Bioinformatic challenges for using deep sequencing data for diet analysis of a threatened species

Rachel Gray

2189163G

Supervised by Dr Umer Zeeshan Ijaz and Professor Barbara Mable

MSc Quantitative Methods in Biodiversity, Conservation and Epidemiology

School of Medical and Veterinary Sciences

Contents

Abstract

Encroachment of grazing sites by livestock is a major threat to herbivorous wild populations. The Kilombero Valley, Tanzania, is home to the largest population of the threatened Puku *(Kobus vardonii)*, over the past two decades cattle presence has increased drastically in this area. This has caused concern that cattle will deplete the Pukus dietary niche. To date, there has been no classification of the degree of dietary overlap between Puku and cattle in this region. Here, I used a DNA metabarcoding approach to assess competitive exclusion of Puku by Cattle, this is a well-documented approach for studying the diet of elusive species; however, the optimal approach for bioinformatic processing of such samples is lacking. Here, I investigated the use of the novel sequence classification and taxonomic assignment methods to develop a bioinformatic pipeline for the analysis of low-quality sequence data. The pipeline was then used to determine if there is a difference between the diet of Puku in a site that contain cattle, and in a site that does not contain cattle. In the site where both Puku and cattle were present, I determined the degree of dietary overlap between the Puku and cattle. The bioiformatic pipeline recovered higher taxonomic assignment than previously used methods and supported the use of amplicon sequence variants. The Puku diet was found to be consistent across two samples sites, suggesting that cattle encroachment is not occurring. Despite this, multiple dietary items were still found to be shared between the Puku and cattle diet. The results presented here provide the first solid understanding of the Puku diet in the Kilombero Valley and so enhance understanding of Puku ecology overall, and should be considered by conservation managers when developing long-term conservation strategies.

Keywords:Metabarcoding, trnL, Puku, Bioinformatics

Acknowledgements

Thank you to Dr Umer Zeeshan Ijaz for the help and guidance regarding the bioinformatics aspect of this project. Thanks to Professor Barbara Mable for support in understanding the topics covered within this study. Many thanks to Koggani Koggani for providing the data set used in this project and providing indispensable knowledge on the Puku populations.

1 Background and Aims

1.1 Diet analysis as a conservation tool

Around the world, natural populations are facing increasing threats to their persistence, largely due to direct and indirect effects of human activities (Ceballos et al., 2015). One such anthropogenic driver of population declines is increasing livestock presence in wildlife areas (Prins, 2000). This is particularly a threat for herbivorous wildlife, whose dietary niche is at risk of encroachment and depletion by livestock. To combat the devastating loss of biodiversity that is predicted from these interactions, conservation managers must obtain accurate and precise dietary information for both domestic and wild populations within an ecosystem (Valentini et al., 2009a). A sound understanding of dietary overlap between livestock and wildlife populations will aid in the protection of wildlife by providing the necessary information to develop robust conservation strategies (Valentini et al., 2009a).

Classifying the diet of wild animals has traditionally been difficult and so extents of dietary overlap between wildlife and livestock are generally poorly resolved (Prins, 2000). Direct observation of the target species whilst it forages has been employed in some studies, however, this approach is limited in that it requires tracking of often elusive species and good visibility (Valentini et al., 2009a). Other studies have analysed stomach contents of the target animal, where undigested dietary items are identified using reference keys and digested items are investigated under the microscope (Balestrieri et al., 2011). This approach requires the animal to be deceased or for the stomach contents to be obtained whilst the animal is under anaesthesia (Valentini et al., 2009a) and therefore is unsuitable for threatened wild populations. Non-invasive sampling through the collection of animal faeces reduces some of these limitations and has been conducted in various ways (Carroll et al., 2018; Valentini et al., 2009a). Firstly, microhistological analyses have been undertaken where plant cuticle fragments are identified microscopically. This approach is extremely effort intensive and requires expert knowledge to identify dietary components (Kartzinel et al., 2015; Pompanon et al., 2012). Secondly, natural alkanes in plant wax have been investigated to identify taxa present; however this approach is also limited in scale, in that it is very difficult to determine taxa in complex samples (Dove and Mayes, 1996). Lastly, over the past few decades sequencing approaches have been developed to investigate diet (Carroll et al., 2018). Taking a sequencing

approach is held to be the most promising of the possible methods to date, yet it also faces challenges in terms of sample quality. Namely, faecal samples are prone to DNA degradation as a result of the length of time exposed to the environment, as well as the environmental conditions, and so quality and quantity of DNA obtained is often low (Wultsch et al., 2015). The samples, once collected, need to be stored appropriately to prevent further degradation before sequencing (Choo et al., 2015), but in remote field sites such storage facilities are often not available (Wultsch et al., 2015). The quality and quantity of DNA that can be obtained from the faecal samples determines the downstream analyses that can be undertaken and subsequent insights that can be gained (Ogden et al., 2009). With these challenges in mind, it is essential that the workflow for using sequence data for diet inference is optimised.

1.2 Next generation sequencing approaches for diet analysis

Early diet analysis studies that utilised sequencing were conducted using traditional methods, where DNA was subject to PCR, the resulting PCR products were then cloned, and clones were subject to Sanger sequencing by capillary electrophoresis (Shendure and Ji, 2008; Shehzad et al., 2012). Such studies provided a basis for the field of diet analysis by sequencing, but are limited in throughput (Hunter et al., 2018; Pegard et al., 2009). Next generation sequencing (NGS) technologies became widely available in the early-2010s (Valentini et al., 2009b), and revolutionised the field of diet analysis by introducing amplicon (PCR products) sequencing approaches. Amplicon sequencing approaches enable specific genetic regions of interest to be investigated (Callahan et al., 2016). There are several differences between traditional electrophoretic sequencing and NGS, but the crucial difference is the unprecedented amount of data generated by NGS as a result of multiplexing samples. Template copies to be sequenced are generated by in-vitro amplification, rather than cloning, and this is typically done via one of two methods; bridge amplification is the simplest approach, where a complex template library with primers immobilised on a surface and amplified, in this way copies of each template are tightly clustered together for sequencing; the second approach is emulsion PCR (ePCR), where copies of each template to be sequenced are immobilised on beads, then arrayed on a surface (Shehzad et al., 2012). Currently, the market leader for NGS platforms is Illumina due to its unprecedented accuracy, low relative cost and flexibility (Shendure et al., 2017). For diet investigation using amplicons the Illumina Hi-Seq platform is widely used due to

its high-throughput and production of short read lengths that are favourable for amplicon analysis as they allow for an accurate targeted approach (Porter and Hajibabaei, 2018).

Amplicon-based DNA metabarcoding is one such commonly used approach for diet analysis. In short, metabarcoding consists of targeting specific taxonomically informative DNA regions to be amplified by PCR and subsequently sequenced, such DNA regions are known as the DNA barcode (Alberdi et al., 2019; Valentini et al., 2009a). The resulting amplicons are then sequenced on NGS platforms, such as the Illumina HiSeq and produce FASTQ files that are then analysed using bioinformatic pipelines (Swift et al., 2018), see section 1.3 for more information on bioinformatic amplicon analysis. Barcode sequences are then identified by being matched to a DNA reference database (i.e such as those provided by BOLD or Genbank). (Valentini et al., 2009a). Many samples can therefore be sequenced and analysed in parallel in metabarcoding. In addition to this, the bioinformatic analysis of metabarcoding data has a low computational cost (Alberdi et al., 2019) and so does require expensive, powerful computers, making it a more widely accessible approach.

The choice of barcode used in a metabarcoding study is a key part of the experimental design (Taberlet et al., 2007). A good barcode should have discriminatory power and be conserved at the species level to effectively deduce sequences (Taberlet et al., 2007). Within the literature, the rbcL and matK barcodes have been widely used to resolve plant taxa (Mallott et al., 2018). Taberlet et al. (2007) investigated the use of the chloroplast trnL (UAA) intron as another plant barcode and determined that although the trnL barcode returned relatively low resolution in comparison to other plant markers, it also holds much potential. The first advantage of the trnL barcode is that the intraspecific variation of the target region is low between samples, which makes the barcode withstand amplification well. Secondly, the short length of the target region of the trnL (UAA) intron means that DNA can be amplified even highly degraded and so holds much promise for identifying plant species within non-invasively collected faecal samples. The chloroplast trnL (UAA) intron approach has since been incorporated into a wide range of studies to non-invasively infer the diet of many species including: red deer (Flojgaard et al., 2017), subterranean rodents (Lopes et al., 2015) and African herbivores (Kartzinel et al., 2015).

1.3 Amplicon Sequence Variants

Traditionally, NGS marker gene data generated by the DNA metabarcoding approach have been analysed by the construction of operational taxonomic units (OTUs) (Callahan et al., 2017). An OTU is a cluster of sequences that have been grouped together by statistical software as sequences that are similar by a specified threshold, which is most commonly defined at 97% (Westcott and Schloss, 2015). Recently, amplicon sequencing variants (ASVs) have been popularised as an alternative, as they do not require an arbitrary similarity threshold and so better represent true sequences (Callahan et al., 2016).The determination of ASVs requires the use of an error-model programmae, such as DADA2 (Callahan et al., 2016). DADA2 takes raw FASTQ files, processes them through a statistical model of amplicon sequencing error and identifies distinct true sequences from those that are likely to have been generated by error (Callahan et al., 2016). The ASVs are then complied in an ASV abundance table (Callahan et al., 2019; Porter and Hajibabaei, 2018). Other advantages of ASVs over OTUs include that they are consistent between studies and so can be directly compared and reproduced in future datasets, also, ASVs can be distinguished by differences as small as a single nucleotide (Callahan et al., 2016).

1.4 Novel approaches to data analysis

. Amplicon sequence variant data is multifaceted and high dimensional and as a result there are many potential downstream analyses available, many of which are tailored towards ASV analysis for microbiome research. It is therefore essential that these novel ASV analysis techniques are explored for use in diet inference using ASVs. Specifically, novel ASV analysis methods that take a null modelling approach allow for environmental processes that are impacting communities observed to be understood (Vass et al., 2020) (see Methods 2.5 for a full description). The microbiome literature also promotes community assemblage investigation from a phylogenetic point of view through a Lottery model approach (Verster and Borenstein, 2018), which has the power to identify the clades responsible for changes in the community (see Methods 2.6 for a full description). Both null modelling and the lottery model, have to the best of my knowledge not been applied to dietary analysis studies previously and so it is of great interest to determine the power they have in enhancing dietary inference from ASVs.

1.5 Case Study: Diet Analysis of Puku and Domestic Cattle

The Puku *(Kobus vardonii)* is one such species that is in urgent need of in-depth dietary classification (Rduch, 2016). Briefly, Puku are thought to have once been widely present in Savannah grasslands and floodplains of southern and central Africa, but are now isolated to just eight African countries (Jenkins et al., 2003). To date, previous studies that have investigated the Puku diet across Africa have either taken a microhistological approach, where fragments of plant epidermal tissue in faeces were investigated microscopically in an attempt to classify diet (Rduch, 2016), or a direct observational approach, where Puku were observed foraging and then the plant species within the forage patch were identified once they had moved on (O'Shaughnessy et al., 2014; Rosser, 1992). These studies provide a good general idea of the Puku diet and agreed on the main components of the diet are grasses and monocotyledon plants (Rduch, 2016; O'Shaughnessy et al., 2014). Currently, the largest Puku population is located in the Kilombero Valley in Tanzania, which has been suggested to be the most important population to direct conservation interventions towards (Jenkins et al., 2003). The diet of Puku population in the Kilombero Valley has been greatly understudied despite its importance and so could benefit greatly from in-depth diet investigation by sequencing. This is particularly the case, as over the past few decades livestock presence has increased hugely in this area (Jenkins et al., 2003). Previous studies that have investigated cattle diet in similar ecosystems report grasses and forbs to be the main taxa present (Kartzinel et al., 2015) and so there is evidence to suggest that Puku and cattle may overlap in diet (Bonnington et al., 2007). Cattle presence in the Kilombero Valley therefore poses a potential threat to Puku, in that cattle may indirectly adversely affect the Puku population by habitat changes due to over grazing (Jenkins et al., 2003). If domestic cattle have out-competed Puku for grazing, Puku in areas with a high livestock presence are expected to consume different food items than Puku in livestock free areas. Overall, a better understanding of the diet of the Puku in the Kilombero Valley will aid in developing robust conservation management plans as well as improving understanding of Puku ecology.

1.6 Aims and Objectives

The overall objective of this study is to produce an optimal bioinformatic pipeline that can infer diet from non-invasively collected and low-quality metabarcoding samples, and ultimately can be used to inform

conservation strategies for threatened species. Specifically, the pipeline will be used to determine the diet of Puku and Cattle in the Kilombero Valley region of Tanzania and investigate:

1. Whether there is a difference in the vegetation that Puku are eating in two different sites.

2. Whether there is overlap in the diets of Puku and domestic Cattle in a shared site.

2 Methods

2.1 Data description

Data used within this project was collected and provided by Koggani Koggani. Puku and cattle faecal samples were collected in the Kilombero Valley, Tanzania. The Kilombero Valley was divided into two sites based on the degree of cattle encroachment and agricultural activities within the area (Figure 1). The ND site contained both Puku and cattle and has high levels of ongoing agriculture. The ME site is a protected area and so is purely Puku, as cattle have no access, and there are low level of agricultural activities in this area. In total 143 faecal samples were collected, of which, 65 were domestic cattle from the ND site, 33 were Puku from the ND site and 45 were Puku from the ME site. Samples were collected from November 2018 to April 2019. Samples were amplified using the trnL barcode approach and subsequently sequenced following Illumina Hi-seq protocols to produce FASTQ files for analysis.

In addition to genetic data, transect data was also provided for the two sites that detailed plant species abundance. For the ND site, one transect survey was conducted by point sampling at 30 points. For the ME site, three surveys were conducted, of which two were sampled at 30 points and one was sampled at 25 points.

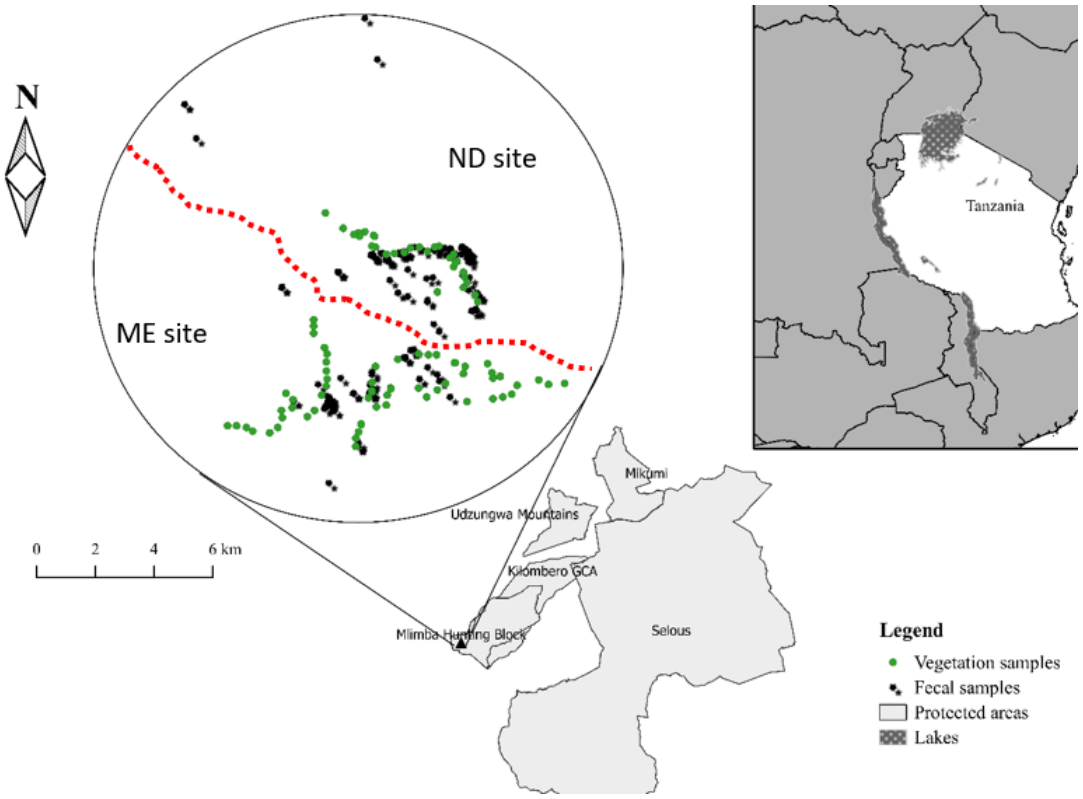


Figure 1: Kilombero Valley map. Provided by Koggan Koggani and edited to show the division of the landscape into the ND and ME site, represented by the red dashed line. Points of faecal sample collection are shown in black and transect point surveys shown in green.

2.2 Taxonomic Identification Comparison

Two alternative taxonomic assignment methods were investigated to determine which provides the best resolution for the trnL barcode; the naive bayes classifier approach (Wang et al., 2007) and the bayesian lowest common ancestor (hereafer, BLCA) approach (Gao et al., 2017). The naive classifier approach has been widely used in the literature to resolve amplicon sequence data, and is based on a simple Bayes network model that predicts the probability distribution of taxonomic assignments from kingdom through to genus level (Wang et al., 2007).The main drawbacks of the naive bayes classifier approach is that it is often non-informative for functional genes, such as trnL (Gao et al., 2017). The BLCA approach uses pairwise sequence alignments to calculate the true sequence similarity between each query sequence and database hits (Gao et al., 2017). Taxonomy is then assigned from the species level up to the phylum level, based on the lowest common ancestor of the multiple database hits for each query sequence (Gao et al., 2017). The BLCA approach only requires a reference database to align the query sequence, whereas the naive approach requires training of a classifier to the dataset before taxonomy can be determined (Wang et al., 2007), which is timely and requires bioinformatic expertise (Gao et al., 2017). To the best of my knowledge, the BLCA approach has not yet been used in diet analysis studies.

There is currently no published trnL reference library available and so we constructed one from the trnL repository available at: https://github.com/RTRichar/MetabarcodeDBsV2 . This database contains 75,406 unique sequences which includes 107 orders, 454 families, 8,478 genera and 51,033 species. A classifier was also generated from this library for use in the naive bayesian classifier approach. For a full list of steps involved in generating the classifier please see AppendixI. The bioinformatic workflow for the naive classifier and BLCA approach followed the same structure and differed only slightly (see Methods 2.3). The output of both the naive bayesian classifier and BLCA approach were compared for quality of taxonomic assignments first before continuing with further analyses.

2.3 Qiime2 bioinformatics pipeline with DADA2

The workflow for Qiime2 (Bolyen et al., 2019) with DADA2 (Callahan et al., 2016) workflow used here to create an abundance table of amplicons sequence variants was adapted from https://github.com/umerijaz/tutorials/blob/master/qiime2_tutorial.md . For a complete methodology of the steps performed in my analysis please see Appendix II. In brief, FASTQ files were imported into Qiime2 in Earth Microbiome Project Paired-end Sequencing Format. This required organisation of forward and reverse FASTQ files for each samples into a 'Raw' folder for each sample, as well as generation of fictitious barcodes before assembly. The forward and reverse reads for each sample were then assembled together and imported into Qiime2 and demultiplexed. The reads were then exported to the Qiime2 viewer (https://view.qiime2.org/) for visual quality assessment. For the forward reads quality was observed to drop at 110bp and for reverse reads at 100bp. The DADA2 algorithm was then performed on the truncated reads. A pylogenetic tree was then created within Qiime2 with allign-to-tree-mafft-fasttree using MAFFT v7.310 (Katoh and Standley, 2013) and Fastree v2.1.10(Price et al., 2010), which was then exported in NEWICK format. At this stage the methodology then split for the two separate taxonomic assignment approaches.

For the naive classifier approach the taxonomy was generated by using the trnL classifier generated in Methods 2.2. This produced a taxonomy file (taxonomy_naive.tsv) which could then be merged with the ASV abundance table (feature_table.txt) to create a BIOM file (feature_w_tax_naive.biom). The BIOM file format therefore contains the taxonomic identification for each ASV and the number of times the ASV is observed in each sample and is the traditional file format for ASV analysis.

For the BLCA approach the first step was to convert the trnL reference database to the correct format to be used with BLCA software, this was achieved by renaming columns to 'species', 'genus', 'family', 'order', 'class', 'phylum' and 'superkingdom' appropriately. The BLCA software was implemented as a Python package which is available at: https://github.com/qunfengdong/BLCA . The BLCA software produced a taxonomy file (taxonomy.tsv), which had to be reformatted from the BLCA software format so that it could then be converted to BIOM format. This consisted of replacing the 'species', 'genus', 'family', 'order', 'class', 'phylum' and 'superkingdom' column names with 'D_6_', 'D_5_', 'D_4_', 'D_3_', 'D_2_' and 'D_1_' respectively. Lastly, the ASV abundance table (feature_table.txt) was

combined with the BLCA taxonomy file (taxonomy_BLCA.tsv) to generate a BLCA approach BIOM file (feature_w_tax_BLCA.biom).

2.4 Statistical analysis

Statistical analysis was performed in R v 4.0.2 (RCoreTeam, 2020), using both the BIOM and NEWICK files generated from the DADA2 bioinformatic pipeline described in Methods 2.3, and the associated metadata for the study which included information on the site, species and month that the sample was collected. All R scripts are provided in accompanying folder and named according to analysis.

2.4.1 Diversity Patterns: Alpha Diversity

Alpha diversity (within population variation) was calculated for Puku ME, Puku ND and Cattle ND respectively using the vegan package (Oksanen et al., 2019). It is becoming increasingly recognised within the literature that a single alpha diversity measure is does not completely characterise diversity within populations and so five alternative measures were calculated; species richness, which identifies the number of species present and is the simplest metric for representing diversity(Whittaker, 1972); shannon entropy, which measures the balance of communities, where a higher index is a more balanced community (Shannon, 1948), pielou's index, measures evenness within communities (Pielou, 1966); fisher alpha, a diversity index that compares among communities varying in number of individuals (Fisher, 1972) and simpson diversity which is another measure of evenness within communites and ranges from 0 to 1(Simpson, 1949).

2.4.2 Diversity Patterns: NRI and NTI

Environmental filtering was investigated at the genera level for Puku ND, Puku ME and Cattle ND, to identify phylogenetic overdispersion or clustering within each population. Phylogenetic distances were characterised by calculating the nearest taxa index (hereafter NTI) and nearest relatedness index (hereafter NRI). The NTI and NRI reflect environmental filtering processes that are occurring at different parts of the phylogenetic tree (Cooper et al., 2008). The NTI is based on the mean nearest taxon distance

(MNTD), which is defined as the mean distance between each ASV in a sample, and its closest relative in the phylogenetic tree. NTI reflects overdispersion or clustering near the tips of the phylogeny (Cooper et al., 2008). The NRI is based on the mean phylogenetic distance (MPD) of samples, which is defined as the mean phylogenetic distance among all pairs of ASVs within a sample and reflects clustering or overdispersion across the whole phylogeny (Cooper et al., 2008). Here, the NTI and NRI were calculated using the picante package (Kembel et al., 2010). Specifically, to calculate NRI I used the mpd() and ses.mpd() functions and to calculate NTI I used the mntd() and ses.mntd() functions. The NRI and NTI values reported here are the MPD and MNTD values multiplied by -1, as it is more intuitive for clustered values to be positive and over-dispersed values to be negative (Cooper et al., 2008). Values of NRI or NTI recovered as >0 indicate that there is strong phylogenetic clustering, as a result of environmental filtering. NRI or NTI values <0, indicate phylogenetic over dispersion, where the environment plays a small role in the ASVs observed. Here, phylogenetic overdispersion refers to when competition in the environment affects the ASVs within a sample, and so species identified in samples will be more distantly related than what is expected by chance (Campbell O. Webb and Donoghue, 2002). In contrast, phylogenetic clustering refers to when ASVs present within a sample is a result of environmental filtering, and so the species observed within a sample will be more closely related than what is expected by chance (Campbell O. Webb and Donoghue, 2002). .

2.4.3 Diversity Patterns: Beta Diversity

To investigate the differences in diet composition between populations, beta diversity was measured via three alternative distance metrics. Firstly, Bray-Curtis distances (Bray and Curtis, 1957) were calculated, which test if populations differ significantly in terms of abundance counts of ASVs. Next, Unifrac distances (Lozupone and Knight, 2005) were calculated to determine if populations are significantly different based purely on phylogeny. In the Unifrac method, phylogenetic distance is calculated between pairs of ASVs in the phylogeny. A branch that leads to an ASV from both samples is classed as a shared branch, and branches which lead to ASVs only present in one sample are classed as unshared (Lozupone and Knight, 2005). The phylogenetic distance between samples is then calculated as the total of all unshared branch lengths over the total of all shared branch lengths. (Lozupone and Knight, 2005) The

Weighted unifrac distances (Lozupone et al., 2007) were calculated, where both abundance counts of ASVs and phylogenetic distances are taken into account. (Lozupone et al., 2007).Unifrac and weighted unifrac were calculated using the phyloseq package (McMurdie and Holmes, 2013).

A PERMANVOA analysis was performed alongside all the beta diversity measures to identify sources of variation within the dataset, using the adonis() function in Vegan. Here, both Population and Month were tested for causing variation. If predictors are recovered as significant from the PERMANOVA analysis, then R^2 values represent the percentage variability explained by the predictor.

2.4.4 Diversity Patterns: Observation of the top-25 most abundant taxa

The 25 most abundant genera were calculated for Puku ME, Puku ND and Cattle ND, to provide a visual comparison of how vegetation taxa abundance within the diets differed across the three populations.

2.4.5 Identifying the key drivers of diet variation in terms of beta diversity: Subset analysis

A subset analysis was performed to identify the ASVs driving differences in beta diversity observed between populations. In brief, this consisted of calculating pairwise Bray-Curtis distances between all possible sample combinations, then permuting through all the possible combinations and retaining the subset of ASVs that has the maximum correlation with the full set of ASVs. The remaining subset of ASVs therefore explains roughly the same beta diversity as the full set, but with reduced complexity and contains only those ASVs that are causing differences in beta diversity between samples. This was conducted via the ”BVSTEP” routine using the bvStep() function from the sinkr package (Taylor, 2017). The subset analysis was performed for all population comparison combinations (Puku ND + Puku ME, Cattle ND + Puku ND, Cattle ND + Puku ME). In addition, the two Puku sites were grouped together to investigate the general difference between Puku and Cattle, the subset of ASVs that were recovered from this analysis were then constructed into a phylogentic tree for visualisation. The phylogenetic tree was created using the ape (Paradis and Schliep, 2019) and phangorn (Schliep et al., 2017) packages and then edited within the Evolview environment : https://www.evolgenius.info/evolview/ .

2.4.6 Identifying the key drivers of diet variation in terms of beta diversity: DeSeq and Heat tree

The subset of genera returned by the subset analysis for the Puku ND + Puku ME, Puku ND + Cattle ND and Puku ME + Cattle ND pairs were then investigated further to identify which genera are causing the differences in beta diversity between populations. This was first done via a DeSeq analysis, using the DeSeqDataSetFromMatrix() function in the DeSeq2 package (Love et al., 2014). This analysis applies a negative binomial GLM to the dataset to obtain the maximum likelihood estimates for ASVs log fold change between two populations, here the log fold change was set at 2 the adjusted p-value significance was cut-off at 0.005. The CSV file obtained from this analysis identified genera up-regulated in populations (i.e. the genera that are identified at an increased presence in one population, compared to the other). Next, a heat tree analysis was conducted to identify important clades causing differences in beta diversity between populations. This was achieved using the metacoder package (Foster et al., 2017). First, a labelled taxa tree was produced that shows the evolutionary distribution of ASVs within all samples. Then, a colour coded tree that shows pairwsie comparisons between populations was generated and allowed for visualisation of the taxa that were up-regulated in populations using Wilcoxin p-value adjusted with multiple comparisons.

2.4.7 Identifying the key drivers of diet variation in terms of beta diversity: investigating core diet between Puku and cattle

The ASVs that were prevalent in $>85\%$ of samples (a traditional high prevalence threshold in microbiome literature (Shetty et al., 2017)) were identified by a core vegetation analysis. To perform this analysis the microbiome package (Lahti and Shetty, 2017) was used. This analysis was performed on all pairwise combinations of populations, as well as across all three populations at once.

2.4.8 Null Modelling Approaches: Calculating Quantitative Process Estimate and incidence-based (Raup-Crick) beta-diversity

A null modelling approach was taken to explore the ecological drivers of the vegetation community assemblages of each population. To the best of my knowledge, null modelling approaches have not been

applied to diet analysis studies before, and so here I am also testing their power and performance for dietary samples. For this analysis, both the Puku ME and Puku ND were grouped into one Puku sample, as the previous analyses have identified them to be indifferent. Here, we investigated null modelling approaches for diet analysis in two ways.The first of which was to calculate quantitative process estimates (QPE), which quantify the assembly processes involving phylogeny and abundance-based (Raup-Crick) beta-diversity (Vass et al., 2020). The QPE approach aimed to investigate the relative importance of each assembly process and here, the assembly processes investigated were: dispersal limitation, ecological drift, homogenising dispersal, homogenising selection and variable selection. The second approach taken was to investigate incidence-based (Raup-Crick) beta-diversity (βRC) that tests for stochastic and deterministic processes within samples by looking at the presence and absence of ASVs in an abundance table, and does not account for phylogeny (Vass et al., 2020). If incidence- basedβRC is recovered to not differ significantly from 0, then the community is classified as stochastically assembled. In the case that incidence-based βRC is recovered close to +1 then this indicates that deterministic processes are favouring dissimilar communities. If incidence-based βRC values are recovered close to -1, then communities are deterministically assembled and more similar to each other than would be expected by chance (Vass et al., 2020). The QPE and incidence-based βR were calculated using the picante (Kembel et al., 2010), ape (Paradis and Schliep, 2019) and ecodistance (Goslee and Urban, 2007) packages.

2.4.9 The Lottery Model

A lottery model approach was taken to further understand vegetation community assembly at both the family and genus levels. The lottery model is traditionally used in microbiology to describe how organisms colonise a niche (Verster and Borenstein, 2018). Essentially organisms from a pool of species all have the potential to colonise the same niche, but the lottery model posits that first species to arrive at the niche will exclude all the other possible organisms (Verster and Borenstein, 2018). Here, I am investigating this in terms of dietary niche and the lottery winners recovered are the ASVs that are driving changes in the dietary community composition. The methodology followed here was adapted from (Verster and Borenstein, 2018). In brief, both winner prevalence and winner diversity were investigated. Winner prevalence is the proportion of samples in which a genus or family that occupies >90% of total

samples is observed (Verster and Borenstein, 2018). Winner diversity is the frequency that each ASV is selected as a lottery winner, considering all the samples observed lottery winners. Here, winner diversity was calculated as the Shannon diversity of the distribution of winner ASVs, across all samples (Verster and Borenstein, 2018). I then normalised the winner diversity so that values range from 0 to 1 to make graphical outputs easier to interpret, and this was achieved by log2 of the number of winners. A low winner diversity, suggests that not many groups are being selected as lottery winners, and so it is typically the same groups that are selected in all samples. A high winner diversity suggests that the groups chosen by lottery winners is even across all samples (Verster and Borenstein, 2018). The phyloseq (McMurdie and Holmes, 2013) package was used to calculate winner diversity and winner prevalence.

2.4.10 Comparing Transect and Genetic Data

Transect data was provided as counts of species present at a point within a transect. This was transformed to proportions of each species present in a transect. Three transects were from the ME site and a single transect from the ND site. The three ME site transects were combined together so that species comparison could be made between sites. The genetic data where ASVs were identified to species was also transformed to proportions for comparison. Firstly, the transect and genetic were compared by producing a Venn diagram using the draw.quad.venn() function within the VennDiagram package (Chen, 2018). The purpose of this was to determine how many taxa that could be identified to species level were shared between the transect and genetic approach. Then, a rank abundance plot was produced for: transect data at ME site, transect data at ND site, genetic data from cattle ND, genetic data from Puku ND and genetic data from Puku ME. For the genetic plots only the top 5% most abundant taxa were plotted due to large amount of taxa present. For all plots the top 5 most abundant taxa were labelled down to the species level. The purpose of this analysis was to determine if the most abundant taxa identified within the diets of the study species corresponded with the most abundant taxa present in the field.

3 Results

3.1 Taxonomic Identification Comparison

The abundance table produced from the Qiime2 with DADA2 workflow contained 4382 ASVs for 143 samples. For the BLCA approach, 77.9% of ASVs were returned as unidentified, 19.9% were identified to genus and 17.02% were identified to the ASV level. For the naive classifier approach 100% of ASVs were identified to kingdom, 99.61% were identified to the phylum, 27.43% to class, 26.74% to order, 25.01% to family, 6.89% to genus and 2.69% to ASV level. The majority of downstream analyses implemented here required classification to genus level and so subsequent analyses were performed solely on the BLCA approach output as it yielded the highest classification of taxa at Genus and ASV level.

3.2 Diversity Patterns

3.2.1 Alpha Diversity

Investigations of alpha diversity measures revealed how diversity of dietary components varied within Puku ME, Puku ND and Cattle ND populations (Figure 2a). Firstly, the two Puku populations recovered similar variation in terms of the dietary components for all measures. Within the cattle population, a wider range of ASVs were identified, meaning that the cattle population shows more variation in terms of the diversity of dietary components. Overall, all populations recovered similar mean values for all five measures.

3.2.2 NRI and NTI

For all three populations, the NRI and NTI were recovered as >0, which indicates that strong phylogenetic clustering, driven by environmental filtering is present throughout the whole phylogney as well as at the tips (Figure 2a). Therefore, in all populations the ASVs within each population are more closely related to each other than is expected by chance and ASV presence based on stochastic processes can be ruled out. For the cattle population, the mean NTI was higher than that of both populations, suggesting that dietary items within the cattle diet are more strongly clustered at the tips of the phylogeny than Puku.

3.2.3 Beta Diversity

In terms of differences between populations driven by abundance counts, sequences from Puku ME and Puku ND are clustered closely together, whereas Cattle ND sequences are quite far off from these (Figure 2b). The PERMANOVA identified month as the predictor explaining the most beta diversity in terms of abundances ($p<0.001$, R2=0.193, df=6), and population explained less variation ($p<0.001$, R2=0.154, df=2). The Unifrac (phylogeny) in terms of phylogentic distance, also identifies the sequences of the two Puku populations to be very clustered very close together and mostly overlapping and the Cattle ND is again quite far off. The PERMANOVA for the unifrac distance metric identified both month and population to explain similar amounts of variation (population = $p<0.001$, R2=0.148, df=2, month=$p<0.001$, R2=0.147, df=6). Wunifrac (phylogeny and abundance) measures were also considered but are not shown in the main text (see Appendix IV).

3.2.4 Top-25 Most abundant taxa identified for each population

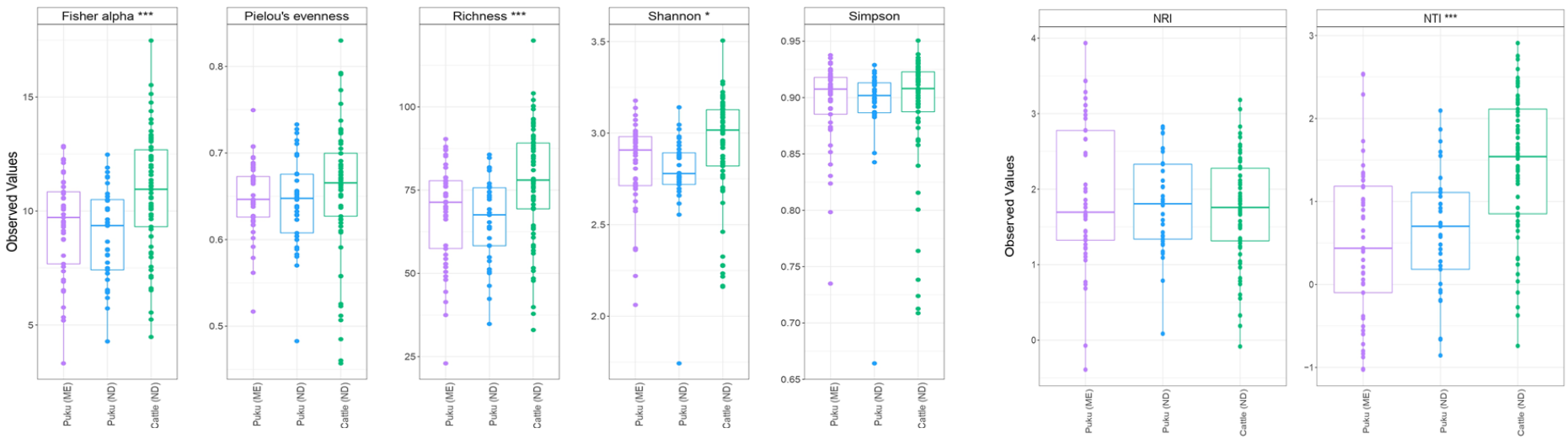
The top-25 most abundant genera are shown for each population (Figure 2b). The breakdown of taxa at finer levels (the ASV level) are shown in Appendix V. Observational comparisons of the top-25 most abundant genera between the three populations gave the first insight into similarities and differences of the dietary composition at genera level of each population.

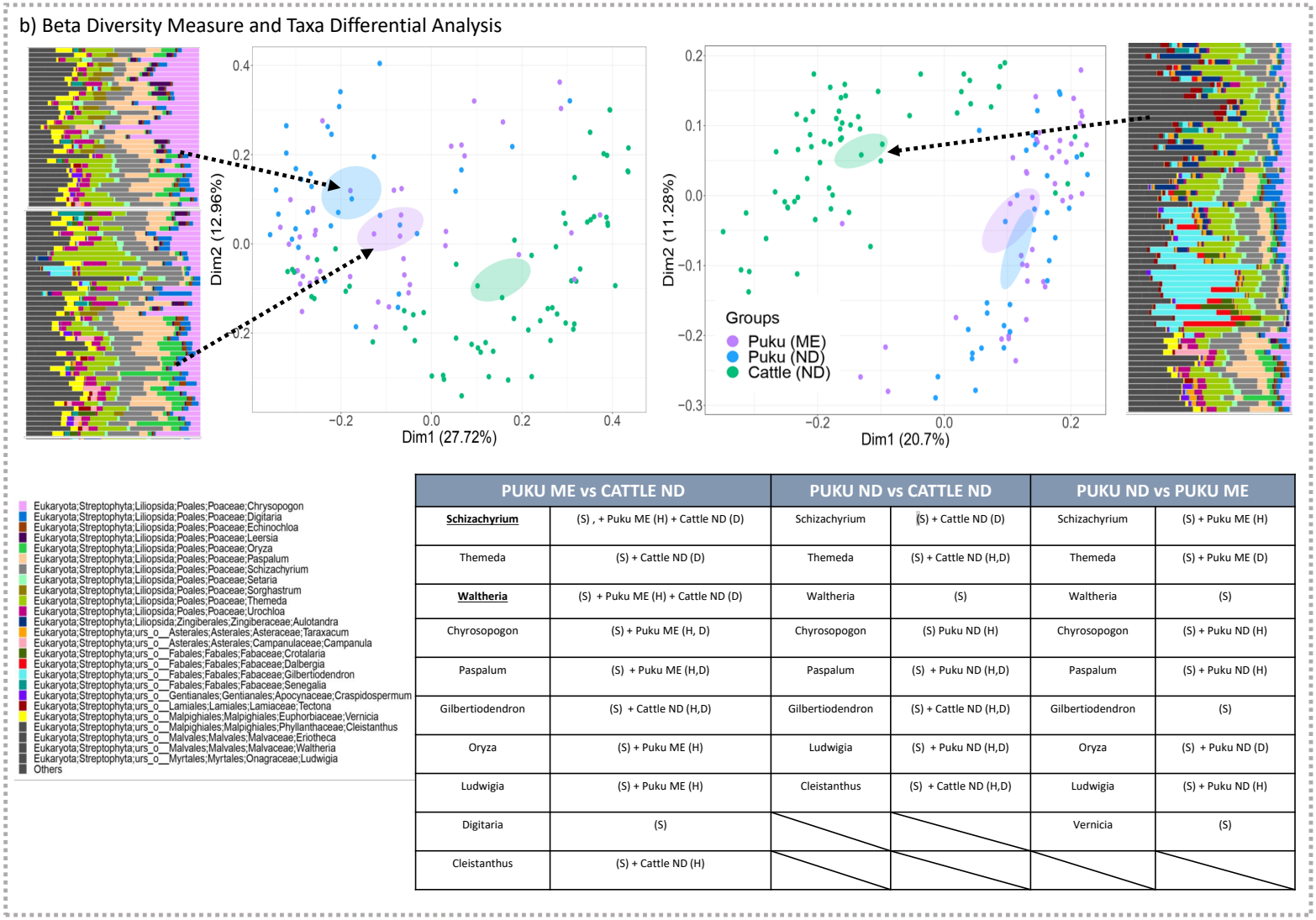
Firstly, comparison of the Puku ND and Puku ME populations revealed very few differences. *Chyrosopogon*, *Paspalum*,*Themeda* and *Schizachyrium* were recovered as highly abundant genera in both populations. These genera are all members of the Poaceae family (Clayton and Williamson, 2006 onwards), which is commonly known as the grass family. The clearest difference between Puku ND and Puku ME diet composition that is observed is the presence of *Oryza* in Puku ME at higher abundance. *Oryza* is also a member of the Poceae family, but it is of note as it contains the food crop rice.

Next, comparing the Puku ND and Puku ME populations with the Cattle ND population, a few differences can be observed. Most noticeably, there is a high abundance of *Gilbertiodendron* in the Cattle ND populations diet, that is only observed at very low levels in both the Puku populations. *Gilbertiodendron* is a member of the Fabaceae family, commonly known as the legume family, which contains flowering

plants (Estrella et al., 2014). Members of the *Gilbertiodendron* genus are primarily found in dry-land and gallery forests(Estrella et al., 2014). The Cattle diet also has a higher abundance of the *Dalbergia* genus than either Puku population, which is also a member of the Fabaceae family and is pan-tropical (Vatanparast et al., 2013). Both *Paspalum* and *Chyrosopogn* are present at higher abundances in Puku ME and Puku ND, than in the Cattle ND population. Additionally of note, both *Themeda* and *Schizachryrium* were also present at high abundance in Cattle ND, at what appears to be similar abundance levels to both Puku ND and Puku ME.

a) Alpha Diversity Measures





PUKU ME vs CATTLE ND		PUKU ND vs CATTLE ND		PUKU ND vs PUKU ME	
Schizachyrium	(S) , + Puku ME (H) + Cattle ND (D)	Schizachyrium	(S) + Cattle ND (D)	Schizachyrium	(S) + Puku ME (H)
Themeda	(S) + Cattle ND (D)	Themeda	(S) + Cattle ND (H,D)	Themeda	(S) + Puku ME (D)
Waltheria	(S) + Puku ME (H) + Cattle ND (D)	Waltheria	(S)	Waltheria	(S)
Chyrosopogon	(S) + Puku ME (H, D)	Chyrosopogon	(S) Puku ND (H)	Chyrosopogon	(S) + Puku ND (H)
Paspalum	(S) + Puku ME (H,D)	Paspalum	(S) + Puku ND (H,D)	Paspalum	(S) + Puku ND (H)
Gilbertiodendron	(S) + Cattle ND (H,D)	Gilbertiodendron	(S) + Cattle ND (H,D)	Gilbertiodendron	(S)
Oryza	(S) + Puku ME (H)	Ludwigia	(S) + Puku ND (H,D)	Oryza	(S) + Puku ND (D)
Ludwigia	(S) + Puku ME (H)	Cleistanthus	(S) + Cattle ND (H,D)	Ludwigia	(S) + Puku ND (H)
Digitaria	(S)			Vernicia	(S)
Cleistanthus	(S) + Cattle ND (H)				

Figure 2: a) The alpha diversity and NRI/NTI measures for each of the three populations b) The beta diversity using Bray-curtis (left) and Unifrac (right) distance measures are shown by the PCoA plots, where coloured ellipsis are the standard error. The top-25 most abundant genera for each population are also shown around the PCoA plot, with a taxa key on the bottom left hand side. The table within the figure represents the genera that were found to be significant based on the subset analysis (i.e. that explain roughly the same distance between samples as all the genera).The tables also includes information as to whether the taxa was found to be significant based on the DeSeq (represented by D) and Differential heat tree analysis (represented by H) and the direction is represented by a + for up-regulation. For example, in Puku ND vs Cattle ND the Gilbertiodendron genus was selected as significant by both the DeSeq and heat tree analysis.

3.3 Identifying the key drivers of diet variation in terms of beta diversity

3.3.1 Subset Analysis

Those genera identified as driving the differences in beta diversity between populations for pairwise comparisons are highlighted in the table in figure 2b. This table also contains information as to which populations these genera were up-regultated in, if any, as reported by the DeSeq and Heat tree analysis. Please see section 3.3.3 for a full description of this table. Since the composition of dietary components present in Puku ME and Puku ND showed very little differentiation from each other, they were also combined into one general Puku population for a separate subset analysis. This subset analysis aimed to identify the drivers at ASV level of beta diversity between Puku in general and cattle. The reduced feature set selected by the subset analysis is shown as a phylogenetic tree (Figure 3). A heat map is included to show how the abundance of the selected ASVs differs between the Puku and Cattle. Overall, 10 ASVs were identified, of which 8 were identified down to the species level. In two cases, multiple ASVs were resolved to one species, which is to be expected when working with the BLCA approach as databases are typically accurate to the genus level. The phylogenetic tree labels are colour coded according to the Family level. Firstly, the most prevalent family identified to be driving differences in beta diversity between Puku and Cattle was the Poaceae family: 6 ASVs were identified, of which 5 were resolved to species level. Of note, *Zea Mays* was identified in both the Puku and Cattle population which is Maize, a domestic is a cereal grain. *Zea mays* was present in similar abundances in both Puku and Cattle, suggesting it is not abundance of this ASV that is causing changes. Additionally, *Vaseyochloa multinervosa* was recovered; however, this species is known only to be found in Texas, USA and so this result is likely due to a database error (See Discussion).

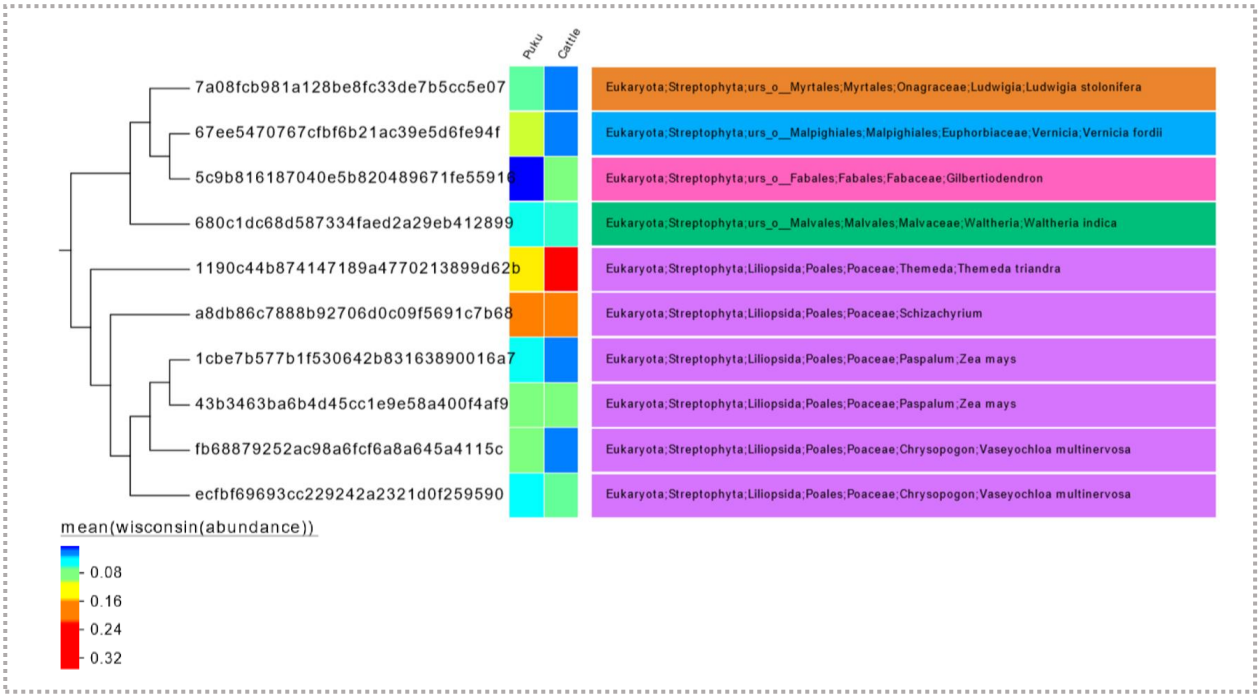


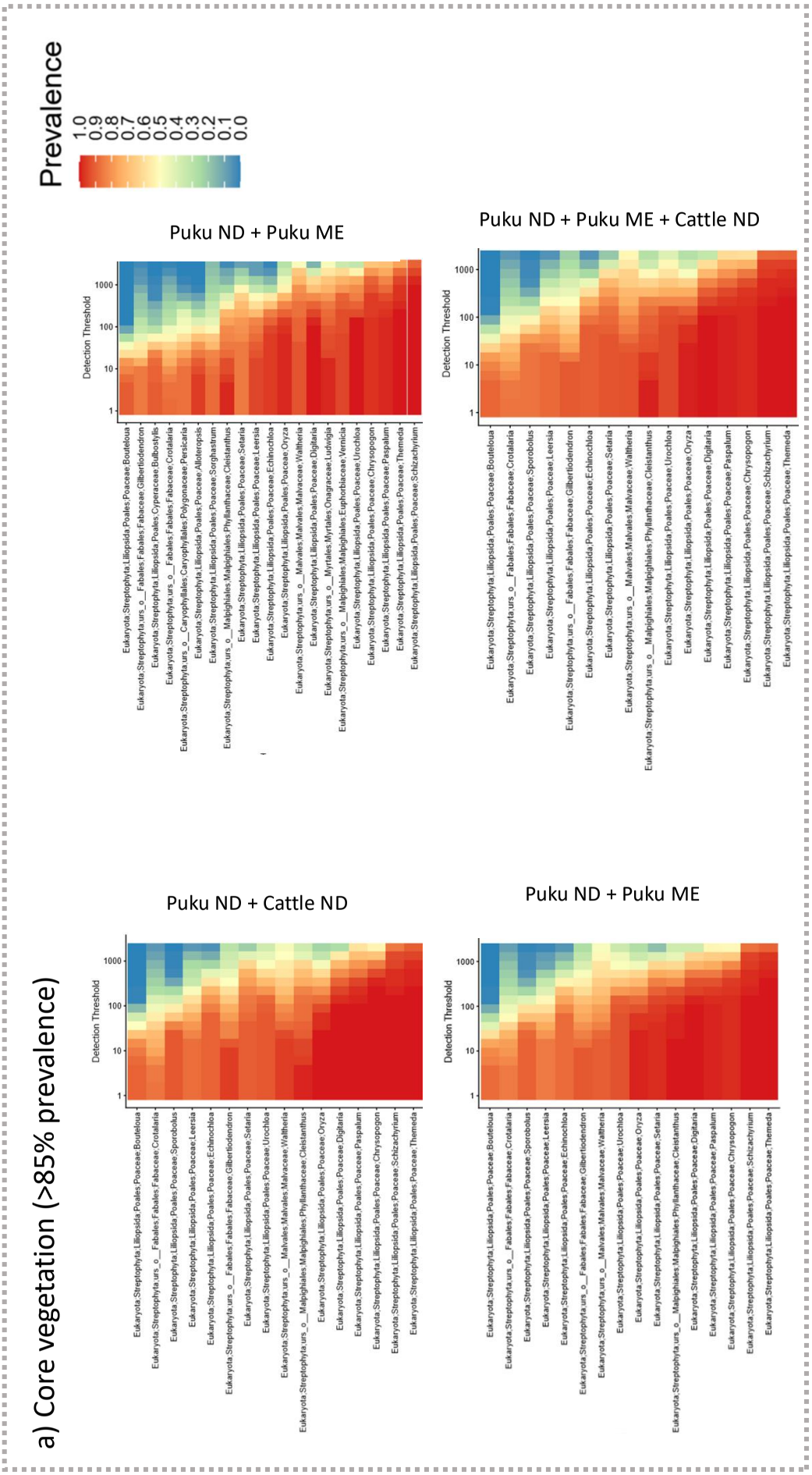
Figure 3: Subset analysis for Puku (all samples) against Cattle ND. The heat squares show how the abundance of the ASVs identified varies between Puku and Cattle populations. This subset shows the ASVs that are responsible for differences in beta diversity between Puku (in general) and Cattle.

3.3.2 Investigating core diet between Puku and cattle

The core vegetation, where genera have at least 85% prevalence, identified the dietary components common between all combinations of the three populations, as well as between Puku ME, Puku ND and Cattle ND altogether (Figure 4a). Genera are sorted by their abundance in the heat map figures, where those on top of the heat map are low abundant prevalent genera, and those at the bottom are highly abundant prevalent genera. First, looking at the core dietary items identified between the Puku ND and Puku ME populations, *Themeda, Schizachyrium, Chryospogon, Paspalum* and *Digitaria* are the top 5 most abundant prevalent genera, all of which are genera of plants within the Poaceae family. Of interest, the top 5 most abundant ASVs for the Puku ND and Cattle ND comparison and Puku ME, Puku ND and Cattle ND comparison were the same as the Puku ME and Puku ND. The Puku ME and Cattle ND comparison of the top 5 most abundant prevalent genera only had one difference in that *Urochloa* was found instead of *Digitaria. Urcochloa* is also a grass genus in the Poaceae family (Clayton and Williamson, 2006 onwards). Overall, the majority of genera identified to be of high prevalence and abundance that are common in all Puku ME, Puku ND and Cattle ND are grass types.

3.3.3 DeSeq and Heat tree

For each population comparison, the tables in Figure 2b show the genera selected by the subset analysis (annotated with S), as well as detailing whether the DeSeq and heat tree analysis also selected this genera (annotated with D and H respectively), and the direction of regulation ('+' for up-regulation and '-' for down-regulation). In some cases, genera selected by the subset analysis were not supported by either the DeSeq or heat tree analyses. In two cases, the population that was selected as unregulated by DeSeq and heat tree analysis was contradictory, and these are highlighted in bold and underlined. In this case, a conclusion could not be reached as to which population the taxa was up-regulated in. Firstly, 9 genera were identified by the subset analysis to be driving the differences between the Puku ME and Puku ND populations. Of which, 4 were supported by either the DeSeq or heat tree as being up-regulated in Puku ND. These were *Chrysopogon, Paspalum, Oryza* and *Ludwigia*. The *Ludwigia* genus contains aquatic plants and is a member of the Onagraceae family (Smith, 1987),whereas the rest are grass types of the Poaceae family. A further two genera were identified as up-regulated in the Puku ME population, the *Schizachyrium* and *Themeda grasses*. Observations of the heat tree show (Figure 4b) that there is very few branches that are highlighted as different between Puku Me and Puku ND outside of those selected by the subset analysis. Next, genera were identified that were selected by the subset analysis and up-regulated in Cattle ND in comparison to both Puku ME and Puku ND. This resulted in three genera : *Cleistanthus, Gilbertiodendron* and *Themeda.* *Cleistanthus* is a genus in the Phyllanthaceae family, a family of flowering plants (Smith, 1987). Next, comparisons of the heat trees (Figure 2b) showed that the branches up-regulated in Cattle ND were very similar when compared with both Puku ME and Puku ND. For example, the Fabales order, Lamiales order, Rubiaceae family, Asteraceae family and Vitaceae family are all highlighted as up-regulated in Cattle ND and are all orders/families of flowering plants (Chase et al., 2016). This suggests the differences between the Puku and Cattle diet are mostly driven by the higher number of flowering plants within the Cattle diet.


a) Core vegetation (>85% prevalence)
Prevalence
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Puku ND + Puku ME
Puku ND + Puku ME + Cattle ND
Puku ND + Cattle ND
Puku ND + Puku ME
Detection Threshold
1000
100
10
1
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Bouteloua
Eukaryota;Streptophyta;urs_o__Fabales;Fabaceae;Gilbertiodendron
Eukaryota;Streptophyta;Liliopsida;Poales;Cyperaceae;Bulbostylis
Eukaryota;Streptophyta;urs_o__Fabales;Fabaceae;Crotalaria
Eukaryota;Streptophyta;urs_o__Caryophyllales;Polygonaceae;Persicaria
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Alloteropsis
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Sorghastrum
Eukaryota;Streptophyta;urs_o__Malpighiales;Phyllanthaceae;Cleistanthus
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Setaria
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Leersia
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Echinochloa
Eukaryota;Streptophyta;urs_o__Malvales;Malvaceae;Waltheria
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Digitaria
Eukaryota;Streptophyta;urs_o__Myrtales;Onagraceae;Ludwigia
Eukaryota;Streptophyta;urs_o__Malpighiales;Euphorbiaceae;Vernicia
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Urochloa
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Chrysopogon
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Paspalum
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Themeda
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Schizachyrium
Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Oryza

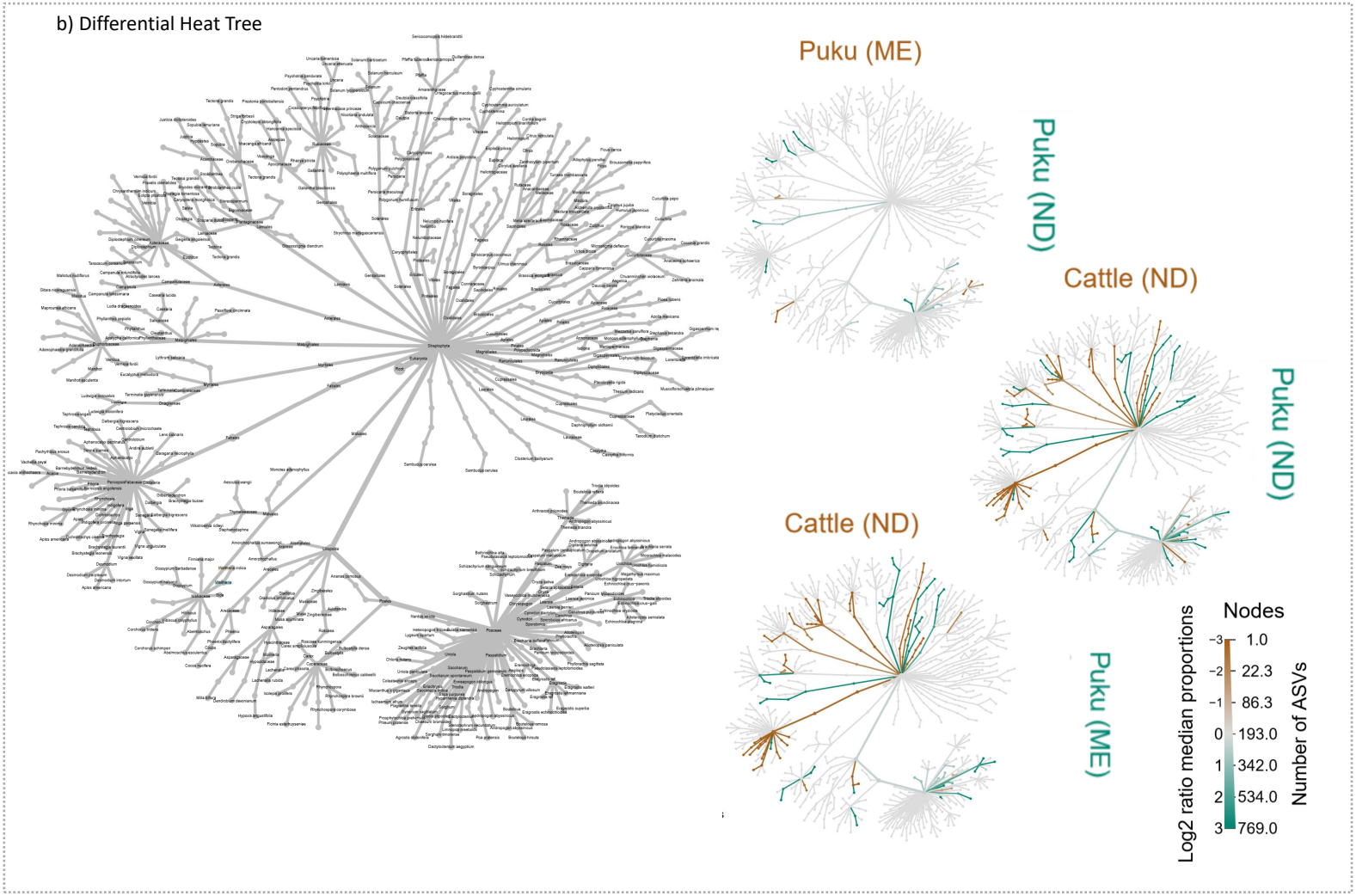


Figure 4: Similarities and differences between populations. Note to reader: The labels on the core maps and heat trees may look small but please zoom in and you will find they are clear to read and compare. (a) Core vegetation that persist in >85% of the samples for comparisons of all populations separately as well as the three different populations compared at once. ASVs are sorted by their abundance, where those at the top of heat tree are low abundance prevalent ASVs, and those at the bottom are highly abundant prevelant ASVs. (b) Differential heat trees on the right hand side with taxonomy key given on the left hand side. The differential heat trees show if taxa were selected to be up-regulated by the heat tee analysis shown by the colour. For example, in the bottom tree that compares Cattle ND and Puku ME, those branches highlighted in brown are up-regulated in the Cattle ND population and the blue up-regulated in the Puku ME population.

3.4 Null Modelling Approaches: Calculating Quantitative Process Estimate and incidence-based (Raup-Crick) beta-diversity

Ecological drivers of dietary components were found to be almost entirely as a result of ecological drift in both Puku and Cattle populations (Figure 5a). Here, high ecological drift suggest that the species abundances observed within the diet will fluctuate stochastically as a result of the environment (Stegen et al., 2015). Dispersal limitation explained the next highest amount of assembly processes. For the Cattle population, dispersal limitation accounted for more approximately three times the amount of total assembly processes that in the Puku population. Here, dispersal limitation describes a situation where a low dispersal rate is the main cause of high dietary composition turnover (Stegen et al., 2015). This suggests that the higher variation observed within the Cattle diet is a result of low dispersal rate within the environment. Homogenising selection, variable selection and homogenising dispersal accounted for very little of the total assembly processes observed in Puku and Cattle.

β RC for both Puku and Cattle were returned at close to -1 (Figure 4b). Overall, this suggests that both the Puku and Cattle populations are deterministically assembled and similar. The βRC for Puku was closer to -1, suggesting that the vegetation community found within the Puku diet was more similar to each other than that of the cattle, additionally adding support for evidence that the Puku diet is more conserved in the species present than the cattle.

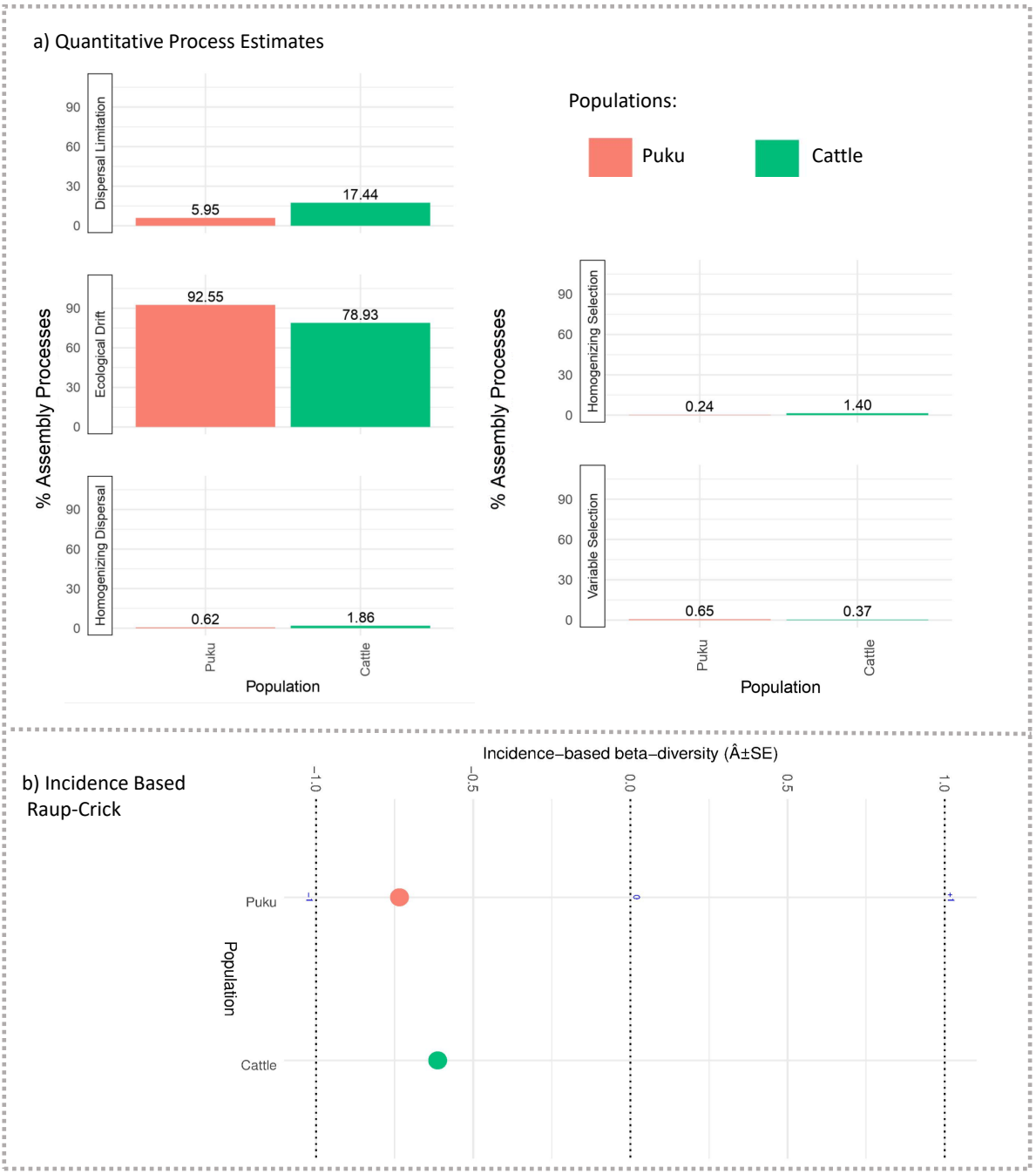
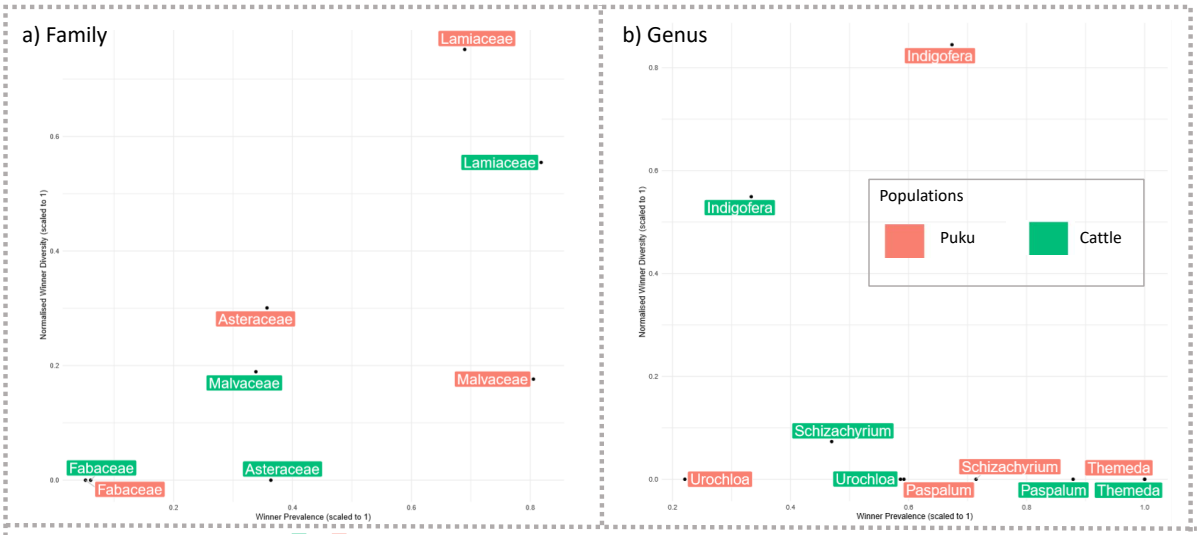


Figure 5: a) The results of the QPE. For each assembly process the percentage contribution is shown. b) The β results for both Puku and Cattle.

3.5 The Lottery Model

Lottery winners were recovered at both Family and Genus level (Figure 6a and 6b respectively), with the lottery winning ASVs detailed for each in the below table. Verster and Borenstein (2018) define those taxa with high winner diversity as $>0.25\%$ and high winner prevalence as $>0.75\%$ as genera that are more lottery-like than others. First looking at the family level, only Lamiceae is a recovered at a high winner prevalence and winner diversity for both the Puku and Cattle populations. Multiple ASVs were recovered for the Puku and Cattle population within the Lamiceae family, for both Puku and Cattle these were: *Tectona grandis* and *Otostegia tomentosa*. *Tectona grandis* is commonly known as Teak, and is a tropical decidous tree that is found in mixed forests, it has small flowers. *Otostegia tomentose* is also a flowering plant. At the genera level, there were no ASVs recovered that met the conditions for both high winner diversity and high winner prevalence. Despite this, Indigofera was recovered at fairly high prevalence within the Puku population and at very high diversity. There were four ASVs within this clade identified, two of which were identified down to species level as *Indigofera circinella*, which is also a floral plant. Overall, ASVs within the genus and families recovered by the lottery model represent those which are driving the changes in composition of the diet.



Family	Population	Winner ASV
Malvaceae	Puku	680c1dc68d587334faed2a29eb412899:Eukaryota;Streptophyta;urs_o__Malvales;Malvales;Malvaceae;Waltheria;Waltheria indica, ab8cc8735c076d7c84ad555cce87baa0:Eukaryota;Streptophyta;urs_o__Malvales;Malvales;Malvaceae;Hibiscus;Hibiscus trionum, fea82d809b29d06b2c42fae02c50afc7:Eukaryota;Streptophyta;urs_o__Malvales;Malvales;Malvaceae;Eriotheca;Eriotheca pentaphylla
	Cattle	680c1dc68d587334faed2a29eb412899:Eukaryota;Streptophyta;urs_o__Malvales;Malvales;Malvaceae;Waltheria;Waltheria indica, fea82d809b29d06b2c42fae02c50afc7:Eukaryota;Streptophyta;urs_o__Malvales;Malvales;Malvaceae;Eriotheca;Eriotheca pentaphylla
Fabaceae	Puku	0fbdddbab0d8fbec1fe41a2613686c84:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Inga;Inga paraensis
	Cattle	5c9b816187040e5b820489671fe55916:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Gilbertiodendron;
Asteraceae	Puku	8b5b626fd8c9aa07798070774422760d:Eukaryota;Streptophyta;urs_o__Asterales;Asterales;Asteraceae;Diplostephium;Diplostephiumcinereum, dca8910abe85a4366be2a321a6014643:Eukaryota;Streptophyta;urs_o__Asterales;Asterales;Asteraceae;Taraxacum;Taraxacum officinale
	Cattle	dca8910abe85a4366be2a321a6014643:Eukaryota;Streptophyta;urs_o__Asterales;Asterales;Asteraceae;Taraxacum;Taraxacum officinale
Lamiaceae	**Puku**	c282fde44e285238e8baf4c29afad063:Eukaryota;Streptophyta;urs_o__Lamiales;Lamiales;Lamiaceae;Tectona;Tectona grandis, 04b1fbdf5ce67ae60d22fef8935d8cbc:Eukaryota;Streptophyta;urs_o__Lamiales;Lamiales;Lamiaceae;Tectona;Tectona grandis, b1481dc1411c69d1d38224215b5ea873:Eukaryota;Streptophyta;urs_o__Lamiales;Lamiales;Lamiaceae;Tectona;Tectona grandis, 1f8efde167f644248bc615bda0c2c627:Eukaryota;Streptophyta;urs_o__Lamiales;Lamiales;Lamiaceae;Otostegia;Otostegia tomentosa
	Cattle	04b1fbdf5ce67ae60d22fef8935d8cbc:Eukaryota;Streptophyta;urs_o__Lamiales;Lamiales;Lamiaceae;Tectona;Tectona grandis, b1481dc1411c69d1d38224215b5ea873:Eukaryota;Streptophyta;urs_o__Lamiales;Lamiales;Lamiaceae;Tectona;Tectona grandis, 1f8efde167f644248bc615bda0c2c627:Eukaryota;Streptophyta;urs_o__Lamiales;Lamiales;Lamiaceae;Otostegia;Otostegia tomentosa, c282fde44e285238e8baf4c29afad063:Eukaryota;Streptophyta;urs_o__Lamiales;Lamiales;Lamiaceae;Tectona;Tectona grandis
Genus	**Population**	**Winner ASV**
Schizacyrium	Puku	a8db86c7888b92706d0c09f5691c7b68:Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Schizachyrium;
	Cattle	a8db86c7888b92706d0c09f5691c7b68:Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Schizachyrium;, ac413c663bf8b6b6f7234acc38efe823:Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Schizachyrium;Schizachyrium sanguineum
Themeda	Puku	1190c44b874147189a4770213899d62b:Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Themeda;Themeda triandra
	Cattle	1190c44b874147189a4770213899d62b:Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Themeda;Themeda triandra
Paspalum	Puku	43b3463ba6b4d45cc1e9e58a400f4af9:Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Paspalum;Zea mays
	Cattle	43b3463ba6b4d45cc1e9e58a400f4af9:Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Paspalum;Zea mays
Urochola	Puku	3092806ba9719e62d83378627c39d65e:Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Urochloa;Megathyrsus maximus
	Cattle	3092806ba9719e62d83378627c39d65e:Eukaryota;Streptophyta;Liliopsida;Poales;Poaceae;Urochloa;Megathyrsus maximus
Indigofera	Puku	e59d7b3c3c69d814f6b7c254fdce0cec:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Indigofera;Indigofera circinella, 71673ac8e49d45715ca31bab5cf2d878:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Indigofera;, 93a1d4cbd26b6aa5c01282a87373ec7a:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Indigofera;, fa8791c14c823feed3ac2fb4898f856e:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Indigofera;Indigofera circinella
	Cattle	fa8791c14c823feed3ac2fb4898f856e:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Indigofera;Indigofera circinella, 93a1d4cbd26b6aa5c01282a87373ec7a:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Indigofera;, 71673ac8e49d45715ca31bab5cf2d878:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Indigofera;, e59d7b3c3c69d814f6b7c254fdce0cec:Eukaryota;Streptophyta;urs_o__Fabales;Fabales;Fabaceae;Indigofera;Indigofera circinella

Figure 6: a) The Lottery winners at the Family level for both Puku and Cattle. b) The Lottery winners at the Genus level for both Puku and Cattle. The table showing the particular lottery winning ASVs for each clade is shown below the figures

3.6 Comparing Field and Genetic Data

There was very little overlap between transect data and genetic data, which imposed limitations on the comparisons that could be performed (Figure 6). Overall, 743 ASVs were resolved to species level classification via the genetic approach, and only 23 plants were identified in the transects to species level. For the ME site, there were 19 speices identified in total by the transect approach, and only 3 were found to be supported with genetic data. For the ND site, 15 species were identified, with 2 supported by genetic data. The genetic approach recovered 349 species to be commonly found in both ME and ND, whereas the transect approach recovered only 10 to be shared between the sites.

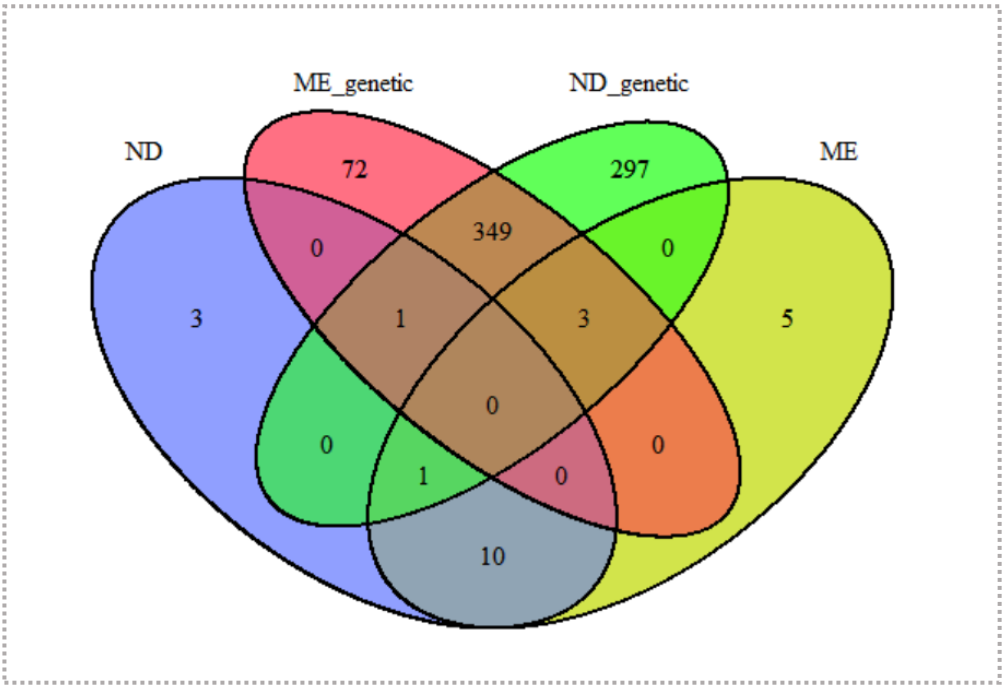


Figure 7: Overlap of species identified by the transect and genetic approach at site ND and ME, showing the genetic approach to have yielded a higher number of species classified.

For each method and site, rank abundance plots revealed that species were identified at variable abundances (Figure 7). Comparison of the five most abundant species present in the ND and ME site by the genetic and transect approach showed no commonalities. The transect approach identified two species, *Hyparrhenia filipendula* and *Panicum maxium*, as highly abundant in both ND and ME sites, which are both grasses within the Poaceae family. The genetic approach identified two species to be highly abundant in all three ecosystems, *Themeda triandria* and *Zea mays* (commonly known as maize), again both Poaceae family members. Of note, *Oryza sativa*, the rice crop, was recovered within the ME site diet as

highly abundant, however; the ME site is classified as far away from agricultural activities so the presence of cultivated crops is surprising.

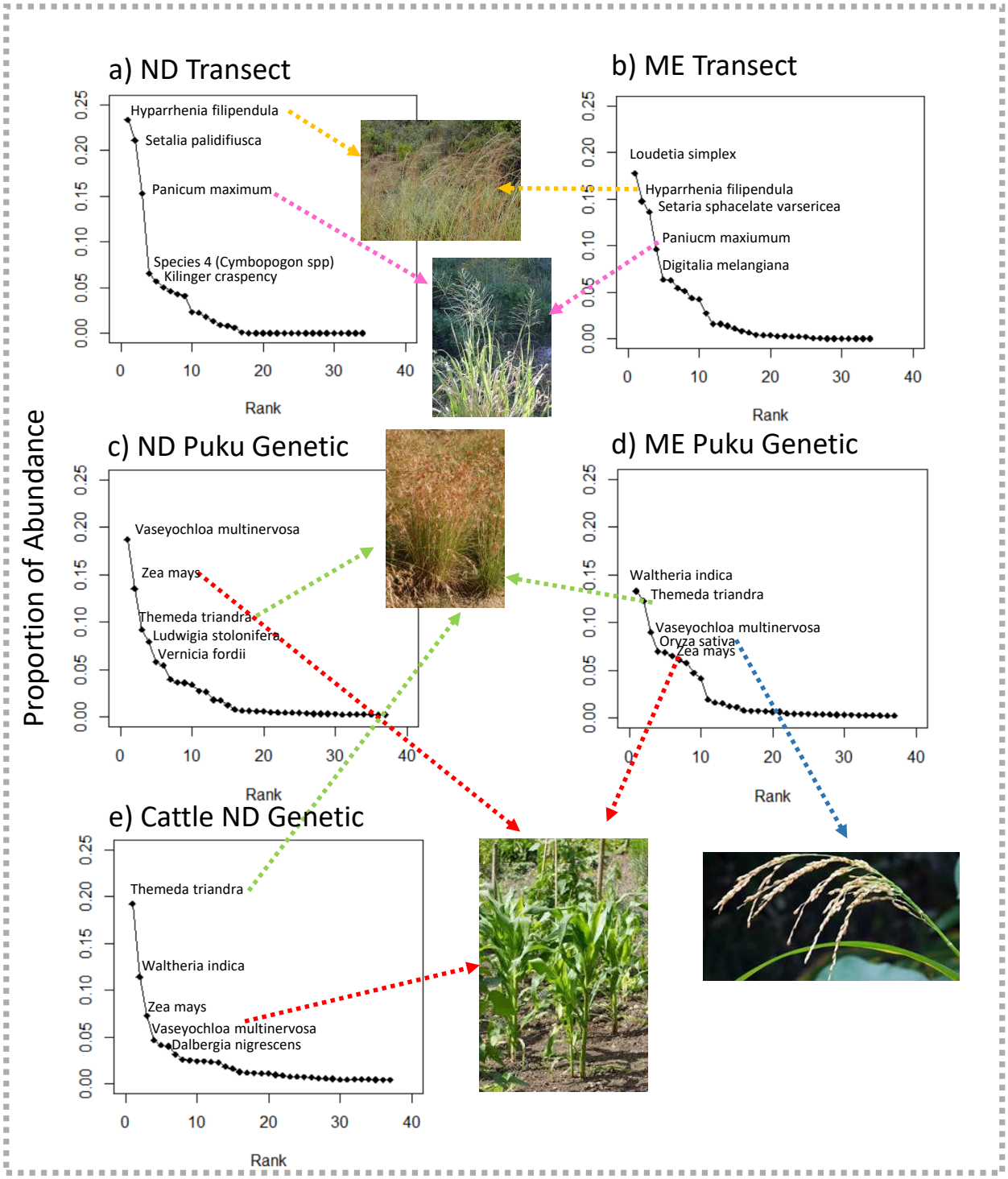


Figure 8: Top 5 most abundant species recovered in ME and ND by transect and genetic approach. Images displayed represent species that were found identified by multiple methods or are of specific note.

4 Discussion

4.1 Optimal bioinformatic approach for diet analysis

4.1.1 Performance of trnL as a barcode

The trnL (UAA) intron used here performed well overall as a barcode for identifying plant taxa in degraded DNA samples; however there were some limitations to its use. Taberlet et al. (2007) described the trnL (UAA) intron as recovering lower resolution results than other non-coding chloroplast regions, and therefore is limited in its ability to characterise closely related plant species and for phylogenetic analyses. Evidence of this limitation was found within the analyses presented here as on few occasions, species identifications were made for species that are known to not be found within the ecosystem samples sites. For example *Vaseyochloa multinervosa* was recovered at high abundance in all three populations, but this grass is known to be native to Texas(Clayton and Williamson, 2006 onwards). *Vernicia fordii*, which is native to Vietnam and China (Clayton and Williamson, 2006 onwards) was also recovered. These results are therefore likely an artefact of trnL being unable to discriminate between closely related taxa. This highlights the importance of checking the results obtained by sequencing methods to ensure they make sense in terms of the ecosystem studied. The workflow developed here is highly flexible and would work with any plant taxa barcode that has an extensive database. In future studies, it would therefore be of interest to conduct this workflow with the same samples sequenced at multiple barcodes, to allow for comparison of taxonomic inferences between variable barcode types. For example, both the rbcL and matK barcode are popular within plant metabarocding literature and so would be of interest to investigate (Kang et al., 2017). The main advantage that these barcodes provide over trnL is that they both have more extensive databases and (Mallott et al., 2018) additionally, rbcL has been widely used for phylogenetic analysis and so may provide better resolution than trnL for the phylogeny analyses described here (Kang et al., 2017).

4.1.2 Amplicon Sequence Variants

Using ASVs over OTUs had the advantage that we did not have to assign an arbitrary similarity threshold to deduce sequences. Here, I used the DADA2 software to deduce ASV sequences and found it worked

well at reducing ASVs that are not expected. The ASV approach is gaining popularity in within microbiome studies; however the field of diet analysis has been slower in adapting to it and so there are a lack of studies within the current literature which have taken as ASV approach. Ideally, in this study I would have been able to perform a comparison between the ASV and OTU results, but time constraints did not allow for the completion of two pipelines. Future studies could therefore aim to investigate how end taxonomic assignments may vary between the ASV and OTU approach. such as in (Callahan et al., 2017).

4.1.3 Taxonomic assignment

Taxonomic assignment of sequences is one of the main goals in diet analysis studies but there is lack of uniformity in the literature for the most appropriate method and the choice of assignment method can alter the overall biological inference (O'Rourke et al., 2020). The BLCA approach used here resulted in a three-fold increase in taxonomic assignment at the genus level, and eight-fold increased at the species ASV level when compared with the naive classifier method. To the best of my knowledge, this is the first time a BLCA approach has been applied to degraded faecal samples for diet analysis. The biggest limitation of the BLCA approach, was the lack of a complete trnL database for African taxa. Approximately 80% of ASVs were unassigned using the BLCA approach, suggesting that a lot of the taxa present in the ecosystem are not present within the database. I used the trnL reference database provided on: https://github.com/RTRichar/MetabarcodeDBsV2 which was primarily a database for North American plant species, to my knowledge a trnL database is not currently widely available that focuses purely on African plant taxa. O'Rourke et al. (2020) evaluated a range of classifiers for dietary analysis on biological mock communities and concluded that overall the best approach taken to assign taxonomy is by using a range of assignment methods. Therefore, although BLCA approach performed well it would be of interest to have included a wider range of taxonomic assignment methods to compare against.
Additionally, this study had limited transect data available and so comparisons between genetic and field data were difficult due to the inequality in the number of taxa recovered by each method (24 species identified via observation vs 743 identified using genetics). Although this demonstrates the power of using genetics to infer diet over observation, it would be useful to have more information on the plant species

within the ecosystem to be able to decipher if Puku are preferentially foraging on certain items. This also made it difficult to validate the species recovered by genetics and so would provide further support for the inferences made by the genetic approach if we could provide evidence that the plant species recovered are present within the ecosystem.

4.1.4 Inclusion of Null Modelling, Lottery Modelling

Null modelling was an informative tool for determining the environmental processes that are influencing the dietary components recovered. To the best of my knowledge, this project is the first to report on the potential applications of null modelling in dietary studies. I specifically focused on investigating quantitative process estimates and RC to determine the assembly processes influencing the diet as well as the degree of stochasicity/determinstic processes. Additionally, lottery modelling has typically been used in habitat selection or microbiome studies (Cooper et al., 2008) to determine the ASVs that are key to habitat assembly and has to date not been investigated for use in dietary studies. I therefore suggest that diet analysis studies should aim to incorporate these methods, as the ability to compare results from diet analysis studies from various species will be very useful for determining key species and informing conservation strategies.

4.2 Puku diet in different sites

This is the first dietary analysis study that investigates Puku diet via a deep sequencing approach. This project provides evidence that the Puku diet is consistent across the two sample sites investigated here in the Kilombero Valley, despite one site having a high cattle and agricultural presence. Specifically, both the ME and ND site recovered a high abundance of genera belonging to the Poaceae family, suggesting the Puku diet in these regions are primarily grass based. Some of the grass species identified here are of particular note, such as *Oryza sative* (rice crop) and *Zea mays*(maize crop) as their presence at high abundance indicates Puku are commonly foraging in agricultural areas. It is particularly interesting that agricultural items are recovered in the diet of the Puku at the ME site as this site has very low agricultural presence. One possibility could be that Puku within this site are travelling to forage; however to date there is a lack of research on Puku movements for foraging to confirm this. Additionally, we identified *Tectona*

gondis (Teak) within the Puku diet, which is of interest, as the establishment of Teak plantations in the Kilombero Valley is thought to impact the Puku population, but previous studies that used spoor transects failed to confirm this (Bonnington et al., 2009). There are no studies within the literature that focus specifically on Puku diet within the Kilombero Valley to date, and so comparisons have to be made with studies which have been undertaken in other Puku sites in Africa. Mircohistological analyses of Puku faecal samples were conducted by Rduch (2016), in the Kasanka National Park, Zambia, an ecosystem that is characterised primarily by Mimbio woodlands and has a small proportion of grasslands and floodplain environment. Rduch (2016) characterised the Puku diet to mainly consist of monoctyledons and identified few grass taxa present. Of the grass taxa that were identified, *Panicum spp., Brachiaria spp., Sporobolus spp.*, and *Hyparrhenia/Andropogon* were classified as the most abundant, which were identified in the Puku diet in my study but are very low abundance. In the analysis presented here, several flowering plant genera were identified at high abundance, but the abundance of grass taxa were far more. One potential explanation for the difference in taxonomic dietary classifications of my study and that of Rduch (2016) could be the use of the different methods. In microhistological analyses, there is a tendency for shrub and floral plants to be overestimated, as these plants are digested less well than grass types and so may be over-represented in faecal samples (King and Schoenecker, 2019). Another potential explanation could be due to differences in plants available in the environment. The Kilombero Valley fringes onto the Mimbio Forest, but is primarily grasslands (Jenkins et al., 2003) and agriculture in the ND site. To determine if these differences are due to technique or if Puku diets differ across countries, future studies could compare Puku across multiple, sites.

In the PERMANOVA analysis conducted here we found month to be a significant predictor of variation in the ASVs observed within the diet of each population. The data used here was collected between November and April, which, in Tanzania is primarily the dry season, with March and April being the beginning of the wet season (Jenkins et al., 2003). Floodplain ungulates, such as Puku, are thought to face the greatest limitation of food during the late wet season, when large proportions of flood-plains are cut off, making food resources unavailable and so they move on to boundary zone habitats, where cattle are found (O'Shaughnessy et al., 2014; Jenkins et al., 2003). It would therefore be of interest to apply the workflow presented here to samples collected throughout the year that can then investigate the difference

in Puku diet as the seasons change and Puku are forced further into boundary zones. O'Shaughnessy et al. (2014) found that Puku dietary overlap with bovids in Botswanna varied between the wet and dry season via an observational approach.

4.3 Overlaps and differences between Cattle and Puku diet

This is also the first time, to my knowledge that the cattle diet within the Kilombero Valley has been determined using sequencing. Cattle diet was found to primarily consists of monocots and grasses. Specifically, *Gilbertiodendron*, and *Themeda* were consistently recovered as highly abundant. Kartzinel et al. (2015) conducted a study using a similar trnL metabarcoding approach in the Kenyan savanna to classify the diet of many herbivores including Cattle; they also found that grasses and flowering plants were the main components of the Cattle diet.

The comparison between diets of Puku and Cattle revealed several similarities and differences. Firstly, in terms of diversity, the cattle diet was found to be much more variable, in that the number of ASVs recovered was consistently higher than the Puku. This was supported by the null modelling approach which suggested that the dietary community of the Puku was more similar than in the cattle and that a low dispersal rate is the main cause of high dietary composition turnover in cattle. Altogether, this suggests that the dietary niche of the cattle is broader than that of the Puku. The main overlaps between the Puku and Cattle populations were found between grass genera consumed. Namely; *Themeda, Schizachyrium, Chyrosopogon, Paspalum* and *Digitaria* were found to be both highly present and abundant in Puku and Cattle. Identification of these genera gives a starting point for land managers to begin to better understand how cattle presence is affecting Puku. It is also of note, that the genera that are shared between Puku and Cattle are also those genera which were found to be the main dietary components for Puku, but not for cattle and so even though an effect of cattle encroachment was not found in this study, there is still evidence to suggest cattle could eventually deplete shared resources if the presence of cattle in this area continues to increase. Bonnington et al. (2007) investigated effects of livestock encroachment on Puku on the Kilombero Valley by observing puku and cattle presence in stoor transects; they observed that areas that are heavily grazed by livestock are used to a lesser extent by Puku, and that areas previously used by Puku will be avoided following livestock encroachment. With this in mind, there is the poten-

tial that Puku could abandon sites that are shared with cattle, such as those we have identified here, and therefore there overall range will decrease. The main differences observed between the Puku populations and cattle were in terms of the abundance of monocots present within the diet, where cattle were typically found to eat a higher abundance and a wider range of genera. Identifying that cattle are consuming mostly monocots is of interest as this could be a primary reason as why I have not observed Puku diet to consists of mostly monocots, as the (Rduch, 2016) study took place in the absence of cattle. Until, further knowledge is obtained on the overall range of Puku diet and preferences that exist between populations it is difficult to determine based on the current literature if cattle are out-competing Puku for monocots in this area.

Of note, variation in cattle presence throughout the year as a result of the hunting season was not accounted for in this study. In the Kilombero Valley, the hunting season takes place from July to December; during this time, wildlife areas are patrolled by guards which protect the areas from livestock encroachment (Bonnington et al., 2007). For the remaining year, there is no trophy hunting and so few guards are present within the Kilombero valley and do not protect against livestock encroachment during this time (Bonnington et al., 2007). Therefore, for two of the six months that faecal samples were collected for in this study, the Puku were guarded. Future studies should therefore ensure that guard presence and overall cattle density is accounted for within the study as the presence of guards protecting against encroachment could be causing a change in the Puku diet.

5 Conclusions

Here, I have presented a workflow that demonstrates the power and limitations of novel bioinformatic techniques for the study of diet in elusive herbivorous species by non-invasively collected low quality samples. Specifically, this project provides further evidence that trnL as a barcode is limited in discriminatory power when it comes to identifying between closely related taxa, as all classifications returned to genus level were accurate for the ecosystem, but at species level some results were returned that did not match the environment. I have demonstrated the power that the BLCA taxonomic assignment approach has over the traditional naive classifier approach for achieving higher resolution down to genus and species level. I have also provided support for using DADA2 software for ASV inference over the

traditional OTU approach and suggest that this method of sequence determination should be incorporated into more diet analysis studies. This was also the first diet analysis workflow that incorporated null modelling and lottery modelling and I found they provide informative results as to how the environment is affecting dietary components.

Analysis of Puku diet in two sites in Kilombero Valley using this approach revealed that the diet across the two sites does not differ. This suggests, that cattle encroachment is not negatively affecting the Puku population. By also analysing the cattle diet using this workflow I was able to identify components of the diets that do overlap between the Puku and cattle, which gives an indication into potential range overlap areas. This was the first study which has investigated the Puku diet using a sequencing approach and has classified more genera than any previous study.

Overall, the bioinformatic pipeline I have presented here is widely applicable to other herbivorous species as a non-invasive tool to investigate diet and classify dietary overlaps. Within the Kilombero Valley, there are numerous other species that could benefit from dietary classification from this method such as; the elephant *(Loxondonta africana)* classified as vulnerable; buffalo *(Syncerus caffer)* classified as near threatened, common zebra *(Equus quagga burcheli)* classified as least concern, sable antelope *(Hippotragus niger)* classified as least concern; and the eland *(Taurotragus oryx)* classified as least concern (East, 1999). In addition to extending the wildlife populations included in this study, there is also evidence to suggest that other livestock within the Kilombero Valley can can encroach wildlife areas and so could be incorporated. For example, Bonnington et al. (2007) investigated how both Cattle and Goats affected the presence of Puku, Buffalo and Elephant by inspecting spoors within transect areas, and found that both cattle and goats presence was negatively correlated with wildlife presence. By extending the study presented here to include more species will ultimately lead to a more in-depth understanding of the ecosystem functioning within the Kilombero Valley and ensure that conservation interventions will not negatively affect any species. In addition, this method could additionally be applied to any ecosystem where herbivorous species are present.

In summary, this project highlights the need for further case studies that investigate bioinformatic tools that could enhance the resolution of data obtained from low-quality samples and thus the inferences that can be made. By harnessing the optimal power of available techniques, a more in-depth understanding

of diet, and overall ecosystem functioning of endangered species will be achieved. This is particularly important now as many species are on the brink of loosing habitats as a result of rapidly changing environments. Pipelines, like the one I have outlined here, will play an essential role in determining the key areas and dietary items within an ecosystem that need to be conserved to protect animal species.

References

Alberdi A, Aizpurua O, Bohmann K, Gopalakrishnan S, Lynggaard C, Nielsen M, Gilbert MTP (2019) Promises and pitfalls of using high-throughput sequencing for diet analysis. Molecular Ecology Resources 19:327-348

Balestrieri A, Remonti L, Prigioni C (2011) Assessing carnivore diet by faecal samples and stomach contents: A case study with Alpine red foxes. Central European Journal of Biology 6: 283-292

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CCe (2019) Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. nature biotechnology 37:852-857

Bonnington C, Weaver D, Fanning E (2007) Livestock and large wild mammals in the Kilombero Valley, in southern Tanzania. African Journal of Ecology 45:658–663

Bonnington C, Weaver D, Fanning E (2009) The use of teak (tectona grandis) plantations by large mammals in the kilombero valley, Southern Tanzania. African Journal of Ecology 47:138–145

Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern wisconsin. Ecological Monographs 27:325–349

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13: 581–583

Callahan BJ, McMurdie PJ, Holmes SP (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11: 2639–2643

Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK (2019) High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Research 47:e103

Campbell O Webb MAM David D Ackerly, Donoghue MJ (2002) Phylogenies and community ecology. Annual Review of Ecology and Systematics 33:475-505

Carroll EL, Bruford MW, DeWoody JA, Leroy G, Strand A, Waits L, Wang J (2018) Genetic and genomic mnitoring with minimally invasive sampling methods. Evolutionary Applications 11:1094–1119. Evolutionary Applications

Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM (2015) Accelerated modern human-induced species losses: Entering the sixth mass extinction. Science Advances 1:1400253

Chase MW, Christenhusz MJ, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF, Briggs B, Brockington S, Chautems A, Clark JC, Conran J, Haston E, Möller M, Moore M, Olmstead R, Perret M, Skog L, Smith J, Tank D, Vorontsova M, Weber A (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society 181:1–20

Chen H (2018) VennDiagram: Generate High-Resolution Venn and Euler Plots. BMC Bioinformatics 12, 35

Choo JM, Leong LE, Rogers GB (2015) Sample storage conditions significantly influence faecal microbiome profiles. Scientfic Reports 5:16350

Clayton VMHK WD, Williamson H (2006 onwards) Grassbase - the online world grass flora. http://www.kew.org/data/grasses-db.html. accessed 16th august 2020

Cooper N, Rodríguez J, Purvis A (2008) A common tendency for phylogenetic overdispersion in mammalian assemblages. Proc. R. Soc. B.2752031–2037

Dove H, Mayes RW (1996) Plant wax components: A new approach to estimating intake and diet composition in herbivores. The Journal of Nutrition. 126:13-26

East R (1999) African antelope database 1998

Estrella MM D, Wieringa JJ, Mackinder B, van der Burgt X, Devesa JA, Bruneau A (2014) Phylogenetic analysis of the african genus gilbertiodendron J. Léonard and related genera (Leguminosae-Caesalpinioideae-Detarieae). International Journal of Plant Sciences 175:975–985

Fisher ASW R A; Corbet (1972) The relation between the number of species and the number of individuals in a random sample of an animal population. journal of animal ecology 12: 42–58

Flojgaard C, De Barba M, Taberlet P, Ejrnaes R (2017) Body condition, diet and ecosystem function of red deer (cervus elaphus) in a fenced nature reserve. global ecology and conservation 11:312-323

Foster Z, Sharpton T, Grünwald N (2017) Metacoder: An r package for visualization and manipulation of community taxonomic diversity data. PLOS Computational Biology 13:1–15

Gao X, Lin H, Revanna K, Dong Q (2017) A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. BMC Bioinformatics 18:247

Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. Journal of Statistical Software 22:1–19

Hunter ME, Hoban SM, Bruford MW, Segelbacher G, Bernatchez L (2018) Next-generation conservation genetics and biodiversity monitoring. Evolutionary Applications. 11:1029–1034.

Jenkins RK, Maliti HT, Corti GR (2003) Conservation of the puku antelope (Kobus vardoni, Livingstone) in the Kilombero Valley, Tanzania. Biodiversity Conservation 12:787–797

Kang Y, Deng Z, Zang R, Long W (2017) DNA barcoding analysis and phylogenetic relationships of tree species in tropical cloud forests. Scientific Reports 7:12564

Kartzinel TR, Chen PA, Coverdale TC, Erickson DL, Kress WJ, Kuzmina ML, Rubenstein DI, Wang W, Pringle RM (2015) DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. Proceedings of the National Academy of Sciences 112:8019-8024

Katoh K, Standley D (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30:772-80

Kembel S, Cowan P, Helmus M, Cornwell W, Morlon H, Ackerly D, Blomberg S, Webb C (2010) Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464

King SR, Schoenecker KA (2019) Comparison of Methods To Examine Diet of Feral Horses from Non-invasively Collected Fecal Samples. Rangeland Ecology and Management 72:661–666

Lahti L, Shetty S (2017) microbiome r package. tools for microbiome analysis in r. version 2.1.26. url: http://microbiome.github.com/microbiome.

Lopes CM, De Barba M, Boyer F, Mercier C, Da Silva Filho PJ, Heidtmann LM, Galiano D, Kubiak BB, Langone P, Garcias FM, Gielly L, Coissac E, De Freitas TR, Taberlet P (2015) DNA metabarcoding diet analysis for species with parapatric vs sympatric distribution: A case study on subterranean rodents. Heredity 114: 525–536

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology 15:550

Lozupone C, Knight R (2005) UniFrac: A new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71:8228–8235

Lozupone CA, Hamady M, Kelley ST, Knight R (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Applied and environmental microbiology 73:1576-85

Mallott EK, Garber PA, Malhi RS (2018) Trnl outperforms rbcl as a DNA metabarcoding marker when compared with the observed plant component of the diet of wild white-faced capuchins (cebus capucinus, primates). PLoS ONE 13:e0199556

McMurdie PJ, Holmes S (2013) phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8:e61217

Ogden R, Dawnay N, McEwing R (2009) Wildlife DNA forensics - Bridging the gap between conservation genetics and law enforcement. Endangered Species Resources, 9: 179–195

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2019) vegan: Community Ecology Package, R package version 2.5-6

O'Rourke DR, Bokulich NA, MacManes MD, Foster JT (2020) A total crapshoot? Evaluating bioinfor-

matic decisions in animal diet metabarcoding analyses. Ecology and Evolution: Early View. Accessed on 19/08/2020

O'Shaughnessy R, Cain JW, Owen-Smith N (2014) Comparative diet and habitat selection of puku and lechwe in northern Botswana. Journal of Mammology 95:933–942

Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528

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

Pielou EC (1966) The measurement of diversity in different types of biological collections. Journal of Theoretical Biology 13:131–144

Pompanon F, Deagle BE, Symondson WO, Brown DS, Jarman SN, Taberlet P (2012) Who is eating what: Diet assessment using next generation sequencing. Molecular Ecology 21:1931-1950

Porter TM, Hajibabaei M (2018) Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Molecular Ecology 27:313-338

Price M, Dehal P, Arkin A (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Prins HHT (2000) Competition Between Wildlife and Livestock in Africa. In: Wildlife Conservation by Sustainable Use

RCoreTeam (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Rduch V (2016) Diet of the puku antelope (Kobus vardonii) and dietary overlap with selected other bovids in Kasanka National Park, Zambia. Mammal Research 61: 289–297

Rosser AM (1992) Resource distribution, density, and determinants of mate access in puku. Behavioral Ecology 3:13–24

Schliep K, Potts AJ, Morrison DA, Grimm GW (2017) Intertwining phylogenetic trees and networks. Methods in Ecology and Evolution 8:1212-1220

Shannon CE (1948) A mathematical theory of communication. the bell system technical journal, 27, 379–423 and 623–656

Shehzad W, Riaz T, Nawaz MA, Miquel C, Poillot C, Shah SA, Pompanon F, Coissac E, Taberlet P (2012) Carnivore diet analysis based on next-generation sequencing: Application to the leopard cat (prionailurus bengalensis) in pakistan. molecular ecology 21:1951-65

Shendure J, Ji H (2008) Next-generation DNA sequencing, Nature Biotechnology 26: 1135–1145

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH (2017) DNA sequencing at 40: Past, present and future. Nature 550: 345–353

Shetty SA, Hugenholtz F, Lahti L, Smidt H, de Vos WM (2017) Intestinal microbiome landscaping: Insight in community assemblage and implications for microbial modulation strategies. FEMS microbiology reviews 41:182-199. FEMS Microbiology Reviews

Simpson EH (1949) Measurement of diversity. nature. 163 (4148): 688

Smith A (1987) Flora of tropical East Africa. Euphorbiaceae.-(Part 1). Balkema, Rotterdam, Netherlands

Stegen JC, Lin X, Fredrickson JK, Konopka AE (2015) Estimating and mapping ecological processes influencing microbial community assembly. Frontiers in Microbiology 6:370

Swift JF, Lance RF, Guan X, Britzke ER, Lindsay DL, Edwards CE (2018) Multifaceted DNA metabarcoding: Validation of a noninvasive, next-generation approach to studying bat populations. Evolutionary Applications 11:1120–1138.

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

Taylor M (2017) sinkr: Collection of functions with emphasis in multivariate data analysis. Available: https://github.com/menugget/sinkr [Accessed]

Valentini A, Christian M, Muhammad Ali N, Eva B, Eric C, François P, Ludovic G, Corinne C, Giuseppe N, Patrick W, Jon E S, Pierre T (2009a) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. Molecular Ecology Resources 9:51-60

Valentini A, Pompanon F, Taberlet P (2009b) DNA barcoding for ecologists. Trends in Ecology and Evolution 24:110-7

Vass M, Székely AJ, Lindström ES, Langenheder S (2020) Using null models to compare bacterial and microeukaryotic metacommunity assembly under shifting environmental conditions. Scientific Reports 10:2455

Vatanparast M, Klitgård BB, Adema FA, Pennington RT, Yahara T, Kajita T (2013) First molecular phylogeny of the pantropical genus Dalbergia: Implications for infrageneric circumscription and biogeography. South African Journal of Botany 89:143–149

Verster AJ, Borenstein E (2018) Competitive lottery-based assembly of selected clades in the human gut microbiome. Microbiome 6:186

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73:5261-5267

Westcott SL, Schloss PD (2015) De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ 8:e1487

Whittaker RH (1972) Evolution and measurement of species diversity. taxon, 21: 213–251

Wultsch C, Waits L, Hallermna E, Kelly M (2015) Optimizing collection methods for noninvasive genetic sampling of neotropical felids. wildlife society bulletin 39: 403-412

Appendix I: Reference Library Construction

The below steps show how the reference library was constructed in the Linux environment.

Step 1: Enable Qiime2 on the Orion Cluster

```
[MScBioinf@becker /shared5/Rachel/trnL]$ export PATH=/home/opt/
miniconda2/bin:$PATH
[MScBioinf@becker /shared5/Rachel/trnL]$ source activate qiime2-
2019.7
```

Step 2: Import all the sequences from the trnL database to Qiime2:

```
(qiime2-2019.7) [MScBioinf@backer /shared5/Rachel/trnL]$ qiime tools
import --type
'FeatureData[Sequence]' --input-path trnL.fa --output-path trnL.qza
```

Step 3: The trnL database contained duplicates in taxonomy which Qiime2 complains about and so these were removed.

```
(qiime2-2019.7) [MScBioinf@becker /shared5/Rachel/trnL]$ awk
'!seen[$1]++' trnL.tax
> trnL_removed_duplicates.tax
```

Step 4: The taxonomy was then imported in Qiime2 format:

```
(qiime2-2019.7) [MScBioinf@becker /shared5/Rachel/trnL]$ qiime tools
import --type
'FeatureData[Taxonomy]' --input-format HeaderlessTSVTaxonomyFormat
--input-path
trnL_removed_duplicates.tax --output-path trnL-taxonomy.qza
```

Step 5: Generate the classifier to be used with Qiime2

```
(qiime2-2019.7) [MScBioinf@becker /shared5/Rachel/trnL]$ qiime
feature-classifier
fit-classifier-naive-bayes --i-reference-reads trnL.qza --i-
reference-taxonomy
trnL-taxonomy.qza --o-classifier trnL-classifier.qza
```

Appendix II: Bioinformatic Workflow for Qiime2 with DADA2

The full Qiime2 with DADA2 workflow is shown below. This workflow is adapted from that available at: https://github.com/umerijaz/tutorials/blob/master/qiime2_tutorial.md . Steps 1-7 were common between the two taxonomy approach. Then steps 8-10 are specific to the naive baysesian classifier approach and steps 11 - 18 are specific to the BLCA approach.

FASTQ files were first organised so that they were in the correct format needed for analysis in the Qiime2 and DADA2 workflow . To achieve this both forward and reverse FASTQ files were moved into a 'Raw' folder for each sample in the Linux environment via the steps shown below:

```
[MScBioinf@becker /shared5/Rachel/KogganiResult/00.RawData/D1]$ mkdir
Raw
[MScBioinf@becker /shared5/Rachel/KogganiResult/00.RawData/D1]$ mv
*.fq Raw
[MScBioinf@becker /shared5/Rachel/KogganiResult/00.RawData/D1]$ cd
Raw
[MScBioinf@becker /shared5/Rachel/KogganiResult/00.RawData/D1/Raw]$
ls FDMP2000H000282-1a_L1_D1_1.fq FDMP20H000282-1a_L1_D1_2.fq
```

Then, once files for all 143 samples had been organised the workflow was undertaken.

Step 1: Copy FASTQ files

```
[MScBioinf@becker /shared5/Rachel/KogganiResult/00.RawData]$
d="/shared5/Rachel/KogganiResult/00.RawData/";
```

Step 2: Make fictitious barcodes

```
[MScBioinf@becker /shared5/Rachel/KogganiResult/00.RawData]$
(echo -e "sample-id\tforward-absolute-filepath\treverse-absolute-
filepath";for i in $(ls $d); do R1=$(ls ${d}$i/Raw/*_1.fq); R2=$(ls
${d}$i/Raw/*_2.fq); echo -e "$i $R1 $R2"; done) > pe-33-manifest
```

Step 3: Activate Qiime2 on Limux

```
[MScBioinf@becker /shared5/Rachel/KogganiResult/00.RawData]$ export
PATH=/home/opt/miniconda2/bin:$PATH
```

```
[MScBioinf@becker /shared5/Rachel/KogganiResult/00.RawData]$ source
activate qiime2-2019.7
```

Step 4: Move paired end data

```
(qiime2-2019.7) [MScBioinf@becker /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2] tools import --type 'Sample-
Data[PairedEndSequencesWithQuality]' --input-path
pe-33-manifest --output-path demux.qza --input-format PairedEndFastq-
ManifestPhred33V2
```

Step 5: Demultiplex

```
(qiime2-2019.7) [MScBioinf@becker /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ qiime demux summarize --i-data ./demux.qza
--o-visualization ./demux.qzv
(qiime2-2019.7) [MScBioinf@becker /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ qiime tools export --input-path demux.qzv
--output-path output
```

Next the forward and reverse reads were downloaded, and then dragged and dropped on to the Qiime2 viewer available at: https://view.qiime2.org. Reads were visually checked for quality and decided where to cut off the forward and reverse reads where the quality drops down significantly.

Figure 9: Qiime2 Viewer demultiplexed sequence counts summary

Table 1: The Sequence count for each Sample as was shown in the Qiime2 viewer

Sample Name	Sequence Count	Sample Name	Sequence Count	Sample Name	Sequence Count
D20	168129	D134	167724	D37	166844
D23	165957	D48	164028	D59	163743
D44	162519	D168	161862	D45	160741
D97	160741	D170	159986	D24	158786
D47	159405	D72	159051	D85	155747
D13	155376	D196	154690	D21	153589
D38	153452	D158	152898	D87	152420
D84	151009	D194	151009	D75	143446
D110	142300	D60	141277	D73	140756
D14	138920	D116	137261	D140	135888
D43	134991	D19	134289	D35	133261
D180	131768	D146	131440	D93	124851
D169	120078	D62	117834	D182	116073
D40	113577	D80	112848	D81	112311
D141	112154	D86	110509	D142	107930
D39	107653	D15	107574	D92	106553
D16	106382	D50	105992	D74	105896
D22	102526	D200	100789	D46	100165
D36	97491	D77	94927	D94	93208
D88	91556	D89	90932	D166	90126
D156	89169	D82	88952	D76	88827
D135	88109	D111	87977	D69	87474
D61	86120	D120	84873	D117	83366
D78	82890	D49	80023	D25	79986

Sample Name	Sequence Count	Sample Name	Sequence Count	Sample Name	Sequence Count
D56	78977	D90	78579	D33	78076
D70	77784	D144	77639	D68	77581
D32	76936	D55	76442	D164	76134
D63	75473	D65	72915	D118	72915
D119	71518	D26	69345	D51	68824
D52	68023	D57	67697	D165	66134
D31	64651	D173	64361	D161	61361
D167	60745	D41	59538	D66	59345
D53	58941	D11	58352	D157	58186
D159	57911	D17	57278	D160	57148
D28	55093	D27	53989	D178	53692
D95	52639	D123	52414	D138	50818
D34	50636	D83	49830	D12	47109
D71	44733	D154	42313	D148	41027
D179	40979	D8	40039	D184	39167
D149	38480	D147	38212	D58	37731
D1	37698	D128	36867	D9	36836
D133	35319	D114	33945	D190	33878
D155	33019	D183	32750	D152	32689
D137	32656	D42	31246	D153	31216
D7	30525	D191	26703	D4	25735
D136	25261	D10	24842	D131	19981
D162	17791	D127	17510	D125	14232
D5	13187	D150	10288		

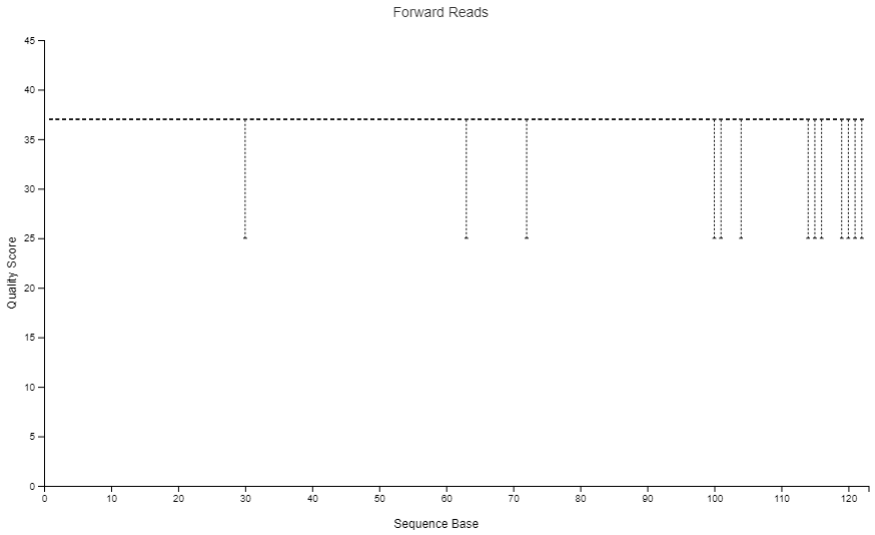


Figure 10: Qiime2 view of the forward read quality.

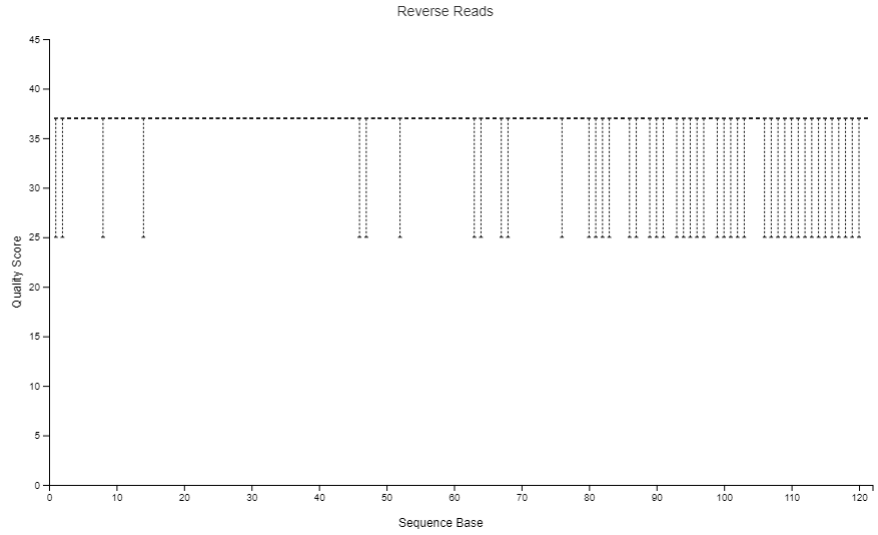


Figure 11: Qiime2 view of the reverse read quality

Step 6: Trim the reads using DADA2

```
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ qiime dada2 denoise-paired --i-
demultiplexed-seqs demux.qza --p-trim-left-f 0
--p-trim-left-r 0 --p-trunc-len-f 110 --p-trunc-len-r 100 --p-n-
threads 0 --o-table table.qza --o-representative-sequences rep-
seqs.qza
```

```
--o-denoising-stats denoising-stats.qza --verbose
```

Step 7: Generating Phylogenetic tree for the ASV

```
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ unset MAFFT_BINARIES
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ qiime phylogeny align-to-tree-mafft-
fasttree --i-sequences rep-seqs.qza
--o-alignment aligned-rep-seqs.qza --o-masked-alignment masked-
aligned-rep-seqs.qza
--p-n-threads 0 --o-tree unrooted-tree.qza --o-rooted-tree rooted-
tree.qza
```

Step 8: Generating taxonomy using the **Naive Classifier approach**

```
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ qiime feature-classifier classtify-sklearn
--i-classifier/ --i-reads reps-seqs.qza --o-classicaation taxon-
omy_naive.qza
```

Step 9: Export all the files that Qiime2 has generated for the **Naive classifier approach**

```
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ qiime tools export --input-path table.qza
--output-path output
Exported table.qza as BIOMV210DirFmt to directory output
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ qiime tools export --input-path rep-
seqs.qza --output-path output
Exported rep-seqs.qza as DNASequencesDirectoryFormat to directory
output
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ qiime tools export --input-path rooted-
```

```
tree.qza --output-path output
Exported rooted-tree.qza as NewickDirectoryFormat to directory output
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ qiime tools export --input-path taxon-
omy.qza --output-path output
Exported taxonomy.qza as TSVTaxonomyDirectoryFormat to directory out-
put
```

Step 10: Making the BIOM file compatible with the R phyloseq package for statistical analysis. To do so need to attach the abundance table of ASVs with their corresponding taxonomy. For the **Naive classifier approach**

```
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2 /output]$ biom convert -i feature-table.biom
-o feature-table.tsv --to-tsv
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2 /output]$ sed -i s/Taxon/taxonomy/ taxon-
omy.tsv | sed -i s/Feature ID/FeatureID/ taxonomy.tsv
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2
/output]$ biom add-metadata -i feature-table.tsv -o feature_w_tax.biom
--observation-metadata-fp taxonomy.tsv --observation-header Fea-
tureID,taxonomy,Confidence --sc-separated taxonomy --float-fields
Confidence
```

Step 11: Now, moving onto **The BLCA approach**. The first step was to export the sequences from Qiime2 to the output folder

```
[MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2/output]$ export PATH=/home/opt/miniconda2/bin:$PATH
[MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2/output]$ source activate qiime2-2019.7
```

Step 12: Convert traditional Qiime2 taxonomy file format to **BLCA format**:

```
(qiime2-2019.7) [MScBioinf@becker /shared5/Rachel/trnL]$ awk
'gsub("\tk__|;p__|;c__|;o__|;f__|;g__|;s__","\t",$0);gsub
(";$","",$0)1' trnL_removed_duplicates.tax | awk -F"\t" 'print
$1"\tspecies:"$8";genus:"$7";family:"$6";order:"$5";class:"$4";phylum:"
$3";superkingdom:"$2' > trnL_Reference_Taxonomy_BLCA.txt
```

Step 13: Enable **BLCA** on Orion cluster:

```
(qiime2-2019.7) [MScBioinf@becker /shared5/Rachel/trnL]$ source
/home/opt/BLCA/enable_BLCA.sh
```

Step 14: Create reference database to use it with **BLCA** software. It complains about duplicate sequences and so thet are removed.

```
(python3) [MScBioinf@becker /shared5/Rachel/trnL]$ bioawk -cfastx
'!seen[$1]++print ">"$1"\n"$2' trnL.fa > trnL_removed_duplicates.fa
(python3) [MScBioinf@becker /shared5/Rachel/trnL]$ makeblastdb
-in trnL_removed_duplicates.fa -dbtype nucl -parse_seqids -out
trnL_removed_duplicates.fa
```

Step 15: Use **BLCA** software on the sequences against the reference database

```
(python3) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2/output]$ 2.blca_main.py -i dna-
sequences.fasta -q
[MScBioinf@becker /shared5/Rachel/trnL/trnL_removed_duplicates.fa -r
[MScBioinf@becker /shared5/Rachel/trnL/trnL_Reference_Taxonomy_BLCA.txt
clustalo is located in your PATH!
> > Fasta file read in!
> > Reading in taxonomy information!  ....
blastn is located in your PATH!
> > Running blast!!
> > Blastn Finished!!
```

```
> > read in blast file...
> > blastn file opened
> > blast output read in
> > Start aligning reads...
> > Taxonomy file generated!!
Time elapsed:  16 minutes
```

Step 16: Then reformatted the results, so that we can attach these to the abundance table in biom format. For this purpose, I generated a taxonomy file that biom software accepts. I have written one line that gets rid of intermediate confidences at higher up the taxonomic level, and retains the latest one

```
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2 /output]$ awk -F"\t" 'BEGINprint
"FeatureID\tTaxon\tConfidence"gsub("superkingdom:
","D_0__",$2);gsub("phylum:","D_1__",$2);gsub("class:","D_2__",$2);
gsub("order:","D_3__",$2);gsub("family:","D_4__",$2);gsub("genus:",
"D_5__",$2);gsub("species:","D_6__",$2);gsub(";$|\t$","",$2)
;gsub(";"," ;",$2);gsub(" ;[0-9]+(\.[0-9]+)?
;",";",$2);gsub("Unclassified","Unassigned\t100.0",$2);gsub("
+;","\t",$2);print $1"\t"$2' dna-sequences.fasta.blca.out > taxon-
omy.tsv
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2 /output]$ cd ..
```

Step 17: Activate Qiime2, and export abundance table as a TSV format

```
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult
/qiime.analysis.attempt2]$ qiime tools export --input-path table.qza
--output-path output
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2]$ cd output (qiime2-2019.7) [MScBioinf@becker
/shared5/Rachel/KogganiResult/
```

```
qiime.analysis.attempt2
/output]$ biom convert -i feature-table.biom -o feature-table.tsv --
to-tsv
```

Step 18: Connect feature-table.tsv with taxonomy.tsv to generate feature_w_tax.biom file

```
(qiime2-2019.7) [MScBioinf@becker  /shared5/Rachel/KogganiResult/
qiime.analysis.attempt2
/output]$ biom add-metadata -i feature-table.tsv -o feature_w_
tax.biom --observation-metadata-fp taxonomy.tsv --observation-header
FeatureID,taxonomy,Confidence --sc-separated taxonomy --float-fields
Confidence
```

Appendix III: Alternative Beta Diversity Measures

As well as the bray-curtis distance measure reported in the main text, I also considered the weighted unifrac distances. Overall, all three measures of beta diversity considered supported that the Puku ND and Puku ME were similar in terms of abundance, phylogeny and when both were considered at once, whereas Cattle ND were not as similar.

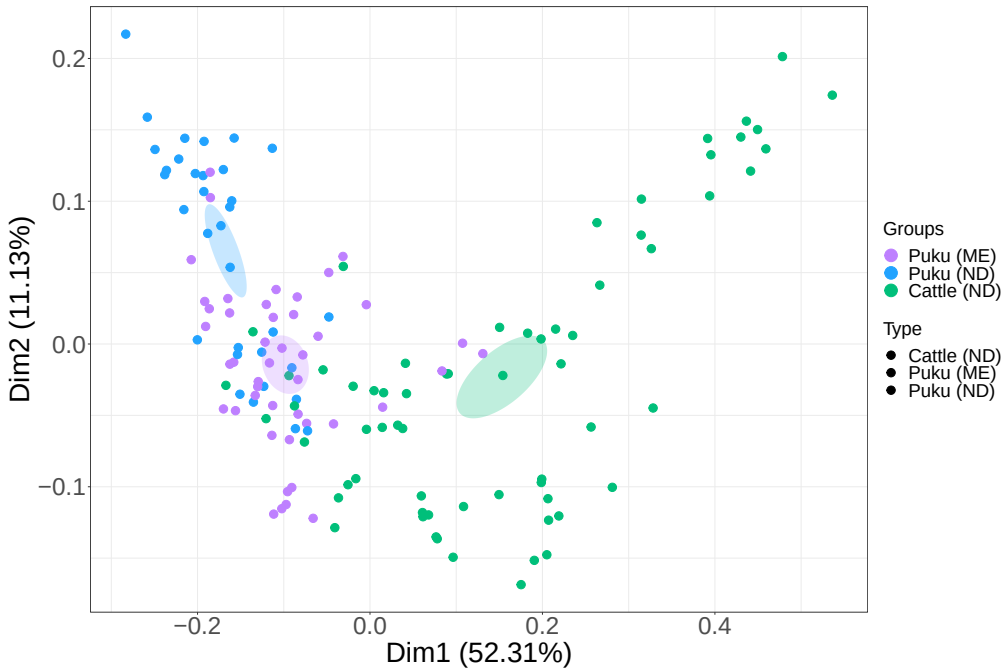


Figure 12: Weighted Unifrac: The clustering here is based on both phylogeny and abundance. This time, all three population clustering are quite distinct with some overlap. This suggests that between the Puku although the taxa present are similar they are different in abundance.The PERMOVA results for this analysis found both Population and Month to be significant in explaining variation ($p < 0.001$), with population explaining 31.9% of variation and month explaining 21.5% of variation.

Appendix IV: Top-25 Abundant Taxa

As well, as the top-25 most abundant taxa at genera level, shown in the main text, the top-25 most abundant ASVs were also investigated. Please note that in the figures the colours do not correspond between the ASV and Genus plots, so please take each taxa independently. Since the BLCA output either returned ASVs as unassigned, genus level or ASV level I could not investigate the top-25 most abundant at any other level. Also of note, at the ASV level there are a lot more informative results present in the abundance plot that an the genus level, which is why the genus level is used within the main text.

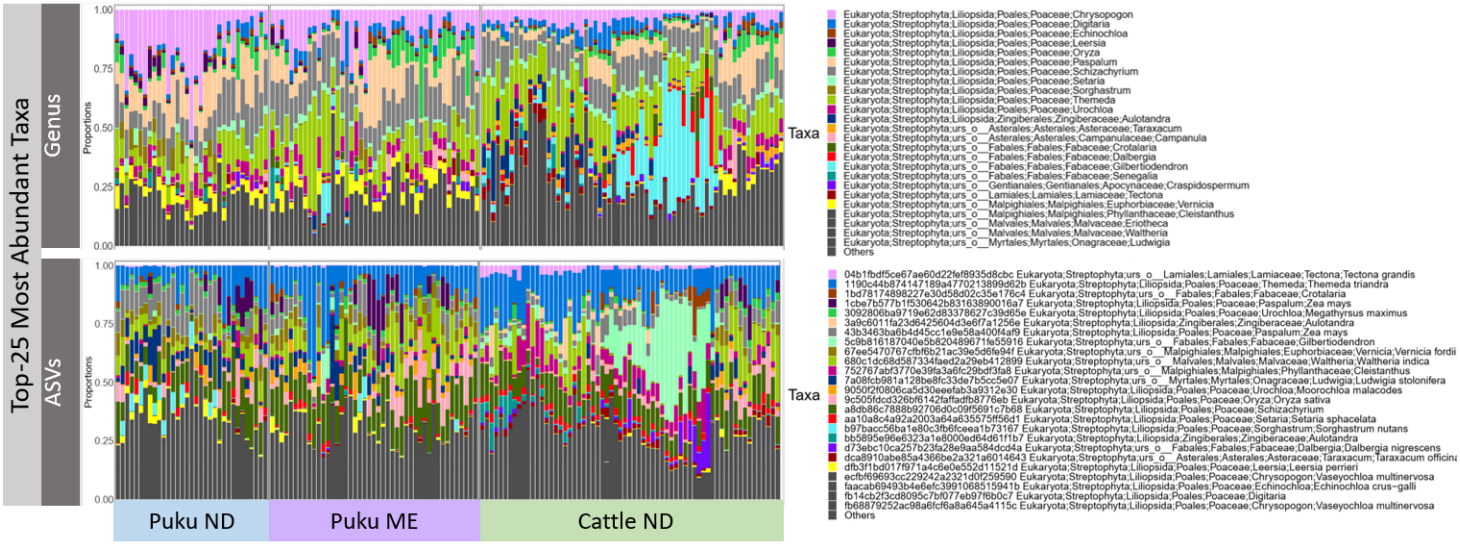


Figure 13: The top-25 most abundant taxa at both Genera and ASV level for the BCLA approach

Appendix VI: Abundance of plants identified in each site via observation of transects.

There were a total of 34 samples recorded in the field, most to species level but some only to genus level. To investigate and get an idea for what is present in the field abundance plots were generated for the proportions of plants. Both sites look

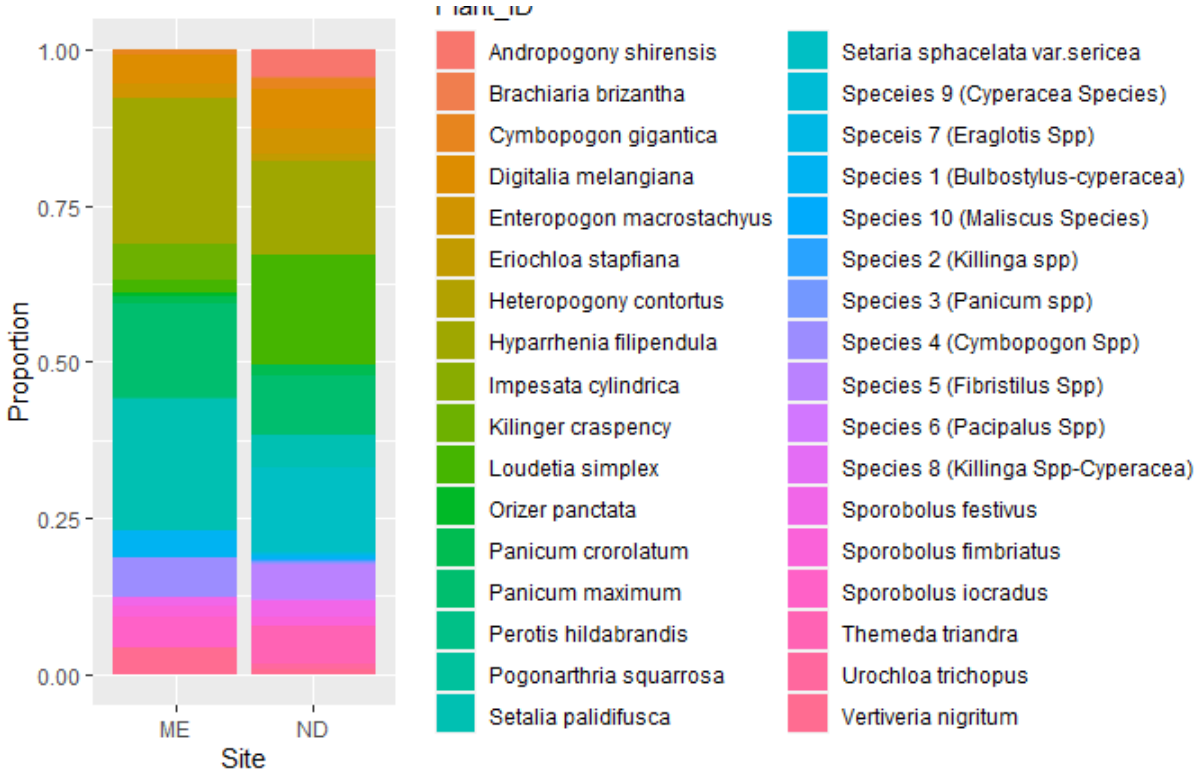


Figure 14: Abundance plots of species present in the ME and ND site, based on proportions