Developmental validation of Oxford Nanopore Technology MinION sequence data and the NGSpeciesID bioinformatic pipeline for forensic genetic species identification

Forensic Science International: Genetics 53 (2021) 102493

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Developmental validation of Oxford Nanopore Technology MinION sequence data and the NGSpeciesID bioinformatic pipeline for forensic genetic species identification

Nina Vasiljevic [a,*], Marisa Lim [b], Emily Humble [c], Adeline Seah [b], Adelgunde Kratzer [a], Nadja V. Morf [a], Stefan Prost [e,f,1], Rob Ogden [c,d,1]

[a] Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland
[b] Wildlife Conservation Society, Zoological Health Program, Bronx, NY, USA
[c] Royal (Dick) School of Veterinary Studies and the Roslin Institute, University of Edinburgh, UK
[d] TRACE Wildlife Forensics Network, Edinburgh, UK
[e] LOEWE-Centre for Translational Biodiversity Genomics, Senckenberg, Frankfurt, Germany
[f] South African National Biodiversity Institute, National Zoological Garden, Pretoria, South Africa

ARTICLE INFO

Keywords:
High-throughput sequencing (HTS)
MtDNA
Validation
DNA barcoding
Species identification
MinION
Bioinformatic pipeline
NGSpeciesID

ABSTRACT

Species identification of non-human biological evidence through DNA nucleotide sequencing is routinely used for forensic genetic analysis to support law enforcement. The gold standard for forensic genetics is conventional Sanger sequencing; however, this is gradually being replaced by high-throughput sequencing (HTS) approaches which can generate millions of individual reads in a single experiment. HTS sequencing, which now dominates molecular biology research, has already been demonstrated for use in a number of forensic genetic analysis applications, including species identification. However, the generation of HTS data to date requires expensive equipment and is cost-effective only when large numbers of samples are analysed simultaneously. The Oxford Nanopore Technologies (ONT) MinION™ is an affordable and small footprint DNA sequencing device with the potential to quickly deliver reliable and cost effective data. However, there has been no formal validation of forensic species identification using high-throughput (deep read) sequence data from the MinION making it currently impractical for many wildlife forensic end-users. Here, we present a MinION deep read sequence data validation study for species identification. First, we tested whether the clustering-based bioinformatics pipeline NGSpeciesID can be used to generate an accurate consensus sequence for species identification. Second, we systematically evaluated the read variation distribution around the generated consensus sequences to understand what confidence we have in the accuracy of the resulting consensus sequence and to determine how to interpret individual sample results. Finally, we investigated the impact of differences between the MinION consensus and Sanger control sequences on correct species identification to understand the ability and accuracy of the MinION consensus sequence to differentiate the true species from the next most similar species. This validation study establishes that ONT MinION sequence data used in conjunction with the NGSpeciesID pipeline can produce consensus DNA sequences of sufficient accuracy for forensic genetic species identification.

1. Introduction

1.1. Non-human DNA forensics

Non-human biological evidence can inform criminal investigations in three ways. Most commonly, animals and plants may be the victims of crime, in cases ranging from animal persecution to illegal harvest and subsequent trafficking of protected species. Second, trace biological evidence may contribute indirectly to reconstructing events at a crime scene, for example through the analysis of shed hairs, or profiling of plant or microbial communities from evidence recovered in relation to almost any type of crime. Third, animals may be the primary

* Corresponding author.
E-mail address: nina.vasiljevic@irm.uzh.ch (N. Vasiljevic).
[1] contributed equally

https://doi.org/10.1016/j.fsigen.2021.102493
Received 19 November 2020; Received in revised form 9 February 2021; Accepted 3 March 2021
Available online 12 March 2021

perpetrators of unlawful acts, for example dog attacks on humans. In each case, it is usually necessary to establish the biological species of the evidence, either as the point to prove in an investigation, or as a precursor to subsequent analytical testing. Species identification may be achieved using a range of scientific approaches, including morphology and mass spectrometry; however, since its initial application to law enforcement in the early 1990s, DNA sequencing has gradually developed to become the preferred method of forensic analysis [1]. This is particularly true where evidence has lost its morphological features or where enforcement agencies lack access to traditional taxonomic expertise.

1.2. DNA species identification – current methods and limitations

The gold standard for forensic genetic species identification is a four-step process consisting of DNA extraction, PCR amplification, conventional Sanger DNA nucleotide sequencing, and sequence similarity analysis against a reference database. This is a well-established technique that uses short, conserved DNA sequence markers (so called "DNA barcodes"), [2]) that are species-diagnostic, meaning that within-species sequence variation should not create any overlap among closely related species. This "break" in the distribution of pairwise sequence divergence, from intra-specific to inter-specific variation may be referred to as the "barcoding gap" [3]. Regions within several different mitochondrial DNA (mtDNA) genes such as cytochrome b [4], cytochrome oxidase I [5,6], and 12S rRNA [7] exhibit such gaps in different taxonomic groups and have been tested and validated for the identification of species in forensic casework. DNA sequences from unknown samples are identified against reference sequences, through a process of sequence similarity matching, or in some cases, phylogenetic analysis. Forensic genetic species identification is more robust if longer sequences are generated and compared at multiple DNA markers, however in practice, species identification is often performed using only 300–700 base pair (bp) sequences from a single gene region. A barcoding gap between the true species of origin and its closest relative of 2% over a 300 bp sequence equates to a 6 bp divergence between species, requiring an accurate, reliable sequencing method.

Sanger sequencing, which produces a single sequence read output (usually duplicated through forward and reverse sequencing reactions), has proven highly successful, with a typical sequence error rate of just ~0.001% [8]. Despite this performance record, there are limitations to Sanger sequencing and more recent high-throughput sequencing (HTS) technologies are starting to replace Sanger sequencing platforms in many molecular genetic laboratories. Sanger sequencing only generates a single sequence read, or electropherogram, for each sample PCR amplification product. In cases of co-amplification, where a contaminated or mixed species template generates PCR products from two or more donors, the electropherogram is typically unreadable and the individual component sequences of the co-amplified donors cannot be distinguished [9,10]. Although Sanger sequencing will remain a useful forensic genetic approach capable of validating HTS output as its popularity as a primary sequencing method declines its availability is likely to reduce and it is therefore necessary to consider HTS alternatives for DNA species identification.

HTS platforms are capable of producing millions of individual sequence reads from hundreds or thousands of samples simultaneously. Each read is generated from a single DNA template molecule (with or without PCR amplification) and thus can more effectively distinguish contamination or mixed species samples than Sanger sequencing. After the HTS run, consensus sequences can be constructed from a multiple sequence read alignment using sequence alignment software [11–13]. During this process, sequence reads are clustered based on sequence homology with the aim of generating a single consensus sequence for every source of input DNA. In the case of species identification, this means that one consensus sequence should be generated per distinct taxon present within a sample. Such consensus sequences can then be compared against a reference database to identify the species, just as for Sanger sequence data. Consensus sequences generated from contaminants can be easily excluded and the principle donor source of the evidence can be determined. In cases where there is interest in identifying multiple species components within a mixed sample, for example plant and animal DNA mixtures found in traditional medicines (TM), individual reads can be clustered to form multiple consensus sequences which are then identified individually to their taxonomic origin [14,15].

HTS platforms such as Illumina MiSeq, IonTorrent and PacBio platforms have already been established for DNA barcoding within the research community [16–19] and some have been subsequently assessed and validated for forensic applications. Their error rates vary from as high as 0.2–16% in PacBio, ~1% for PGM IonTorrent and as low as 0.01% on Illumina platforms [20–22]. The latter technology is the next most accurate one after the Sanger sequencing and has therefore been most widely subjected to validation studies for forensic purposes. For example, the Illumina MiSeq FGx Forensic Genomics System has been validated for human STR profiling [23,24] and the MiSeq utilized in conjunction with the PowerSeq™ CRM Nested System is now used for mitotyping [25]. Similarly, a multi-locus DNA metabarcoding method based on Illumina MiSeq amplicon sequencing has been validated for identification of endangered species in mixed samples for non-human forensic purposes [26]. However, despite their potential use in species identification, the production of HTS data evaluated to date requires expensive equipment and is cost-effective only when large numbers of samples are analysed simultaneously. As such most HTS platform have been inaccessible or impractical for many low-throughput end-users, such as wildlife forensic scientists.

Here, we explore and validate an alternative HTS DNA sequencing method for species identification using the MinION™ DNA sequencing device from Oxford Nanopore Technologies (ONT) that has the potential to quickly deliver reliable and cost-effective results without needing access to big sequencing facilities. Despite its potential as a field-deployable system, for the purposes of forensic analysis, the method we present using ONT's MinION platform would still need to be conducted in a quality-assured laboratory environment.

1.3. MinION sequencing - potential benefits and limitations

ONT's MinION sequencer is a small and inexpensive nanopore-based DNA sequencing platform. This relatively new technology has several important advantages over other HTS platforms. It has long-read output; a low initial startup cost ($1000); fairly simple and quick library preparation protocols, and it allows for rapid real-time analysis and data transfer via a single USB connection to a standard laptop computer. The ease of use and rapid processing time is especially beneficial when rapid identification of an evidence sample is required [27]. The limitation of this platform is the high error rate, spanning from 5% to 25% in raw reads [28–30]. Nonetheless, the error rate is constantly decreasing; first, due to updates in sequencing chemistry and improvements to the nanopores released by ONT [31] and second, because of continuously evolving bioinformatic tools that are specifically developed to handle the nature of sequencing error from the MinION. These computational tools are consequently tested and evaluated by the scientific community [28,32]. To date, several studies have shown that data produced by the MinION are sufficiently accurate to generate a consensus sequence from a single species sample for species identification with >99% accuracy [33–36].

Nevertheless, generation of consensus sequences using MinION data is often laborious requiring multiple software programs, frequent and cumbersome reformatting of data and advanced bioinformatics skills. A new program that addresses these issues has been recently developed. The NGSpeciesID program was specifically built as a user-friendly tool that generates a highly accurate consensus sequences from long-read amplicon-based high-throughput sequencing platforms [37]. It includes clustering of the reads to filter out contaminants or reads with

high error rates and employs error correction strategies specific to the MinION sequencing platform. This raises the possibility that definitive DNA sequences suitable for forensic application can be generated despite errors present in individual sequence reads or reads showing variation.

1.4. Scope and purpose of developmental validation study

In order to use the MinION for forensic purposes we need to investigate the intrinsic differences between Sanger and Minion sequence data. The two critical differences are that first, as with all HTS platforms, the MinION produces thousands or millions of reads for the same gene region and the same individual, as opposed to a single sequence read from Sanger sequencing; second, sequencing error rates for individual reads are much higher for the MinION, meaning that potentially all observed reads may differ from each other and from the true biological sequence. Consequently, the consensus sequence is used as the analytical result, as an estimation of the biological sequence. This creates an almost unique situation in forensic science where we would report to the courtroom a result that we have never directly observed from our analytical measurements. Rather, we are deducing the true biological sequence based on a consensus approach where the resulting sequence is generated from a very high number of slightly inaccurate reads.

It is anticipated that the presentation of such consensus sequence data as DNA evidence in court will be challenged and its acceptance will require clear explanation supported by an appropriate validation study. Species identification using the MinION platform requires a bioinformatic pipeline to generate the consensus sequence, which is then compared against a reference sequence database to infer the species of origin. The validation study design must examine the steps in the pipeline to determine possible sources of analytical variation, and assess the accuracy and precision of the result data for determining the true species origin. Here we present the results of a developmental validation study of the NGSpeciesID pipeline, a new bioinformatic tool for identifying species from MinION data.

1.5. Specific aim of the study

To meet the identified need for an affordable, accessible, reliable and rapid method for forensic species identification we set out to design and implement a developmental validation protocol to assess whether or not we can use the MinION platform to generate species-diagnostic DNA sequence data for use as evidence in court. To achieve this, we sought to address the following question: Given the sequence variation observed among reads within a single sample attributed to MinION sequencing error, is it possible to generate a reliable consensus sequence that accurately estimates the true biological sequence for use in forensic genetic species identification?

2. Material and methods

2.1. Data production

2.1.1. Amplification and library preparation

MinION sequence data sets were generated from individual tissue samples of five mammal species; wild boar (*Sus scrofa*), roe deer (*Capreolus capreolus*), chamois (*Rupicapra rupicapra*), Euroasian lynx (*Lynx lynx*) and snow leopard (*Panthera uncia*) and one bird species, the Inca tern (*Larosterna inca*). A species-diagnostic region of the mitochondrial cytochrome b (mtDNA cyt b) gene approximately 421 nucleotide base pairs (bp) long was amplified using the mcb primers [38] previously shown to be applicable to forensic genetic species identification. In the second round of PCR, dual ONT PCR barcodes were attached to the cyt b amplicons using the ONT PCR Barcoding Expansion kit (EXP-PBC001). After both PCR rounds, PCR products were purified and tested for purity and quantity. Samples were grouped into two libraries and prepared using the ONT Ligation Sequencing kit (SQK-LSK108 and SQK-LSK109) and Josh Quick's One-pot ligation protocol for ONT libraries [39].

2.1.2. Sequencing and demultiplexing

After adapter ligation and purification, libraries were prepared and loaded onto two separate flow cells: FLO-MIN106 R9 (Edinburgh) and FLO-MIN106D R9.4.1 chemistry flow cells (ONT, Oxford, UK). The first library was sequenced for 12 h and obtained between 250.000 and 750.000 reads per sample. The second library was run for approximately 1 h to obtain at least 100.000 reads per sample.

MinKNOW (ONT) was used for sequencing and the raw sequence data were basedcalled using Guppy v3.5.1 (ONT) with basecalling model "dna_r9.4.1_450bps_fast.cfg". The qcat software v1.1.0 (ONT) specifically developed for demultiplexing reads barcoded with ONT's barcode kits was used to demultiplex ONT barcoded reads and assign to the correct sample. The same primers were also used to generate DNA nucleotide control sequences via Sanger sequencing of the same set of samples. See Supplementary methods for a full description of the molecular genetic laboratory analysis.

2.2. Data analysis

2.2.1. Sanger sequencing

The bi-directional Sanger sequence data was analysed using Geneious software (Biomatters Inc. New Zealand) to align the two reads and edited by eye to generate a single consensus sequence result, hereafter referred to as the *Sanger control*.

2.2.2. Consensus sequence generation

To generate consensus sequences for each of the replicate subsets, we used the NGSpeciesID pipeline. Briefly, NGSpeciesID is a five-step process that takes raw (demultiplexed) and/or filtered MinION read data and outputs one or more consensus sequences for each sample (Fig. 1). In Step 1, multiple MinION output sequence reads are grouped into self-similar clusters using isONclust software v0.0.4 [40]. In Step 2, a consensus sequence is formed for each cluster containing more reads than an abundance threshold (10% of the total number of reads by default) with SPOA v3.0.1 (https://github.com/rvaser/spoa), which is based on a partial order alignment (POA) algorithm [12,41]. The default abundance threshold of 10% within NGSpeciesID was chosen to monitor for the presence of mixed samples, remove sample contamination and filter out reads showing high error rates. It is designed to be low enough to discriminate contamination from minor species contributions, but high enough to avoid calling clusters based on single or very low read numbers.

In Step 3, the pipeline merges reverse complement clusters if any are present, using pairwise alignment in Parasail [42]. In Step 4, the resulting consensus sequence(s) is/are polished with ONT's Medaka software v0.10.0 (https://github.com/nanoporetech/medaka). Finally in Step 5, the tool removes primer sequences from the consensus sequence and reruns the reverse-complement removal and polishing steps to identify any remaining redundant consensus sequences that were not removed due to presence of primers. The end result of the NGSpeciesID pipeline is one or more consensus sequences for each sample, hereafter referred to as the *MinION consensus*.

2.2.3. Validation study datasets

To standardize the dataset size across all validation steps and to investigate the effect of read depth on the consensus calling, 50 replicates of randomly selected reads were generated (sampled with replacement) at depths of 50, 100, 300, 500, 1000 and 5000 reads. This was performed for each of five mammal and one bird species from the filtered and demultiplexed sequence read files, resulting in a total of 300 datasets per species (filtered and demultiplexed datasets before subsampling contained between 9400 and 114,000 reads per species).

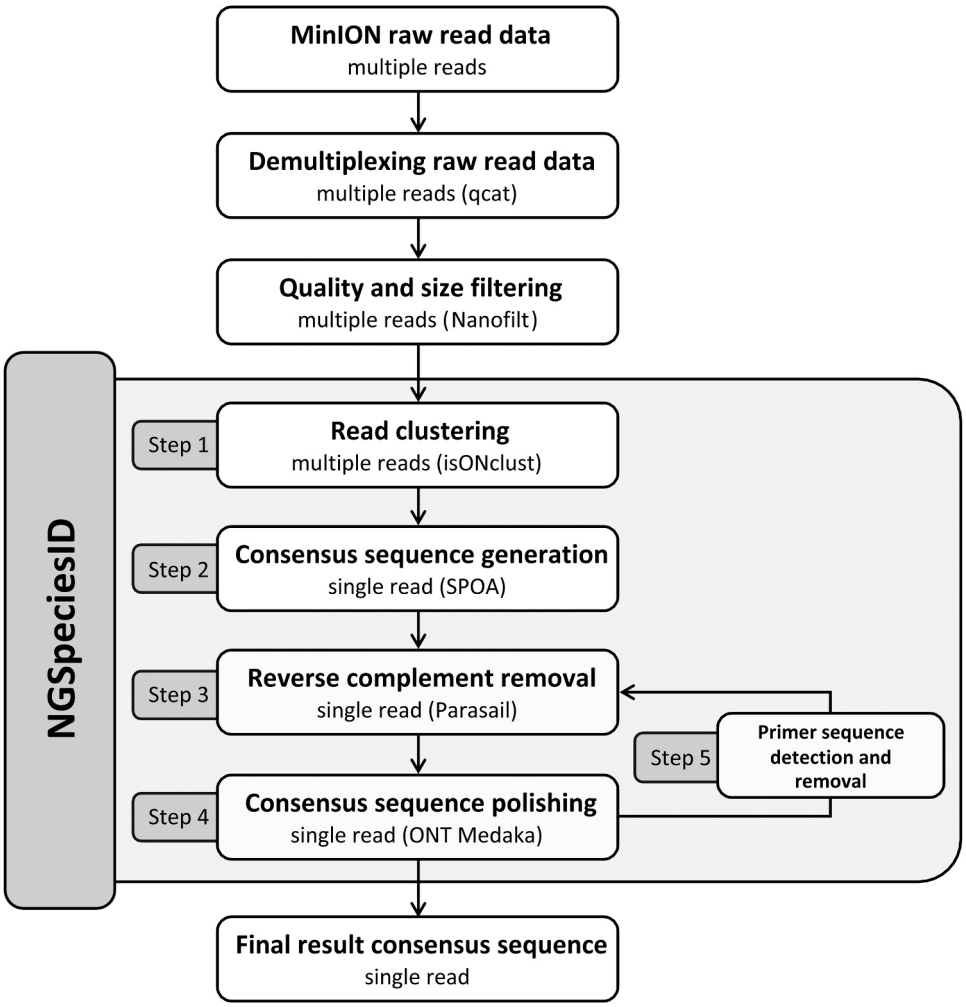

Fig. 1. Flowchart of the steps for generation of a consensus sequence result from the quality and size filtered MinION sequencing data, using the bioinformatic program NGSpeciesID. Individual bioinformatic software at each step is shown in parentheses.

2.3. Pre-validation study – MinION sequence data quality score

A sequence quality score, known as a Phred score, is produced for each base of a raw MinION sequence read, which indicates the level of confidence assigned to each sequence base call and is dependent on the version of the base calling software. Filtering the sequence data based on the Phred score should remove poor quality reads while retaining sufficient read numbers (or *depth*) for subsequent data analysis. The retention of too much poor quality read data can lead to the NGSpeciesID pipeline generating two or more consensus sequences for a single sample, complicating subsequent interpretation and typically rendering the resulting data unsuitable for forensic casework. Thus, in this pre-validation step we evaluated the impact of data quality on the production of multiple consensus sequences, to determine a Phred score quality threshold.

We selected and tested raw reads over a range of Phred quality scores typical for MinION data, from 7 to 12, and read lengths filtered for between 300 and 450 bp, using NanoFilt v2.5.0 [43]. A Phred quality score of 11 was found to be optimal for minimizing multiple consensus sequence results while maximizing good quality read retention for the tested dataset; consequently all data analysis was performed following initial filtering at this quality threshold (Table S1, See Supplementary Methods).

2.4. Validation study design

The validation study was designed to address the overarching issues of whether it is possible to generate an accurate, robust consensus sequence from MinION data and if so, whether this consensus sequence is comparable to the use of Sanger sequencing for forensic genetic species identification. To achieve this aim, the validation process was divided into three stages to assess the performance of the NGSpeciesID pipeline with MinION data (For a visual overview see Table S2). Each validation stage focused on a different aspect of analytical accuracy to investigate: i) variation in reads used to generate the MinION consensus sequence; ii) variation between the MinION consensus and the Sanger control sequence; and iii) variation in the percentage pairwise sequence similarities between these two sequence types and the reference data in the subsequent species identification results.

In terms of analytical precision, in the case of discrete DNA sequence data, the limit of precision for reported measurement is one nucleotide. As this level of precision does not vary, measurement precision was not evaluated in the validation study.

2.4.1. Validation stage 1

The first validation stage examined the distribution of variation among individual reads *within a single replicate*. As multiple reads are used to create a single consensus sequence representing the true sequence of the replicate, it is important to understand the level of variation among consensus composite reads. This was achieved by

addressing two questions:

Validation focus: Measuring the divergence of reads from the consensus.

Q1. *What is the mean pairwise sequence divergence between individual reads and the consensus sequence, at a given read depth?*

Validation focus: Measuring the spread of the data.

Q2. *What is the standard deviation of the pairwise sequence divergence between every read and the consensus sequence, at a given read depth?*

The analysis of per sample read variation serves primarily as an internal validation step that can be conducted for every casework sample. It allows the development of general guidance regarding confidence in the accuracy of the resulting consensus sequence and helps determine how to interpret individual sample results.

To establish a threshold for the within-sample read variation we need to understand the distribution of this variation across multiple replicates. We therefore repeated the calculations of mean and standard deviation for 50 replicates subsampled from the parent dataset, to create a population dataset representing 50 individual samples. From this dataset we investigated uncertainty around the empirical mean estimate of pairwise sequence divergence.

Validation focus: Measuring average variation between the reads and consensus sequences among samples.

Q3. *What is the mean of the mean pairwise divergence of reads around the MinION consensus sequences?*

Validation focus: Measuring the distribution of the sample means.

Q4. *What is the standard error of the result (mean of mean divergences) among replicates?*

We conducted these analyses at each read depth (50, 100, 300, 500, 1000 and 5000) across the six species datasets.

2.4.2. Validation stage 2

In the next part of the validation study we investigated how the MinION consensus sequence compares to the Sanger control sequence generated from the same sample. While Sanger sequencing might be considered as simply another approach for estimating the true biological sequence, it was used here for two important reasons: first, it is the current standard for DNA sequencing in forensic genetics; second, it was used to generate the majority of sequence data in international species reference databases, which are used in comparative sequence similarity searches when identifying unknown evidence samples. MinION consensus sequence replicates were compared to Sanger control sequences from the same specimens using a nucleotide BLAST search v.2.8.1+ [11] to address three questions.

First we examined the typical level of Sanger-MinION sequence divergence across read depths and species.

Validation focus: Comparison of MinION consensus sequence to the reference Sanger sequence.

Q5. *What is the scale and distribution of pairwise sequence divergence between the Sanger control sequence and consensus sequence across read depths and species?*

Second, we wanted to assess a reasonable worst-case scenario of the effects of MinION consensus sequence error by assessing Sanger-MinION divergence at the upper-end of the divergence distribution.

Validation focus: Assessment of maximum likely divergence.

Q6. *What is the maximum pairwise divergence of consensus sequence replicates from the reference Sanger sequence?*

The upper limit of divergence represents the greatest level of divergence from the Sanger data that we expect to see in MinION consensus sequences. This is an important measure, as for example, if the mean divergence is 1% from the Sanger sequence, but 2.5% of the time (our upper limit – 97.5%) we might expect divergence to be 3%, we need to recognize this is as an occasional risk to accurate identification. However, it should be noted that 3% divergence between MinION and Sanger sequences does not necessarily lead to the wrong species identification, even if the next closest species is only 3% diverged from the true species. This is because we do not expect MinION sequence error to reflect the very specific sequence changes at phylogenetically informative nucleotide positions required to transition from one species to another [44].

Lastly, if variation in individual sequence reads affects the accuracy of the resulting consensus sequence we would expect to see a correlation between the read sequence variation to the consensus sequence (Q2) and the deviation of that consensus sequence from the Sanger reference sequence (Q5). With poor quality MinION data or insufficient read depth it is expected that the resulting consensus sequence may not accurately reflect the true DNA sequence of the sample. To assess whether or not this becomes an issue within our pipeline we tested for a relationship between read variation and deviation of the resulting consensus from Sanger sequence:

Q7. *Is there a relationship between the per sample (replicate) read variation (sequence error) and the accuracy of the resulting MinION consensus sequence measured as divergence from the Sanger control?*

At this point we used the available result data to identify an optimal read depth (= 500 reads) to use in the final stage of validation (see Results).

2.4.3. Validation stage 3

In the final validation stage, we examined the impact of sequence differences between MinION consensus and Sanger control sequences on the accuracy of species identification. We compared the results of BLAST analysis using the GenBank database for Sanger control sequence and the 50 replicate MinION consensus sequences for each species.

Validation focus: Qualitative species identification comparison.

Q8. *Does the MinION consensus sequence return the same species assignment as the Sanger control sequence in a BLAST analysis?*

Lastly, we assessed the specificity of the MinION consensus sequence for species identification. In forensic genetic species identification, the degree of sequence divergence between the two highest ranked species in the BLAST result (the barcoding gap) is an indicator of how much confidence we have in obtaining a specific species identification result. To address this issue, if MinION and Sanger sequences returned the same qualitative species result in the BLAST search (Q8), we then examined how the MinION compares to the Sanger sequence in terms of the pairwise sequence divergence to the next closest species. All MinION consensus and Sanger sequences across species were aligned with Mafft v7.450 in Geneious® 11.1.5 to identify common regions with sequence differences.

Validation Criteria: Species specificity comparison.

Q9. *Does any observed difference between the level of Sanger control sequence divergence from its two highest ranked sequence similarity results, and the level of MinION consensus sequence divergence from its two highest ranked sequence similarity results, affect the resulting species identification?*

All statistics for the validation study were calculated using packages dplyr and plyr in R Studio version 1.1.463[45] and modules statistics and math in Python. Graphs were plotted in R, using ggplot package.

3. Results

3.1. Validation stage 1

Variation in reads within a sample replicate (Q1 and Q2).

Q1. *What is the mean pairwise sequence divergence between individual reads and the consensus sequence, at a given read depth?*

Q2. *What is the standard deviation of the pairwise sequence divergence between every read and the consensus sequence, at a given read depth?*

Results of mean pairwise sequence divergence between individual reads and the consensus sequence ranged between 0.044 and 0.064 across all read depths and all species, with associated standard deviation estimates ranging from 0.0091 to 0.0281. This is in line with the typical per read sequence error rate of ~5% observed for the ONT MinION. No distinct pattern was observed in mean pairwise divergence across species and read depths. The spread of data around the mean decreased with increasing read depth (Fig. 2 (wild boar) and Fig. S1 (all species)).

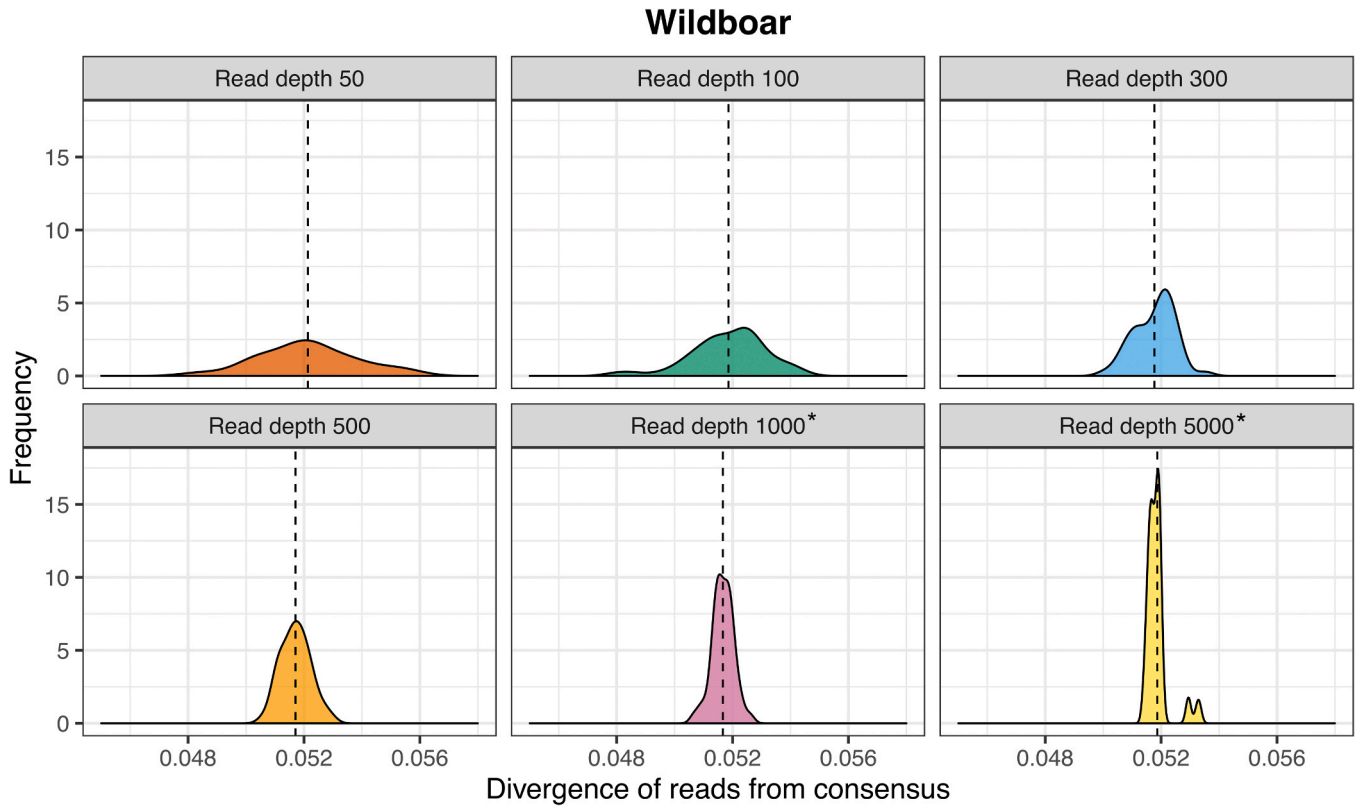

Fig. 2. Mean divergence of individual reads from consensus sequence result (n = 50) across six different reads depths. Results are shown for a single species – wild boar, across 50 replicates at each of six different read depths. The density plots display the spread of mean divergence of reads from the consensus sequence across six read depths. The means of the mean of divergence of different read depths are presented as dashed lines. *the 1000 and 5000 read depth results are probably not representative as in some species we were sampling the same reads more often.

For example, mean pairwise sequence divergence of reads from the consensus sequence for wild boar at read depth of 500 (suggested read depth for casework) ranged from 0.0506 to 0.0530 across the 50 replicates (Fig. 2, Fig. S1 (all species), Table S3 (all data)).

Variation among sample replicates (Q3 and Q4).

Q3. *What is the mean of the mean pairwise divergence of reads around the MinION consensus sequences?*

Q4. *What is the standard error of the result (mean of mean divergences) among replicates?*

The mean of the mean divergence of reads (overall mean divergence) from the consensus sequences among 50 replicates varied slightly across species and was the smallest in the Inca tern (read depth of 300 = 0.0482) and highest in the roe deer (read depth 5000 = 0.0582) (Table S4), with associated standard error of mean estimates ranging from 0.00175 to 0.00245 (Table S4).

3.1.1. Validation stage 2

Q5. *What is the scale and distribution of pairwise sequence divergence between the Sanger control sequence and consensus sequence across read depths and species?*

Q6. *What is the maximum pairwise divergence of consensus sequence replicates from the reference Sanger sequence?*

In the case of wild boar, pairwise divergence between the Sanger control and consensus sequences ranged from 0 to 0.00487 at the 97.5th percentile across all read depths (Table 1 (wild boar), Table S5(all species)). The percent of identical consensus sequences to the Sanger control sample ranged from 88% to 100% at the read depths of 500 and 1000 (Table 1 (wild boar), Table S5 (all species)). Maximum pairwise divergence of consensus sequences (replicates) from the Sanger reference sample ranged between 0.0027 and 0.0055 (Table 1 (wild boar), Table S5 (all species)).

Q7. *Is there a relationship between the per sample (replicate) read variation (sequence error) and the accuracy of the resulting MinION consensus sequence measured as divergence from Sanger control?*

Within the observed range of read variation around the consensus sequence there was no correlation between the level of read variation and the level of consensus sequence deviation from the Sanger reference sequence across all read depths (Fig. 3 (wild boar), Fig. S2 (all species)).

Table 1
Divergence of consensus sequence from Sanger reference sample. There was no variation from the Sanger control sequence at read depths of 500 and 1000 for the wild boar sample.

Sample	Read depth	Pairwise divergence from Sanger at different quantiles				Max observed divergence	% of identical consensus seq. to Sanger
		q = 25	q = 50	q = 75	q = 97.5		
Wild boar	50	0	0	0	0.00487	0.0055	88
	100	0	0	0	0.0027	0.0027	88
	300	0	0	0	0	0.0027	98
	500	0	0	0	0	0	100
	1000	0	0	0	0	0	100
	5000	0	0	0	0.0027	0.0055	94

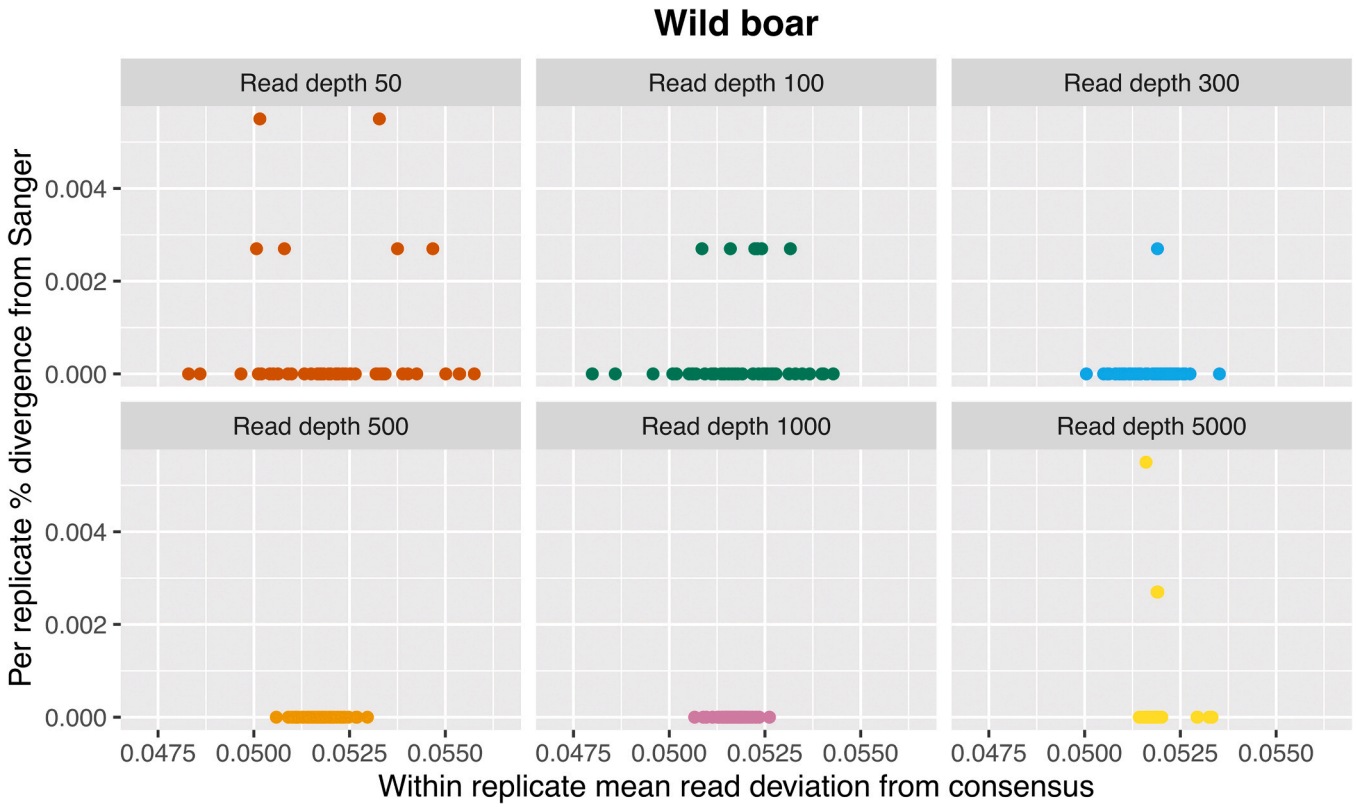

Fig. 3. The relationship between within-replicate mean read variation and consensus percent divergence from the Sanger sequence. No correlation was observed between the read variation around consensus sequence and the consensus deviation from the Sanger sequence across 50 replicates at six different read depths. The result at each read depth is shown in a separate box and in a different colour.

Based on the data generated to this point, we selected a single read depth to use for subsequent stages of the validation study. A read depth of n = 500 was selected as being optimal in terms of the observed consistency in generating a single consensus sequence, with minimal variation in read divergence from the consensus and in divergence from the Sanger control sequences, across all species.

3.1.2. *Validation stage 3*

Q8. *Does the MinION consensus sequence return the same species result as Sanger under a BLAST analysis?*

In the BLAST analysis, all consensus sequences and their Sanger reference sequences returned the same species across all species datasets at a read depth of 500 (Table 2, Tables S6-S11 (full result for all consensus sequence replicates and Sanger reference samples).

Q9. *Does any observed difference between the level of Sanger control sequence divergence from its two highest ranked species sequence similarity results, and the level of consensus MinION sequence divergence from its two highest ranked species sequence similarity results, affect the resulting species identification?*

Wild boar had all 50 consensus sequences identical to the Sanger reference sequence and thus no change was observed in the first species similarity or the barcoding gap (Table 2).

In **roe deer**, 23 of the consensus sequences were identical to the Sanger sequence and were 100% identical to the first species in GenBank (Table 2). The remaining 27 consensus sequences were identical to each other (consensus 2). They deviated from the first species in GenBank by 0.24% (1 bp) and this led to an overall reduction in the barcoding gap from 1.90% to 1.66% between the first and the second most closely related species.

Chamois had 47 consensus sequences identical to the Sanger (consensus 1; see Table 2). The three remaining consensus sequences each had 1 bp change and deviated from the first species in GenBank by 0.24% without leading to a big difference (0.01% decrease) in the barcoding gap between first and second species (consensus 3 and 4).

In the case of the **lynx**, there were 49 consensus sequences identical to the Sanger sequence and one consensus sequence that differed by 1 bp (Table 2). That single consensus sequence had a 0.24% difference from the first GenBank species and this had no impact on the barcoding gap between the first and second species.

For **snow leopard** 49 of 50 consensus sequences were identical to the Sanger sequence (Table 2). For one consensus sequence there was a 1 bp change but this had no impact on the barcoding gap between the first species and second species (distances from the Sanger and MinION consensus sequences were identical).

The **Inca tern** had 47 identical consensus sequences (consensus 1) to the Sanger, two identical sequences with the same 1 bp change (consensus 2) and one with a unique 1 bp change in the sequence (consensus 3; see Table 2). The difference of 1 bp in the third consensus sequence led to a very slight increase (0.04% higher) in the barcoding gap between the first and the second species.

4. Discussion

4.1. *Validation stage 1*

The purpose of validation stage 1 was to examine the level of divergence among individual sequence reads and how they contribute to the generation of a single consensus sequence for each sample.

The value of the mean pairwise sequence divergence of a single read from the consensus is approximately equivalent to the observed MinION sequencing error. In our datasets, the observed mean divergence varied between 4.4% and 6.4% which is within expected boundaries for ONT MinION sequencing error [28]. There was little variation around the mean, either among reads within a sample, or among replicate samples,

Table 2
Species specificity comparison**.** All resulting consensus sequences and Sanger control sequences returned the same species at a read depth of 500 under the BLAST analysis. Differences between consensus sequences and Sanger control sequence had no impact on the species identification result in GenBank. The barcoding gap is the difference in the percent similarity of the first species minus percent similarity of the second species. Overall reduction in the barcoding gap when using MinION consensus sequences was negligible (indicated in bold).

Species	Sequence type	No. of replicates (out of 50)	% Divergence from Sanger	First species name	% sim to 1st species	2nd species name	% sim to 2nd species	1st to 2nd species gap (barcoding gap)	Sequence length (bp)	Sequence divergence details
Wild boar	Sanger (control)			Sus scrofa	100	Sus barbatus	97.15	2.85	421	
	Consensus 1	50	0.00	Sus scrofa	100	Sus barbatus	97.15	2.85	421	
Roe deer	Sanger (control)			Capreolus capreolus	100	Capreolus pygargus	98.3	1.9	420	
	Consensus 1	23	0.00	Capreolus capreolus	100	Capreolus pygargus	98.3	1.9	420	
	Consensus 2	27	0.24	Capreolus capreolus	99.76	Capreolus pygargus	98.1	**1.66**	421	Insertion (G homopolymeric region,129 position)
Chamois	Sanger (control)			R.rupicapra	99.27	R.pyrenaica	95.62	3.65	411	
	Consensus 1	47	0.00	R.rupicapra	99.27	R.pyrenaica	95.62	3.65	411	
	Consensus 2	1	0.24	R.rupicapra	99.03	R.pyrenaica	95.38	3.65	410	Deletion (G,homopolymeric region,47 position)
	Consensus 3	1	0.24	R.rupicapra	99.03	R.pyrenaica	95.39	**3.64**	412	Insertion (G, 393 position)
	Consensus 4	1	0.24	R.rupicapra	99.03	R.pyrenaica	95.39	**3.64**	412	Insertion (G, homopolymeric region, 71 position)
Lynx	Sanger (control)			Lynx lynx	100	Lynx pardinus	94.92	5.08	413	
	Consensus 1	49	0.00	Lynx lynx	100	Lynx pardinus	94.92	5.08	413	
	Consensus 2	1	0.24	Lynx lynx	99.76	Lynx pardinus	94.67	5.09	413	Substitution (G->C, 345 position)
Snow leopard	Sanger (control)			Panthera uncia	100	Panthera pardus	91.9	8.1	421	
	Consensus 1	49	0.00	Panthera uncia	100	Panthera pardus	91.9	8.1	421	
	Consensus 2	1	0.24	Panthera uncia	100	Panthera pardus	91.89	8.11	421	Substitution (T->C, 420 position)
Inca tern	Sanger (control)			Larosterna inca	99.52	Gelochelidon nilotica	91.45	8.07	421	
	Consensus 1	47	0.00	Larosterna inca	99.52	Gelochelidon nilotica	91.45	8.07	421	
	Consensus 2	1	0.24	Larosterna inca	99.52	Gelochelidon nilotica	91.45	8.07	421	Substitution (A->G, 1st position)
	Consensus 3	2	0.24	Larosterna inca	99.52	Gelochelidon nilotica	91.41	8.11	422	insertion (C, 419 position)

suggesting that outliers with unusually high levels of sequence error are very rare. Similarly, there was very little variation among read depths and across species; the lowest mean pairwise divergence of reads from the consensus were consistently observed at read depths between 300 and 1000. As read depth decreases (e.g. around n = 50) we expect to see mean divergence values increase due to stochastic variation and an insufficient number of reads being sampled to accurately estimate the mean. At much higher read depths (e.g. n = 5000), consistent sequence errors may occur at a frequency that prevents the bioinformatic algorithm from filtering such errors out, leading to their retention in the datasets and marginally increasing mean divergence values. Our results lead to an optimal read-depth being observed at an intermediate value (300–1000 reads). We should note however that in some species, replicate samples at read depths of n = 1000 and n = 5000 were subsampled from datasets with a low number of reads (lynx: ~9400, roe deer: ~22,500, chamois: ~15,900 and wild boar: ~22,800 - after filtering) which will have caused a degree of pseudo-replication as many reads are shared between the replicate data sets. This resulted in similar results across replicate samples (although with higher mean divergence values) and consequently yielded artificially narrower distributions of mean divergence values.

4.2. Validation stage 2

The overall purpose of this validation stage was to compare the MinION consensus sequence with the Sanger control sequence from the same individual sample and to characterize any differences between them.

The MinION consensus sequencing results displayed either no difference, or extremely small differences, when compared to the Sanger control samples. No MinION consensus sequence replicate deviated from the Sanger control sequence by more than 1 bp over the ~420 bp sequence length. The maximum sequence divergence between MinION consensus and Sanger control sequences, across replicates and species at read depth n = 500 was 0.0024%, 1 base per 420 bp sequence. This compares to a Sanger sequence error rate 0.001% (0.42 bases per 420 bp sequence). There was no observed relationship between the read variation within a sample (mean read divergence from the consensus) and the accuracy of the resulting consensus sequence against the Sanger control sample for our data. As the MinION sequencing error increases, at some point we would expect the mean read deviation from the consensus to reduce the accuracy of the resulting consensus sequence, with a correlated increase in divergence from the Sanger control sequence. However, this was not observed across the range of sequence error values within the dataset, indicating that per read error rates were sufficiently low and did not compromise the consensus sequence accuracy.

4.3. Validation stage 3

The overall purpose of validation stage three was to investigate the impact of differences between the MinION consensus and Sanger control sequence on the subsequent species identification result.

For this assessment we examined the impact of using individual observed consensus sequences on species identification results, rather than mean consensus sequence divergence, in order to evaluate the worst-case scenario in terms of the least accurate consensus sequence data. Nevertheless, every MinION consensus sequence replicate returned the same species identification result as the Sanger control.

The impact of MinION consensus divergence from Sanger controls on the power of the consensus sequence to differentiate the true (1st ranked) species from the next most similar species was either very small or not observed at all, demonstrating that the barcoding gap is effectively maintained when using MinION consensus sequences. As the nature of the divergence between the two sequence types is not associated with phylogenetic variation, there is no expectation that, say, a 1% divergence of the MinION consensus from the Sanger control would shift the consensus sequence closer to the next most phylogenetically similar species, ranked second in sequence similarity search results. Indeed, in the case of the lynx and the Inca tern, the difference observed in one MinION consensus sequence replicate marginally increased the barcoding gap. The largest reduction in the barcoding gap was observed in the roe deer (from 1.9% to 1.66%) due to a 1 bp insertion in a G homopolymeric region. The potential impact of such as result is to require the identification result to be interpreted more cautiously; it does not, however, elevate the risk of misidentification.

4.4. Implementation – per sample validation

In addition to performing developmental validation of the use of MinION sequence data for species identification, this study also provides the basis for case-by-case internal validation to assess whether individual sample results can be considered within the validated scope. There are two important parameters to consider: 1) the sequence quality filter Phred score which we set at a minimum threshold value of 11 and 2) the read depth: for which, based on our results, we recommend a sequence depth of 500 reads. Given these parameters we can assess the validity of per sample sequence data by examining the sequence read variation around the consensus sequence. If that variation is below the highest variation observed in this study (6% error rate) then we can be confident that the resulting consensus sequence will be sufficiently accurate for species identification. Beyond this value, there is a risk that sequence error will affect the accuracy of the consensus sequence, leading to divergence from the true sample sequence. Where sequence quality is poor and a threshold Phred score of 11 cannot be met, it will still be possible to derive an accurate consensus sequence for a sample; however, in these instances the practitioner needs to be very cautious in using and interpreting the consensus sequence as it may be prone to higher levels of deviation from the true sequence than were observed in this study.

4.5. Potential applications

The results of this validation study support the use of MinION sequencing data for forensic genetic species identification. This expands on a number of existing applications in fields such as on-site food authentication [46] and biodiversity assessment [32], to enable its use in non-human forensic genetics.

Species identification using ONT's MinION is of particular interest to the wildlife forensic community, where there is an urgent need for cost-effective laboratory-based sequencing solutions in countries where access to traditional Sanger sequencing and more recent larger HTS platforms is severely restricted. This causes delays in casework processing time, as it is typically necessary for samples to be transported abroad for analysis. The international shipping of biological material can be substantially delayed by conservation and wildlife trade laws and regulations and, furthermore, may be subject to legal challenge if those involved in the analysis are not available to provide witness testimony in court. The low cost and relatively easy implementation of MinION sequencing offers great advantages in terms of accessibility for forensic laboratories with low purchase and maintenance costs and no need for changes to existing infrastructure. In addition, the availability of the more cost effective Flongle cell will lower costs substantially in future.

5. Conclusions

- The experiments performed in this validation study demonstrate that it is possible to produce an accurate and reliable single consensus sequence using the ONT MinION sequencer.
- The use of the NGSpeciesID pipeline with appropriate filters to generate a single consensus sequence enables a species identification

result that is considered robust enough for forensic genetic species identification.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

This work was financially supported by the ELK ┼Emma-Louise Kessler Fund┼, (Switzerland) [grant number F-41813-04, 2018]. Funding for the WCS work was provided by the G. Unger Vetlesen Foundation, (USA). The authors would like to thank the two anonymous reviewers for their comments and suggestions in the improvement of this manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2021.102493.

References

[1] R. Ogden, N. Dawnay, R. McEwing, Wildlife DNA forensics - bridging the gap between conservation genetics and law enforcement, Endanger. Species Res. 9 (2009) 179–195, https://doi.org/10.3354/esr00144.
[2] P.D.N. Hebert, A. Cywinska, S.L. Ball, J.R. DeWaard, Biological identifications through DNA barcodes, Proc. R. Soc. B Biol. Sci. 270 (2003) 313–321, https://doi.org/10.1098/rspb.2002.2218.
[3] C.P. Meyer, G. Paulay, DNA barcoding: error rates based on comprehensive sampling, PLoS Biol. 3 (2005) 1–10, https://doi.org/10.1371/journal.pbio.0030422.
[4] W. Parson, K. Pegoraro, H. Niederstätter, M. Föger, M. Steinlechner, Species identification by means of the cytochrome b gene, Int. J. Leg. Med. 114 (2000) 23–28, https://doi.org/10.1007/s004140000134.
[5] N. Dawnay, R. Ogden, R. McEwing, G.R. Carvalho, R.S. Thorpe, Validation of the barcoding gene COI for use in forensic genetic species identification, Forensic Sci. Int. 173 (2007) 1–6, https://doi.org/10.1016/j.forsciint.2006.09.013.
[6] L. Wilson-Wilde, J. Norman, J. Robertson, S. Sarre, A. Georges, Current issues in species identification for forensic science and the validity of using the cytochrome oxidase I (COI) gene, Forensic Sci. Med. Pathol. 6 (2010) 233–241, https://doi.org/10.1007/s12024-010-9172-y.
[7] T. Melton, C. Holland, Routine forensic use of the mitochondrial 12S ribosomal RNA gene for species identification, J. Forensic Sci. 52 (2007) 1305–1307, https://doi.org/10.1111/j.1556-4029.2007.00553.x.
[8] J. Shendure, H. Ji, Next-generation DNA sequencing, Nat. Biotechnol. 26 (2008) 1135–1145, https://doi.org/10.1038/nbt1486.
[9] N.L. Vollmer, A. Viricel, L. Wilcox, M.K. Moore, P.E. Rosel, The occurrence of mtDNA heteroplasmy in multiple cetacean species, Curr. Genet. 57 (2011) 115–131, https://doi.org/10.1007/s00294-010-0331-1.
[10] M.A. Smith, C. Bertrand, K. Crosby, E.S. Eveleigh, J. Fernandez-Triana, B.L. Fisher, J. Gibbs, M. Hajibabaei, W. Hallwachs, K. Hind, J. Hrcek, D.W. Huang, M. Janda, D.H. Janzen, Y. Li, S.E. Miller, L. Packer, D. Quicke, S. Ratnasingham, J. Rodriguez, R. Rougerie, M.R. Shaw, C. Sheffield, J.K. Stahlhut, D. Steinke, J. Whitfield, M. Wood, X. Zhou, Wolbachia and DNA barcoding insects: patterns, potential, and problems, PLoS One 7 (2012), e36514, https://doi.org/10.1371/journal.pone.0036514.
[11] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2.
[12] C. Lee, Generating consensus sequences from partial order multiple sequence alignment graphs, Bioinformatics 19 (2003) 999–1008, https://doi.org/10.1093/bioinformatics/btg109.
[13] R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads, Genome Res 27 (2017) 737–746, https://doi.org/10.1101/gr.214270.116.
[14] M.L. Coghlan, J. Haile, J. Houston, D.C. Murray, N.E. White, P. Moolhuijzen, M. I. Bellgard, M. Bunce, Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns, PLoS Genet. 8 (2012), e1002657, https://doi.org/10.1371/journal.pgen.1002657.
[15] J. Chen, Z. Jiang, C. Li, X. Ping, S. Cui, S. Tang, H. Chu, B. Liu, Identification of ungulates used in a traditional Chinese medicine with DNA barcoding technology, Ecol. Evol. 5 (2015) 1818–1825, https://doi.org/10.1002/ece3.1457.
[16] S. Shokralla, J.F. Gibson, H. Nikbakht, D.H. Janzen, W. Hallwachs, M. Hajibabaei, Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens, Mol. Ecol. Resour. 14 (2014) 892–901, https://doi.org/10.1111/1755-0998.12236.
[17] S. Shokralla, T.M. Porter, J.F. Gibson, R. Dobosz, D.H. Janzen, W. Hallwachs, G. B. Golding, M. Hajibabaei, Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform, Sci. Rep. 5 (2015) 9687, https://doi.org/10.1038/srep09687.
[18] M.A. Forin-Wiart, M.L. Poulle, S. Piry, J.F. Cosson, C. Larose, M. Galan, Evaluating metabarcoding to analyse diet composition of species foraging in anthropogenic landscapes using Ion Torrent and Illumina sequencing, Sci. Rep. 8 (2018) 1–12, https://doi.org/10.1038/s41598-018-34430-7.
[19] P.D.N. Hebert, T.W.A. Braukmann, S.W.J. Prosser, S. Ratnasingham, J.R. DeWaard, N.V. Ivanova, D.H. Janzen, W. Hallwachs, S. Naik, J.E. Sones, E.V. Zakharov, A sequel to sanger: amplicon sequencing that scales, BMC Genom. 19 (2018) 1–14, https://doi.org/10.1186/s12864-018-4611-3.
[20] A.M. Wenger, P. Peluso, W.J. Rowell, P.C. Chang, R.J. Hall, G.T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N.D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.S. Chin, A.M. Phillippy, M.C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F.J. Sedlazeck, J.M. Zook, H. Li, S. Koren, A. Carroll, D.R. Rank, M.W. Hunkapiller, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome, Nat. Biotechnol. 37 (2019) 1155–1162, https://doi.org/10.1038/s41587-019-0217-9.
[21] T.C. Glenn, Field guide to next-generation DNA sequencers, Mol. Ecol. Resour. 11 (2011) 759–769, https://doi.org/10.1111/j.1755-0998.2011.03024.x.
[22] S. Ardui, A. Ameur, J.R. Vermeesch, M.S. Hestand, Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics, Nucleic Acids Res 46 (2018) 2159–2168, https://doi.org/10.1093/nar/gky066.
[23] A.C. Jäger, M.L. Alvarez, C.P. Davis, E. Guzmán, Y. Han, L. Way, P. Walichiewicz, D. Silva, N. Pham, G. Caves, J. Bruand, F. Schlesinger, S.J.K. Pond, J. Varlaro, K. M. Stephens, C.L. Holt, Developmental validation of the MiSeq FGx Forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories, Forensic Sci. Int. Genet. 28 (2017) 52–70, https://doi.org/10.1016/j.fsigen.2017.01.011.
[24] R. England, S. Harbison, A review of the method and validation of the MiSeq FGxTM forensic genomics solution, wiley interdiscip, Rev. Forensic Sci. 2 (2020) 1–19, https://doi.org/10.1002/wfs2.1351.
[25] M.D. Brandhagen, R.S. Just, J.A. Irwin, Validation of NGS for mitochondrial DNA casework at the FBI Laboratory, Forensic Sci. Int. Genet. 44 (2020), 102151, https://doi.org/10.1016/j.fsigen.2019.102151.
[26] A.J. Arulandhu, M. Staats, R. Hagelaar, M.M. Voorhuijzen, T.W. Prins, I. Scholtens, A. Costessi, D. Duijsings, F. Rechenmann, F.B. Gaspar, M.T. Barreto Crespo, A. Holst-Jensen, M. Birck, M. Burns, E. Haynes, R. Hochegger, A. Klingl, L. Lundberg, C. Natale, H. Niekamp, E. Perri, A. Barbante, J.P. Rosec, R. Seyfarth, T. Sovova, C. Van Moorleghem, S. van Ruth, T. Peelen, E. Kok, Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples, Gigascience 6 (2017) 1–18, https://doi.org/10.1093/gigascience/gix080.
[27] J.M. Butler, The future of forensic DNA analysis, Philos. Trans. R. Soc. B Biol. Sci. 370 (2015), https://doi.org/10.1098/rstb.2014.0252.
[28] F.J. Rang, W.P. Kloosterman, J. de Ridder, From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy, Genome Biol. 19 (2018) 1–11, https://doi.org/10.1186/s13059-018-1462-9.
[29] R.R. Wick, L.M. Judd, K.E. Holt, Deepbinner: demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks, PLoS Comput. Biol. 14 (2018) 1–11, https://doi.org/10.1371/journal.pcbi.1006583.
[30] J.L. Weirather, M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X. Wang, D. Buck, K.F. Au, Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis, F1000Research 6 (2017) 100, https://doi.org/10.12688/f1000research.10571.1.
[31] C. Brown, Oxford Nanopore Technologies, Nanopore Community Meeting 2019, New York, December 2019. https://nanoporetech.com/resource-centre/nanopore-community-meeting-2019-technology-update.
[32] H. Krehenwinkel, A. Pomerantz, S. Prost, Genetic biomonitoring and biodiversity assessment using portable sequencing technologies: current uses and future directions, Genes 10 (2019) 858, https://doi.org/10.3390/genes10110858.
[33] A. Pomerantz, N. Peñafiel, A. Arteaga, L. Bustamante, F. Pichardo, L.A. Coloma, C. L. Barrio-Amorós, D. Salazar-Valenzuela, S. Prost, Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building, Gigascience 7 (2018) 1–14, https://doi.org/10.1093/gigascience/giy033.
[34] A. Seah, M.C.W. Lim, D. McAloose, S. Prost, T.A. Seimon, MinION-based DNA barcoding of preserved and non-invasively collected wildlife samples, Genes 11 (2020) 445, https://doi.org/10.3390/genes11040445.
[35] S.T. Calus, U.Z. Ijaz, A.J. Pinto, NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform, Gigascience 7 (2018) 1–16, https://doi.org/10.1093/gigascience/giy140.
[36] A. Srivathsan, B. Baloğlu, W. Wang, W.X. Tan, D. Bertrand, A.H.Q. Ng, E.J.H. Boey, J.J.Y. Koh, N. Nagarajan, R. Meier, A MinIONTM-based pipeline for fast and cost-effective DNA barcoding, Mol. Ecol. Resour. 18 (2018) 1035–1049, https://doi.org/10.1111/1755-0998.12890.
[37] K. Sahlin, M.C.W. Lim, S. Prost, NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data, Ecol. Evol. 11 (2021) 1392–1398, https://doi.org/10.1002/ece3.7146.
[38] S.K. Verma, L. Singh, Novel universal primers establish identity of an enormous, Mol. Ecol. Notes 3 (2003) 28–31, https://doi.org/10.1046/j.1471-8286.2003.0340.x.
[39] J. Quick, One-pot ligation protocol for Oxford Nanopore libraries Protoc. Io. 2018 1 5 doi: 10.17504/protocols.io.k9acz2e.

[40] K. Sahlin, P. Medvedev, De novo clustering of long-read transcriptome data using a greedy, quality-value based algorithm, BioRxiv (2018) 1–16, https://doi.org/10.1101/463463.
[41] C. Lee, C. Grasso, M.F. Sharlow, Multiple sequence alignment using partial order graphs, Bioinformatics 18 (2002) 452–464.
[42] J. Daily, Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments, BMC Bioinform. 17 (2016) 1–11, https://doi.org/10.1186/s12859-016-0930-z.
[43] W. De Coster, S. D'Hert, D.T. Schultz, M. Cruts, C. Van Broeckhoven, NanoPack: visualizing and processing long-read sequencing data, Bioinformatics 34 (2018) 2666–2669, https://doi.org/10.1093/bioinformatics/bty149.
[44] R. Krishnakumar, A. Sinha, S.W. Bird, H. Jayamohan, H.S. Edwards, K.D. Patel, S. S. Branda, M.S. Bartsch, Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias, Sci. Rep. (2018) 1–13, https://doi.org/10.1038/s41598-018-21484-w.
[45] Team RStudio, RStudio: integrated development for R, RStudio, Inc., Boston, MA URL http//www.rstudio.com. 42, 2015. 14.
[46] K. Kappel, I. Haase, C. Käppel, C.G. Sotelo, U. Schröder, Species identification in mixed tuna samples with next-generation sequencing targeting two short cytochrome b gene fragments, Food Chem. 234 (2017) 212–219, https://doi.org/10.1016/j.foodchem.2017.04.178.