RAMBO: Resolving Amplicons in Mixed Samples for Accurate DNA Barcoding with Oxford Nanopore

Andreas Kolter, Paul Hebert

Andreas Kolter: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, writing – review and editing, visualization

Paul Hebert: resources, writing – review and editing, supervision, project administration, funding acquisition.

Abstract

DNA barcoding, the use of short genetic markers to identify and differentiate species, is a foundational tool in ecological and taxonomic research. The method has scaled rapidly with next-generation sequencing technologies, enabling the processing of thousands of specimens in parallel. Nanopore sequencing offers a flexible, low-cost alternative to other platforms, producing full-length reads in real time and supporting applications in remote settings. However, its comparatively high error rate complicates downstream processing, particularly when PCR co-amplifies multiple templates from a single specimen, reflecting pseudogenes, paralogs, or contaminants. We present a novel pipeline for DNA barcoding that resolves mixed sequence signals from Nanopore reads using unsupervised clustering and staged consensus generation, without relying on curated reference databases, taxonomic priors, or error models. Unlike existing approaches to curate Nanopore sequence data that assume a single dominant amplicon per sample or require large sequence divergence among amplicons, our method distinguishes variants differing by 0.15 percent. It combines column-weighted encodings, UMAP projection, and HDBSCAN clustering, followed by conservative consensus refinement. The pipeline was benchmarked and validated using datasets with known composition, including high-fidelity PacBio sequences. The results show that Nanopore barcoding, when paired with appropriate analysis tools, can recover biologically meaningful variation even in technically complex samples. The pipeline is particularly suited for specimens where divergent templates are often co-amplified, including mitochondrial pseudogenes or multicopy nuclear regions like ITS. As such, it provides a generalizable framework for high-resolution Nanopore analysis of complex amplicon data.

Introduction

Understanding global biodiversity is essential for tracking its loss, managing ecosystems, and informing conservation policy (IPBES et al. 2019; Sandall et al. 2023). Yet identifying species remains a major challenge in ecology, taxonomy, and environmental monitoring (Bickford et al. 2007; Goldberg et al. 2016; Li and Wiens 2023; Löbl et al. 2023). Molecular tools, particularly those employing DNA, offer scalable and replicable alternatives to traditional morphology-based approaches and have transformed how biodiversity is documented (Adams et al. 2023; Deiner et al. 2017; Goldberg et al. 2016; Pawlowski et al. 2021).

DNA barcoding identifies species by amplifying and sequencing standardized genetic markers, such as the mitochondrial cytochrome *c* oxidase subunit I (COI) gene in animals or the nuclear ribosomal internal transcribed spacer (ITS) region in fungi and plants from an individual specimen (Chen et al. 2010; Hebert et al. 2003; Schoch et al. 2012). As Sanger sequencing provides sufficient read length for COI and ITS barcodes, it shaped the development of early protocols. However, it only produces clean data (chromatograms) when a single dominant PCR product is amplified, as co-amplified templates are typically masked or result in an entangled sequencing signal (Al-Shuhaib and Hashim 2023; Kommedal et al. 2008).

Next-generation sequencing (NGS) platforms enable greater scalability and produce millions of reads per sample (Porter and Hajibabaei 2018), potentially enabling the separation of target amplicons from occasionally co-amplified by-products (Shokralla et al. 2015). While short-read NGS platforms like Illumina are widely used for metabarcoding due to their high throughput and low error rates, they cannot recover full-length 658 bp COI barcodes without assembly or truncation (Elbrecht and Leese 2017; Shokralla et al. 2015). Although the Sequel II platform from Pacific Biosciences (PB) delivers highly accurate long reads, its high upkeep and size restrict its use to core facilities (Wenger et al. 2019; Huo et al. 2021).

Sequencers from Oxford Nanopore Technologies (ONT) support full-length barcode recovery and their compact size and low-cost makes them suitable for distributed workflows (Seah et al. 2020; Srivathsan et al. 2024). However, their comparatively high

error rate of 1–2.5% (R10.4, dorado v0.31+) reflecting both miscalled bases and indels, particularly in homopolymer regions, requires the combination of multiple reads to generate an accurate consensus sequence (Seah et al. 2020).

Consensus sequence generation per sample is effective when all reads originate from the same underlying PCR template, allowing errors to be averaged out while retaining the true nucleotide sequence (Figure 1). This condition is met in DNA barcoding when a single-specimen sample features just one dominant amplicon (Srivathsan et al. 2024). However, complexity is introduced when multiple templates within a sample are co-amplified (Figure 1). It can be resolved if the templates derive from distantly related taxa, such as organisms from a different family. In such cases, clustering approaches at 3% or higher can effectively identify all underlying templates, as the divergence between them exceeds the error rate of Nanopore by a substantial margin (Hebert et al. 2025). However, discriminating closely related sequences generated by co-amplification of mitochondrial pseudogenes, heteroplasmic variants, or contamination by congeneric species within a sample is challenging, as their sequence divergence is similar to the Nanopore error rate (Bensasson 2001; Hebert et al. 2023; Bonin et al. 2023). These cases result in ambiguous consensus sequences or, more problematically, to the merging of distinct biological entities into a single consensus that obscures the true biological signal, particularly when functional mitochondrial sequences are conflated with ancestral pseudogenes (Parr et al. 2006; Song et al. 2008).

Computational strategies are central to addressing the limitations of Nanopore sequencing for mixed-template amplicon pools (Karst et al. 2021; Rodríguez-Pérez et al. 2021; Sahlin et al. 2021; Srivathsan et al. 2024). Reference guided genome analysis methods such as variant calling and phasing infer variation relative to a reference genome and can be affected by reference bias when the reference is not representative. In biodiversity applications, comprehensive reference resources are often unavailable across taxa and regions, which limits the applicability and interpretability of reference-based analyses (Zhang et al. 2010). Read-to-read self-correction methods likewise assume long molecules with deep coverage, typically over 10 kilobases in length (Y. Liu et al. 2024), and are not applicable to barcode-length fragments. By contrast, barcoding workflows such as ONTbarcoder are tuned for short amplicons but assume a single

dominant sequence per sample, an assumption that is violated in mixed-template contexts (Srivathsan et al. 2024). Together, these constraints motivate analytical approaches that separate and denoise closely related sequences directly from the read set, without relying on genome assemblies, long-read self-correction, single-haplotype assumptions, externally trained error models, or curated taxonomic databases.

These specific requirements are currently unmet and motivated the development of RAMBO (Resolving Amplicons in Mixed Samples for Accurate DNA Barcoding with Oxford Nanopore). As a resolution-aware denoising pipeline for Nanopore amplicon reads, it explicitly quantifies its resolution, defined as the minimal divergence at which variants remain reliably distinguishable in the Nanopore read data. Its performance is validated against ground truth data from high-fidelity PacBio amplicons. The pipeline operates directly on demultiplexed raw reads and yields denoised sequence variants that are generated upstream of, and fully decoupled from, any taxonomic assignment.

The computational approach of combining UMAP (McInnes et al. 2018) with HDBSCAN (McInnes et al. 2017) builds on earlier pipelines (Rodríguez-Pérez et al. 2021). Additionally, the parallel development of PIKE, also arriving at a closely related UMAP–HDBSCAN strategy, provides convergent support for the suitability of this approach to Nanopore amplicon clustering (Krivonos et al. 2025). RAMBO is implemented in R and integrates into existing bioinformatics workflows. We evaluated its performance on three datasets. The first contained pooled COI amplicons from multiple specimens of the same species and was designed to test if low (<1%) sequence divergence can be resolved. The second comprised COI barcode reads from a published study in which consensus calling failed. We reanalyzed these data with our pipeline to resolve base-call ambiguities and recover the correct barcode in the presence of extensive co-amplification. The third dataset consisted of nuclear ribosomal DNA amplicons from Euglossini bees (ITS1, 5.8S, ITS2, and part of 26S rRNA) that had been validated against high-fidelity PacBio reference sequences. Together, these datasets show that the pipeline can resolve sequence diversity within barcoding markers even when divergent templates are co-amplified.

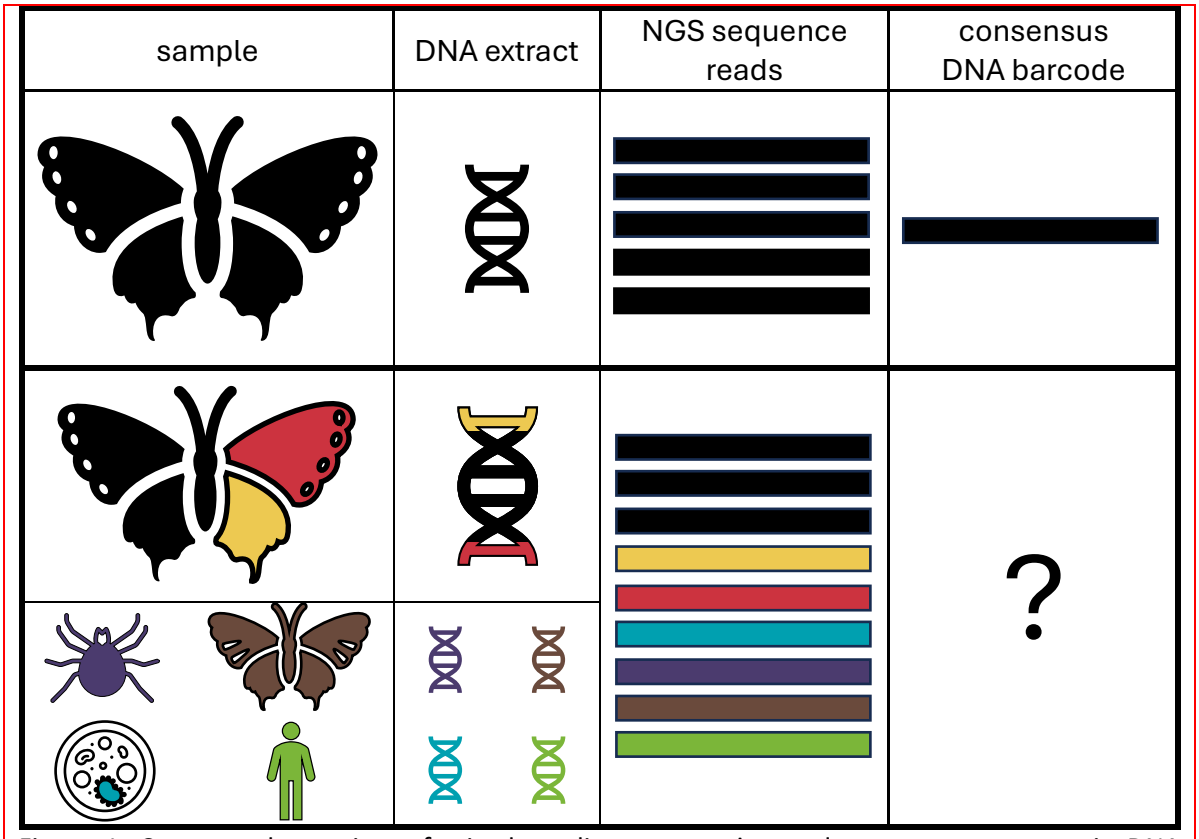

Figure 1: Conceptual overview of mixed-amplicon generation and consensus recovery in DNA barcoding.

Ideal barcoding scenario (row 1) in which DNA extracted from a single specimen yields sequence reads from one template derived from the target organism, allowing straightforward consensus sequence generation.
Realistic barcoding scenario (row 2) illustrating complications. DNA extracts may contain multiple co-amplified templates, including heteroplasmic variants (yellow), mitochondrial pseudogenes (red), host-associated parasites (purple), microbial symbionts (blue), and external contaminants from other species (brown) and humans (green). These templates can be sequenced simultaneously and produce overlapping read pools, complicating clustering and consensus calling.

Material and Methods

Overview and Implementation

The pipeline is implemented as a R workflow for clustering and denoising by consensus generation of Nanopore amplicon reads without reliance on reference sequences. It produces one or multiple consensus sequences from mixed-input amplicon samples where each sample is represented by a FASTQ file. Performance-critical steps such as pairwise Hamming calculations were accelerated using custom C++ functions via Rcpp (Eddelbuettel et al. 2008). The pipeline depends on publicly available packages including Biostrings, dbscan, dplyr, FNN, ggplot2, igraph, magrittr, plotly, ShortRead, stringr and

uwot (H. Pagès 2017; Hahsler et al. 2019; Wickham et al. 2014; Beygelzimer et al. 2010; Wickham 2009; Csárdi et al. 2025; Bache and Wickham 2014; Sievert 2020; Morgan et al. 2009; Wickham 2016; Melville 2019; Corporation and Weston 2011; Microsoft and Weston 2009). Throughout the pipeline, read-level quality was expressed as a log10 based Phred score corresponding to the mean per base error probability.

In short, the pipeline performs the following steps:

(1) The pipeline accepts primer-trimmed, demultiplexed, unaligned FASTQ reads and aligns them using MAFFT in FFT-NS-2 mode (Figure 2), preserving the quality scores. All downstream calculations operate on these exact aligned strings with no column rearrangements.

(2) Homopolymer regions are very susceptible to sequencing errors and are therefore masked to prevent technical artifacts from distorting downstream clustering. Homopolymers are restored after the clustering step, prior to consensus calling.

Homopolymer free sequence alignments were processed in three main stages: feature detection & feature encoding (3), distance calculations (4), distance adaptive merging (5) and density clustering (6):

(3) Informative features were defined at the level of individual alignment columns. For each column, every distinct nonconsensus base (A, C, G or T) that satisfied joint thresholds on minimum count and minimum frequency was treated as a candidate feature. To control false positives, candidate features were evaluated with a one-sided binomial test against a smoothed background error estimate. Features passing this test were encoded as binary presence or absence across reads.

(4) Based on these features two distance metrics were constructed:

(4.1) First, a pairwise Hamming distance matrix was computed across all alignment columns containing features and subjected to hierarchical clustering with average linkage (UPGMA) using the R function hclust, yielding a dendrogram. This dendrogram was cut at multiple non-uniform height intervals, chosen from densely populated regions of the merge height distribution to capture frequent branching events. For each cut, each disjoint clade passing a set of thresholds (default: > 10 sequences and > 1% of total sample sequence count)was analyzed independently. For each clade, column wise

nucleotide frequencies were computed for the clade versus all other sequences in the sample. As a contrast measure we used the total variation distance, defined as one half of the sum of absolute differences between the two base frequency distributions. Columns whose total variation distance exceeded a preset threshold for the contrast between the clade and the remainder of the sample (default 0.25) were marked as informative and received an additive score equal to this distance. Because this procedure was repeated across all clades and cut heights, we applied a diminishing returns scheme in which the incremental weight assigned to a column at each additional hit was divided by the number of times that column had previously been flagged as informative. This prevented early, deep splits in the dendrogram from dominating the final column weights relative to finer scale differences near the tips. The column weights were applied in a one hot encoding matrix, where each nucleotide at each alignment position was represented by a value which was the product of its Phred derived confidence (%) and the corresponding column weight. UMAP embedding reduced the complexity down to 5 dimensions.

(4.2) Second, a weighted Jaccard type distance was then calculated on the binary feature matrix. For each pair of reads, the weighted overlap of shared features (intersection) and features present in at least one read (union) was converted into a similarity score that emphasised shared minor allele patterns while downweighting features that were common across many reads. Agreement on joint absence of features was captured by a complementary term that corrected the effective denominator in sparse regions. In addition, a signed cosine component on a two hot encoding of presence and absence was included to distinguish pairs with opposing feature patterns from pairs that were simply uninformative. The two regimes were blended adaptively: one that strengthened links between reads sharing rare, specific features, and one that penalised links dominated by ubiquitous or weak features, with mixing weights depending on per pair feature overlap and per read feature coverage. The final distance between two reads was defined as one minus this hybrid similarity.

(5) The two distance measures were combined into a single mixed distance that interpolates between Jaccard and UMAP based geometry. Jaccard distances on the binary feature matrix and Euclidean distances in UMAP space were each divided by their

median over all finite entries, so that typical distances had similar numerical magnitude and neither component dominated purely by scale. For each read, an adaptive Jaccard weight was derived from its feature coverage: reads with few informative features received a higher Jaccard weight so that rare markers contributed more strongly, whereas feature rich reads received a lower Jaccard weight and a correspondingly higher UMAP weight to suppress bridge forming dense feature patterns. For any read pair, the effective Jaccard weight was taken as the larger of the two read specific weights, the UMAP weight as the complementary fraction with a small minimum contribution, and the mixed distance was defined as the weighted average of the rescaled Jaccard and UMAP distances.

(6) Density based hierarchical clustering (HDBSCAN) was applied to the mixed distance matrix from the non homopolymer masked alignments while scanning a grid of candidate minPts values (which determine the effective neighbourhood size used for density estimation). For each candidate, a single clustering run yielded four metrics: the fraction of reads in non noise clusters, the mean membership probability, a parsimony score reflecting the gap between the minimal allowed and largest observed cluster size, and an ambiguity penalty from the number of alignment columns with substantial secondary nucleotides within clusters. These metrics were rescaled to the interval [0,1] across candidates and combined by a fixed weighted sum favouring low noise, high membership probability, low ambiguity and moderate parsimony, and the neighbourhood size with the highest score was retained. A final HDBSCAN run with this setting on the same mixed distance provided definitive cluster labels and outlier scores; reads assigned to the noise cluster or exceeding a fixed outlier threshold were flagged as outliers. For each remaining cluster, an IUPAC-aware consensus sequence (IUPAC threshold default: 25%) was computed from the non homopolymer masked alignment, and clusters whose consensus sequences differed by at most a user defined Hamming distance were merged (default threshold: 0, implying no post hoc merging). Optionally, a UMAP implementation limited to three dimensions is capable of generating an interactive 3D cluster map (Supplementary Material 1a).

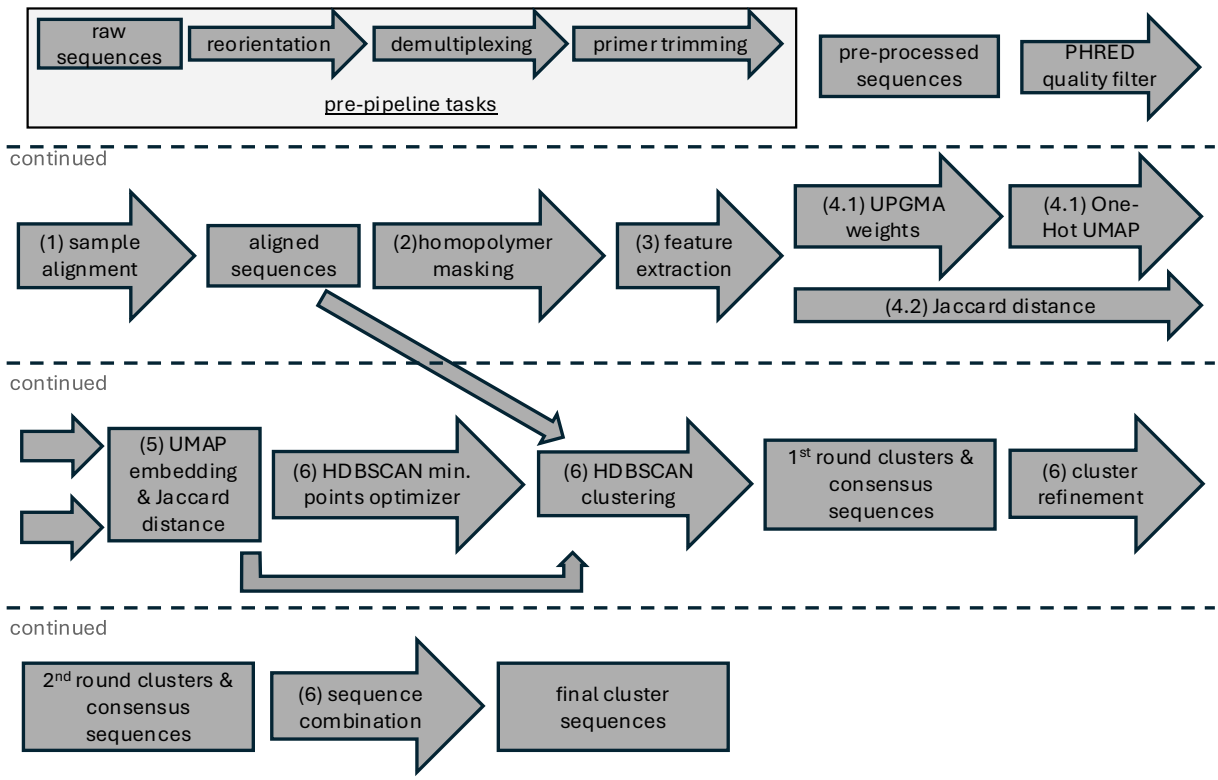

Figure 2: Workflow of the RAMBO amplicon clustering pipeline.

The pipeline starts from preprocessed long-read amplicon sequences and applies per-sample alignment, homopolymer-aware masking, and feature extraction to construct a weighted distance representation. Dimensionality reduction and density-based clustering are performed using UMAP and HDBSCAN, with automatic optimization of clustering parameters. Consensus sequences are generated for each cluster in a first pass, followed by cluster refinement and a second consensus round to obtain final cluster sequences and summary metrics. Numbered steps correspond to the workflow stages described in the Methods.

Code and data availability

The full pipeline, including installation instructions, test data, and parameter documentation, is publicly available at https://github.com/Andreas-Bio/RAMBO. Benchmark datasets are deposited under 10.5281/zenodo.17940933.

Dataset 1: Processing of Low-Divergence Barcode Samples

The first test dataset consisted of sequences recovered from 23 *Phyllocnistis populiella* (Lepidoptera, BIN BOLD:ACX5473) specimens represented in „Barcode 100K Specimens: In a Single Nanopore Run“ (Hebert et al. 2025). These samples were chosen because their COI consensus sequences differed by between 1 and 10 nucleotides in pairwise Hamming distance, corresponding to 0.15% to 1.5% divergence, with a median of 4

differences (Supplementary Material 1b). This narrow divergence range was used to assess the pipeline's ability to resolve fine-scale sequence variation.

Raw sequencing reads and consensus sequences were obtained from the Dryad repository (Prosser et al. 2024). Basecalling was performed using Dorado v0.3.1 in FAST mode, yielding reads with an average estimated error rate of 2.2%, corresponding to a Phred score of 16.6 (Supplementary Material 1c). Reads were reoriented into 5'–3' direction and subjected to demultiplexing, primer trimming, and length filtering to retain fragments between 650 and 665 base pairs. Read counts per sample after filtering ranged from 90 to 379 with a median of 231 and a standard deviation of 57 (Supplementary Material 1d). Demultiplexing information was removed from the sequences and embedded in the read headers, and the 23 sample-specific read sets were pooled into a single FASTQ file for downstream analysis. Sample identifiers were retained from the original publication.

The denoising pipeline was executed using standard parameters. The average Phred quality threshold for alignment was set to 16 to accommodate the lower accuracy of FAST basecalls.

Finally, the same dataset has been analyzed with the PIKE pipeline in single mode, adjusting the maximum sequence length filter to 700, and the results were compared (Krivonos et al. 2025).

Dataset 2: Resolving Mixed Templates to Remove Consensus Ambiguities in Nanopore Reads

The second test dataset consists of sequences from 66 insect samples (Supplementary Material 1e): 42 Diptera, 13 Hymenoptera and 11 belonging to other orders. The most abundant families were Phoridae (Diptera, 14 specimens) and Sciaridae (Diptera, 11 specimens). Samples were selected from the same study as dataset 1 based on their high number of ambiguous nucleotides (N >= 8) in the published COI barcode consensus sequence (Hebert et al. 2025). Raw ONT reads for these samples were reprocessed with RAMBO to evaluate its ability to recover unambiguous consensus sequences from challenging input. The RAMBO consensus calling used the same IUPAC threshold of 25%. These 66 samples comprised 9,519 raw reads in total, with read depths ranging from 89

to 243 (Supplementary Material 1e). Basecalling was performed using Dorado v0.3.1 in high-accuracy (HAC) mode, resulting in an average estimated read error rate of 1.8%. All remaining preprocessing steps and pipeline parameters were identical to those applied to the *Phyllocnistis* dataset.

Dataset 3: Resolving Intragenomic ITS Variants: A Strategy for Accurate Marker Recovery from Multicopy Ribosomal Repeats

The third dataset was generated in this study and comprises nrDNA sequences of *Euglossini* (Apidae, Hymenoptera), more specifically the ITS1 & 5.8S & ITS2 & 28S (partial, approx. 2050 bp) regions. First, samples were collected between 500 masl (meters above sea level) and 1440 masl in Kosñipata, Cusco, Peru (Supplementary Material 1f). One hind leg per specimen was exported to the Centre of Biodiversity Genomics, University of Guelph, Canada for DNA extraction, PCR and sequencing. PCR reactions were set up using GXL polymerase (Takara Bio Inc., Japan; Cat. No. R050A) and performed in two rounds, targeted template amplification and indexing (Supplementary Material 1e & 1k). After pooling the indexed PCR reactions, each pool was sequenced on both a Nanopore PromethION flowcell and on a PacBio Sequel II SMRT 8M flowcell, following standard protocols with negative size selection of 3 kb and below during library preparation. Preprocessing involved demultiplexing, primer trimming, and removal of a so-far undescribed repetitive ITS1 region using Tandem Repeats Finder (Benson 1999), to avoid artefactual inflation of genetic distances caused by uncharacterized sequence patterns (Supplementary Material 1i). Due to the complex pre-processing operations, all steps were performed in FASTA format. After preprocessing, samples were filtered by read count, retaining 369 paired samples (738 total), each with more than 100 reads from both PacBio (PB) and Oxford Nanopore Technologies (ONT) data. For samples with more than 500 reads, a random subset of 500 was used, and when read counts differed between technologies, the larger set was downsampled to match the smaller. Reads shorter than 5,000 bp were excluded. The total number of reads for each set of 369-samples was 135,703. The median read depth was 436 for both technologies, with median read lengths of 5,614 bp (ONT) and 5,639 bp (PB). The pipeline was run with default parameters, except that FASTA files were used as input.

Results

Dataset 1: Benchmarking Fine-Scale Clustering on Closely Related Barcode Sequences

To evaluate the performance of our pipeline on a realistic but tightly constrained dataset, we analyzed a *Phyllocnistis populiella* (Lepidoptera) dataset (Hebert et al. 2025). A total of 5,463 reads were provided as input to the pipeline. After the initial low-quality filter, 4,937 reads remained (-9.6%). Subsequent alignment-based filtering used an average Phred score threshold of 16, further reducing the dataset to 3,351 reads (-32.1%). Of these, 3,105 reads (92.6%) were successfully assigned to clusters, while 246 reads (7.3%) were classified as noise by the clustering algorithm and were excluded from further interpretation.

Each of the 23 samples was associated with a single dominant cluster (Figure 3). Dominant cluster sizes ranged from 55 to 216 reads, with a median of 132. Cluster purity, defined as the proportion of reads contributed by the dominant sample, ranged from 97.8% to 100%, with a median of 100%, indicating that clusters were effectively sample specific (Figure 3). The elevated proportions (> 40%) of noisy reads in BIOUG94608-F10 and BIOUG94608-C06 are explained by the combination of very small sequence divergence between the two specimens (1 bp) and a large set of variable pseudogene-like COI amplicons, identifiable by lengths of 622 to 638 bp.

Subsampling down to 600 reads (~18% of 3351) still preserved the pattern observed earlier, with the 23 samples each occupying a unique cluster (Supplementary Material 1j).

Running the PIKE pipeline returned 59 raw clusters and 41 after size filtering with an average cluster purity of 66% (Supplementary Material 1k). Most notably, 12 of the 23 BIOUG samples were found in mixed clusters which also contained a large proportion of reads from other samples. In all of these cases, the Hamming distance between consensus sequences of affected samples was either 1 or 2.

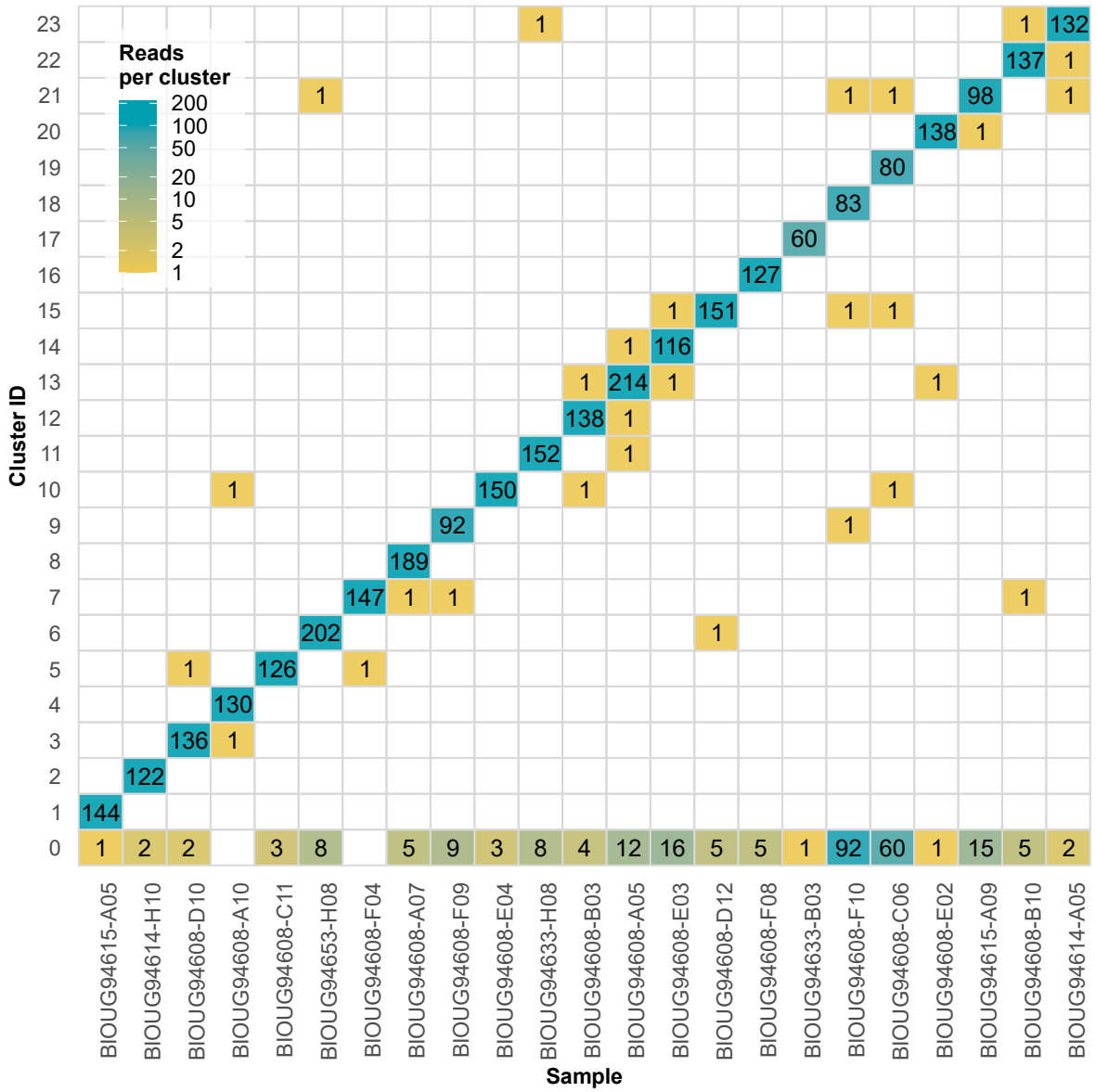

Figure 3: Sample-to-cluster assignments in dataset 1 (*Phyllocnistis populiella*)

The 23 *Phyllocnistis populiella* samples (x-axis) are each assigned to one cluster (y-axis) by the RAMBO pipeline. Reads designated as noise (Cluster ID 0) by HDBSCAN generally only have small counts (color) in relation to the total reads per sample.

Dataset 2: Accurate Separation of Co-Amplified Sequences in Problematic Barcoding Cases

The 66 samples, selected for their high barcoding difficulty, from the 100K dataset (Hebert et al. 2024) initially featured 686 Ns, while after processing with the RAMBO pipeline only 17 Ns (-97.5%) remained in consensus sequences (Figure 4). The median number of Ns dropped from 10 to 0 (Figure 4). The median number of reads included in the consensus decreased from 141 to 61 (Supplementary Material 1e). The RAMBO dominant cluster

consensus had a similarity of 99.98% to previously published results when ambiguous nucleotides were interpreted based on IUPAC definitions (Supplementary Material 1e). Manual inspection of the non-dominant clusters revealed highly variable COI sequences to be the root cause of the ambiguous consensus.

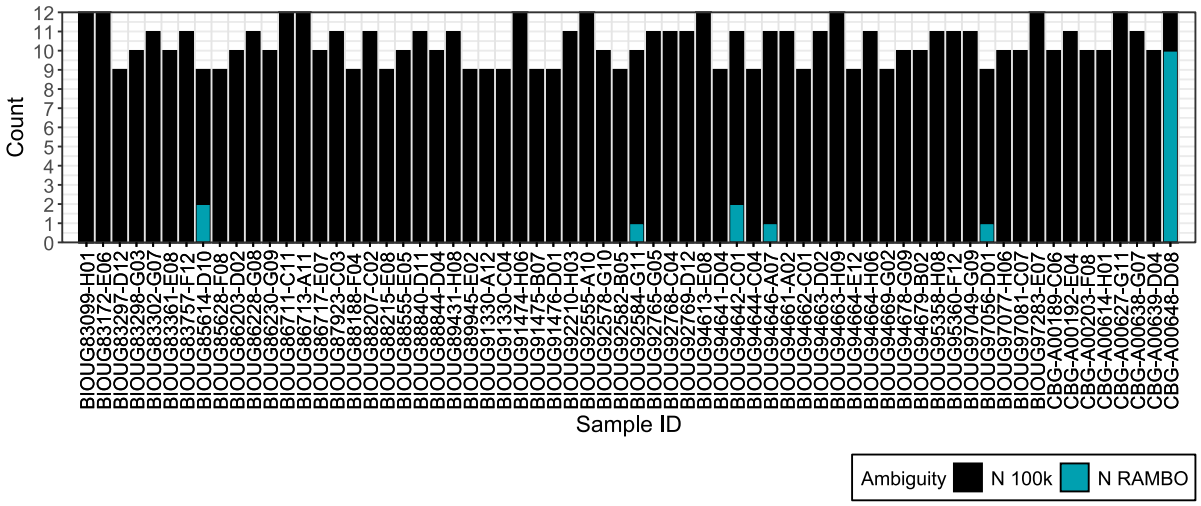

Figure 4: Comparison of Ambiguous Base Counts: 100K Study vs. RAMBO Output

Barplot illustrating the number of ambiguous bases in consensus sequences from the 100K study by Hebert *et al*. (2024) in black and the same raw reads reprocessed using the RAMBO pipeline (cyan). Each bar represents a unique sample, with the y-axis showing the count of ambiguous nucleotides.

Dataset 3: Multicopy nrDNA Marker Recovery across ONT and PacBio Platforms

We processed 738 samples in parallel (15 threads) in 407 min on an AMD Ryzen 5900X using < 64 GB RAM. Initial alignment lengths ranged from 5,728–10,682 bp (median 7,408 bp), consistent with expected insertions.

To test robustness, we compared consensus sequences generated from ONT and PacBio reads of the same PCR amplicon pools, processed identically and treating PacBio as a high-accuracy reference with starting quality near Q40. For each sample, instead of a strict comparison between the dominant ONT and PacBio clusters, candidate clusters for cross-platform comparison were selected from the two most abundant clusters per sequencing technology based on read count to avoid over interpreting platform dependent fluctuations in cluster dominance among closely competing clusters. The

ONT–PacBio cluster pair with the highest sequence concordance was selected as the corresponding cluster pair for cross-platform comparison.

The mean sequence identity across platforms was 99.98% (Figure 5). Differences were driven primarily by indels, with a median of 1.73 gap openings per dominant cluster and a mean of 0.14 substitutions per cluster, corresponding to roughly one substitution per 41,500 bp. When gap openings were ignored, the mean identity increased to 99.998% while manual inspection indicated indel hotspots in and around homopolymer tracts. Within clusters, per-read divergence was higher for ONT than for PacBio (1.60% versus 0.18%), yet ONT consensuses remained stable and accurate, corresponding to an effective accuracy near Q35 and representing a major improvement over raw ONT reads, which had an effective quality near Q22 (Supplementary Material 1c). Ambiguity in the dominant consensus was low, with ONT averaging 0.27 ambiguous nucleotides per sequence, about one per 21,000 bp, versus 0.18 for PacBio, about one per 33,000 bp (Figure 5). Median read support for dominant clusters was 115 reads for ONT and 108 for PacBio, and dominant consensus lengths had a median of 5632 bp with a range from 5353 to 6291 bp.

When we extended the analysis to all clusters across 369 paired samples, PacBio produced more clusters per sample, mainly in the low-support regime. ONT yielded 1371 clusters in total with a median of three clusters per sample, whereas PacBio yielded 2111 clusters with a median of five clusters per sample. The paired difference in total cluster counts (ONT − PacBio) had a mean of −2.005 (Wilcoxon signed-rank $V = 10{,}307$, $p = 1.8 \times 10^{-18}$). For well supported clusters with at least 50 reads, ONT showed slightly more clusters per sample than PacBio, with a mean paired difference of +0.154 and a median of 0 ($V = 10{,}478$, $p = 3.6 \times 10^{-3}$). For clusters with fewer than 50 reads, PacBio produced more clusters, with a mean paired difference of −2.16 ($V = 10{,}766$, $p = 1.5 \times 10^{-19}$). Across all clusters, ambiguity remained low but was higher for ONT than for PacBio, with mean numbers of ambiguous IUPAC positions of 1.58 versus 0.32 per consensus, but a median of 0 for both. High-ambiguity clusters with more than 0.5% IUPAC positions were rare: we only observed 15 for ONT and none for PacBio, with median read depths of 23.

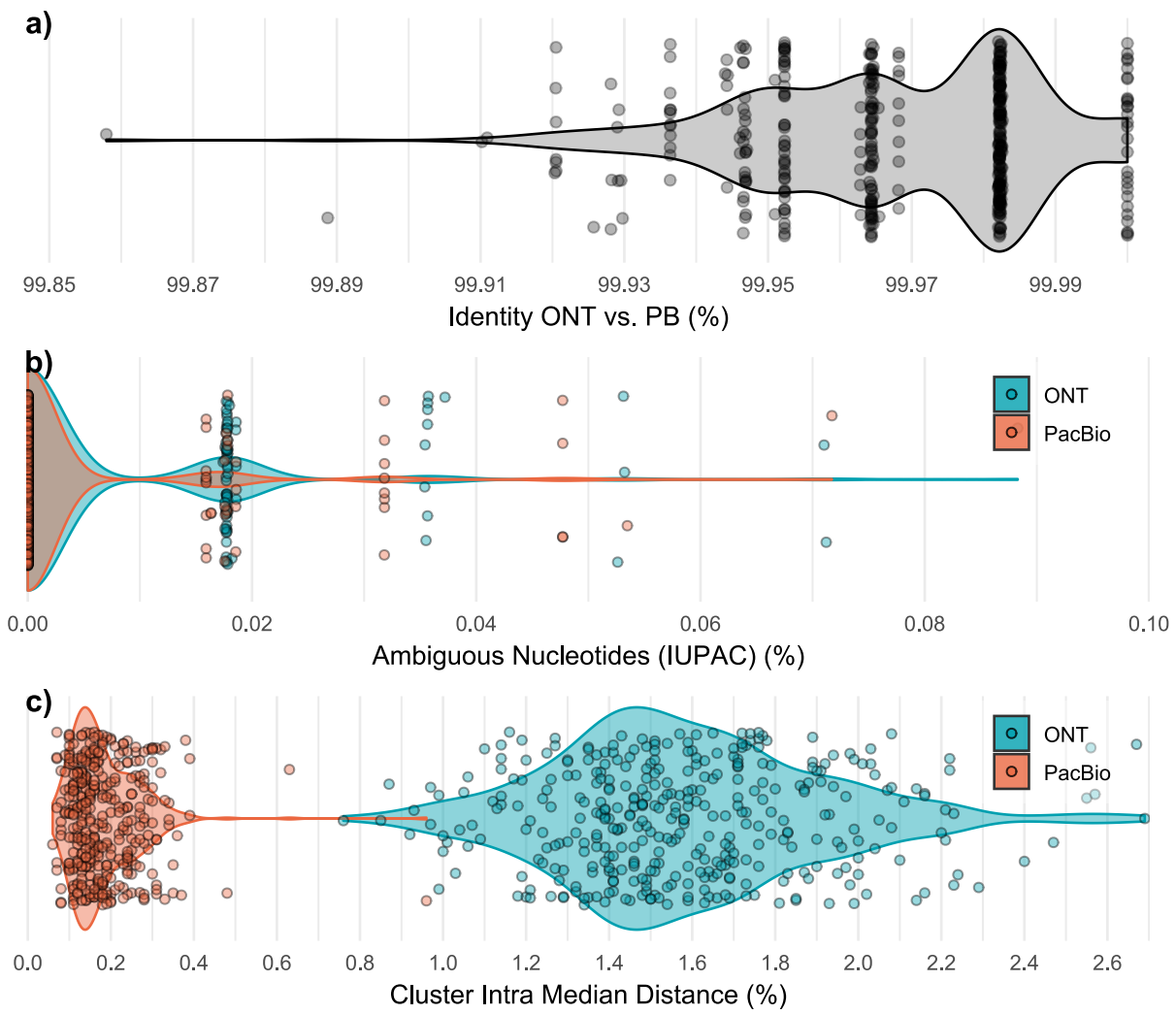

Figure 5: Euglossini ITS sequenced by ONT vs. PB

Violin plots show kernel density with width scaled by count. Points represent individual cluster values. a) Sequence identity between the consensus of the dominant ONT cluster and the consensus of the dominant PacBio cluster from the same specimen. Identity was computed on aligned sequences with IUPAC aware matching and terminal gaps ignored. The banded pattern is caused by discrete mismatch values (N=0,1,2,3...), offset by varying sequence lengths. b) Proportion of ambiguous nucleotides IUPAC in the ONT and PacBio consensuses of dominant clusters. Two outliers were removed for readability. c) Intra cluster median distance for ONT and PacBio, calculated as the median pairwise Hamming distance among inlier reads within each cluster. Higher values indicate greater within cluster divergence. One outlier was removed for readability.

Discussion

Reliable species identification depends on transparent, regimented data processing. For researchers, it is essential that observed sequence diversity reflects biological reality

rather than methodological artifacts. For regulatory bodies and conservation practitioners, reproducibility and clearly defined performance thresholds are necessary for integrating molecular methods into monitoring frameworks. Without rigorous benchmarking and resolution metrics, it remains unclear whether a pipeline can distinguish closely related taxa or deliver consistent results across sample types.

Dataset 1: Clustering Accuracy, Variant Recovery, and Pseudogene Detection

That error-prone Nanopore reads complicate the discrimination of closely related taxa is well established, particularly when sequence divergence approaches the scale of residual sequencing error (Pomerantz et al. 2022; Ohta et al. 2023). To mitigate this limitation, earlier approaches often relied on prior knowledge of the expected amplicon sequence, assuming the availability of a high-fidelity reference that is present in the sample and can be used to model systematic Nanopore error patterns during consensus generation (Espada et al. 2022). While effective in targeted settings, such strategies are not reference-free per definition and therefore do not directly address resolution limits in closely related, unknown samples, as often present in biodiversity studies.

By contrast, the present framework operates without prior knowledge of sample identity or composition. Sample identities were not used during clustering or consensus generation. Ground truth information was used exclusively for downstream validation of the inferred clusters, not for guiding their formation. A subset of 23 *Phyllocnistis populiella* samples with low pairwise divergence was used to benchmark the resolution of the clustering and consensus framework. Despite the small pairwise genetic Hamming distances (1–10), the pipeline accurately recovered all of the original sample identities. Discrimination down to 0.15% sequence differences demonstrates the method's ability to retain fine-scale variation while minimizing both oversplitting and erroneous cluster merging. This is more than an order of magnitude improvement from values in literature, such as 2% in a fungal metabarcoding study (Dierickx et al. 2024), 5% for Amplicon_sorter and "Barcode 100K Specimens" (Vierstraete and Braeckman 2022; Prosser et al. 2024), or the default cluster setting of 10% for Natrix2 (Deep et al. 2023) or PRONAME (Dubois et al. 2024).

Dataset 2: Improved Consensus Accuracy in Difficult Barcoding Cases

Of the 66 challenging samples from Hebert et al. (2024), 60 produced a fully unambiguous consensus sequence with the RAMBO pipeline, while the remaining showed a reduction in the number of ambiguous bases. This demonstrates the pipeline's ability to isolate the dominant COI signal by separating divergent or noisy reads that would otherwise introduce ambiguity, as reflected by the reduced sequence count contributing to each final consensus.

In the original 100K Nanopore barcoding study (Prosser et al. 2024), contamination and pseudogenes were mitigated by splitting sequences with more than 5% divergence into separate contigs. The persistence of ambiguous nucleotides in the selected samples therefore implies that the interfering reads in these cases diverged from the target sequence by less than 5%. RAMBO addresses this challenge by applying dynamic similarity-based clustering and selecting the most internally consistent cluster with the highest read count for consensus generation, thereby avoiding reliance on fixed divergence thresholds.

In practical terms, these findings suggest that high-throughput barcoding datasets contain a recoverable fraction of sequences previously discarded or flagged due to ambiguity, and that unsupervised read clustering provides an effective route to restoring them. This not only increases the yield of usable barcodes for ecological and taxonomic analyses but also opens the door to recovering accurate sequences in challenging cases involving co-amplified pseudogenes, particularly in taxa such as Orthoptera and Formicidae where conventional methods frequently fail (Song et al. 2014; Cristiano et al. 2014; X. Liu et al. 2024).

Dataset 3: Cross-platform consistency in nrDNA marker recovery

Long-read barcoding has largely relied on PacBio because circular consensus sequencing yields very high per read accuracy and comparatively low indel burden in long amplicons (Wenger et al. 2019). In fungi, sequencing the nrDNA operon has already been established using PacBio, even mentioning the possibility to extend the workflow to

Nanopore (Runnel et al. 2022). However, the authors also mentioned inherent problems, such as widespread contamination in older samples and described intragenomic polymorphisms below 2% divergence. Current Nanopore nrDNA workflows lack mock communities composed of closely related taxa, leaving their sensitivity insufficiently characterized (Ohta et al. 2023).

We selected the nrDNA region rather than COI because it poses a more difficult benchmark due to its greater length and longer homopolymer stretches which stress alignment, gap placement, and subsequently consensus rules. Additionally, intragenomic variation potentially introduces divergent intragenomic variants (Feliner and Rosselló 2007). Homopolymer tracts were masked before clustering during distance calculation, but not during consensus calling. Differences between consensus sequences in ONT and PB are best explained by indel errors, consistent with literature (Liu-Wei et al. 2024).

Overall, PB establishes a stringent reference, ONT reaches comparable consensuses under the same pipeline, and the residual differences show where support is lowest. For routine marker work this means that PB and ONT produce very similar dominant clusters, but that clusters with a low read count, especially below 20 reads, should be viewed with caution as some of their signal might be masked by noise. Alternative options to our approach are Unique Molecular Identifer workflows when experimental goals justify the substantial added complexity and costs (Baloğlu et al. 2021).

Comparison with PIKE & NanoCLUST

PIKE, NanoCLUST, and RAMBO all perform unsupervised clustering of Nanopore amplicon reads using UMAP followed by HDBSCAN as the central step of sequence inference. A direct performance comparison with NanoCLUST was not possible because the pipeline is incompatible with read lengths shorter than 1,000 bp, owing to an external dependency designed for genome polishing (Rodríguez-Pérez et al., 2021). This limitation motivated the development of the PIKE pipeline, which was developed independently and in parallel to RAMBO and is included in this study for comparative purposes (Krivonos et al. 2025).

PIKE and NanoCLUST represent each cluster by a single polished sequence generated using Medaka (Oxford Nanopore Technologies Ltd 2025), whereas RAMBO retains and reports intracluster sequence variation during refinement rather than collapsing clusters to a single Medaka-polished sequence. In this study, we show that the underlying assumption of one amplicon per cluster, which allows sequence collapse at the cluster level, does not hold for PIKE in datasets containing closely related sequences (see Results). This distinction is relevant for the interpretation of results and for applications in which single nucleotide differences are informative, such as plant barcoding, where individual SNPs can determine species-level assignment (Chen et al. 2013; Poliakova et al. 2020; Kolter and Gemeinholzer 2021), as well as population-level analyses using COI (Matumba et al. 2020; Mohamadzade et al. 2023; Toli et al. 2023). Retaining information of heterogeneous clusters is pivotal for informing the user about cluster purity and for preventing overinterpretation of the results. It also allows irregular intragenomic nucleotide variation of multicopy barcoding markers, such as variation commonly observed in telomere-proximal regions of nrDNA repeats (Chen et al. 2023), to be incorporated into the cluster signal without imposing prior assumptions about whether such variation is phylogenetically informative or not.

A further limitation concerns how pipeline performance is evaluated. In both PIKE and NanoCLUST, performance is primarily assessed using mock communities and reported in terms of taxonomic completeness or correctness (Rodríguez-Pérez et al. 2021; Krivonos et al. 2025). Although mock communities are a valuable validation tool, taxonomy-based metrics conflate sequence inference with primer performance and matching logic, such as user-defined minimum coverage or maximum allowed divergence thresholds. Consequently, such evaluations do not provide a general benchmark for sequence reconstruction performance and do not allow direct assessment of the minimal divergence. The only metric that directly reflects pipeline performance is the sequence identity of the inferred sequences relative to the ground truth, which is in practice a high-fidelity proxy, such as PacBio-derived sequences.

In summary, all three pipelines produce sequence data suitable for ecological applications. However, when closely related taxa must be resolved separately, caution is

warranted because the method of consensus sequence generation determines how much informative variation is being retained.

Limitations

The datasets analyzed in this study were intentionally selected to have a low read depth (approximately 100 reads/sample), although we tested the pipeline with samples between 20 and 10,000 reads per sample. While other pipelines produce results from as few as 5 reads (Prosser et al. 2024), they require that the read set not contain secondary amplicons. Despite the higher read requirement, the cost remains minimal at scale. A PromethION flow cell can conservatively produce at least 50 million reads for approximately USD 1000, corresponding to 500 reads per $0.01.

The high resolution of RAMBO arises from its capacity to preserve site-specific informative positions throughout clustering. This strategy separates closely related variants but assumes that sequences are globally alignable and relies on pairwise alignment, which increases computational cost. For classical barcode libraries with broadly uniform reads, the resulting gap structure remains well behaved and within design limits.

By contrast, highly diverse mixtures with roughly equal contributions from multiple families are outside the intended use. In such cases, many alignment columns become dominated by gaps, including biologically valid positions, and are removed by the global gap filter. This behaviour matches metabarcoding scenarios rather than single specimen barcoding, and RAMBO was not designed for that setting. The algorithm has been tuned for samples with up to about 10,000 reads with downsampling recommended for much deeper libraries.

The pipeline explicitly supports multicopy markers. Not every divergent alignment column is forced into a separate cluster. Heuristics protect the integrity of markers such as ITS by allowing true within sample heterogeneity to appear as IUPAC codes in the consensus. For example, scattered divergent positions without a coherent pattern are not used to split clusters into a quadratic number of minor groups (all vs. all combinations). Conversely, highly divergent reads that fail to form well supported clusters can be collected into small residual clusters that often show elevated IUPAC counts and are easy

to flag or discard. In principle, there is no hard constraint on amplicon length, and both long and short reads can be processed as long as they admit a reliable global alignment.

Conclusions and Outlook

Across the trial datasets, RAMBO met three demanding goals. First, it separated closely related sequences down to roughly 0.15% divergence while avoiding oversplitting, showing that fine-scale variation can be retained without fragmenting true biological units. Second, it recovered the dominant barcode from noisy mixtures that contained co-amplified pseudogenes, so that contaminating templates did not inflate ambiguity in the final sequence. Third, when coverage was adequate, it produced ONT consensus sequences that were highly concordant with their PacBio counterparts, indicating that Nanopore barcoding is close to PacBio-level accuracy at the consensus level.

These results clarify where RAMBO sits within the growing ecosystem of tools for the analysis of amplicon pools analyzed on ONT. For COI barcoding taxa with no pseudogenes, ONTbarcoder is adequate because it assumes a single dominant mitochondrial haplotype per specimen (Srivathsan et al. 2024). When libraries combine strongly divergent barcoding markers in a single sample, Amplicon_sorter can perform well because the markers are separated by large sequence distance (Vierstraete and Braeckman 2022). RAMBO instead targets the common but more subtle cases where several closely related templates co-occur in a barcoding sample. These include nuclear mitochondrial pseudogenes and plastid paralogs co-amplifying with functional copies, multicopy nuclear markers such as ITS with genuine intra individual heterogeneity, and host tissues that carry endophytic fungi, parasites or gut contents that are readily amplified by universal primers (Figure 1). Large barcoding campaigns also introduce mixed individuals, low level contamination, and tag jumping. In all these situations the reads form a small community of closely and more distantly related templates rather than a single homogeneous amplicon, and RAMBO is designed to resolve that structure.

Looking ahead, RAMBO provides a path to high-resolution metabarcoding on Nanopore. A light front end that bins reads into broad taxonomic groups, for example at the family level using reference matches, would separate deep divergence between lineages from the shallow variation within species that RAMBO is designed to resolve. Once reads are

partitioned in this way, the same clustering and consensus strategy can be applied to mixed environmental samples while still distinguishing closely related species. Because family-level assignment is already reliable at Nanopore error rates, this extension is conceptually simple yet carries large practical impact. It would turn low cost long-read barcoding into a general metabarcoding tool that can follow fine-scale changes in community composition, unlike current Nanopore metabarcoding pipelines.

Funding statement & Competing Interests

Funding was provided by the Gordon and Betty Moore Foundation (Andes Amazon Program). The authors declare no competing interests, financially or otherwise.

Data Accessibility Statement

Scripts and data used to create the results presented in this publication are available from: https://github.com/Andreas-Bio/RAMBO. All output files for all pipeline steps and the final cluster sequences can be downloaded at DOI: 10.5281/zenodo.17940933. Insect specimens were collected and exported with permissions MIDAGRI-SERFOR-DGGSPFFS-DGSPFS: D000061-2022 & D000052-2023 & D000358-2023 & D000120-2024.

Adams, Clare I M, Gert-Jan Jeunen, Hugh Cross, et al. 2023. "Environmental DNA Metabarcoding Describes Biodiversity across Marine Gradients." *ICES Journal of Marine Science* 80 (4): 953–71. https://doi.org/10.1093/icesjms/fsad017.

Al-Shuhaib, Mohammed Baqur S., and Hayder O. Hashim. 2023. "Mastering DNA Chromatogram Analysis in Sanger Sequencing for Reliable Clinical Analysis." *Journal of Genetic Engineering and Biotechnology* 21 (1): 115. https://doi.org/10.1186/s43141-023-00587-6.

Bache, Stefan Milton, and Hadley Wickham. 2014. "Magrittr: A Forward-Pipe Operator for R." CRAN: Contributed Packages. The R Foundation, February 25. https://doi.org/10.32614/cran.package.magrittr.

Baloğlu, Bilgenur, Zhewei Chen, Vasco Elbrecht, Thomas Braukmann, Shanna MacDonald, and Dirk Steinke. 2021. "A Workflow for Accurate Metabarcoding Using Nanopore MinION Sequencing." *Methods in Ecology and Evolution* 12 (5): 794–804. https://doi.org/10.1111/2041-210X.13561.

Bensasson, D. 2001. "Mitochondrial Pseudogenes: Evolution's Misplaced Witnesses." *Trends in Ecology & Evolution* 16 (6): 314–21. https://doi.org/10.1016/S0169-5347(01)02151-6.

Benson, G. 1999. "Tandem Repeats Finder: A Program to Analyze DNA Sequences." *Nucleic Acids Research* 27 (2): 573–80. https://doi.org/10.1093/nar/27.2.573.

Beygelzimer, Alina, Sham Kakadet, John Langford, Sunil Arya, David Mount, and Shengqiao Li. 2010. "FNN: Fast Nearest Neighbor Search Algorithms and Applications." March 17. https://doi.org/10.32614/CRAN.package.FNN.

Bickford, David, David J. Lohman, Navjot S. Sodhi, et al. 2007. "Cryptic Species as a Window on Diversity and Conservation." *Trends in Ecology & Evolution* 22 (3): 148–55. https://doi.org/10.1016/j.tree.2006.11.004.

Bonin, Aurélie, Alessia Guerrieri, and Gentile Francesco Ficetola. 2023. "Optimal Sequence Similarity Thresholds for Clustering of Molecular Operational Taxonomic Units in DNA Metabarcoding Studies." *Molecular Ecology Resources* 23 (2): 368–81. https://doi.org/10.1111/1755-0998.13709.

Chen, Jian, Zijian Wang, Kaiwen Tan, et al. 2023. "A Complete Telomere-to-Telomere Assembly of the Maize Genome." *Nature Genetics* 55 (7): 1221–31. https://doi.org/10.1038/s41588-023-01419-6.

Chen, Shilin, Hui Yao, Jianping Han, et al. 2010. "Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species." *PLoS ONE* 5 (1): e8613. https://doi.org/10.1371/journal.pone.0008613.

Chen, Xiaochen, Baosheng Liao, Jingyuan Song, Xiaohui Pang, Jianping Han, and Shilin Chen. 2013. "A Fast SNP Identification and Analysis of Intraspecific Variation in the Medicinal Panax Species Based on DNA Barcoding." *Gene* 530 (1): 39–43. https://doi.org/10.1016/j.gene.2013.07.097.

Corporation, Microsoft, and Steve Weston. 2011. "doParallel: Foreach Parallel Adaptor for the 'parallel' Package." CRAN: Contributed Packages. The R Foundation, December 7. https://doi.org/10.32614/cran.package.doparallel.

Cristiano, Maykon Passos, Danon Clemes Cardoso, and Tânia Maria Fernandes-Salomão. 2014. "Could Pseudogenes Be Widespread in Ants? Evidence of Numts in the Leafcutter Ant Acromyrmex Striatus (Roger, 1863) (Formicidae: Attini)." *Comptes Rendus. Biologies* 337 (2): 78–85. https://doi.org/10.1016/j.crvi.2013.11.007.

Csárdi, Gábor, Tamás Nepusz, Kirill Müller, et al. 2025. *Igraph for R: R Interface of the Igraph Library for Graph Theory and Network Analysis*. V. v2.1.4. Zenodo, released January 25. https://doi.org/10.5281/ZENODO.7682609.

Deep, Aman, Dana Bludau, Marius Welzel, et al. 2023. "Natrix2 – Improved Amplicon Workflow with Novel Oxford Nanopore Technologies Support and Enhancements in Clustering, Classification and Taxonomic Databases." *Metabarcoding and Metagenomics* 7 (October). https://doi.org/10.3897/mbmg.7.109389.

Deiner, Kristy, Holly M. Bik, Elvira Mächler, et al. 2017. "Environmental DNA Metabarcoding: Transforming How We Survey Animal and Plant Communities." *Molecular Ecology* 26 (21): 5872–95. https://doi.org/10.1111/mec.14350.

Dierickx, Glen, Lowie Tondeleir, Pieter Asselman, Kris Vandekerkhove, and Annemieke Verbeken. 2024. "What Quality Suffices for Nanopore Metabarcoding? Reconsidering Methodology and Ectomycorrhizae in Decaying Fagus Sylvatica Bark as Case Study." *Journal of Fungi* 10 (10): 708. https://doi.org/10.3390/jof10100708.

Dubois, Benjamin, Mathieu Delitte, Salomé Lengrand, Claude Bragard, Anne Legrève, and Frédéric Debode. 2024. "PRONAME: A User-Friendly Pipeline to Process Long-Read Nanopore Metabarcoding Data by Generating High-Quality Consensus Sequences." *Frontiers in Bioinformatics* 4 (December): 1483255. https://doi.org/10.3389/fbinf.2024.1483255.

Eddelbuettel, Dirk, Romain Francois, Jj Allaire, et al. 2008. "Rcpp: Seamless R and C++ Integration." CRAN: Contributed Packages. The R Foundation, November 6. https://doi.org/10.32614/cran.package.rcpp.

Elbrecht, Vasco, and Florian Leese. 2017. "Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment." *Frontiers in Environmental Science* 5 (April). https://doi.org/10.3389/fenvs.2017.00011.

Espada, Rocío, Nikola Zarevski, Adèle Dramé-Maigné, and Yannick Rondelez. 2022. "Accurate Gene Consensus at Low Nanopore Coverage." *GigaScience* 11 (November): giac102. https://doi.org/10.1093/gigascience/giac102.

Feliner, Gonzalo Nieto, and Josep A. Rosselló. 2007. "Better the Devil You Know? Guidelines for Insightful Utilization of nrDNA ITS in Species-Level Evolutionary Studies in Plants." *Molecular Phylogenetics and Evolution* 44 (2): 911–19. https://doi.org/10.1016/j.ympev.2007.01.013.

Goldberg, Caren S., Cameron R. Turner, Kristy Deiner, et al. 2016. "Critical Considerations for the Application of Environmental DNA Methods to Detect Aquatic Species." *Methods in Ecology and Evolution* 7 (11): 1299–307. https://doi.org/10.1111/2041-210X.12595.

H. Pagès, P. Aboyoun. 2017. *Biostrings*. Bioconductor, released. https://doi.org/10.18129/B9.BIOC.BIOSTRINGS.

Hahsler, Michael, Matthew Piekenbrock, and Derek Doran. 2019. "Dbscan: Fast Density-Based Clustering with R." *Journal of Statistical Software* 91 (1). https://doi.org/10.18637/jss.v091.i01.

Hebert, Paul D. N., Dan G. Bock, and Sean W. J. Prosser. 2023. "Interrogating 1000 Insect Genomes for NUMTs: A Risk Assessment for Estimates of Species Richness." *PloS One* 18 (6): e0286620. https://doi.org/10.1371/journal.pone.0286620.

Hebert, Paul D. N., Alina Cywinska, Shelley L. Ball, and Jeremy R. deWaard. 2003. "Biological Identifications through DNA Barcodes." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270 (1512): 313–21. https://doi.org/10.1098/rspb.2002.2218.

Hebert, Paul D. N., Robin Floyd, Saeideh Jafarpour, and Sean W. J. Prosser. 2025. "Barcode 100K Specimens: In a Single Nanopore Run." *Molecular Ecology Resources* 25 (1). https://doi.org/10.1111/1755-0998.14028.

Huo, Wenxing, Wei Ling, Zilun Wang, et al. 2021. “Miniaturized DNA Sequencers for Personal Use: Unreachable Dreams or Achievable Goals.” *Frontiers in Nanotechnology* 3 (February): 628861. https://doi.org/10.3389/fnano.2021.628861.

IPBES, Eduardo Brondizio, Sandra Diaz, Josef Settele, and Hien T. Ngo. 2019. *Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. Version 1. Zenodo. https://doi.org/10.5281/ZENODO.6417333.

Karst, Søren M., Ryan M. Ziels, Rasmus H. Kirkegaard, et al. 2021. “High-Accuracy Long-Read Amplicon Sequences Using Unique Molecular Identifiers with Nanopore or PacBio Sequencing.” *Nature Methods* 18 (2): 165–69. https://doi.org/10.1038/s41592-020-01041-y.

Kolter, Andreas, and Birgit Gemeinholzer. 2021. “Plant DNA Barcoding Necessitates Marker-Specific Efforts to Establish More Comprehensive Reference Databases.” *Genome* 64 (3): 265–98. https://doi.org/10.1139/gen-2019-0198.

Kommedal, Øyvind, Bjarte Karlsen, and Øystein Sæbø. 2008. “Analysis of Mixed Sequencing Chromatograms and Its Application in Direct 16S rRNA Gene Sequencing of Polymicrobial Samples.” *Journal of Clinical Microbiology* 46 (11): 3766–71. https://doi.org/10.1128/JCM.00213-08.

Krivonos, Danil V., Dmitry E. Fedorov, Dmitry N. Konanov, et al. 2025. “Pike: OTU-Level Analysis for Oxford Nanopore Amplicon Metagenomics.” *International Journal of Molecular Sciences* 26 (9): 4168. https://doi.org/10.3390/ijms26094168.

Li, Xin, and John J Wiens. 2023. “Estimating Global Biodiversity: The Role of Cryptic Insect Species.” *Systematic Biology* 72 (2): 391–403. https://doi.org/10.1093/sysbio/syac069.

Liu, Xuanzeng, Nian Liu, Xuan Jing, et al. 2024. “Genomic and Transcriptomic Perspectives on the Origin and Evolution of NUMTs in Orthoptera.” *Molecular Phylogenetics and Evolution* 201 (December): 108221. https://doi.org/10.1016/j.ympev.2024.108221.

Liu, Yuansheng, Yichen Li, Enlian Chen, et al. 2024. “Repeat and Haplotype Aware Error Correction in Nanopore Sequencing Reads with DeChat.” *Communications Biology* 7 (1). https://doi.org/10.1038/s42003-024-07376-y.

Liu-Wei, Wang, Wiep van der Toorn, Patrick Bohn, Martin Hölzer, Redmond P. Smyth, and Max Von Kleist. 2024. “Sequencing Accuracy and Systematic Errors of Nanopore Direct RNA Sequencing.” *BMC Genomics* 25 (1): 528. https://doi.org/10.1186/s12864-024-10440-w.

Löbl, Ivan, Bernhard Klausnitzer, Matthias Hartmann, and Frank-Thorsten Krell. 2023. “The Silent Extinction of Species and Taxonomists—An Appeal to Science Policymakers and Legislators.” *Diversity* 15 (10): 1053. https://doi.org/10.3390/d15101053.

Matumba, Tshifhiwa G., Jody Oliver, Nigel P. Barker, Christopher D. McQuaid, and Peter R. Teske. 2020. “Intraspecific Mitochondrial Gene Variation Can Be as Low as That of Nuclear rRNA.” *F1000Research* 9 (August): 339. https://doi.org/10.12688/f1000research.23635.2.

McInnes, Leland, John Healy, and Steve Astels. 2017. “Hdbscan: Hierarchical Density Based Clustering.” *The Journal of Open Source Software* 2 (11): 205. https://doi.org/10.21105/joss.00205.

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. “UMAP: Uniform Manifold Approximation and Projection.” *Journal of Open Source Software* 3 (29): 861. https://doi.org/10.21105/joss.00861.

Melville, James. 2019. “Uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction.” CRAN: Contributed Packages. The R Foundation, April 6. https://doi.org/10.32614/cran.package.uwot.

Microsoft, and Steve Weston. 2009. “Foreach: Provides Foreach Looping Construct.” CRAN: Contributed Packages. The R Foundation, June 27. https://doi.org/10.32614/cran.package.foreach.

Mohamadzade, Saeed, Jiaxing Huang, Jiandong An, and Chuleui Jung. 2023. “Genetic Variation and Phylogenetic Relationships of Commercial Populations of Bombus Ignitus (Hymenoptera, Apidae) with Wild Populations in Eastern Asia.” *Journal of Hymenoptera Research* 96 (June): 495–506. https://doi.org/10.3897/jhr.96.102569.

Morgan, Martin, Simon Anders, Michael Lawrence, Patrick Aboyoun, Hervé Pagès, and Robert Gentleman. 2009. “ShortRead: A Bioconductor Package for Input, Quality Assessment and Exploration of High-Throughput Sequence Data.” *Bioinformatics* 25 (19): 2607–8. https://doi.org/10.1093/bioinformatics/btp450.

Ohta, Atsufumi, Kenichiro Nishi, Kiichi Hirota, and Yoshiyuki Matsuo. 2023. “Using Nanopore Sequencing to Identify Fungi from Clinical Samples with High Phylogenetic Resolution.” *Scientific Reports* 13 (1): 9785. https://doi.org/10.1038/s41598-023-37016-0.

Oxford Nanopore Technologies Ltd. 2025. *Medaka*. Python, C. V. 2.2.0. Oxford Nanopore Technologies Ltd, released December 15. https://github.com/nanoporetech/medaka.

Parr, Ryan L, Jennifer Maki, Brian Reguly, et al. 2006. “The Pseudo-Mitochondrial Genome Influences Mistakes in Heteroplasmy Interpretation.” *BMC Genomics* 7 (1): 185. https://doi.org/10.1186/1471-2164-7-185.

Pawlowski, Jan, Aurélie Bonin, Frédéric Boyer, Tristan Cordier, and Pierre Taberlet. 2021. “Environmental DNA for Biomonitoring.” *Molecular Ecology* 30 (13): 2931–36. https://doi.org/10.1111/mec.16023.

Poliakova, T. A., E. V. Banaev, and M. A. Tomoshevich. 2020. “Identification of Species in the Genus Nitraria L. (Nitrariaceae) Based on Nucleotide Variability of Nuclear Ribosomal DNA.” *Vavilov Journal of Genetics and Breeding* 24 (5): 481–88. https://doi.org/10.18699/VJ20.640.

Pomerantz, Aaron, Kristoffer Sahlin, Nina Vasiljevic, et al. 2022. “Rapid in Situ Identification of Biological Specimens via DNA Amplicon Sequencing Using Miniaturized Laboratory Equipment.” *Nature Protocols* 17 (6): 1415–43. https://doi.org/10.1038/s41596-022-00682-x.

Porter, Teresita M., and Mehrdad Hajibabaei. 2018. “Scaling up: A Guide to High-throughput Genomic Approaches for Biodiversity Analysis.” *Molecular Ecology* 27 (2): 313–38. https://doi.org/10.1111/mec.14478.

Prosser, Sean, Robin Floyd, Saeideh Jafarpour, and Paul Hebert. 2024. “Data from: Barcoding 100K Specimens in a Single Nanopore Run.” Version 4. Dryad, May 30. 16066321257 bytes. https://doi.org/10.5061/DRYAD.41NS1RNP1.

Rodríguez-Pérez, Héctor, Laura Ciuffreda, and Carlos Flores. 2021. “NanoCLUST: A Species-Level Analysis of 16S rRNA Nanopore Sequencing Data.” *Bioinformatics* 37 (11): 1600–1601. https://doi.org/10.1093/bioinformatics/btaa900.

Runnel, Kadri, Kessy Abarenkov, Ovidiu Copoț, et al. 2022. “DNA Barcoding of Fungal Specimens Using PacBio Long-read High-throughput Sequencing.” *Molecular Ecology Resources* 22 (8): 2871–79. https://doi.org/10.1111/1755-0998.13663.

Sahlin, Kristoffer, Marisa C. W. Lim, and Stefan Prost. 2021. “NGSpeciesID: DNA Barcode and Amplicon Consensus Generation from Long-read Sequencing Data.” *Ecology and Evolution* 11 (3): 1392–98. https://doi.org/10.1002/ece3.7146.

Sandall, Emily L., Aurore A. Maureaud, Robert Guralnick, et al. 2023. "A Globally Integrated Structure of Taxonomy to Support Biodiversity Science and Conservation." *Trends in Ecology & Evolution* 38 (12): 1143–53. https://doi.org/10.1016/j.tree.2023.08.004.

Schoch, Conrad L., Keith A. Seifert, Sabine Huhndorf, et al. 2012. "Nuclear Ribosomal Internal Transcribed Spacer (ITS) Region as a Universal DNA Barcode Marker for *Fungi*." *Proceedings of the National Academy of Sciences* 109 (16): 6241–46. https://doi.org/10.1073/pnas.1117018109.

Seah, Adeline, Marisa C.W. Lim, Denise McAloose, Stefan Prost, and Tracie A. Seimon. 2020. "MinION-Based DNA Barcoding of Preserved and Non-Invasively Collected Wildlife Samples." *Genes* 11 (4): 445. https://doi.org/10.3390/genes11040445.

Shafin, Kishwar, Trevor Pesout, Pi-Chuan Chang, et al. 2021. "Haplotype-Aware Variant Calling with PEPPER-Margin-DeepVariant Enables High Accuracy in Nanopore Long-Reads." *Nature Methods* 18 (11): 1322–32. https://doi.org/10.1038/s41592-021-01299-w.

Shokralla, Shadi, Teresita M. Porter, Joel F. Gibson, et al. 2015. "Massively Parallel Multiplex DNA Sequencing for Specimen Identification Using an Illumina MiSeq Platform." *Scientific Reports* 5 (1): 9687. https://doi.org/10.1038/srep09687.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Chapman and Hall/CRC the R Series. CRC Press, Taylor and Francis Group.

Song, Hojun, Jennifer E. Buhay, Michael F. Whiting, and Keith A. Crandall. 2008. "Many Species in One: DNA Barcoding Overestimates the Number of Species When Nuclear Mitochondrial Pseudogenes Are Coamplified." *Proceedings of the National Academy of Sciences* 105 (36): 13486–91. https://doi.org/10.1073/pnas.0803076105.

Song, Hojun, Matthew J. Moulton, and Michael F. Whiting. 2014. "Rampant Nuclear Insertion of mtDNA across Diverse Lineages within Orthoptera (Insecta)." *PLoS ONE* 9 (10): e110508. https://doi.org/10.1371/journal.pone.0110508.

Srivathsan, Amrita, Vivian Feng, Daniel Suárez, Brent Emerson, and Rudolf Meier. 2024. "ONTBARCODER 2.0: Rapid Species Discovery and Identification with Real-time Barcoding Facilitated by Oxford Nanopore R10.4." *Cladistics* 40 (2): 192–203. https://doi.org/10.1111/cla.12566.

Toli, Elisavet-Aspasia, Panoraia Zacharopoulou, Anastasios Bounas, Vassiliki Kati, and Konstantinos Sotiropoulos. 2023. "Conservation Genetics of the Steno-Endemic Chorthippus Lacustris (Orthoptera: Acrididae)." *Journal of Insect Conservation* 27 (6): 907–16. https://doi.org/10.1007/s10841-023-00507-6.

Vierstraete, Andy R., and Bart P. Braeckman. 2022. "Amplicon_sorter: A Tool for Reference-free Amplicon Sorting Based on Sequence Similarity and for Building Consensus Sequences." *Ecology and Evolution* 12 (3): e8603. https://doi.org/10.1002/ece3.8603.

Wenger, Aaron M., Paul Peluso, William J. Rowell, et al. 2019. "Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome." *Nature Biotechnology* 37 (10): 1155–62. https://doi.org/10.1038/s41587-019-0217-9.

Wickham, Hadley. 2009. "Stringr: Simple, Consistent Wrappers for Common String Operations." CRAN: Contributed Packages. The R Foundation, November 9. https://doi.org/10.32614/cran.package.stringr.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Second edition. With Carson Sievert. Use R! Springer international publishing.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2014. "Dplyr: A Grammar of Data Manipulation." CRAN: Contributed Packages. The R Foundation, January 16. https://doi.org/10.32614/cran.package.dplyr.

Zhang, A.B., L.J. He, R.H. Crozier, C. Muster, and C.-D. Zhu. 2010. "Estimating Sample Sizes for DNA Barcoding." *Molecular Phylogenetics and Evolution* 54 (3): 1035–39. https://doi.org/10.1016/j.ympev.2009.09.014.