NEW ZEALAND JOURNAL OF ECOLOGY

RESEARCH

Nanopore sequencing of metagenomic DNA from rat stomach contents to infer diet

Nikki E. Freed[1,2]*, Adam N. H. Smith[2], Georgia Breckell[2,3], James Dale[2] and Olin K. Silander[2]

[1]School of Biological Sciences, University of Auckland, Auckland, New Zealand
[2]School of Natural Sciences, Massey University, Auckland 0745, New Zealand
[3]Ministry of Primary Industries, Auckland, New Zealand
*Author for correspondence (Email: olinsilander@gmail.com)

Published online: 29 November 2023

Abstract: Accurate determination of animal diets is difficult. Methods such as molecular barcoding or metagenomics offer a promising approach allowing quantitative and sensitive detection of different taxa. Here we show that rapid and inexpensive quantification of animal, plant, and fungal content from stomach contents is possible through metagenomic sequencing with the portable Oxford Nanopore Technologies (ONT) MinION. Using an amplification-free approach, we profile the stomach contents from 24 wild-caught rats. We conservatively identify stomach contents from over 50 taxonomic orders, ranging across nine phyla, including plants, vertebrates, invertebrates, and fungi. This highlights the wide range of taxa that can be identified using this approach. We calibrate the accuracy of this method by comparing the characteristics of reads matching the ground-truth host genome (rat) to those matching non-rat non-microbial taxa (i.e. stomach content) and show that at the family-level taxon assignments are approximately 97.5% accurate. Some inaccuracies may arise from biases in sequence databases, for example due to overrepresentation of DNA sequences from commonly studied species. We suggest a means to decrease the effects of database biases on inferring taxon membership when using metagenomic approaches. Finally, we implement a constrained ordination analysis and show that it is possible to identify the sampling location of an individual rat within tens of kilometres based on stomach contents alone. This work establishes proof-of-principle for long-read metagenomic methods in quantitative analysis of the stomach contents of a terrestrial mammal. We show that stomach content can be quantified even with limited expertise using a simple, amplification free workflow and a relatively inexpensive and accessible next generation sequencing method. Continued increases in the accuracy and throughput of ONT sequencing, along with improved genomic databases, suggests that a metagenomic approach for quantification of stomach contents, and by proxy animal diets, will become an important method in the future.

Keywords: diet analysis, next generation sequencing, *Rattus rattus*

Introduction

Accurate quantification of animal diets can yield critical insights into ecosystem and food web dynamics. However, unbiased and sensitive assessment of diet content is difficult to achieve, largely due to the limited accuracy of many current methods. Such methods include (1) visual inspection of gut contents (Daniel 1973; Pierce & Boyle 1991), which presents bias against items are most easily degraded (for example, soft-bodied species), (2) stable isotope analysis (Major et al. 2007; Carreon-Martinez & Heath 2010), which yields only broad information on diet, such as whether diet items are terrestrial or marine (Hobson 1987; Basha et al. 2016), and (3) time-lapse video (Dunlap & Pawlik 1996; Brown et al. 2008), for which species identification is difficult for small prey items or in low-light conditions.

To circumvent these issues, DNA-based methods (King et al. 2008; Soininen et al. 2009) have become popular. Perhaps the most widely applied DNA-based method is metabarcoding. This approach relies on PCR amplification and sequencing of conserved regions from nuclear, mitochondrial, or plastid genomes (King et al. 2008). With adequate primer selection for PCR amplification this method can detect a wide range of species and does not require taxon-specific expertise, which is often necessary for other methods.

However, DNA metabarcoding is also not free from bias. This is primarily because PCR primers must be specifically tailored to particular sets of taxa or species (Jarman et al. 2002). Although universal PCR primer pairs have been developed, for example targeting all bilaterians or even all eukaryotes (Jarman et al. 2004), all primer sets exhibit bias towards certain taxa. Even when using different sets of fungal-specific PCR primer pairs, five-fold differences in fungal operational taxonomic units (OTU) estimates have been found (Tedersoo et al. 2015). It has also been shown that published universal primer pairs are only capable of amplifying between 57% and

DOI: https://doi.org/10.20417/nzjecol.47.3554

91% of tested metazoan species with as few as 33% of species in some phyla being amplified at all (e.g. cnidarians) (Leray et al. 2013). Primers directed at different genomic loci can exhibit up to 2000-fold differences in detection (Pawluczyk et al. 2015). The choice of polymerase can also bias diversity metrics when using metabarcoding (Pereira et al. 2018). For these reasons, an approach that circumvents PCR amplification and thus avoids these biases is desirable.

Metagenomic sequencing is an amplification-free method that aims to directly sequence all the DNA in a sample. Although there are still biases with this approach, for example due to nucleotide content affecting the likelihood of a molecule being sequenced, these are inherently less than those introduced by amplification steps during metabarcoding. Metagenomic approaches have long been used to yield insights into microbial diversity and function (Tyson et al. 2004; Fierer et al. 2012; Xu & Knight 2015; Anantharaman et al. 2016; Hover et al. 2018), while metagenomic applications aimed at eukaryotic taxa identification are less common. Several diet studies have performed metagenomic sequencing and then implemented computational filtering steps to select only mitochondrial DNA sequence or metabarcode regions, or have used abridged databases before data analysis to decrease biases in sequence databases, for example due to overrepresentation of commonly studied species. (Arribas et al. 2016; Paula et al. 2016; Srivathsan et al. 2016). However, to our knowledge, very few studies have used unfiltered metagenomic sequence analysis to infer diet (Srivathsan et al. 2015; Søe et al. 2018).

Here, we establish a proof-of-principle methodology to accurately classify metagenomic sequences from eukaryotic taxa and infer diet content using low-accuracy, long-read sequencing, Oxford Nanopore (ONT). Toward this aim, we quantify rat stomach contents from several locations in the North Island of New Zealand.

Using rat stomach samples provides three distinct advantages. First, rats are extremely omnivorous. As such, they serve as an excellent means to quantify the breadth of taxa that can be detected using a metagenomic long read approach.

Second, the use of stomach samples means that a significant number of reads will be host reads. This allows us to assess the characteristics of true positive sequence reads (reads derived from DNA isolated from rat cells that also match rat database sequences), as well as false positive reads (rat-derived reads that match non-rat database sequences). We can then determine whether reads matching non-host, non-bacterial taxa (hypothetical diet items or other stomach contents) have similar characteristics to known true positive reads. This use of host reads to establish false positive taxon assignments is exactly analogous to feeding the rats a diet of known content (i.e. rat) and testing whether the contents of the known diet can be accurately identified.

Third, rat stomach samples may provide a proxy for rat diet content. Knowledge of rat diets is critical in understanding the ramifications of rats as ecosystem players. It is well-established that the relatively recent introduction of mammalian predators to New Zealand has had significant negative effects on many of the native animal populations. This has ranged from insects (Gibbs 1998), to reptiles (Towns et al. 2001), molluscs (Stringer et al. 2003), and birds (Diamond & Veitch 1981; Dowding & Murphy 2001) with downstream effects on terrestrial and aquatic ecosystems (Graham et al. 2018). To counteract the effects of mammalian predators, an ambitious plan is currently being put into place that aims for the eradication of all mammalian predators from New Zealand (including possums, rats, stoats, and hedgehogs), by 2050 (http://www.doc.govt.nz/predator-free-2050; Russell et al. 2015). A useful step toward this goal would be to prioritise the management of predators and establish in which locations native species experience the highest levels of predation. To do so requires establishing the diet content of local mammalian predators. Here, we use a proxy for rat diet: rat stomach content.

Finally, we note that using DNA sequencing to profile stomach contents (diet) should not require extensive or deep sequencing to accurately determine the type and number of organisms consumed by an animal. Here we aim to quantify all organismal tissue present in the stomach at a fraction of 1% or more. If we assume that read counts are Poisson distributed, with only 2000 sequencing reads we will quantify the number of reads from within 2-fold of their true amount 99% of the time. However, it is overly optimistic to assume that one could calculate with any certainty the amount of tissue from this read number; undeniably there are different amounts of DNA per gram of tissue and there are strong biases in DNA isolation from different tissues.

Methods

Study areas

We trapped rats from three locations near Auckland, New Zealand. Each location comprised a different type of habitat: undisturbed inland native forest (Waitākere Regional Park: WP); native bush surrounding an estuary (Okura Bush Walkway: OB); and restored coastal wetland (Long Bay Regional Park: LB) (Fig. 1). Snap traps in OB and LB were baited with peanut butter, apple, and cinnamon wax pellets; or bacon fat and flax pellets.

Traps in WP were baited with chicken eggs, rabbit meat, or cinnamon scented poison pellets. From 16 November to 16 December 2016, traps were surveyed by established conservation groups at each site every 48 hours. A total of 36 rats were collected from these locations. Three of the rats from WP had poison in their stomachs, which may have influenced their foraging behaviour. However, all three of these were killed by baited snap traps, suggesting their behaviour was not completely abnormal. In addition, they are unlikely to have swallowed any bait, as none of these rats were identified as having chicken or rabbit in their stomachs. Most rats collected (34/36) were determined to be male *Rattus rattus* by visual inspection. These 34 rats were selected for further analysis.

DNA isolation

Within 48 hours of trapping, rats were stored at either −20°C or −80°C until dissection. We dissected out intact stomachs from each animal and removed the contents using a sterile spatula. After snap freezing in liquid nitrogen, we homogenised the stomach contents using a sterile mini blender to ensure sampling was representative of the entire stomach.

We purified DNA from 20 mg of homogenised stomach contents using the Promega Wizard Genomic DNA Purification Kit, with the following modifications to the Animal Tissue protocol: after protein precipitation, we transferred the supernatant to a new tube and centrifuged a second time to minimise protein carryover. The DNA pellet was washed twice with ethanol. These modifications were performed to improve DNA purity. We rehydrated precipitated DNA by incubating overnight in molecular biology grade water at 4°C and stored

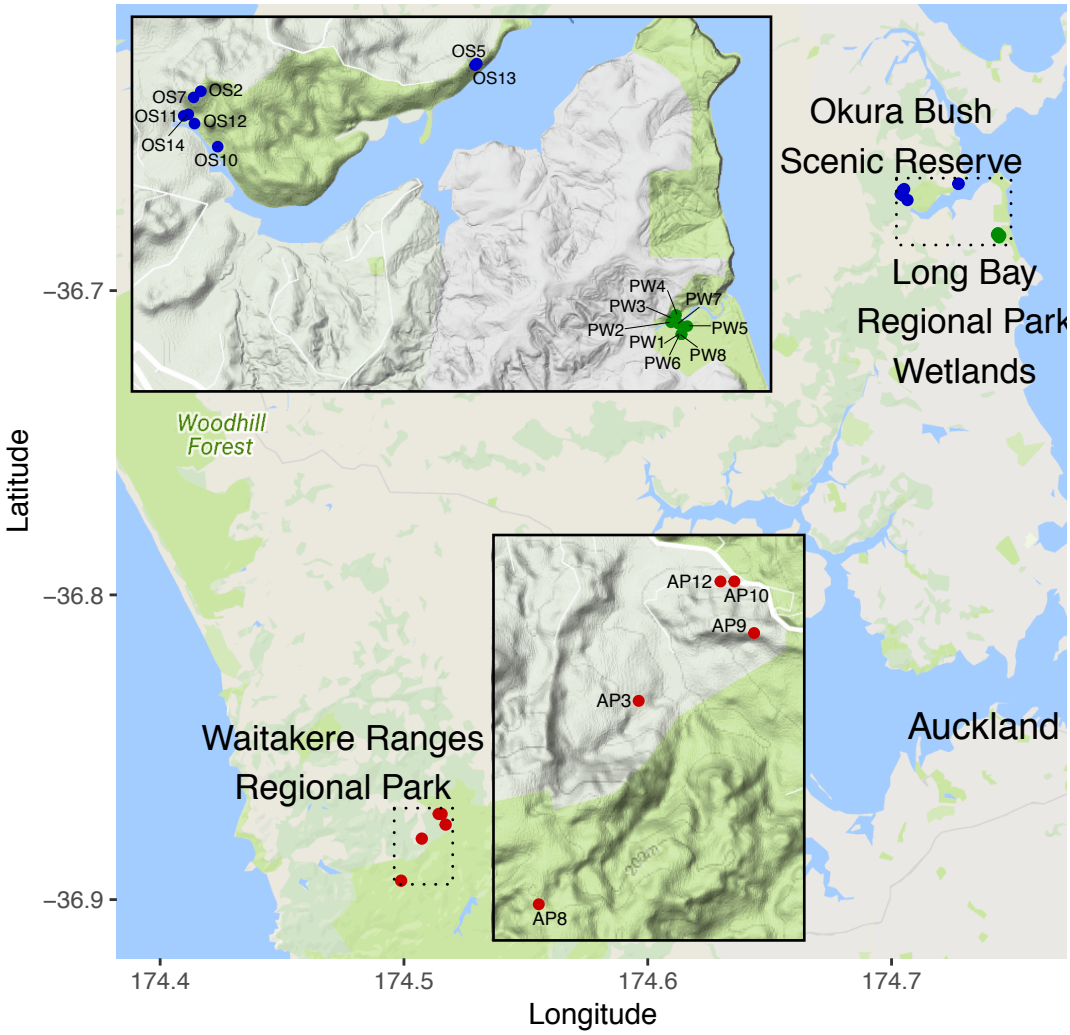

Figure 1. Location of rat sampling sites in the greater Auckland area in the North Island of New Zealand. Each point indicates a trap where one rat was captured, with the colour of the points indicating the three broad locations: the native estuarine bush habitat of Okura Bush (OB), the restored wetland of Long Bay (LB), and the native forest of Waitākere Regional Park (WP). The two insets show the three locations in higher resolution with topographical details. Green indicates park areas. Precise geographical coordinates were only

the DNA at −20°C. DNA quantity, purity, and quality was ascertained by nanodrop and agarose gel electrophoresis. The DNA samples were ranked according to quantity and purity (based on A260/A280 and, secondarily, A230/A280 ratios). The eight highest quality DNA samples from each of the three locations were selected for sequencing. We did no size selection on the purified DNA.

DNA sequencing

Sequencing was performed on two different dates (24 January 2017 and 17 March 2017) using a MinION Mk1B device and R9.4 chemistry. For each sequencing run, DNA from each rat was barcoded using the 1D Native Barcoding Kit (Barcode expansion kit EXP-NBD103 with sequencing kit SQK-LSK108) following the manufacturer's instructions. This included an AMPure bead purification step to remove adaptors, which also likely removed very short reads (less than 200 bp; see Fig. 2a). Twelve samples were pooled and run on each flow cell for a total of 24 individual rats. The flow cells had 1373 active pores (January 2017; Appendix S1 in Supplementary Material) and 1439 active pores (March 2017; Appendix S2). Both runs were re-basecalled after data collection using Albacore 2.2.7 with demultiplexing performed in Albacore and filtering disabled (options --barcoding --disable_filtering).

Sequence classification

All sequences were BLASTed (blastn v2.6.0+) against a locally compiled database consisting of the NCBI other_genomic (RefSeq chromosome records for non-human organisms), and NCBI nt databases (the partially non-redundant nucleotide sequences from all traditional divisions of GenBank excluding genome survey sequence, EST, high-throughput genome, and whole genome shotgun). Both were downloaded on 13 June 2018 from NCBI. Default blastn parameters were used (match 2, mismatch -3, gapopen -5, gapextend -2). Due to the predominance of short indels present in nanopore sequence data, we tested whether changing these default penalties affected the results (gapopen -1, gapextend -1). We found that these adjusted parameters did not qualitatively change our results.

We assigned sequence reads to specific taxon levels using MEGAN6 (v.6.11.7 June 2018) (Huson et al. 2016). We only used reads with BLAST hits having an e-value of 1×10^{-20} or

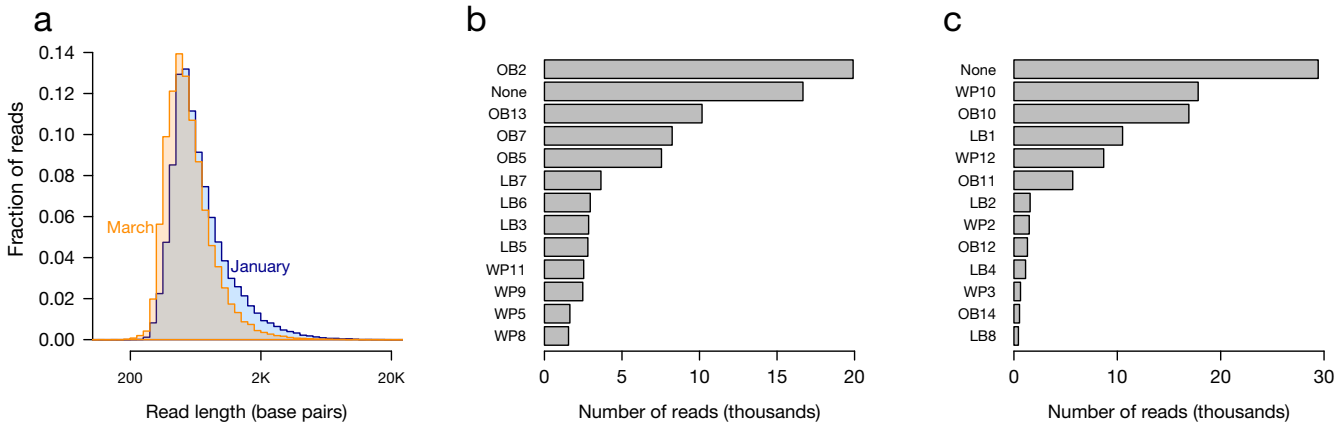

Figure 2. Run statistics of nanopore metagenomic sequencing of rat stomach contents. Barcode distributions for January (a) and March (b) runs. We multiplexed the samples on the flow cells, using 12 barcodes per flow cell. The distribution of read numbers across barcodes varied by up to 40-fold. 20% (January) and 30% (March) of all reads could not be assigned to a barcode ("None"). The inability to assign these reads to a barcode is due primarily to their lower quality. (c) Read length distribution for January and March nanopore runs. 90% of the reads were between 350 bp and 1580 bp in length, with only 0.55% being longer than 4,000 bp.

lower (corresponding to a bit score of 115 or higher given the databases we used) and an alignment length of 100 base pairs or more. MEGAN6 employs a cut-off and lowest common ancestor (LCA) algorithm to assign reads to a taxon. For example, if a read has BLAST hits to five different species, three of which have bit scores within 20% of the best hit, the read will be assigned to the genus, family, order, or higher taxon level that is the LCA of those best-hit three species (Huson et al. 2007). If a read matches one species far better than any other, by definition, the LCA is that species. Here, to assign reads to taxon levels, we considered all hits having bit scores within 20% of the bit score of the best hit (MEGAN parameter Top Percent).

Multivariate analyses

Multivariate analyses were done using the software PRIMER v7 (Clarke & Gorley 2015). The data used in the multivariate analyses were in the form of a sample- (i.e. individual rat) by-family matrix of read counts. All bacteria, rodent, and primate families were removed as these are not diet items, but microbiome, host (rat), or contamination, apparent as most primate hits (32 in total) were assigned to Hominidae (Appendix S3).

The read counts were converted to proportions per individual rat by dividing by the total count for each rat, to account for the fact that the number of reads varied substantially among rats. The proportions were then square root transformed so that subsequent analyses were informed by the full range of taxa, rather than just the most abundant families (Clarke & Green 1988). We then calculated a matrix of Bray-Curtis dissimilarities, which quantified the difference in the gut DNA of each pair of rats based on the square root transformed proportions of read counts across families (Clarke et al. 2006).

We applied an unconstrained ordination method, non-metric multidimensional scaling (nMDS) to the dissimilarity matrix to examine the overall patterns in the stomach content composition among rats. To assess the degree to which the stomach compositions of rats were distinguishable among the three locations, we applied canonical analysis of principal coordinates (CAP) (Anderson & Willis 2003) to the dissimilarity matrix. CAP is a constrained ordination which aims to find axes through multivariate data that best separates a priori delineated groups of samples (in this case, the groups were the locations from which the rats were sampled). CAP is akin to linear discriminant analysis, but it can be used with any resemblance matrix. The out-of-sample classification success was evaluated using a leave-one-out cross-validation procedure (Anderson & Willis 2003).

We used similarity percentage analysis (SIMPER; Clarke 1993) to characterise stomach composition and distinguish between the locations. This allowed us to identify the taxa with the greatest percentage contributions to (1) the Bray-Curtis similarities of stomach composition within each location (Appendix S4) and (2) the Bray-Curtis dissimilarities between each pair of locations (Appendix S5).

Results

DNA sequencing

We selected eight rats from each of the three locations for diet quantification. Each location comprised a different type of habitat: undisturbed inland native forest (Waitākere Regional Park; WP); native bush surrounding an estuary (Okura Bush Walkway; OB); and restored coastal wetland (Long Bay Regional Park; LB). We isolated DNA from whole homogenised stomach contents from each rat. We sequenced these DNA samples on two dates, multiplexing the samples on each date. We obtained a total of 82 977 reads (January 2017) and 96 150 reads (March 2017). These numbers were not far below expectations given the flow cell and kit chemistry and MinKNOW software versions available at that time (Jain et al. 2018). However, these read numbers are considerably below those expected for current ONT flow cells and software, which has improved per flow cell output approximately 100-fold.

After de-multiplexing the reads, we found large variation in the numbers of reads per barcode (i.e. animal): approximately 10-fold variation across the stomach samples sequenced in January, and up to 40-fold for those samples sequenced in March (Fig. 2b and 2c). We hypothesise that this is due to the highly variable quality of DNA in each sample. This did not appear to have strong effects on read accuracy, as the median

quality scores per read ranged from 7–12 (0.80–0.94 accuracy) for both runs. The degradation of the DNA during digestion in the stomach, as well as fragmentation during DNA isolation (Deagle et al. 2006) and sequencing library preparation led to relatively short median read lengths of 606 bp and 527 bp for the January and March datasets, respectively (Fig. 2a). However, there was wide variation in read length, with almost 10% of all reads being longer than 1200 bp. Longer reads, although more error prone, are often more accurate in determining taxonomy than shorter reads (Pearman et al. 2020).

Assignment of reads to taxa

To quantify stomach contents, we first BLASTed all sequences against a DNA sequence database containing a large number of eukaryotic taxa. We used BLAST as it is generally viewed as the gold standard method in metagenomic analyses (McIntyre et al. 2017). Of the 133 022 barcoded reads, 30 535 (23%) hit a sequence in the combined nt and other_genomic database at an e-value cut-off of 1e–2.

We first aimed to assess the quality of these hits. We found a bimodal distribution of alignment lengths and e-values (marginal histograms in Fig. 3a). We also noticed that mean read quality had substantial effects on the likelihood of a read yielding a BLAST hit, with almost 40% of high accuracy reads (quality scores greater than 92%) having hits in the March dataset, compared to 1% of low accuracy reads (quality scores less than 75%; Fig. 3b). We found the same pattern in the January dataset, although to a lesser extent.

We hypothesised that many of the short alignments with high e-values were false positives given the reported taxa. We thus first filtered the BLAST results, only retaining hits with e-values less than 1e–20 and alignment lengths greater than 100 bp. Similar quality filters based on length and identity have been imposed previously (Srivathsan et al. 2015). A total of 22 154 hits (73%) passed this e-value filter.

We next used MEGAN6 (Huson et al. 2016) to assign reads to specific taxa. 16 820 reads (76%) were assigned to a taxon. Of these, 31% were assigned by MEGAN as being bacterial, and 55% of these were *Lactobacillus* spp. These results match previous studies on rat stomach microbiomes (Brownlee & Moss 1961; Horáková et al. 1971; Maurice et al. 2015; Li et al. 2017). Plant-associated *Pseudomonas*, as well as *Lactococcus* taxa, were also common, at 7% and 6%, respectively.

MEGAN also assigned reads to a wide range of eukaryotic taxa. To conservatively infer taxon presence, we first reclassified all MEGAN species-level assignments to the level of genus. However, after this, several clear false positive taxon assignments remained (e.g. hippo and naked mole rat). These matches were generally short and of low identity. To reduce such false positive taxon inferences, we used information from reads assigned to the genera *Rattus* (rat) and *Mus* (mouse), using the following strategy.

We inferred that the reads assigned to *Rattus* (2696 reads in total) were true positive genus-level assignments (deriving from DNA isolated from *Rattus* cells that were collected during the acquisition of the stomach contents), and that the reads assigned to *Mus* (2798 reads in total) were false positive genus-level assignments (i.e. they were derived from *Rattus* cells and not *Mus*-derived). Relying on these true-positive and

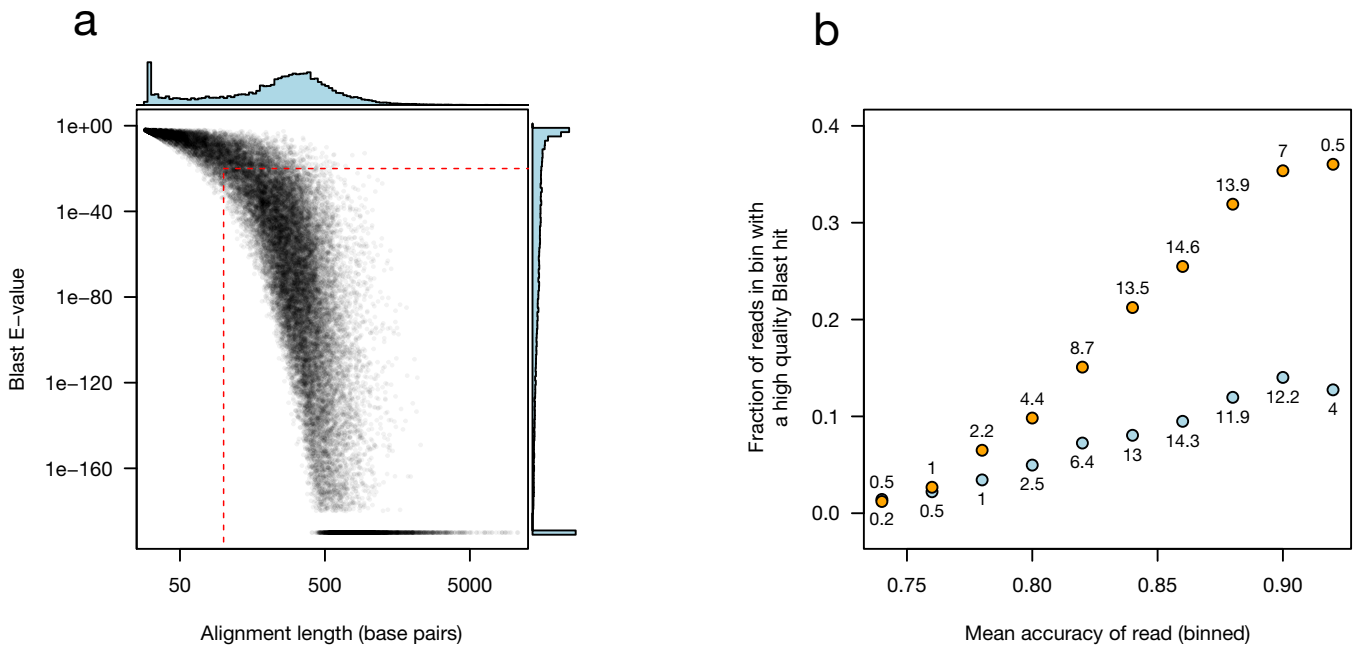

Figure 3. BLAST hits of metagenomic reads. (a) Alignment lengths and e-values were bimodally distributed. The y-axis is plotted on a log scale, with zero e-values suppressed by adding a small number (1e–90) to each e-value. The horizontal red dotted line indicates the e-value cut-off we implemented, and the vertical red dotted line indicates the length cut-off (e-value < 1e–20 and alignment length of 100, respectively) to decrease false positive hits. (b) The fraction of reads with high quality BLAST hits (e-value < 1e–20) increased as a function of read accuracy. We binned the data according to mean read accuracy (bin width = 0.02) and calculated the fraction of reads within each bin that have a high-quality BLAST hit (alignment length greater than 100bp and e-value less than 1e–20) for the January and March runs separately (blue and orange points, respectively). The number of reads in each bin is indicated above each point (in thousands). There is a clear positive correlation between mean accuracy and the likelihood of a high-quality BLAST hit, reaching almost 40% for high accuracy reads (> 92.5%) for the March dataset.

false positive read sets, we can implement filtering strategies to determine the characteristics of true positive and false positive taxon assignments. Having these read sets provides the same advantages as feeding a diet of known content and testing whether the contents of the known diet can be accurately identified. Here, the known diet is *Rattus*.

First, it is important to establish that the reads assigned to *Mus* are false positive taxon assignments. If these were true positive *Mus* reads, then they would necessarily be due to mouse predation. Although rats are known to prey on mice (Bridgman et al. 2013), if this had recently occurred in some rats, we would expect that (1) the ratio of mouse to rat reads would be higher than in rats that had not predated mice, and (2) the percent identity of the reads assigned to *Mus* would be higher than in rats that had not predated mice. However, we found that the ratio of mouse to rat reads and percent identity of reads assigned to *Mus* was similar for all rats. This suggested either that all rats had predated similar amounts of mice very recently (which we view as unlikely), or that these *Mus* hits were indeed false positives. Thus, we use the *Mus* hits to delineate false positive and true positive genus-level assignment using the specific read identity and alignment length characteristics of each read set.

We first noted that the mean percent identity values of the best BLAST hits for *Rattus* and *Mus* reads differed, with reads matching *Rattus* having a median identity of 86.4%, and reads matching *Mus* having 81.0% median identity (Fig. 4a). The mean percent identity for *Rattus* reads corresponds very well to that expected given the mean quality scores of the reads (86.4% identity corresponds to a mean quality score of 8.7, similar to what we observed; Appendix S1a–c).

A second characteristic we considered was the ratio of alignment length to read length; if a read fully aligns this ratio is one. Generally, higher quality alignments should have higher ratios. Indeed, we found this ratio differed substantially between the *Rattus*- and *Mus*-assigned reads: the median ratio of alignment length to read length was longer for *Rattus* (0.57) than *Mus* (0.52; Fig. 4b).

We used these read characteristics to select cut-off values for assigning reads as being true positive or false positive genus-level taxon assignments. For genus-level assignments, we required at least 82.5% alignment identity, and an alignment length to read length ratio of at least 0.55. For any alignment of lower quality, we assign reads at the family level. These cut-offs excluded 88% of the reads assigned to *Mus*, instead assigning them one taxon level higher to the family Muridae (which contains the genus *Rattus*).

However, the remaining fraction of false positive assignments at the genus level (12%) was still relatively high, and there may frequently be instances in which a lower rate of false positives is desired. One means of decreasing false positive taxon assignments is to only consider higher-level taxon classifications. We thus also quantified the rate of false positive assignment at the level of family rather than genus. We first identified all reads classified as being from the order Rodentia. Within this order, we assumed that reads assigned to the family Muridae were true positives, and that reads assigned to any other family in the order Rodentia were false positives (i.e. no rats had predated animals from other families in Rodentia). This assumption is conservative when calculating the fraction of false positives, as rats may indeed have predated some families.

We then again implemented specific percent identity and read length ratio cut-offs, requiring reads assigned at the Family level to have database matches of at least 77.5% identity, an alignment length to read length ratio of at least 0.1, and a total alignment length of at least 150 bp. With these cut-offs, 97.3% of all reads assigned to the order Rodentia were classified in the family Muridae (which contains the genus *Rattus*). The remaining 2.7% were assigned to the family Cricetidae (voles and lemmings), except for four reads assigned to Spalacidae (mole-rats) (Appendix S8). All these family assignments

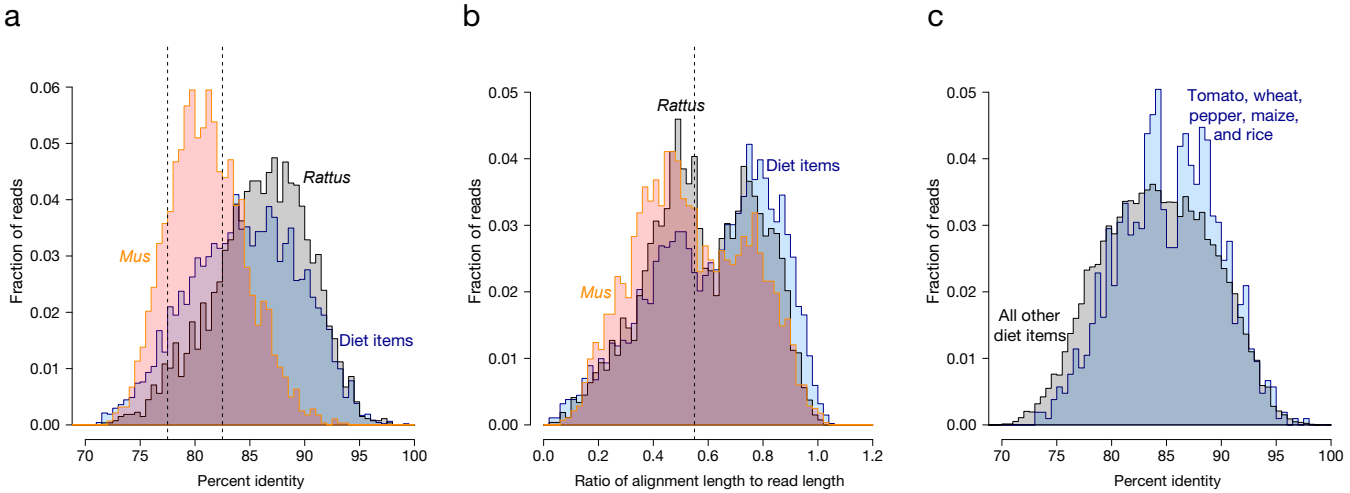

Figure 4. Distributions of percent identity and length for alignments of reads matching *Rattus* (rat), *Mus* (mouse), and hypothetical diet items. (a) The percent identity for alignments of *Rattus* and hypothetical diet items is much higher than for *Mus*. Histograms of the percent identity of the alignment of the top BLAST hit with the read. *Mus* matches have substantially lower percent identity compared to both *Rattus* and diet items. The dotted lines indicate the cut-offs that we implemented for inferring reads as belonging to a specific genus (above 82.5% identity) or family (above 77.5% identity). (b) Ratios of alignment lengths to read lengths of *Rattus* and sequence inferred as diet items are higher than sequences assigned to the genus *Mus*. This plot is analogous to that in (a). The dotted line indicates the cut-off that we implemented for inferring reads as belonging to a specific genus (above 0.55). (c) Percent identity for plant stomach contents that are non-native (tomato, wheat, pepper, maize, and rice) are as high or higher than sequences from other taxa. This suggests that these are not false positives due to database bias, but true positive assignments due to rats consuming human food stuffs.

are clear false positives, as it is highly unlikely that these families were predated. However, these results establish that by implementing specific cut-off values for alignments, we can ensure a low rate of false positive assignments at the family level.

However, it is still possible that database bias may still make the quantification of stomach contents inaccurate. For example, *Mus* and *Rattus* sequences are among the most common in the database that we matched against, whereas diet-related taxa may be extremely rare and thus much less likely to match to. However, if reads are assigned to the wrong taxon (as above), one expectation is that these false positive assignments will have lower percent identities and low alignment lengths. We thus checked whether read alignments of all inferred diet items had percent identities and alignment lengths similar to the true positive *Rattus* alignments, or instead whether they were more similar to the false positive *Mus* alignments (or worse). We found that most sequences from stomach contents had alignment percent identities that overlapped with the *Rattus* reads. Furthermore, the alignment length to read length ratios often exceeded those for the *Rattus* reads. This suggested that the taxa assignments were often correct down to the level of genus (as the *Rattus*-assigned reads are correct to the level of genus) and were not false positive assignments. Despite this indication of genus-level accuracy, here we conservatively report taxa at the level of family.

For reads that did not pass the above cut-offs, we placed taxon assignments at the level of order, or used the taxon level assigned by MEGAN. Using these cut-offs, 16% of all reads were classified at the genus-level (although for the analyses below we consider these at the Family-level); 71% were classified at the family-level or genus-level; 89% were classified at the order-level or below; and 98% were classified at the phylum-level or below.

There were few clear false positive taxon assignments after filtering steps, and most clear false positives had alignment characteristics just above our cut-offs (Appendix S6). The exception to this were three reads from two rats matching Buthidae (scorpions), which had alignment lengths of 762 bp, 664 bp, and 298 bp with identities of 83%, 88%, and 79%, respectively. It is unlikely these are true positives, and instead we hypothesise that these rats predated harvestmen (Opiliones), a closely related sister taxon within Arachnida, but lacking significant amounts of genomic data in the database. Despite the presence of these false positive taxa, we did not further increase the stringency of our filters, as the fractions were very small.

A much larger number of reads were also assigned to genera that are clearly non-native to New Zealand, especially plant genera: *Triticum*, *Solanum*, *Zea*, *Capsicum*, and *Oryza*. We inferred that these were true positive assignments arising from human-associated food items: wheat, tomato, maize, pepper, and rice, respectively. There are two pieces of evidence supporting this. First, these sequences had as high, or higher, percent identities to the database matches as other diet items (Fig. 4c). Second, sequences from these genera were usually present in only a single rat, as opposed to appearing across rats, which is what would be expected if they were rare and random false positive assignments. For example, out of the 232 sequences assigned to *Triticum*, 207 came from only two rats (with 105 and 102 reads respectively).

Taxon variability

Within each rat, we identified a wide variety of plant, animal, and fungal orders, ranging from two to 25 orders per rat (mean 8.7; Fig. 5). In total, we identified taxa from 68 different families, 55 different orders, 15 different classes, and eight different phyla (Fig. 6). These results highlight the taxonomic breadth that can be achieved using this approach. However, this breadth comes at the sacrifice of specificity, as we are only able to consistently identify taxa down to the level of family.

Plant taxa were the majority of stomach contents,

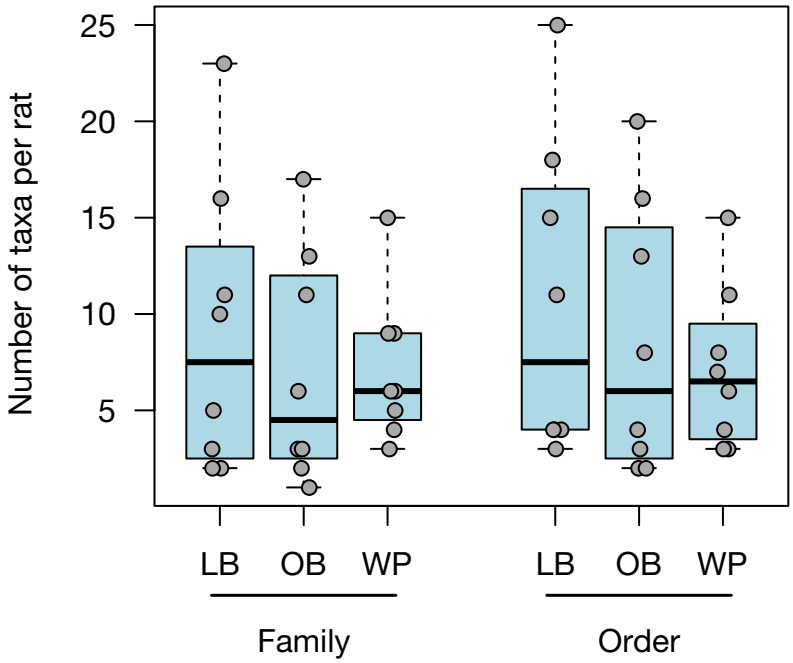

Figure 5. Numbers of taxa in individual rats. Each boxplot indicates the range of families (left boxes), or orders (right boxes) consumed by each rat in each location (OB: Okura Bush, native bush; LB: Long Bay Park, restored wetland; WP: Waitākere Regional Park, native forest). The numbers for individual rats (eight per location) are plotted in grey.

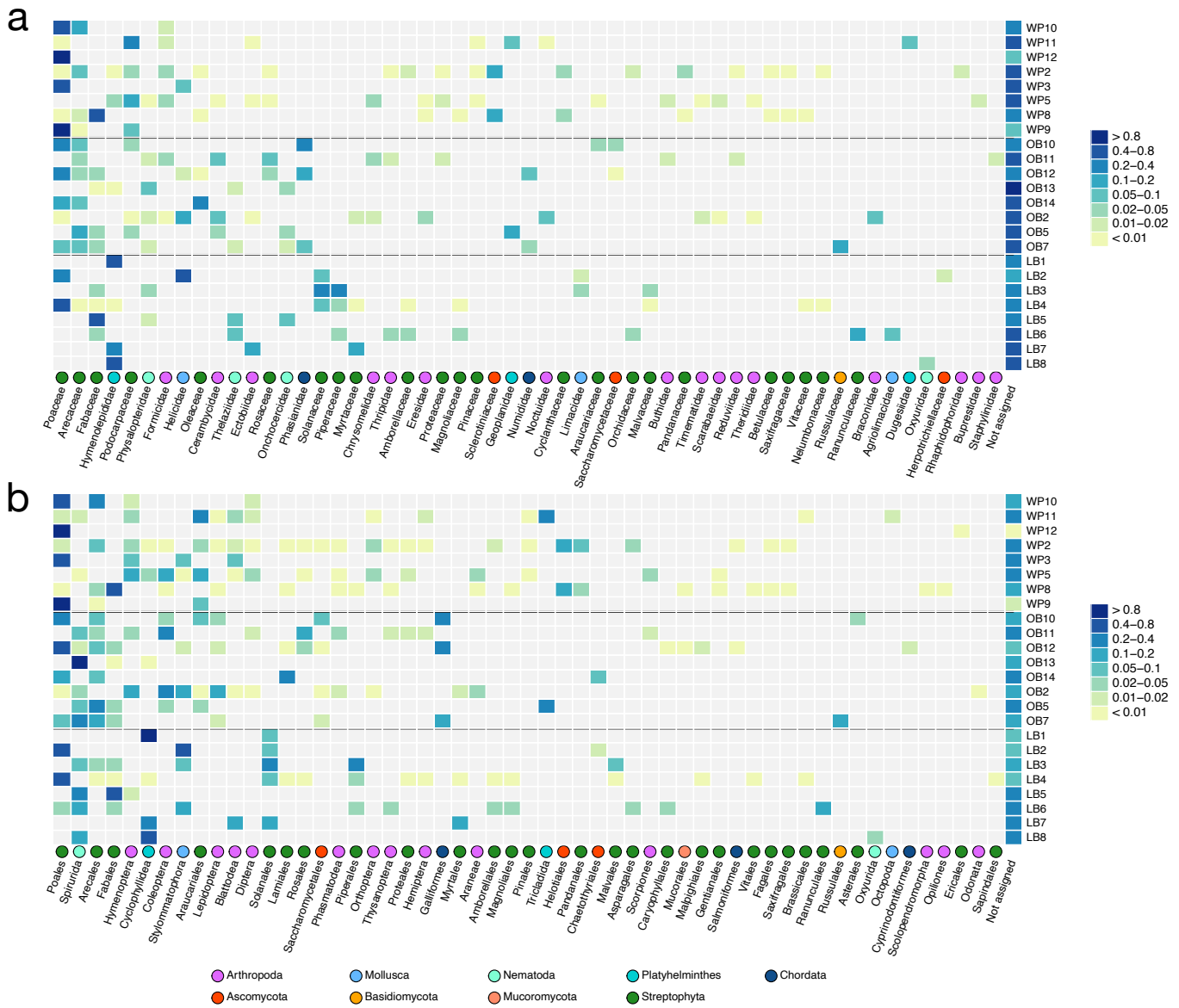

Figure 6. Proportions of taxa in the stomach contents of individual rats. (a) Reads assigned to taxa at the family and (b) order level. The rows correspond to a single rat, with the proportions of reads for that rat assigned to each family or order indicated in shades of blue and yellow. Reads that were not assigned to a specific family or order are indicated at the right side of the figure. The families and orders have been sorted so that the most common taxa appear on the left. Only the 55 most common families are shown. Note that the colour gradations presented on the scale are not linear.

with four predominant orders: Poales (grasses), Fabales (legumes), Arecales (palms), and Araucariales (specifically, Podocarpaceae, a common native New Zealand tree family). The dominance of plant matter (fruits and seeds) in rat diets has been established previously (Sweetapple & Nugent 2007; Riofrío-Lazo & Páez-Rosas 2015). Animal taxa made up a smaller component of each rat's stomach contents, with Insecta dominating: Hymenoptera (bees, wasps, and ants), Coleoptera (beetles), Lepidoptera (moths and butterflies), Blattodea (cockroaches), Diptera (flies), and Phasmatodea (stick insects). In addition, Stylommatophora (slugs and snails) were present in substantial numbers (Fig. 6a and 6b). Fungi were only a small component of the rats' stomach contents, although several orders were present: Sclerotiniales (commonly plant pathogens), Saccharomycetales (budding yeasts), Mucorales (pin moulds), Russulales (brittlegills and milk-caps), and Chytotheriales (black yeasts). Finally, for many rats, a substantial proportion of the stomach contents were parasitic worms, primarily Spirurida (nematodes) and Hymenolepididae (tapeworms), and are almost certainly not diet items, but parasitic infections.

It is important to note that due to our metagenomic approach, the fraction of each element of the rats' diets may be distorted by biases in genomic databases: whole genome data exists for only a few taxa, while mtDNA, rDNA, metabarcode loci, and microsatellite sequence data are present in the database for many animal and plant genera. However, we propose it is possible to decrease this bias.

To quantify database-driven bias for taxa that we consider diet items, for each taxon we determined the fraction of hits that mapped to mtDNA, rDNA, microsatellites, or EST libraries (we refer to this as non-genomic, as these data are not

from genomic sequencing projects). We also determined the fraction of hits that mapped to DNA sequences arising from genome sequencing projects. We expect that for animals with sequenced genomes, these two fractions should be primarily determined by the relative amounts of mtDNA and nuclear DNA in a diet item, rather than database bias. If a diet item consists of cells that have large numbers of mitochondria (or if the organism being consumed has a small genome), we expect a large fraction of reads will map to mtDNA sequences even if there is genomic sequence in the database. Alternatively, if a diet item consists of cells with few mtDNA then most reads will map to genomic sequence (if there is genomic sequence in the database). In contrast, for animals without sequenced genomes most database matches will be from mtDNA, plastid DNA, rDNA, and microsatellites (non-genomic sequence), with few (if any) genomic hits. This will be true regardless of the relative amounts of mtDNA and nuclear DNA in the diet item. However, by comparing the relative fractions of non-genomics and genomic DNA for taxa that we know to have complete genome sequences in the database to taxa without complete genomes, we propose it is possible to decrease the effects of this database bias.

For animal genera with at least one species that had sequenced genomes in the database, we found that the fraction of reads that mapped to non-genomic sequence (mtDNA, rDNA, microsatellite, plastid, and EST library) ranged from 0% (*Coturnix*, quail) to 39% (*Gallus*, chicken) (Fig. 7). This is a considerable range and we hypothesise that variation in the fraction of non-genomic reads is due to the type of tissue sequenced (affecting mtDNA content). However, it is also possible that the results are biased by the relative frequency of specific types of studies. For example, there may be very few mtDNA or microsatellite studies in *Coturnix* relative to *Gallus*. For *Rattus* we found that 27% of all reads were non-genomic. In this case, the sequenced DNA tissue was likely to be from stomach muscle cells that were scraped out during removal of the stomach contents. As muscle has a relatively high fraction of mtDNA, this perhaps explains the large fraction of reads mapping to non-genomic DNA.

For plant genera with at least one species having a sequenced genome, the fraction of reads matching non-genomic sequence (mostly mtDNA, plastid, and rDNA) was generally lower: between 2% (*Cenchrus*, buffelgrass) and 12% (*Zea*, corn). On average, for animals with sequenced genomes present in the database, approximately 30% of all reads mapped to non-genomic sequences; for plants, approximately 5% mapped to non-genomic sequences. Again, this difference may arise from there being fewer mtDNA or microsatellite studies in plants compared to animals.

For taxa with little or no genomic sequence in the database, the vast majority of matches were non-genomic (mtDNA, plastid, rDNA, or microsatellite loci): 90% of *Phoenix* (date palm) hits; all *Helix* (snail); and all Rhaphidophorae (endemic cave weta) hits. All Arthurdendyus (endemic New Zealand flatworm) hits were solely to rDNA loci.

These ratios are in strong contrast to animals with sequenced genomes, for which an average of only 30% of all reads mapped to non-genomic sequence. This suggests that for animal taxa with little or no genomic sequence data, we have underestimated the actual number of sequences (or proportion of stomach content) from that taxon by two- to three-fold. For plant taxa with little or no genomic sequence data, we may have underestimated read abundance by approximately 20-fold. Although there is considerable uncertainty in both of these estimates, they yield some insight into how database bias (and cellular DNA content) can affect estimates of organismal proportions during metagenomic sequencing.

Differences in stomach contents between locations

Our read classification results indicated that specific taxa were overrepresented in the stomach contents of rats from particular locations. For example, six out of eight rats from

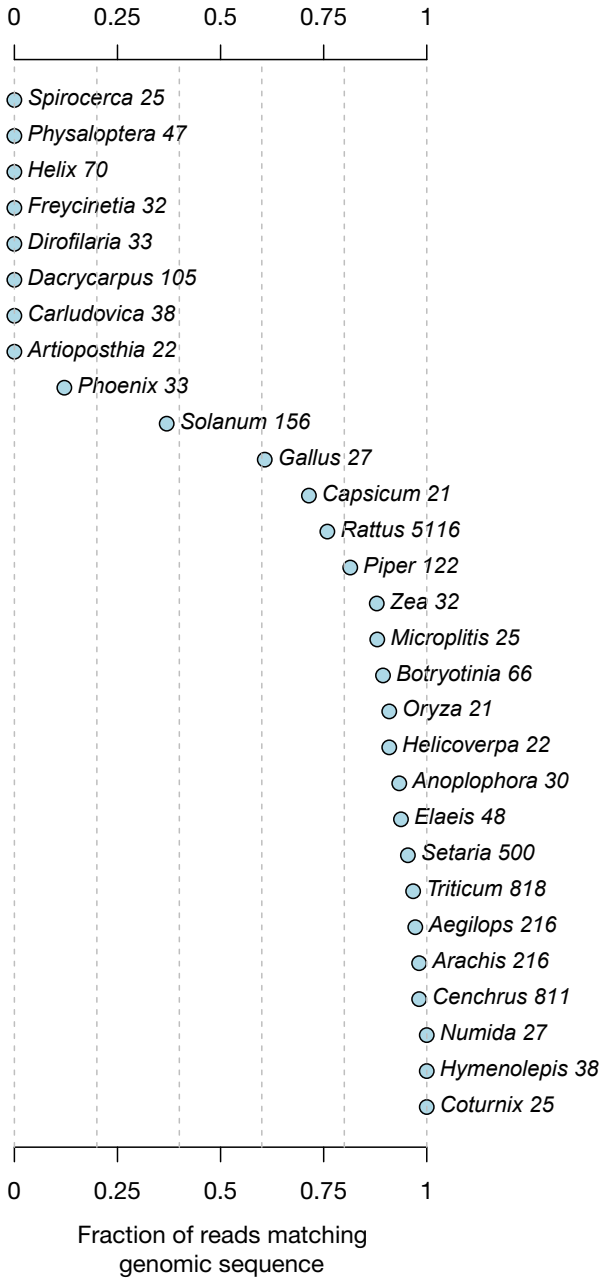

Figure 7. Fractions of reads matching genomic and non-genomic sequence for the best BLAST hit of each read. For the species with complete genomes, the fraction of reads matching genomic sequence ranges from 40% (*Solanum*) to 100%. This large range is likely due to the tissue from which the DNA was isolated. For example, muscle tissue has a higher fraction of mtDNA to total DNA than leaf tissue. For species without fully sequenced genomes, this fraction ranges from 0% to 20% (*Phoenix*, which has a small amount of genomic data present in the database).

the native estuarine bush habitat (OB) had stomach contents containing Arecaceae (palms), while only one in the restored wetland area (LB) did. All three rats that had stomach contents with Phaseanidae (pheasants and quail) were from the native estuarine habitat (OB). All five rats that consumed Solanales (tomatoes, petunias, and allied families) were from the restored wetland area. These patterns suggested that it might be possible to use the read classification results from stomach items to pinpoint the habitat from which each rat was sampled.

To determine if the stomach content of the rats differed consistently between locations, we first performed an unconstrained analysis using non-metric multidimensional scaling (nMDS) on taxa assigned at the family level. The input for the nMDS was the dissimilarity matrix (Bray-Curtis distance of stomach content at the family level). Non-metric multidimensional scaling uses rank-based distances to cluster samples that are most similar.

The family-level unconstrained ordination (nMDS) showed no obvious grouping of rats with respect to the locations (Fig. 8a), indicating that locations did not correspond to the predominant axes of variation among the stomach contents. We next performed a constrained ordination method, Canonical Analysis of Principal coordinates (CAP), which identifies axes of variation, if any, that may distinguish the stomach contents of rats from pre-specified categories (here, different locations) (Fig. 8b). We found that the CAP axes correctly classified the locations of 19 out of 24 (79%) rats using a leave-one-out procedure. The families having the largest correlations with the first two principal coordinates, and thus most responsible for the separation between groups, were primarily plants: Arecaceae (palms), Podocarpaceae (podocarps), Piperaceae (peppers), and Pinaceae (pines). In addition, insect groups (Cerambycids, longhorn beetles and Formicids, ants) and birds (Phaseanidae, pheasants and quails and Numididae, a likely false positive family-level assignment but a taxon in the superfamily Phasianoidea) played a role (Fig. 8c).

The families driving similarity within the three locations (i.e. those that had the greatest within-location SIMPER scores) varied among locations: LB had average Bray-Curtis within-location similarity of 13%; mostly attributable to Hymenolepidae (tapeworms, accounting for 51% of the within-group similarity), Solanaceae (tomatoes, petunias, and allied families, 11%), and Fabaceae (peas, 11%). The average similarity for OB was 21%, with the greatest contributing taxa being Arecaceae (33%), Poaceae (grasses, 23%), Fabaceae (9%), and Phasianidae (8%). The average similarity for WP was 24%, with the greatest contributing taxon being Poaceae (72%) (Appendix S6**)**.

Discussion

Accuracy and sensitivity

Here we have shown that using a simple metagenomic approach with error-prone long reads allows rapid and accurate classification of rat stomach content (approximately 2.7% error in taxon assignments at the family-level). We expect that this technique can be used to infer diet for a wide variety of animal and sample types, including samples that use less invasive collection methods, such as faecal matter. The accuracy of this approach will likely improve as the accuracy and yield of ONT sequencing continues to increase. The analysis here is based on fewer than 200 000 reads from two flow cells. Current yields for similar read length distributions are in excess of five million reads per flow cell. In addition, ONT modal sequencing accuracy is currently just above 99%, and continues to improve. This increase in read accuracy will clearly improve the accuracy of taxon assignment, illustrated by the fact that the fraction of reads yielding BLAST hits increases substantially for high accuracy reads, approaching 40% for high quality reads in our dataset (reads with greater than 92.5% accuracy, Fig. 3b; with current ONT sequencing techniques, 92.5% is at the lower end of read accuracy.

Furthermore, as the species sampling of genomic databases increases (Lewin et al. 2018), the taxon-level precision of this method will improve. Given the current rate of genomic sequencing, with careful sampling, the vast majority of multicellular plant and animal families (and even genera) will likely have at least one type species with a sequenced genome within the next decade. Continued advancement in sequence database search algorithms as compared to current methods (Wood & Salzberg 2014; Kim et al. 2016; Nasko et al. 2018)

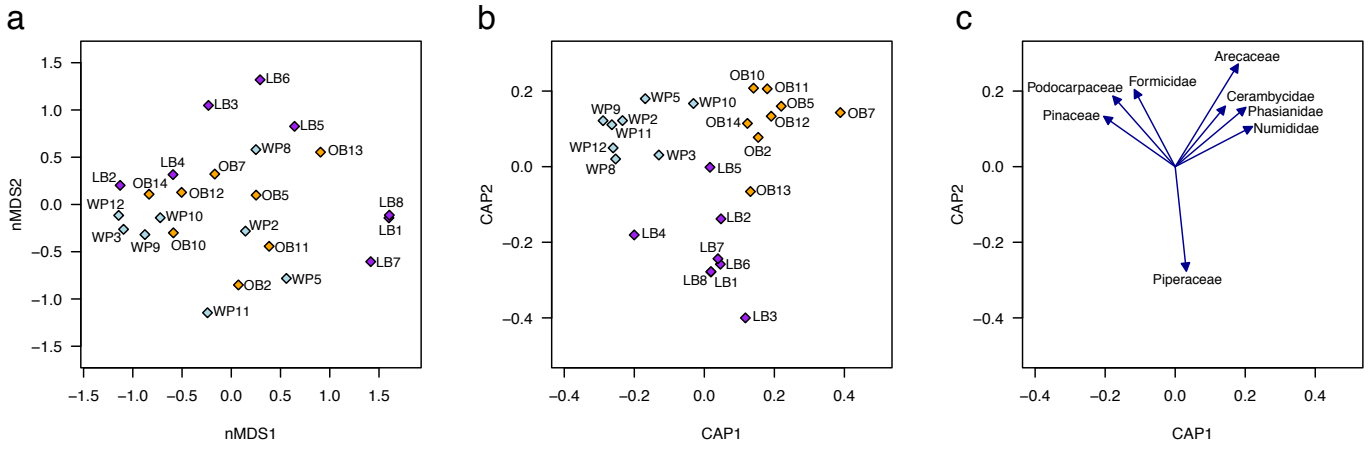

Figure 8. (a) Unconstrained nMDS and (b) constrained CAP ordinations of the stomach contents of rats from three locations. Both ordinations were based on Bray-Curtis dissimilarities of square root transformed proportions of reads attributed to each family. The locations were a native estuarine bush (OB, orange); a restored marine wetland (LB, purple); and a native forest (WP, light blue). (c) The CAP ordination from panel (b) plotted as a biplot with the rats omitted to show the Pearson correlations between families and the first two CAP axes. The eight families with the strongest correlations are shown, indicating the taxa associated with each location.

should considerably decrease the computational workload necessary to find matching sequences.

Methodological advantages

As genomic databases become more complete, metagenomic approaches will offer significant advantages due to decreased bias as compared to other methods. We found that rats consumed many soft-bodied species (e.g. mushrooms, flatworms, slugs, and lepidopterans) that would be difficult to identify using visual inspection of stomach contents. Achieving data on such a wide variety of taxa (across multiple phyla) would also be difficult to quantify using metabarcoding, as there are no universal 18S or COI universal primers capable of amplifying sequences in all these taxa. While it might be possible to use several different primer sets targeted at different phyla or orders, quantitatively comparing diet components across these using sequences amplified with different primer sets is extremely difficult due to differences in primer binding and PCR efficiency.

The ONT-based sequencing method has several unique advantages. Perhaps the most obvious is the accessibility of the platform. Compared to other high throughput sequencing technologies (e.g. Illumina, IonTorrent, or PacBio), for the MinION platform, there is no initial capital investment required. For the P2 PromethION platform, with approximately 10-fold greater output, there is only a $10 000 USD capital investment. Library preparation and sequencing can be extremely cheap and rapid, going from DNA sample to sequence in less than two hours (Zaaijer et al. 2017). Furthermore, the MinION sequencing platform itself is highly portable (the higher output P2 is portable but the computational resources required complicate portability). Given (1) that ONT-based methods are now similar in cost-per-read as the most accessible Illumina method (on the P2 platform we estimate $1500 USD for up to 50 million reads using ONT, versus $1300 USD for 20 million reads using MiSeq); and (2) that even marginal increases in read length are likely to significantly improve species identification, we expect that ONT-based methods should soon become useful for routine ecological monitoring of species (Kamenova et al. 2017).

Methodological challenges

A significant problem in this approach is determining the taxonomic level at which organismal identification becomes inaccurate. In some cases, clear false positives or contaminants are apparent. For example, some taxa may live within the eco-region, but are highly unlikely to have been consumed (e.g. octopus, Appendix S6). In other cases, the reverse might be true - items could have been consumed but do not usually live within the eco-region (e.g. scorpions). One more nuanced example in the data presented here are reads classified into the family *Numididae* (guinea fowl). These are not native to New Zealand, nor are they known to exist in substantial numbers. However, *Numididae* sit in the superfamily *Phasionoidea* and the order Galliformes, which contains at least two species common to New Zealand (although not native), the common pheasant and the California quail.

We note that in the case of diet, it is very often true that within a single rat, reads from a single taxon will all have arisen from a single diet item. Thus, one method of disentangling false positive read assignments from true positives is that, largely speaking, taxon assignments within a sample should be the same. Reads from only three rats were assigned to the order Galliformes (Fig 6b). In one of these rats, seven of seven reads were assigned to the family *Phasianidae*. In the other two, 25 of 37 and four of seven reads were assigned to *Phasianidae*, with the remainder assigned to *Numididae*. This suggests that the assignment to Numididae was a false positive assignment, as the Numididae reads are not from a single rat that may have consumed Numididae, but spread across two rats. It is unlikely that both rats consumed both Phasianidae and Numididae. It is more likely that both consumed only Phasianidae, and a fraction of reads were falsely assigned to Numididae. Frequently, such detailed examination may help to distentangle false positive and true positive assignments.

In addition, some modifications to our approach might further increase the precision of our ability to accurately infer community composition. Any error-prone long read dataset (i.e. PacBio or ONT) has both short (e.g. 500 bp) and long (e.g. 5000 bp) reads, as well as high quality (e.g. mean accuracy greater than 98%) and low-quality reads. When inferring community composition, a null expectation is that true positive taxa should be equally represented by long, high quality reads as they are by short, low-quality reads. If some taxa are represented only by short, low-quality reads, this suggests that these taxa may be false positive inferences. In fact, the difficulty in correctly mapping short inaccurate reads could be mitigated by weighting the probability of taxon presence by the number of long, accurate reads that map to certain taxa. Thus, the fact that not all reads are extremely long and accurate does not mean that they cannot all be used to infer taxon presence in metagenomic analyses.

It is critical to note that for many diet studies, the aim is to resolve biomass, nutritional content, or prey numbers. However, estimating these numbers is constrained by the fact that different tissues and different taxa have different amounts of DNA (both nuclear and mitochondrial) per gram of biomass. It is nearly impossible to fully account for this variation using any DNA-based method. Regardless, there is considerable utility in using DNA-based approaches for diet assessment, not least because it is one of the few methods that allows the full breadth of the diet to be observed, as illustrated here by the number of different orders we find.

It is also difficult to determine to what extent the quantitative nature of DNA-based methods is influenced by timing; DNA from some diet items may exist at low levels not because less was consumed, but because these items were consumed in the past. However, we expect that the majority of DNA will come from items that are currently in the stomach; the amount of DNA from previous items should be vanishingly small. More problematically, some types of diet items may be more prone to remaining in the stomach, especially those that are difficult to digest (e.g. plants). This emphasises that DNA-based approaches may best be used to quantitatively compare diet across sampled individuals, rather than quantify diet fractions within samples.

Here we have shown that using a rapid error-prone long read metagenomic approach we can accurately characterise stomach content taxa at the family level, and distinguish between the stomach contents of rats according to the locations from which they were sourced. This information may be used to guide conservation efforts toward specific areas and habitats in which native species are most at risk from this highly destructive introduced predator.

Acknowledgements

Thanks to W. Pearman for assisting with DNA isolation and compiling the BLAST database; and Friends of Okura Bush, Mary Stewart from Auckland Council, and Gillian Wadams and the volunteers at the Waitākere Ranges for collecting rat samples and aiding in rat species identification.

Additional information and declarations

Author contributions: JD, NF, and OS conceived the project. NF optimised the genomic DNA isolation and library preparation. NF performed the nanopore sequencing. GB and OS processed and performed quality control on the sequencing data. OS performed the sequence classification. ANHS, NF, and OS analysed the data. NF, ANHS, and OS wrote the paper, with input from all authors.

Funding: This work was supported by a Massey University Research Fund to NF, a Marsden Fund Grant (15-MAU-136) to JD and Marsden Fund Grant (MAU1703) to OS.

Data and code availability: Sequence data are available in the SRA archive (accession number PRJEB27647). The code for analysis and generation of figures is available from https://github.com/osilander/rat-diet-analysis.

Ethics: Sample collection was performed under (Auckland Council Permit to Undertake Research WS1064).

Conflicts of interest: The authors report no conflicts of interest.

References

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature Communications 7: 13219.

Anderson MJ, Willis TJ 2003. Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. Ecology 84(2): 511–525.

Arribas P, Andújar C, Hopkins K, Shepherd M, Vogler AP 2016. Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. Methods in Ecology and Evolution 7(9): 1071–1081.

Basha WA, Chamberlain AT, Zaki ME, Kandeel WA, Fares NH 2016. Diet reconstruction through stable isotope analysis of ancient mummified soft tissues from Kulubnarti (Sudanese Nubia). Journal of Archaeological Science: Reports 5: 71–79.

Bridgman LJ, Innes J, Gillies C, Fitzgerald N, King CM 2013. Do ship rats display predatory behaviour towards house mice? Animal Behaviour 86(2): 257–268.

Brown KP, Moller H, Innes J, Jansen P 2008. Identifying predators at nests of small birds in a New Zealand forest. The Ibis 140(2): 274–279.

Brownlee A, Moss W 1961. The influence of diet on lactobacilli in the stomach of the rat. The Journal of Pathology 82(2): 513–516.

Carreon-Martinez L, Heath DD 2010. Revolution in food web analysis and trophic ecology: diet analysis by DNA and stable isotope analysis. Molecular Ecology 19(1): 25–27.

Clarke KR 1993. Non-parametric multivariate analyses of changes in community structure. Austral Ecology 18: 117–143.

Clarke KR, Gorley RN 2015. PRIMER v7: User manual/tutorial. Plymouth, PRIMER-E. 296 p.

Clarke KR, Green RH 1988. Statistical design and analysis for a 'biological effects' study. Marine Ecology Progress Series 46: 213–226.

Clarke KR, Robert Clarke K, Somerfield PJ, Gee Chapman M 2006. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. Journal of Experimental Marine Biology and Ecology 330: 55–80.

Daniel MJ 1973. Seasonal diet of the ship rat (*Rattus rattus*) in lowland forest in New Zealand. Proceedings of the New Zealand Ecological Society 20: 21–30.

Deagle BE, Eveson JP, Jarman SN 2006. Quantification of damage in DNA recovered from highly degraded samples--a case study on DNA in faeces. Frontiers in Zoology 3: 11.

Diamond JM, Veitch CR 1981. Extinctions and introductions in the New Zealand avifauna: cause and effect? Science 211(4481): 499–501.

Dowding JE, Murphy EC 2001. The impact of predation by introduced mammals on endemic shorebirds in New Zealand: a conservation perspective. Biological Conservation 99: 47–64.

Dunlap M, Pawlik JR 1996. Video-monitored predation by Caribbean reef fishes on an array of mangrove and reef sponges. Marine Biology 126: 117–123.

Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proceedings of the National Academy of Sciences of the United States of America 109(52): 21390–21395.

Gibbs GW 1998. Why are some weta (Orthoptera: Stenopelmatidae) vulnerable yet others are common? Journal of Insect Conservation 2(3-4): 161–166.

Graham NAJ, Wilson SK, Carr P, Hoey AS, Jennings S, MacNeil MA 2018. Seabirds enhance coral reef productivity and functioning in the absence of invasive rats. Nature 559(7713): 250–253.

Hobson KA 1987. Use of stable-carbon isotope analysis to estimate marine and terrestrial protein content in gull diets. Canadian Journal of Zoology 65(5): 1210–1213.

Horáková Z, Zierdt CH, Beaven MA 1971. Identification of lactobacillus as the source of bacterial histidine decarboxylase in rat stomach. European Journal of Pharmacology 16(1): 67–77.

Hover BM, Kim S-H, Katz M, Charlop-Powers Z, Owen JG, Ternei MA, Maniko J, Estrela AB, Molina H, Park S, Perlin DS, Brady SF 2018. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. Nature Microbiology 3(4): 415–422.

Huson DH, Auch AF, Qi J, Schuster SC 2007. MEGAN analysis of metagenomic data. Genome Research 17(3): 377–386.

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh H-J, Tappu R 2016. MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Computational Biology 12(6): e1004957.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA,

Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology 36(4): 338–345.

Jarman SN, Gales NJ, Tierney M, Gill PC, Elliott NG 2002. A DNA-based method for identification of krill species and its application to analysing the diet of marine vertebrate predators. Molecular Ecology 11(12): 2679–2690.

Jarman SN, Deagle BE, Gales NJ 2004. Group-specific polymerase chain reaction for DNA-based analysis of species diversity and identity in dietary samples. Molecular Ecology 13(5): 1313–1322.

Kamenova S, Bartley TJ, Bohan DA, Boutain JR, Colautti RI, Domaizon I, Fontaine C, Lemainque A, Le Viol I, Mollot G, Perga M-E, Ravigné V, Massol F 2017. Invasions Toolkit: Current methods for tracking the spread and impact of invasive species. Advances in Ecological Research 56: 85–182.

Kim D, Song L, Breitwieser FP, Salzberg SL 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Research 26(12): 1721–1729.

King RA, Read DS, Traugott M, Symondson WOC 2008. Molecular analysis of predation: a review of best practice for DNA-based approaches. Molecular Ecology 17(4): 947–963.

Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10: 34.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, Goldstein MM, Grigoriev IV, Hackett KJ, Haussler D, Jarvis ED, Johnson WE, Patrinos A, Richards S, Castilla-Rubio JC, van Sluys M-A, Soltis PS, Xu X, Yang H, Zhang G 2018. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences of the United States of America 115(17): 4325–4333.

Li D, Chen H, Mao B, Yang Q, Zhao J, Gu Z, Zhang H, Chen YQ, Chen W 2017. Microbial biogeography and core microbiota of the rat digestive tract. Scientific Reports 8: 45840.

Major HL, Jones IL, Charette MR, Diamond AW 2007. Variations in the diet of introduced Norway rats (*Rattus norvegicus*) inferred using stable isotope analysis. Journal of Zoology 271(4): 463–468.

Maurice CF, Knowles SCL, Ladau J, Pollard KS, Fenton A, Pedersen AB, Turnbaugh PJ 2015. Marked seasonal variation in the wild mouse gut microbiota. The ISME Journal 9(11): 2423–2434.

McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, Minot SS, Danko D, Foox J, Ahsanuddin S, Tighe S, Hasan NA, Subramanian P, Moffat K, Levy S, Lonardi S, Greenfield N, Colwell RR, Rosen GL, Mason CE 2017. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biology 18: 182.

Nasko DJ, Koren S, Phillippy AM, Treangen TJ 2018. RefSeq database growth influences the accuracy of k-mer-based species identification. Genome Biology 19: 165.

Paula DP, Linard B, Crampton-Platt A, Srivathsan A, Timmermans MJTN, Sujii ER, Pires CSS, Souza LM, Andow DA, Vogler AP 2016. Uncovering trophic interactions in arthropod predators through DNA shotgun-sequencing of gut contents. PloS One 11(9): e0161841.

Pawluczyk M, Weiss J, Links MG, Egaña Aranguren M, Wilkinson MD, Egea-Cortines M 2015. Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. Analytical and Bioanalytical Chemistry 407(7): 1841–1848.

Pearman WS, Freed NE, Silander OK 2020. Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. BMC Bioinformatics 21(1): 220.

Pereira RPA, Peplies J, Brettar I, Hoefle MG 2018. Impact of DNA polymerase choice on assessment of bacterial communities by a *Legionella* genus-specific next-generation sequencing approach. bioRxiv247445 p.

Pierce GJ, Boyle 1991. A review of methods for diet analysis in piscivorous marine mammals. Oceanography and Marine Biology: An Annual Review 29: 409–486.

Riofrío-Lazo M, Páez-Rosas D 2015. Feeding habits of introduced black rats, *Rattus rattus*, in nesting colonies of Galapagos petrel on San Cristóbal Island, Galapagos. PloS One 10(5): e0127901.

Russell JC, Innes JG, Brown PH, Byrom AE 2015. Predator-free New Zealand: Conservation country. Bioscience 65(5): 520–525.

Søe MJ, Nejsum P, Seersholm FV, Fredensborg BL, Habraken R, Haase K, Hald MM, Simonsen R, Højlund F, Blanke L, Merkyte I, Willerslev E, Kapel CMO 2018. Ancient DNA from latrines in Northern Europe and the Middle East (500 BC–1700 AD) reveals past parasites and diet. PloS One 13(4): e0195481.

Soininen EM, Valentini A, Coissac E, Miquel C, Gielly L, Brochmann C, Brysting AK, Sønstebø JH, Ims RA, Yoccoz NG, 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: 16.

Srivathsan A, Sha JCM, Vogler AP, 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.

Srivathsan A, Ang A, Vogler AP, Meier R 2016. Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. Frontiers in Zoology 13: 17.

Stringer IAN, Bassett SM, McLean MJ, McCartney J, Parrish GR 2003. Biology and conservation of the rare New Zealand land snail *Paryphanta busbyi watti* (Mollusca, Pulmonata). Invertebrate Biology 122(3): 241–251.

Sweetapple PJ, Nugent G 2007. Ship rat demography and diet following possum control in a mixed podocarp—hardwood forest. New Zealand Journal of Ecology 31(2): 186–201.

Tedersoo L, Anslan S, Bahram M, Põlme S, Riit T, Liiv I, Kõljalg U, Kisand V, Nilsson H, Hildebrand F, Others 2015. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MycoKeys 10: 1.

Towns DR, Daugherty CH, Cree A 2001. Raising the prospects for a forgotten fauna: a review of 10 years of conservation effort for New Zealand reptiles. Biological Conservation 99: 3–16.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978): 37–43.

Wood DE, Salzberg SL 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15(3): R46.

Xu Z, Knight R 2015. Dietary effects on human gut microbiome diversity. The British Journal of Nutrition 113 Suppl: S1–5.

Zaaijer S, Gordon A, Speyer D, Piccone R, Groen SC, Erlich Y 2017. Rapid re-identification of human samples using portable DNA sequencing. eLife 6.

Received: 11 August 2022; Accepted: 8 September 2023
Editorial board member: Natalie Forsdick

Supplementary material

Additional supporting information may be found in the supplementary material file for this article:

Appendix S1. Read numbers and total base pairs for each barcode in the January 2017 sequencing run.

Appendix S2. Read numbers and total base pairs for each barcode in the March 2017 sequencing run.

Appendix S3. Characteristics of alignments for reads assigned to the Primate family.

Appendix S4. SIMPER analysis of family contributions to group similarities.

Appendix S5. SIMPER analysis of family contributions to group dissimilarities.

Appendix S6. Characteristics of alignments for reads that are likely false positive assignments.

Appendix S7. Correlation of read accuracy with alignment characteristics.

Appendix S8. Alignment characteristics of true positive and false positive taxon assignments at the family level.

Appendix S9. Table of read BLAST hits and assigned MEGAN taxa with reads reclassified at the family or order level by filtering on read length to alignment length ratio and percent identity.

Appendix S10. Table of read BLAST hits and assigned MEGAN taxa with no filters applied.

The New Zealand Journal of Ecology provides supporting information supplied by the authors where this may assist readers. Such materials are peer-reviewed and copy-edited but any issues relating to this information (other than missing files) should be addressed to the authors.