Combining NCBI and BOLD databases for OTU assignment in metabarcoding and metagenomic datasets: The BOLD_NCBI _Merger

Jan-Niklas Macher‡, Till-Hendrik Macher‡, Florian Leese‡

‡ Aquatic Ecosystem Research, University of Duisburg-Essen, Essen, Germany

Corresponding author: Jan-Niklas Macher (jan.macher@uni-due.de)

Academic editor: Masaki Miya

Abstract

Background

Metabarcoding and metagenomic approaches are becoming routine techniquesfor use in biodiversity assessment and in ecological studies. The assignment of taxonomic information to millions of sequences obtained via high-throughput sequencing is challenging, as many DNA reference libraries are lacking information on certain taxonomic groups and can contain erroneous sequences. Combining different reference databases is therefore a promising approach for maximising taxonomic coverage and reliability of results.

New information

The “BOLD_NCBI_Merger” bash script is introduced, which combines sequence data obtained from the National Centre for Biotechnology Information (NCBI) GenBank and the Barcode of Life Database (BOLD) and prepares it for taxonomic assignment with the software MEGAN.

Keywords

Biodiversity, High-throughput sequencing, Operational taxonomic unit, software, MEGAN, script, taxonomic assignment

Introduction

Background

High-throughput biodiversity assessment techniques such as metagenomics (Yu et al. 2012) and metabarcoding (Taberlet et al. 2012) produce millions of sequences in a short amount of time. These techniques are becoming standard in many fields of research (Deiner et al. 2015, Choo et al. 2017, Macher et al. 2017), as well as application (Elbrecht et al. 2017). One of the challenges connected to the analyses of millions of DNA sequences is the assignment of the obtained Operational Taxonomic Units (OTUs) to taxonomic names. Taxonomic information is often needed, especially in ecological studies and for the assessment of ecosystem health, which is largely based on the knowledge of species' ecological traits (Gayraud et al. 2003, Hering et al. 2006). Several existing databases contain millions of DNA reference sequences, which can be used to assign taxonomic names to OTUs (Santamaria et al. 2012). However, these databases are often specialised, each containing mostly data for certain genetic markers (e.g. rRNA: SILVA (Quast et al. 2012) or selected taxonomic groups (e.g. fungi: UNITE Kõljalg et al. 2005). Two of the largest reference databases are the Barcode of Life Database (BOLD, Ratnasingham and Hebert 2007), which contains mostly cytochrome c oxidase I (COI) sequences of Metazoa and the National Centre for Biotechnology Information (NCBI) GenBank database (Benson et al. 2012), which contains reference sequences of taxa from all domains of life. Sequence data is available for download via websites and/or command line applications and can be used for taxonomic assignment via different tools. This is a standard approach in metabarcoding and metagenomic studies, as it is not feasible to identify millions of sequences one by one. For the identification of sequences from metabarcoding studies targeting metazoan taxa, the BOLD Identification API (http://www.boldsystems.org/index.php/resources/api?type=idengine) is often used (e.g. Elbrecht and Leese 2015, Prosser et al. 2017, Kranzfelder et al. 2015). BLAST+ (Camacho et al. 2009) searches against the NCBI GenBank are often used for the identification of non-metazoan sequences obtained through metagenomic approaches (Hasan et al. 2014, Shi et al. 2013), as well as confirming results of the BOLD API (Kranzfelder et al. 2015, Elbrecht and Leese 2015). Web tools and APIs remotely accessing databases tend to be rather slow, making fast identification of millions of sequences and OTUs a time-consuming task. In addition, the BOLD database is somewhat restricted and does not contain all sequences that are deposited in the NCBI GenBank, which is due to the focus on genetic barcodes of metazoan taxa and of a certain length (several hundred basepairs). On the other hand, reliability of information in the curated BOLD database is expected to be higher than that in the NCBI database, although errors do occur (e.g. Lis et al. 2016). The NCBI GenBank, however, does not include all sequences available in the BOLD database, as not all scientists submit their sequences to both databases.

Studies have shown that both databases can be used to successfully identify metazoan taxa (Sonet et al. 2013), but uncertainties remain. Metagenome sequencing studies and metabarcoding studies using degenerated primers are known to produce data not only

from either microbial or metazoan taxa, but also from all trees of life (Capra et al. 2016, Macher and Leese 2017, Horton et al. 2017). For such studies, taxonomic assignment with the BOLD database only will result in the loss of information, as many non-metazoan taxa cannot be identified. Using only the NCBI GenBank can circumvent this problem, but at the cost of losing information on metazoan taxa and lowered accuracy. Combining information from both databases therefore improves both speed of identification, reliability of results and taxonomic coverage. However, although theoretically possible, studies are currently not directly combining databases in order to improve speed and accuracy of analyses. This might be partly due to the large amount of data that needs to be downloaded on to a local hard drive and the needed reformatting of data in order to make it compatible, which requires basic bioinformatic skills. Several tools for analyses and taxonomic assignment of sequences downloaded from reference databases are available and could theoretically be used with combined databases, e.g. RDP Classifier (Wang et al. 2007), KRAKEN (Wood and Salzberg 2014), SPINGO (Allard et al. 2015) and MEGAN (Huson et al. 2007).

The "BOLD_NCBI_Merger" is introduced, a bash-script that builds databases containing sequence data from both BOLD and NCBI GenBank. In the tutorial accompanying the script (Suppl. material 1), the method used to download and prepare data for analyses in the MEGAN software is explained. The built database can also be used for analyses and software other than MEGAN. MEGAN implements a lowest common ancestor (LCA) approach for taxonomic assignment of sequences and was originally developed for analyses of metagenomic datasets (Huson et al. 2007), but the LCA approach can also be used for taxonomic assignment of sequences obtained through metabarcoding (Hänfling et al. 2016, Horton et al. 2017).

Technical specification

Prior to analyses BLAST+ (v. 2.6), vsearch (Rognes et al. 2016) and MEGAN need to be installed. All analyses described in the tutorial were conducted on a Mac with OS Yosemite 10.10.5.

The bash script "BOLD_NCBI_Merger" concatenates multiple files downloaded from BOLD and NCBI, respectively. Then, COI sequences are extracted from the downloaded BOLD fasta file. COI is the most widely used gene for barcoding of metazoan taxa and most sequences deposited in the BOLD database are COI sequences. However, few sequences of other markers (e.g. 18S rRNA) are also deposited in BOLD. These sometimes get downloaded together with COI sequences and need to be removed in order for the script to work properly. Headers of both BOLD and NCBI files are formatted so that vsearch can dereplicate the sequences without cutting the header. Then, vsearch is used to dereplicate the sequences in order to prevent over-representation of sequences in the final database. In the next step, the headers are formatted so that MEGAN can identify species names. A local BLAST database is built from the concatenated BOLD and NCBI dataset. Finally, a BLAST search against the database is performed with a metabarcoding or metagenomics dataset. The resulting txt file can be imported into MEGAN and taxonomic assignments can be exported subsequently.

The detailed tutorial including all commands can be found in supplementary material 1. The package including the script used for processing and preparing sequence files can be found in supplementary material 2. Sequence data for the tutorial can be obtained from BOLD and NCBI GenBank, respectively. All Trichoptera sequences (used here as an example) can be downloaded as one fasta file from BOLD via the Public Data Portal (http://www.barcodinglife.org/index.php/Public_BINSearch?searchtype=records; search term: “Trichoptera”, “Public Data”). All Trichoptera sequences from GenBank can be downloaded from the nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/; search term: "Trichoptera AND (COI OR CO1 OR COX1 OR COXI; sequence length: 1-1000 bp)" and saved on a local hard drive. Sequences other than COI can be processed as long as the data format is the same as for the COI data.

For ease of use, a dataset containing few sequences (Trichoptera, COI barcoding region) was used for this tutorial, but it should be noted that, for reliable results and real analyses, a larger reference database containing as many taxa as possible should be used in order to prevent erroneous assignments (Porter et al. 2014, Garcia-Etxebarria et al. 2014, Ueno et al. 2014). In-depth studies, comparing different software usable for taxonomic assignment and different combinations of databases, should be conducted in order to quantify the benefits and possible pitfalls of combining data from several databases. It should also be mentioned that the approach of assigning taxonomy to OTUs by using local databases has limitations. As the created database is stored on a local hard drive, it does not receive automated updates and will age. Thus, the databases need to be updated on a regular basis. This requires some effort, since several gigabytes of data need to be downloaded from NCBI and BOLD databases, a process which can take several hours. Processing large amounts of data on a local hard drive also requires machines powerful enough to complete the task within a reasonable amount of time. Still, the approach of combining databases will be worth the efforts for many studies targeting diverse biological communities, as taxonomic assignment is fast and reliable once the local databases have been constructed and the gained information can help improve results.

Project description

Title: Combining NCBI and BOLD databases for OTU assignment in metabarcoding and metagenomic datasets: The BOLD_NCBI _Merger

Study area description: Metabarcoding, metagenomics and bioinformatics

Web location (URIs)

Download page: https://peerj.com/preprints/3133/

Technical specification

Platform: Unix

Programming language: Bash

Operational system: Linux, macOS

Usage licence

Acknowledgements

The script was developed in the context of the European Cooperation in Science and Technology (COST) Action DNAqua-Net (CA15219).

Author contributions

Conceived and designed the study: JNM; Wrote the script: JNM, THM; Analysed the data: JNM, THM, FL; Wrote the paper: JNM, THM, FL

References

- Allard G, Ryan F, Jeffery I, Claesson M (2015) SPINGO: a rapid species-classifier for microbial amplicon sequences. BMC Bioinformatics 16 (1). https://doi.org/10.1186/s12859-015-0747-1
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. Nucleic acids research 41 (Database issue): 36-42. https://doi.org/10.1093/nar/gks1195
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC bioinformatics 10: 421. https://doi.org/10.1186/1471-2105-10-421
- Capra E, Giannico R, Montagna M, Turri F, Cremonesi P, Strozzi F, Leone P, Gandini G, Pizzi F (2016) A new primer set for DNA metabarcoding of soil Metazoa. European Journal of Soil Biology 77: 53-59. https://doi.org/10.1016/j.ejsobi.2016.10.005
- Choo LQ, Crampton-Platt A, Vogler A (2017) Shotgun mitogenomics across body size classes in a local assemblage of tropical Diptera: Phylogeny, species diversity and mitochondrial abundance spectrum. Molecular Ecology 26 (19): 5086-5098. https://doi.org/10.1111/mec.14258
- Deiner K, Walser J, Mächler E, Altermatt F (2015) Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. Biological Conservation 183: 53-63. https://doi.org/10.1016/j.biocon.2014.11.018
- Elbrecht V, Leese F (2015) Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. PLOS ONE 10 (7): e0130324. https://doi.org/10.1371/journal.pone.0130324

- Elbrecht V, Vamos EE, Meissner K, Aroviita J, Leese F (2017) Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods in Ecology and Evolution 8 (10): 1265-1275. https://doi.org/10.1111/2041-210x.12789
- Garcia-Etxebarria K, Garcia-Garcerà M, Calafell F (2014) Consistency of metagenomic assignment programs in simulated and real data. BMC Bioinformatics 15 (1): 90. https://doi.org/10.1186/1471-2105-15-90
- Gayraud S, Statzner B, Bady P, Haybachp A, Scholl F, Usseglio-Polatera P, Bacchi M (2003) Invertebrate traits for the biomonitoring of large European rivers: an initial assessment of alternative metrics. Freshwater Biology 48 (11): 2045-2064. https://doi.org/10.1046/j.1365-2427.2003.01139.x
- Hänfling B, Handley LL, Read D, Hahn C, Li J, Nichols P, Blackman R, Oliver A, Winfield I (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. Molecular Ecology 25 (13): 3101-3119. https://doi.org/10.1111/mec.13660
- Hasan N, Young B, Minard-Smith A, Saeed K, Li H, Heizer E, McMillan N, Isom R, Abdullah AS, Bornman D, Faith S, Choi SY, Dickens M, Cebula T, Colwell R (2014) Microbial Community Profiling of Human Saliva Using Shotgun Metagenomic Sequencing. PLoS ONE 9 (5): e97699. https://doi.org/10.1371/journal.pone.0097699
- Hering D, Johnson R, Kramm S, Schmutz S, Szoszkiewicz K, Verdonschot PM (2006) Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. Freshwater Biology 51 (9): 1757-1785. https://doi.org/10.1111/j.1365-2427.2006.01610.x
- Horton D, Kershner M, Blackwood C (2017) Suitability of PCR primers for characterizing invertebrate communities from soil and leaf litter targeting metazoan 18S ribosomal or cytochrome oxidase I (COI) genes. European Journal of Soil Biology 80: 43-48. https://doi.org/10.1016/j.ejsobi.2017.04.003
- Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome research 17 (3): 377-86. https://doi.org/10.1101/gr.5969107
- Kõljalg U, Larsson K, Abarenkov K, Nilsson RH, Alexander I, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R, Taylor AS, Tedersoo L, Vrålstad T (2005) UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytologist 166 (3): 1063-1068. https://doi.org/10.1111/j.1469-8137.2005.01376.x
- Kranzfelder P, Ekrem T, Stur E (2015) Trace DNA from insect skins: a comparison of five extraction protocols and direct PCR on chironomid pupal exuviae. Molecular Ecology Resources 16 (1): 353-363. https://doi.org/10.1111/1755-0998.12446
- Lis JA, Lis B, Ziaja DJ (2016) In BOLD we trust? A commentary on the reliability of specimen identification for DNA barcoding: a case study on burrower bugs (Hemiptera: Heteroptera: Cydnidae). Zootaxa 4114 (1): 83-6. https://doi.org/10.11646/zootaxa.4114.1.6
- Macher J, Leese F (2017) Environmental DNA metabarcoding of rivers: Not all eDNA is everywhere, and not all the time. bioRxiv https://doi.org/10.1101/164046
- Macher JN, Zizka V, Weigand AM, Leese F (2017) A simple centrifugation protocol increases mitochondrial DNA yield 140-fold and facilitates mitogenomic studies. bioRxiv https://doi.org/10.1101/106583

- Porter T, Gibson J, Shokralla S, Baird D, Golding GB, Hajibabaei M (2014) Rapid and accurate taxonomic classification of insect (class Insecta) cytochromecoxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. Molecular Ecology Resourcesn/a-n/a. https://doi.org/10.1111/1755-0998.12240
- Prosser SJ, deWaard J, Miller S, Hebert PN (2017) DNA barcodes from century-old type specimens using next-generation sequencing. Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow https://doi.org/10.15468/ITPL93
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41: D590-D596. https://doi.org/10.1093/nar/gks1219
- Ratnasingham S, Hebert PN (2007) BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7 (3): 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4: e2584. https://doi.org/10.7717/peerj.2584
- Santamaria M, Fosso B, Consiglio A, Caro GD, Grillo G, Licciulli F, Liuni S, Marzano M, Alonso-Alemany D, Valiente G, Pesole G (2012) Reference databases for taxonomic assignment in metagenomics. Briefings in Bioinformatics 13 (6): 682-695. https://doi.org/10.1093/bib/bbs036
- Shi P, Jia S, Zhang X, Zhang T, Cheng S, Li A (2013) Metagenomic insights into chlorination effects on microbial antibiotic resistance in drinking water. Water Research 47 (1): 111-120. https://doi.org/10.1016/j.watres.2012.09.046
- Sonet G, Jordaens K, Braet Y, Bourguignon L, Dupont E, Backeljau T, Meyer Md, Desmyter S (2013) Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of forensically important Diptera from Belgium and France. ZooKeys 365: 307-328. https://doi.org/10.3897/zookeys.365.6027
- Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21 (8): 2045-2050. https://doi.org/10.1111/j.1365-294x.2012.05470.x
- Ueno K, Ishii A, Ito K (2014) ELM: enhanced lowest common ancestor based method for detecting a pathogenic virus from a large sequence dataset. BMC Bioinformatics 15 (1): 254. https://doi.org/10.1186/1471-2105-15-254
- Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology 73 (16): 5261-5267. https://doi.org/10.1128/aem.00062-07
- Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15 (3): R46. https://doi.org/10.1186/gb-2014-15-3-r46
- Yu D, Ji Y, Emerson B, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3 (4): 613-623. https://doi.org/10.1111/j.2041-210x.2012.00198.x

Supplementary material

Suppl. material 1: BOLD_NCBI_Merger script & tutorial

Authors: Jan-Niklas Macher, Till-Hendrik Macher, Florian Leese
Data type: BOLD_NCBI_Merger script & tutorial
Brief description: The supplementary material contains the BOLD_NCBI_Merger script, the needed folder structure and the tutorial explaining how to use the script
Filename: Supplementary material 1_Tutorial and script.zip - Download file (50.00 kb)