# NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data

Kristoffer Sahlin[1], Marisa Lim[2], and Stefan Prost[3]

[1]Stockholm University
[2]University of California Davis
[3]Senckenberg Research Institutes and Natural History Museums

December 4, 2020

**Abstract**

Third generation sequencing technologies, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), have gained popularity over the last years. These platforms can generate millions of long read sequences. This is not only advantageous for genome sequencing projects, but also for amplicon-based high-throughput sequencing experiments, such as DNA barcoding. However, the relatively high error rates associated with these technologies still pose challenges for generating high quality consensus sequences. Here we present NGSpeciesID, a program which can generate highly accurate consensus sequences from long-read amplicon sequencing technologies, including ONT and PacBio. The tool includes clustering of the reads to help filter out contaminants or reads with high error rates and employs polishing strategies specific to the appropriate sequencing platform. We show that NGSpeciesID produces consensus sequences with improved usability by minimizing preprocessing and software installation and scalability by enabling rapid processing of hundreds to thousands of samples, while maintaining similar consensus accuracy as current pipelines

## NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data

Kristoffer Sahlin[1], Marisa C.W. Lim[2], Stefan Prost[3,4,#]

[1]Department of Mathematics, Science for Life Laboratory, Stockholm University, 106 91 Stockholm, Sweden

[2] Department of Population Health and Reproduction, University of California, 1 Garrod Dr, Davis, CA 95616, USA

[3]LOEWE-Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt, Germany

[4]South African National Biodiversity Institute, National Zoological Garden, Pretoria 0001, South Africa

[#]Correspondence:

*stefanprost.research@protonmail.com*

**Summary**

Third generation sequencing technologies, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), have gained popularity over the last years. These platforms can generate millions of long read sequences. This is not only advantageous for genome sequencing projects, but also for amplicon-based high-throughput sequencing experiments, such as DNA barcoding. However, the relatively high error rates associated with these technologies still pose challenges for generating high quality consensus sequences. Here we present NGSpeciesID, a program which can generate highly accurate consensus sequences from long-read

amplicon sequencing technologies, including ONT and PacBio. The tool includes clustering of the reads to help filter out contaminants or reads with high error rates and employs polishing strategies specific to the appropriate sequencing platform. We show that NGSpeciesID produces consensus sequences with improved usability by minimizing preprocessing and software installation and scalability by enabling rapid processing of hundreds to thousands of samples, while maintaining similar consensus accuracy as current pipelines

**Keywords:** DNA barcoding, amplicon sequencing, third generation sequencing, sequence clustering

## Introduction

We are in the middle of a biodiversity crisis, in which anthropogenic change is driving many species to extinction, often faster than they can be characterized (see e.g. Ceballos et al., (2020)). The identification of species in our environments is paramount to informing conservation policy and practice. The development of DNA barcoding (Hebert et al., 2003) was a major step towards large-scale characterizations of biodiversity. This technique utilizes amplification of standardized genetic regions to characterize species present within biological samples. Besides the documentation of biodiversity, this method and other amplicon-sequencing technologies have been widely used for monitoring of invasive species, detection of pathogens in environmental samples, and many other applications in taxonomy, medicine or evolutionary biology (e.g. reviewed in Kress et al., (2015)).

Third-generation sequencing is able to sequence millions of single molecules up to several Mbs in lengths (Jain et al., 2018). Currently, two platforms are readily available for DNA barcoding efforts, PacBio's Sequel II and ONT's MinION. These platforms offer the advantage of longer reads, at the cost of sequencing errors. While ONT's MinION still shows higher error rates >5% (Wick et al., 2018), the new PacBio HiFi mode allows for the generation of read with <1% error (Wenger et al., 2019), which will greatly improve the generation of accurate DNA barcodes. Early on, researchers identified the potential of third-generation sequencing platforms for sequencing much longer DNA barcodes than previously possible (see e.g. Krehenwinkel et al., (2019a); Tedersoo et al., (2018); Wurzbacher et al., (2019)). Beside the longer amplicon length, ONT's MinION also offers the advantage that sequencing can be carried out almost anywhere in the world, due to its small size and affordability (reviewed in Krehenwinkel et al., 2019b). While there has been a considerable software development effort to assemble high-quality amplicon consensus sequences from error-prone ONT MinION reads (see e.g. Maestri et al., 2019; Seah et al., 2020; Srivathsan et al., 2019; reviewed in Krehenwinkel et al., 2019b), only a few software solutions are available for PacBio-based DNA barcodes (see e.g. Wurzbacher et al., 2019). To our knowledge, of these, only the pipeline presented in Wurzbacher et al., 2019 is able to handle both PacBio and ONT sequencing reads.

Here, we present NGSpeciesID a one-software solution for reconstructing high-quality amplicon consensus sequences for both PacBio and ONT sequencing reads. We also investigate the performance of ONT's Medaka polishing software compared to Racon (Vaser et al., 2017) for MinION based DNA barcoding. Compared to other programs, NGSpeciesID can be easily installed with conda, does not require any specific file name structures, can handle data from both third-generation sequencing types, includes different consensus polishing options and only needs fastq files as input. We show that our tool produces consensus sequences of a similar quality than other software solutions, while reducing the burden to users by requiring little to no additional tools or data reformatting.

## Software description

NGSpeciesID is a program developed in python that wraps a set of tools for read clustering, consensus forming, and consensus polishing (Figure 1). It is a one-software solution and extension of the Saiga pipeline, we developed previously (Seah et al., 2020). It can be easily installed using the free and open-source Anaconda distribution. Briefly, NGSpeciesID clusters amplicon sequencing reads (in fastq format) and forms a consensus sequence for each cluster. Next, it merges reverse complement clusters. Finally, the remaining consensus sequence(s) is/are polished. Optionally, the tool can also remove primer sequences from the consensus after the polishing step. In the following sections we describe the workflow of NGSpeciesID, which is freely available at*https://github.com/ksahlin/NGSpeciesID* . For more details see Supplementary File 1

and Figure 1.

Clustering of reads

NGSpeciesID first clusters the reads based on expected sequence similarity for ONT or PacBio reads. The clustering algorithm used in NGSpeciesID is the isONclust algorithm which is described in detail in Sahlin and Medvedev (2019).

Forming draft consensus

Next, a draft consensus is formed for each cluster that contains more reads than an abundance threshold (default: 10% of the total number of reads). The draft consensus sequences are formed with spoa (Vaser et al., 2017).

Reverse complement detection and removal

NGSpeciesID then detects and merges any consensus sequences classified as reverse complement sequences using pairwise alignment with parasail (Daily, 2016). All consensus sequences are aligned to the reverse complements of the other sequences, and the respective raw read clusters are combined. Finally, all draft consensus sequences passing this step, together with the original reads, are sent to polishing.

Polishing

The remaining consensus sequences are polished with either Medaka (https://github.com/nanoporetech/medaka) or Racon (Vaser et al., 2017). The polished consensus sequences are the final output of NGSpeciesID.

Primer detection and removal

Many basecalling and de-multiplexing tools do not remove primers from the amplicon sequences (but see Minibar (Krehenwinkel et al., 2019a)). NGSpeciesID, therefore, implements an optional primer removal step by searching the forward and reverse complement of each primer within a window at each end of the read. This is carried out for the polished sequences. If no primer is found, the polished consensus sequence(s) remain the final output of NGSpeciesID. If primer(s) have been detected and trimmed, NGSpeciesID reruns the reverse-complement removal and polishing steps to identify any remaining redundant consensus sequences that were not removed due to primers.

**Use cases and comparison to other tools**

We tested our software on publicly available data from Maestri et al. (2019) and Wurzbacher et al. (2019), and compared the accuracy of respective consensus sequences generated in the two studies to those reconstructed with NGSpeciesID. To measure accuracy, we aligned the consensus sequences to the Sanger sequence using BLAST (Altschul et al., 1990) and calculated accuracy as the sum of all matches in the alignment divided by the alignment length. We chose the software solution presented in Wurzbacher et al., (2019) for our comparison as it is currently, to our knowledge, the only one that can handle both PacBio and ONT sequencing reads. We further compared our result to the ONTrack software (Maestri et al., 2019) developed for ONT data specifically. In both comparisons we carried out polishing within NGSpeciesID using Medaka (https://github.com/nanoporetech/Medaka) and Racon (Vaser et al., 2017).

Comparison to Mothur + Consension

We randomly selected five out of the 61 fungi datasets from Wurzbacher et al., (2019), ranging from 201 to 447 reads per dataset (Supplementary Table 1). These cover five fungi species of the genus *Inocybe* for ribosomal DNA (rDNA) and the full ribosomal tandem repeat region (TR). We provide alignments of the corresponding Sanger sequences with our consensus sequences in the Supplementary (Supplementary files 2-6). In their approach, Wurzbacher et al., (2019) first perform operational taxonomic unit (OTU) clustering on the read data using Mothur (Schloss et al., 2009). Next, they create consensus sequences using Consension (Wurzbacher et al., 2019).

In general, we see that for both ONT and PacBio data NGSpeciesID and the Mothur + Consension pipeline perform equally well, generating consensus sequences with 98.6% to 100% accuracy (Table 1). In three out of the five cases, the two pipelines produced consensus sequences with the same accuracy, while in one case each software slightly outperformed the other (Table 1). Medaka polishing outperformed Racon polishing in four out of five cases (Table 1).

Comparison to ONTrack

Next, we compared the performance of NGSpeciesID to the pipeline ONTrack from Maestri et al. (2019). This pipeline first clusters all reads using VSEARCH (Rognes et al., 2016), then randomly selects 200 reads, aligns those with Mafft (Katoh and Standley, 2013), calls the consensus with EMBOSS cons (http://emboss.sourceforge.net/apps/cvs/emboss/apps/cons.html), and lastly carries out polishing with 200 randomly selected reads using Nanopolish (https://github.com/jts/nanopolish.). We generated consensus sequences for all seven DNA barcodes from Maestri et al. (2019), which comprise *Cytochrome C Oxidase Subunit 1* (COI) sequences of two snails and five beetles (Supplementary Table 1). We provide the respective alignments in the Supplementary (Supplementary files 7-13).

Previously, Krehenwinkel et al., (2019a) showed that consensus accuracy can decrease when too many reads (in the realm of a few hundred reads, depending on the error rate of the individual reads) are selected for the consensus generation, likely due to an increase in the signal to noise ratio. We thus randomly subsampled 300 reads using seqtk (https://github.com/lh3/seqtk), a number which has been shown to work well with Nanopore data (Krehenwinkel et al., 2019a). We see that the consensus quality is comparable between the two tools (Table 2), with accuracy of 99.8% to 100%. In five out of the seven DNA barcode sets both tools performed equally well, while in one each the two tools outperformed each other, however, differing by only 1 basepair (Table 2).

Mixed samples

We tested NGSpeciesID's performance on mixed samples *in silico* by combining 300 reads of each of the seven barcodes from Maestri et al. (2019). To do so, we set the cluster abundance ratio to 5% (–abundance_ratio 0.05). We recovered seven consensus sequences corresponding to the seven DNA barcodes, ranging from 99.3% to 100% similarity to the corresponding Sanger sequence (Table 2). In four out of the seven cases, we recovered the same percentage similarity to the Sanger sequence in the mixed analysis as in the respective single barcode processing. In three cases the accuracy was slightly lower with two and four basepair differences, respectively.

## Discussion

Consensus Quality

Here we present NGSpeciesID, an easy-to-use, one-software solution for the generation of high-quality consensus sequences for the long-read sequencing technologies from ONT and PacBio. We compared NGSpeciesID against results obtained with Mothur + Consension and ONTrack. In general, all three software solutions produced consensus sequences of a very high quality, reaching 99-100% accuracy in almost all cases. We show that NGSpeciesID performs comparably to the other tools. Throughout all comparisons, we see that consensus sequences based on ONT data polished with Racon usually show lower percent similarities to the Sanger sequence than consensus sequences polished with Medaka. NGSpeciesID carries out 2 rounds of Racon polishing by default. Increasing or decreasing the number of rounds might increase the consensus quality. We chose Medaka as the default error corrector in NGSpeciesID as it includes up to date error models. We did not include an option to use Nanopolish in NGSpeciesID, which is used in ONTrack, as this tool requires fast5 files, which are often not available for published Oxford Nanopore data. Furthermore, it requires preprocessing to generate the appropriate header structure in the corresponding fastq files, which makes it much more time consuming to use.

As the generation of consensus sequences for DNA barcoding takes only a few seconds for each sample (depending on the number of reads), we did not compare run times between the different pipelines.

Easy-use

NGSpeciesID was designed to be straightforward to use. It works on individual read files, outputted either directly from the basecalling or after demultiplexing (e.g. using Minibar (Krehenwinkel et al. 2019a) or qcat (https://github.com/nanoporetech/qcat)), but can quickly be adjusted to run in a loop over multiple fastq files using a bashscript (see Supplementary File 14). It only requires fastq files as input. In contrast, ONTrack requires the input reads in three formats (fast5, fasta and fastq), which requires additional preprocessing of the sequencing data. Furthermore, NGSpeciesID allows fastq files to have any naming structure, thus making it easy for the user to run and to identify samples and replicates. This saves time on preprocessing of the read data compared to other software solutions.

NGSpeciesID employs quality filtering of the reads based on read phred scores. However, we recommend also removing reads much shorter or longer than the intended target, which often represent chimeras or contaminations using NanoFilt (De Coster et al., 2018) before running NGSpeciesID. While our tool can handle unfiltered data, this might result in the generation of multiple consensus sequences. NGSpeciesID also offers the option to remove priming sites from the amplicon sequences. As many universal primers include ambiguity codes, primer regions can potentially include incorrect bases, and should thus be removed. We further found that primer regions can cause issues for the reverse-complement matching. We thus included an additional reverse-complement matching step after primer removal, in case NGSpeciesID outputs multiple consensus sequences. Our tool outputs multiple consensus sequences in case the clustering results in multiple clusters over a certain percentage of the total reads (by default this is set to 10%). Each consensus sequence is only polished with the corresponding reads from the clustering. This feature is very useful as it allows the user to explore potential contaminant reads or mixed samples through the generating of multiple consensus sequences.

NGSpeciesID and the Mothur + Consension software solution both can handle ONT and PacBio long-read data. While both tools produce consensus sequences of similar accuracy, Mothur + Consension requires an in-depth knowledge of the pipeline requiring (i) preprocessing of the input data, (ii) individual components of the pipeline to be run separately and (iii) has parameter settings that are difficult to interpret, while NGSpecies is designed to be user friendly and packaged as a one command solution.

Mixed samples

While NGSpeciesID was not designed specifically for metabarcoding data, the flexibility of the algorithmic steps in the pipeline enables the tool to handle mixed samples. We recovered seven consensus sequences corresponding to the seven DNA barcodes pooled in the mixed sample analysis. NGSpeciesID generated highly accurate consensus sequences for all barcodes, ranging from 99.2% to 100%. For the mixed sample test we adjusted the read abundance ratio for the clusters to 5%, since the seven barcodes at equal abundance are each present in only 14% of the reads in the sample. Therefore, the default abundance cutoff of 10% would require 210 out of the 300 reads to be used per cluster, which might not be the case. Three out of seven barcodes showed a slightly lower consensus accuracy than in the respective single species analysis, which is likely due to the presence of some reads from other barcodes in the clusters that might have affected the polishing accuracy, and the random selection of the 300 reads for each barcode (as individual read error rates can differ). We expect some cross-contamination (reads assigned to the wrong cluster), especially for closely related species. However, this should improve with the continued improvement of third-generation sequencing read accuracy. This experiment shows that NGSpeciesID, even though it was not developed for mixed samples, can recover highly accurate consensus sequences from metabarcoding data. However, its performance on metabarcoding data will need to be investigated separately with mock datasets of varying ratios and sample relationships (taxonomic divergences).

**Conclusion and Future directions**

We present NGSpeciesID, an easy-to-use and flexible one-software solution to generate high-quality consensus sequences for both ONT and PacBio sequencing data. It performs equally well as other pipelines and software solutions tested here, but offers advanced usability as it is simple to use and does not require pre-processing

of the data before running. Portable devices such as the inexpensive MinION sequencer have started to democratize the process of molecular biodiversity monitoring (see eg. Krehenwinkel et al., (2019b). Here we add to this, by the development of a simple to install and run bioinformatic software, that should further enable students and citizen-scientists without a formalized bioinformatic training to carry out biodiversity monitoring and assessment studies.

**Conflict of Interest**

The authors have declared that no competing interests exist.

**Authors contributions**

KS, MCWL and SP developed the software tool. SP carried out the comparisons. All authors wrote and commented on the manuscript.

**Acknowledgement**

We thank Tilman Schell and Aaron Pomerantz for their valuable comments on the paper draft, and Christian Wurzbacher for the consensus and Sanger sequences from their study. The authors declare no conflict of interest.

**References**

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ceballos, G., Ehrlich, P.R., Raven, P.H., 2020. Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. PNAS 117, 13596–13602. https://doi.org/10.1073/pnas.1922686117

Daily, J., 2016. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. BMC Bioinformatics 17, 81. https://doi.org/10.1186/s12859-016-0930-z

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., Van Broeckhoven, C., 2018. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

Hebert, P.D.N., Ratnasingham, S., de Waard, J.R., 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London. Series B: Biological Sciences 270, S96–S99. https://doi.org/10.1098/rsbl.2003.0025

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H.E., Pedersen, B.S., Rhie, A., Richardson, H., Quinlan, A.R., Snutch, T.P., Tee, L., Paten, B., Phillippy, A.M., Simpson, J.T., Loman, N.J., Loose, M., 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36, 338–345. https://doi.org/10.1038/nbt.4060

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30, 772–780. https://doi.org/10.1093/molbev/mst010

Krehenwinkel, H., Pomerantz, A., Henderson, J.B., Kennedy, S.R., Lim, J.Y., Swamy, V., Shoobridge, J.D., Graham, N., Patel, N.H., Gillespie, R.G., Prost, S., 2019a. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. Gigascience 8. https://doi.org/10.1093/gigascience/giz006

Krehenwinkel, H., Pomerantz, A., Prost, S., 2019b. Genetic Biomonitoring and Biodiversity Assessment Using Portable Sequencing Technologies: Current Uses and Future Directions. Genes 10, 858. https://doi.org/10.3390/genes10110858

Kress, W.J., García-Robledo, C., Uriarte, M., Erickson, D.L., 2015. DNA barcodes for ecology, evolution, and conservation. Trends in Ecology & Evolution 30, 25–35. https://doi.org/10.1016/j.tree.2014.10.008

Maestri, S., Cosentino, E., Paterno, M., Freitag, H., Garces, J.M., Marcolungo, L., Alfano, M., Njunjić, I., Schilthuizen, M., Slik, F., Menegon, M., Rossato, M., Delledonne, M., 2019. A Rapid and Accurate MinION-Based Workflow for Tracking Species Biodiversity in the Field. Genes 10, 468. https://doi.org/10.3390/genes10060468

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

Sahlin, K., Medvedev, P., 2019. De Novo Clustering of Long-Read Transcriptome Data Using a Greedy, Quality-Value Based Algorithm, in: Cowen, L.J. (Ed.), Research in Computational Molecular Biology, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 227–242. https://doi.org/10.1007/978-3-030-17083-7_14

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Horn, D.J.V., Weber, C.F., 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. 75, 7537–7541. https://doi.org/10.1128/AEM.01541-09

Seah, A., Lim, M.C.W., McAloose, D., Prost, S., Seimon, T.A., 2020. MinION-Based DNA Barcoding of Preserved and Non-Invasively Collected Wildlife Samples. Genes 11, 445. https://doi.org/10.3390/genes11040445

Srivathsan, A., Hartop, E., Puniamoorthy, J., Lee, W.T., Kutty, S.N., Kurina, O., Meier, R., 2019. Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. BMC Biology 17, 96. https://doi.org/10.1186/s12915-019-0706-9

Tedersoo, L., Tooming-Klunderud, A., Anslan, S., 2018. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytologist 217, 1370–1385. https://doi.org/10.1111/nph.14776

Vaser, R., Sović, I., Nagarajan, N., Šikić, M., 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746. https://doi.org/10.1101/gr.214270.116

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 37, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wick, R.R., Judd, L.M., Holt, K.E., 2018. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. PLoS Comput Biol 14. https://doi.org/10.1371/journal.pcbi.1006583

Wurzbacher, C., Larsson, E., Bengtsson-Palme, J., Wyngaert, S.V. den, Svantesson, S., Kristiansson, E., Kagami, M., Nilsson, R.H., 2019. Introducing ribosomal tandem repeat barcoding for fungi. Molecular Ecology Resources 19, 118–127. https://doi.org/10.1111/1755-0998.12944

**Figure Legend**

**Figure 1. Steps involved in DNA barcode consensus calling of long-read data.** The respective software tools used in the different steps are provided in brackets. In the first step long-read data is usually demultiplexed. After demultiplexing, the reads are filtered for read length and quality. This step can also be carried out before demultiplexing if the respective amplicons do not differ in length. Next, consensus sequences for the individual read files can be generated using NGSpeciesID. If multiple read files need to be processed, NGSpeciesID can be run in a pipeline (see Supplementary File 14). Within the tool, reads are first clustered according to similarity. Next, consensus sequences are generated for each cluster larger than an abundance threshold (default: $> 10\%$ of all reads). In the third step NGSpeciesID checks the generated consensus sequences for reverse complementary. If consensus sequences are reverse complement, then the

respective clusters are merged. In step four, the consensus sequences are polished using the reads from the respective clusters (this step is optional). In the last optional step, primers can be removed if this was not already carried out by the demultiplexing or basecalling tools. If primers are removed, NGSpeciesID will carry out step 3 - 4 again.

**Tables**

| | | *17075* | *17078* | *16416* | *16427* | *16483* |
|---|---|---|---|---|---|---|
| **NGSpeciesID** | **ONT** | **98.6%** | **99.2%** | 99.5% (4/731) | **99.6%** | **100%** |
| | **Medaka** | **(10/726)** | **(6/741)** | | **(3/790)** | **(0/709)** |
| | **ONT Racon** | 98.5% (11/726) | 99.1% (7/741) | 99.7% (2/731) | 99.1% (7/790) | 99.9% (1/709) |
| | **PB Racon** | **98.6%** | 99.1% (7/741) | **99.9%** | **99.6%** | **100%** |
| | | **(10/726)** | | **(1/731)** | **(3/790)** | **(0/709)** |
| **Mothur + Consension** | **ONT** | **98.6%** | **99.2%** | **99.9%** | 99.5% (4/790) | **100%** |
| | | **(10/726)** | **(6/741)** | **(1/731)** | | **(0/709)** |
| | **PB** | **98.6%** | **99.2%** | 99.7% (2/731) | **99.6%** | **100%** |
| | | **(10/726)** | **(6/741)** | | **(3/790)** | **(0/709)** |

**Table 1. Percent similarity to the respective Sanger sequence for the datasets 17075, 17078, 16416, 16427 and 16483 from Wurzbacher et al. (2019).** The highest similarity scores are highlighted in bold.

| | | *BC1** | *BC2* | *BC3* | *BC4** | *BC5* | *BC6** | *BC7** |
|---|---|---|---|---|---|---|---|---|
| **NGSpeciesID** | **ONT** | **100%** | **100%** | 99.9% | **100%** | **100%** | **99.8%** | **100%** |
| | **Medaka** | **(0/651)** | **(0/658)** | (1/649) | **(0/606)** | **(0/658)** | **(1/576)** | **(0/536** |
| | **ONT** | 99.5% | 99.5% | 98.9% | 99.2% | 99.8% | 99.7% | 99.4% |
| | **Racon** | (3/651) | (3/658) | (7/649) | (5/606) | (1/658) | (2/576) | (3/536) |
| **ONTrack** | **ONT** | 99.9% | **100%** | **100%** | **100%** | **100%** | **99.8%** | **100%** |
| | | (1/651) | **(1**/658)** | **(2**/649)** | **(0/606)** | **(2**/658)** | **(1/576)** | **(0/536** |
| **Mixed** | | | | | | | | |
| **NGSpeciesID** | **ONT** | **100%** | **100%** | 99.7% | 99.3% | **100%** | **99.8%** | 99.6% |
| | **Medaka** | **(0/651)** | **(0/658)** | (2/649) | (4/606) | **(0/658)** | **(1/576)** | (2/536) |

**Table 2. Percent similarity to the respective Sanger sequence for the datasets B1 to BC7 from Maestri et al. (2019).** For the mixed samples, 300 reads of each of the seven DNA barcodes were combined into a single file, from which NGSpecesID generated multiple consensus sequences. NGSpeciesID was run using Medaka polishing. * Here the Sanger sequence from Maestri et al. 2019 was shorter than the expected fragment length and all the consensus sequences. In these cases we only calculated the percentage similarity for the region covered by the respective Sanger sequence. ** The consensus sequences from Maestri et al. 2019 are missing one or two bases at the start, which could be due to a consensus calling error, or deletion of one additional base during the primer removal. For the percentage accuracy we assumed them to be incorrectly trimmed. The highest similarity scores are highlighted in bold.