Received: 13 August 2020 | Revised: 8 April 2021 | Accepted: 4 May 2021

DOI: 10.1111/1755-0998.13425

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Metagenomics: A viable tool for reconstructing herbivore diet

Physilia Y. S. Chua[1,2] | **Alex Crampton-Platt**[3] | **Youri Lammers**[4] | **Inger G. Alsos**[4] | **Sanne Boessenkool**[5] | **Kristine Bohmann**[1]

[1]Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark

[2]Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark

[3]NatureMetrics, Cabi Site, Egham, UK

[4]Tromsø Museum, UiT – The Arctic University of Norway, Tromsø, Norway

[5]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway

Correspondence
Physilia Y. S. Chua, Section for Evolutionary Genomics, Globe Institute, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen K, Denmark.
Email: physilia.chua@sund.ku.dk

Funding information
H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 765000

Abstract

Metagenomics can generate data on the diet of herbivores, without the need for primer selection and PCR enrichment steps as is necessary in metabarcoding. Metagenomic approaches to diet analysis have remained relatively unexplored, requiring validation of bioinformatic steps. Currently, no metagenomic herbivore diet studies have utilized both chloroplast and nuclear markers as reference sequences for plant identification, which would increase the number of reads that could be taxonomically informative. Here, we explore how *in silico* simulation of metagenomic data sets resembling sequences obtained from faecal samples can be used to validate taxonomic assignment. Using a known list of sequences to create simulated data sets, we derived reliable identification parameters for taxonomic assignments of sequences. We applied these parameters to characterize the diet of western capercaillies (*Tetrao urogallus*) located in Norway, and compared the results with metabarcoding *trnL* P6 loop data generated from the same samples. Both methods performed similarly in the number of plant taxa identified (metagenomics 42 taxa, metabarcoding 43 taxa), with no significant difference in species resolution (metagenomics 24%, metabarcoding 23%). We further observed that while metagenomics was strongly affected by the age of faecal samples, with fresh samples outperforming old samples, metabarcoding was not affected by sample age. On the other hand, metagenomics allowed us to simultaneously obtain the mitochondrial genome of the western capercaillies, thereby providing additional ecological information. Our study demonstrates the potential of utilizing metagenomics for diet reconstruction but also highlights key considerations as compared to metabarcoding for future utilization of this technique.

KEYWORDS
environmental DNA, faeces, grouse, plants, shotgun sequencing

Inger G. Alsos, Sanne Boessenkool and Kristine Bohmann contributed equally to this research.

1 | INTRODUCTION

Disentangling trophic relationships is integral to understanding ecosystem functions (Duffy et al., 2007). Given the central role that herbivores play in shaping the structure and species diversity of terrestrial ecosystems (Danell et al., 2006), reconstructing their diet has practical implications for conservation biology (Shipley et al., 2009; Valdés-Correcher et al., 2018). To reconstruct the diet of herbivores, methods such as analysis of browsing signs (Salas & Fuller, 1996) and microhistological examinations of plant remains in gut and crop contents (Borchtchevski, 2009; Gayot et al., 2004; Wegge & Kastdalen, 2008) or in faecal samples (González et al., 2012; Greve Alsos et al., 1998; Iversen et al., 2013; Steinheim et al., 2005) have traditionally been used. However, these methods are often time-consuming and labour-intensive (Ait Baamrane et al., 2012), and more recently, the emergence of high-throughput sequencing (HTS) technologies has enabled DNA sequencing of ingested plants in herbivore faecal samples to become a valuable tool in diet reconstruction (e.g., Ait Baamrane et al., 2012; Soininen et al., 2009).

When using HTS to assess herbivore diet from faecal samples, two main approaches can be used: metabarcoding and metagenomics. The most common approach is metabarcoding, a PCR-based method that employs primers designed to amplify a taxonomically informative marker for one or more taxonomic groups (Valentini et al., 2009). A second approach is metagenomics, in which the total extracted DNA from a sample is sequenced without PCR enrichment of specific markers prior to sequencing (Noonan et al., 2005). When studying the diet of species in cases where we have little prior information on their diet, metagenomics is valuable as in contrast to metabarcoding, it does not require a priori knowledge of which taxonomic groups to target. Also, the most-used plant metabarcoding primer set in diet studies amplifying the *trnL* P6 loop (Taberlet et al., 2007) is not among the standard plant barcodes in public DNA reference databases such as BOLD (Hollingsworth, 2011). This means that the availability of the *trnL* P6 loop reference sequences limits metabarcoding studies in most geographical regions, increasing the potential of utilizing metagenomics for herbivore diet reconstruction as it can utilize plant barcode markers that are well represented in DNA reference databases. Further, given that the total DNA of a sample is sequenced in metagenomics, it offers the potential to simultaneously retrieve a vast wealth of information in addition to diet, such as host mitochondrial genome, gut parasites and host-microbiome, without additional laboratory work which would increase time and costs (Srivathsan et al., 2015, 2016). The ability to rapidly generate data is a great advantage in time-sensitive research where urgent conservation intervention is needed, and samples should be studied exhaustively to maximize ecological information (e.g., for endangered or elusive species).

Taxonomic assignment in DNA-based herbivore diet analyses can be hindered by the limited availability of plant reference sequences. Taxonomically informative plant marker sequences in DNA reference databases are generally short and mostly limited to chloroplast regions (*trnL-F* ~ 994 bp, *rbcL* ~ 654 bp, *matK* ~ 889 bp), and there is a lack of whole genome sequence availability in reference databases (Ford et al., 2009; Hollingsworth, Graham, et al., 2011; Kress & Erickson, 2007; Lahaye et al., 2008). Moreover, bioinformatic pipelines for metagenomic diet analyses are relatively unexplored, while those for metabarcoding are well validated and documented, such as the ecotag program from the OBITools pipeline used for taxonomic assignment (De Barba et al., 2014; Boyer et al., 2016; Hibert et al., 2013; Pegard et al., 2009; Quéméré et al., 2013). In the only two metagenomic diet analyses published to date, which focused on mammals, the taxonomic assignment of plants in herbivore diet was based on matching metagenomic paired-end (PE) reads to chloroplast marker sequences (Srivathsan et al., 2015, 2016). However, limiting taxonomic assignment to chloroplast markers may result in false negatives due to the differences in rates of DNA degradation of genomic regions within the ingested plants; that is, chloroplast genomes in plants may be more prone to degradation than plant nuclear genomes because of the instability of the chloroplast genomes (Xin et al., 2018). Bioinformatically, metagenomics for diet reconstruction remains in an exploratory phase and the lack of an optimized and validated bioinformatic pipeline can dissuade researchers from utilizing this tool.

The use of both chloroplast and nuclear markers as reference sequences for plant identification in herbivore metagenomic studies could increase the number of taxonomically informative sequence reads. Despite this, no metagenomic herbivore diet studies have explored the use of both chloroplast and nuclear markers as reference sequences for plant identification. Further, validation of taxonomic assignments from metagenomics data in herbivore diet studies has relied on comparison with data collected from field observations, and from metabarcoding outputs, as demonstrated by Srivathsan et al. (2016), Srivathsan et al. (2015). For metagenomics to become a valuable tool in herbivore diet studies, we must therefore (i) test and validate the use of bioinformatic parameters that combine chloroplast and nuclear markers for taxonomic assignment, and (ii) develop methods with which to validate bioinformatic steps without field observations. The latter is especially important when studying endangered or elusive animals where behavioural observations in the field are often challenging. Simulation of metagenomic data *in silico* presents a way to meet these needs, and it has become an increasingly popular tool for testing and validating bioinformatic strategies (Escalona et al., 2016; Haiminen et al., 2019). However, this has not been applied to diet studies. Through the generation of *in silico* metagenomic data sets resembling the faecal samples sequenced from the studied herbivore, bioinformatic parameters and taxonomically informative marker combinations used in taxonomic assignment steps can be assessed for accuracy. Subsequently, the validated taxonomic assignment steps can be applied to real data sets (Escalona et al., 2016; Gourlé et al., 2019).

In this study, we demonstrate how *in silico* simulation of metagenomic data sets that resemble faecal samples collected from western capercaillies (*Tetrao urogallus*) can be used to assess and validate taxonomic assignment steps for plant identification in herbivore diet. We then: (i) apply the validated taxonomic assignment steps from the *in silico* simulation to real data obtained from metagenomic shotgun sequencing of faecal samples collected from eight western

capercaillies in Norway, and (ii) compare the metagenomic output with metabarcoding data amplified from the same samples using the P6 loop of the chloroplast *trnL* intron, and analysed the impact of sample freshness with both techniques. Further, we also assemble a capercaillie mitochondrial genome to show that other ecological information can be retrieved in parallel with metagenomics. Our study demonstrates the value and potential of metagenomics for future diet research, and the challenges to consider when utilizing this technique, which is also relevant beyond the study of herbivore diet.

2 | METHODS

2.1 | *In silico* simulation

To test and validate taxonomic assignment steps for accurate plant identification in *Tetrao urogallus* faecal samples, we used the INSILICOSEQ software package (Gourlé et al., 2019) to carry out *in silico* simulation of metagenomic data sets that resembled reads expected from *T. urogallus* faecal samples. Three main metagenomic data sets that combined sequences from three known genome types (plants, bacteria and *T. urogallus*) with different plant species compositions were created (*in silico* test 1.1, *in silico* test 2.1 and *in silico* test 3.1) (Table 1). The genome types included were bacteria genomes, a *T. urogallus* mitogenome and plant genomes. To generate bacterial genomes, we used INSILICOSEQ to download 10 random complete bacterial genomes from NCBI (Table S1). The *T. urogallus* mitogenome used was generated and assembled in this study (Tables S2 and S3). Plastid and ribosomal sequences from the PhyloNorway project were used for generating the plant genome mix (sequences unpublished and not available in NCBI, permitted for use only in the *in silico* simulation, metadata available in Alsos et al., 2020). In the plant genome mix, we used sequences from known diet items of *T. urogallus* that are represented by both chloroplast and nuclear sequences in the PhyloNorway database, and also sequences from plant species that have not been recorded as known diet but are present in the study area and could be a potential source of diet. For plant species that lacked full chloroplast genomes in the PhyloNorway database, we used fragmented chloroplast sequences in the plant genome mix. For each data set, three subsets (repeats) with different proportions of genome types were generated (1.2, 1.3, 2.2, 2.3, 3.2, 3.3). Each data set was simulated twice to create two replicates (labelled "a" and "b"). We also generated a negative control which did not include any plant genomes, and a positive control which only consisted of plant genomes (Table S4). The inclusion of a simulated negative control was used to check for false positives, while the simulated positive control was used to check for misidentifications and false negatives.

The proportion of sequences assigned to plants, bacteria and *T. urogallus* were based on subsampling the sequenced metagenomic faecal samples. We subsampled 10,000 reads from each of our sequenced metagenomic samples, and queried the reads against the GenBank database using BLAST (subject_besthit). The BLAST output was exported into MEGAN to view the taxonomic contents (Huson et al., 2007; Huson & Weber, 2013). Between 75% and 98% of sequences belonged to bacterial sequences, 2%–20% of sequences belonged to capercaillie and around 1% of sequences belonged to plants. As environmental contamination of faecal samples may increase the proportion of bacteria found in faecal metagenomes, the highest proportion of sequences was assigned to bacteria (Hawlitschek et al., 2018). We varied the species composition of plants from *in silico* tests 1 to *in silico* tests 3 to see if increasing the plant diversity would have any effect on the number of detected plant taxa. For *in silico* test 2, we varied the proportion of both the *T. urogallus* and bacteria sequences while keeping the plant sequence proportion the same. This is to assess if changing the other DNA proportions present in the metagenome would have any effect on plants detected. Additionally, for *in silico* test set 3, we also increased the proportion of plant sequences from ~1% to ~7% to assess whether increasing the proportion of plant sequences would also affect the number of detected plant taxa. For each simulated data set, INSILICOSEQ (using a precomputed error model based on a HiSeq instrument) generated around 1 million fragmented Illumina PE reads (125 bp).

TABLE 1 Proportion of each genome type and number of plant species included in each *in silico* test

In silico tests	Plant species composition	Plant sequence proportion (%)	*Tetrao urogallus* sequence proportion (%)	Bacteria sequence proportion (%)
In silico test 1.1a,b	11 families, 16 genera, 26 species	3.2	2.9	93.9
In silico test 1.2a,b		1.2	18.6	80.2
In silico test 1.3a,b		1	1.5	97.5
In silico test 2.1a,b	22 families, 37 genera, 50 species	1.4	19.8	78.8
In silico test 2.2a,b		1.4	5.8	92.8
In silico test 2.3a,b		1.4	12.8	85.8
In silico test 3.1a,b	8 families, 11 genera, 16 species	6.5	2.8	90.7
In silico test 3.2a,b		7	1.5	91.5
In silico test 3.3a,b		9.8	14.7	75.5
In silico positive control	7 families, 7 genera, 7 species	100	0	0
In silico negative control	NA	0	5	95

2.2 | Testing taxonomic assignment of plants in simulated metagenomic data set

For the *in silico* tests, we assessed the accuracy of taxonomic identification parameters using five plant markers commonly used in plant identification: three chloroplast markers (*rbcL*, *matK* and *trnL-F*), and two nuclear markers (*ITS1* and *ITS2*). Following the methods used by Hunt et al. (2007) and adapted by Srivathsan et al. (2015), we generated reference databases from GenBank (downloaded on March 25, 2019) (Appendix S1) for the five plant markers as local reference databases for these markers were unavailable. Whole genomes were not utilized as a possible candidate for the reference database due to the lack of representation of our target species (20% species representation with RefSeq sequences as at October 25, 2020) (Table S5) and high false positive assignment rates (43%–100%) in our initial tests (Table S6). Additionally, localized organeller and ribosomal reference data for plants are currently only limited to the flora in areas such as Australia (Nevill et al., 2020), or not published as in the Alps and Norway (Alsos et al., 2020). Due to the better availability of standard barcode regions, these five plant markers were chosen to give the best results based on sequence availability. Species-level identification using these five plant markers was possible for all plant species included in the *in silico* simulation, with the exception of *Omalotheca norvegica*, *Urtica dioica dioica* and *Urtica dioica sondenii*. Genus-level identification was possible for all taxa (Table S5).

Following the reads filtering step employed by the only other two metagenomics herbivore diet studies to date (Srivathsan et al., 2015, 2016), we conducted MEGABLAST searches (word size =28, percentage identity =98%) for the forward and reverse reads generated in the simulated metagenomic data set against the generated plant marker reference databases. Any reads not receiving a hit were discarded. Our choice of using MEGABLAST to classify reads was to reduce the number of variables that require testing, as this classifier has already been validated for metagenomics herbivore diet studies (Srivathsan et al., 2015, 2016). Additionally, based on recent benchmarking studies, while this classifier is not the fastest, it performs equally well in terms of precision to other metagenomics classifiers (McIntyre et al., 2017; Ye et al., 2019).

For the taxonomic assignment steps, we first assign each read to the lowest taxonomic rank based on the lowest common ancestor (LCA) algorithm using READSIDENTIFIER (version 1.0) (Huson & Weber, 2013; Srivathsan et al., 2015), with a minimum of 63 bp (half the length of reads generated) or 85 bp (two-thirds the length of reads generated) reads overlap for the initial round of *in silico* tests with the three main data sets and its replicates (1.1, 2.1, 3.1) (details of LCA assignment are given in Appendix S1). We then repeated this taxonomic assignment step using only 85 bp reads overlap with the other simulated metagenomic data sets (1.2, 1.3, 2.2, 2.3, 3.2 and 3.3) as no errors were observed with this length (Table S7). Plant identifications from the *in silico* simulation outputs were compared to the list of plants used in each of the simulated metagenomic data sets to obtain the most optimal marker combinations for accurate taxonomic assignment (Table S8, details of choosing marker combination in Appendix S1). From the *in silico* tests, we derived the following criteria for accurate taxonomic assignment: (i) PE reads with at least 98% sequence identity matched against the reference database in MEGABLAST search, (ii) 85 bp overlap using READSIDENTIFIER, and (iii) taxonomic assignment for each read assigned by READSIDENTIFIER must be based on one of the predetermined optimal marker combinations (Table S8). If these three criteria were not met, reads were discarded.

2.3 | Sample collection, DNA extraction and sequencing

Eight faecal samples were collected from *T. urogallus* between September and November 2018 in Norway, including one from a captive male located at the Namsskogan Familiepark (Table S9). Each faecal sample consisted of two droppings deposited by one individual. Upon collection, faecal samples were placed in sterile airtight tubes filled with Merck silica gel (with indicator, granulate size 1–3 mm; Merck KGaA). Faecal samples collected immediately after defecation were labelled as "fresh," otherwise, they were labelled as "old." Faecal samples were stored at −20°C before DNA extraction with a Qiagen PowerFaecal DNA Isolation Kit. DNA extracts of each sample were split into two sets for metagenomics and metabarcoding. For metagenomics, 50 µl of each DNA extract and one extraction blank were fragmented (475 bp fragment size), built into Illumina libraries using the Blunt-End Single-Tube (BEST) protocol (Carøe et al., 2018; Mak et al., 2017), indexed and pooled for sequencing on two lanes of an Illumina HiSeq 4000 (150 bp PE). For metabarcoding, instead of using all five markers (*rbcL*, *matK*, *trnL-F*, *ITS1* and *ITS2*) utilized for the metagenomics taxonomic assignment step, we chose only the *trnL* P6 loop as a molecular marker, which is a shorter fragment of the *trnL-F* gene. This is due to practical reasons such as costs, time and primer suitability in amplifying degraded DNA. To amplify the *trnL* P6 loop from the DNA extracts of each sample (three PCR replicates per sample), including positive controls (*Cinchona officinalis* extract), extraction blanks and PCR blanks, we used the 5′ nucleotide tagged primer sets *trnL*-g and *trnL*-h for amplification (Binladen et al., 2007; Coissac, 2012; Taberlet et al., 2007). Amplicons were pooled and built into libraries using the TagSteady protocol (Carøe & Bohmann, 2020), before sequencing on the Illumina MiSeq V3 (150 bp PE). All sequencing was carried out at the National High-throughput DNA Sequencing Centre, University of Copenhagen. All details of the metagenomics and metabarcoding laboratory workflow can be found in Appendix S1.

2.4 | Bioinformatics analyses

2.4.1 | Metagenomics

Adapter sequences and low-quality reads were removed using TRIM GALORE version 0.5.0, with Phred score 30 (Krueger, 2012). We also

removed reads shorter than 100 bp using CUTADAPT version 1.11 (Martin, 2011). Quality checks were carried out using FASTQC version 0.11.7 (Andrews, 2010). To identify the composition of plants in *T. urogallus* diet from the faecal samples collected in our study, we used the criteria for accurate taxonomic assignment determined from the *in silico* simulation with minor modification during taxonomic assignment using READSIDENTIFIER. As the reads generated by Illumina HiSeq were longer than the simulated reads in the *in silico* data sets generated by INSILICOSEQ, we used the identification parameters of a minimum 100 bp overlap (two-thirds the length of reads generated). The combination of markers used for identification was based on the predetermined optimal marker combinations from the *in silico* simulation (Table S8). Additionally, we checked the extraction blank for any possible contamination, and removed any false positives (reads that came from contamination, sequencing artefacts or inaccurate database matches) (Alsos et al., 2018; Ficetola et al., 2015), by comparing the data to (i) the regional flora (species present in the Norwegian study sites as listed in the Global Biodiversity Information Facility [GBIF; https://www.gbif.no), (ii) a list of *T. uroagallus* diet recorded from field observations from previous studies (Table S10), and (iii) a list of diet items fed to the captive male *T. urogallus* (Table S11). False positives were removed if they fulfilled all three of the following conditions: plants not found in the region, not in the list of previously recorded diet, and not part of the diet fed to the captive *T. urogallus*. The removal of false positive reads was carried out at each hierarchical level, starting with the species level. Excluded reads were then reclassified at the genus level, and this was repeated at the family level if unclassified reads remained. The diet fed to the captive capercaillie was included to test if the methods employed would be able to retrieve its full diet. Further, to demonstrate that metagenomics can be used to simultaneously retrieve other ecological information, we also assembled a capercaillie mitochondrial genome (mitogenome) (Appendix S1).

2.4.2 | Metabarcoding

The OBITOOLS package was used for data processing (Boyer et al., 2016) (details in Appendix S1). The resulting sequences were identified using two different *trnL* P6 loop reference databases: the global EMBL reference database (r142), and a local, high-quality, reference database containing 2,445 sequences from arctic and boreal vascular plants, as well as bryophytes (Soininen et al., 2015; Sønstebø et al., 2010; Willerslev et al., 2014).

3 | RESULTS

3.1 | *In silico* simulation

For each of the simulated metagenomic data sets and the simulated positive control, BLAST searches against the plant marker databases yielded between 17 and 62 reads (0.003%–0.01%) assigned to the plant reference database (Table S12). Other than the positive control, inaccurate identifications were observed for all the *in silico* tests when there were hits to only one taxonomic marker. More errors were observed with hits to a single chloroplast marker than to any nuclear markers, and with the least stringent identification parameter of 63 bp read overlap (Table S13). After applying our derived marker combination for taxonomic identification to the lowest taxonomic level, all the plants included in the data set were detected at the family level while six out of seven genera and one species were detected in the positive control (Table S14). No differences in identification were observed between *in silico* replicates, no inaccurate identifications were made for the simulated positive control, and no identifications were made for the simulated negative control.

Reads were assigned to the family (31%) or genus (31%) level, and a small proportion of the reads could be assigned to the species level (11%) (Table S14). Using nuclear markers often resulted in more identifications and higher resolution than chloroplast markers (Table S15), while the use of both chloroplast and nuclear markers combined increased plant identifications at all taxonomic levels (ANOVA TYPE III sum of square test = $p < .05$; *post hoc* Tukey multiple pairwise comparison test chloroplast:both = $p < .05$, nuclear:both = $p = .16$, nuclear:chloroplast = $p < .005$; Figure 1a). When we increased the diversity of plant species used in the simulated data sets, the proportion of plants identified decreased (ANOVA TYPE III sum of square test = $p < .05$; *post hoc* Tukey multiple pairwise comparison test 7:50 species = $p < .05$, 16:50 species = $p < .05$, 26:50 species = $p < .05$, 16:7 species = $p = .67$, 26:7 species = $p < .05$, 26:16 species = $p = .11$; Figure 1b). Furthermore, some of the plants included in the simulated metagenomic data set were not detected in all of the nine *in silico* tests, including plants that could be part of the diet of *Tetrao urogallus* (Table S16).

3.2 | Diet identification

For the metagenomic data set, Illumina HiSeq sequencing generated between ~59 and ~178 million PE reads per faecal sample, while ~3 million PE reads were generated for the extraction blank (Table S9). To reconstruct *T. urogallus* diet from metagenomic data, BLAST searches against the plant marker databases yielded between one and 1,005 reads (<0.0001%–0.001%) per sample assigned to plants (mean =222, *SD* =325). For the extraction blank, 37 reads (0.001%) were assigned to plants (Table S17). The reads from the extraction blank only matched to the class Liliopsida, and were excluded from subsequent analysis. After filtering, 34% of the reads could be assigned to family (683 reads, mean =85, *SD* =148), 46% to genus (918 reads, mean =32, *SD* =82) and 11% to species level (230 reads, mean =20, *SD* =23). The remaining 9% of reads that could not be taxonomically identified to at least the family level were discarded (170 reads, mean =17, *SD* =15). Analysis of the extraction blank and sample N47 did not result in any identifications (Table S18). Plant identifications from all samples using chloroplast markers alone included plants from seven families, 10 genera and four

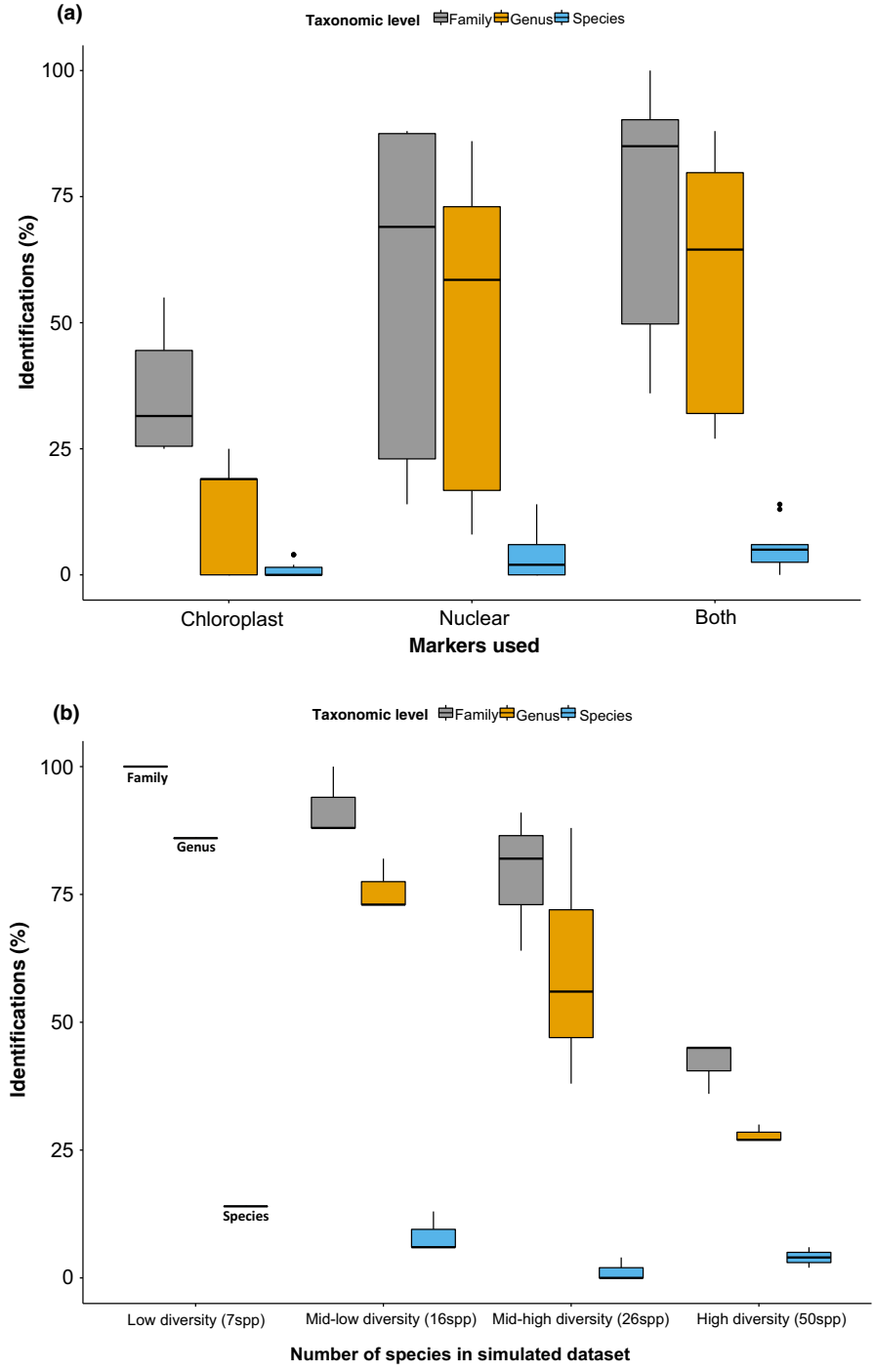


FIGURE 1 Identification of plants (%) at each taxonomic level (family, genus and species) for the *in silico* tests. (a) Percentage of taxa identified using different sets of markers (chloroplast, nuclear or both). (b) Percentage of taxa identified with an increasing diversity of plant species used in the simulated metagenomic data set. The *p*-value stated for both graphs is between each category on the *x*-axis, independent of specific taxonomic levels

species. Nuclear markers identified plants from 10 families, 16 genera and five species. The use of both marker types combined increased the numbers of identifications at all taxonomic levels (12 families, 20 genera and 10 species; Table S19).

For the metabarcoding data set, Illumina MiSeq sequencing generated between ~60,000 and ~78,000 PE reads for each of the three PCR replicates per sample. A total of 30 sequences were retained after OBITOOLS classification, varying from one to 14 sequences per sample. Of these, 24% (seven sequences) could be identified to the family level, 33% (10 sequences) to the genus level and 43% (13 sequences) to the species level (Tables S20 and S21). From these 30 sequences, a total of 25 plant taxa were identified when a combination of both local and global databases was used. Matches to the local reference database alone retrieved 88% (22 out of 25) of the plant taxa, at 40% species resolution, with no sequence misidentification. For matches to the global EMBL database, three sequences had no taxonomic identification resolved minimally to the family level. For the remaining sequences with taxonomic identifications, 11% of the sequences were misidentified (three out of 27 sequences), and 84% (21 out of 25) of the plant taxa were retrieved but at a reduced taxonomic resolution of 4% species resolution (Table S22).

From the combined metagenomics and metabarcoding results (Table S23), we retrieved a total of 28 plant taxa from seven wild capercaillies (two to eight plant taxa retrieved per individual, mean =5, *SD* =3) (Table S24). From the single captive capercaillie, 23 plant taxa were retrieved of which 14 taxa were detected only in the captive capercaillie and not in the wild capercaillies (Table S25). All diet items fed to the captive capercaillie were identified to at least the family level, as well as items not fed but found in its enclosure (Table S11). From both the captive and wild capercaillies, we identified several plant taxa from different genera that were previously unknown or unrecorded in the diet of the western capercaillies such as *Alnus*, *Astragalus*, *Athyrium*, *Avenella*, *Brassica*, *Delphinium* and *Poa*. Three of the potential new diet items (*Avenella*, *Poa* and *Delphinium*) were growing in the enclosures of the captive capercaillie, and in our data set, these were present only in the faecal samples obtained from the captive capercaillie. However, *Avenella* and *Poa* are common plants found all over Scandinavia. Around 87% of the families, 56% of the genera and 77% of the species identified are known diet items.

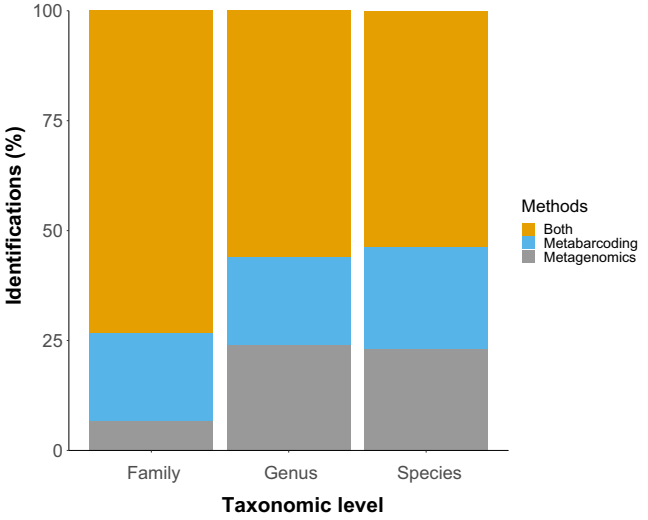


FIGURE 2 Proportion of distinct identifications at the three taxonomic levels (family, genus and species) across all samples using metagenomics and metabarcoding

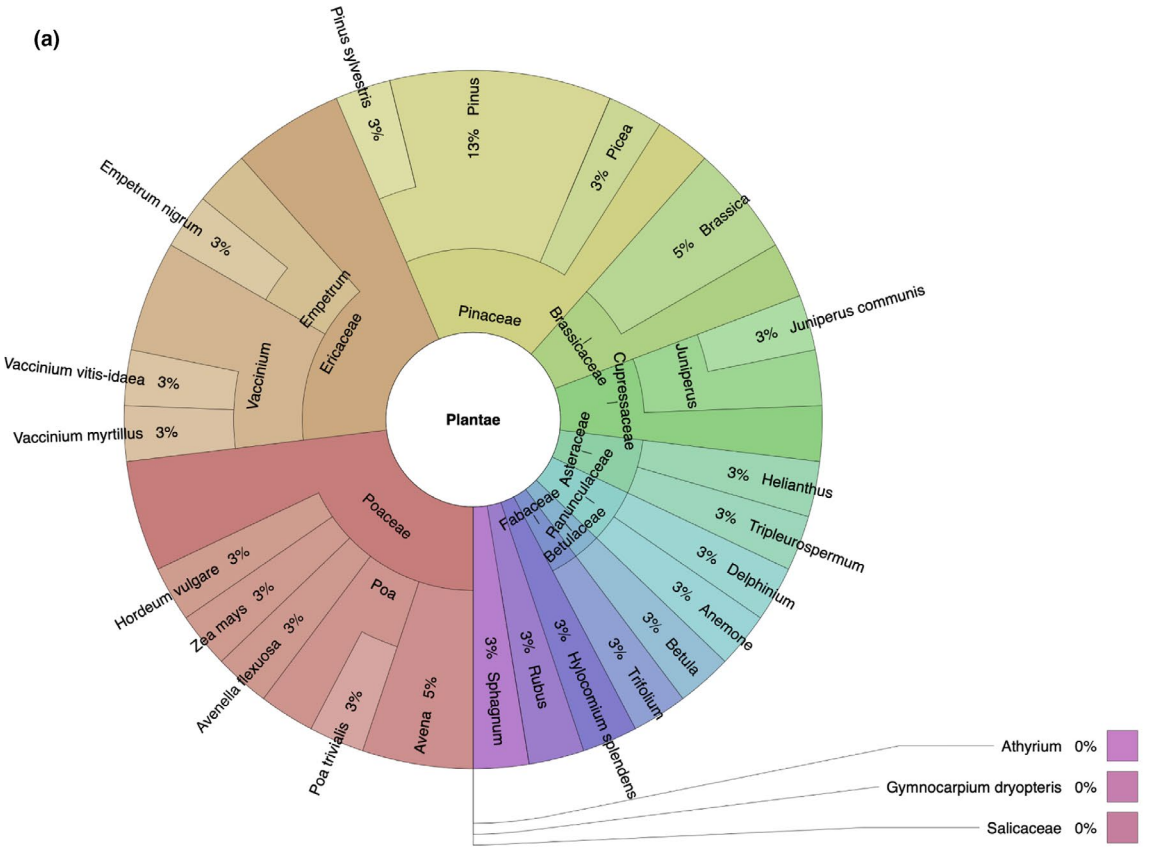
3.3 | Metagenomics vs. metabarcoding

At all taxonomic levels (family, genus and species), there was no significant difference (two-sample *t* test, $p > 0.05$) in the total number of identified plant taxa between the metagenomic and metabarcoding data sets (Table S26; database comparison in Tables S27 and S28). Metagenomics identified 42 plant taxa while metabarcoding identified 43 plant taxa. For metagenomics, 12 of the 42 taxa were at the family level (0–9 families per sample, mean 2.4 ± SE 0.9, 28% resolution), 20 of the 42 taxa were at the genus level (~0 to 14 genera per sample, mean 3.3 ± SE 1.5, 48% resolution), and 10 of the 42 taxa were at the species level (~0–7 species per sample, mean 1.3 ± SE 0.8, 24% resolution). For metabarcoding, 14 out of 43 taxa were at the family level (~1–7 families per sample, mean 3.3 ± SE 0.6, 33% resolution), 19 out of 43 were at genus level (~1–8 genera per sample, mean 3.4 ± SE 0.7, 44% resolution), and 10 out of 43 were at the species level (~0–5 species per sample, mean 2 ± SE 0.5, 23% resolution). At the family level, the congruence of plants identified between the two HTS methods was 73%, at the genus level it decreased to 56%, and for the species level it was 54% (Figure 2). Hylocomiaceae was identified only with metagenomics, whereas Athriaceae, Cystopteridaceae and Salicaceae were identified only with metabarcoding (Table S23). At the genus level, known diet items such as *Hylocomium* (six reads) and *Rubus* (one read) were only identified using metagenomics, whereas *Ranunculus* was only identified using metabarcoding. The three main families of *T. urogallus* diet items overlapped between the two methods (Figure 3). With metagenomics, the most frequently occurring plant families across all samples were Poaceae (23%), followed by Ericaceae (21%) and Pinaceae (18%). For metabarcoding, the most frequently occurring families were Ericaceae (30%), followed by Poaceae (15%) and Pinaceae (13%).

More taxa were identified at all taxonomic levels in fresh compared to old faecal samples from the results of both HTS methods (for fresh faecal samples: 21 families, 32 genera and 16 species; whereas for old faecal samples: 13 families, 11 genera and six species, two-way ANOVA $p < 0.05$, Figure 4). Even though metagenomics identified more plant taxa in fresh faecal samples, when compared with metabarcoding (metagenomics 44 taxa, metabarcoding 32 taxa), the difference was not significant (two-way ANOVA $p > .05$). By contrast, for old faecal samples, metabarcoding performed significantly better than metagenomics (metabarcoding 37 taxa, metagenomics 11 taxa, two-way ANOVA $p < .05$).

4 | DISCUSSION

The use of metagenomics to reconstruct the diet of herbivores is a relatively new tool. To date, testing for suitable bioinformatic parameters to use in taxonomic assignments has been based on comparisons with both metabarcoding data and field observations (Srivathsan et al., 2015, 2016). However, field observations are extremely time-consuming, as demonstrated by Srivathsan et al. (2016), and field observations can also be challenging in habitats that are hard to access such as the forest canopy. Instead of relying on field observations, our study demonstrates how simulation of metagenomic data *in silico* can be used in testing and validating steps used for taxonomic assignment in metagenomics herbivore diet studies. Our study also highlights issues to consider when choosing between the metagenomic or metabarcoding approach in future herbivore diet studies. Additionally, to our knowledge, this is the first metagenomic diet analysis conducted on birds. As differences between the mammalian and avian gastrointestinal tract may result in differences in faecal metagenomes, our study has important implications for


(a)
Plantae
Pinaceae
Pinus
Pinus sylvestris 3%
13% Pinus
3% Picea
Brassicaceae
5% Brassica
Cupressaceae
Juniperus
3% Juniperus communis
Ericaceae
Empetrum
Empetrum nigrum 3%
Vaccinium
Vaccinium vitis-idaea 3%
Vaccinium myrtillus 3%
Poaceae
Poa
Hordeum vulgare 3%
Zea mays 3%
Avenella flexuosa 3%
Poa trivialis 3%
Avena 5%
Asteraceae
3% Helianthus
Ranunculaceae
3% Tripleurospermum
Fabaceae
3% Delphinium
Betulaceae
3% Anemone
3% Betula
3% Trifolium
3% Sphagnum
3% Rubus
3% Hylocomium splendens
Athyrium 0%
Gymnocarpium dryopteris 0%
Salicaceae 0%

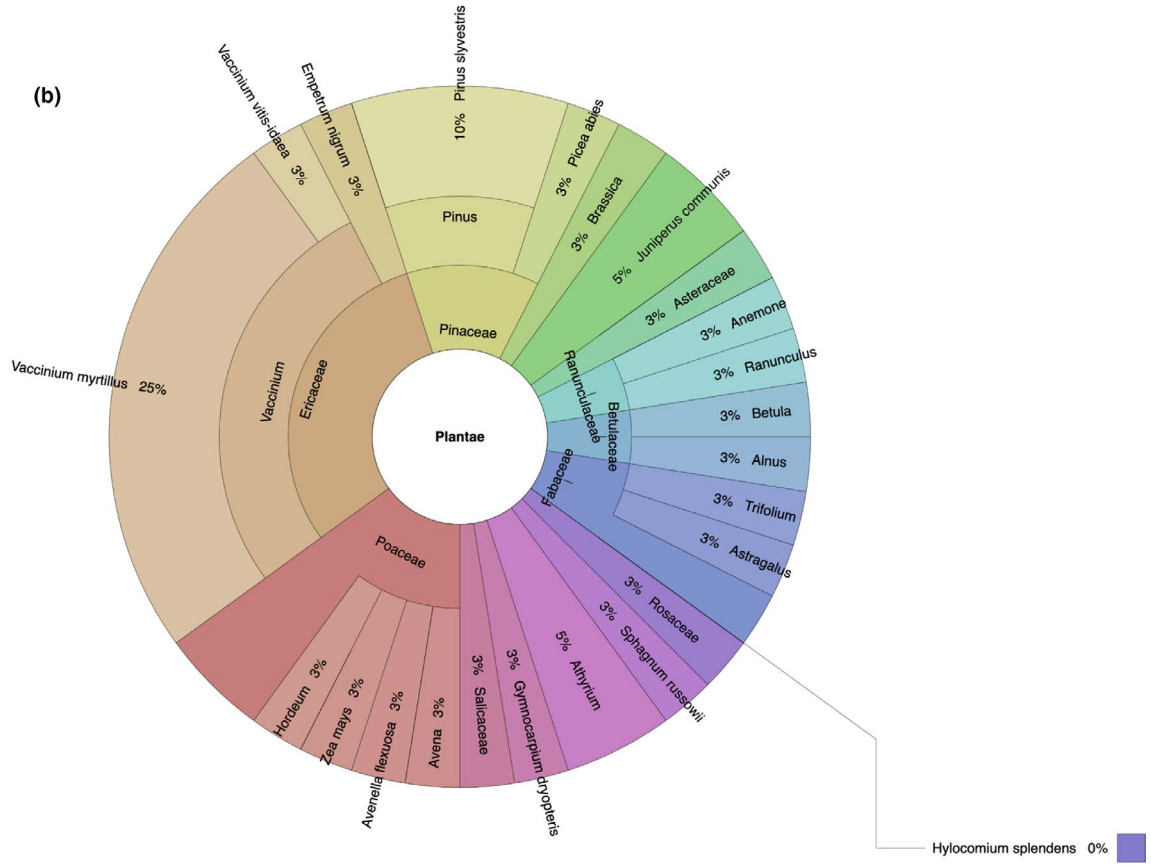

(b)
Plantae
Pinaceae
Pinus
Pinus sylvestris
10% Pinus sylvestris
Picea abies
3% Picea abies
Brassica
3% Brassica
Juniperus communis
5% Juniperus communis
Ericaceae
Vaccinium
Empetrum nigrum 3%
Vaccinium vitis-idaea 3%
Vaccinium myrtillus 25%
Poaceae
Hordeum 3%
Zea mays 3%
Avenella flexuosa 3%
Avena 3%
Asteraceae
3% Asteraceae
Anemone
3% Anemone
Ranunculaceae
Ranunculus
3% Ranunculus
Betulaceae
3% Betula
Alnus
3% Alnus
Fabaceae
3% Trifolium
3% Astragalus
Rosaceae
3% Rosaceae
3% Sphagnum russowii
5% Athyrium
3% Salicaceae
3% Gymnocarpium dryopteris
Hylocomium splendens 0%

FIGURE 3 Taxonomic contents of plants found in all *Tetrao urogallus* faecal samples collected in Norway, with percentage value assigned according to the number of times it is identified across all samples. (a) Taxonomic identity of plants using metagenomic reads. (b) Taxonomic identity of plants using unique metabarcoding sequences. Plots were made with KRONA (Ondov et al., 2011)

understanding the proportion of plant sequences available in the faecal metagenome of birds.

4.1 | Utility of *in silico* simulation of metagenomic data

Here, we show that by simulating realistic data sets that resemble sequences derived from shotgun sequencing of faecal samples of *Tetrao urogallus*, we were able to test and validate bioinformatic parameters used in taxonomic classification. From the *in silico* simulation outputs, one can obtain an understanding of the sensitivity of the bioinformatic parameters used, and which potential diet items may not be picked up even though they might be present in the samples. The three repeats of the *in silico* tests with different numbers of reads generated per species showed broad congruence in the identified taxa, and no differences were observed between replicates. This reproducibility and replicability give confidence in the bioinformatic approach used to employ metagenomics for diet studies.

In our simulated metagenomic data set, common capercaillie diet items including sedges such as *Carex* sp. and heathers such as *Calluna* sp. were not detected by utilizing chloroplast markers alone. They were also not identified in some of the simulated metagenomic data sets, even when using a combination of both chloroplast and nuclear markers, which in principle should increase species discrimination and taxon identification (pending the completeness of the reference database used). The two undetected genera were also subsequently not detected in our real metagenomic data set. Given that the number of plants identified decreased with increasing species diversity, only a small proportion of reads could be assigned to the species level, which led to many undetected species. This could be due to a number of issues, including a reduced proportion of sequences per species when the number of simulated species increased, reads that matched to only one marker and thus failed in the multimarker criterion we applied, and our use of a global reference database where standard barcodes have low discriminatory power between sister species (Hollingsworth et al., 2016). If a well-represented and complete regional reference database had been available, we expect higher taxonomic resolution could be obtained. The generation of larger amounts of metagenomic data sets (with more repeats and replicates) through *in silico* simulation could also potentially be used in future studies to further explore how variations in species composition may affect the recovery of individual plant species. The utility of this approach would open up avenues on how to better curate bioinformatics pipelines best suited to the data set. Additionally, other types of metagenomics classifier other than the MEGABLAST approach which we have used here can similarly be validated through *in silico* simulation. These alternative classifiers such as KRAKEN2 (Wood et al., 2019) and CENTRIFUGE (Kim et al., 2016) can speed up the identification of metagenomic data sets and increase computational efficiency (Ye et al., 2019). Hence, future metagenomic diet studies could explore the use of alternative metagenomics classifiers, depending on research questions and data set sizes. Overall, despite our small sample set, simulating metagenomic reads *in silico* provides a valuable starting point for applying this approach to real data sets.

4.2 | Diet identification

In our metagenomic data set, there were almost three times more reads assigned to chloroplast markers as compared to nuclear markers. Some families, such as Pinaceae and Sphagnaceae, were identified by only chloroplast markers, while some families, such as Asteraceae, Betulaceae, Fabacae, Hylocomiaceae and Rosaceae, were only identified by nuclear markers. This marker bias towards the identification of certain family taxa was overcome by using a combination of both marker types. More reads were also assigned to plants when combining both marker types (1,352 reads assigned for chloroplast markers, 463 reads assigned for nuclear markers, 1,831 reads assigned for both markers). Due to the larger combined reference length of chloroplast markers, more reads were assigned to them as compared to nuclear markers. The difference in read numbers may also reflect which parts of the plant are preferentially consumed, since ingested leaves are richer in chloroplast DNA than other parts of the plants such as roots, fruits or seeds (Valentini et al., 2009). Chloroplast sequences are also generally present in higher copy numbers than nuclear sequences (Tonti-Filippini et al., 2017). Additionally, there may be a difference in digestive rates between different parts of the plant which could be affected by plant age (Pompanon et al., 2012), and the possible differences in rates of DNA degradation between chloroplast and nuclear genomes (Xin et al., 2018).

As more markers used for taxonomic identification in our study resulted in more informative reads, this suggests that future studies should consider using both chloroplast and nuclear markers to gain a more complete overview of the subject's diet even if feeding preference is known. On the other hand, should whole organelle plant reference genomes become available, it has been suggested that the use of whole organelle genomes might result in a 50× increase of available data for taxonomic identification (Srivathsan et al., 2015). This would overcome the need to choose combinations of markers and greatly ease data generation and downstream analyses because only chloroplast sequences would be needed. The development and availability of whole organelle genome databases such as the Flora of China (Li et al., 2019),

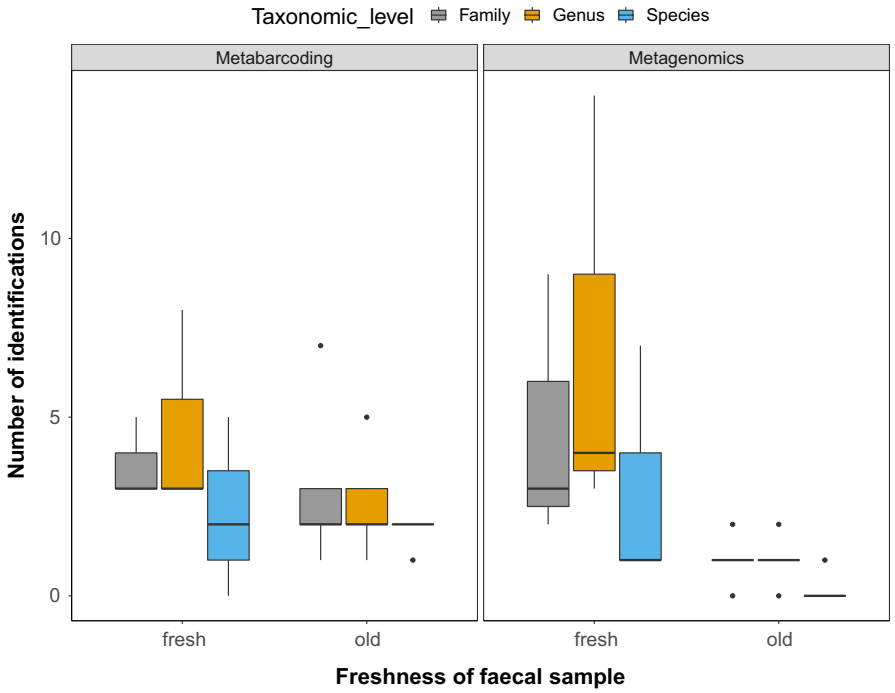


FIGURE 4 Number of plants identified for fresh and old faecal samples at each taxonomic level (family, genus and species) using metagenomics or metabarcoding

Australia (Nevill et al., 2020), PhyloAlps/PhyloNorway (Alsos et al., 2020), and DNAmark (www.dnamark.ku.dk) will greatly improve both the detection and resolution of metagenomic data sets for future studies.

From both the metagenomic and metabarcoding approaches, we identified a few potential new diet items that have not been recorded in wild capercaillies or fed to captive capercaillies. The new diet items comprised two plant families and eight genera. Four of the genera (*Astragalus*, *Avenella*, *Delphinium* and *Poa*) were found only in faecal samples obtained from the captive capercaillie. This captive capercaillie fed on a wider range of plant taxa as compared to wild capercaillies, including 14 plant taxa that were not detected in wild capercaillies, suggesting that in captivity, capercaillies may develop a wider diet than they would have had in the wild. This is also consistent with studies that show grouse specialization in food resources is largely dependent on its habitat type, and only a few common plant species are important diet items across its geographical range (Sedinger, 1997). The HTS methods applied in the present study allowed us to identify previously unknown diet items for wild capercaillies from a sample of just seven individuals. We were thus able to recover a greater breadth in and better species resolution of the capercaillie diet items than would have been possible by using field observations and macroscopic identification of plant remains in faecal samples. However, our methods did not detect some common diet items that are usually detected from field observations or microhistological analysis, such as *Calluna* sp., *Carex* sp., *Melampyrum* sp., *Salix* sp., *Vaccinium uliginosum* and *Vaccinium oxycoccos* (Picozzi et al., 1996; Picozzi et al., 1999; Wegge & Kastdalen, 2008), even though they are found in the sampling regions and commonly found in metabarcoding data (Alsos et al., 2018) (https://www.biodiversity.no/). This may be due to the limited number of samples analysed. As molecular identification is more time-efficient than field or histological studies, future fieldwork may be able to dispense with feeding observations and rather focus on the collection of large numbers of faecal samples.

4.3 | Metagenomics vs. metabarcoding

The results showed little difference between the metagenomics and metabarcoding approaches. Theoretically, metagenomics is expected to obtain better species resolution than metabarcoding, due to its ability to utilize more and longer markers for the matching of reads (Srivathsan et al., 2016). In reality, this is strongly dependent on the completeness of the reference database used, as well as the taxonomic resolution of the markers utilized.

A higher resolution could have been achieved from the metagenomic data sets if a comprehensive local database had been available. With a minibarcode commonly used in metabarcoding studies such as *trnL*, utilizing a local database is advantageous as it mitigates the issue of low species-level resolution by matching reads to plant sequences found only in the study area. For example, the *trnL* marker has a taxonomic resolution of 33% to the species level on a circumarctic scale (Sønstebø et al., 2010), but within a catchment area, this may increase to 77%–93% (Alsos et al., 2018). Thus, by using a local reference database with biogeographical criteria to narrow down species identification as we did here for the metabarcoding data set, a higher taxonomic resolution may be obtained. However, in many biodiverse regions, comprehensive species reference databases and inventories for local flora are not readily available. This would necessitate the use of global reference databases without biogeographical criteria to resolve sequence identity. Additionally, in the presence of invasive plant species or changes in the range of the studied animal, utilizing a local reference database for diet reconstruction may

potentially result in missing important findings. Thus, the choice of reference database used (depending on availability) would also be directly influenced by the research objective of the study.

Based on the comparison between the local and global reference databases used in our study for metabarcoding, utilizing global reference databases can result in reduced species resolution. However, the resolution of metabarcoding data can be improved by multiplex PCRs with multiple short markers, but there is currently no specific combination of markers that allows for universal species identification in plants (Li et al., 2015). When there is incomplete species representation in the reference database used, which currently is the case in most environmental DNA studies, both methods will always be biased by the available reference database. However, metabarcoding is additionally biased by preferential amplification due to factors such as primer binding sites and sequence length (Berry et al., 2011; Deagle et al., 2009; Pompanon et al., 2012). For example, even though the plant taxa identified using the two methods in our study are broadly congruent, particularly at the family level with the three most frequently occurring plant families overlapping (Ericaceae, Poaceae, and Pinaceae), it is of note that some taxa were uniquely detected by using only either metabarcoding (Athyriaceae, Cystopteridaceae and Salicaceae) or metagenomics (Hylocomiaceae). This discrepancy may be due to the marker amplification bias for metabarcoding, which is not as important in metagenomics as the method does not involve a marker amplification step prior to sequencing (Paula et al., 2016; Srivathsan et al., 2015).

We also observed that, with fresh faecal samples, metagenomics performed best, but with old faecal samples, metabarcoding outperformed metagenomics. This is shown for instance in sample N47, where metagenomics did not yield any identification, while metabarcoding was able to pick up a few diet items. Even though PCR amplification used in metabarcoding would enrich for the degraded DNA of interest, we had expected metagenomics to perform better in older samples because short fragments that do not include primer binding sites can be analysed, and DNA damage has been shown to inhibit the PCR extension step used in metabarcoding (Deagle et al., 2006). Additionally, metagenomics has been successfully used for reconstructing plant communities from ancient lake sediments (Parducci et al., 2019; Pedersen et al., 2016). However, based on our current metagenomic setup, the identification workflow would have to be adjusted for shorter reads along with the expectation that fewer reads can be identified and that the obtained identification may be limited to higher taxonomic levels. Future metagenomic studies could also consider library insert sizes of 300 bp or shorter, to test how fragmentation might affect the obtained diet composition. There should be a potential for obtaining good results for old faecal samples with metagenomics, possibly by using less stringent length cut-off identification parameters during the taxonomic assignment step. This can be achieved by utilizing a more complete reference database where a more relaxed length cut-off could be set without increasing the risk of false positives. The availability of fresh faecal samples is therefore one of the limiting factors of metagenomics when used with a less comprehensive reference database. Sequencing old faecal samples may not yield any diet information, which would be a waste of time and resources. Another possible explanation for why metagenomics performed worse in old faecal samples could be due to the presence of free DNAses in avian faecal samples, which are a major cause of DNA degradation (Regnaut et al., 2006). This reduces the proportion of longer diet DNA fragments that could be more informative for taxonomic assignment. Hence, currently, metabarcoding is better suited to situations where the age of the faecal samples is unknown, when no fresh faecal samples are available, particularly in instances where a comprehensive reference database is unavailable for metagenomics, or when working with avian samples due to DNA degradation. This consideration and the knowledge of marker-associated amplification bias is something which users need to be aware of.

Based on sequencing costs alone, metabarcoding is able to sequence at least 10× more samples than metagenomics. With more samples sequenced for the same costs, metabarcoding can provide a broader overall diet profile, which could be used for ecological inferences such as habitat and resource partitioning. However, if additional experiments are required to retrieve other information (e.g., insect diet as is the case for capercaillie chicks), this would increase the cost for metabarcoding. By contrast, no additional sequencing would be required with a metagenomics approach. Metagenomics can provide additional ecological information which can be retrieved simultaneously. This includes animal diet analysis, gut parasites, population genetics and microbiome (Hicks et al., 2018; Srivathsan et al., 2015, 2016, 2019). Thus, metagenomics provides researchers with a multidimensional ecological characterization of taxonomic groups. Therefore, decisions on which method to use may also depend on the funding available, and the nature of the research question to be addressed. Ideally, when no information is available on the local vegetation or diet of the studied species, metagenomics should be the preferred approach as it is less biased even when used with a global reference database. Despite the advantages of using metagenomics to re-create the diet profile of herbivores, diet studies may still continue to utilize metabarcoding as it has the added advantages of lower costs and sequencing effort, lower demands on computational power, well-validated bioinformatics pipelines, better availability of reference databases and well-documented bias as compared to metagenomics.

4.4 | Future outlook

By providing a reliable computational environment to test and validate bioinformatic parameters, *in silico* simulation of metagenomic data reduces the need to carry out field observations for herbivore diet reconstruction. As the number of diet studies that use metagenomics on a variety of species with different dietary profiles increases, dependency on comparison with metabarcoding results should decrease. Metagenomics has the potential to become the go-to technique for diet reconstruction in the future, particularly when a localized reference database is not available for minibarcodes used in metabarcoding or additional information is needed, for example

on parasites, host genetics and microbiome. However, there are several considerations which may limit the use of metagenomics such as (i) cost, (ii) bioinformatic challenges, (iii) type of research question, (iv) availability of fresh faecal samples and, most importantly, (v) the completeness of the reference database used.

ACKNOWLEDGEMENTS

We thank NatureMetrics, especially Dr Kat Bruce, for collaborating on this paper. Thanks to Professor Torbjørn Ekrem for valuable inputs and organizing fieldwork. We would like to thank the Danish National High-throughput Sequencing Centre for generating the Illumina data. We are grateful to the staff at Namsskogan Familiepark, Norway, for access to the captive male capercaillie. We would like to thank Arne Flor, Bjørn Morten Baardvik, Bjørn-Roar Hagen, Pål Fossland Moa, Per Gustav Thingstad and Stein Nilsen for assisting with sample collection or providing information on study sites. Lastly, we thank Dr Christopher Barnes for providing the *Cinchona* DNA extracts. We also thank the reviewers for their constructive comments. This research is part of the H2020 MSCA-ITN-ETN Plant.ID network, and has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement No. 765000.

AUTHOR CONTRIBUTIONS

P.Y.S.C., K.B., S.B. and I.G.A. designed and conceived the research; K.B., S.B. and I.G.A. supervised all parts of the research; P.Y.S.C. collected samples; P.Y.S.C. performed laboratory work; P.Y.S.C., A.C.P. and Y.L. analysed data; A.C.P., Y.L. and I.G.A. validated data, I.G.A. provided data for *in silico* tests; P.Y.S.C. wrote the manuscript with inputs from all coauthors. All authors contributed to the manuscript and approved the final version.

DATA AVAILABILITY STATEMENT

The assembled western capercaillie metagenome is available at GenBank (accession: MT787324). The plant genomes used for the in silico simulation are part of the PhyloNorway research project and the metadata are available in Alsos et al. (2020). These genomes are deposited in the European Nucleotide Archive (ENA) under the study AC: ERP127839: bioproject id PRJEB43865. The genomic sequences have been submitted for publication in another study, and will be released upon its publication. All other *in silico* simulation, metagenomics and metabarcoding data, as well as scripts used for this study have been deposited in the Dryad Digital Repository, Chua et al. (2021) https://doi.org/10.5061/dryad.hqbzkh1d0, and will be released upon publication of this paper.

ORCID

Physilia Y. S. Chua https://orcid.org/0000-0001-7229-4480
Alex Crampton-Platt https://orcid.org/0000-0002-8096-615X
Youri Lammers https://orcid.org/0000-0003-0952-2668
Inger G. Alsos https://orcid.org/0000-0002-8610-1085
Sanne Boessenkool https://orcid.org/0000-0001-8033-1165
Kristine Bohmann https://orcid.org/0000-0001-7907-064X

REFERENCES

Ait Baamrane, M. A., Shehzad, W., Ouhammou, A., Abbad, A., Naimi, M., Coissac, E., Taberlet, P., & Znari, M. (2012). Assessment of the food habits of the Moroccan dorcas gazelle in M'Sabih Talaa, west central Morocco, using the trnL approach. *PLoS ONE*, *7*(4), e35643. https://doi.org/10.1371/journal.pone.0035643

Alsos, I. G., Lammers, Y., Yoccoz, N. G., Jørgensen, T., Sjögren, P., Gielly, L., & Edwards, M. E. (2018). Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. *PLoS ONE*, *13*(4), 1–23. https://doi.org/10.1371/journal.pone.0195403

Alsos, I. G., Lavergne, S., Merkel, M. K. F., Boleda, M., Lammers, Y., Alberti, A., Pouchon, C., Denoeud, F., Pitelkova, I., Puşcaş, M., Roquet, C., Hurdu, B.-I., Thuiller, W., Zimmermann, N. E., Hollingsworth, P. M., & Coissac, E. (2020). The treasure vault can be opened: Large-scale genome skimming works well using herbarium and silica gel dried material. *Plants*, *9*(4), 432. https://doi.org/10.3390/plants9040432

Andrews, S. (2010). *FastQC: A Quality control tool for high throughput sequence data*. Retrieved February 14, 2020, from Babraham Bioinfo website: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Berry, D., Mahfoudh, K. B., Wagner, M., & Loy, A. (2011). Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environmental Microbiology*, *77*(21), 7846–7849. https://doi.org/10.1128/AEM.05220-11

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, *2*(2), 1–9. https://doi.org/10.1371/journal.pone.0000197

Borchtchevski, V. (2009). The May diet of Capercaillie (Tetrao urogallus) in an extensively logged area of NW Russia. *Ornis Fennica*, *86*(1), 18–29.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Carøe, C., & Bohmann, K. (2020). Tagsteady: A metabarcoding library preparation protocol to avoid false assignment of sequences to samples. *Molecular Ecology Resources*, *20*(6), 1620–1631. https://doi.org/10.1111/1755-0998.13227

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., Wales, N., Sicheritz-Ponten, T., & Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution*, *9*(2), 410–419. https://doi.org/10.1111/2041-210X.12871

Coissac, E. (2012). OligoTag: A program for designing sets of tags for next-generation sequencing of multiplexed samples. *Methods in Molecular Biology*, *888*, 13–31. https://doi.org/10.1007/978-1-61779-870-2_2

Danell, K., Bergström, R., Duncan, P., & Pastor, J. (2006). *Large herbivore Ecology, Ecosystem Dynamics and Conservation*. Cambridge University Press. https://doi.org/10.5860/choice.44-2102

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Molecular Ecology Resources*, *14*(2), 306–323. https://doi.org/10.1111/1755-0998.12188

Deagle, B. E., Eveson, J. P., & Jarman, S. N. (2006). Quantification of damage in DNA recovered from highly degraded samples - A case study on DNA in faeces. *Frontiers in Zoology*, *3*, 1–10. https://doi.org/10.1186/1742-9994-3-11

Deagle, B. E., Kirkwood, R., & Jarman, S. N. (2009). Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, *18*(9), 2022–2038. https://doi.org/10.1111/j.1365-294X.2009.04158.x

Duffy, J. E., Cardinale, B. J., France, K. E., McIntyre, P. B., Thébault, E., & Loreau, M. (2007). The functional role of biodiversity in ecosystems: Incorporating trophic complexity. *Ecology Letters*, *10*(6), 522–538. https://doi.org/10.1111/j.1461-0248.2007.01037.x

Escalona, M., Rocha, S., & Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature Reviews Genetics*, *17*, 459–469. https://doi.org/10.1038/nrg.2016.57

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G., & Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*(3), 543–556. https://doi.org/10.1111/1755-0998.12338

Ford, C. S., Ayres, K. L., Toomey, N., Haider, N., Van alphen stahl, J., Kelly, L. J., Wikström, N., Hollingsworth, P. M., Duff, R. J., Hoot, S. B., Cowan, R. S., Chase, M. W., & Wilkinson, M. J. (2009). Selection of candidate coding DNA barcoding regions for use on land plants. *Botanical Journal of the Linnean Society*, *159*(1), 1–11. https://doi.org/10.1111/j.1095-8339.2008.00938.x

Gayot, M., Henry, O., Dubost, G., & Sabatier, D. (2004). Comparative diet of the two forest cervids of the genus Mazama in French Guiana. *Journal of Tropical Ecology*, *20*(1), 31–43. https://doi.org/10.1017/S0266467404006157

González, M. A., Olea, P. P., Mateo-tomás, P., García-tejero, S., De frutos, Á., Robles, L., Purroy, F. J., & Ena, V. (2012). Habitat selection and diet of Western Capercaillie Tetrao urogallus in an atypical biogeographical region. *Ibis*, *154*(2), 260–272. https://doi.org/10.1111/j.1474-919X.2012.01217.x

Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., & Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, *35*(3), 521–522. https://doi.org/10.1093/bioinformatics/bty630

Greve Alsos, I., Elvebakk, A., & Wing Gabrielsen, G. (1998). Vegetation exploitation by barnacle geese Branta leucopsis during incubation on Svalbard. *Polar Research*, *17*(1), 1–14. https://doi.org/10.3402/polar.v17i1.6603

Haiminen, N., Edlund, S., Chambliss, D., Kunitomi, M., Weimer, B. C., Ganesan, B., Baker, R., Markwell, P., Davis, M., Huang, B. C., Kong, N., Prill, R. J., Marlowe, C. H., Quintanar, A., Pierre, S., Dubois, G., Kaufman, J. H., Parida, L., & Beck, K. L. (2019). Food authentication from shotgun sequencing reads with an application on high protein powders. *Npj Science of Food*, *3*(1), 1–11. https://doi.org/10.1038/s41538-019-0056-6

Hawlitschek, O., Fernández-González, A., Balmori-de la Puente, A., & Castresana, J. (2018). A pipeline for metabarcoding and diet analysis from fecal samples developed for a small semi-aquatic mammal. *PLoS ONE*, *13*(8), e0201763. https://doi.org/10.1371/journal.pone.0201763

Hibert, F., Taberlet, P., Chave, J., Scotti-Saintagne, C., Sabatier, D., & Richard-Hansen, C. (2013). Unveiling the diet of elusive rainforest herbivores in next generation sequencing era? The tapir as a case study. *PLoS ONE*, *8*(4), e60799. https://doi.org/10.1371/journal.pone.0060799

Hicks, A. L., Lee, K. J., Couto-Rodriguez, M., Patel, J., Sinha, R., Guo, C., Olson, S. H., Seimon, A., Seimon, T. A., Ondzie, A. U., Karesh, W. B., Reed, P., Cameron, K. N., Lipkin, W. I., & Williams, B. L. (2018). Gut microbiomes of wild great apes fluctuate seasonally in response to diet. *Nature Communications*, *9*(1), 1–18. https://doi.org/10.1038/s41467-018-04204-w

Hollingsworth, P. M. (2011). Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 19451–19452. https://doi.org/10.1073/pnas.1116812108

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE*, *6*(5), e19254. https://doi.org/10.1371/journal.pone.0019254

Hollingsworth, P. M., Li, D. Z., Van Der Bank, M., & Twyford, A. D. (2016). Telling plant species apart with DNA: From barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150338. https://doi.org/10.1098/rstb.2015.0338

Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., John, O. S., Wild, R., Hammond, P. M., Ahrens, D., Balke, M., Caterino, M. S., Gomez-Zurita, J., Ribera, I., Barraclough, T. G., Bocakova, M., Bocak, L., & Vogler, A. P. (2007). A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science*, *318*(5858), 1913–1916. https://doi.org/10.1126/science.1146954

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Huson, D. H., & Weber, N. (2013). Microbial community analysis using MEGAN. *Methods in Enzymology*, *531*, 465–485. https://doi.org/10.1016/B978-0-12-407863-5.00021-6

Iversen, M., Aars, J., Haug, T., Alsos, I. G., Lydersen, C., Bachmann, L., & Kovacs, K. M. (2013). The diet of polar bears (Ursus maritimus) from Svalbard, Norway, inferred from scat analysis. *Polar Biology*, *36*(4), 561–571. https://doi.org/10.1007/s00300-012-1284-2

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. https://doi.org/10.1101/gr.210641.116

Kress, W. J., & Erickson, D. L. (2007). A Two-locus global DNA barcode for land plants: The coding rbcL gene complements the non-coding trnH-psbA spacer region. *PLoS ONE*, *2*(6), e508. https://doi.org/10.1371/journal.pone.0000508

Krueger, F. (2012). *Trim Galore*. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/trimgalore/

Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T. G., & Savolainen, V. (2008). DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(8), 2923–2928. https://doi.org/10.1073/pnas.0709936105

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., Gitzendanner, M. A., Fritsch, P. W., Cai, J., Luo, Y., Wang, H., van der Bank, M., Zhang, S.-D., Wang, Q.-F., Wang, J., Zhang, Z.-R., Fu, C.-N., Yang, J., Hollingsworth, P. M., … Li, D.-Z. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants*, *5*(5), 461–470. https://doi.org/10.1038/s41477-019-0421-0

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: from gene to genome. *Biological Reviews of the Cambridge Philosophical Society*, *90*, 157–166. https://doi.org/10.1111/brv.12104

Mak, S. S. T., Gopalakrishnan, S., Carøe, C., Geng, C., Liu, S., Sinding, M.-H., Kuderna, L. F. K., Zhang, W., Fu, S., Vieira, F. G., Germonpré, M., Bocherens, H., Fedorov, S., Petersen, B., Sicheritz-Pontén, T., Marques-Bonet, T., Zhang, G., Jiang, H., & Gilbert, M. T. P. (2017). Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *GigaScience*, *6*(8), 1–13. https://doi.org/10.1093/gigascience/gix049

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., Minot, S. S., Danko, D., Foox, J., Ahsanuddin, S., Tighe, S., Hasan, N. A., Subramanian, P., Moffat, K., Levy, S., Lonardi, S., Greenfield, N., Colwell, R. R., Rosen, G. L., & Mason, C. E. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, *18*(1), 182. https://doi.org/10.1186/s13059-017-1299-7

Nevill, P. G., Zhong, X., Tonti-Filippini, J., Byrne, M., Hislop, M., Thiele, K., van Leeuwen, S., Boykin, L. M., & Small, I. (2020). Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. *Plant Methods*, *16*(1), 1. https://doi.org/10.1186/s13007-019-0534-5

Noonan, J. P., Hofreiter, M., Smith, D., Priest, J. R., Rohland, N., Rabeder, G., & Rubin, E. M. (2005). Paleontology: Genomic sequencing of pleistocene cave bears. *Science*, *309*(5734), 597–600. https://doi.org/10.1126/science.1113485

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*(1), 385. https://doi.org/10.1186/1471-2105-12-385

Parducci, L., Alsos, I. G., Unneberg, P., Pedersen, M. W., Han, L. U., Lammers, Y., Salonen, J. S., Väliranta, M. M., Slotte, T., & Wohlfarth, B. (2019). Shotgun environmental DNA, pollen, and macrofossil analysis of Lateglacial lake sediments from Southern Sweden. *Frontiers in Ecology and Evolution*, *7*, 189. https://doi.org/10.3389/fevo.2019.00189

Paula, D. P., Linard, B., Crampton-Platt, A., Srivathsan, A., Timmermans, M. J. T. N., Sujii, E. R., Pires, C. S. S., Souza, L. M., Andow, D. A., & Vogler, A. P. (2016). Uncovering trophic interactions in arthropod predators through DNA shotgun-sequencing of gut contents. *PLoS ONE*, *11*(9), 1–14. https://doi.org/10.1371/journal.pone.0161841

Pedersen, M. W., Ruter, A., Schweger, C., Friebe, H., Staff, R. A., Kjeldsen, K. K., Mendoza, M. L. Z., Beaudoin, A. B., Zutter, C., Larsen, N. K., Potter, B. A., Nielsen, R., Rainville, R. A., Orlando, L., Meltzer, D. J., Kjær, K. H., & Willerslev, E. (2016). Postglacial viability and colonization in North America's ice-free corridor. *Nature*, *537*(7618), 45–49. https://doi.org/10.1038/nature19085

Pegard, A., Miquel, C., Valentini, A., Coissac, E., Bouvier, Frédéric, François, D., Taberlet, P., Engel, E., & Pompanon, François (2009). Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from feces. *Journal of Agricultural and Food Chemistry*, *57*(13), 5700–5706. https://doi.org/10.1021/jf803680c

Picozzi, N., Moss, R., & Catt, D. C. (1996). Capercaillie habitat, diet and management in a Sitka spruce plantation in central Scotland. *Forestry*, *69*(4), 373–388. https://doi.org/10.1093/forestry/69.4.373

Picozzi, N., Moss, R., & Kortland, K. (1999). Diet and survival of capercaillie Tetrao urogallus chicks in Scotland. *Wildlife Biology*, *5*, 11–23. https://doi.org/10.2981/wlb.1999.004

Pompanon, F., Deagle, B. E., Symondson, W. O. C., Brown, D. S., Jarman, S. N., & Taberlet, P. (2012). Who is eating what: Diet assessment using next generation sequencing. *Molecular Ecology*, *21*(8), 1931–1950. https://doi.org/10.1111/j.1365-294X.2011.05403.x

Quéméré, E., Hibert, F., Miquel, C., Lhuillier, E., Rasolondraibe, E., Champeau, J., Rabarivola, C., Nusbaumer, L., Chatelain, C., Gautier, L., Ranirison, P., Crouau-Roy, B., Taberlet, P., & Chikhi, L. (2013). A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. *PLoS ONE*, *8*(3), e58971. https://doi.org/10.1371/journal.pone.0058971

Regnaut, S., Lucas, F. S., & Fumagalli, L. (2006). DNA degradation in avian faecal samples and feasibility of non-invasive genetic studies of threatened capercaillie populations. *Conservation Genetics*, *7*(3), 449–453. https://doi.org/10.1007/s10592-005-9023-7

Salas, L. A., & Fuller, T. K. (1996). Diet of the lowland tapir (Tapirus terrestris L.) in the Tabaro River valley, southern Venezuela. *Canadian Journal of Zoology*, *74*(8), 1444–1451. https://doi.org/10.1139/z96-159

Sedinger, J. S. (1997). Adaptations to and consequences of an herbivorous diet in grouse and waterfowl. *Condor*, *99*(2), 314–326. https://doi.org/10.2307/1369937

Shipley, L. A., Forbey, J. S., & Moore, B. D. (2009). Revisiting the dietary niche: When is a mammalian herbivore a specialist. *Integrative and Comparative Biology*, *49*(3), 274–290. https://doi.org/10.1093/icb/icp051

Soininen, E. M., Gauthier, G., Bilodeau, F., Berteaux, D., Gielly, L., Taberlet, P., Gussarova, G., Bellemain, E., Hassel, K., Stenøien, H. K., Epp, L., Schrøder-Nielsen, A., Brochmann, C., & Yoccoz, N. G. (2015). Highly overlapping winter diet in two sympatric lemming species revealed by DNA metabarcoding. *PLoS ONE*, *10*(1), 1–18. https://doi.org/10.1371/journal.pone.0115335

Soininen, E. M., Valentini, A., Coissac, E., Miquel, C., Gielly, L., Brochmann, C., Brysting, A. K., Sønstebø, J. H., Ims, R. A., Yoccoz, N. G., & Taberlet, P. (2009). Analysing diet of small herbivores: The efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, *6*(1), 1–9. https://doi.org/10.1186/1742-9994-6-16

Sønstebø, J. H., Gielly, L., Brysting, A. K., Elven, R., Edwards, M., Haile, J., Willerslev, E., Coissac, E., Rioux, D., Sannier, J., Taberlet, P., & Brochmann, C. (2010). Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular Ecology Resources*, *10*(6), 1009–1018. https://doi.org/10.1111/j.1755-0998.2010.02855.x

Srivathsan, A., Ang, A., Vogler, A. P., & Meier, R. (2016). Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. *Frontiers in Zoology*, *13*(1), 1–13. https://doi.org/10.1186/s12983-016-0150-4

Srivathsan, A., Nagarajan, N., & Meier, R. (2019). Boosting natural history research via metagenomic clean-up of crowdsourced feces. *PLoS Biology*, *17*(11), 1–10. https://doi.org/10.1371/journal.pbio.3000517

Srivathsan, A., Sha, J. C. M., Vogler, A. P., & Meier, R. (2015). Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (Pygathrix nemaeus). *Molecular Ecology Resources*, *15*(2), 250–261. https://doi.org/10.1111/1755-0998.12302

Steinheim, G., Wegge, P., Fjellstad, J. I., Jnawali, S. R., & Weladji, R. B. (2005). Dry season diets and habitat use of sympatric Asian elephants (Elephas maximus) and greater one-horned rhinoceros (Rhinocerus unicornis) in Nepal. *Journal of Zoology*, *265*(4), 377–385. https://doi.org/10.1017/S0952836905006448

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C., & Willerslev, E. (2007). Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, *35*(3), e14. https://doi.org/10.1093/nar/gkl938

Tonti-Filippini, J., Nevill, P. G., Dixon, K., & Small, I. (2017). What can we do with 1000 plastid genomes? *Plant Journal*, *90*(4), 808–818. https://doi.org/10.1111/tpj.13491

Valdés-Correcher, E., Rodriguez, E., Kemp, Y. J. M., Wassen, M. J., & Cromsigt, J. P. G. M. (2018). Comparing the impact of a grazing regime with European bison versus one with free-ranging cattle on coastal dune vegetation in the Netherlands. *Mammal Research*, *63*(4), 455–466. https://doi.org/10.1007/s13364-018-0373-1

Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J. E., & Taberlet, P. (2009). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: The trnL approach. *Molecular Ecology Resources*, *9*(1), 51–60. https://doi.org/10.1111/j.1755-0998.2008.02352.x

Wegge, P., & Kastdalen, L. (2008). Habitat and diet of young grouse broods: Resource partitioning between Capercaillie (Tetrao urogallus) and Black Grouse (Tetrao tetrix) in boreal forests. *Journal of Ornithology*, *149*(2), 237–244. https://doi.org/10.1007/s10336-007-0265-7

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M. E., Lorenzen, E. D., Vestergård, M., Gussarova, G., Haile, J., Craine, J., Gielly, L., Boessenkool, S., Epp, L. S., Pearman, P. B., Cheddadi, R., Murray, D., Bråthen, K. A., Yoccoz, N., … Taberlet, P. (2014). Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, *506*(7486), 47–51. https://doi.org/10.1038/nature12921

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Xin, T., Su, C., Lin, Y., Wang, S., Xu, Z., & Song, J. (2018). Precise species detection of traditional Chinese patent medicine by shotgun metagenomic sequencing. *Phytomedicine*, *47*, 40–47. https://doi.org/10.1016/j.phymed.2018.04.048

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*, *178*, 779–794. https://doi.org/10.1016/j.cell.2019.07.010

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Chua PY, Crampton-Platt A, Lammers Y, Alsos IG, Boessenkool S, Bohmann K. Metagenomics: A viable tool for reconstructing herbivore diet. *Mol Ecol Resour.* 2021;21:2249–2263. https://doi.org/10.1111/1755-0998.13425