The Quantitative Impact of an Environmental Intervention

NGN Team

4/26/2022

Abstract

Environmental laws around the world require some version of an environmental impact assessment surrounding construction projects and other discrete instances of human development. Information requirements for these assessments vary by jurisdiction, but nearly all require an analysis of the living elements of affected ecosystems. Because it is possible to sample and amplify the genetic material of many species present in those environments, amplicon-sequencing — also called metabarcoding or eDNA analysis — is a tractable, powerful, and increasingly common way of doing environmental impact analysis for development projects. Here, we analyze a 12-month time-series of water samples taken before, during, and after a construction project in a salmonid-bearing freshwater stream. We use an asymmetrical BACI design with four control streams to develop a robust null expectation against which to evaluate the impact of this discrete human intervention on the fish fauna of the treatment stream. After accounting for seasonal variability, we find XXX effects on YY species of fish as of ZZ months post-construction. We therefore demonstrate a rigorous, quantitative method for environmental-impact reporting using eDNA that is broadly applicable in environments worldwide. More specifically, in the context of billions of dollars of court-mandated road-culvert replacements taking place in the coming decade in Washington State, USA, our results suggest [salmonids are likely to recover quickly, etc., or whatever our actual results do, in fact, suggest]

Introduction

[much of the below from the grant proposal; need to update/smooth, obviously]

At present, it is difficult or impossible to measure the environmental impacts of discrete human activities, despite such assessment often being required by law.

Sequencing environmental DNA (eDNA) is means of surveying many species in a consistent and scaleable way. Sampling eDNA before, during, and after a development project would be a new and powerful way of assessing that project's impacts on the local biological communities, and conceivably could become the standard way to do such impact assessment.

All methods of environmental sampling are biased, in the sense that they capture a selective portion of the biodiversity present. Net samples for fish, for example, fail to capture species too small or too large to be caught in the net; bacterial cultures capture only those species that can be cultured on available media, and so forth. Despite the pleasing simplicity of the idea, there is no one way to survey the world and just "see what is there." Environmental DNA, however, comes as close to this goal as any method yet developed: a sample of water, soil, or even air, contains the genetic traces of many thousands of species, from microbes to whales.

Surveying the natural world by amplifying and sequencing DNA from environmental sources such as water, air, or soil has long been commonplace in microbial ecology (Rondon et al. 2000; Ogram, Sayler, and Barkay 1987; Turnbaugh et al. 2007) , but has recently become popular for characterizing ecological communities of eukaryotes (Port et al. 2016; Stat et al. 2017; R. P. Kelly et al. 2014; Valentini et al. 2016; Taberlet et al. 2012; De Vargas et al. 2015). Because the source of samples is the environment itself rather than

specific target organisms, the data resulting from such studies have become known as environmental DNA (eDNA) (Taberlet et al. 2012); the ultimate source of genetic material in the environment may be living or waste cells or extracellular DNA (Taberlet et al. 2012). Techniques that take advantage of such data may include non-PCR-based methods such as hybridization, but generally include an amplification step such as quantitative PCR, digital- droplet PCR, or traditional PCR from mixed templates followed by high-throughput sequencing. This last technique is known as metabarcoding, eDNA amplicon-sequencing, or more generally, marker-gene analysis.

In the metabarcoding context, broad-spectrum PCR primers capture many taxa across a very wide diversity of the tree of the life (Leray et al. 2013), but nevertheless the absence of a taxon from a sequenced sample does not indicate the absence of that taxon from the environment. Instead, the unsampled species simply may not have been susceptible to that set of PCR primers, and so failed to amplify. The result is often a dataset that represents hundreds or thousands of taxa, but these taxa are a fraction of a larger (and perhaps taxonomically broad) pool of species present. Using multiple, independent primer sets increases taxonomic scope by drawing from overlapping pools of taxa (Kelly et al. 2017), maximizing the likelihood of detecting any given taxon present. In virtually all comparisons, metabarcoding recovers far more taxa from an area than any other sampling method (Port et al. 2016; Kelly et al. 2017).

The results of metabarcoding studies differ dramatically from those of traditional, non- PCR-based sampling methods as a result of the PCR process itself. This exponential process means that 1) small changes in laboratory technique can yield large differences in outcomes, 2) PCR-based assays likely act differently on every target species, 3) there is consequently no one-to-one correspondence between the number of assigned reads in an eDNA study and the abundance of the source organism, and 4) neither might we expect a universally strong correlation in estimates of taxon-richness between eDNA and traditional methods. By understanding these process differences, we can correct for taxon-specific biases in ampification effiency to yield quantitative estimates of the community composition prior to PCR (Shelton et al. submitted).

The resulting dataset is compositional, revealing the proportions of each species' DNA present in each sample, but importantly this contains no information about the absolute abundance of DNA present. We can tie these proportional estimates to absolute abundances using additional data such as a qPCR assay for one of the taxa present.

. . .

Here, we report the results of a yearlong sampling effort before, during, and after a small construction project in our experimental creek, assessing the impact of that project on the salmonid species present. We do so using a combination of metabarcoding (12s mtDNA) and qPCR to yield quantitative estimates of the concentrations of DNA present at each time point, and we use parallel samples from an additional four control creeks to develop a causal analysis of changes in these concentrations.

As a result of a federal court ruling [US v Washington, culvert case, J. Martinez], Washington State is under a court order to replace many of the culverts that allow water to pass under roads and highways. These culverts, at present, collectively prevent or hinder anadomous salmon species from using [thousands of square kilometers] of habitat, which in turn violates the treaty rights of the region's indigenous tribes. Because replacing culverts can require substantial intervention – for example, diverting the water from a creek segment and rebuilding the road with a redesigned culvert – and because these replacements occur serially according to a schedule, they present an attractive experimental design.

. . .

A clear opportunity for policy-relevant eDNA work is in using its power to survey many species at a time to improve the way we assess the impacts of human activities. Within the United States, both state and federal laws often require a form of environmental-impact assessment for medium- to large-scale projects (i.e., those that might have a significant impact on the environment). Outside the US, many nations have their own versions of these same laws. Environmental assessments have begun to make use of eDNA for such work [CITE NatureMetrics, few published examples] around the world; in the U.S., however, assessments generally continue to rely on literature reviews or field measurements of a few key species, selected beforehand.

[Moreover, they often lack post-project sampling, given that the goals of a development project normally focus on construction itself, etc.]

Methods

Sites

Water samples were collected monthly for one year in five creeks in northwest Washington (add map). Sampling in each creek occurred both downstream and upstream of culverts with varying levels of fish passability under roadways of various sizes. Sample design followed a BACI (Before-After-Control-Impact, add citations) design. [Ez to add info here.] Padden Creek was the experimental creek as over the course of the year-long sampling, the previous, impassable culvert was removed and replaced with a new culvert. [RPK: Open question about what to do with two different culverts here; we have to decide if this is experimentally important, or if we can just look at downstream and up5, keeping the experimental design cleaner. EA: I think we should just do one. I think it would have to be either Dn/Up11 or Up11/Up5 - the time frame works for Dn/Up11 so I think that is what is best (despite knowing now that the biggest difference really is with Up5).] Two of the four other creeks had culverts allowing fish passibility, and two had culverts blocking fish passage. Fish passibility was determined by the Washington Department of Transportation [add info here].

*STOCKING IN LAKE PADDEN

For this study, we refer to the upstream and downstream locations in each creek where water sampling was conducted as a "station" (i.e., Padden-Downstream, Portage-Upstream, etc.), and each biological replicate (3 per station) a "sample". We therefore collected 3 samples at each station, 2 stations per creek, at 5 creeks, for 12 months, for a total of approximately 360 samples.

```
{r FIG1_map, fig.cap="\\label{fig:map} Figure 1: Map of sampling
locations in Bellingham, Washington. Locations of stream gauges
are shown for three of the five creeks sampled.", fig.path=here("Output"
echo=F, message=F, warning=F, include = T, fig.height=4, dpi =
300} # knitr::include_graphics(here("Output","Figures","map.png"),
auto_pdf = TRUE) #
```

Water Sampling

At each station, triplicate biological water samples (2 L each) were collected using Smith Root's eDNA Backpack. For most months, a trident sampler was used to collect all 3 biological replicates at the exact same time, for a total sampling time of about 5 minutes. Otherwise, the three replicates were collected consecutively, for a total sampling time of about 15 minutes. The backpack also monitored pressure and flow rate over the course of sampling. The backpack was set to have a target flow rate of 1 L/min and a max pressure of 12 psi. In some months, less than 2 L of water was filtered due to clogging (Supplemental Table X). Water samples were filtered using single use inlet tubes onto 5 um pore-size self-preserving filters (Smith Root, Vancouver, WA). After filtering, the filters were dried, brought back to the laboratory, and kept at room temperature until DNA extraction within 1 month (check longest we waited) of collection (cite self-preserving filter paper).

DNA Extraction, Amplification, Sequencing

All molecular work was performed at the University of Washington. Benchtops were cleaned with 10% bleach for 10 minutes and then wiped with 70% ethanol. Molecular work was separated onto pre- and post-PCR benches; all DNA extractions and PCR preparation was conducted on a bench where no PCR product was handled.

Filters were removed from their housing with sterile tweezers and were cut in half using sterile razor blades. Half of each filter was archived at -20C. DNA was extracted off the second half of the filter using a Qiashredder column (Qiagen, XX) and the DNEasy Blood and Tissue Kit (Qiagen, XX) with an overnight incubation (Supplemental File X). Extracts were eluted in 100 uL of molecular grade water, quantified via Qubit (Invitrogen, XX) and stored at -20C until PCR amplification within X months of extraction (check longest time waited).

We used a published fish-specific primer set (MiFish) targeting a ~170 bp hypervariable region of the mitochondrial DNA 12S rRNA gene for PCR amplification [cite MiFish]. Primer sequences were slightly modified from the original publication based on Praebel and Wangensteen [cite]. Primer sequences also included the Nextera overhang sequences to add unique indices after the first PCR amplification. The primers used were as follows: F 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCGGTAAAACTCGTGCCAGC 3', R 5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATAGTGGGGTATCTAATCCCAGTTTG 3' (italics indicates Nextera overhang). PCR reactions included 10 uL of 5X Platinum ii Buffer, 0.4 uL of Platinum ii Taq, 1.25 uL of 8 mM dNTPS, 1.25 uL of 10 uM F primer, 1.25 uL of 10 uM R primer, 5 uL of template, and 30.85 uL of molecular grade water, for a total reaction volume of 50 uL. Cycling conditions were as follows: 95C for 2 min, 35 cycles of 95C for 30 sec, 60C for 30 sec, 72C for 30 sec, followed by a final extension of 72C for 5 min. Each month of samples was run on a single plate with the addition of a no template control (NTC; molecular grade water in lieu of template) and a positive control (genomic DNA from kangaroo). After PCR amplification, PCR products were visualized on a 1-2% gel. If no band was present, extracts were diluted 1:10 until a band was detected. Following PCR and gel visualization, PCR products were size selected and cleaned using MagBind Beads (Omega Biotek, XX) at a sample:beads ratio of 1.2. Bead-cleaned PCR products were eluted in 30 uL of molecular grade water and quantified via Qubit.

A second PCR amplification (or indexing PCR) added a unique index to each sample using Nextera indices (Illumina, XX) to allow pooling multiple samples onto the same sequencing run. For the second PCR, 10 ng of the first PCR product was used as tempalte at a final volume of 11.25 uL. For samples with concentrations less than 0.88 ng/uL, 11.25 uL was added despite being less than 10 ng of amplicon. Each sample had a unique mastermix with the primers being the complementary Nextera overhang and the unique index. Sets A and B were used, making sure to never use the same index for more than one sample on one sequencing run. The PCR reaction included the 11.25 uL of template, 12.5 uL of Kapa HiFi MMX (Roche, XX), and 1.25 uL of indexed primer. Cycling conditions were as follows: 95C for 5 min, 8 cycles of 98C for 20 sec, 56C for 30 sec, 72C for 3 min, and a final extension of 72C for 5 min. Indexed PCR products were also size-selected and purified using MagBind Beads (Omega Biotek, XX) at a sample:beads ratio of 0.8. Bead-cleaned PCR products were eluted in 30 uL of molecular grade water and quantified via Qubit.

Samples were randomized in 3-month blocks and split across 3 sequencing runs, for a total of 11 MiSeq runs. Indexed and bead-cleaned products were normalized before pooling. Libraries were quantified via Qubit and also were run on a Bioanalyzer (Agilent, XX) before sequencing. The loading concentration of each library was 4 nM and 5-20% PhiX was included depending on the composition of the run (Supplemental Table X).

Bioinformatics

Pipeline description and online availability. N sequences, etc., with table of pipeline steps and surviving N sequences at each of those steps.

After sequencing, bioinformatic analyses were conducted in R [cite R]. Primer sequences were removed using Cutadapt (Version 1.18) [cite cutadapt] before dada2 [cite] trimmed, filtered, merged paired end reads, and generated amplicon sequence variants (ASVs). Taxonomic assignment was conducted via the insect package

[cite] using a tree generated by the developers for the MiFish primers that was last updated in November 2018. Only species level assignments from insect were retained and ASVs not annotated or not annotated to species level were then checked against the NCBI nucleotide database using BLAST+ [cite]. Query sequences that matched a single species at >95% identity were retained.

Quantitative PCR

We used two recently published qPCR primer sets and probes to quantify cutthroat trout (Oncorhynchus clarkii) and Coho salmon (Oncorhynchus kisutch). The O. clarkii assay targeted a 114 bp fragment of the cytochrome b gene. The primers/probe sequences were: F 5' CCGCTACAGTCCTTCACCTTCTA 3', R 5' GATCTTTGTATGAGAAGTAAGGATGGAA 3', P 5' 6FAM-TGAGACAGGATCCAAC-MGB-NFQ 3'. The O. kisutch assay also targeted a 114 bp fragment of the cytochrome b gene. The primers/probe sequences were: F 5' CCTTGGTGGCGGATATACTTATCTTA 3', R 5' GAACTAGGAAGATGGCGAAGTAGATC 3', P: 5' 6FAM-TGGAACACCCATTCAT-MGB-NFQ 3'. All samples were run with the O. clarkii assay. The O. kisutch assay was only used for samples from Padden Creek and for samples from the other four creeks where O. clarkii was not detected.

The O. clarkii assay was multiplexed with TaqMan Exogenous Internal Positive Control Reagents (EXO-IPC) (Applied Biosystems) to check for the presence of PCR inhibitors. Each DNA sample was run in triplicate using Gene Expression Mastermix (XX), a final concentration of 0.375 uM F primer, 0.375 uM R primer, and 0.105 uM probe, as well as 1X EXO-IPC mix, 1X EXO-IPC DNA, 3.5 uL of template for a final reaction volume of 12 uL. The EXO-IPC mix includes the primers and probe for the EXI-IPC DNA, with the probe having a VIC reporter, allowing it to be multiplexed with the O. clarkii assay, which has a FAM reporter. The O. kisutch assay was also run in triplicate using Gene Expression Mastermix (XX), a final concentration of 0.375 uM F primer, 0.375 uM R primer, and 0.105 uM probe, 3.5 uL of template, and molecular grade water for a final reaction volume of 12 uL (without the EXO-IPC kit). For both assays, the thermocycling was as follows: 50C for 2 min, 95C for 10 min, followed by 45 cycles of 95C for 15 sec, 60C for 1 min.

The cycle threshold (Ct) value determined for the EXO-IPC assay from the NTC was compared to the Ct value for the EXO-IPC assay in each of the enviornmental samples. If the Ct value was >0.5 Ct values from the mean Ct for the NTCs, the sample was deemed inhibited and diluted 1:10 until the Ct value fell within the accepted range. For inhibited samples, the final dilution determined was then used for the O. kisutch assay and after converting Ct values to DNA concentrations using the standard curve, the concentration was multiplied by the dilution factor.

Each plate included a 8-point standard curve created using synthetic DNA (gBlocks) at the following concentrations: 100,000 copies/uL, 10,000 copies/uL, 1,000 copies/uL, 100 copies/uL, 10 copies/uL, 5 copies/uL, 3 copies/uL, 1 copy/uL. Additionally, six NTCs were included on each plate: 3 with the IPC DNA mix and 3 with molecular grade water instead of template or IPC DNA mix. All qPCRs were conducted on an Applied Biosystems StepOnePlus thermocycler. Plates were re-run if efficiency as determined by the standard curve was outside of the range of 90-110%.

All qPCR data was processed in R using Stan. Absolute concentrations from environmental samples were estimated by building a model to incorporate uncertainty from the standard curve for each plate. The Bayesian model estimates the

Quantitative Metabarcoding Model

Calibration with mock community; description of Stan model with reference to Shelton et al.

Tying proportions to absolute concentrations using qPCR; these absolute concentrations are a derived quantity in the Stan model, and they incorporate uncertainty from the qPCR standard curve.

```
{r FIG2_conceptualframework, fig.cap="\\label{fig:map} Figure
2: Conceptual framework outlining data and models used for
analyses.", fig.path=here("Output","Figures"), echo=F, message=F,
warning=F, include = T, fig.height=4, dpi = 300} # knitr::include_graphi
auto_pdf = TRUE) #
```

Estimating the Effect of Culverts

Estimating the Effect of Culvert Replacement

Consistent with the asymmetrical BACI study design, we use observations from our four control creeks to develop a null expectation, against which we compare the observations in Padden Creek, our experimental site. Here, we use the difference in DNA concentration upstream vs. downstream of the existing culvert as our outcome variable of interest. We calculate this difference – treating our biological replicates (within a site) as being drawn from a common distribution, and treating upstream and downstream sites within a creek as independent [ALTHOUGH WE COULD CHANGE THIS] – independently for each month in each creek. [WE COULD do this as a time-series, where months within creeks aren't independent, etc... let's discuss.]

The result is a null distribution of [deltaSalmonDNA] for each species, in the absence of construction or other major intervention. This null varies by month, and so accounts for the expected strong seasonal patterns of salmonids. We then assess the effect of the culvert replacement [construction/intervention] in our experimental creek, comparing our observed [deltaSalmonDNA] in Padden to our null for each species in each month. Prior to culvert replacement, we expect no difference between experimental and control creeks; observed differences during and after construction are attributable directly to the construction project.

Results

Metabarcoding Results

Environmental samples and mock communities were sequenced across 12 Illumina MiSeq runs. The total number of reads generated was ~42 million across XX environmental samples (12 months x 11 stations x 3 biological replicates = 396 filters) and 27 mock community samples (3 communities x 9 replicates [6 even, 3 skewed proportions]). After quality filtering and merging all runs, XX reads remained from XX amplicon sequence variants (ASVs), of which XX% of reads and XX% of reads were annotated to species level.

Importantly, the ASVs assinging to salmonid species in the mock communities were found in environmental samples, providing confidence in the species level assignments and application of the quantitative metabarcoding model to correct for different amplification efficiencies across similar species.

The most common salmonid species found in the environmental samples was XX (cutthroat trout?), with XX% of samples across times, creeks, and stations having at least 50% of reads assigned to *O. clarkii*.

Cutthroat Trout Quantitative PCR Results

All environmental samples (n = 356, should be 357, one set of triplicates missing from Sqm Up August, otherwise 3x2x5x12 = 360) were quantified for absolute concentrations of cutthroat trout DNA across 30 plates, resulting in 280 samples with a positive detection in at least 1 of 3 technical replicates. A subset of samples (n = 111, all of Padden Creek and then any samples that were not detected using the cutthrat assay) were quantified for coho salmon across 12 plates, resulting in 56 samples with a positive detection in at least 1 of 3 technical replicates. For all 42 plates, efficiencies ranged from X to X%.

NOTE TO SELF - go back and find why 0222_1Prt_Up_2 has two sets of triplicates and 0222_1Prt_Up_1 is missing in cut – and coho has too many reps for 0222.3Chk.Up.1, 0621.2Brn.Up.2, and 1121.2Brn.Dn.3.

The majority of environmental samples (65%) were determined to be inhibited. The highest dilution factor to remove inhibition was 1:1000, but the most common dilution factor was 1:10 (75% of inhibited samples).

The modeled output of cutthroat trout DNA concentrations after converting Ct values to copies/L of water (after accounting for dilution and volume filtered), ranged from 10 copies/L to 1,377,656.67 copies/L, with a mean value of 57,529 copies/L. Coho DNA concentrations ranged from 27 copies/L to 1,214,611 copies/L with a mean value of 46,132 copies/L.

```
{r FIG3_qpcr, fig.cap="\\label{fig:qpcr} Figure 3: Absolute
concentration (copies/L of water) of cutthroat trout (Panel A)
and coho salmon DNA (Panel B) as measured by qPCR.", fig.path=here("Outp
echo=F, message=F, warning=F, include = T, fig.height=4, dpi =
300} # knitr::include_graphics(here("Output","Figures","modeled_cut_qpcr
auto_pdf = TRUE) #
```

Quantitative Metabarcoding

The intercalibration of the mock community samples demonstrates the rank order of ampliciation efficiencies for salmonids (Supplemental Figure X). Of the seven salmonids included in the mock community, *O. gorbuscha* had the highest amplification efficiency and *O. kisutch* had the lowest amplification efficiency (Supplemental Figure X).

In the environmental samples, proportions of reads were then corrected to account for these varying amplification efficiencies to obtain the initial proportions of DNA per species prior to PCR amplification (Figure X).

Using the results from the qPCR assay for *O. clarkii* and the corrected proportions of starting DNA concentrations from the quantitative metabarcoding model, the total concentration of DNA of salmonids can be inferred by dividing the concentration of *O. clarkii* (copies/L) by the proportion of *O. clarkii* in each environmental sample. The total DNA concentration of salmonid DNA in environmental samples ranged from XX to XX copies/L (Supplemental Figure X), with the highest concentrations found in X month and in X creek.

```
{r FIG4_metabarcoding, fig.cap="\\label{fig:metabarcoding}
Panel A) Compositions of salmonid DNA as determined by metabarcoding
after correction for amplification bias. B) Absolute concentrations
of salmonid DNA using total DNA derived from cutthroat qPCR
results.", fig.path=here("Output","Figures"), echo=F, message=F,
warning=F, include = T, fig.height=4, dpi = 300} # knitr::include_graphi
auto_pdf = TRUE) #
```

Culvert Effects

Construction Effects

Discussion

Environmental DNA can provide quantitative measurements of environmental impacts

A clear seasonal pattern occurred for all the salmonids detected in the study. The time series model uses shared information across creeks to include the change in eDNA concentrations due to time, whether a sample was collected below or above a barrier (i.e., culvert), and whether or not there was construction occurring.

In this was, we could isolate the changes in eDNA concentrations as a result of the intervention (i.e., construction) while accounting for the variance due to time and station (i.e., culvert).

Not all culverts are barriers

By measuring DNA concentrations of species above and below culverts on a small spatial scale, we were able to determine how much of a barrier each culvert was (or was not) to fish passage. We found that four of the five creeks sampled were not actually major barriers to fish passage. The only creek that was determined to be a barrier to fish passage was Barnes Creek, which was also the largest distance between the downstream and upstream station.

Though eDNA can move downstream with water flow, here, we were measuring if culverts were barriers to fish moving upstream as we were focused on the impact of culverts on migratory salmon. In our case, we were comparing if downstream stations had higher DNA concentrations than upstream stations as a result of fish being unable to get upstream.

Salmonids can survive a bulldozer in a creek

The intervention (i.e., construction) in Padden Creek occurred over about two months and included the "de-watering" of the creek, removal of the existing culvert, installation of the new culvert, and then the "re-watering" of the creek.

Conclusion