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

Jan- Niklas Macher[1,2*], Till- Hendrik Macher[1,2], Florian Leese[1,2]

[1]*Aquatic Ecosystem Research, Faculty of Biology, University of Duisburg-Essen, Universitätsstraße 5, 45141 Essen, Germany*

[2]*Centre for Water and Environmental Research, University of Duisburg-Essen, Universitätsstraße 2, 45141 Essen, Germany*

JNM: jan.macher@uni-due.de, *+49.201.183-6710*

Abstract

Metabarcoding and metagenomic approaches are becoming routine techniques in biodiversity assessment and ecological studies. The assignment of taxonomic information to sequences is challenging, as many reference libraries are lacking information on certain taxonomic groups and can contain erroneous sequences. Combining different reference databases is therefore a promising approach for maximizing taxonomic coverage and reliability of results. This tutorial shows how to use the "BOLD_NCBI_Merger" script to combine sequence data obtained from the National Center for Biotechnology Information (NCBI) GenBank and the Barcode of Life Database (BOLD) and prepare it for taxonomic assignment with the software MEGAN.

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 (Macher et al. 2016; Choo et al., 2017; Deiner et al., 2015), but also 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 largely relies on the knowledge of species' ecological traits (Gayraud et al., 2003; Hering et al., 2006). Several databases containing millions of reference sequences exist, which can be used to assign taxonomic names to OTUs (Santamaria et al., 2012). These databases are often specialised, containing mostly data for certain genetic markers (e.g. rRNA: SILVA (Quast et al., 2013) 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 & Hebert, 2007), which contains mostly metazoan sequences, and the National Center for Biotechnology Information (NCBI) GenBank database (Benson et al., 2012), which contains reference sequences for all domains of life. Sequence data is usually available for download via websites and/or command line applications and can be used for taxonomic assignment. This is a standard approach in metabarcoding and metagenomic studies, as manual blasting and identification of millions of sequences is not feasible. 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 & Leese, 2015; Prosser et al., 2016; Kranzfelder, Ekrem & Stur, 2016). Blast+ (Camacho et al., 2009) searches against the NCBI GenBank are often used for the identification of microbial sequences obtained through metagenomic approaches (Hasan et al., 2014; Shi et al., 2013), but also to confirm results of the BOLD API when metazoan taxa are studied (Kranzfelder, Ekrem & Stur, 2016; Elbrecht & Leese, 2015). Web tools and APIs remotely accessing databases tend 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, due to the focus on genetic barcodes 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 occur (e.g. Lis, Lis & Ziaja, 2016). The NCBI GenBank, however, does not include all sequences available in the BOLD database, as many scientists do not submit data to both databases, and data needs to be downloaded to a local hard drive in order to speed up blast searches.

Studies have shown that both databases can be used to successfully identify metazoan taxa (Sonet et al., 2013), but uncertainties remain. Metagenomic studies and metabarcoding studies have been shown to produce data not only from either microbial or metazoan taxa, but all trees of life (Capra et al., 2016; Macher & Leese, 2017; Horton, Kershner & Blackwood, 2017). For such studies, taxonomic assignment with the BOLD database only will result in the loss of information, as many microbial taxa cannot be identified. Using the NCBI GenBank only can circumvent this problem, but at the cost of losing information on metazoan taxa and lowered accuracy. Thus, combining information from both databases 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 several gigabytes of data that need to be downloaded onto 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 & Salzberg, 2014), SPINGO (Allard et al., 2015) and MEGAN (Huson et al., 2007).

Here we introduce our bash-script called "BOLD_NCBI_Merger" that builds databases containing sequence data from both BOLD and NCBI GenBank. In the tutorial coming along with the script we explain how to prepare data for analyses in the MEGAN software. The built database is stored separately and can also be used for other 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 be used for taxonomic assignment of sequences obtained through metabarcoding (Hänfling et al., 2016; Horton, Kershner & Blackwood, 2017).

Technical specification

Prior to analyses Blast+ (v. 2.6), vsearch (Rognes et al., 2016) and MEGAN need to be installed. All analyses for this tutorial were conducted on a Mac running Yosemite 10.10.5.

The bash script "BOLD_NCBI_Merger" concatenates multiple filed downloaded from BOLD and NCBI, respectively. Then, COI sequences are extracted from the downloaded BOLD fasta file. Headers of both BOLD and NCBI files are formatted so that vsearch can dereplicate the sequences without cutting the header, and files are concatenated. Then, vsearch is used to dereplicate the sequences in order to prevent overrepresentation 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 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.

For the 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, Garcia-Garcerà & Calafell, 2014; Ueno, Ishii & Ito, 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, respectively, 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.

References

Allard G., Ryan FJ., Jeffery IB., Claesson MJ. 2015. SPINGO: a rapid species-classifier for microbial amplicon sequences. BMC bioinformatics 16:324.

Benson DA., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman DJ., Ostell J., Sayers EW. 2012. GenBank. Nucleic acids research 41:D36–D42.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009. BLAST+: architecture and applications. BMC bioinformatics 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.

Choo, L. Q., Crampton-Platt, A., Vogler, A. P. 2017. Shotgun mitogenomics across body size classes in a local assemblage of tropical Diptera: Phylogeny, species diversity and mitochondrial abundance spectrum. Molecular Ecology. doi: 10.1111/mec.14258

Deiner K., Walser J-C., Mächler E., Altermatt F. 2015. Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. Biological conservation 183:53–63.

Díaz S., Fargione J., Chapin FS 3rd., Tilman D. 2006. Biodiversity loss threatens human well-being. PLoS biology 4:e277.

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:e0130324.

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 / British Ecological Society. DOI: 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 15:90.

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 48:2045–2064.

Hänfling B., Lawson Handley L., Read DS., Hahn C., Li J., Nichols P., Blackman RC., Oliver A., Winfield IJ. 2016. Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. Molecular ecology 25:3101–3119.

Hasan NA., Young BA., Minard-Smith AT., Saeed K., Li H., Heizer EM., McMillan NJ., Isom R., Abdullah AS., Bornman DM., Faith SA., Choi SY., Dickens ML., Cebula TA., Colwell RR. 2014. Microbial community profiling of human saliva using shotgun metagenomic sequencing. PloS one 9:e97699.

Hebert PDN., Cywinska A., Ball SL., deWaard JR. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society B: Biological Sciences 270:313–321.

Hering D., Johnson RK., Kramm S., Schmutz S., Szoszkiewicz K., Verdonschot PFM. 2006. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. Freshwater biology 51:1757–1785.

Horton DJ., Kershner MW., Blackwood CB. 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.

Huson DH., Auch AF., Qi J., Schuster SC. 2007. MEGAN analysis of metagenomic data. Genome research 17:377–386.

Kõljalg U., Larsson K-H., Abarenkov K., Henrik Nilsson R., Alexander IJ., Eberhardt U., Erland S., Høiland K., Kjøller R., Larsson E., Pennanen T., Sen R., Taylor AFS., Tedersoo L., Vrålstad T. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. The New phytologist 166:1063–1068.

Kranzfelder P., Ekrem T., Stur E. 2016. Trace DNA from insect skins: a comparison of five extraction protocols and direct PCR on chironomid pupal exuviae. Molecular ecology resources 16:353–363.

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 4114:83.

Macher J-N., Leese F. 2017. Environmental DNA metabarcoding of rivers: Not all eDNA is everywhere, and not all the time. bioRxiv, doi: https://doi.org/10.1101/164046

Macher, J.-N., Zizka, V., Weigand, A. M.,Leese, F. 2017. A simple centrifugation protocol increases mitochondrial DNA yield 140-fold and facilitates mitogenomic studies. bioRxiv, https://doi.org/10.1101/106583

Marchesi JR., Weightman AJ., Cragg BA., Parkes RJ., Fry JC. 2001. Methanogen and bacterial diversity and distribution in deep gas hydrate sediments from the Cascadia Margin as revealed by 16S rRNA molecular analysis. FEMS microbiology ecology 34:221–228.

Porter TM., Gibson JF., Shokralla S., Baird DJ., Brian Golding G., 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 resources. DOI: 10.1111/1755-0998.12240.

Prosser SWJ., deWaard JR., Miller SE., Hebert PDN. 2016. DNA barcodes from century-old type specimens using next-generation sequencing. Molecular ecology resources 16:487–497.

Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J., Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research 41:D590–6.

Ratnasingham S., Hebert PDN. 2007. BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular ecology notes 7:355–364.

Rockström J., Steffen W., Noone K., Persson A., Chapin FS 3rd., Lambin EF., Lenton TM., Scheffer M., Folke C., Schellnhuber HJ., Nykvist B., de Wit CA., Hughes T., van der Leeuw S., Rodhe H., Sörlin S., Snyder PK., Costanza R., Svedin U., Falkenmark M., Karlberg L., Corell RW., Fabry VJ., Hansen J., Walker B., Liverman D., Richardson K., Crutzen P., Foley JA. 2009. A safe operating space for humanity. Nature 461:472–475.

Rognes T., Flouri T., Nichols B., Quince C., Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.

Rondon MR., August PR., Bettermann AD., Brady SF., Grossman TH., Liles MR., Loiacono KA., Lynch BA., MacNeil IA., Minor C., Tiong CL., Gilman M., Osburne MS., Clardy J., Handelsman J., Goodman RM. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Applied and environmental microbiology 66:2541–2547.

Santamaria M., Fosso B., Consiglio A., De Caro G., 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 13:682–695.

Shi P., Jia S., Zhang X-X., Zhang T., Cheng S., Li A. 2013. Metagenomic insights into chlorination effects on microbial antibiotic resistance in drinking water. Water research 47:111–120.

Sonet G., Jordaens K., Braet Y., Bourguignon L., Dupont E., Backeljau T., De Meyer M., 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:307–328.

Taberlet P., Coissac E., Pompanon F., Brochmann C., Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology 21:2045–2050.

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 15:254.

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 73:5261–5267.

Wood DE., Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology 15:R46.

Yu DW., Ji Y., Emerson BC., 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 / British Ecological Society 3:613–623.