is hosted on the server of the Institute of Biology at Irkutsk State University at <http://bioinformatics.isu.ru:3838/speCOIdent/>. Furthermore, the source code of speCOIdent has been made publicly available on the GitHub web service at (<https://github.com/MutinAndrei/Shiny-web-application-for-analyzing-genotyping-results>).

Funding: The study was financially supported by grant Ministry of Science and Higher Education of Russia No. FZZE-2024-0008.

Список литературы/ References

1. Timoshkin O.A. Lake Baikal: diversity of fauna, problems of its immiscibility and origin, ecology and “exotic” communities. In: Index of animal species inhabiting Lake Baikal and its catchment area. Vol. I. Lake Baikal. Novosibirsk: Nauka, 2001;74-113
2. Gurkov A. et al. Indication of ongoing amphipod speciation in Lake Baikal by genetic structures within endemic species. *BMC Evol Biol.* 2019;19:1-16
3. Drozdova P. et al. Experimental crossing confirms reproductive isolation between cryptic species within *Eulimnogammarus verrucosus* (Crustacea: Amphipoda) from Lake Baikal. *Int J Mol Sci.* 2022;23(18):10858
4. Chang W. et al. shiny: Web application framework for R. 2023. <https://CRAN.R-project.org/package=shiny>
5. Hahsler M., Nagar A. rBLAST: R interface for the basic local alignment search tool. 2019. <https://github.com/mhahsler/rBLAST>
6. Wickham H. ggplot2: Elegant graphics for data analysis. Berlin: Springer, 2016
7. Appelhans T., Detsch F., Reudenbach C., Woellauer S. mapview: Interactive viewing of spatial data in R. 2023. <https://cran.rproject.org/web/packages/mapview/index.html>
8. Cheng J., Schloerke B., Karambelkar B., Xie Y. leaflet: Create interactive web maps with the javaScript “Leaflet” library. 2023. <https://cran.rproject.org/web/packages/leaflet/index>