1. Title: New tools for diet analysis: nanopore sequencing of metagenomic DNA from rat stomach contents to quantify diet

2. Authors: William S. Pearman, Adam N. H. Smith, Georgia Breckell, James Dale, Nikki E. Freed, Olin K. Silander

3. Addresses of all authors: Institute of Natural and Mathematical Sciences, Massey University, Auckland 0745, New Zealand

4. Keywords: Metagenomics, nanopore sequencing, diet analysis, rat, minion

5. Communicating authors: Olin K. Silander, Institute of Natural and Mathematical Sciences, Massey University, Auckland 0745, New Zealand, olinsilander@gmail.com, +64 9 213 6618; Nikki E. Freed, Institute of Natural and Mathematical Sciences, Massey University, Auckland 0745, New Zealand, freednikki@gmail.com, +64 9 213 6639

6. Running title: Quantifying rat diets by nanopore sequencing

Abstract

Using metagenomics to determine animal diet offers a new and promising alternative to current methods. Here we show that rapid and inexpensive diet quantification is possible through metagenomic sequencing with the portable Oxford Nanopore MinION. Using a simple amplification-free approach, we profiled the stomach contents from wild-caught rats. We conservatively identified diet items from over 50 taxonomic orders, ranging across nine phyla that include plants, vertebrates, invertebrates, and fungi. This highlights the wide range of taxa that can be identified using this simple approach. We calibrate the accuracy of this method by comparing the characteristics of reads matching the ground-truth host genome (rat) to those matching diet items. We also suggest a means to correct for biases in metagenomic approaches that arise due to the paucity of genomic sequence in databases as compared to mitochondrial DNA or rDNA. Finally, we implement a constrained ordination analysis to show that it is possible to identify the sampling location of an individual rat within tens of kilometres based on diet content alone. This work establishes long-read metagenomic methods as a straightforward and robust approach for diet quantification. It considerably simplifies the workflow and avoids many inherent biases as compared to metabarcoding. Continued increases in the accuracy and throughput of Nanopore sequencing, along with improved genomic databases, means that this approach will continue to improve in accuracy.

Introduction

Bias in current methods

Accurate information about what organisms are eating informs many aspects of our understanding of ecosystems and food web dynamics, however unbiased and sensitive assessment of diet content is extremely difficult to achieve due to the limited accuracy of available methods. A variety of methods have been applied to quantify diet components in animals, including visual inspection of gut contents (Daniel, 1973; Pierce & Boyle, 1991) stable isotope analysis (Carreon-Martinez & Heath, 2010; Major, Jones, Charette, & Diamond, 2007), and time-lapse video (Brown, Moller, Innes, & Jansen, 2008; Dunlap & Pawlik, 1996). However, these methods can be biased and imprecise. Identification of prey items using visual examination of stomach contents is strongly affected by which items are most easily degraded (for example, soft-bodied species). Stable isotope analysis yields only broad information on diet such as relative consumption of protein and plant matter, as well as information on whether prey items are terrestrial or marine in origin (Basha, Chamberlain, Zaki, Kandeel, & Fares, 2016; Hobson, 1987). Time-lapse video (Dunlap & Pawlik, 1996; Volpov et al., 2015) requires identification of the specific prey item, often difficult or impossible for small prey items or in low-light conditions. To circumvent these issues, DNA-based methods (King, Read, Traugott, & Symondson, 2008; Soininen et al., 2009) are becoming more 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, this method can detect a wide range of species, and does not require specific expertise necessary for other methods (for example identifying degraded prey items).

However, DNA metabarcoding is not free from bias. PCR primers must be specifically tailored to particular sets of taxa or species (Jarman, Gales, Tierney, Gill, & Elliott, 2002). Although more “universal” PCR primer pairs have been developed (for example targeting all bilaterians or even all eukaryotes; (Jarman, Deagle, & Gales, 2004), all primer sets exhibit bias towards certain taxa. Tedersoo et al. (2015) (Tedersoo et al., 2015) found five-fold differences in fungal operational taxonomic units (OTU) estimates when using different sets of fungal-specific PCR primer pairs. Leray et al. (2013) (Leray et al., 2013) found that published universal primer pairs (i.e. those that do not target specific taxa) were capable of amplifying only between 57% and 91% of tested metazoan species, with as few as 33% of species in some phyla being amplified at all (e.g. cnidarians). Deagle et al. (2014) argued that in general, COI regions are simply not sufficiently conserved, and thus should not be used for metabarcoding studies at all (Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). Finally, Pawluczyk et al. (2015) showed that different loci from the same species exhibit up to 2,000-fold differences in qPCR-estimated DNA quantity within samples (Pawluczyk et al., 2015). It has even been shown that the polymerase itself can bias diversity metrics when using metabarcoding methods (Pereira, Peplies, Brettar, & Hoefle, 2018). For these reasons, a less biased method is desirable.

Metagenomic sequencing for diet

Metagenomic sequencing, in which all of the DNA in the sample is directly sequenced, offers an attractive alternative to metabarcoding for several reasons. Metagenomic approaches have most frequently been used to yield insights into microbial diversity and function (Anantharaman et al., 2016; Fierer et al., 2012; Hover et al., 2018; Xu & Knight, 2015). Recent advances in computational methods (Breitwieser & Salzberg, 2018; Huson, Mitra, Ruscheweyh, Weber, & Schuster, 2011; Kim, Song, Breitwieser, & Salzberg, 2016; Wood & Salzberg, 2014) now allow routine rapid quantification of microbial taxa in metagenomic samples. However, metagenomic approaches have rarely been used to quantify eukaryotic taxa. An important application of such a method would be for diet analysis, as many diet items are difficult to identify based on macro- or microscopic analysis.

Here, we quantify rat diet composition using a novel metagenomic approach based on long-read nanopore sequencing (Oxford Nanopore Technologies). This study shows for the first time that low-accuracy long-read sequences can be used to accurately classify eukaryotic metagenomic data. As a test case, we quantify rat diet using stomach contents. Using such samples is opportune for both methodological and ecological reasons.

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 (rat-derived reads that match rat database sequences), as well as false negative and false positive reads (rat-derived reads that match non-rat database sequences). We can then determine whether reads matching diet items have similar characteristics to known true positive (host) reads.

Finally, understanding rat diets has important ecological implications. It is well-established that the relatively recent introduction of mammalian predators to New Zealand and other islands has had significant negative effects on many of the native animal populations. This ranges from insects (Gibbs, 1998), to reptiles (Towns, Daugherty, & Cree, 2001), to molluscs (Stringer, Bassett, McLean, McCartney, & Parrish, 2003), to birds (Diamond & Veitch, 1981; Dowding & Murphy, 2001), and can have detrimental effects for entire terrestrial and aquatic ecosystems (Graham et al., 2018). Currently, an ambitious plan is 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, Innes, Brown, & Byrom, 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.

Materials and 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 (Waitakere Regional Parklands, WP); native bush surrounding an estuary (Okura Bush Walkway, OB); and restored coastal wetland (Long Bay Regional Park, LB) (**Fig. 1**). 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. The majority of 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 removed intact stomachs from each animal and removed the contents. 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 10-20 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 improved DNA purity. We rehydrated precipitated DNA by incubating overnight in molecular biology grade water at 4°C, and stored the DNA at -20°C. DNA quantity, purity, and quality was ascertained by nanodrop and agarose gel electrophoresis. The DNA samples were ranked according 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 DNA sequencing.

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. 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) and 1439 active pores (March). Sequencing was performed using local base calling in MinKnow v1.3.25 (January) or MinKnow v1.5.5 (March), but 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 combined NCBI other_genomic and nt databases (downloaded on 13th June 2018 from NCBI). Default blastn parameters were used (gapopen 5, gapextend 2), and only hits with an e-value of 1e-2 or less were saved. Due to the predominance of short indels present in nanopore sequence data, we used an initial set of basecalled data to test 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 $1x10^{-20}$ or lower (corresponding to a bit score of 115 or higher) and an alignment length of 100 base pairs or more. 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 (K. R. 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. The majority of rodent hits were to rat and mouse, resulting from the

rats' own DNA (see below). The majority of the primate hits were to human sequences, which likely resulted from sample contamination.

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 (K. Robert Clarke, Robert Clarke, Somerfield, & Gee Chapman, 2006). 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 (K. R. 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 (K. Robert Clarke et al., 2006).

We used unconstrained ordination--specifically, non-metric multidimensional scaling (nMDS) applied to the dissimilarity matrix--to examine the overall patterns in the diet composition among rats. To assess the degree to which the diet 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* groups of samples (in this case, the groups are 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 (SIMPER; (K. R. Clarke, 1993)) to characterise and distinguish between the locations. This allowed us to identify the families with the greatest percentage contributions to (1) the Bray-Curtis similarities of diets within each location (**Table S3**) and (2) the Bray-Curtis dissimilarities between each pair of locations (**Table S4**).

Results

DNA sequencing and assignment of reads to taxa

After DNA isolation and sequencing, we obtained a total of 82,977 reads from the January run and 96,150 reads from the March run. Median read lengths were 606 bp and 527 bp for the January and March datasets, respectively (**Fig. 2A**). These lengths are considerably shorter than other nanopore sequencing results from both our and others work (Jain, Olsen, Paten, & Akeson, 2016). This is most likely due to degradation of the DNA during digestion in the stomach as well as fragmentation during DNA isolation (Deagle, Eveson, & Jarman, 2006) and sequencing library preparation. The median phred quality scores per read ranged from 7-12 (0.80 - 0.94 accuracy) for both runs (**Fig. S1**). The number of reads per barcoded rat sample varied by 10-fold for January and up to 40-fold in March (**Fig. 2B** and **2C**). This is due mostly to the highly variable quality of DNA in each sample. However, read length and quality were similar for all samples (**Fig. S1**).

To quantify diet contents we first BLASTed all sequences against a combined database of the NCBI nt database (the partially non-redundant nucleotide sequences from all

218 traditional divisions of GenBank excluding genome survey sequence, EST, high-
219 throughput genome, and whole genome shotgun
220 (ftp://ftp.ncbi.nlm.nih.gov/blast/db/README)) and the NCBI other_genomic database
221 (RefSeq chromosome records for non-human organisms
222 (ftp://ftp.ncbi.nlm.nih.gov/blast/db/README)). We used BLAST as it is generally viewed
223 as the gold standard method in metagenomic analyses (McIntyre et al., 2017). Of the
224 133,022 barcoded reads, 30,535 (23%) hit a sequence in the combined nt and
225 other_genomic database at an e-value cutoff of 1e-2.

226 As an initial assessment of the quality of these hits, we examined the alignment lengths
227 and e-values. We found a bimodal distribution of alignment lengths and a highly skewed
228 distribution of e-values (**Fig. 3A**). We hypothesized that many of the short alignments
229 with high e-values were false positives. We thus first filtered this hit set, only retaining
230 BLAST hits with e-values less than 1e-20 and alignments greater than 100 bp. Similar
231 quality filters have been imposed previously (Srivathsan, Sha, Vogler, & Meier, 2015). A
232 total of 22,154 hits passed this filter (**Datafile S1)**. Mean read quality had substantial
233 effects on the likelihood of a read yielding a BLAST hit, with almost 40% of high
234 accuracy read having hits in the March dataset, as compared to 1% of low accuracy hits
235 (**Fig. 3B**).

236 To specifically assign each sequence read to a taxon, we analysed the BLAST results in
237 MEGAN6 (Huson et al., 2016). The algorithm employed in MEGAN6 assigns reads to a
238 most recent common ancestor (MRCA) taxon level. For example, if a read has BLAST
239 hits to five 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 MRCA of those best-hit three species (Huson, Auch, Qi, & Schuster, 2007). If a read matches one species far better than to any other, by definition, the MRCA is that species.

5,334 reads (24%) were not assigned to any taxon by Megan. Of the remainder, 31% were assigned by MEGAN as being bacterial. 55% of these were *Lactobacillus spp*. These results match previous studies on rat stomach microbiomes, which have found lactobacilli to be the dominant taxa (Brownlee & Moss, 1961; Horáková, Zierdt, & Beaven, 1971; Li et al., 2017; Maurice et al., 2015). Plant-associated *Pseudomonas* and *Lactococcus* taxa were also common, at 7% and 6%, respectively.

MEGAN assigned reads to a wide range of eukaryotic taxa. To conservatively infer taxon presence, we first reclassified MEGAN species-level assignments to the level of genus. However, after this, many clear false positive 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). We inferred that the reads assigned to *Rattus* (2,696 reads in total) were true positive genus-level assignments and that the reads assigned to *Mus* (2,798 reads in total) were false positive genus-level assignments (and not true positive *Mus*-derived reads). Although rats are known to prey on mice (Bridgman, Innes, Gillies, Fitzgerald, & King, 2013), if this had occurred, we would expect that (1) the ratio of mouse to rat reads would be higher in the subset of rats that had predated mice; (2) in those same rats, 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 was similar for all rats. In addition, there was no evidence of higher percent identities for *Mus* reads from rats that had higher ratios.

Notably, the mean percent identity values of the best BLAST hits for *Rattus* and *Mus* reads differed substantially, with *Rattus* reads having a median identity of 86.4%, and *Mus* 81.0% (**Fig. 4A**). The mean percent identity for *Rattus* reads corresponds very well to that expected given the mean quality scores of the reads (assuming the true sequence of the read is 100% identical to *Rattus*, 86.4% identity corresponds to a mean quality score of 8.7; **Fig. S2A-C**). There was also a clear difference in the alignment lengths: the median ratio of alignment length to read length was 0.57 for *Rattus* and 0.52 for *Mus* (**Fig. 4B**). We note that read identity and the ratio of alignment length to read length are positively correlated (**Fig. S2G-I**). There is little correlation between read identity and alignment length alone (**Fig. S2D-F**).

Importantly, the majority of diet items have percent identities that overlap with the *Rattus* reads, and alignment length to read length ratios that often exceed the *Rattus* reads. This suggests that many diet taxa assignments are correct down to the level of genus (as the *Rattus*-assigned reads are correct to the level of genus). However, to further decrease false positive taxon assignments of diet items, we implemented cut-offs based on the characteristics of the *Mus*- and *Rattus*-assigned reads. For genus-level assignment, we required at least 82.5% identity and an alignment length to read length ratio of at least 0.55. These cutoffs exclude 88% of the reads falsely assigned to *Mus*, instead assigning them correctly to one taxon level higher, the Family *Muridae*.

283 For family-level assignments, we required 77.5% identity, an alignment length to read
284 length ratio of at least 0.1, and a total alignment length of at least 150 bp. Using higher
285 cutoffs for the ratio of alignment length to read length excluded a large number of likely
286 true positive taxa for which only short mtDNA or rDNA database sequences were
287 present in the databases. For all other read-to-taxon assignments, we placed the read
288 at the level of Order, or used the taxon level assigned by MEGAN. Using these cutoffs,
289 16% of all reads were classified at the Genus level; 71% were classified at the Family-
290 level or below; 89% were classified at the Order-level or below; and 98% were classified
291 at the Phylum-level or below.

292 After filtering out bacterial, host, and contaminant reads (matching primate DNA), 4,719
293 reads remained (28% of all classified reads) (**Datafile S2**). Within these, we observed
294 that a small number of likely false positive taxa remained. Most were single reads with
295 short alignments: *Poeciliidae* (177 bp); *Salmonidae* (172 bp); *Cyprinodontiformes* (140
296 bp and 177 bp); and *Octopodidae* (151 bp). The exception to this were three reads from
297 two rats matching *Buthidae* (scorpions), which had alignment lengths of 762 bp, 664 bp,
298 and 298 bp. It is unlikely these are true positives, and instead we hypothesise that these
299 rats predated harvestmen (*Opiliones*), a closely related sister taxon within *Arachnida* but
300 lacking significant amounts of genomic data. Despite the presence of these false
301 positive taxa, we did not further increase the stringency of our filters, allowing us to
302 resolve most taxa at the level of family, with a small rate of false positive inference
303 (here, eight clear instances out of almost 5,000 reads).

Identification of diet

Within each rat, a wide variety of plant, animal, and fungal orders were discernible, 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**). Plants were the primary diet item, with the largest fraction of rats consuming four predominant orders: *Poales* (grasses), *Fabales* (legumes), *Arecales* (palms), and *Araucariales* (podocarps). The dominance of plant matter (fruits and seeds) in rat diets has been established previously (Riofrío-Lazo & Páez-Rosas, 2015; Sweetapple & Nugent, 2007). Animal taxa made up a smaller component of each rat's diet, with *Insecta* dominating: *Hymenoptera*, *Coleoptera*, *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' diet, although several orders were present: *Sclerotiniales* (plant pathogens), *Saccharomycetales* (budding yeasts), *Mucorales* (pin molds), *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)).

Due to our metagenomic approach, the fraction of each element of the rats' diets is distorted by biases in genomic databases: whole genome data exists for only a few taxa, while mtDNA and rDNA sequence data are present in the database for the vast majority of animal and plant genera. To quantify this bias, we determined the fraction of hits that mapped to non-genomic database sequences relative to the fraction of hits that mapped

to genomic DNA. By quantifying this fraction for species with complete genome sequences in the database and species without complete genomes we aimed to assess the effects of this bias.

For the majority of animals with sequenced genomes in the database, we found that the fraction of reads that mapped genomic sequence ranged from 61% (*Gallus*) to 73% (*Rattus*) to 100% (*Coturnix* and *Numida*) (**Fig. 7**). We hypothesise that this variation is likely due to the type of tissue sequenced. For *Rattus* the sequenced tissue was primarily stomach muscle, which has a relatively high fraction of mtDNA; for *Coturnix* and *Numida* it may have been eggs. For plants with sequenced genomes, the fraction of reads matching genomic sequence was generally higher: between 88% (*Zea*) and 98% (*Cenchrus*).

In contrast, for genera with little or no genomic sequence in the database, the vast majority of matches were solely to mtDNA, rDNA, or microsatellite loci: 90% of *Phoenix* (date palm) hits; all *Helix* (snail); and all *Rhaphidophora* **(**cave weta) hits. All *Artioposthia* (New Zealand flatworm) hits were to rDNA. These results indicate that for genera with no genomic sequence data, we have underestimated the actual number of sequences from that taxon by approximately three- to twenty-fold (for animals and plants, respectively). It is difficult to determine how these numbers correlate with biomass.

Close examination of the sequence classification data suggested that specific families (and orders) were overrepresented in the diets of rats from particular locations. For example, six out of eight rats from the native estuarine bush habitat (OB) consumed *Arecaceae*, while only one in the restored wetland area (LB) did. All three rats that

consumed *Phaseanidae* were from the native estuarine habitat (OB). All five rats that consumed *Solanales* were from the restored wetland area. These patterns suggested that it might be possible to use diet components alone to pinpoint the habitat from which each rat was sampled.

nMDS and CAP analysis by location

In order to determine if diet composition of the rats differed consistently between locations, we first performed an unconstrained analysis using nMDS on taxa assigned at the family level. Using family rather than order or genus provides a balance between how precisely we identify the taxon of diet item (genus, family, order), and whether we assign a taxon at all. While family-level assignments are less precise than genus-level, only 16% of all reads were classified at the genus level, while 71% were classified at the family level.

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 diets. However, a constrained ordination analysis (CAP) identified axes of variation that distinguished the diets of rats from 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 most responsible for the separation between groups, were primarily plants: *Arecaceae*,

368 *Podocarpaceae*, *Piperaceae*, and *Pinaceae*. In addition, insect groups (*Cerambycids*
369 and *Formicids*) and birds (*Phaseanidae* and *Numididae*) played a role (**Fig. 8c**).

370 The families driving similarity within the three locations (i.e., had the greatest within-
371 location SIMPER scores) varied among locations. LB had average Bray-Curtis within-
372 location similarity of 13%; mostly attributable to *Hymenolepidae* (accounting for 51% of
373 the within-group similarity), *Solanaceae* (11%), and *Fabaceae* (11%). The average
374 similarity for OB was 21%, with the greatest contributing taxa being *Arecaceae* (33%),
375 *Poaceae* (23%), *Fabaceae* (9%), and *Phasianidae* (8%). The average similarity for WP
376 was 24%, with the greatest contributing taxa being *Poaceae* (72%) (**Table S4**).

377 # Discussion

378 ## Accuracy and sensitivity

379 Here we have shown that using a simple metagenomic approach with error-prone long
380 reads allows rapid and accurate classification of rat diet components. We expect that
381 this technique can be used to infer diet for a wide variety of animal and sample types,
382 including samples that use less invasive collection methods, such as fecal matter. The
383 sensitivity of this approach will likely improve as the accuracy and yield of Oxford
384 Nanopore sequencing increases. The analysis here is based on less than 200,000 reads
385 from two flow cells. The rapid improvement of this technology is such that current
386 yields are often far in excess of two million reads per flow cell. The method will also

387 improve as the diversity of taxa in genomic sequence databases increases. Several
388 aspects of the data support this.

389 First, we note that we did not find BLAST hits for the majority of reads. This is partially
390 due the relatively low accuracy of the Oxford Nanopore sequencing platform at the time
391 these data were collected (approximately 87%). However, the fraction of reads yielding
392 hits in the database increased substantially for higher quality reads, approaching 40%
393 for very high quality reads (**Fig. 3b)**. Other factors also likely reduce the numbers of
394 BLAST hits, such as the paucity of genome sequence data for many taxa. This is
395 convincingly illustrated by comparing across taxa the fraction of genomic hits to
396 mitochondrial or rDNA sequence hits.

397 As the species sampling of genomic databases increases (Lewin et al., 2018), the
398 taxon-level precision of this method will improve. Given the current rate of genomic
399 sequencing, with careful sampling, the vast majority of multicellular plant and animal
400 families (and even genera) will likely have at least one type species with a sequenced
401 genome within the next decade. Continued advancement in sequence database search
402 algorithms as compared to current methods (Kim et al., 2016; Nasko, Koren, Phillippy,
403 & Treangen, 2018; Wood & Salzberg, 2014) should considerably decrease the
404 computational workload necessary to find matching sequences.

405 Although metagenomic approaches decrease the bias arising from PCR amplification of
406 specific DNA regions, additional biases can arise, as the presence or absence of
407 species and genera can only be inferred for those species or genera present in

genomic databases. Although this is similarly true for metabarcoding approaches, metabarcode databases are rapidly becoming more comprehensive in terms of species representation as compared to genomic databases. Importantly, genomic sequence databases are rapidly increasing in species diversity, as are the methods to query these large databases(Kim et al., 2016; Wood & Salzberg, 2014)

To decrease biases in genomic databases, some previous studies have performed metagenomic classification using mitogenome data alone. Using such methods, Srivathsan et al and Paula et al. (2016) (Srivathsan, Ang, Vogler, & Meier, 2016); (Paula et al., 2016) found between 0.004% and 0.008% of all metagenomic reads matched mitogenomes from diet taxa. Limiting database searches to mitogenomes partially ameliorates biases in terms of taxon field in terms of taxon representation (i.e. most taxa will have similar levels of genomic representation in the databases). However, it considerably decreases diet resolution given that for some taxa, only a small percentage of sequence reads derive from the mitochondria as opposed to the nuclear genome.

It is also important to note that our interest in diet also includes resolving relative biomass and relative numbers of each prey species, neither of which necessarily correlate well with the amount of DNA (either mitochondrial or nuclear) purified from a sample. Even a simple correction for the fraction of reads matching mitochondrial versus nuclear genomes is difficult, as different plant and animal tissues differ considerably in the relative amounts of mitochondrial versus nuclear DNA (e.g. leaf versus fruit).

Methodological advantages

We found that rats consumed many soft-bodied species (e.g. mushrooms, flat worms, slugs, and lepidopterans) that would be difficult to identify using visual inspection of stomach contents. Achieving data on such a wide variety of taxa would be difficult to quantify using other molecular methods, 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 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 nanopore MinION-based sequencing method used in this simple metagenomic approach has several advantages. Compared to other high throughput sequencing technologies (e.g. Illumina, IonTorrent, or PacBio), there is no initial capital investment required to use the platform. On a per-sample basis, data generation is inexpensive (approximately $150 USD per barcoded sample, and approximately half this price if reagents are purchased in bulk). Library preparation and sequencing can be extremely rapid, going from DNA sample to sequence in less than two hours (Zaaijer et al., 2017). Furthermore, the sequencing platform itself is highly portable. As the cost of nanopore-based sequencing continues to decrease (both per sample and per base pair), it should become possible to use molecular methods for routine ecological monitoring of species presence or absence in field settings, without significant investment in infrastructure (Kamenova et al., 2017). Finally, we suggest that our approach of standardising the read counts by sample, followed by an optional transformation such as square root and

450 dissimilarity-based multivariate ordination, offers a useful analytical pipeline for
451 analysing metagenomic diet-composition data.

452 We note that modifications to our approach might further increase the precision of our
453 ability to infer community composition. Any error-prone long read dataset (i.e. PacBio or
454 ONT) has both short (e.g. 500 bp) and long (e.g. 5000 bp) reads, as well as high quality
455 (e.g. mean accuracy greater than 90%) and low quality (e.g. mean accuracy less than
456 80%) reads. When inferring community composition, a null expectation is that taxa
457 should be equally represented by long, high quality reads as they are by short, low
458 quality reads. If some taxa are represented only by short, low quality reads, this
459 suggests that these taxa may be false positive inferences. Similarly, the difficulty in
460 correctly mapping short inaccurate reads could be mitigated by weighting the probability
461 of taxon mapping by the number of long, accurate reads that map to certain taxa. Thus,
462 the fact that not all reads are extremely long and accurate does not mean that they
463 cannot all be used to infer taxon presence in metagenomic analyses.

464 Conclusion

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

Acknowledgements

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. Thanks to Friends of Okura Bush, Mary Stewart from Auckland Council, and Gillian Wadams and the volunteers at the Waitakere Ranges for collecting rat samples and aiding in rat species identification. Sample collection was performed under (Auckland Council Permit to Undertake Research WS1064).

References

1. Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., … Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*, 13219.
2. Anderson, M. J., & Willis, T. J. (2003). Canonical Analysis Of Principal Coordinates: A Useful Method Of Constrained Ordination For Ecology. *Ecology*, *84*(2), 511–525.
3. Basha, W. A., Chamberlain, A. T., Zaki, M. E., Kandeel, W. A., & Fares, N. H. (2016). Diet reconstruction through stable isotope analysis of ancient mummified soft tissues from Kulubnarti (Sudanese Nubia). *Journal of Archaeological Science: Reports*, *5*, 71–79.
4. Breitwieser, F. P., & Salzberg, S. L. (2018). KrakenHLL: Confident and fast metagenomics classification using unique k-mer counts. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2018/06/06/262956.abstract
5. Bridgman, L. J., Innes, J., Gillies, C., Fitzgerald, N., & King, C. M. (2013). Do ship rats display predatory behaviour towards house mice? *Animal Behaviour*, *86*(2)), 257–268.
6. Brown, K. P., 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.
7. 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.
8. Carreon-Martinez, L., & Heath, D. D. (2010). Revolution in food web analysis and trophic ecology: diet analysis by DNA and stable isotope analysis. *Molecular Ecology*, *19*(1), 25–27.
9. Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in

503 community structure. *Austral Ecology*, *18*(1), 117–143.
504 10. Clarke, K. R., & Gorley, R. N. (2015). *PRIMER v7: User Manual/Tutorial* (p. 296).
505 PRIMER-E, Plymouth.
506 11. Clarke, K. R., & Green, R. H. (1988). Statistical Design and Analysis for a
507 "biological Effects" Study. *Marine Ecology Progress Series*, *46*, 213–226.
508 12. Clarke, K. R., Robert Clarke, K., Somerfield, P. J., & Gee Chapman, M. (2006).
509 On resemblance measures for ecological studies, including taxonomic
510 dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded
511 assemblages. *Journal of Experimental Marine Biology and Ecology*, *330*(1), 55–
512 80.
513 13. Daniel, M. J. (1973). Seasonal Diet Of The Ship Rat (Rattus Rattus) In Lowland
514 Forest In New Zealand. *Proceedings* , *20*, 21–30.
515 14. Deagle, B. E., Eveson, J. P., & Jarman, S. N. (2006). Quantification of damage in
516 DNA recovered from highly degraded samples--a case study on DNA in faeces.
517 *Frontiers in Zoology*, *3*, 11.
518 15. Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014).
519 DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a
520 perfect match. *Biology Letters*, *10*(9). https://doi.org/10.1098/rsbl.2014.0562
521 16. Diamond, J. M., & Veitch, C. R. (1981). Extinctions and introductions in the new
522 zealand avifauna: cause and effect? *Science*, *211*(4481), 499–501.
523 17. Dowding, J. E., & Murphy, E. C. (2001). The impact of predation by introduced
524 mammals on endemic shorebirds in New Zealand: a conservation perspective.
525 *Biological Conservation*, *99*(1), 47–64.
526 18. Dunlap, M., & Pawlik, J. R. (1996). Video-monitored predation by Caribbean reef
527 fishes on an array of mangrove and reef sponges. *Marine Biology*, *126*(1), 117–
528 123.
529 19. Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., …
530 Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial
531 communities and their functional attributes. *Proceedings of the National*
532 *Academy of Sciences of the United States of America*, *109*(52), 21390–21395.
533 20. Gibbs, G. W. (1998). Why are some weta (Orthoptera: Stenopelmatidae)
534 vulnerable yet others are common? *Journal of Insect Conservation*, *2*(3-4), 161–
535 166.
536 21. Graham, N. A. J., Wilson, S. K., Carr, P., Hoey, A. S., Jennings, S., & MacNeil, M.
537 A. (2018). Seabirds enhance coral reef productivity and functioning in the
538 absence of invasive rats. *Nature*, *559*(7713), 250–253.
539 22. Hobson, K. A. (1987). Use of stable-carbon isotope analysis to estimate marine
540 and terrestrial protein content in gull diets. *Canadian Journal of Zoology*, *65*(5),
541 1210–1213.
542 23. Horáková, Z., Zierdt, C. H., & Beaven, M. A. (1971). Identification of lactobacillus
543 as the source of bacterial histidine decarboxylase in rat stomach. *European*
544 *Journal of Pharmacology*, *16*(1), 67–77.
545 24. Hover, B. M., Kim, S.-H., Katz, M., Charlop-Powers, Z., Owen, J. G., Ternei, M.

A., … Brady, S. F. (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.

25. Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386.
26. Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., … Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology*, *12*(6), e1004957.
27. Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., & Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, *21*(9), 1552–1560.
28. Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239.
29. Jarman, S. N., Deagle, B. E., & Gales, N. J. (2004). Group-specific polymerase chain reaction for DNA-based analysis of species diversity and identity in dietary samples. *Molecular Ecology*, *13*(5), 1313–1322.
30. Jarman, S. N., Gales, N. J., Tierney, M., Gill, P. C., & Elliott, N. G. (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.
31. Kamenova, S., Bartley, T. J., Bohan, D. A., Boutain, J. R., Colautti, R. I., Domaizon, I., … Massol, F. (2017). Chapter Three - Invasions Toolkit: Current Methods for Tracking the Spread and Impact of Invasive Species. In D. A. Bohan, A. J. Dumbrell, & F. Massol (Eds.), *Advances in Ecological Research* (Vol. 56, pp. 85–182). Academic Press.
32. 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.
33. King, R. A., Read, D. S., Traugott, M., & Symondson, W. O. C. (2008). Molecular analysis of predation: a review of best practice for DNA-based approaches. *Molecular Ecology*, *17*(4), 947–963.
34. Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … Machida, R. J. (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.
35. Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., … 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.
36. Li, D., Chen, H., Mao, B., Yang, Q., Zhao, J., Gu, Z., … Chen, W. (2017). Microbial Biogeography and Core Microbiota of the Rat Digestive Tract. *Scientific Reports*, *8*, 45840.

37. Major, H. L., Jones, I. L., Charette, M. R., & Diamond, A. W. (2007). Variations in the diet of introduced Norway rats (Rattus norvegicus) inferred using stable isotope analysis. *Journal of Zoology*, *271*(4), 463–468.
38. Maurice, C. F., Knowles, S. C. L., Ladau, J., Pollard, K. S., Fenton, A., Pedersen, A. B., & Turnbaugh, P. J. (2015). Marked seasonal variation in the wild mouse gut microbiota. *The ISME Journal*, *9*(11), 2423–2434.
39. McIntyre, A. B. R., Ounit, R., Afshinnekoo, E., Prill, R. J., Hénaff, E., Alexander, N., … Mason, C. E. (2017). Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, *18*(1), 182.
40. Nasko, D. J., Koren, S., Phillippy, A. M., & Treangen, T. J. (2018). RefSeq database growth influences the accuracy of k-mer-based species identification, 1–21.
41. Paula, D. P., Linard, B., Crampton-Platt, A., Srivathsan, A., Timmermans, M. J. T. N., Sujii, E. R., … Vogler, A. P. (2016). Uncovering Trophic Interactions in Arthropod Predators through DNA Shotgun-Sequencing of Gut Contents. *PloS One*, *11*(9), e0161841.
42. Pawluczyk, M., Weiss, J., Links, M. G., Egaña Aranguren, M., Wilkinson, M. D., & 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.
43. Pereira, R. P. A., Peplies, J., Brettar, I., & Hoefle, M. G. (2018). *Impact of DNA polymerase choice on assessment of bacterial communities by a Legionella genus-specific next-generation sequencing approach*. *bioRxiv*. https://doi.org/10.1101/247445
44. Pierce, G. J., & Boyle. (1991). A review of methods for diet analysis in piscivorous marine mammals. *Oceanography and Marine Biology: An Annual Review*, *29*, 409–486.
45. 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.
46. Russell, J. C., Innes, J. G., Brown, P. H., & Byrom, A. E. (2015). Predator-Free New Zealand: Conservation Country. *Bioscience*, *65*(5), 520–525.
47. Soininen, E. M., Valentini, A., Coissac, E., Miquel, C., Gielly, L., Brochmann, C., … 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.
48. 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*, 17.
49. 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.

50. Stringer, I. A. N., Bassett, S. M., McLean, M. J., McCartney, J., & Parrish, G. R. (2003). Biology and conservation of the rare New Zealand land snail Paryphanta busbyi watti (Mollusca, Pulmonata). *Invertebrate Biology: A Quarterly Journal of the American Microscopical Society and the Division of Invertebrate Zoology/ASZ*, *122*(3), 241–251.
51. Sweetapple, P. J., & 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.
52. Tedersoo, L., Anslan, S., Bahram, M., Põlme, S., Riit, T., Liiv, I., … Others. (2015). Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys*, *10*, 1.
53. Towns, D. R., Daugherty, C. H., & 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*(1), 3–16.
54. Volpov, B. L., Hoskins, A. J., Battaile, B. C., Viviant, M., Wheatley, K. E., Marshall, G., … Arnould, J. P. Y. (2015). Identification of Prey Captures in Australian Fur Seals (Arctocephalus pusillus doriferus) Using Head-Mounted Accelerometers: Field Validation with Animal-Borne Video Cameras. *PloS One*, *10*(6), e0128789.
55. Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46.
56. Xu, Z., & Knight, R. (2015). Dietary effects on human gut microbiome diversity. *The British Journal of Nutrition*, *113 Suppl*, S1–S5.
57. Zaaijer, S., Gordon, A., Speyer, D., Piccone, R., Groen, S. C., & Erlich, Y. (2017). Rapid re-identification of human samples using portable DNA sequencing. *eLife*, *6*. https://doi.org/10.7554/eLife.27798

Data Accessibility

Sequence data are available in the SRA archive (accession number PRJEB27647)

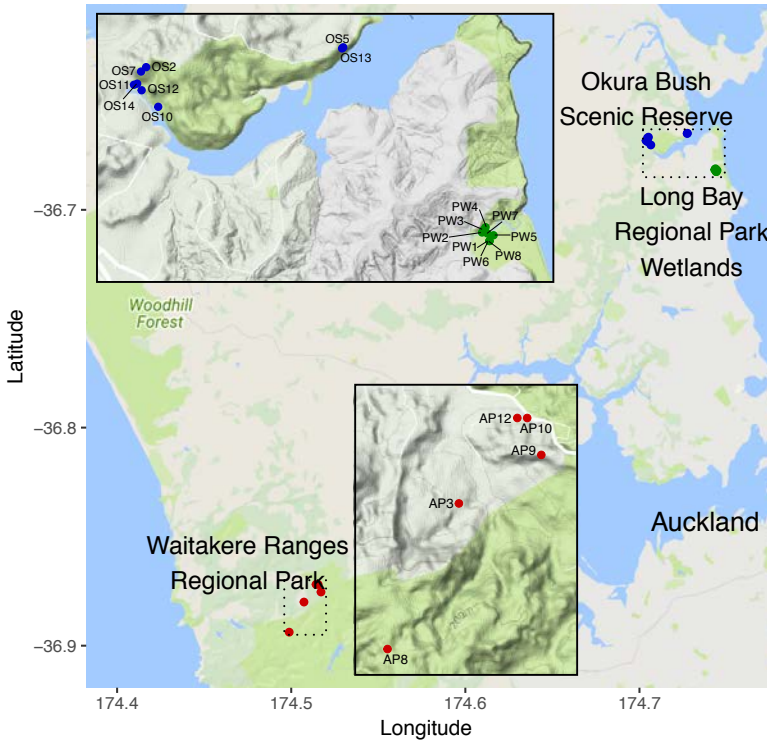
Author Contributions

WP, JD, NF, and OS conceived the project. WP performed the stomach dissections. WP and NF optimised the genomic DNA isolation and library preparation. NF performed the nanopore sequencing. GB and OS processed and performed quality control on the

665 sequencing data. WP and OS performed the sequence classification. WP, AS, NF, and

666 OS analysed the data. WP, NF, AS, and OS wrote the paper, with input from all authors.

667 ## Tables and Figures



668

669 **Fig. 1. Location of rat sampling sites** in the greater Auckland area in the North Island
670 of New Zealand. Each point indicates a trap where one rat was captured, with the
671 colour of the points indicating the three broad locations: the native estuarine bush
672 habitat of Okura Bush (OB), the restored wetland of Long Bay (LB), and the native
673 forest of Waitakere Park (WP). The two insets show the three locations in higher
674 resolution with topographical details. Green indicates park areas. Precise geographical
675 coordinates were only available for five out of eight rats in WP.

676

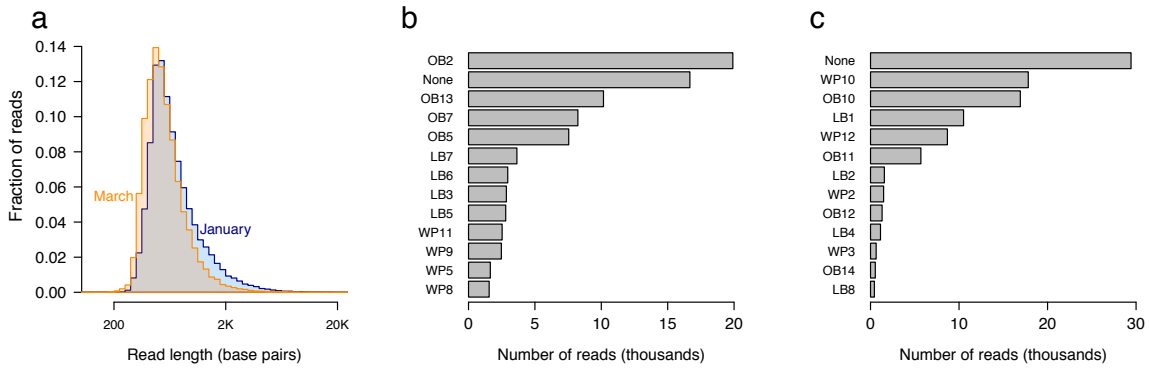


Fig. 2. Results of nanopore metagenomic sequencing of rat stomach contents. (a) Read length distribution for January and March nanopore runs. Read lengths varied between ~300 and 3,000 bp, with a small number greater than 10,000 bp. **(b) and (c) Barcode distributions for January and March runs, respectively.** We multiplexed the samples on the flow cells, using 12 barcodes per flow cell. The distribution of read numbers across barcodes was quite uneven, varying by up to 40-fold in some cases. 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.

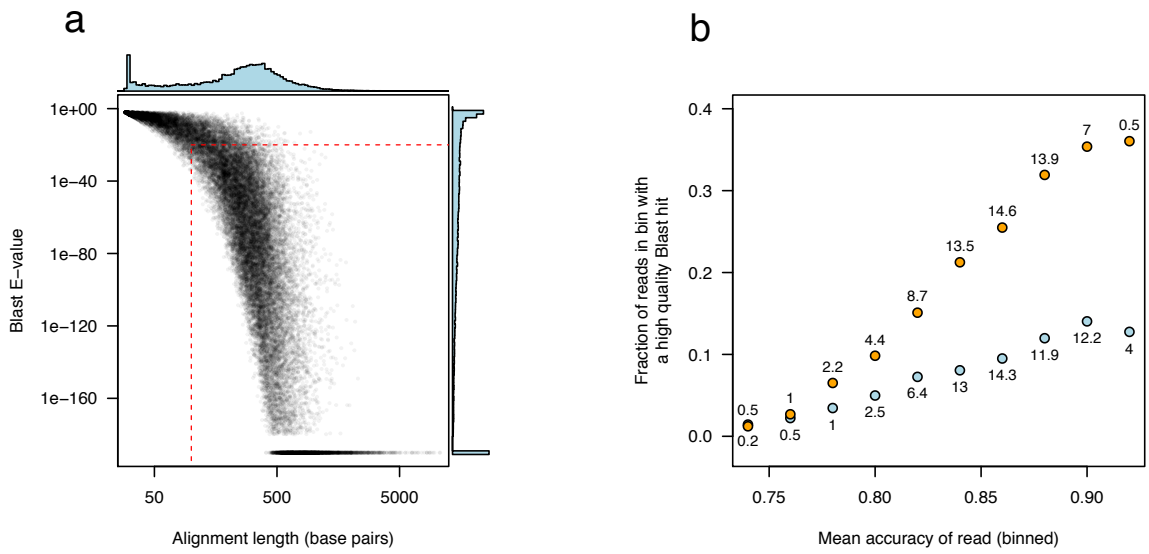


Fig. 3. BLAST hits of metagenomic reads. (a) Plot with marginal histograms showing the e-value and alignment length of the top BLAST hit for each read. We observed bimodal distributions of alignment lengths and e-values. The y-axis is plotted on a log scale, with zero e-values suppressed by adding a small number (1e-190) to each e-value. The horizontal red dotted line indicates the e-value cutoff we implemented and the vertical red dotted line indicates the length cutoff (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) increases 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 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 very high quality reads (accuracy >92.5%).

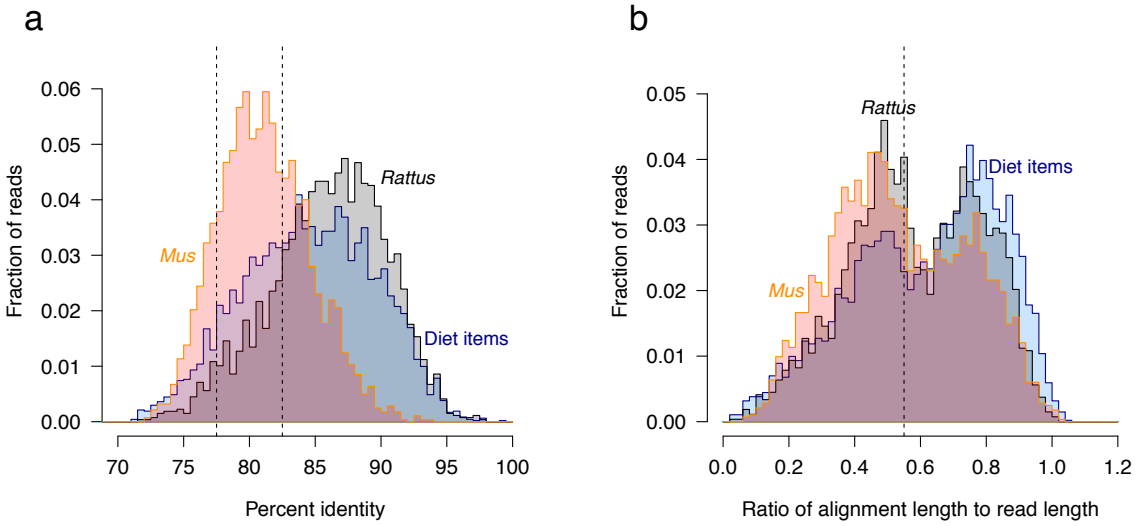
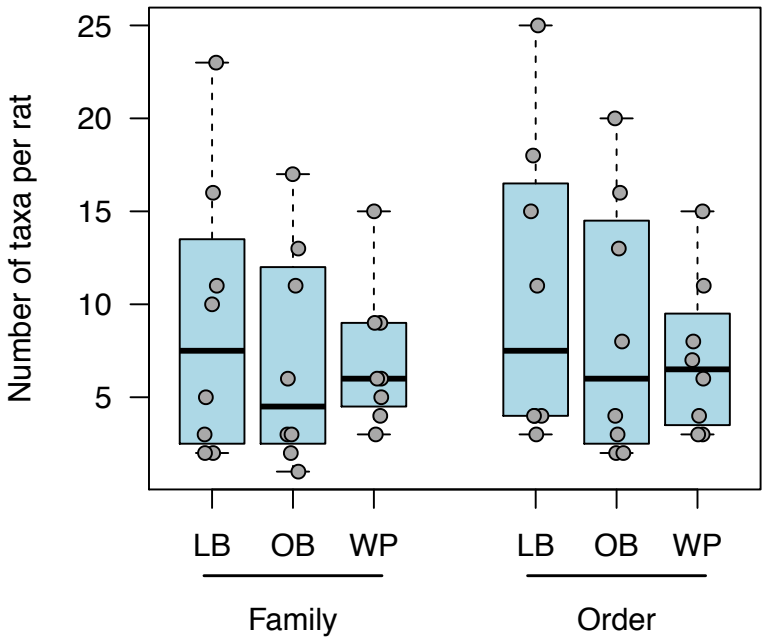


Fig. 4. Distributions of percent identity and length for alignments of reads matching *Rattus* (rat), *Mus* (mouse), and diet items. (a) Percent identity for alignments of rat (*Rattus*) and diet items is much higher than for mouse (*Mus*). Histograms are shown for the percent identity of the alignment of the top BLAST hit with the read. *Mus* matches show a clear shift to the left (lower percent identity) as compared to *Rattus* and diet items. Although different genera, *Mus* and *Rattus* are in the same family (*Muridae*). 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 rat (*Rattus*) and diet items are higher than for mouse (*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).



720

721 **Fig. 5. Numbers of taxa in individual rats.** Each boxplot indicates the range of families
722 (left boxes) or orders (right boxes) consumed by each rat in each location (OB: Okura
723 Bush; LB: Long Bay Park; WP: Waitakere Park). The numbers for individual rats (eight
724 per location) are plotted in grey.
725

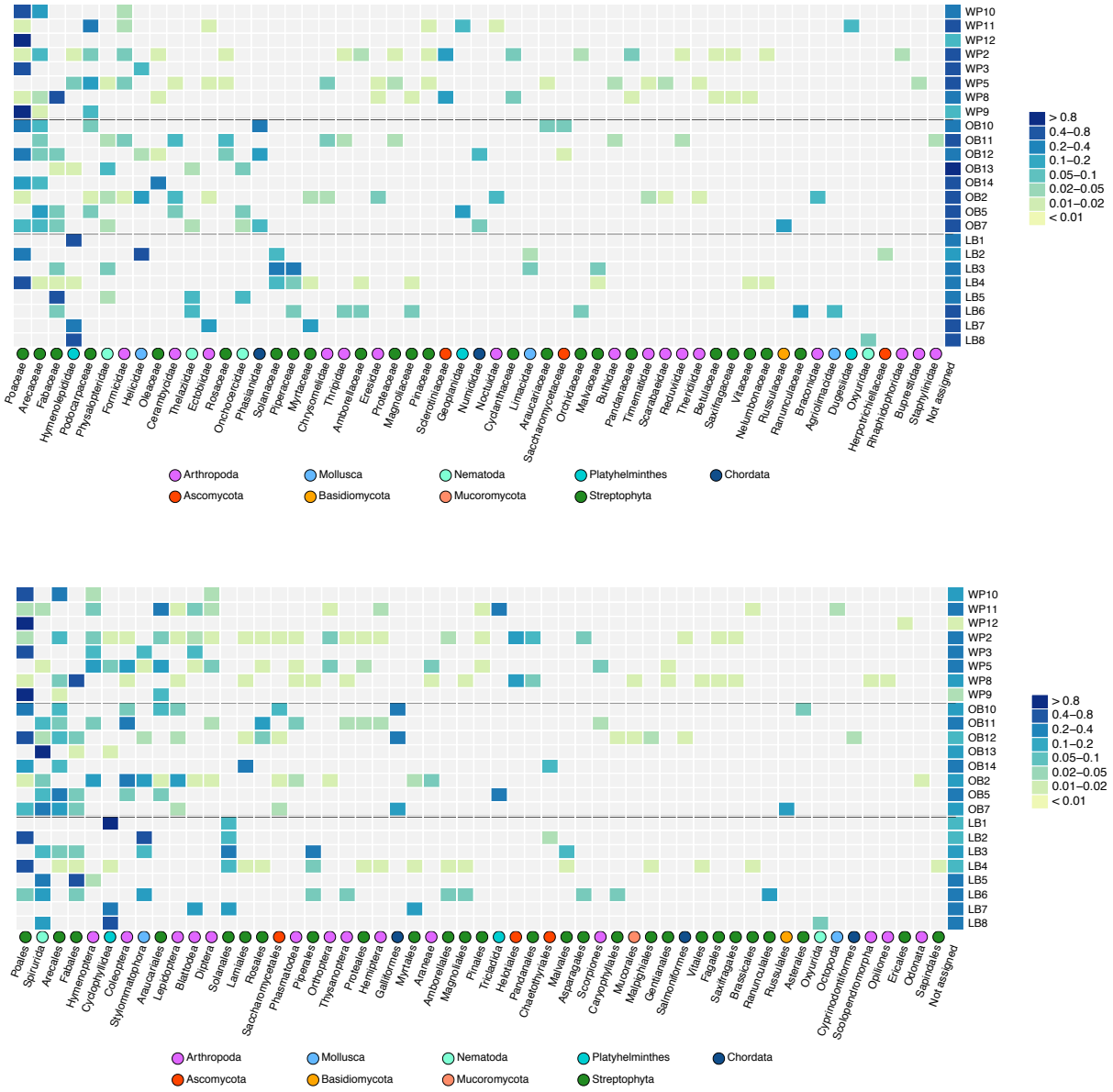
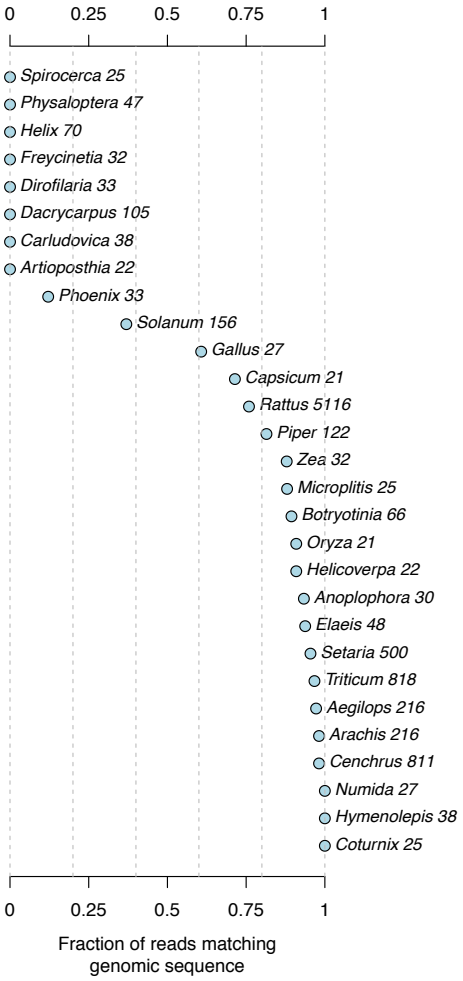


Fig. 6. Proportions of taxa in the diets of individual rats. (a) Reads assigned to taxa at the family and (b) order level. The rows correspond to a single rat, with the proportions of reads for that rat assigned to each family or order indicated in shades of blue and yellow. Reads that were not assigned 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 diet components appear on the left. Only the 55 most common families are shown. Note that the color gradations presented on the scale are not linear.



737

738 **Fig. 7.** Fractions of reads matching genomic and non-genomic sequence for the best
739 BLAST hit of each read. For the species with largely complete genomes, the fraction of
740 reads matching genomic sequence ranges from 60% to 100%. This large range is likely
741 due to the tissue from which the DNA was isolated. For example, muscle tissue has a
742 higher fraction of mtDNA to nuclear DNA than egg. For species without fully sequenced
743 genomes, this fraction ranges from 0% to 20% (for species with a small amount of
744 genomic data present in the database).

745

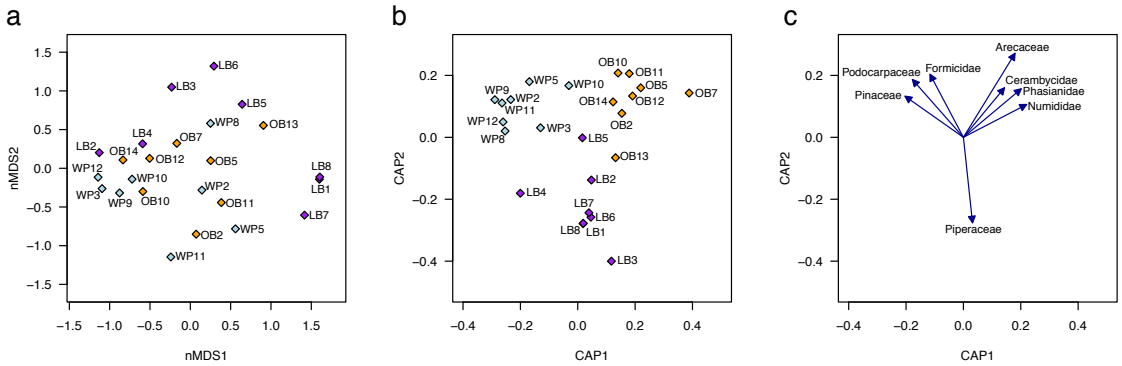


Fig. 8. Unconstrained nMDS (a) and constrained CAP (b) ordinations of the diets of rats from three locations. Both ordinations were based on Bray-Curtis dissimilarities of square root transformed proportions of reads attributed to each family. The locations were a native estuarine bush (OB, orange); a restored marine wetland (LB, purple); and a native forest (WP, light blue). The CAP ordination is repeated in panel **(c)** 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.

Supplemental Tables

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

Rat	Total reads	Total Mbp	Mean length
OB2	19907	14.62	734
WP11	10164	9.63	947
WP5	8237	6.78	823
LB7	7548	7.04	933
OB13	3644	3.63	995
WP9	2954	2.4	814
OB5	2850	2.06	721
WP8	2801	2.32	827
LB6	2531	1.6	632
OB7	2473	1.87	756
LB5	1641	1.16	705
LB3	1554	0.99	636
None	16673	13.01	781
Total	82977	67.1	N/A

762 **Table S2**. Read numbers and total base pairs for each barcode in the March
763 sequencing run.

Barcode	Total reads	Total Mbp	Mean length
LB1	17820	9.21	517
LB8	16923	13.13	776
WP2	10511	7.00	666
LB4	8684	4.92	567
OB11	5689	3.40	598
WP10	1563	0.99	633
OB12	1479	0.89	604
WP12	1309	0.78	596
LB2	1127	0.76	676
WP3	637	0.73	1141
OB14	541	0.37	683
OB10	435	0.24	555
None	29432	21.33	725
Total	96150	63.75	N/A

764

765

766 **Table S3.** SIMPER analysis of family contributions to group similarities.

Family	Average Abundance	Average Similarity	Similarity/SD	Percentage contribution	Group
Hymenolepididae	3.37	6.87	0.34	51.2	LB
Solanaceae	1.57	1.48	0.34	11.1	LB
Fabaceae	1.74	1.41	0.44	10.5	LB
Arecaceae	2.86	7.11	1	33.4	OB
Poaceae	2.87	4.82	0.55	22.7	OB
Fabaceae	1.17	1.98	0.51	9.3	OB
Phasianidae	1.79	1.67	0.34	7.9	OB
Poaceae	5.08	17.61	0.62	72.1	WP

767

768

769 **Table S4.** SIMPER analysis of family contributions to group dissimilarities.

Species	Average Abund Group1	Average Abund Group2	Avg. Dis-similarity	Dis-similarity /SD	% contrib	Group1	Group2
Poaceae	1.95	5.08	15.15	1.04	16.74	LB	WP
Poaceae	2.87	5.08	11.29	1.26	13.78	OB	WP
Hymenolepidae	3.37	0.48	10.8	0.73	11.93	LB	WP
Hymenolepidae	3.37	0.29	9.37	0.79	10.32	LB	OB
Poaceae	1.95	2.87	8.37	1.1	9.22	LB	OB
Arecaceae	0.05	2.86	6.99	1.41	7.7	LB	OB
Arecaceae	2.86	1.31	5.92	1.29	7.23	OB	WP
Fabaceae	1.74	1.05	6.14	0.67	6.78	LB	WP
Podocarpaceae	0	2.38	5.34	0.83	5.9	LB	WP
Podocarpaceae	0.71	2.38	4.82	0.99	5.88	OB	WP
Fabaceae	1.74	1.17	4.87	0.81	5.37	LB	OB
Fabaceae	1.17	1.05	4.31	0.84	5.26	OB	WP

770

771 **Datafile S1**. Table of read BLAST hits and assigned MEGAN taxa with no filters
772 applied.

773 **Datafile S2**. Table of read BLAST hits and assigned MEGAN taxa for diet items, with
774 reads reclassified at the family or order level by filtering on read length to alignment
775 length ratio and percent identity.

776 Supplemental Figures

777 **Fig S1. Biplots of read lengths and qualities for each barcode in the January and**
778 **March runs.**

779 **Fig S2. Correlation of read accuracy with alignment characteristics**. **(a-c)** Read
780 accuracy is positively correlated with the percent identity of the top BLAST hit. Points
781 show a subsample of reads; orange line indicates a running median; red dotted line is

782 the y=x line, which is expected if accuracy corresponds exactly to percent identity. (a)
783 indicates the relationship for diet items; (b) for rats; and (c) for mice. **(d-f)** Read accuracy
784 and alignment length show no significant relationship. Plots again are (d) diet items; (e)
785 rats; and (f) mice. **(g-i)** Read accuracy and the ratio of read length to alignment length
786 are positively correlated: more accurate reads are more likely to have long alignments
787 relative to read length. Plots again are (g) diet items; (h) rats; and (i) mice.

788