MOLECULAR ECOLOGY
RESOURCES

Minimizing polymerase biases in metabarcoding

Journal:	*Molecular Ecology Resources*
Manuscript ID	MER-17-0405.R1
Manuscript Type:	Resource Article
Date Submitted by the Author:	12-Feb-2018
Complete List of Authors:	Nichols, Ruth; University of California Santa Cruz, Ecology & Evolutionary Biology Vollmers, Christopher; University of California Santa Cruz, Biomolecular Engineering Newsom, Lee; Flagler College, Social Sciences Wang, Yue; University of Wisconsin-Madison, Geography Heintzman, Peter; University of California, Santa Cruz, Ecology and Evolutionary Biology; UiT - The Arctic University of Norway Leighton, McKenna; University of California Santa Cruz, Biomolecular Engineering Green, Richard; University of California Santa Cruz, Biomolecular Engineering Shapiro, Beth; University of California Santa Cruz, Ecology & Evolutionary Biology
Keywords:	eDNA, Environmental DNA, soil, trnL P6 loop, metabarcoding, bias

SCHOLARONE™
Manuscripts

Minimizing polymerase biases in metabarcoding

Ruth V. Nichols[1], Christopher Vollmers[2], Lee A. Newsom[3], Yue Wang[4], Peter D. Heintzman[1,5], McKenna Leighton[1], Richard E. Green[2], Beth Shapiro[1]

Author Affiliations:
[1]Department of Ecology and Evolutionary Biology, University of California Santa Cruz
[2]Department of Biomolecular Engineering, University of California Santa Cruz
[3]Department of Social Sciences, Flagler College, St. Augustine, Florida
[4]Department of Geography, University of Wisconsin-Madison
[5]Tromsø University Museum, UiT - The Arctic University of Norway, 9037 Tromsø, Norway

Corresponding Author: Ruth V. Nichols, email: ruthvnichols@gmail.com

Abstract

DNA metabarcoding is an increasingly popular method to characterize and quantify biodiversity in environmental samples. Metabarcoding approaches simultaneously amplify a short, variable genomic region, or "barcode", from a broad taxonomic group via the polymerase chain reaction (PCR), using universal primers that anneal to flanking conserved regions. Results of these experiments are reported as *occurrence* data, which provide a list of taxa amplified from the sample, or *relative abundance* data, which measure the relative contribution of each taxon to the overall composition of amplified product. The accuracy of both occurrence and relative abundance estimates can be affected by a variety of biological and technical biases. For example, taxa with larger biomass may be better represented in environmental samples than those with smaller biomass. Here, we explore how polymerase choice, a potential source of technical bias, might influence results in metabarcoding experiments. We compared potential biases of six commercially available polymerases using a combination of mixtures of amplifiable synthetic sequences and real sedimentary DNA extracts. We find that polymerase choice can affect both occurrence and relative abundance estimates, and that the main source of this bias appears to be polymerase preference for sequences with specific GC contents. We further

For Review Only

recommend an experimental approach for metabarcoding based on results of our synthetic experiments.

Keywords

Environmental DNA; eDNA; soil; trnL P6 loop; metabarcoding; bias

Introduction

Metabarcoding, which is erroneously described as barcoding or metagenomics in some literature, is the technique in which a universal primer pair is used to amplify multiple templates from a mixture of many different taxa or haplotypes. Metabarcoding is often used in conjunction with environmental DNA (eDNA), or DNA that is collected from environmental sources such as water, sediment, air, and feces (Deiner *et al.* 2017). Metabarcoding is an increasingly popular tool in ecological and paleoecological research, mainly due to its simplicity and low cost. eDNA can be used, for example, to characterize biodiversity of a particular taxonomic group (Ushio *et al.* 2017) or to estimate the ranges of rare, extinct, or cryptic species (Haile *et al.* 2009; Jerde *et al.* 2011; Pedersen *et al.* 2016; Rees *et al.* 2017). Additionally, metabarcoding has been used to calculate differences in haplotype or allele frequency between populations of the same species (Sigsgaard *et al.* 2016), and to link changes in community composition over time to climatic shifts (Willerslev *et al.* 2003, 2007, 2014; Haile *et al.* 2007). These latter examples analyze both the *occurrence* and *relative abundance* of each unique sequence in the amplification product, where abundance is estimated as the proportion of the total number of sequences generated matching each taxon or haplotype.

While metabarcoding is a promising approach to characterize biodiversity both quickly and inexpensively, few studies have validated the method experimentally by, for example, testing the extent to which the true community or population is reconstructed. It is generally

For Review Only

accepted that taxon occurrence can be inferred via metabarcoding, provided that a sufficient number of PCR replicates–amplifying DNA multiple times from the same soil extract using the same amplification conditions–are performed (Piñol *et al.* 2015; Shaw *et al.* 2016) and false positives have been accounted for (Lahoz-Monfort *et al.* 2016). The first eDNA metabarcoding studies used replication (Cooper & Poinar 2000), where DNA extraction and amplification were both replicated, to help confirm their results (Willerslev *et al.* 2003), but many subsequent studies did not replicate experiments (Valentini *et al.* 2009; Soininen *et al.* 2009; Sønstebø *et al.* 2010). After a detailed exploration of the utility of replication in metabarcoding (Darling & Mahon 2011), the use of replication increased, but the number of replicates performed per experiment varied widely. Most studies used between two and five PCR replicates per sample (Andersen *et al.* 2012; Jørgensen *et al.* 2012; Willerslev *et al.* 2014; De Barba *et al.* 2014) and some as many as eight (Giguet-Covex *et al.* 2014). Recently, the use of site occupancy models has been proposed as a tool to estimate how many replicates are needed; with most recommendations ranging from six to 12 replicates per sample (Schmidt *et al.* 2013; Ficetola *et al.* 2015; Lahoz-Monfort *et al.* 2016), depending on the number and abundance of rare taxa. Another approach to estimate the amount of replication required is rarefaction, whereby the number of new taxa identified per replicate PCR is used to estimate the probability that most rare taxa have been recovered (Sanders 1968; Hsieh *et al.* 2016).

Whether relative abundance can be estimated accurately from metabarcoding data is a more contentious issue. Some researchers routinely interpret the relative abundance of sequences post-PCR as indicative of real relative biomass estimates (Kowalczyk *et al.* 2011; Willerslev *et al.* 2014; Niemeyer *et al.* 2017). Others argue against this approach, citing challenges that include differential DNA degradation, different primer binding efficiencies, and sequencing errors as confounding factors that might influence the utility of relative abundance

For Review Only

data collected from metabarcoding loci (Deagle *et al.* 2007, 2013; Pawluczyk *et al.* 2015; Piñol *et al.* 2015; Marcelino & Verbruggen 2016).

Biases that might influence the likelihood of a taxon being detected during metabarcoding can be both biological and technical in origin. Biological differences include organism size, seasonal presence and senescence, preservation, and dispersal strategy, amongst others. Larger taxa, taxa that are present year-round, or taxa whose DNA is readily transported across long distances by wind or water, may be more likely to be observed in environmental samples than smaller, seasonal, and sedentary taxa (Andersen *et al.* 2012; Barnes & Turner 2016; Buxton *et al.* 2017; Rees *et al.* 2017; Hemery *et al.* 2017; Dunn *et al.* 2017). Even when the same number of cells are present in an environmental sample, the starting copy number of target loci may vary between taxa and tissue-type. Chloroplast DNA, for example, is a common target for metabarcoding, but can differ in copy number between taxa, individuals, and cell tissue-types within the same plant (Morley & Nielsen 2016). Taphonomic factors may also influence DNA preservation, for example by affecting the rate of degradation. Lignified structures in plants may slow the rate of DNA degradation (Yoccoz *et al.* 2012), as may anoxic environments (Corinaldesi *et al.* 2011). In some environments, soil leaching and post-depositional mixing may move DNA up or down sediment columns or horizontally over space (Anderson-Carpenter *et al.* 2011; Andersen *et al.* 2012; Rawlence *et al.* 2014; Pedersen *et al.* 2015).

Technical biases can be introduced during DNA extraction and PCR amplification. DNA extraction protocols can be more or less optimized for soil chemistry, which can influence the extent to which DNA is recovered (Zielińska *et al.* 2017). Soils rich in clays or humic acids may bind DNA, for example, reducing DNA recovery (Direito *et al.* 2012). PCR is a highly stochastic process, that is further complicated by the presence of variable templates, with many opportunities for the introduction of bias (Suzuki & Giovannoni 1996; Polz & Cavanaugh 1998;

For Review Only

Aird *et al.* 2011; Pinto & Raskin 2012). Although the universal primers used in metabarcoding are designed to anneal to conserved genomic regions, slight variation in binding site sequences may affect primer binding efficiency, resulting in bias (Elbrecht & Leese 2015; Pinol *et al.* 2014). For example, Fahner et al (2016) used four plant-specific primers to infer community composition from the same soil samples, and found that each primer pair produced a different result. This result may also be related to amplicon length whereby shorter amplicons amplify more readily than longer amplicons. Template secondary structures can also bias PCR when molecules with secondary structures bind to themselves and inhibit their own amplification. In addition, templates with suboptimal GC contents can be disfavored during amplification, although some polymerases are known to have reduced GC-bias and additives such as dimethyl sulfoxide (DMSO) for GC-rich templates or betaine for AT-rich templates can reduce this bias (Baskaran *et al.* 1996; Kozarewa *et al.* 2009; van Dijk *et al.* 2014). Finally, the number of PCR cycles has also been shown to influence results: while a higher number of PCR cycles might increase the likelihood that rare molecules are observed, it could also skew abundance estimates by amplifying the biases described above (Casbon *et al.* 2011; Weyrich *et al.* 2017), but this can vary (Krehenwinkel *et al.* 2017; Vierna *et al.* 2017).

Here, we explore the potential of polymerase choice to influence the results of metabarcoding analyses, with particular reference to polymerase GC-bias. We selected the *trnL* g/h primer set (Taberlet *et al.* 2007) as our universal barcoding primers for this evaluation, as the target *trnL* (P6 loop) locus of the chloroplast genome is commonly used for plant metabarcoding studies (Valentini *et al.* 2009; Sønstebø *et al.* 2010; Pornon *et al.* 2016). Additionally, amplicons derived from this primer set are within the range of 50 and 150 base-pairs (bp) which is suitable for degraded environmental DNA and also sequenceable using short-read sequencing technologies. We performed metabarcoding on DNA extracted from soil collected from St. Paul Island, Alaska, and on mixtures of synthetic oligonucleotides whose inserts varied by GC

For Review Only

content, using six polymerases, including those commonly used in metabarcoding. Using these experiments, we asked three questions: (1) Does polymerase GC preference affect relative abundance estimates in metabarcoding data? (2) Are some polymerases more appropriate for metabarcoding-derived estimates of relative abundance than others? And (3) Does GC bias affect occurrence estimates in metabarcoding experiments?

Materials and Methods

1. Experimental Design Overview

We designed our experiment to ask three questions. First, *Does polymerase GC preference affect relative abundance estimates in metabarcoding data?* To answer this, we performed metabarcoding analyses of sedimentary DNA samples collected from St. Paul Island, Alaska. We performed two separate tests. First, we performed *trnL* (P6 loop) metabarcoding from nine samples, and compared DNA-derived biodiversity estimates and biodiversity estimates based on above-ground survey data from the same sites. Next, for four of these nine sedimentary DNA samples, we explored whether relative abundance changed during the course of PCR amplification, following the design depicted in Figure 1. In both of these tests, we found that polymerase GC preference did affect relative abundance estimated. Our second question was therefore *Are some polymerases more appropriate for metabarcoding-derived estimates of relative abundance than others?* To answer this question, we amplified pools of synthetic oligonucleotides with a range of GC contents using six different polymerases, and measured the precision with which each polymerase reconstructed the starting concentrations of each oligonucleotide pool. Our third question was, *Does GC bias also affect occurrence estimates in metabarcoding experiments?* To answer this question, we again used the sedimentary DNA samples from St. Paul Island, Alaska, but this time performed metabarcoding using the

For Review Only

polymerase identified in Question 2 as the least biased. We estimated the reproducibility of occurrence data using rarefaction analysis of ten replicate PCRs per sample.

2. Data Generation

Environmental DNA from St Paul Island, Alaska

We collected soil samples from St. Paul Island, Alaska. This small (~114 km^2), isolated island is situated ~450 km west of the coast of Alaska in the Bering Sea (~50.2°N, 170.2°W). St. Paul is the largest and most northerly island of the Pribilof Islands (Mungoven 2005), and has a low diversity of plants and terrestrial mammals (Preble & McAtee 1923; Colinvaux 1981), and completely lacks trees. We selected nine sampling sites that were spatially separate from each other, geologically distinct, and appeared to be colonized by different vegetative communities. At each site, a 1×1 m quadrat was chosen. We removed a ~15×15×10 cm (L×W×D) volume of surface soil from the center of each quadrat using a knife and trowel that we cleaned with ethanol between uses. We transferred ~10-20 g of soil to a sterile 50 mL falcon tube for eDNA analyses.

In addition to collecting sediment, we performed surveys of above-ground vegetation. We photographed the surface vegetation in each quadrat and performed a census of each taxon growing within the unit. We counted stems from each representative of each plant taxon and tallied the total for each unit (no counts exceeded 50). For very widespread and ubiquitous taxa, including spreading mat-forming types (e.g. mosses growing at the ground surface) and oversized plants with wide crowns, we estimated relative abundance based on percent coverage within the unit. We identified the majority of common taxa in the field by comparison with a local collection curated at the St. Paul Public School, and verified taxonomic assignments using Hultén's floras (Hultén 1960, 1968). We collected representative samples of distinct or unknown taxa for later taxonomic verification, which we carried out using the relevant published floras

For Review Only

179 along with online keys and floristics data (Hultén 1960; Talbot & Talbot 1994; Mungoven 2005;
180 Stotler & Crandall-Stotler 2005; Walker *et al.* 2005). We converted the count data and the
181 proportion of ground covered as a rank order (1 = 1-20% cover or <10 count; 2 = 21-40% or 10-
182 24 count; 3 = 41-60% or 25-50 count; 4 = 61-80%; 5 = 81-100%) as a proxy for plant abundance
183 at each sampling location.

184 We extracted environmental DNA from all nine soil samples using the MoBio PowerSoil
185 DNA Isolation kit (now called Qiagen DNeasy PowerSoil kit), following the manufacturer's
186 instructions. To avoid contamination, we performed all steps in a clean laboratory that is
187 physically isolated from other molecular biology research, while wearing sterile suits, face-
188 masks, and gloves for DNA extractions and PCR set-up. To monitor cross-contamination, we
189 extracted and processed the samples alongside two negative extraction controls, but did not use
190 a positive control.

191

192 Synthetic oligonucleotide pools

193 We designed and synthesized 12 oligonucleotides with inserts of 47 base-pairs (bp)
194 flanked by the *trnL* g/h primer binding sites with no mismatches (total length: 83 bp;
195 Supplementary Table 1). This set included two oligonucleotides with 13% average GC content,
196 two with 26% average GC content, two with 51% average GC content, two with 63% average
197 GC content, and four oligonucleotides with 38% average GC content. We then created six
198 mixtures of these 12 oligonucleotides in which each oligonucleotide was included at different, but
199 known, concentrations. We then diluted each mixture to 10 fM, which qPCR indicated was
200 similar to the concentrations in our eDNA extracts. To verify pooling accuracy, we amplified each
201 mixture using an approach that adds unique molecular identifiers (MIDs) to each starting
202 molecule (Cole *et al.* 2016; Hoshino *et al.* 2017). Briefly, we first performed two cycles of PCR
203 using modified versions of the *trnL* g/h primers that contained a 5' molecular identifier (which

For Review Only

comprised five random nucleotides, followed by AT, followed by another three random nucleotides: NNNNNATNNN) and the Nextera adapter sequence (Supplementary Figure 1). This two-cycle PCR, which is performed using the permissive Phusion polymerase (New England Biosystems), adds to each starting molecule a uniquely identifying barcode that can be used to reconstruct bioinformatically the true starting relative abundance of molecules. After a clean-up step, we then amplified the product of this two-cycle PCR for an additional 30 cycles with standard Nextera indexing primers and the higher fidelity polymerase in Kapa HiFi ReadyMix (Kapa Biosystems). After sequencing, we counted the number of unique MIDs for each amplicon to verify the starting relative abundance of molecules in the pool.

PCR amplification, library preparation, sequencing, and bioinformatics

We performed PCR using the *trnL* g/h primers and six different polymerases (Table 1). We performed gradient PCR as necessary to determine optimal annealing temperatures for each of the different polymerases. For Platinum HiFi Taq, AmpliTaq Gold and Phusion, we used reagent mixes that are described in previous publications (De Barba *et al.* 2014; Cole *et al.* 2016; Graham *et al.* 2016). All final recipes and cycling conditions are provided in the supplement (Supplementary Table 2). We confirmed that amplification products were in the expected size range (50-150 bp) via gel electrophoresis, which also confirmed that all extraction and PCR negative controls lacked visible amplification products. We purified amplification products using a SPRI bead protocol (Rohland & Reich 2012).

We transformed PCR amplicons into sequenceable libraries using two different approaches. Initially (for questions one and two), we used a lengthy protocol described by Meyer and Kircher (2010) (MK) that involves blunt-end repair, phosphorylation, adapter ligation and fill-in, and indexing PCR. To answer question three, we compared the MK protocol to a shorter and less expensive approach that amplifies DNA using *trnL* g/h primers with 5' overhangs containing

For Review Only

the Illumina TruSeq adapter sequences. This made it possible to proceed directly to indexing PCR following the initial metabarcoding PCR, allowing library preparation to be completed in two steps (two PCR set-ups). To assess whether the two-step protocol performed differently from the MK protocol, we performed a comparative experiment in which we amplified DNA and sequenced libraries generated from a common master mix of Qiagen Multiplex Master Mix, water and template (consisting of an equimolar mixture of synthetic oligonucleotides). After sequencing, we found there was no significant difference between the two methods (Standard Least Squares Test: Whole Model F Ratio = 0.55, P = 0.58, Supplementary Figure 2). While we find no difference between these two library preparation approaches, additional comparative analyses of prepared libraries that more finely sample, for example, different GC-content binning strategies, will be necessarily to explore fully whether one library preparation approach is superior by all metrics to another.

For all experiments, we sequenced libraries on the Illumina MiSeq platform using 2×75 v3 chemistry, targeting 150,000 reads per sample. We used rarefaction to confirm that sequencing depth was sufficient to recover all amplified molecules (Hsieh *et al.* 2016).

After sequencing, we processed each data set using an in-house bioinformatics pipeline. Briefly, we removed adapters and merged overlapping reads using SeqPrep v2 (https://github.com/jstjohn/SeqPrep), with the following flags: minimum length of reads (-L) 37 (combined length of the primer sequences plus one), overlap required to merge read1 and read2 (-o) 10, minimum length of adapter to consider trimming (-O) 8, and quality threshold (-q) 15. We filtered the merged reads and retained sequences containing either an exact match to the forward primer and the reverse complement of the reverse primer (correct orientation), or an exact match to the reverse primer and the reverse complement of the forward primer (incorrect orientation). We then reverse-complemented the data in the incorrect orientation using the FASTX toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/) and concatenated these data

For Review Only

with those in the correct orientation. We trimmed any remaining adapter and PCR primer sequences from the ends of the filtered reads, and removed any reads that retained any primer sequences or that were shorter than six base pairs using PRINSEQ-lite v0.20.4 (Schmieder & Edwards 2011). We created a single file with all unmerged reads, so that read1 and read2 were on the same line, and processed this file as described above. We then split this file back into read1 and read2 files. We did not remove sequences that contained a mismatch to known synthetic oligonucleotide insert sequences (see below). For the sequence data derived from the St. Paul soil samples, all amplicons were short enough that the sequences could be merged. We used the obitools software defaults (Boyer et al. 2016) to group identical sequences (obiuniq), remove singletons and PCR artifacts (obiclean –H) and compare the sequences to the arctic, boreal, and embl reference libraries (Sønstebø *et al.* 2010; Willerslev *et al.* 2014) to identify the reads to their best-associated plant taxa. Because we used three reference libraries, three separate result files were created for each sample (one for each reference library). We parsed the three files using a script that compared the results in each file and extracted only the entries with the highest percent identity and lowest taxonomic rank. If two species of the same genus were seen, that sequence was classified to the genus level. We set a cut-off value of 98% identity and removed reads at proportions less than 0.001. The number of raw and merged reads and number of identified taxa per sample are listed in the Supplementary Materials (Supplementary Table 3).

For the synthetic oligonucleotide pools, we used grep to pull out the known sequences and their reverse complements and count how many times they occurred within each fasta file. As the obitools and grep methods both provided count data, we converted these counts to relative abundances.

3. Data Analysis

For Review Only

279 *Question 1: Does polymerase GC preference affect relative abundance estimates in*
280 *metabarcoding data?*
281 For the nine St. Paul samples, we performed ten replicate PCRs per sample using Platinum HiFi
282 Taq polymerase (Invitrogen) following the protocol found in Graham *et al*. (2016). After
283 sequencing and read processing as detailed above, we used standard least squares to test the
284 effects of above-ground vegetation abundance and amplicon average GC content on DNA
285 relative abundance, both separately and interactively.
286 To test the effect of PCR cycle number on the relative abundances of different plant taxa,
287 we chose four St. Paul soil eDNA extracts and two PCR controls, scaled up the PCR to 100 µL,
288 and collected 1 µL aliquots at five-cycle intervals from cycles 10 to 60 (Figure 1). We used a
289 large reaction volume to minimize the impact of aliquot removal and cooled the reaction to 20 C
290 during each collection step to avoid evaporation. Large numbers of cycles are often used in
291 metabarcoding experiments because the target loci are at very low abundances relative to the
292 total amount of extracted DNA and eDNA extracts often have PCR inhibitors (Kennedy *et al.*
293 2013). We used 60 cycles to be sure that all PCRs had reached the plateau phase. Each aliquot
294 was made into an Illumina sequencing library individually using a library preparation protocol
295 based on Meyer & Kircher (2010) (as detailed above). We called this our amplicon competition
296 experiment (Figure 1).
297
298 *Question 2: Are some polymerases more appropriate for metabarcoding-derived estimates of*
299 *relative abundance than others?*
300 We assessed whether six polymerases (Table 1) could individually maintain the starting ratio of
301 oligonucleotides (relative abundance) in mixtures after 35 cycles of PCR. For each polymerase,
302 we performed six experiments in which synthetic oligonucleotides were combined at different
303 ratios based on sequence GC content. The oligonucleotides were combined (1) in equimolar

For Review Only

ratios (two experiments), (2) by increasing proportion with GC content, (3) by decreasing proportion with GC content, (4) with extreme GC contents being most abundant, and (5) with extreme GC contents being least abundant. For each experiment, we performed metabarcoding PCRs in triplicate using the *trnL* g/h primers. After obtaining relative abundance estimates for each oligonucleotide in each pool, we plotted expected abundances (relative abundance prior to amplification) versus observed abundances (relative abundance after amplification) for each polymerase. We then calculated the Pearson correlation coefficient between observed and expected abundance values for each enzyme.

Question 3: Does GC bias affect occurrence estimates in metabarcoding experiments?

We again performed metabarcoding on the nine St. Paul soil eDNA extracts as described for Question 1, but used the Qiagen Multiplex Master Mix (Qiagen), which our results indicated is the least biased of the six polymerases tested (see below). As with the experiment described in Question 1 using Platinum HiFi Taq (Invitrogen), we performed ten replicate PCRs for each sample. We assigned amplicons to taxa as described above. We then performed rarefaction for each replicate set from both polymerases using iNEXT (Hsieh *et al.* 2016) in R v3.4.2 (http://www.R-project.org/).

Results

Question 1: Does polymerase preference for certain GC contents affect relative abundance estimates in metabarcoding data?

For this question we used the above-ground vegetation abundance data, that was collected prior to the DNA work, and the Platinum HiFi Taq-amplified metabarcoding data. Both data sets were

For Review Only

generated from the same nine localities on St. Paul. Using both of these, we plotted all plant taxa that were identified using both above-ground and eDNA at all locations on the same plot but split into GC content bins (Figure 2). The x-values are above-ground ranked abundances and the y-values are mean eDNA abundance across replicates. When we compared relative abundance estimates from the metabarcoding experiments to the relative abundance inferred from above-ground biomass, we found that whether or not these two estimates agreed depended on average GC content of the plant's *trnL* (P6 loop) locus (Standard Least Squares, whole model: $F = 34.25$, $P<0.0001$; effect tests: Average GC, $t = 1.54$, $P = 0.124$, above-ground abundance, $t = 12.27$, $P<0.0001$, Average GC*above-ground abundance, $t = 4.39$, $P<0.0001$). Figure 2 shows above-ground and eDNA-based estimates of abundance are correlated most strongly in middle GC content bins, but this relationship decreases or disappears completely at the more extreme GC contents. This pattern is consistent with the previously reported optimal GC content of 34-38% for Platinum HiFi Taq polymerase (Dabney & Meyer 2012).

While this pattern observed in Figure 2 supports the hypothesis that sequences with certain GC contents are preferentially amplified via PCR, it does not exclude the possibility that biological factors, such as differences in above- versus below-ground biomass, are influencing the results. We therefore performed an additional experiment in which we measured changes in DNA-based relative abundance estimates directly during the course of PCR for four St. Paul eDNA extracts (Figure 1). Figure 3A shows the changes in relative abundance of the twelve most abundant taxa in each of the four samples during cycles 20 through 60 of the PCR. Libraries from cycles 10 and 15 had no sequenceable molecules. Exponential amplification appears to start at cycle 30 for all samples and this is confirmed in the qPCR plots (Supplementary Figure 3). We calculated the fold change from cycle 30 to 60, and used this to quantify the increase or decrease in the relative abundance of each amplicon. We then recorded

For Review Only

the number of primer mismatches and barcode length for each amplicon. We found that neither primer mismatches nor amplicon length explained the increase or decrease in relative abundance (primer mismatches, R^2: 0.011; sequence length, R^2: 0.095). However, we found a positive correlation with average GC content and fold change from cycle 30 to 60 (R^2: 0.474, Linear fit P=0.002); Figure 3B).

Question 2: Are some polymerases more appropriate for metabarcoding-derived estimates of relative abundance than others?

Results from Question 1 suggest that Platinum HiFi Taq polymerase preferentially amplifies sequences with 34-38% GC. To identify polymerases that might be more appropriate for metabarcoding than Platinum HiFi Taq, we performed metabarcoding on mixtures of synthetic oligonucleotides with different GC contents using six commonly used polymerases (Table 1). We found that the correlation between observed and expected oligonucleotide proportions differed between enzymes (Figure 4). Among the polymerases tested, the Qiagen Multiplex Master Mix polymerase most accurately reconstructed the known starting relative abundances (Figure 4A), and varied the least in accuracy by GC content (Figure 4B). However, the Qiagen Multiplex Master Mix polymerase also had the highest proportion of sequences with at least one error (Figure 4C). Figure 5 shows the differences between observed and expected relative abundance using the most quantitatively accurate (Qiagen Multiplex Master Mix polymerase) and least quantitatively accurate (Phusion polymerase) enzymes. Detailed plots for the other four enzymes are provided in the Supplementary Materials (Supplementary Figures 4-7).

Question 3: Does GC bias affect occurrence data?

For Review Only

The results above show that polymerase biases can influence eDNA-based estimates of relative abundance. To test whether polymerase bias may also influence the accuracy of occurrence estimates, we performed an additional experiment in which we PCR-amplified the *trnL* (P6 loop) locus from the same nine St. Paul eDNA extracts that were amplified for Question 1, however this time using the best-performing enzyme as identified by the synthetic oligonucleotide experiment above, Qiagen Multiplex Master Mix. As with Platinum HiFi Taq polymerase, we performed 10 replicate PCRs for each of the nine eDNA samples, and used rarefaction to confirm that sequencing depth of each PCR library was sufficient to recover all amplified molecules (Hsieh *et al.* 2016). We then performed additional rarefaction analyses, this time asking whether additional PCR replicates were contributing significantly toward biodiversity estimates, i.e. were sampling taxa that had not yet been sampled. We found that after 10 replicates, mean sample coverage (the probability that all rare taxa have been recovered) was not significantly different when using the Qiagen Multiplex Master Mix compared to Platinum HiFi Taq ($t = -0.66$, $df=15.76$, $p=0.52$; Figure 6). In addition, despite the fact that St. Paul has low plant diversity (Preble & McAtee 1923; Colinvaux 1981), only one site appears to have reached a rarefaction plateau, which would suggest that the majority of species present have been sequenced, after 10 replicates. However, when we compared this to the data generated using Platinum HiFi Taq, this sample had not yet reached a rarefaction plateau. Given the small sample size, it is not possible to know whether this difference is due to polymerase choice or to chance.

Discussion

Our results show polymerase GC bias can dramatically alter the relative abundance of molecules during PCR. It is important, therefore, to use an experimental approach in

For Review Only

metabarcoding that limits the influence of polymerase GC bias. Molecular Identifier (MID), also called Unique Molecular Identifier (UMI), methods (Cole *et al.* 2016) offer a possible solution, as they allow each starting molecule to be disambiguated bioinformatically after PCR. In this way, GC bias that manifests during PCR can be effectively ignored. However, these methods are not yet optimized for the mixed, low concentration samples that are most often available for metabarcoding. While we successfully tested a UMI approach for the analysis of synthetic mixtures of oligonucleotides, the approach often failed to produce sequencing libraries when analyzing actual eDNA samples. This may be due to inhibitors and/or very low concentrations of target DNA compared to all extracted DNA. Because polymerases vary in the degree to which they are biased toward GC content, another approach is to simply choose the least biased polymerase. Of the six polymerases evaluated here, our data show that the Qiagen Multiplex Master Mix is the least biased and effectively retains abundance ratios throughout the PCR (R^2: 0.95). Qiagen Multiplex Master Mix (but not the enzyme, HotStarTaq, itself) was originally engineered for experiments that targeted multiple templates simultaneously, which may explain why it performs well here (Qiagen 2013).

If a biased polymerase is used in metabarcoding, the DNA results may not reflect the true relative abundance of target taxa. For the plant *trnL* (P6 loop) locus, for example, GC content varies considerably among major plant growth forms (Figure 7). The GC content of forbs, or low-lying herbaceous flowering plants, falls mainly within the range preferred by most polymerases (Dabney & Meyer 2012). Our DNA-based relative abundance estimates of plants from St. Paul (Figure 8) and those previously published from Siberia and Alaska (Supplementary Figure 8) (Willerslev et al. 2014) were both generated using Platinum HiFi Taq polymerase targeting the *trnL* P6-loop locus, and showed that graminoids (grasses and sedges) were less abundant than forbs. Because this pattern falls within the biases of Platinum HiFi Taq polymerase, these results may simply reflect polymerase bias rather than true biological signal.

For Review Only

Although our results indicate that GC bias can confound metabarcoding-based relative abundance estimates, other potential sources of bias may also influence amplicon competition during PCR. For example, differences in the number of mismatches between the sequence and the primer at the primer binding site and differences in template length will also affect the efficiency with which an amplicon is copied (Stadhouders *et al.* 2010). While we did not find that the number of primer mismatches affected the efficiency of replication, few taxa have mismatches to the *trnL* g/h primers (Taberlet *et al.* 2007). Primer mismatches have been shown, however, to influence relative abundance for other metabarcoding loci (Piñol *et al.* 2015). Additionally, shorter molecules tend to amplify more readily than longer molecules during PCR (Shagin *et al.* 1999), and, while most sequences amplified by the *trnL* g/h primers in this study tended to be around the same length, other metabarcoding loci vary considerably in barcode length between amplified taxa. Another source of bias during PCR is homopolymer repeats (Kieleczawa 2006). In our amplicon competition experiment using Platinum HiFi Taq, the plant taxa Anthemideae and *Pedicularis* decreased in abundance in all four samples despite having optimal (Anthemideae has a GC content of 36%) and close to optimal (*Pedicularis* is 31%) GC contents, which may be because these barcodes contain 8 and 9 bp-long homopolymer runs respectively. In comparison to Platinum HiFi Taq, we noted that Anthemideae and *Pedicularis* had increased abundances when using Qiagen Multiplex MasterMix (Supplementary Figures 9-12), suggesting that Qiagen Multiplex MasterMix was not deterred by the homopolymer repeats. Finally, polymerase error rates are a potential source of error in metabarcoding experiments, and our results showed that HotStarTaq in the Qiagen Multiplex Master Mix had the highest error rate of the six polymerases used (Figure 4C). Polymerase error has the potential to produce false positive results when barcoding loci differ by one or a few base-pairs, although this may be ameliorated by bioinformatic pipelines capable of identifying potential sequencing errors.

For Review Only

Our results suggest that occurrence data, which has been believed to be largely reliable from metabarcoding experiments, can also be challenging to interpret. While it is understood that rare taxa may be more difficult to identify than common taxa, recommendations within the field have been to perform replicate PCRs, with little guidance as to how many PCRs are necessary. Our experiments from St. Paul suggest, however, that more than 10 replicate PCRs would be necessary to sample the breadth of taxa within our extracts, regardless of polymerase GC bias. In many instances, it may be more practical to combine DNA-based surveys with other data types, such as pollen and identification of macroscopic remains (Birks & Birks 2015). While site occupancy models offer a potential solution to estimate the number of replicates required to identify rare taxa (Schmidt *et al.* 2013; Dorazio & Erickson 2017), these are constructed for single species, and would not be practical for experiments that aim to describe an entire community. We note, however, that the most abundant taxa were recovered in all PCR replicates for all sites and both polymerases, suggesting that DNA metabarcoding is a reasonable approach to identify at least the most abundant taxa in an environment, even if only a single replicate PCR is performed (Leray & Knowlton 2017).

While our current work has identified an experimental approach to reduce the influence of GC content on relative abundance estimates in metabarcoding, it is important also to consider other sources of potential biases and error when interpreting results. For example, errors such as template switching, where sample-specific barcodes are associated to the incorrect sample during either library preparation (Schnell *et al.* 2015) or sequencing (Kircher *et al.* 2012), may influence both occurrence and relative abundance data. Fortunately the latter problem can be mitigated by adding indices to both ends of the molecule (Kircher *et al.* 2012). The choice of bioinformatic pipeline can also influence results. For example, in a recent analysis of the metagenome of fresh basil, three out of four pipelines identified *Salmonella* but, because *Salmonella* was not identified via qPCR, the authors concluded that the bioinformatic results

For Review Only

were erroneous (Ceuppens *et al.* 2017). While public databases containing metabarcoding loci continue to expand in taxonomic depth (Bell *et al.* 2017), some lineages are more poorly represented. Finally, biological differences between species, including variation per cell or tissue type in the number of amplifiable loci (Morley & Nielsen 2016), differences in organism size, seasonal senescence, and behavior, may all influence the probability that an organism will be represented in a particular environmental sample. Although work remains to be done to better understand the consequences of these various types of bias and error, metabarcoding remains a powerful approach to quickly and inexpensively characterize communities.

Conclusion

Despite the rapid growth of metabarcoding as a technique for characterizing communities from eDNA samples, relatively little attention has been given to validating the methodology and understanding its limitations. Polymerase GC bias is a known challenge for applications that rely on PCR (Kozarewa *et al.* 2009; Aird *et al.* 2011; Dabney & Meyer 2012). With the advent of next-generation sequencing approaches, PCR-free methods have been developed to convert extracted DNA into sequenceable molecules (Kozarewa *et al.* 2009). PCR remains the most useful approach to catalogue diversity in environmental samples, however, as the number of target molecules is small relative to the total extracted DNA. For this reason, it is important to understand the influence of GC bias in metabarcoding approaches and, if possible, mitigate these biases. Here, we showed that many commonly used PCR protocols are not appropriate for generating reliable estimates of relative abundance. In these cases, our results show that the relative abundance of amplified sequences changes during PCR cycling, and that these changes are related to the GC content of the target. Of the six polymerases and mixtures tested, Qiagen Multiplex Master Mix provided the most accurate estimates of relative abundance, but also

For Review Only

502 generated the highest error rate. However, we found no evidence that occurrence data was
503 influenced by polymerase bias.

For Review Only

Acknowledgements

We thank Soumaya Belmecheri for fieldwork assistance, Brendan O'Connell and Joshua Kapp for assistance with lab work and for writing scripts, and Duane Froese for discussion. We thank the two reviewers whose comments helped improve this manuscript. This work was funded by awards from the Gordon and Betty Moore Foundation (GBMF-3804), the National Science Foundation (ARC-1203990), and the University of California Office of the President (20160713SC).

Data Accessibility

The processed sequence data and above-ground survey data are available in the Dryad repository (10.5061/dryad.k129c). Raw amplicon sequence data have been made available on the NCBI Short Read Archive (BioProject: PRJNA433185).

Author Contributions

BS, PDH, LAN and YW designed and executed the collection of sediment samples. LAN and YW identified the plants. RN, REG, PDH and BS designed the amplicon competition experiment and the synthetic oligonucleotides. CV and RN designed the experiments testing different polymerases. RN and ML performed the laboratory work. RN and CV analyzed the data. RN, BS, CV and PDH wrote the paper, with critical input from all remaining authors.

References

Aird D, Ross MG, Chen W-S *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**, R18.

Andersen K, Bird KL, Rasmussen M *et al.* (2012) Meta-barcoding of "dirt" DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Anderson-Carpenter LL, McLachlan JS, Jackson ST *et al.* (2011) Ancient DNA from lake sediments: bridging the gap between paleoecology and genetics. *BMC Evolutionary Biology*, **11**, 30.

Barnes MA, Turner CR (2016) The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, **17,**1,1-17.

Baskaran N, Kandpal RP, Bhargava AK *et al.* (1996) Uniform amplification of a mixture of deoxyribonucleic acids with varying GC content. *Genome Research*, **6**, 633–638.

Bell KL, Loeffler VM, Brosi BJ (2017) An rbcL reference library to aid in the identification of plant species mixtures by DNA metabarcoding. *Applications in plant sciences*, **5**.

Birks HJB, Birks HH (2015) How have studies of ancient DNA from sediments contributed to the reconstruction of Quaternary floras? *The New Phytologist*, **209**, 499–506.

Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E (2016) OBITOOLS: a unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, **16**, 176-182.

Buxton AS, Groombridge JJ, Zakaria NB, Griffiths RA (2017) Seasonal variation in environmental DNA in relation to population size and environmental factors. *Scientific reports*, **7**, 46294.

Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR

For Review Only

template molecules with application to next-generation sequencing. *Nucleic Acids Research*, **39**, e81.
Ceuppens S, De Coninck D, Bottledoorn N, Van Nieuwerburgh F, Uyttendaele M (2017) Microbial community profiling of fresh basil and pitfalls in taxonomic assignment of enterobacterial pathogenic species based upon 16S rRNA amplicon sequencing. *International Journal of Food Microbiology*, **257**, 148–156.
Cole C, Volden R, Dharmadhikari S, Scelfo-Dalbey C, Vollmers C (2016) Highly Accurate Sequencing of Full-Length Immune Repertoire Amplicons Using Tn5-Enabled and Molecular Identifier-Guided Amplicon Assembly. *Journal of Immunology*, **196**, 2902–2907.
Colinvaux P (1981) Historical ecology in beringia: the south land bridge coast at st. paul island. *Quaternary Research*, **16**, 18–36.
Cooper A, Poinar HN (2000) Ancient DNA: do it right or not at all. *Science*, **289**, 1139.
Corinaldesi C, Barucca M, Luna GM, Dell'Anno A (2011) Preservation, origin and genetic imprint of extracellular DNA in permanently anoxic deep-sea sediments. *Molecular Ecology*, **20**, 642–654.
Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*, **52**, 87–94.
Darling JA, Mahon AR (2011) From molecules to management: adopting DNA-based methods for monitoring biological invasions in aquatic environments. *Environmental Research*, **111**, 978–988.
De Barba M, Miquel C, Boyer F *et al.* (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular ecology resources*, **14**, 306–323.
Deagle BE, Gales NJ, Evans K *et al.* (2007) Studying seabird diet through genetic analysis of faeces: a case study on macaroni penguins (Eudyptes chrysolophus). *Plos One*, **2**, e831.
Deagle BE, Thomas AC, Shaffer AK, Trites AW, Jarman SN (2013) Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular ecology resources*, **13**, 620–633.
Deiner K, Bik HM, Mächler E *et al.* (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology*.
van Dijk EL, Jaszczyszyn Y, Thermes C (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Experimental Cell Research*, **322**, 12–20.
Direito SOL, Marees A, Röling WFM (2012) Sensitive life detection strategies for low-biomass environments: optimizing extraction of nucleic acids adsorbing to terrestrial and Mars analogue minerals. *FEMS Microbiology Ecology*, **81**, 111–123.
Dorazio RM, Erickson RA (2017) eDNAoccupancy: An R Package for Multi-scale Occupancy Modeling of Environmental DNA Data. *Molecular ecology resources*.
Dunn N, Priestley V, Herraiz A, Arnold R, Savolainen V (2017) Behavior and season affect crayfish detection and density inference using environmental DNA. *Ecology and evolution*, **7**, 7777–7785.
Elbrecht V, Leese F (2015) Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PloS One*, **10**,7,e0130324-16.
Fahner NA, Shokralla S, Baird DJ, Hajibabaei M (2016) Large-Scale Monitoring of Plants through Environmental DNA Metabarcoding of Soil: Recovery, Resolution, and Annotation of Four DNA Markers. *Plos One*, **11**, e0157505.
Ficetola GF, Pansu J, Bonin A *et al.* (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular ecology*

For Review Only

resources, **15**, 543–556.
Giguet-Covex C, Pansu J, Arnaud F *et al.* (2014) Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nature Communications*, **5**, 3211.
Graham RW, Belmecheri S, Choy K *et al.* (2016) Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 9310–9314.
Haile J, Froese DG, Macphee RDE *et al.* (2009) Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 22352–22357.
Haile J, Holdaway R, Oliver K *et al.* (2007) Ancient DNA chronology within sediment deposits: are paleobiological reconstructions possible and is DNA leaching a factor? *Molecular Biology and Evolution*, **24**, 982–989.
Hemery LG, Politano KK, Henkel SK (2017) Assessing differences in macrofaunal assemblages as a factor of sieve mesh size, distance between samples, and time of sampling. *Environmental Monitoring and Assessment*, **189**, 413.
Hsieh TC, Ma KH, Chao A (2016) iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in ecology and evolution / British Ecological Society*.
Hoshino T, Inagaki F (2017) Application of Stochastic Labeling with Random-Sequence Barcodes for Simultaneous Quantification and Sequencing of Environmental 16S rRNA Genes. PLoS ONE 12(1): e0169431.
Hultén E (1960) *Flora of the Aleutian Islands and Westernmost Alaska Peninsula: With Notes on the Flora of Commander Islands*. Hafner Pub. Co., New York.
Hultén E (1968) *Flora of Alaska and Neighboring Territories: a Manual of the Vascular Plants*. Stanford University Press, Stanford.
Jerde CL, Mahon AR, Chadderton WL, Lodge DM (2011) “Sight-unseen” detection of rare aquatic species using environmental DNA. *Conservation letters*, **4**, 150–157.
Jørgensen T, Kjaer KH, Haile J *et al.* (2012) Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA meta-barcoding. *Molecular Ecology*, **21**, 1980–1988.
Kennedy S, Callahan H, Carlson M (2013) *Tips and Tricks for Isolation of DNA & RNA from Challenging Samples*. MO BIO Laboratories, Inc., Carlsbad, CA.
Kieleczawa J (2006) Fundamentals of sequencing of difficult templates--an overview. *Journal of Biomolecular Techniques*, **17**, 207–217.
Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, **40**, e3.
Kowalczyk R, Taberlet P, Kaminski T, Wojcik J (2011) Influence of management practices on large herbivore diet–Case of European bison in Bialowieza Primeval Forest (Poland). *Forest ecology and management*, **261**, 821–828.
Kozarewa I, Ning Z, Quail MA *et al.* (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, **6**, 291–295.
Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie R (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, 1-12.
Lahoz-Monfort JJ, Guillera-Arroita G, Tingley R (2016) Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular ecology resources*, **16**, 673–685.
Leray M, Knowlton N (2017) Random sampling causes the low reproducibility of rare eukaryotic

For Review Only

OTUs in Illumina COI metabarcoding. PeerJ, 5, e3006–27.
Marcelino VR, Verbruggen H (2016) Multi-marker metabarcoding of coral skeletons reveals a rich microbiome and diverse evolutionary origins of endolithic algae. *Scientific reports*, **6**, 31508.
Morley SA, Nielsen BL (2016) Chloroplast DNA Copy Number Changes during Plant Development in Organelle DNA Polymerase Mutants. *Frontiers in plant science*, **7**, 57.
Mungoven M (2005) *Soil Survey of Saint Paul Island Area, Alaska*. National Resources Conservation Service, United States Department of Agriculture.
Niemeyer B, Epp LS, Stoof-Leichsenring KR, Pestryakova LA, Herzschuh U (2017) A comparison of sedimentary DNA and pollen from lake sediments in recording vegetation composition at the Siberian treeline. *Molecular ecology resources*.
Pawluczyk M, Weiss J, Links MG *et al.* (2015) Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples. *Analytical and Bioanalytical Chemistry*, **407**, 1841–1848.
Pedersen MW, Overballe-Petersen S, Ermini L *et al.* (2015) Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **370**, 20130383.
Pedersen MW, Ruter A, Schweger C *et al.* (2016) Postglacial viability and colonization in North America's ice-free corridor. *Nature*, **537**, 45–49.
Piñol J, Mir G, Gomez-Polo P, Agustí N (2015) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular ecology resources*, **15**, 819–830.
Pinto AJ, Raskin L (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *Plos One*, **7**, e43093.
Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, **64**, 3724–3730.
Pornon A, Escaravage N, Burrus M *et al.* (2016) Using metabarcoding to reveal and quantify plant-pollinator interactions. *Scientific reports*, **6**, 27282.
Preble EA, McAtee WL (1923) *A Biological Survey of the Pribilof Islands, Alaska, North American Fauna, No. 46*. Department of Agriculture, Bureau of Biological Survey, Government Printing Office, Washington DC, USA.
Qiagen (2013) Qiagen Multiplex PCR Kit: Product Details.
Rawlence NJ, Lowe DJ, Wood JR *et al.* (2014) Using palaeoenvironmental DNA to reconstruct past environments: progress and prospects. *Journal of Quaternary Science*, **29**, 610–626.
Rees HC, Baker CA, Gardner DS, Maddison BC, Gough KC (2017) The detection of great crested newts year round via environmental DNA analysis. *BMC Research Notes*, **10**, 327.
Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.
Sanders HL (1968) Marine benthic diversity: A comparative study. *The American Naturalist*, **102**, 243–282.
Schmidt BR, Kéry M, Ursenbacher S, Hyman OJ, Collins JP (2013) Site occupancy models in the analysis of environmental DNA presence/absence surveys: a case study of an emerging amphibian pathogen. *Methods in ecology and evolution / British Ecological Society*, **4**, 646–653.
Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**(6), 863-864.
Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated--reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular ecology resources*, **15**, 1289–1303.

For Review Only

Shagin DA, Lukyanov KA, Vagner LL, Matz MV (1999) Regulation of average length of complex PCR product. *Nucleic Acids Research*, **27**, e23.
Shaw JLA, Clarke LJ, Wedderburn SD *et al.* (2016) Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biological Conservation*, **197**, 131–138.
Sigsgaard EE, Nielsen IB, Bach SS *et al.* (2016) Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution*, **1**, 0004.
Soininen EM, Valentini A, Coissac E *et al.* (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in zoology*, **6**, 16.
Stadhouders R, Pas SD, Anber J *et al.* (2010) The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5' nuclease assay. *The Journal of Molecular Diagnostics*, **12**, 109–117.
Stotler RE, Crandall-Stotler BJ (2005) Bryophytes. *Department of Plant Biology, Southern Illinois University*.
Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, **62**, 625–630.
Sønstebø JH, Gielly L, Brysting AK *et al.* (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular ecology resources*, **10**, 1009–1018.
Taberlet P, Coissac E, Pompanon F *et al.* (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, **35**, e14.
Talbot SS, Talbot SL (1994) Numerical classification of the coastal vegetation of Attu Island, Aleutian Islands, Alaska. *Journal of Vegetation Science*.
Ushio M, Fukuda H, Inoue T *et al.* (2017) Environmental DNA enables detection of terrestrial mammals from forest pond water. *Molecular ecology resources*.
Valentini A, Miquel C, Nawaz MA *et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular ecology resources*, **9**, 51–60.
Vierna J, Dona J, Vizcaino A, Serrano D, Jovani R (2017). PCR cycles above routine numbers do not compromise high-throughput DNA barcoding results. *Genome*, **60**(10), 868–873.
Walker DA, Raynolds MK, Daniëls FJA, et al. (2005) The circumpolar Arctic vegetation map. *Vegetation Science*.
Weyrich LS, Duchene S, Soubrier J *et al.* (2017) Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature*, **544**, 357–361.
Willerslev E, Cappellini E, Boomsma W *et al.* (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science*, **317**, 111–114.
Willerslev E, Davison J, Moora M *et al.* (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, **506**, 47–51.
Willerslev E, Hansen AJ, Binladen J *et al.* (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, **300**, 791–795.
Yoccoz NG, Bråthen KA, Gielly L *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, **21**, 3647–3655.
Zielińska S, Radkowski P, Blendowska A *et al.* (2017) The choice of the DNA extraction method may influence the outcome of the soil microbial community structure analysis. *MicrobiologyOpen*.

For Review Only

Minimizing polymerase biases in metabarcoding – Nichols et al.

Tables and Figures

Table 1. The six polymerases used in this study. *Platinum HiFi is a blend of two polymerases (one proofreading, one not).

Polymerase/mix	Manufacturer	Proofreading	Hot Start
AmpliTaq Gold, Buffer II	Applied Biosystems	N	Y
Kapa HiFi ReadyMix	Kapa Biosystems	Y	Y
Phusion	New England Biosciences	Y	N
Platinum HiFi	Invitrogen	Y*	Y
Q5 2x MasterMix	New England Biosciences	Y	Y
Qiagen Multiplex MasterMix	Qiagen	N	Y

For Review Only

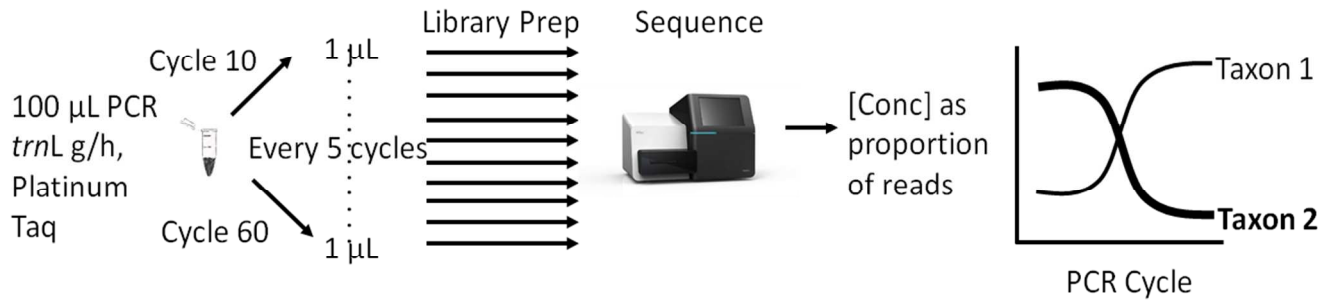


Figure 1. Schematic of the amplicon competition experiment. We chose four eDNA extracts and ran each in a PCR with *trnL* g/h and Platinum Taq using the recipe in Graham et al. 2016. Starting at cycle 10 and every five cycles up to cycle 60, we cooled the reaction to 20 C and removed 1 µL. We converted each 1 µL of PCR product into a sequenceable library individually. After sequencing and processing the reads, we plotted each amplicon as a function of PCR cycle and relative abundance.

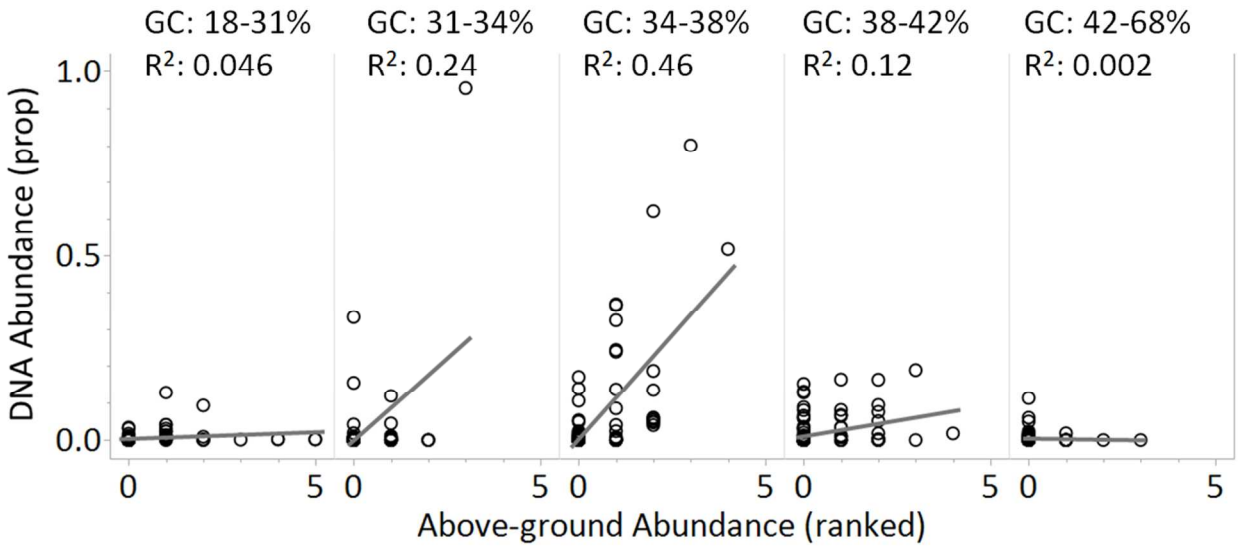


Figure 2. DNA abundance and above-ground abundance across average GC content bins. After collecting the DNA data we took all plant taxa that were identified at all locations, put them in the same plot and split them into GC content bins. Each point is a plant taxa where x is its ranked above-ground abundance and y is its mean DNA abundance across replicates. Some taxa are missing from the DNA data and some are missing from the above-ground data. For above-ground abundance, 5 is the highest rank, meaning the most abundant whereas 0 indicates absence. Lines are linear best fits with P-values > 0.3 for all bins except the middle bin where the P = 0.03.

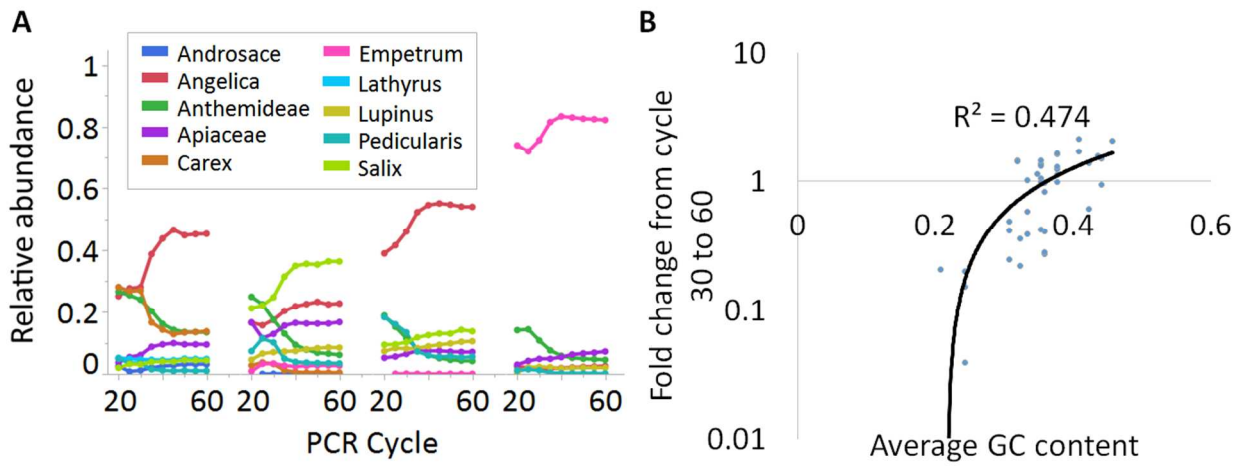
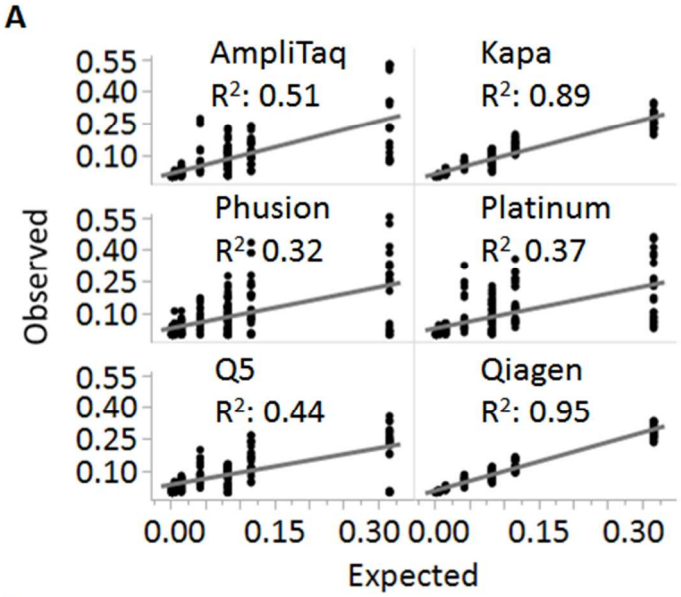
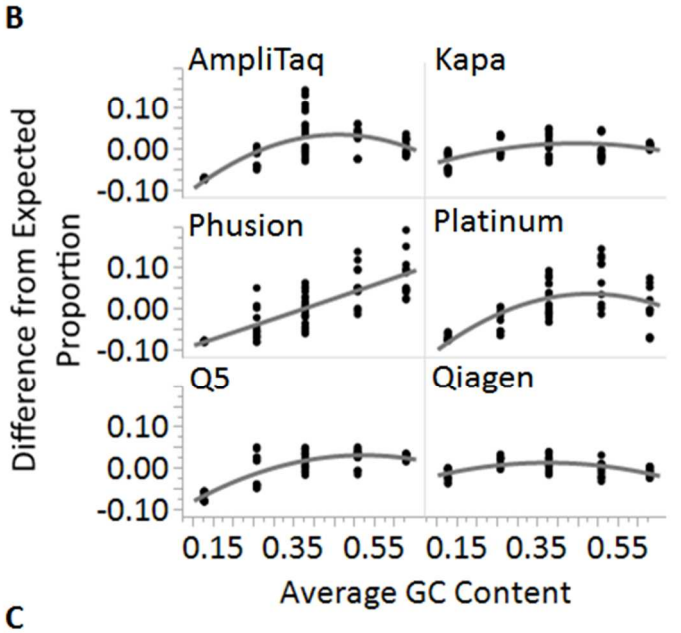
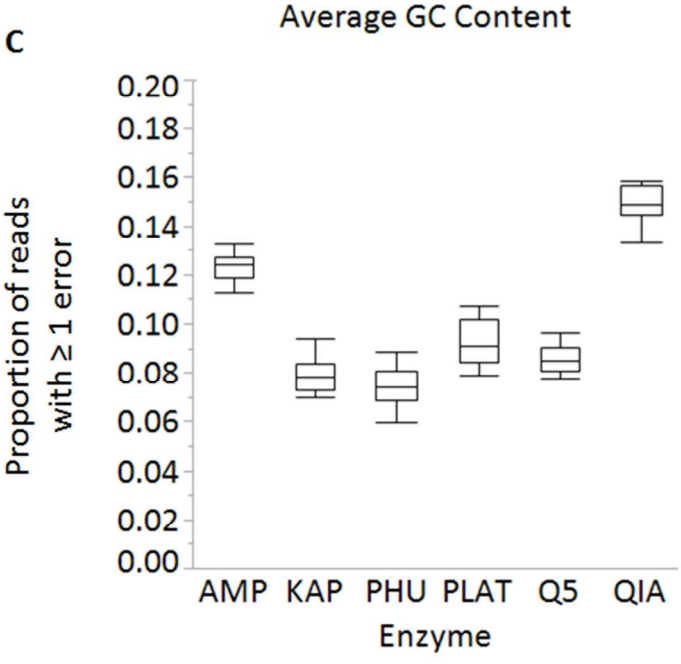


Figure 3. Changes in relative abundance over the duration of a 60-cycle PCR for four St. Paul Island sediment samples. A) Plots showing relative abundance measured at 5-cycle intervals between cycles 20 and 60. Colored lines show relative abundance estimates for the 10 most abundant plant taxa in these samples. B) Plot describing the fold change for each taxon in each experiment between cycle 30 and cycle 60, with a linear line of best-fit (P = 0.002), showing that change in relative abundance correlates with GC content. The y-axis is plotted on a log scale, therefore values above 1 indicate that the amplicon is increasing in abundance from cycle 30 to 60 and values below 1 indicate that the amplicon is decreasing in abundance.

For Review Only







For Peer Review Only

Figure 4. Testing polymerases using pools of synthetic oligonucleotides. A, B, and C combine data from six pools of synthetic oligonucleotides amplified using six polymerases. A) Observed proportions plotted against expected proportions for six polymerases. Each panel contains data for all six pools of oligos; B) Difference from expected proportions plotted against average GC content. Here we only used data from the equimolar pools; C) Proportion of reads with at least one error for each enzyme/mix.

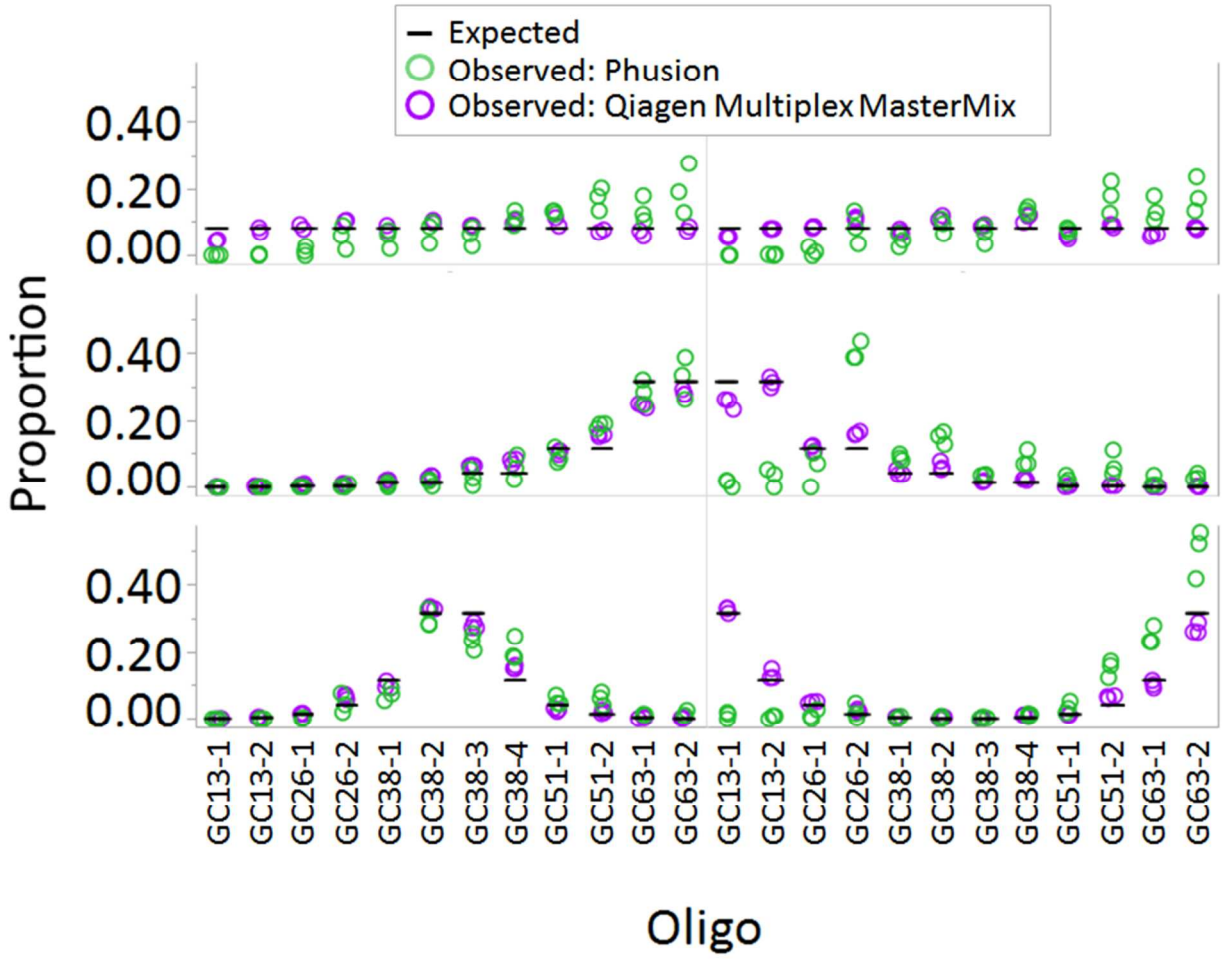


Figure 5. Expected (black lines) and observed abundances of the six synthetic oligonucleotide mixtures using Phusion (green open circles) and Qiagen Multiplex Master Mix (purple open circles) plotted as proportional data. Each oligonucleotide was pooled at 10 µM and then each pool was diluted to 10 fM. Each 10 fM pool underwent PCR using the six different polymerases.

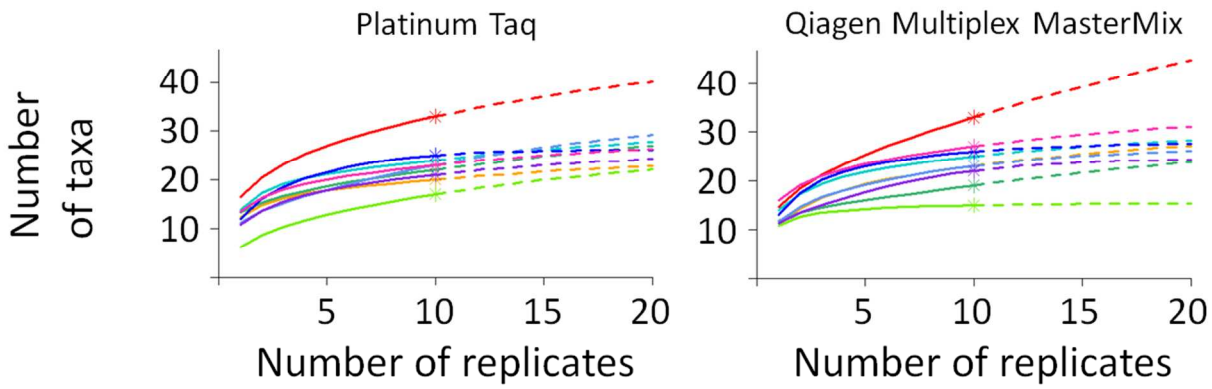


Figure 6. Rarefaction curves resulting from metabarcoding experiments for nine sites on St Paul Island, Alaska, using Platinum HiFi Taq as described in Graham et al. 2016 and the Qiagen Multiplex MasterMix following manufacturer's instructions. For each extract and polymerase, we performed 10 replicate PCRs. Rarefaction plots describe the number of unique taxa added per replicate. Solid lines are results from the 10 experiments, and dashed lines are predicted values calculated using iNEXT (Hsieh *et al.* 2016) in R version 3.4.2.

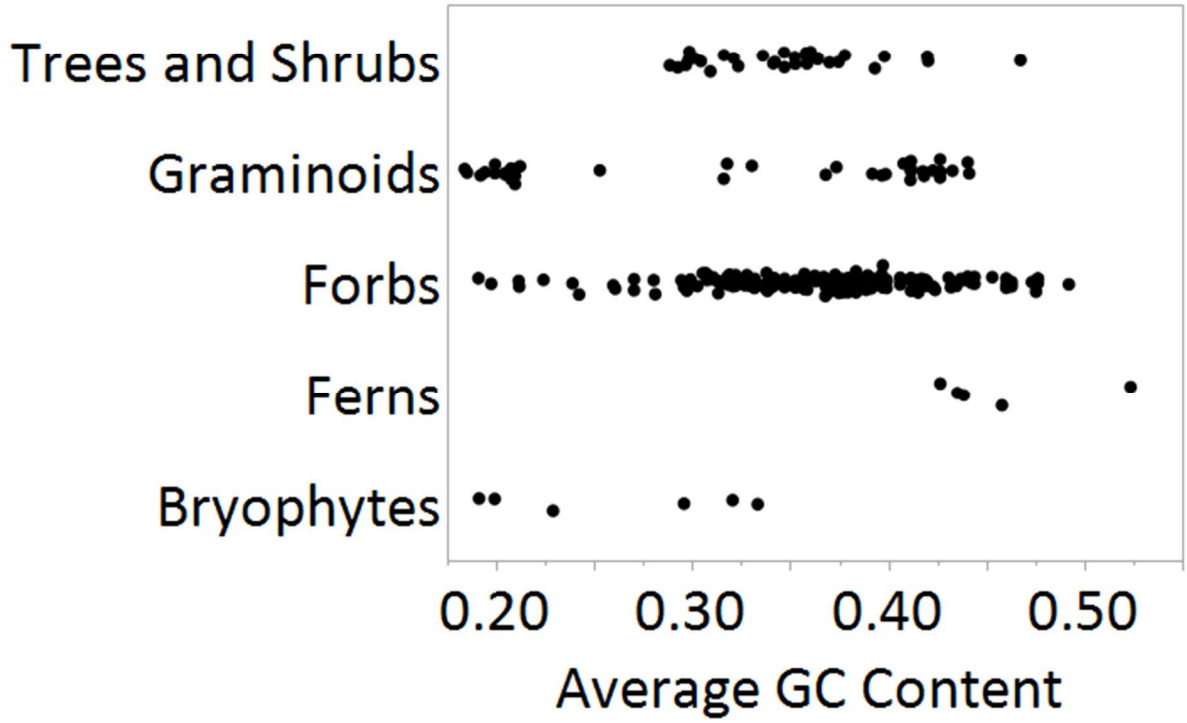


Figure 7. Average GC content across different plant growth forms. The data comes from the current study and Willerslev et al. 2014. Both studies used Platinum HiFi Taq polymerase. Ferns include horsetails.

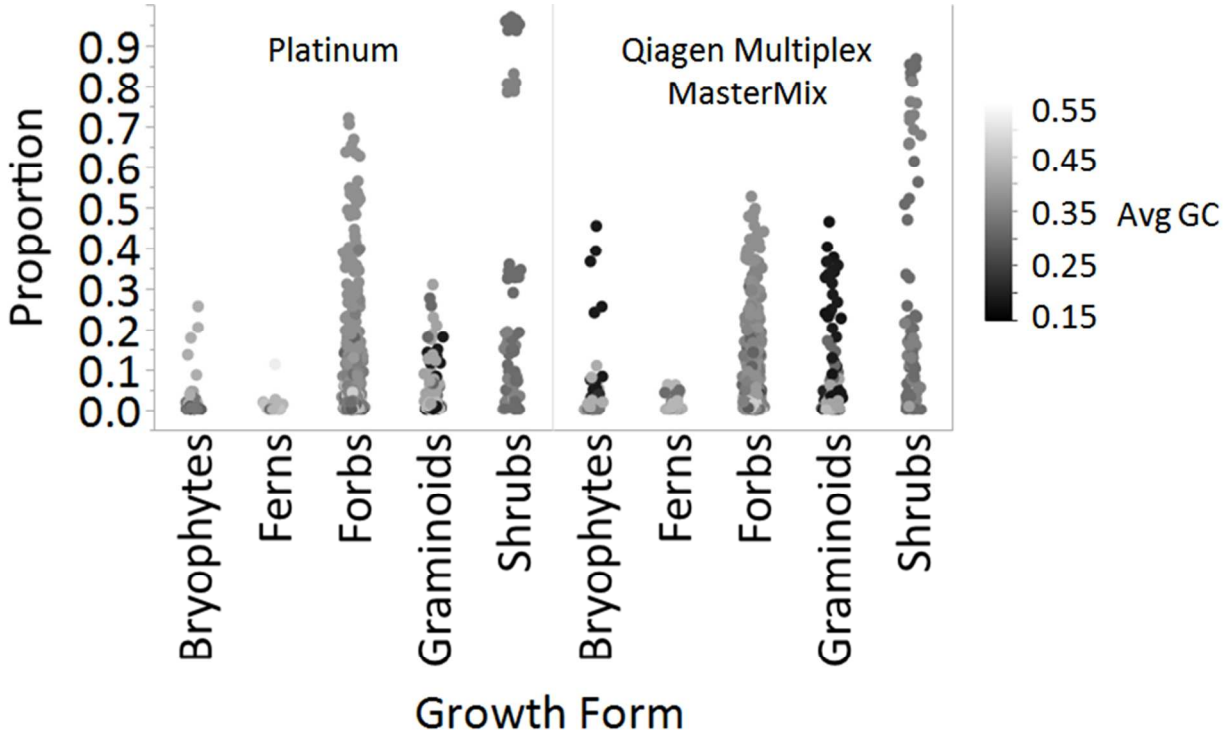


Figure 8. St Paul data generated using eDNA and separated into plant growth forms. Ferns include horsetails. All plant taxa from all samples are plotted both using Platinum HiFi Taq and Qiagen Multiplex MasterMix. Each data point is the relative abundance of a plant taxon from a particular location grouped into its growth form and shaded based on its average *trnL* p6-loop GC content. The darker the point is, the lower the average GC content.

For Review Only