University of Arkansas, Fayetteville
ScholarWorks@UARK

Theses and Dissertations

8-2016

Stream Microbial Communities as Potential Indicators of River and Landscape Disturbance in North-Central Arkansas

Wilson Howard Johnson
University of Arkansas, Fayetteville

Follow this and additional works at: http://scholarworks.uark.edu/etd

 Part of the Fresh Water Studies Commons, Terrestrial and Aquatic Ecology Commons, and the Water Resource Management Commons

Recommended Citation
Johnson, Wilson Howard, "Stream Microbial Communities as Potential Indicators of River and Landscape Disturbance in North-Central Arkansas" (2016). *Theses and Dissertations*. 1624.
http://scholarworks.uark.edu/etd/1624

This Thesis is brought to you for free and open access by ScholarWorks@UARK. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu, ccmiddle@uark.edu.

Stream Microbial Communities as Potential Indicators of River and Landscape Disturbance in North-Central Arkansas

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Biology

by

Wilson H. Johnson
University of Arkansas
Bachelor of Science in Biology, 2013

August 2016
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Dr. Michael E. Douglas Thesis Director	Dr. Jeffrey A. Lewis Co-Thesis Director
Dr. Marlis R. Douglas Committee Member	Dr. Franck Carbonero Committee Member

Abstract

In the past decade, 29 shale basins have been actively developed across 20 states for extraction of natural gas (NG) via horizontal drilling/hydraulic fracturing (=fracking). This includes ~5000 wells within the Fayetteville shale of north-central Arkansas. Development often impacts both river- and landscapes, and management requires catchment-level evaluations over time, with organismal presence/absence as indicators. For this study next-generation sequencing was used to identify/characterize microbial communities within biofilm of eight Arkansas River tributaries, so as to gauge potential catchment influences. Streams spanned a gradient of landscape features and hydrological flows, with four serving as 'potentially impacted catchment zones' (PICZ) and four as 'minimally impacted catchment zones' (MICZ). Overall, 46 bacterial phyla and 141 classes were identified, with 24 phyla (52%) and 54 classes (38%) extending across samples. A principal coordinate analysis arrayed samples according to stream order, suggesting a relationship between communities and gradients. With regard to river- and landscape disturbance, three preliminary indicators emerged: (1) Synechococcophycideae and Oscillatoriophycideae (=Cyanobacteria that act as primary producers exhibiting a positive correlation with increased nitrogen and phosphorus) were significantly more abundant at PICZ sites ($P<0.049$), suggesting elevated nutrient availability; (2) Spartobacteria (a heterotroph negatively associated with salinity) was significantly more abundant at MICZ sites ($p<0.01$), suggesting lower concentrations of brine; (3) Actinobacteria, a bioremediator capable of complex and far-ranging removal of toxic pollutants, was significantly more prevalent at PICZ sites ($p<0.039$). Our results suggest that hydrology and location of NG well pads are potential

covariates in defining microbial communities in study streams. However, long-term follow-up studies are needed to substantiate estimates and provide greater confidence in defining suggested impacts.

Acknowledgments

With immense appreciation, I would like to thank my major professor and co-advisor Dr. Michael Douglas for his genuine support and guidance through my graduate education and research. My progress stems largely from his efforts and advocacy on my behalf. I would also like to thank my other co-advisor, Dr. Jeff Lewis, and Dr. Tara Struecker whose tireless support, instruction, and sharing attitude greatly aided my lab work and completion of this research. I would like to thank Dr. Marlis Douglas for her assistance in lab work, and advice on constructing a cohesive and proper thesis. Also I would like to thank Dr. Franck Carbonero for using his talents in sequencing my samples, and sharing his experience with the bioinformatic process.

I am very grateful to Dr. Michelle Evans-White and Dr. Brad Austin for taking me sampling at their stream ecology sites and for sharing data that directly relates to this research and for encouraging continuing investigation into these important ecological questions.

I would like to thank the office staff of the Biological Sciences program, in particular Ms. Becky Harris, who continuously offered help or answered questions without any hesitation or reservations. Her presence in the department has been an integral part of my success and she is a valuable resource.

Finally, I would like to thank my family and friends who have all encouraged my progress and cheered on my accomplishments. In particular, my mother, who enthusiastically supports my goals and aspirations. Her wisdom and thoughtful guidance has greatly influenced me. I hope my accomplishments will honor her dedication and love.

This research was in part funded by the 21st Century Chair in Global Change Biology and the Bruker Professorship in Life Sciences.

Table of Contents

Page

List of Tables

List of Figures

Abbreviations and Definitions

1**. AOA/AOB ratio** – Ammonia oxidizing archaea versus ammonia oxidizing bacteria ratio as indicator of ecosystem health, with ratio >1 indicating general health.

2. **Epilithic** – Organic organisms growing on surface materials, such as rocks

3. **MICZ** – Minimally impacted catchment zone

4. **$OGIP^{free}$** – Original free natural gas in place (in BCF): Indicates potential gas production

5. **OTU** – Operational taxonomic unit, an operational definition of a group of species, genera, or other qualifying identity used in sequence analysis

6. **Periphytic** – Subsurface microbial populations commonly attached to substrate or sediment

7. **PICZ** – Potentially impacted catchment zone

8. UNG – Unconventional natural gas (i.e. hydraulic fracturing, hydrocarbon extraction using high pressure liquids in shale rock)

Introduction

The quality of water was an early focus of aquatic studies (Duncan and Hoppe-Seyler 1893, Winterstein 1908), with dissolved gases and their effects on aquatic biodiversity an emphasis (Ball 1922, Gutsell 1929). Subsequent work often focused on biodiversity aspects as barometers for ecosystem health, with presence (or absence) of focal species as a driver for appropriate water quality metrics (Cairns et al. 1993). Larger aquatic organisms such as fishes, mussels, and aquatic insects most often served as sentinel species for these studies (Ball 1922, Gutsell 1929, Armitage 1958), given the relative ease with which they could be collected and the taxonomic databases available for their identification (Wiggins and Mackay 1978, Karr 1981).

It is apparent from many of these monitoring studies that aquatic environments are often impacted by agricultural land use (Tong and Chen 2002). These impacts include enhanced nutrient content from fertilizer applications, presence of pesticides, and increased turbidity from soil erosion (Gilliom et al. 2006, Bernot et al. 2006).

Another anthropogenic impact to groundwater and stream environments is an industrial activity, termed 'fracking.' This process extracts trapped hydrocarbons from sedimentary shale formations generally several thousand feet below the surface by injecting fluids at high pressures to fracture and release the trapped elements (Colborn et al. 2011). Fracking has developed into a rapidly growing industry that offers potential economic development and a relatively clean energy source (Springer 2011). However, this industry is largely unregulated and its environmental risks are numerous and well documented (Vidic et al. 2013), to include: accidental spills, release of fracking chemicals into groundwater, wastewater discharge, induced

seismic activity, gas migration, and altered drainage conditions leading to unintended sediment discharge (Lange et al. 2013).

Ecological studies have been conducted in the wake of this burgeoning industry in an attempt to assess potential impacts (Colborn et al. 2011, Austin et al. 2015). The recognition that aquatic microbial communities often form sessile attachments, known as biofilm (Geesey et al. 1978), provides a component to stream monitoring that allows large numbers of OTUs (operational taxonomic units) to be assayed *in toto* (Edgar 2013). The identification and characterization of biofilms provide attractive aspects: an ability to colonize submerged surfaces in a symbiotic composition, a long term exposure to the variability inherent in flowing water, and a means to efficiently evaluate primary production from epilithic or subsurface attached species (Haack and McFeters 1982, Stock and Ward 1989, Proia et al. 2012).

Biofilm communities have primarily been characterized with regards to species diversities and abundances, but their intimate association with environmental factors in streams adds a critical but less studied aspect to their role in ecosystem dynamics (Battin et al. 2003). These associations may include the correspondence of biofilm community structure with a variety of impacts, such as: anthropogenic waste and pollution (Allan et al. 2012), extreme weather events (Pandey and Soupir 2012), natural erosion and alterations (Moslemi et al. 2012), and spatial geography and catchment land-use (Clapcott et al. 2012, Coles et al. 2012). Others studies have suggested effective methods for interpreting and utilizing these data (Poff 1997, Cooper et al. 1998).

Microbial communities in aquatic ecosystems were largely ignored as recently as a decade ago, and for obvious reasons, despite their critical role in decomposing suitable biological substrates and recycling nutrients (Xu et al. 2006). For example, the photosynthetic components

of these communities, such as cyanobacteria, can contribute greater than 80% of primary production, while other components can remediate harmful substances through complex metabolic pathways (Gadd 2010). Microbial communities (*sensu lato*) are thus a vital component of aquatic ecosystems (Cohen 2006).

The primary reason for a dearth of research on the microbial component of aquatic biodiversity was largely due to technical aspects, in that identifications were done optically and with much effort, and the compilation and annotation of community structure(s) for monitoring purposes was quite laborious (Henrici 1933, Karl 1986, Geesey and White 1990). In addition, many constituents of biofilm are difficult or currently impossible to culture in a laboratory (Chiu et al. 2014). These issues have been ameliorated of late with the onset of molecular advances in genomics (Tringe et al. 2005), transcriptomics (Poretsky et al. 2009), and proteomics (Ram et al. 2005), that now provide in tandem the necessary window into the composition and complexities of aquatic microbial communities within biofilms. Genomic approaches can now be used not only to characterize microbial communities but also to interpret their community dynamics.

Ribosomal RNA has been the primary molecular marker used to identify species (Pace et al. 1986), and the 16S ribosomal RNA region has been particularly favored, as it contains both conserved and hyper-variable regions that amplify easily and yield specific identifications (Schütte et al. 2008, Pham et al. 2009, Chiu et al. 2014). Furthermore, next generation sequencing technology such 'sequencing by synthesis,' developed by Illumina for its MiSeq® platforms, reduces the cost of these methods and improves their accuracy (Bokulich 2012, Zarraonaindia 2013).

The current project utilizes a molecular genetic approach to assay biofilm communities in selected streams found within a 360 square mile region located within the Fayetteville shale

region of northwest Arkansas (Bai et al. 2013). Stream reaches are within the foothills of the Boston Mountains, a limestone-based karst topography containing prolific ground and spring-fed streams (Adamski et al. 1995). The objectives of the study were to characterize the microbial biofilm communities in these streams and compare species-composition against a series of abiotic factors that are characteristic of the watershed. Knowledge of resident microbial communities and their abiotic environment will identify links between land-use and stream environmental conditions that are necessary prerequisites for monitoring of aquatic ecosystems. In this sense, alterations of stream conditions can dramatically alter microbial community composition and promote disturbances in higher trophic levels, leading to community disintegration (Schwarzenbach et al. 2006). These data will also provide insight into the manner by which water quality and changing land use impact stream conditions and ecosystem functioning.

Hypotheses and Objectives

Hypotheses for this study are: (a) there are no differences in microbial communities between the potentially impacted catchment zone streams (PICZ) and minimally impacted catchment zone streams (MICZ), and (b) spatial or abiotic factors are not significantly correlated with microbial community composition.

To accept or reject these hypotheses the following objectives will be addressed in this study. First, accurately sample the identified sites in a manner that minimizes bias, strives to limit any introduced errors, and utilizes the best practices as determined by the relevant previous research. Second, extract the collected biofilm samples using a protocol (experimentally determined with additional samples) which maximizes both purity and coverage of DNA sampled. Third, utilize Next Generation Sequencing (Glenn 2011) to accurately target and sequence the hyper-variable V4 region of 16S rRNA to allow for culture-independent characterization of the various microorganisms and their community composition. Fourth, employ a bioinformatics pipeline using tested and reliable software to explore parameters that may determine accurate assessments of community metrics. This pipeline will provide parameters required for determining statistical significance, where possible, using diversity analyses, multivariate statistical analysis, and taxonomic assignments.

Materials and Methods

Sampling Sites and Spatial (GIS) data

Eight sites were chosen from a group that formed part of an ongoing stream ecology research project (Evans-White et al. 2013) (Figures 2, 3, and 4). Global Information System data (GIS) was provided for each sample site by the Nature Conservancy (TNC; Fayetteville). From these, six traditional landscape-scale variables were computed (Table 1), as well as three relating to Unconventional Natural Gas (UNG) activity [i.e., Well Density, Inverse Flow Length (IFL) and Impact]. 'Well Density' is number of UNG well sites within a square kilometer of each sample site. 'IFL' was calculated in ArcGIS for all well sites upstream of each sampling location by using the flow length tool in the 'Spatial Analyst Tool' toolbox within ArcGIS and (corrected for slope) to determine the length of flow from each well site to the stream channel. The inverse of each flow length was summed across all well sites for each catchment area, with wells more proximal having a higher value and thus a greater potential effect. For 'Impact,' a site was scored as '1' based on an IFL value ≥ 0.25 and a Well Density (no. /km^2) ≥ 0.5, otherwise '0.' A Ward hierarchical clustering dendrogram (Ward 1963) was employed to determine the relatedness of each site based upon the GIS variables, save Impact (Figure 1).

Sample collection

Two pools were selected at each site peripheral to the greatest stream flow and sampled at their upstream and downstream boundaries (Figure 5).

Once a suitable pool was identified, the following measures were then collected sequentially so as to replicate coverage and promote an accurate assay: (1) Canopy coverage; (2) horizontal location of sample within pool (i.e., 30% in from bankside pool edge); (3) depth

below the surface (standardized according to the first sample collected); and (4) substrate composition. Two biofilm samples were then obtained from each location, at each site, for a total of four samples per stream.

The sampling site in each pool was approached from downstream, and the biofilm-covered rock lifted by (nitrile-gloved) hand and sampled with a sterile Nasco Whirl-Pak Speci-Sponge™. Previous studies have extracted biofilms by simply scraping surfaces, yet this can promote contamination as well as inconsistent recovery rates (Gagnon and Slawson 1999). Samples were immediately returned to the Whirl-Pak, sealed, placed on dry ice in a cooler, transported to the lab, and stored at -80°C. Photographs were taken of each site and sampling location, as well as data on ambient air temperature, water temperature, weather conditions, depth and size of sample rock, and time of sampling (Table 2).

Biofilm extraction from Nasco Whirl-Pak Speci-Sponges™

Previous studies (Gagnon and Slawson 1999) employed a 'stomacher' (i.e., paddle mixer) to separate biofilm from sponges and suspend it within a solvent. This study instead employed a single wash with a standard PBS buffer solution coupled with five minutes of hand mixing (to simulate the action of a paddle mixer) for extraction of biofilm from sample sponges. Subsequent repetitions (i.e., centrifugation, removal of supernatant, weighing extract) did not yield additional materials. Biofilm was weighed to determine per-sample yield prior to DNA extraction (protocol summarized in Appendix 1).

DNA extraction

Three different protocols were evaluated: (1) a standard phenol-chloroform with ethanol extraction; (2) a commercial kit from Qiagen (QIAamp DNA Stool Mini Kit®); and (3) a commercial kit from MOBIO (PowerBiofilm® DNA Isolation Kit). DNA was quantified using the Thermo Scientific NanoDrop UV-Vis Spectrophotometer and the Life Technology Qubit® 2.0 Fluorometer.

Extractions were subjected to PCR and tested for amplification using four sets of primers (Appendix 1). Amplifications were confirmed using a 1% agarose gel and 1X TBE with a 1K ladder to observe expected fragments (supplemental Figure 1). The MOBIO Kit and included protocol (Appendix 1) were then used to extract the 16 samples.

Sequencing (Illumina MiSeq®)

Amplicon libraries were produced by PCR (Polymerase Chain Reaction). Primers were chosen to amplify the hyper-variable V4 region of the 16S structural subunit rRNA gene as designed by Caporaso et al. (2012b). These primers allow a dual-index and bi-directional sequencing setup on the Illumina® MiSeq sequencer that greatly improves the quality score of each assigned nucleotide and thus increases confidence in a correct sequence assignment.

Amplicons were gel extracted, pooled, and then re-amplified using Solexa primers as described by Klindworth et al. (2012). The raw reads were de-multiplexed with MiSeq Reporter software™ installed within the MiSeq® platform. The de-mulitplexed reads were stored in FASTQ format and made available on the BaseSpace® cloud application (Illumina 2011) for downloading and additional third party processing.

Bioinformatics

The 16 samples yielded 32 FASTQ files (i.e., forward sequence and index [R1], plus reverse sequence and index [R2]). Initial processing and filtering of raw sequences was done with USEARCH v8.0 (Edgar 2015). Forward and reverse paired reads were merged and renamed according to their respective sample site and location (upper or lower pool), using *fastq_mergepairs*. These single contiguous sequences were filtered and converted to FASTA files (*fastq_filter*) using default parameters. Expected errors were set to 0.4, as this is a better measure of sequence quality than quality scores alone (Edgar 2013). Both replicates from each site were combined into one contiguous FASTA file so as to increase sampling depth and improve coverage.

For downstream analyses with QIIME (Caporaso et al. 2012b), the sample ID, metadata, and sequential sequence number were first added to the header of each FASTA file by implementing a custom Perl script (*headerMod.pl*-Appendix 2). Additionally, a mapfile was constructed in Excel using abiotic site data and relevant metadata for each sample site. Both were validated in QIIME using *validate_mapping_file.py* and *validate_demultiplexed_fasta.py*, respectively.

Sequencing the 16S ribosomal RNA structural unit produces a significant number of exact sequence replicates and their removal greatly improves the efficiency of downstream analyses. Thus, each FASTA file was concatenated to create a sequence pool and de-replicated in USEARCH (*derep_fulllength*) with default parameters. Singletons (i.e., sequences found but once in the pooled data) are far more likely to result from sequencing error than are sequences found at least twice. Thus, removal of singletons represents a second quality-control step.

Sequences sorted by abundance using the command *sortbysize* and the option [-minsize 2]. This step also determined the size of each cluster.

Sequencing errors can also stem from 'chimeric sequences' (i.e., hybrid products between multiple parental sequences produced during the PCR reaction) that can be potentially interpreted downstream as novel organisms, thus inflating diversity estimates (Haas et al. 2011). OTUs (Operational Taxonomic Unit) were first generated in USEARCH using *cluster_otus*, with chimeric sequences subsequently identified and eliminated. An additional filtering step was performed in USEARCH using *uchime_ref* to match and eliminate additional chimeras by comparison to a reference FASTA database (rdp_gold.fa). The script *fasta_formatter* (FASTX toolkit; Hannon Lab 2011) was then used to create a single continuous line for proper downstream analysis.

A pipeline developed by the Brazilian Microbiome Project (BMP; Pylro et al. 2014) was employed to solve formatting issues that occur when Illumina data are analyzed with QIIME v1.7 that was originally developed to analyze sequences generated by the Life Sciences 454 Pyrosequencing platform (Roche 2007). Each platform utilizes a different adapter and indexing strategy, as well as a different sequencing technology, and the BMP pipeline corrects these incompatibilities by applying several custom Python scripts (Van Rossum and Drake 2001). The OTUs were renamed so as to be compatible with QIIME by applying the BMP script *bmp-otuName.pl*. Reads were mapped back to the original fasta file containing the assigned OTUs by implementing the USEARCH script *usearch_global*.

OTU taxonomy was assigned with the UCLUST method, as implemented in QIIME (*assign_taxonomy.py*). A representative set of sequences from the GREENGENES database (DeSantis et al. 2006), as well as a separate set of GREENGENES database reference sequences

were used to align sequences (via *align_seqs.py*) via the NAST alignment algorithm (Caporaso et al. 2010). Common gaps and non-conserved regions were removed (*filter_alignment.py*), and a reference tree generated (*make_phylogeny.py*). USEARCH output file (*map.uc*) was then converted into an OTU table file via a BMP script (*python map2qiime.py*), then into ".biom format" (McDonald et al. 2012) using QIIME (*make_otu_table.py*), with error-free and sampling depth validated (*biom summarize-table*).

Alpha and Beta Diversity analyses

Ecological and phylogenetic results were produced from a series of QIIME scripts that were amalgamated into a pipeline. These were: "*core_diversity_analyses.py* -i *otu_table.biom* -m *smap.txt* -c Impact, Slope, Watershed, InvFlowLength, WellDensity, Forest, Pasture, Urban, StrahlerSO -t *rep_set.tre* -e 29829 -o *core_output*". The –c parameter allowed each abiotic factor to be included and also initiated parametric and multivariate statistical analyses. This QIIME pipeline involves the following QIIME scripts: *alpha_rarefaction.py*, *beta_diversity_through_plots.py*, *summarize_taxa_through_plots.py*, plus the (non-workflow) scripts *make_distance_boxplots.py*, *compare_alpha_diversity.py*, and *group_significance.py*.

A measure of alpha diversity (or within community diversity) termed Shannon Entropy (Shannon and Weaver 1948) was computed so as to take into account the number of unique taxa (= richness) of the community and the evenness of its distribution. Using genera-level classifications, Shannon entropy (H') and evenness (J') indices were determined in QIIME© with a rarefaction sampling depth of 29,829 sequences per sample.

Multivariate statistical analysis was conducted to examine alpha diversity at each sample site. All abiotic variables were contrasted against species richness to generate rarefaction curves

based on the number of taxa present in each stream. Chao1, or species richness based upon the amount of rare classes (OTUs) found within a sample (Hortal et al. 2006), was calculated in QIIME and plotted as rarefaction curves that determine whether sampling depth or the number of sequences acquired for each sample was sufficient to capture and accurately characterize the community (Schloss and Handelsman 2005). Repeated subsampling (10 times) of 10 to 29,820 sequences, with steps of 2,981 sequences was conducted to generate rarefaction curves. Analyses were carried out with the default number of Monte-Carlo permutations (=999) and a standard p-value of 0.05.

The QIIME script *beta_diversity_through_plots.py* was used with weighted and unweighted UniFrac analyses to generate beta diversity, or between sample diversity. UniFrac measures phylogenetic distances among various taxa in a data set and can, through ordination and clustering, simultaneously compare several communities concomitant with their landscape scale geographic data (Lozupone and Knight 2005).

Microbial biological indicators

The key role and the quick metabolic processing provided by microbial organisms for chemicals in stream ecosystems makes them ideal to serve as bioindicators (Sims et al. 2013), defined as an organism whose presence, absence, or abundance can reflect a specific environmental condition (Foissner and Berger 1996). To identify potential bioindicator species, two heat maps, one for the four MICZ sites and one for the four PICZ sites, were generated from the most abundant 20 taxonomic classes found at each site (Figure 11). Classes that could serve as bioindicators were then examined at the genus level to specifically identify potential metabolites.

UniFrac neighbor-joining tree

To visualize how sample sites relate to one another according to their identified microbial communities, UniFrac distance matrices were clustered with a neighbor-joining algorithm (Saitou and Nei 1987). Weighted UniFrac phylogenetic distance matrices were generated during QIIME analysis. The UniFrac distance matrices were then clustered in T-REX, a neighbor-joining algorithm available online (Boc et al. 2012), generated with a Kimura 2-parameter substitution model (Kimura 1980) and validation by bootstrapping, and visualized with a 2-dimensional neighbor-joining tree (Figure 12).

Results

Spatial and local data; biofilm and DNA extractions

Rock Creek and Driver Creek, classified as MICZ sites, were most similar based on abiotic variables (Figure 1). Two PICZ-classified sites, Hogans Creek and East Fork Point Remove, were the next most closely related. The remaining four sites clustered together by type, with the two PICZ sites, Black Fork and Sunnyside Creek being slightly more similar to each other than the two remaining MICZ sites, SIS Hollow and Low Cedar. Although sample sites were standardized across drainages, local ecological characteristics varied slightly (Table 2). Water temperature increased as sampling progressed temporally, as did air temperature. Three sites recorded a higher percentage of canopy cover, while site elevation ranged from 353 m (Sis Hollow) to 135 m (Black Fork).

Biofilm averaged 153 mg/sample, with significantly greater amounts from lower sections of pools (average upper pool=131.1 mg; average lower pool=174.4 mg; one-way ANOVA $F_{(1,14)}=7.09$, P=0.0105; Table 3). DNA averaged 39.8 ng/µl, and did not differ by site or pool (average upper pool=35.8 ng/µl; average lower pool=43.8 ng/µl).

Initial processing of Illumina MiSeq sequencing data

Similar results were obtained among samples for the following: Percent sequences converted during merging of paired-end reads; percentage of exact overlaps (forward and reverse reads match exactly); and percentage of reads passing quality filtering threshold for maximum allowable errors (i.e., 0.4 = 99.999% probability of a correct nucleotide call) (Table 4). Means

and variances for merged sequence conversion, and numbers of filtered reads are also provided (Table 4).

De-replication condensed sequences from 682,688 to a unique set of 72,226 (a 90.4% reduction). Elimination of singletons further reduced the total to 33,483 (a 63.6% reduction). Removal of chimeric sequences during clustering eliminated an additional 3,740 (11.2%) with the remaining 29,743 clustered into 6,959 unique OTUs. Replicates within OTUs (22,784 sequences) were also discarded during clustering. A comparison of sequences against a reference database also eliminated an additional 50 chimeric sequences (0.7%). Alignment of sequences with the core set database (DeSantis et al. 2006) also excluded an additional 345 sequences, thus yielding a final total of 6,564 unique OTUs.

Shannon diversity statistics

The number of unique OTUs ranged from 1,048 (East Fork Point Remove) to 547 (Rock Creek) (Table 5). Shannon entropy (H') and evenness (J') were: Highest for East Fork Point Remove (H'=3.22, J'=0.46) and lowest for Black Fork (H'=2.470, J'=0.359) amongst PICZ streams; and highest for Cedar Creek (H'=2.986, J'=0.443) and lowest for Driver Creek (H'=1.88, J'=0.29) amongst MICZ streams. Although the mean values for identified OTUs, Shannon entropy, and Shannon evenness were greater for PICZ sites ($OTUs_{\mu}$=945; H'_{μ}=2.712; J'=0.398) versus MICZ sites ($OTUs_{\mu}$=723; H'_{μ}=2.467; J'=0.375), the differences were not statistically significant.

Taxonomic identifications

The pooled data yielded 141 taxonomic classes, with 54 represented across all samples. The major classes (>2% average abundance) were: Alphaproteobacteria, Betaproteobacteria, Planctomycetia, Oscillatoriophycideae, Synechococcophycideae, Cytophagia, Saprospirae, and Gammaproteobacteria. Of the 14 most abundant (>1%) classes, the most dominant was Alphaproteobacteria, averaging 18.9% across samples, with Betaproteobacteria averaging 8.4% (Figure 6).

Abundances of Oscillatoriophycideae and Synechococcophycideae (Cyanobacteria) were significantly greater at the PICZ sites when compared with MICZ sites (PICZ: t (3) =3.06, P=0.027; MICZ: t (3) = 2.37, P= 0.049). On the other hand, Verrucomicrobiae was significantly more abundant at MICZ sites (t (3) = 3.07, P=0.027).

Gloeobacter sp. (Cyanobacteria) was most abundant across all sites, averaging 26.86%. *Nitrosopumilus sp.* (Archaeon) was the second-most abundant, averaging 4.45%. Both Black Fork and East Fork Point Remove had second-most abundant genera that differed: *Microcystis* (5.54%) in the former instance, and *Zymomonas* (4.86%) in the latter.

Average relative abundance for the top 25 genera was 2.14% and ranged from 0.22% (*Rhodobacter*) to 26.86% (Table 6), with six different phyla represented. These were: Proteobacteria (40%), Bacteroidetes (28%), Cyanobacteria (12%), Plantomycetes (8%), Verrucomicrobia (4%), and Deferribacteres (4%).

Alpha diversity

A total of 46 phyla were represented in the study. Average number per sample was 36 (range: 32 to 39), with 24 found across all samples. Although this represents an excellent

diversity estimate for a substrate biofilm sample (Lyautey et al. 2005, Rundell et al. 2014), several phyla still dominated across all samples (encompassing 86% of samples). These were: Cyanobacteria (37.4%); Proteobacteria (31.7%); Bacteroidetes (7.6%); Planctomycetes (5.3%); Actinobacteria (4%). A total of 310 genera were found across all samples, with 116 (37%) at each, and 297 (95.8%) found in at least 4 or more.

When pooled by sample ID, several rarefaction curves approached asymptotes (Figure 7), suggesting sampling depth was sufficient to capture rare microbes. That being said, more diverse samples (i.e., Black Fork, Sunnyside Creek, and Sis Hollow) displayed a gradually elevating trajectory that suggested the potential for insufficient sampling depth (Figure 7). Samples pooled by impact differed between MICZ and PICZ sites, but not significantly (Figure 8).

Beta diversity

Results from a principal coordinate analysis (PCoA) clustered samples according to site type, either MICZ or PICZ (Figure 8). In the weighted UNIFRAC analysis, MICZ sites clustered along the top of the second PC axis, while PICZ sites clustered along the bottom, with this axis accounting for 18% of the variation within the data (Figure 9). In a weighted UNIFRAC PcoA plot based on Strahler stream order, samples fell in a linear array along the first axis, from highest stream order (Point Remove) at the bottom, to lowest stream order (Driver Creek) at the top. This axis accounted for 56% of the variation within the data (Figure 10).

Microbial biological indicators

Sites were not significantly different when compared across the top five identified taxonomic classes, and the bottom three identified classes also show remarkable similarities

across all sites (Figure 11). However, significant differences between site types exist for the sixth (Synechoccophycideae*)* and ninth (Oscillatoriophycideae) most abundant classes (PICZ: t (3) =3.06, P=0.027; MICZ: t (3) = 2.37, P= 0.049). Synechoccophycideae was represented by six genera (in descending abundance) *Arthronema, Acaryochloris, Leptolyngbya, Pseudanabaena, Paulinella, and Synechococus.* Oscillatoriophycideae was represented by seven genera (in descending abundance) *Microcystis, Chroococcus, Cyanobacterium, Chroococciddoipsis, Phoridium*, and *Planktothrix*. *Microcystis*, was substantially more abundant across PICZ sites and was found particularly high in Black Fork (5.54%). Both species are primary producers (phylum Cyanobacteria) and were more abundant in PICZ sites. The class Spartobacteria*,* a heterotrophic microbe (Herlemann et al. 2013) commonly associated with cyanobacteria and negatively associated with high salinity, was significantly more abundant at MICZ sites (M=0.02, SD=1.9E-04) versus PICZ sites (M=0.0048, SD=6.6E-06) t(3)=9.37, p = 0.01. The class Nostocophycideae, a filamentous cyanobacteria containing heterocysts (Ward et al. 1985), was significantly more abundant at MICZ sites (M=0.007, SD=1.15E-05) versus PICZ sites (M=0.0008, SD=5.1E-07) t(3)=9.28, p = 0.015. The class Actinobacteria, commonly reported as a bioindicator and bioremediation capable species (Lewis et al. 2012, Yergeau et al. 2012, Pascault et al. 2014), was significantly more abundant at PICZ sites (M=0.0186, SD=4.1E-04) versus MICZ sites (M=0.0177, SD=3.67E-05) t(3)=1.08, p = 0.039. Actinobacteria was represented by five genera across all sites (in descending abundance) *Rothia, Streptomyces, Nocardioides, Rhodococcus, and Catellatospora.*

UniFrac neighbor-joining tree

Rock Creek and Driver Creek are MICZ sites that were identified as most similar on abiotic variables in the Ward tree (Figure 1). They also shared the same branch with the other two MICZ sites, Sis Hollow and Low Cedar Creek in the neighbor-joining tree (Figure 12). Two PICZ-classified sites, Black Fork and East Fork Point Remove, were the next most closely related and were also more closely related to the four MICZ sites than the more distantly related PICZ sites, Hogans Creek and Sunnyside Creek.

Discussion

The links between land-use and stream environmental conditions provide insight into the manner by which anthropogenic activities impact stream conditions and potentially alter ecosystem function (Tong and Chen 2002, Gilliom et al. 2006). In this sense, the study area was primarily used for hay fields, cattle farming, or was undeveloped prior to industrial activity (http://www.nass.usda.gov). It has now experienced expansive growth of Unconventional Natural Gas (UNG) wells, and subsequent changes in land use that potentially impact the catchment zones near well sites (Entrekin et al. 2011). Subsequent research (as above) now examines the effects of this new industry, and the potential alterations it may have initiated in catchment zones and associated streams. The current study continues this exploration by utilizing next generation sequencing to characterize fresh water microbial biofilm communities within the Fayetteville Shale region.

Nationwide, freshwater streams have already experienced impacts that have altered and threatened ecosystem health, with over half (55%) in poor condition (Paulsen et al. 2008). The accelerated pursuit of shale resources in previously untapped regions has the potential to likewise exacerbate this decline (Gillen et al. 2012). However, baseline data is either insufficient or unavailable to determine if and where UNG activities have resulted in habitat alterations (Brittingham et al. 2014). One issue is that more conventional anthropogenic impacts on freshwater microbial communities must be compared to those found in streams potentially impacted by UNG activities (Brittingham et al. 2014). Once sufficient data has been generated across varied landscapes and conditions, then potential alterations can be properly quantified, identified, and related back to their most likely sources (Vidic et al. 2013).

Spatial data

Comparisons based solely on landscape scale data (Figure 1) depicted environmental relationships among sites without designated impacts being considered. This, in turn, permitted benchmark species composition metrics to be developed that potentially relate to specific landscape scale catchment variables. These data separated streams into four groups, each containing two sites. Although six landscape-scale variables were utilized, elevated associations among them resulted in a single variable primarily separating sites from one another (Table 1, Figure 1). For example, Rock Creek and Driver Creek reflected the highest percentage of forestation, whereas Hogans Creek and East Fork Point Remove have substantially larger watersheds.

Biofilm and DNA extractions

The significantly greater amounts of biofilm (Table 3) sampled from lower- versus upper-pool sites was not an unexpected result, particularly given the flashy nature these streams display (Johnson et al. 2015), as well as proximity of the upper pool to the upstream riffle. However, these differences did not significantly impact the amounts of DNA extracted.

During the initial optimization steps, substantial quantities of humic substances and polysaccharides, commonly referred to as extracellular polymeric substances (EPS) (Donlan 2002), were found in most samples, again a typical biofilm result (Vu et al. 2009). This aspect also suggested that cyanobacteria and microbial diversity would be elevated in these samples, an aspect subsequently confirmed in taxonomic and diversity analyses (below). Interestingly, samples with high biofilm levels but with unexpectedly lower levels of DNA occurred in three

streams with high levels of microbial alpha diversity (i.e., East Fork Point Remove, Black Fork, and Sis Hollow – Figure 7). Increased microbial diversity, and presence of rare taxa that can metabolize a wider range of substrates, are deemed ecological strategies that counter environmental stress (Pholchan et al. 2013). Thus, greater turnover within diverse biofilms would be expected under conditions of increased stress and increased metabolite availability.

East Fork Point Remove, a PICZ stream, is most likely to have experienced these conditions, given its high inverse flow length value and second-highest well density (Table 1). Black Fork, also a PICZ stream, could be similarly affected. However, Sis Hollow is not impacted with an inverse flow length and well density of zero, and therefore presents an unexpected deviation in this case. Stress and increased variability in metabolites may not be necessary to generate an increased turnover in community diversity as increased nutrient availability can also promote less dominant but rapidly growing species to bloom (Yooseph et al. 2010). Sis Hollow has the highest percentage of pasture among MICZ sites, and the second highest percentage of urbanization. Both are well-documented anthropogenic impacts known to increase freshwater eutrophication (Gilliom et al. 2006). This would also directly increase autotrophic species such as cyanobacteria, which in turn could explain high levels of polysaccharides within samples (Legendre and Rassoulzadegan 1995). Each of these streams is experiencing significantly higher turnover within their resident biofilms—a result which points to anthropogenic pollution in the form of diverse metabolites, stress inducing chemicals, and nutrient enrichment for the two PICZ streams, and nutrient enrichment likely from agriculture for the MICZ stream.

Shannon diversity statistics

Streams with higher numbers of OTUs did not consistently reflect higher H'-values, which in turn suggests the presence of several numerically dominant species in these populations, with many others in much fewer numbers. Microbes most commonly found in these freshwater environments are selected for (Zwart et al. 2002), and therefore evenness is depressed, particularly when compared with soil ecosystems. The streams with lowest evenness for each type [i.e. Driver Creek (MICZ) and Black Fork (PICZ); Table 5] are also most likely to have biological and chemical drivers pushing these results. Driver Creek is a headwater stream (stream order =<1) with the highest percentage of forestation in the study (=95.76%), and thus with an elevated leaf litter input. This promotes competition among bacteria and hyphomycetes (stream fungi) and reduces microbial diversity (Gulis et al. 2003). Black Fork has the highest percentage of pasture in the study (50%) and this in turn elevates available phosphates and nitrates in the stream, again promoting proliferation of a few dominant species (Lear and Lewis 2009). The two least diverse streams in an environmental sense (i.e., Rock Creek and Driver Creek; Figure 1) also had low values for OTUs, suggesting a potential reduction of available niches in these streams.

Taxonomic identifications

Microbial taxa identified in this study predominantly fell within the two most abundant classes of bacteria found within freshwater ecosystems, the Alphaproteobacteria and Betaproteobacteria (Fazi et al. 2005). However, the third most abundant in this study, the Planctomycetia, normally occurs at significantly lower abundances. It has distinctive morphological, metabolic, and genomic characteristics (Youssef and Elshahed 2014), with a

recognized capacity to act as an ammonia-oxidizer. In this sense, the health of an aquatic ecosystem is often reflected by the ratio of ammonia-oxidizing archaea (AOA) to ammonia-oxidizing bacteria (AOB), with an elevated abundance of AOA and thus a ratio-number > 1, reflecting ecosystem health (Sims et al. 2012, Sonthiphand et al. 2013). The presence of Planctomycetia (AOB) at higher than normal abundances for a typical freshwater system lowered the AOA/AOB ratio and may be an indication of deteriorating conditions with respect to ecosystem health and water quality. However, the second most abundant genus in this study was *Nitrosopumilus*, an AOA microbe and the only archaeon identified within the top 100 genera. The presence of this AOA genus when compared to all AOB genera present across sites maintains the AOA/AOB ratio across sites (i.e. lowest=>1.405 (EFPR)) such that streams are still deemed 'healthy.'

Alpha diversity

Freshwater streams are highly diverse microbial ecosystems (Lear et al. 2008), and this translates to their biofilms as well, despite limits on mobility (Lyautey et al. 2005). Stream biofilms herein also reflected this, with alpha diversity elevated across all sites. In the rarefaction plot of species richness (Chao1) versus sequences per sample (Figure 7), most curves (i.e., five of eight samples) approached asymptotes, suggesting that our sampling regime was sufficient to capture most rare microbes.

The greater diversity found in PICZ sites versus MICZ sites (Figure 8) suggests that several diversification factors may be at work in these streams. Most primary is the increased availability and diversity of metabolites (Gibbons et al. 2014), followed by elevated evenness levels (Wittebolle et al. 2009), and lower levels of pollution-induced stress (Girvan et al. 2005).

Given the limited spatial variation demonstrated by the microbial communities in this study (Lear et al. 2013), those streams with similar landscape catchments would be expected to have similar initial evenness values. This in turn suggests that diverse and abundant metabolites and low-level stress may be elevated factors in PICZ sites versus MICZ sites.

Beta diversity

The majority of landscape scale variables did not significantly alter phylogenetic distances or species-metrics among sites. Only two variables, Impact designation and Strahler stream order, yielded a consistent pattern in this regard. Figure 9 suggests that species abundance is impacted by proximity to UNG well activity. It also reflects variability amongst sites, with those identified as PICZ sites more similar to one another than MICZ sites. Because of this segregation along with each PICZ site changing their closest neighbor in the final tree as a result of increased species abundance in those sites with the highest potential impact, it is hypothesized that factors specifically relating to UNG activity have promoted the composition of its biofilm communities.

The linear pattern seen in Figure 10 suggests variation of beta diversity across sites is predominantly influenced by stream order at each sample site, and that taxon richness and diversity are promoted by greater numbers of tributaries feeding study streams. Typically, streams with higher orders (i.e., >5) and larger watershed areas (i.e., >71km^2) yield lower microbial diversity at their confluences (Besemer et al. 2013), whereas those with greater slopes and reduced watershed areas do not (Gillett et al. 2011). Most streams in the current study, despite being mid-level with respect to stream order, have comparably reduced watersheds, steeper slope, and are flashy with predictable dry periods (Johnson et al. 2015). These conditions

can generate surprisingly rich and distinctive microbial communities (Fazi et al. 2013, Meyer et al. 2007). Additionally, catchment areas in our study reflect elevated land-use (Entrekin et al. 2011), and in fact, a catchment area with elevated land-use will positively influence microbial diversity (Lear et al. 2009). The land-use/ flow regime conditions (above) offer anecdotal support for higher diversities at confluences of tributaries, especially with regard to streams within PICZ areas. Streams in the current study have unique landscape scale properties which may have served as mild stressors (i.e. intermittent dry periods and rapid influx of flood waters) leading to increased stress tolerance (Gasith et al. 1999), thus aiding in initial diversity and preservation of enhanced diversity following additional stress.

Although the remaining landscape variables differed among streams, they did not yield consistent patterns. This was somewhat surprising, considering that previous studies have identified catchment land-use as impacting bacterial richness (Winter et al. 2007, Lear et al. 2013), more so than spatially determined gradients such as elevation, latitude, or longitude (herein). Typically greater impacts were found in streams affiliated with greater pasture or grassland areas, which positively correlated with taxon richness. Given this, differences in slope, watershed, and especially land use would be expected to yield consistent patterns of microbial diversity. However, similar studies of benthic invertebrates and bacteria suggest land-use variables must be strongly pervasive before bacterial communities are demonstrably impacted (Lear et al. 2009).

Three sites showed significant environmental differences, but only at single variables. These were: diminished watershed area (Sis Hollow), and greater percentage of forestation (Rock Creek, Driver Creek). The latter also demonstrated a consistent relationship between abiotic and biotic metrics (Figures 1, 12). Both revealed diminished microbial populations that

associated with significantly higher percentage of forested land within their catchments (Table 1). As explanation, natural landscapes with few anthropogenic impacts limit the amount of reduced and dissolved organic matter (DOM) in constituent streams, where DOM serves as a critical substrate for bacterial communities (Kirchman et al. 1991, Williams et al. 2010). Any factor that limits or enhances a potent food source, especially for sessile biofilm, will also impact both community metabolism and composition (Docherty et al. 2006). Although percentage of forestation did not yield a consistent pattern for beta diversity across sites, where it was highest and therefore indicating reduced anthropogenic land-use, there was a correlative reduced diversity (Figure 7). The results from the current study strengthen the hypothesis that landscape scale variables must be significantly divergent to drive consistent and significant differences within microbial biofilm communities (Lear et al. 2009).

In this study, the variability among catchment areas was insufficient to significantly alter the composition of biofilm communities. However, primary drivers in this context are impact designation and stream order. Clearly, additional inputs from upstream tributaries could also promote impact-related conditions in higher-order streams, in this sense both factors could be coinciding for a more substantial influence.

Microbial biological indicators

Many microbial taxa are considered 'biomarker species,' and thus their presence or absence can be a useful criterion for classifying metabolic substrate within a given ecosystem (Foissner and Berger 1996). PICZ and MICZ sites differ significantly with regard to several constituent microbial taxa (Table 7 and Figure 11). Two cyanobacterial classes, Synechococcophycideae and Oscillatoriophycideae, were significantly more abundant at PICZ

sites, suggesting the potential for increased phosphorus and nitrogen inputs, and thus eutrophication, and is consistent with a recent algal biomass study in the study region (Austin et al. 2015). Typically phosphorus would be the dominant nutrient, but the elevated abundances of non-nitrogen fixing cyanobacteria such as *Microcystis* within PICZ catchments (Paerl et al. 2013) (Table 6), suggest the presence of elevated nitrogen, an aspect often associated with anthropogenic land use (Tong and Chen 2002), particularly at/ near UNG well sites (Mouser et al. 2012). Both cyanobacterial classes are effective primary producers capable of surviving in highly varied environments (Rothschild and Mancinelli 2001), and their dominance at PICZ sites suggests anthropogenic alterations in water quality.

Synechococcophycideae can utilize unique metabolic pathways, and often persist with marine sponges in highly acidic environments at volcanic seeps (Morrow et al. 2015). Oscillatoriophycideae is equally diverse and also serves as a bioindicator for organic pollutants (Tanimu et al. 2011). *Microcystis*, for example, utilizes polycyclic aromatic hydrocarbons for growth and metabolism, and these often stem from ongoing pollution. Impressively, *Microcystis* reflects this ability even in the absence of preferred nutrients such as phosphorus and nitrogen (Zhu et al. 2012). Its presence at high levels at PICZ sites, and specifically in Black Fork, may point to elevated polycyclic aromatic hydrocarbons in the water column. These in turn are fracking contaminants found in water sources at close proximity to well pads (Witter et al. 2008).

A Verrucomicrobia class, Spartobacteria, was significantly more abundant at MICZ sites. It is often positively associated with cyanobacteria, but negatively associated with higher salinity (Herlemann et al. 2013). Given the increased abundance of cyanobacterial microbes at PICZ sites, the elevated presence of Spartobacteria would also be expected. However, their significant

decrease could also indicate potentially elevated levels of brine at PICZ sites, in that brine-contaminated waters are also associated with UNG well activity (Myers 2012).

Finally, Actinobacteria (a well-known bioremediator and an additional biomarker class) was also significantly more abundant at PICZ sites, in particular at East Fork Point Remove. Each of the Actinobacterial genera listed in Table 7 is a bioremediator of hydrocarbons, heavy metals, halides, polycyclic aromatic compounds, or other common chemical contaminants (Byss et al. 2008, Chikere et al. 2009, Maldonado et al. 2011, Polti et al. 2011). Although the significant presence of Actinobacteria at PICZ sites could indicate higher levels of contamination due to UNG well activities, it may also be a response to naturally- or agriculturally-introduced contaminants (Villeneuve et al. 2011), in that Actinobacteria are opportunistic and heterotrophic (Berg et al. 2009). Regardless, differences among sites represent important distinctions, as microbial community structure is shaped by environmental factors (Gibbons et al. 2014) that produce a level of 'habitat filtering' (Pontarp et al. 2012). The bioindicator species identified at sites in this study suggest that niche partitioning, or selective competition for diverse resources (Macalady et al. 2008) is also occurring in response to the alteration of available resources.

Conclusions

The current study sought to utilize a reliable method in order to examine the ecosystem health of fresh water streams, namely identifying and characterizing microbial biofilm communities. Based on published research, this is the first study to examine associations of UNG activity occurring within the study area watersheds on biofilm communities and also represents one of a relatively few number of studies examining biofilm communities in fresh waters streams in general. The null hypotheses were rejected by the results, indicating significant differences

were found in resident biofilms, and in most cases these differences seem to be driven by proximity to UNG well activities. Additionally, biofilm communities within two streams were shaped by significantly higher forestation and resulting lower anthropogenic influences. Taken together, the results suggest biofilm communities can serve as reliable indicators for landscape scale perturbations or pollution based impacts within fresh water systems, and increased UNG activities within catchments are either directly or indirectly promoting alterations of these water filtering communities.

Natural conditions involved in this study prevent the possibility of removing unwanted variation, particularly with regard to landscape scale metrics. Additionally, the relatively small sample size involved reduced statistical resolving power. Expanding the available sample sites and increasing the overall quantity of samples examined from these sites would generate higher degrees of confidence in the statistical veracity of results. Future studies will benefit from existing data and could lead to increased appreciation of the critical role that biofilms serve within these important ecosystems. Establishing a biofilm "fingerprint" for healthy streams could serve as an early warning system for the occurrence of significant alterations, such as those from UNG activity or other anthropogenic influences that could threaten the food chain and hence lead to catastrophic collapse.

Literature Cited

Adamski JC, JC Petersen, DA Freiwald, JV Davis. 1995. Environmental and hydrologic setting of the Ozark Plateaus study unit, Arkansas, Kansas, Missouri, and Oklahoma: National Water-Quality Assessment Program.

Allan JD, LL Yuan, P Black, T Stockton, PE Davies, RH Magierowski, and SM Read. 2012. Investigating the relationships between environmental stressors and stream condition using Bayesian belief networks. Freshwater Biology 57(s1): 58-73.

Armitage KB. 1958. Ecology of the riffle insects of the Firehole River, Wyoming. Ecology: 572-580.

Austin BJ, N Hardgrave, E Inlander, C Gallipeau, S Entrekin, MA Evans-White. 2015. Stream primary producers relate positively to watershed natural gas measures in north-central Arkansas streams. Science of the Total Environment 529: 54-64.

Bai B, M Elgmati, H Zhang, M Wei. 2013. Rock characterization of Fayetteville shale gas plays. Fuel 105: 645-652.

Ball GH. 1922. Variation in freshwater mussels. Ecology, 3(2): 93-121.

Battin TJ, LA Kaplan, JD Newbold, and CM Hansen. 2003. Contributions of microbial biofilms to ecosystem processes in stream mesocosms. Nature 426(6965): 439-442.

Berg KA, C Lyra, K Sivonen, L Paulin, S Suomalainen, P Tuomi, J Rapala. 2009. High diversity of cultivable heterotrophic bacteria in association with cyanobacterial water blooms. The ISME journal 3: 314-325.

Bernot MJ, JL Tank, TV Royer, MB David. 2006. Nutrient uptake in streams draining agricultural catchments of the midwestern United States. Freshwater Biology 51: 499-509.

Boc A, Diallo, B Alpha, and V Makarenkov. 2012. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Research 40(W1): W573-W579.

Bokulich NA. 2012. Next-generation approaches to the microbial ecology of food fermentations. Biochemistry and Molecular Biology Reports 45(7): 377-389.

Besemer K, G Singer, C Quince, E Bertuzzo, W Sloan, TJ Battin. 2013. Headwaters are critical reservoirs of microbial diversity for fluvial networks. Proceedings of the Royal Society of London B: Biological Sciences 280: 20131760.

Brittingham MC, KO Maloney, AM Farag, DD Harper, ZH Bowen. 2014. Ecological risks of shale oil and gas development to wildlife, aquatic resources and their habitats. Environmental science & technology 48: 11034-11047.

Byss M, D Elhottová, J Tříska, P Baldrian. 2008. Fungal bioremediation of the creosote-contaminated soil: Influence of Pleurotus ostreatus and Irpex lacteus on polycyclic aromatic hydrocarbons removal and soil microbial community composition in the laboratory-scale study. Chemosphere 73: 1518-1523.

Cairns Jr J, PV McCormick, B Niederlehner. 1993. A proposed framework for developing indicators of ecosystem health. Hydrobiologia 263: 1-44.

Caporaso JG, CL Lauber, WA Walters, D Berg-Lyons, J Huntley, N Fierer, SM Owens, J Betley, L Fraser, M Bauer, N Gormley, JA Gilbert, G Smith, and R Knight. 2012b. Ultra-high-throughput microbial community analysis on the illumina hiseq and miseq platforms. ISME J 6:1621-1624.

Caporaso JG, K Bittinger, FD Bushman, TZ DeSantis, GL Andersen, R Knight. 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26: 266-267.

Chikere C, G Okpokwasili, B Chikere. 2009. Bacterial diversity in a tropical crude oil-polluted soil undergoing bioremediation. African Journal of Biotechnology 8.

Chiu CM, FM Lin, TH Chang, WC Huang, C Liang, T Yang, and HD Huang. 2014. Clinical detection of human probiotics and human pathogenic bacteria by using a novel high-throughput platform based on next generation sequencing. Journal of Clinical Bioinformatics 4(1): 1-13.

Clapcott JE, KJ Collier, RG Death, EO Goodwin, JS Harding, D Kelly, and RG Young. 2012. Quantifying relationships between land-use gradients and structural and functional indicators of stream ecological integrity. Freshwater Biology 57(1): 74-90.

Cohen RR. 2006. Use of microbes for cost reduction of metal removal from metals and mining industry waste streams. Journal of Cleaner Production 14(12): 1146-1157.

Colborn T, C Kwiatkowski, K Schultz, and M Bachran. 2011. Natural gas operations from a public health perspective. Human and Ecological Risk Assessment: An International Journal, 17(5), 1039-1056.

Coles JF, G McMahon, AH Bell, LR Brown, FA Fitzpatrick, BS Eikenberry, and WP Stack. 2012. Effects of urban development on stream ecosystems in nine metropolitan study areas across the United States. US Geological Survey Circular 1373.

Cooper SD, S Diehl, K Kratz, and O Sarnelle. 1998. Implications of scale for patterns and processes in stream ecology. Australian Journal of Ecology 23(1): 27-40.

DeSantis TZ, P Hugenholtz, N Larsen, M Rojas, EL Brodie, K Keller, T Huber, D Dalevi, P Hu, and GL Andersen. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Applied Environmental Microbiology 72:5069-72.

Docherty KM, KC Young, PA Maurice, SD Bridgham. 2006. Dissolved organic matter concentration and quality influences upon structure and function of freshwater microbial communities. Microbial ecology 52: 378-388.

Donlan RM. 2002. Biofilms: microbial life on surfaces. Emerging Infectious Disease: 8.

Duncan FC, and Hoppe-Seyler. 1893. Beitrige note the respiration of fish. Zeitschr. Physiological Chemistry I7: 375-394.

Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 10: 996-998.

Edgar RC, Flyvbjerg H. 2015. Error filtering, pair assembly, and error correction for next-generation sequencing reads. Bioinformatics: btv401

Entrekin S, M Evans-White, B Johnson, and E Hagenbuch. 2011. Rapid expansion of natural gas development poses a threat to surface waters. Frontiers in Ecology and the Environment 9(9), 503-511.

Fazi S, S Amalfitano, J Pernthaler, and A Puddu. 2005. Bacterial communities associated with benthic organic matter in headwater stream microhabitats. Environmental Microbiology, *7*(10), 1633-1640.

Fazi S, E Vázquez, EO Casamayor, S Amalfitano, A Butturini. 2013. Stream hydrological fragmentation drives bacterioplankton community composition.

FASTX Toolkit *http://hannonlab.cshl.edu/fastx_toolkit/index.html* by Hannon Lab. 2011.

Fierer N, JA Jackson, R Vilgalys, RB Jackson. 2005. Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. Applied and environmental microbiology 71: 4117-4120.

Foissner W, H Berger. 1996. A user-friendly guide to the ciliates (Protozoa, Ciliophora) commonly used by hydrobiologists as bioindicators in rivers, lakes, and waste waters, with notes on their ecology. Freshwater Biology 35: 375-482.

Gadd GM. 2010. Metals, minerals and microbes: geomicrobiology and bioremediation. Microbiology 156: 609-643.

Gagnon GA, RM Slawson. 1999. An efficient biofilm removal method for bacterial cells exposed to drinking water. Journal of Microbiological Methods 34: 203-214.

Gardes M, TD Bruns. 1993. ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. Molecular ecology 2: 113-118.

Gasith A, VH Resh. 1999. Streams in Mediterranean climate regions: abiotic influences and biotic responses to predictable seasonal events. Annual review of ecology and systematics: 51-81.

Geesey GG, DC White. 1990. Determination of bacterial growth and activity at solid-liquid interfaces. Annual Reviews in Microbiology 44: 579-602

Geesey GG, R Mutch, JT Costerton, and RB Green. 1978. Sessile bacteria: an important component of the microbial population in small mountain streams. Limnology and Oceanography 23(6): 1214-1223.

Gibbons SM, E Jones, A Bearquiver, F Blackwolf, W Roundstone, N Scott, J Hooker, R Madsen, ML Coleman, JA Gilbert. 2014. Human and environmental impacts on river sediment microbial communities.

Gillen JL, E Kiviat. 2012. Environmental Reviews and Case Studies: Hydraulic Fracturing Threats to Species with Restricted Geographic Ranges in the Eastern United States. Environmental Practice 14: 320-331.

Gillett ND, Y Pan, KM Manoylov, R Stancheva, CL Weilhoefer. 2011. The potential indicator value of rare taxa richness in diatom-based stream bioassessment1. Journal of phycology 47: 471-482.

Gilliom RJ, JE Barbash, CG Crawford, PA Hamilton, JD Martin, N Nakagaki, LH Nowell, JC Scott, PE Stackelberg, GP Thelin. 2006. Pesticides in the nation's streams and ground water, 1992–2001. US Geological Survey Circular 1291: 172.

Girvan M, C Campbell, K Killham, J Prosser, L Glover. 2005. Bacterial diversity promotes community stability and functional resilience after perturbation. Environmental microbiology 7: 301-313.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. Molecular ecology resources 11: 759-769.

Guo X, X Xia, R Tang, J Zhou, H Zhao, K Wang. 2008. Development of a real-time PCR method for Firmicutes and Bacteroidetes in faeces and its application to quantify intestinal population of obese and lean pigs. Letters in applied microbiology 47: 367-373.

Gulis V, K Suberkropp. 2003. Interactions between stream fungi and bacteria associated with decomposing leaf litter at different levels of nutrient availability. Aquatic Microbial Ecology 30: 149-157.

Gutsell, JS. 1929. Influence of certain water conditions, especially dissolved gasses, on trout. Ecology 10(1): 77-96.

Haack, TK, and GA McFeters. 1982. Nutritional relationships among microorganisms in an epilithic biofilm community. Microbial Ecology 8(2): 115-126.

Haas BJ, D Gevers, AM Earl, M Feldgarden, DV Ward, G Giannoukos, D Ciulla, D Tabbaa, SK Highlander, E Sodergren. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome research 21: 494-504.

Henrici, AT. 1933. Studies of freshwater bacteria. 1. A direct microscopic technique. Journal of Bacteriology. 25: 277-286.

Herlemann DP, D Lundin, M Labrenz, K Jürgens, Z Zheng, H Aspeborg, AF Andersson. 2013. Metagenomic de novo assembly of an aquatic representative of the *verrucomicrobial* class *Spartobacteria*. MBio 4: e00569-00512.

Hortal J, PA Borges, C Gaspar. 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. Journal of Animal Ecology 75: 274-287.

Illumina, Inc. 2010. De Novo Assembly Using Illumina Reads. Technical Note: Sequencing available at: http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf, last accessed 28 Feb 2011.

Johnson E, BJ Austin, E Inlander, C Gallipeau, MA Evans-White, S Entrekin. 2015. Stream macroinvertebrate communities across a gradient of natural gas development in the Fayetteville Shale. Science of the Total Environment 530: 323-332.

Karl DM. 1986. Determination of in situ microbial biomass, viability, metabolism, and growth. Bacteria in nature 2: 85-176.

Karr JR. 1981. Assessment of biotic integrity using fish communities. Fisheries 6: 21-27.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111-120.

Kirchman DL, Y Suzuki, C Garside, HW Ducklow. 1991. High turnover rates of dissolved organic carbon during a spring phytoplankton bloom. Nature 352: 612-614.

Klindworth A, E Pruesse, T Schweer, J Peplies, C Quast, M Horn, FO Glöckner. 2012. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research: gks808.

Lange T, M Sauter, M Heitfeld, K Schetelig, K Brosig, W Jahnke, A Kissinger, R Helmig, A Ebigbo, H Class. 2013. Hydraulic fracturing in unconventional gas reservoirs: risks in the geological system part 1. Environmental Earth Sciences 70: 3839-3853.

Lear G, V Washington, M Neale, B Case, H Buckley, and G Lewis. 2013. The biogeography of stream bacteria. Global Ecology and Biogeography 22: 544-554.

Lear G, I Boothroyd, S Turner, K Roberts, G Lewis. 2009. A comparison of bacteria and benthic invertebrates as indicators of ecological health in streams. Freshwater Biology 54: 1532-1543.

Lear G, G Lewis. 2009. Impact of catchment land use on bacterial communities within stream biofilms. Ecological Indicators 9: 848-855.

Lear G, MJ Anderson, JP Smith, K Boxen, GD Lewis. 2008. Spatial and temporal heterogeneity of the bacterial communities in stream epilithic biofilms. FEMS Microbiology ecology 65: 463-473.

Legendre L, F Rassoulzadegan. 1995. Plankton and nutrient dynamics in marine waters. Ophelia 41: 153-172.

Lewis J, N Morley, M Ahmad, GL Challis, R Wright, R Bicker, D Morritt. 2012. Structural changes in freshwater fish and chironomids exposed to bacterial exotoxins. Ecotoxicology and Environmental Safety 80: 37-44.

Lozupone C, R Knight. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71: 8228-8235.

Lyautey E, B Lacoste, L Ten-Hage, J-L Rols, F Garabetian. 2005. Analysis of bacterial diversity in river biofilms using 16S rDNA PCR-DGGE: methodological settings and fingerprints interpretation. Water Research 39: 380-388.

Macalady JL, S Dattagupta, I Schaperdoth, DS Jones, GK Druschel, D Eastman. 2008. Niche differentiation among sulfur-oxidizing bacterial populations in cave waters. The ISME Journal 2: 590-601.

Maldonado J, A Solé, Z Puyen, I Esteve. 2011. Selection of bioindicators to detect lead pollution in Ebro delta microbial mats, using high-resolution microscopic techniques. Aquatic Toxicology 104: 135-144.

McDonald D, JC Clemente, J Kuczynski, JR Rideout, J Stombaugh, D Wendel, A Wilke, S Huse, J Hufnagle, F Meyer. 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience 1: 7.

Morrow KM, DG Bourne, C Humphrey, ES Botté, P Laffy, J Zaneveld, S Uthicke, KE Fabricius, NS Webster. 2015. Natural volcanic CO2 seeps reveal future trajectories for host–microbial associations in corals and sponges. The ISME Journal 9: 894-908.

Moslemi JM, SB Snider, K MacNeill, JF Gilliam, and AS Flecker. 2012. Impacts of an invasive snail (*Tarebia granifera*) on nutrient cycling in tropical streams: the role of riparian deforestation in Trinidad, West Indies. PloS One 7(6), e38806.

Mouser P, M Ansari, A Hartsock, S Lui, J Lenhart. 2012. Microbial community shifts due to hydrofracking: Observations from field-scale observations and laboratory-scale incubations. Page 1196. AGU Fall Meeting Abstracts.

Meyer JL, DL Strayer, JB Wallace, SL Eggert, GS Helfman, NE Leonard. 2007. The contribution of headwater streams to biodiversity in river networks: Wiley Online Library.

Myers T. 2012. Potential contaminant pathways from hydraulically fractured shale to aquifers. Groundwater 50: 872-882.

Pace NR, DA Stahl, DJ Lane, and GJ Olsen. 1986. The analysis of natural microbial populations by ribosomal RNA sequences. Advances in Microbial Ecology 1-55. Springer US.

Paerl HW, and TG Otten. 2013. Harmful cyanobacterial blooms: causes, consequences, and controls. Microbial Ecology 65: 995-1010.

Pandey PK, and M Soupir. 2012. Assessing the impacts of weather pattern on in-stream *Escherichia coli* concentrations. Modeling In-Stream Escherichia coli Concentrations 246.

Pascault N, S Roux, J Artigas, S Pesce, J Leloup, RD Tadonleke, D Debroas, A Bouchez, and J-F Humbert. 2014. A high-throughput sequencing ecotoxicology study of freshwater bacterial communities and their responses to tebuconazole. FEMS microbiology ecology 90: 563-574.

Paulsen SG, A Mayio, DV Peck, JL Stoddard, E Tarquinio, SM Holdsworth, JV Sickle, LL Yuan, CP Hawkins, AT Herlihy. 2008. Condition of stream ecosystems in the US: an overview of the first national assessment. Journal of the north american Benthological society 27: 812-821.

Pham VD, LL Hnatow, S Zhang, RD Fallon, SC Jackson, JF Tomb, and SJ Keeler. 2009. Characterizing microbial diversity in production water from an Alaskan mesothermic petroleum reservoir with two independent molecular methods. Environmental Microbiology 11(1): 176-187.

Pholchan MK, JdC Baptista, RJ Davenport, WT Sloan, TP Curtis. 2013. Microbial community assembly, theory and rare functions. Frontiers in microbiology 4.

Poff NL. 1997. Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. Journal of the North American Benthological society, 16(2): 391-409.

Polti MA, MC Atjián, MJ Amoroso, CM Abate. 2011. Soil chromium bioremediation: synergic activity of actinobacteria and plants. International Biodeterioration & Biodegradation 65: 1175-1181.

Pontarp M, B Canbäck, A Tunlid, P Lundberg. 2012. Phylogenetic analysis suggests that habitat filtering is structuring marine bacterial communities across the globe. Microbial ecology 64: 8-17.

Poretsky RS, S Gifford, J Rinta-Kanto, M Vila-Costa, MA Moran. 2009. Analyzing gene expression from marine microbial communities using environmental transcriptomics. Journal of visualized experiments: JoVE: (24).

Proia L, F Cassió, C Pascoal, A Tlili, and AM Romaní. 2012. The use of attached microbial communities to assess ecological risks of pollutants in river ecosystems: the role of heterotrophs. Emerging and Priority Pollutants in Rivers 55-83. Springer Berlin Heidelberg.

Pylro VS, Roesch LFW, Morais DK, Clark IM, Hirsch PR, Tótola MR. 2014. Data analysis for 16S microbial profiling from different benchtop sequencing platforms. Journal of Microbiological Methods 107: 30-37.

Ram RJ, NC VerBerkmoes, MP Thelen, GW Tyson, BJ Baker, RC Blake, M Shah, RL Hettich, JF Banfield. 2005. Community proteomics of a natural microbial biofilm. Science 308: 1915-1920.

Roche. 2007. Genome sequencer FLX data analysis software manual. Roche Applied Science, Mannheim, Germany.

Rothschild LJ, and RL Mancinelli. 2001. Life in extreme environments. Nature 409: 1092-1101.

Rundell EA, LM Banta, DV Ward, CD Watts, B Birren, DJ Esteban. 2014. 16S rRNA Gene Survey of Microbial Communities in Winogradsky Columns. PLoS One 9: e104134.

Saitou N, M Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4: 406-425.

Schloss PD, J Handelsman. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and environmental microbiology 71: 1501-1506.

Schütte UM, Z Abdo, SJ Bent, C Shyu, CJ Williams, JD Pierson, and LJ Forney. 2008. Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. Applied Microbiology and Biotechnology 80(3): 365-380.

Shannon CE, and W Weaver. 1948. A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423 and 623–656.

Sims A, Y Zhang, S Gajaraj, PB Brown, Z Hu. 2013. Toward the development of microbial indicators for wetland assessment. Water Research 47: 1711-1725.

Sims A, J Horton, S Gajaraj, S McIntosh, RJ Miles, R Mueller, R Reed, Z Hu. 2012. Temporal and spatial distributions of ammonia-oxidizing archaea and bacteria and their ratio as an indicator of oligotrophic conditions in natural wetlands. Water Research 46: 4121-4129.

Sonthiphand P, E Cejudo, SL Schiff, JD Neufeld JD. 2013. Wastewater effluent impacts ammonia-oxidizing prokaryotes of the Grand River, Canada. Applied and environmental microbiology 79: 7454-7465.

Springer, L. 2011. Waterproofing the new fracking regulation: the necessity of defining riparian rights in Louisiana's water law. Louisiana Law Review 72, 225.

Stock MS, and AK Ward. 1989. Establishment of a bedrock epilithic community in a small stream: microbial (algal and bacterial) metabolism and physical structure. Canadian Journal of Fisheries and Aquatic Sciences 46(11): 1874-1883.

Schwarzenbach RP, BI Escher, K Fenner, TB Hofstetter, CA Johnson, U Von Gunten, B Wehrli. 2006. The challenge of micropollutants in aquatic systems. Science 313: 1072-1077.

Tanimu Y, S Bako, J Adakole, J Tanimu. 2011. Phytoplankton as bioindicators of Water quality in saminaka reservoir, Northern Nigeria. International Symposium on Environmental Science and Technology, Dongguan, Guangdong Province, China.

Tong ST, and W Chen. 2002. Modeling the relationship between land use and surface water quality. Journal of Environmental Management 66(4): 377-393.

Tringe SG, C Von Mering, A Kobayashi, AA Salamov, K Chen, HW Chang, M Podar, JM Short, EJ Mathur, JC Detter. 2005. Comparative metagenomics of microbial communities. Science 308: 554-557.

Van Rossum G, and FL Drake (eds), Python Reference Manual, PythonLabs, Virginia, USA, 2001. Available at http://www.python.org.

Vidic RD, SL Brantley, JM Vandenbossche, D Yoxtheimer, and JD Abad. 2013. Impact of shale gas development on regional water quality. Science 340(6134).

Villeneuve A, S Larroudé, and JF Humbert. 2011. Herbicide contamination of freshwater ecosystems: impact on microbial communities. Pesticides-Formulations, Effects, Fate 285-312.

Vu B, M Chen, RJ Crawford, EP Ivanova. 2009. Bacterial extracellular polysaccharides involved in biofilm formation. Molecules 14: 2535-2554.

Ward JH, Jr. 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association. 58: 236–244.

Ward AK, CN Dahm, KW Cummins. 1985. *Nostoc* (Cyanophyta) productivity in Oregon stream ecosystems: Invertebrate influences and differences between morphological types. Journal of phycology 21: 223-227.

Weisburg WG, SM Barns, DA Pelletier, DJ Lane. 1991. 16S ribosomal DNA amplification for phylogenetic study. Journal of Bacteriology 173 (2): 697–703.

Wiggins GB, Mackay RJ. 1978. Some relationships between systematics and trophic ecology in nearctic aquatic insects, with special reference to Trichoptera. Ecology: 1211-1220.

Williams CJ, Y Yamashita, HF Wilson, R Jaffé, MA Xenopoulos. 2010. Unraveling the role of land use and microbial activity in shaping dissolved organic matter characteristics in stream ecosystems. Limnology and Oceanography 55: 1159-1171.

Winter C, T Hein, G Kavka, RL Mach, AH Farnleitner. 2007. Longitudinal changes in the bacterial community composition of the Danube River: a whole-river approach. Applied and environmental microbiology 73: 421-431.

Winterstein H. I908. Contributions to the knowledge of fish respiration. Arch for the whole physiology of humans and animals (Pfluiger's archive) I25: 73-98.

Witter R, K Stinson, H Sackett, S Putter, G Kinney, D Teitelbaum, L Newman. 2008. Potential exposure-related human health effects of oil and gas development: A white paper. Colorado School of public health.

Wittebolle L, M Marzorati, L Clement, A Balloi, D Daffonchio, K Heylen, P De Vos, W Verstraete, N Boon. 2009. Initial community evenness favours functionality under selective stress. Nature 458: 623-626.

Xu P, B Yu, FL Li, XF Cai, CQ Ma. 2006. Microbial degradation of sulfur, nitrogen and oxygen heterocycles. Trends in microbiology 14: 398-405.

Yergeau E, JR Lawrence, S Sanschagrin, MJ Waiser, DR Korber, CW Greer. 2012. Next-generation sequencing of microbial communities in the Athabasca River and its

tributaries in relation to oil sands mining activities. Applied and environmental microbiology 78: 7626-7637.

Yooseph S, KH Nealson, DB Rusch, JP McCrow, CL Dupont, M Kim, J Johnson, R Montgomery, S Ferriera, K Beeson. 2010. Genomic and functional adaptation in surface ocean planktonic prokaryotes. Nature 468: 60-66.

Youssef NH, MS Elshahed. 2014. The Phylum Planctomycetes. The Prokaryotes, Springer: 759-810.

Zarraonaindia I, DP Smith, and JA Gilbert. 2013. Beyond the genome: community-level analysis of the microbial world. Biology and Philosophy 28(2): 261-282.

Zhu X, H Kong, Y Gao, M Wu, F Kong. 2012. Low concentrations of polycyclic aromatic hydrocarbons promote the growth of Microcystis aeruginosa. Journal of hazardous materials 237: 371-375.

Zwart G, BC Crump, MP Kamst-van Agterveld, F Hagen, S-K Han. 2002. Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. Aquatic Microbial Ecology 28.

Table 1: Landscape (GIS) data for each study site in the Fayetteville Shale region of north-central Arkansas, as acquired in 2013 prior to sampling. Stream = sampling location; Forest = % forested land within catchment area; Pasture = % pasture within catchment area; Urban = % urbanization within catchment area; Slope = % stream gradient; Strahler = Stream Order; Watershed area = in km^2; Well density = wells/ km^2; IFL = Inverse flow length; Impact = 0 (MICZ) or 1 (PICZ).

Stream	Forest (%)	Pasture (%)	Urban (%)	Slope (%)	Strahler (stream order)	Watershed area (km^2)	Well Density (no./ km^2)	IFL ($\sum 1/k$)	Impact
Rock Creek	94.49	4.25	1.02	6.3	2	16.11	0.124	0.177	0
Driver Creek	95.76	2.27	1.02	8.59	1	12.28	0	0	0
Low Cedar Creek	89.33	9	0.77	17	5	27.18	0.04	0	0
Sis Hollow	80.94	14	0.89	19	3	6.67	0	0	0
East Fork Point Remove	64	24	2	8	6	68.56	2.32	2.345	1
Sunnyside Creek	49	40	1	6	4	14.41	3.64	0.305	1
Hogans Creek	73	23	3	8	5	55.28	1.773	1.7	1
Black Fork	39	50	2	6	5	32.16	0.69	1.299	1

Table 2: Local data acquired at time of sampling for each site located in the Fayetteville Shale region of north-central Arkansas, with Elevation derived from ArcGIS data. Stream = sampling location; Air Temp = ambient air temperature; Water Temp = temperature of water at sample location; CC = % canopy coverage at site; Time = sampling time; Elevation = elevation above sea level at sampling site; Substrate Depth = depth below water surface of acquired biofilm substrate; SS = length of substrate; Substrate type = composition of substrate (sandstone and siltstone differ only in grain size).

Stream	Air Temp (°C)	Water Temp (°C)	CC (%)	Time (24h)	Elevation (m)	Substrate Depth (cm)	SS (cm)	Substrate type
Rock Creek	28	19.8	64	8:57	256	29	9	Siltstone
Driver Creek	28	21.2	57	9:59	231	27	8.5	Siltstone
Low Cedar Creek	29	23.0	51	11:15	264	30	10	Sandstone
Sis Hollow	29	22.6	67	12:35	353	23	8.1	Siltstone
East Fork Point Remove	30	22.4	78	2:10	218	33	9.1	Sandstone
Sunnyside Creek	30	22.8	68	3:01	182	43	8.6	Sandstone
Hogans Creek	30	23.6	63	4:13	158	41	9.5	Sandstone
Black Fork	31	21.4	74	5:11	135	31	9.8	Siltstone

Table 3: Biofilm and DNA gathered at study sites located in the Fayetteville Shale region of north-central Arkansas. Stream = sampling location; Sample ID = Site and pool location identification; Biofilm extracted = quantity (mg) extracted from each Nasco Whirl-Pak Speci-Sponge™; DNA extracted = quantity (ng/µl) extracted from each biofilm sample.

Stream	Sample ID	Biofilm extracted (mg)	DNA extracted (ng/µl)
Rock Creek	RC^u	78	5.5
	RC^l	129	14.7
Driver Creek	DC^u	92	6.5
	DC^l	199	22.6
Low Cedar Creek	LC^u	108	8.9
	LC^l	166	45.0
Sis Hollow	SH^u	140	41.6
	SH^l	143	28.6
East Fork Point Remove	PR^u	155	49.6
	PR^l	216	36.6
Sunnyside Creek	SS^u	146	61.8
	SS^l	156	82.8
Hogans Creek	HC^u	156	35.6
	HC^l	199	63.0
Black Fork	BF^u	174	77.2
	BF^l	187	57.0

Table 4: Values for raw sequences, merged reads, filtered reads by stream sample, as derived from study sites located in the Fayetteville Shale region of north-central Arkansas. Sample ID = abbreviation assigned to each sample based on site and location of sample from its respective pool; Total raw reads (#) = those sequenced from each sample; Paired-end merging results (%) = reads successfully merged into a contiguous sequence; Paired-end merging (#) = Those successfully merged; Quality Score Filtering (%) = Those passing quality filtering; Quality Score Filtering Discards = Those below <0.001% chance of miscalled nucleotide; Quality Score Filtering (#) reads passed = final number of sequences passed to downstream analysis.

	Total raw reads	Paired-end merging		Quality Score Filtering		
Sample ID	(#)	(%)	(#)	(%)	Discards	(#) Reads passed
RC^u	11676	99.4	11608	89.3	1238	10370
RC^l	22641	99.6	22547	88.9	2513	20034
DC^u	11363	99.5	11309	90.7	1049	10260
DC^l	38462	99.2	38152	91.3	3337	34815
LC^u	23230	99.5	23113	89.9	2571	20542
LC^l	30183	99.4	30007	88.0	3605	26402
SH^u	72246	99.7	72009	91.0	6453	65556
SH^l	48337	99.7	48171	90.3	4687	43484
PR^u	31482	99.6	31353	89.0	3460	27893
PR^l	53447	99.6	53232	88.6	6058	47174
SS^u	56612	99.8	56478	90.0	5634	50844
SS^l	63502	99.8	63371	89.3	6780	56591
HC^u	26650	99.7	26558	89.2	2869	23689
HC^l	71207	99.7	70987	89.2	7699	63288
BF^u	109950	99.7	109671	91.1	9781	99890
BF^l	90926	99.7	90671	90.2	8918	82008

Table 5: Values for Shannon Entropy, total identified species, and Shannon Evenness at each site located in the Fayetteville Shale region of north-central Arkansas. Stream = sampling location; Shannon Entropy (H') = Value based on OTUs at each site; Total Identified OTUs = Number at each site; Shannon Evenness (J') = Values range from (0) = total dominance to (1) = total evenness; Impact Factor = MICZ (0) or PICZ (1).

Stream	Shannon Entropy (H')	Total Identified OTUs	Shannon Evenness (J')	Impact Factor
Rock Creek	2.372	547	0.376	0
Driver Creek	1.881	639	0.291	0
Cedar Creek	2.986	843	0.443	0
Sis Hollow	2.629	861	0.389	0
East Fork Point Remove	3.221	1048	0.463	1
Sunnyside Creek	2.520	800	0.377	1
Hogans Creek	2.737	847	0.406	1
Black Fork	2.470	978	0.359	1

Table 6: Top 25 Genera (by abundance) identified across all 8 sample sites located in the Fayetteville Shale region of north-central Arkansas. RC=Rock Creek; DC=Driver Creek; CC=Low Cedar Creek; SH=Sis Hollow; PR=East Fork Point Remove; HC=Hogans Creek; BF=Black Fork; RA=Relative Abundance.

Phylum	**Genus**	RC	DC	CC	SH	PR	SC	HC	BF	RA
Cyanobacteria	*Gloeobacter*	32.680%	49.246%	11.020%	34.014%	13.739%	31.099%	16.535%	26.569%	26.863%
Crenarchaeota	*Nitrosopumilus*	5.180%	5.740%	3.449%	2.683%	2.203%	6.760%	4.644%	4.947%	4.45%
Proteobacteria	*Sphingobium*	3.986%	3.753%	1.998%	1.895%	2.513%	1.249%	2.449%	2.220%	2.51%
Proteobacteria	*Methylibium*	2.726%	1.587%	3.086%	1.828%	1.357%	3.753%	1.994%	1.763%	2.26%
Bacteroidetes	*Sediminibacterium*	1.837%	1.548%	2.640%	2.816%	3.645%	1.577%	1.831%	1.943%	2.23%
Proteobacteria	*Zymomonas*	0.976%	1.474%	1.900%	2.082%	4.859%	0.788%	0.990%	3.827%	2.11%
Cyanobacteria	*Arthronema*	1.505%	0.384%	0.352%	0.289%	2.796%	5.258%	3.289%	0.901%	1.85%
Proteobacteria	*Balneimonas*	1.177%	0.542%	1.352%	0.729%	2.429%	1.052%	1.741%	1.305%	1.29%
Bacteroidetes	*Fluviicola*	0.168%	0.726%	0.784%	1.748%	0.324%	2.631%	1.379%	1.763%	1.19%
Planctomycetes	*Planctomyces*	0.583%	0.225%	1.206%	1.309%	1.170%	0.553%	1.094%	1.225%	0.92%
Proteobacteria	*Novispirillum*	1.535%	0.694%	2.804%	0.594%	0.808%	0.218%	0.534%	0.378%	0.95%
Bacteroidetes	*Leadbetterella*	2.246%	1.247%	1.219%	0.683%	0.250%	0.429%	0.758%	0.280%	0.89%
Proteobacteria	*Dok59*	1.720%	0.701%	1.321%	0.552%	0.387%	0.427%	0.402%	0.539%	0.76%
Planctomycetes	*Nostocoida*	0.178%	0.348%	0.710%	0.600%	0.587%	0.649%	0.985%	0.476%	0.57%
Bacteroidetes	*Saprospira*	0.409%	0.276%	0.983%	0.996%	0.284%	0.258%	0.466%	0.554%	0.53%
Bacteroidetes	*Chryseobacterium*	0.238%	0.101%	0.590%	0.770%	0.326%	0.912%	0.656%	0.691%	0.54%
Cyanobacteria	*Microcystis*	0.003%	0.000%	0.002%	0.000%	0.001%	2.063%	0.005%	5.539%	0.95%
Proteobacteria	*Rhodoferax*	1.167%	0.978%	0.454%	0.245%	0.402%	0.307%	0.542%	0.186%	0.54%
Verrucomicrobia	*Prosthecobacter*	0.057%	0.022%	0.723%	0.805%	0.471%	0.177%	0.312%	0.450%	0.38%
Bacteroidetes	*Dyadobacter*	0.258%	0.546%	0.413%	0.331%	0.302%	0.629%	0.465%	0.608%	0.44%
Deferribacteres	*Mucispirillum*	0.275%	0.521%	0.140%	0.230%	0.244%	0.431%	0.567%	0.391%	0.35%
Proteobacteria	*Luteimonas*	0.013%	0.049%	0.199%	0.365%	0.472%	0.328%	0.169%	0.413%	0.25%
Bacteroidetes	*Sporocytophaga*	0.694%	0.090%	0.360%	0.405%	0.007%	0.008%	0.420%	0.007%	0.25%
Proteobacteria	*Ramlibacter*	0.500%	0.144%	0.210%	0.099%	0.177%	0.019%	0.527%	0.030%	0.21%
Proteobacteria	*Rhodobacter*	0.144%	0.072%	0.297%	0.319%	0.133%	0.221%	0.152%	0.393%	0.22%

Table 7: Microbial biological indicator and remediator OTUs identified at sample sites. Phylum = representative phylum; Class = representative class; Genus = identified genus; Biomarker type = reported type of biological indicator, either (I) = indicator or (R / I) = remediator / indicator; Remediated substrate or bio-indication = substrate indicated as present, or remediated.

Phylum	Class	Genus	Type	Remediated substrate or bio-indication
Cyanobacteria	Synechococcophycideae	*Arthronema*	I	Eutrophication (enhanced Phosphates and Nitrates)
		Acaryochloris	I	Eutrophication (enhanced Phosphates and Nitrates)
		Leptolyngbya	I	Eutrophication, hydrocarbon presence
		Pseudanabaena	I	Eutrophication (enhanced Phosphates and Nitrates)
		Paulinella	I	Eutrophication (enhanced Phosphates and Nitrates)
		Synechococus	I	Eutrophication (enhanced Phosphates and Nitrates)
Cyanobacteria	Oscillatoriophycideae	*Microcystis*	I	Polycyclic aromatic hydrocarbons
		Chroococcus	I	Heavy metals, especially lead, high salinity
		Cyanobacterium	I	Organic pollution, increased fecal coliform concentration
		Chroococciddoipsis	*	* = Undefined
		Phoridium	R / I	Heavy metals, alkenes, eutrophication
		Planktothrix	R / I	Ammonia, hydrocarbons, eutrophication
Verrucomicrobia	Spartobacteria	*Chthoniobacter*	I	Polysaccharides and low salinity
Actinobacteria	Actinobacteria	*Rothia*	R / I	Phenol and petroleum pollutants
		Streptomyces	R / I	Heavy metals
		Nocardioides	R / I	Herbicides
		Rhodococcus	R / I	Benzene
		Catellatospora	R / I	Arsenic

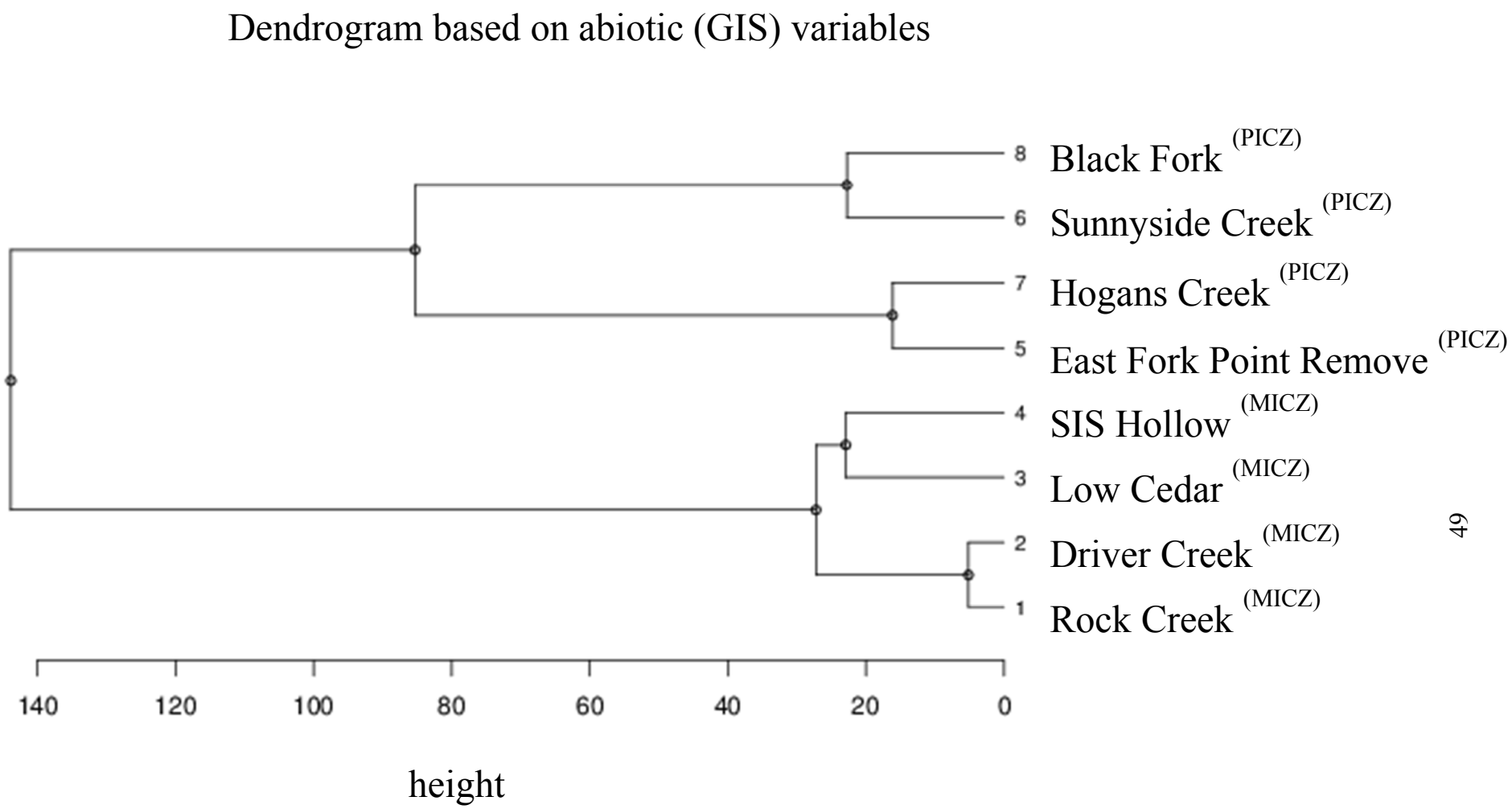

Figure 1: Ward hierarchical dendrogram based on the unweighted pair-group method with arithmetic means (UPGMA) algorithm and derived from eight abiotic variables gathered at study sites located in the Fayetteville Shale region of north-central Arkansas. Driver Creek and Rock Creek were the most similar of all sites, and these MICZ sites clustered with two MICZ sites, SIS Hollow and Low Cedar Creek. The remaining four sites, all PICZ sites, clustered together with Hogans Creek and East Fork Point Remove being most similar.

Figure 2: Map of Arkansas showing the Fayetteville Shale region and drill sites. The majority of sites are located in a seven county region of north-central Arkansas: Pope, Conway, Van Buren, Faulkner, Cleburne, White, and Independence. All sites are within in the Arkansas River drainage basin.

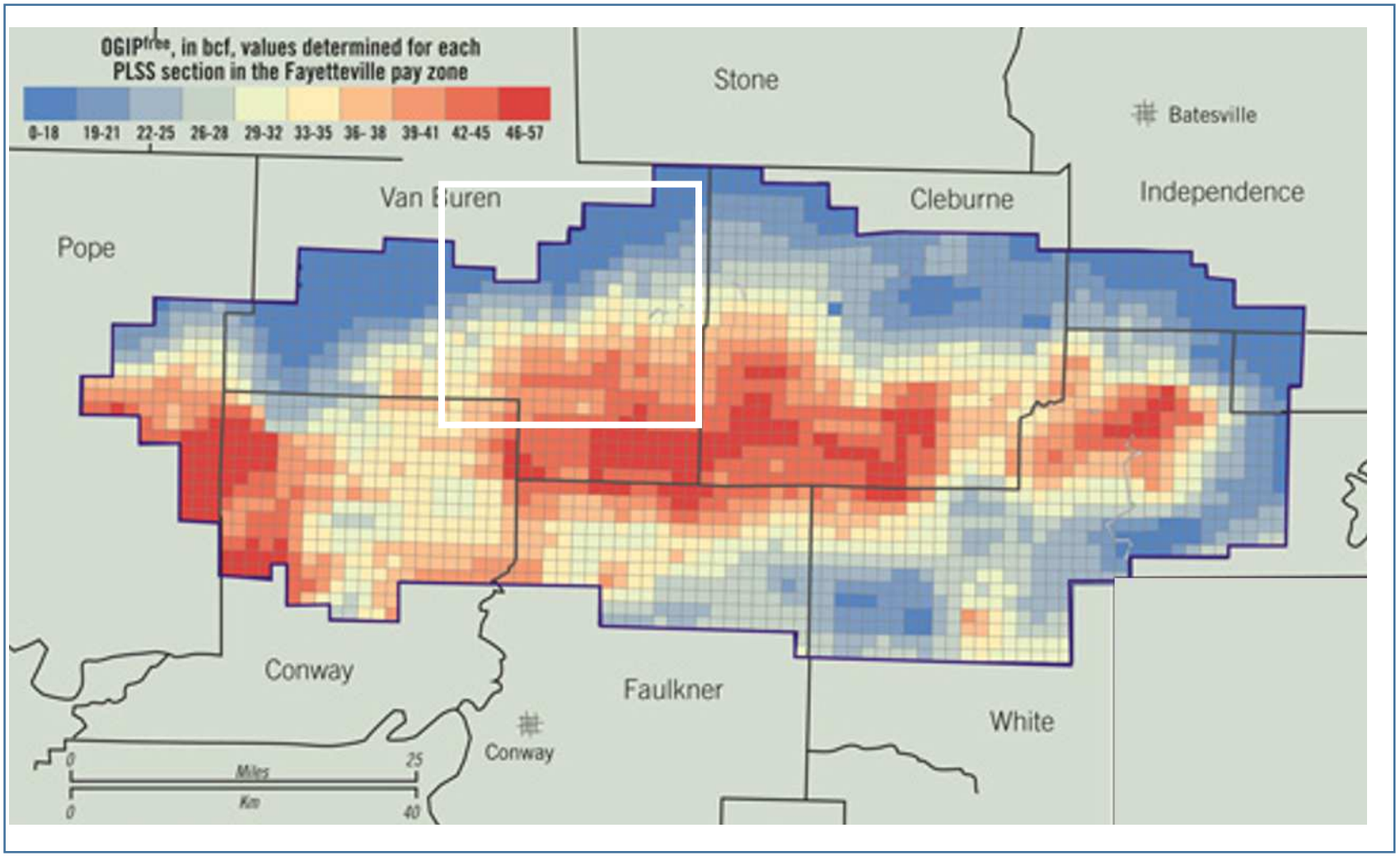

Figure 3: North-central Fayetteville Shale Play with color-coded estimates of 'original natural gas in place', referring to the estimated natural gas that can be extracted from the shale deposit (Browning et al. 2014). White box depicts sample area (Figure 4).

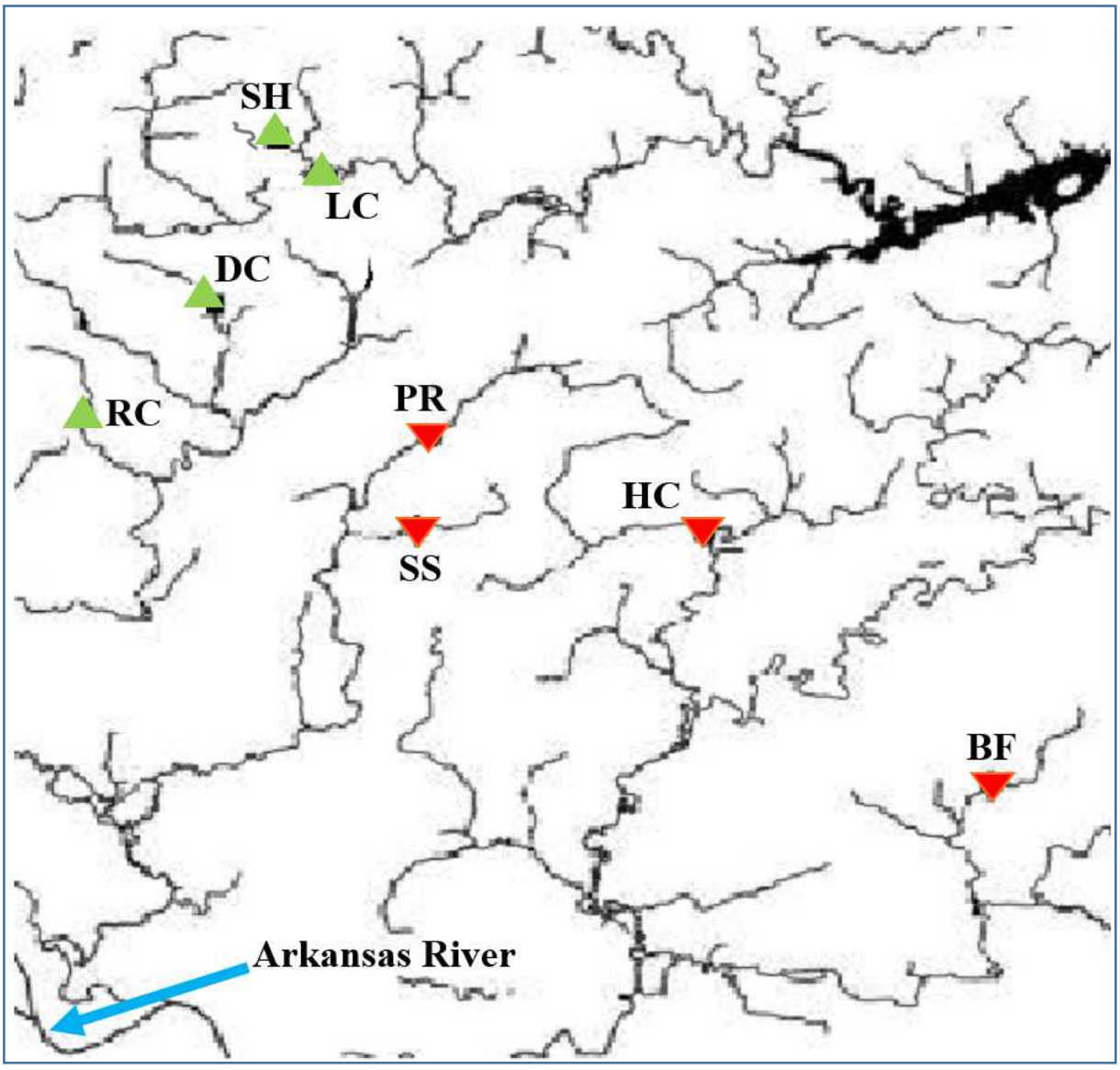

Figure 4: Current study area within the Fayetteville Shale Play region of north-central Arkansas. Site abbreviations are: (RC) Rock Creek, (DC) Driver Creek, (LC) Low Cedar Creek, (SH) Sis Hollow, (SS) Sunnyside Creek, (HC) Hogans Creek, and (BF) Black Fork. Sites marked with (▲) are within the minimally impacted catchment zone (MICZ) and those marked with (▼) fall within potentially impacted catchment zones (PICZ). The Arkansas River, as the primary drainage basin, is seen in the extreme southwest corner.

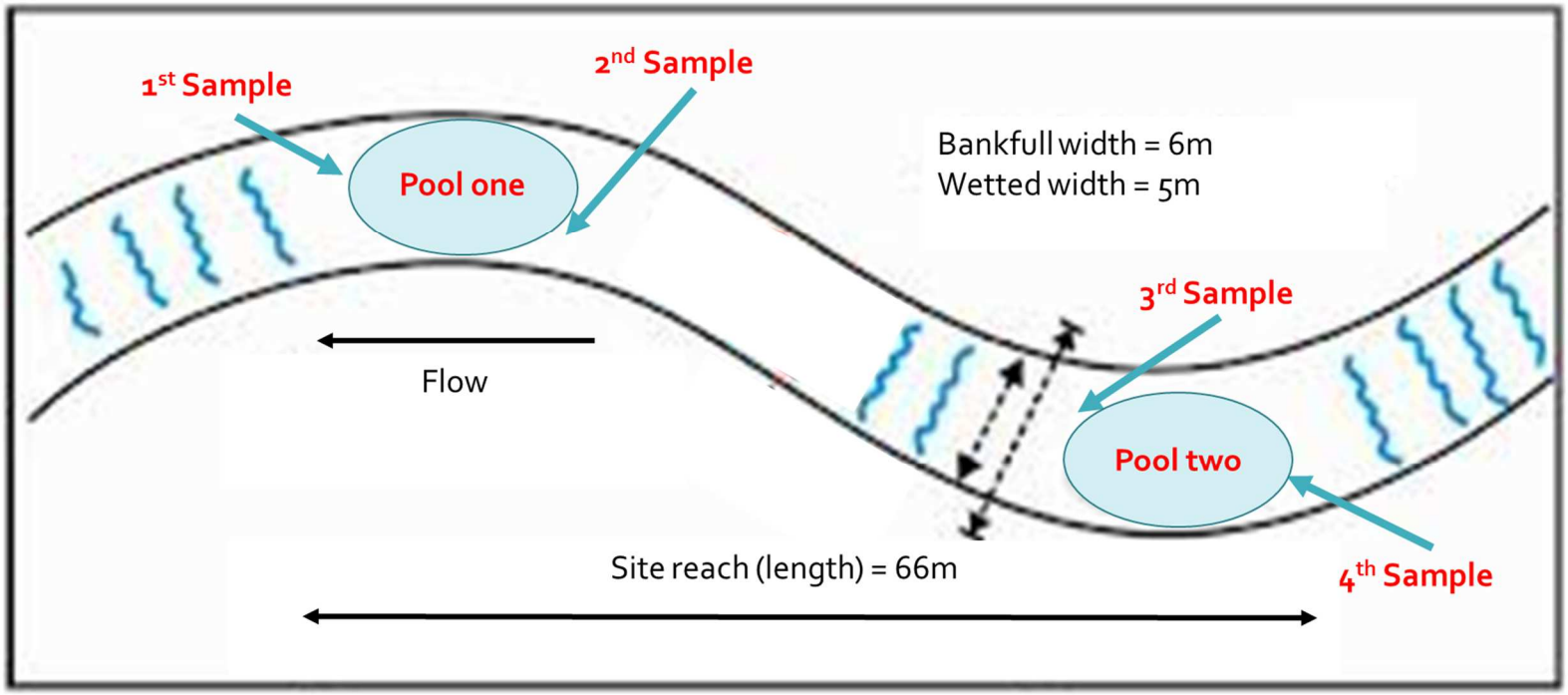

Figure 5: Sampling regime and normalization at Rock Creek in the Arkansas River drainage of Arkansas. Site length was approximately 66m, with average bankfull width = 6m and average wetted width = 5m. Samples were first taken from downstream (right side) then upstream (left side) at each pool, at a distance approximately 30% from edge into transect.

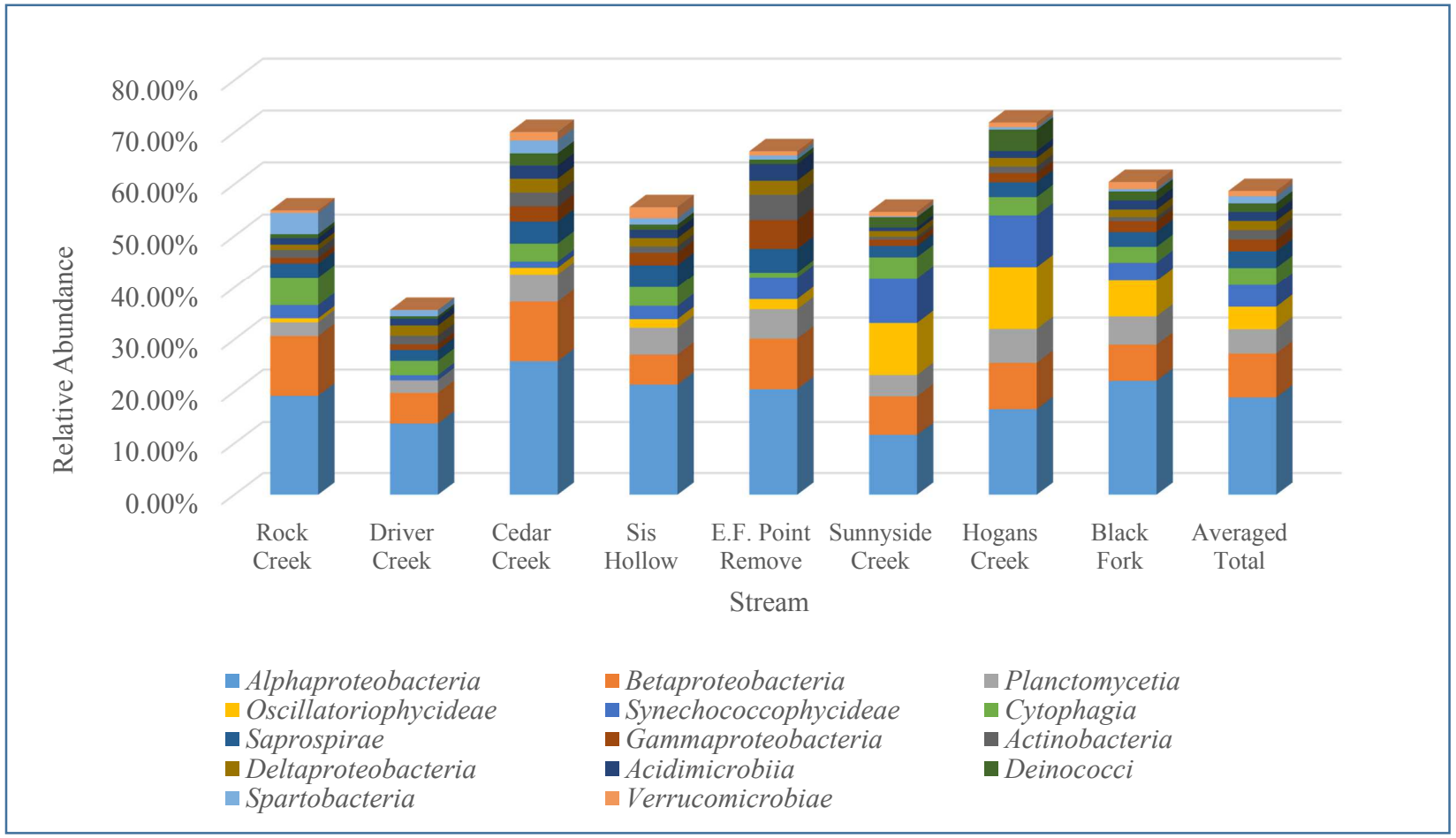

Figure 6: Top 14 classes relative abundance based on percentage of total identified classes at each site with hierarchy determined by an averaged total, as derived from study sites located in the Fayetteville Shale region of north-central Arkansas. Stacked columns have the most abundant at the base and proceed generally to least abundant. Alphaproteobacteria were most abundant (18.9%), Betaproteobacteria were second most abundant (8.45%), and Planctomycetia were third most abundant (4.63%). Overall average abundance was 4.19%.

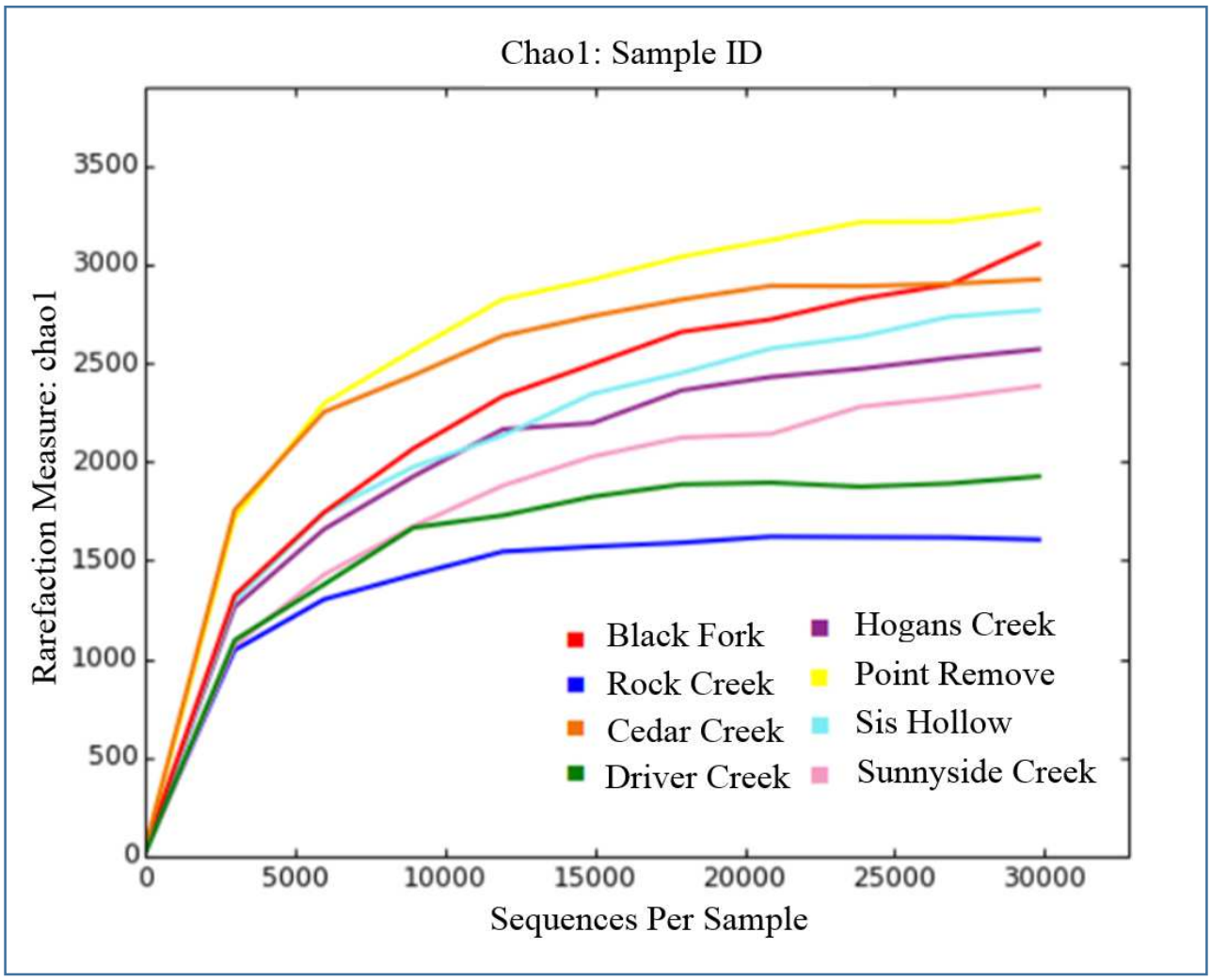

Figure 7: Alpha diversity Chao1 species richness calculated on rarefied samples at a sampling depth of 29,800 for study sites located in the Fayetteville Shale region of north-central Arkansas. Samples were pooled by Sample ID.

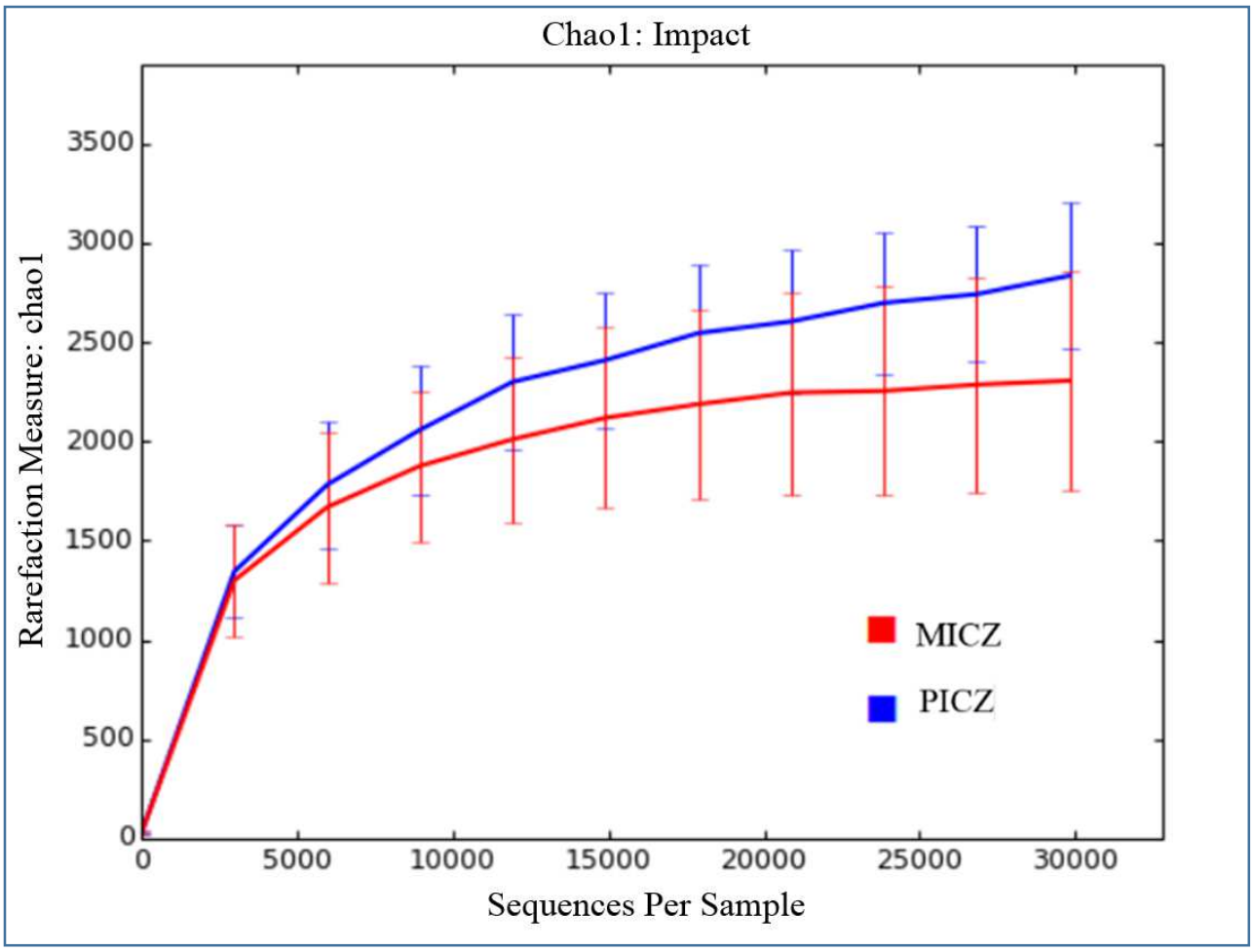

Figure 8: Alpha diversity Chao1 species richness calculated on rarefied samples at a sampling depth of 29,800 for study sites located in the Fayetteville Shale region of north-central Arkansas. Samples were pooled by impact factor. PICZ sites exhibited greater richness than MICZ sites, but this result was not statistically significant.

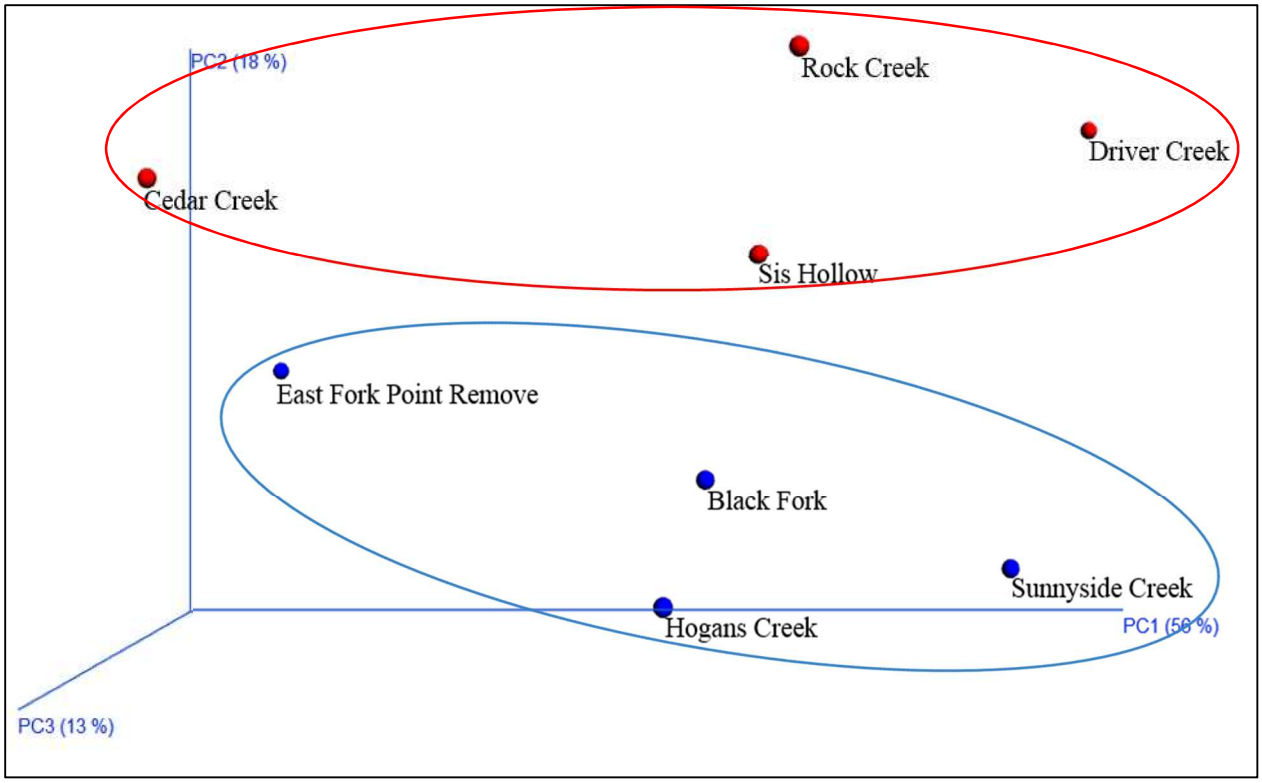

Figure 9. Weighted Unifrac Principal Coordinate plot depicting the relationships among four PICZ sites (in red) and four MICZ sites (in blue) located in the Fayetteville Shale region of north-central Arkansas. Axis1 and Axis2 accounted for 74% of the variation in the data (56% and 18% respectively). MICZ sites cluster together along the top of the second axis, while PICZ sites cluster along the bottom of the second axis.

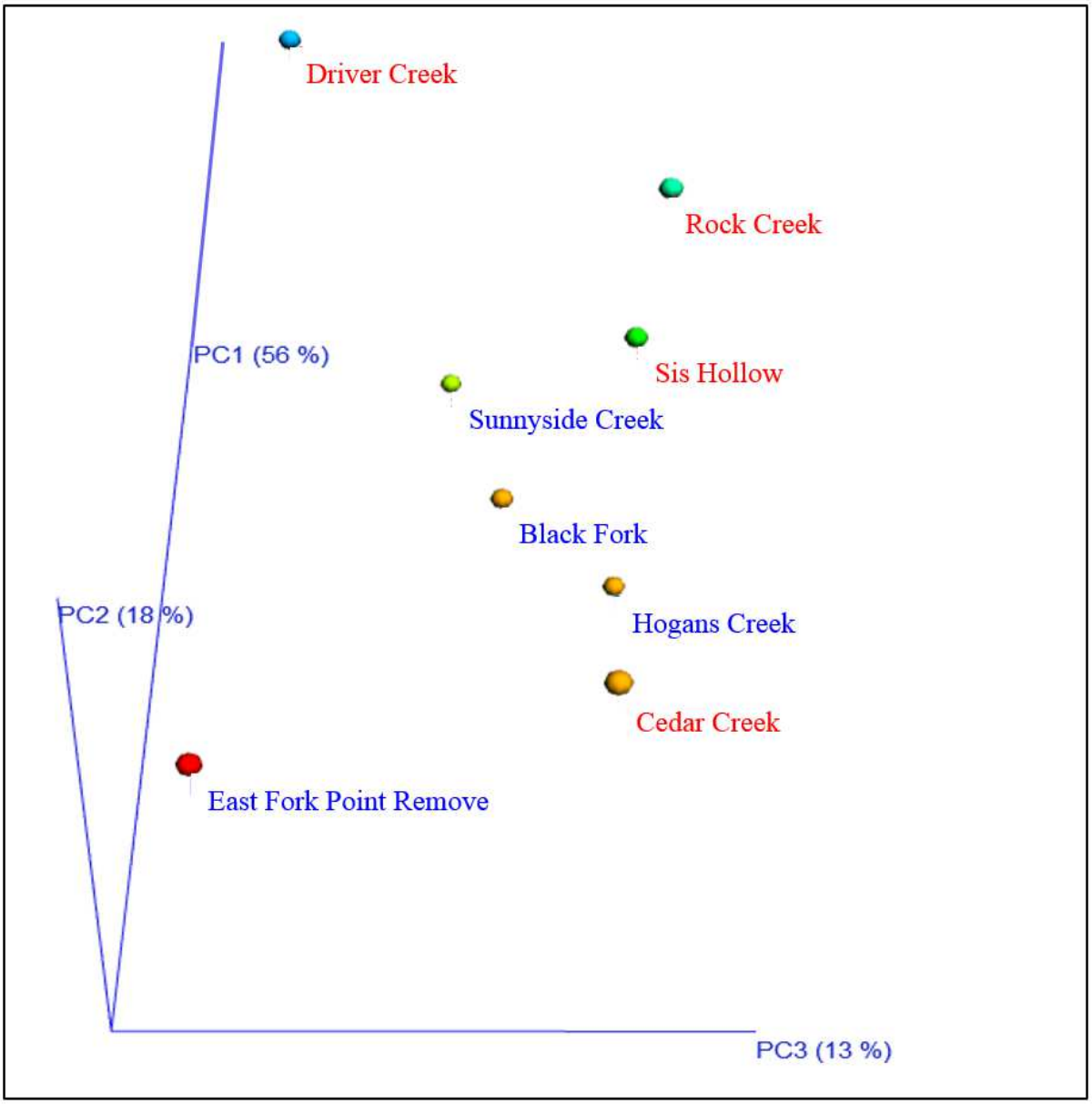

Figure 10. Weighted Unifrac Principal Coordinate plot of each sample site according to Strahler Stream Order (SSO), as derived from study sites located in the Fayetteville Shale region of north-central Arkansas. MICZ sites are labeled in red and PICZ sites are labeled in blue. Colors of each ball are established by each sample site's SSO, and the size of each ball represents its relative position along the second axis. A straight gradient is found along the first axis, from the highest SSO (6 – Point Remove) at the bottom to the lowest SSO (1 – Driver Creek) at the top.

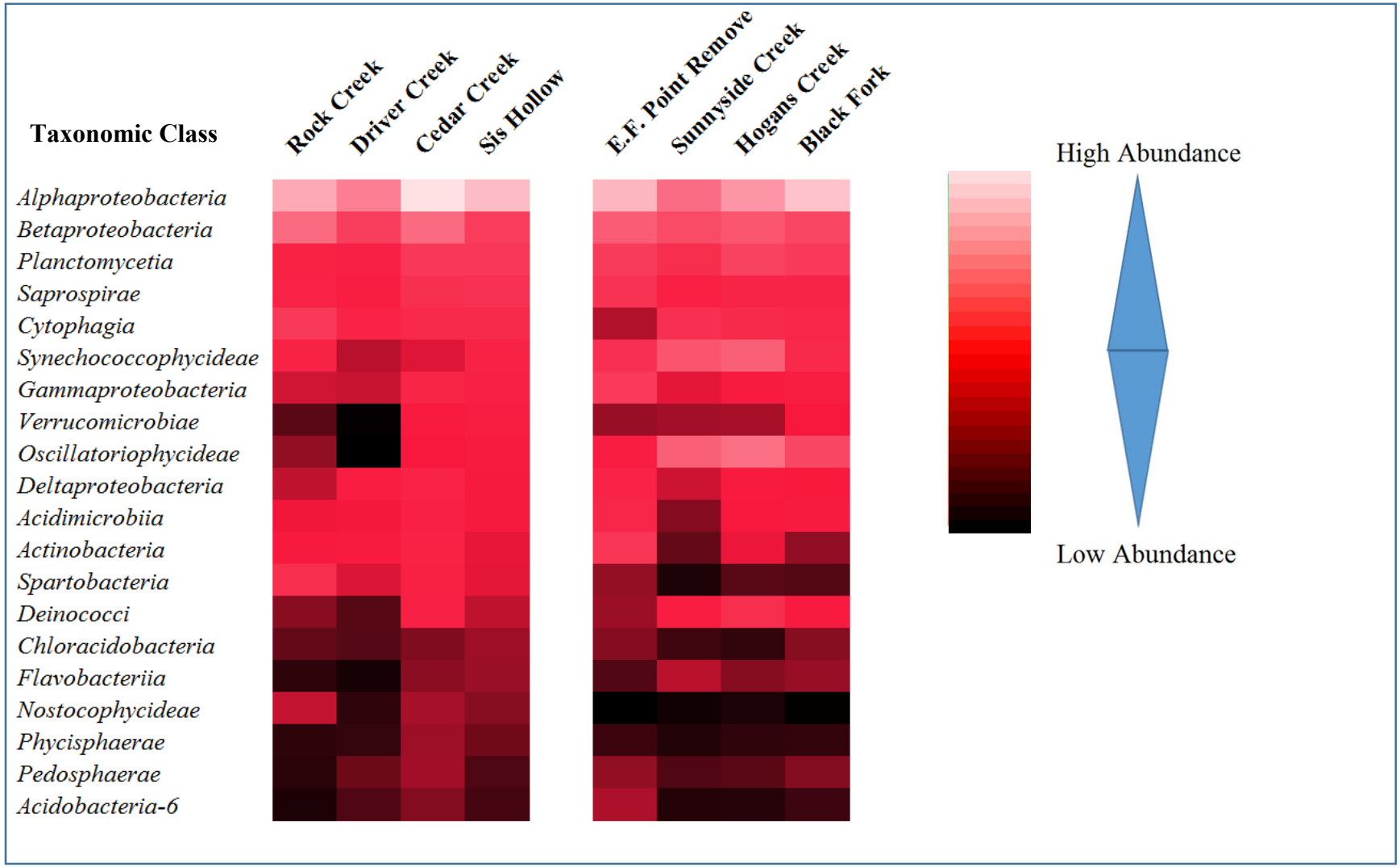

Figure 11. Heat maps of top 20 identified classes based on abundance of microbes at each site located in the Fayetteville Shale region of north-central Arkansas. The heat map on the left side represents the four MICZ sites, and the heat map on the right represents the four PICZ sites.

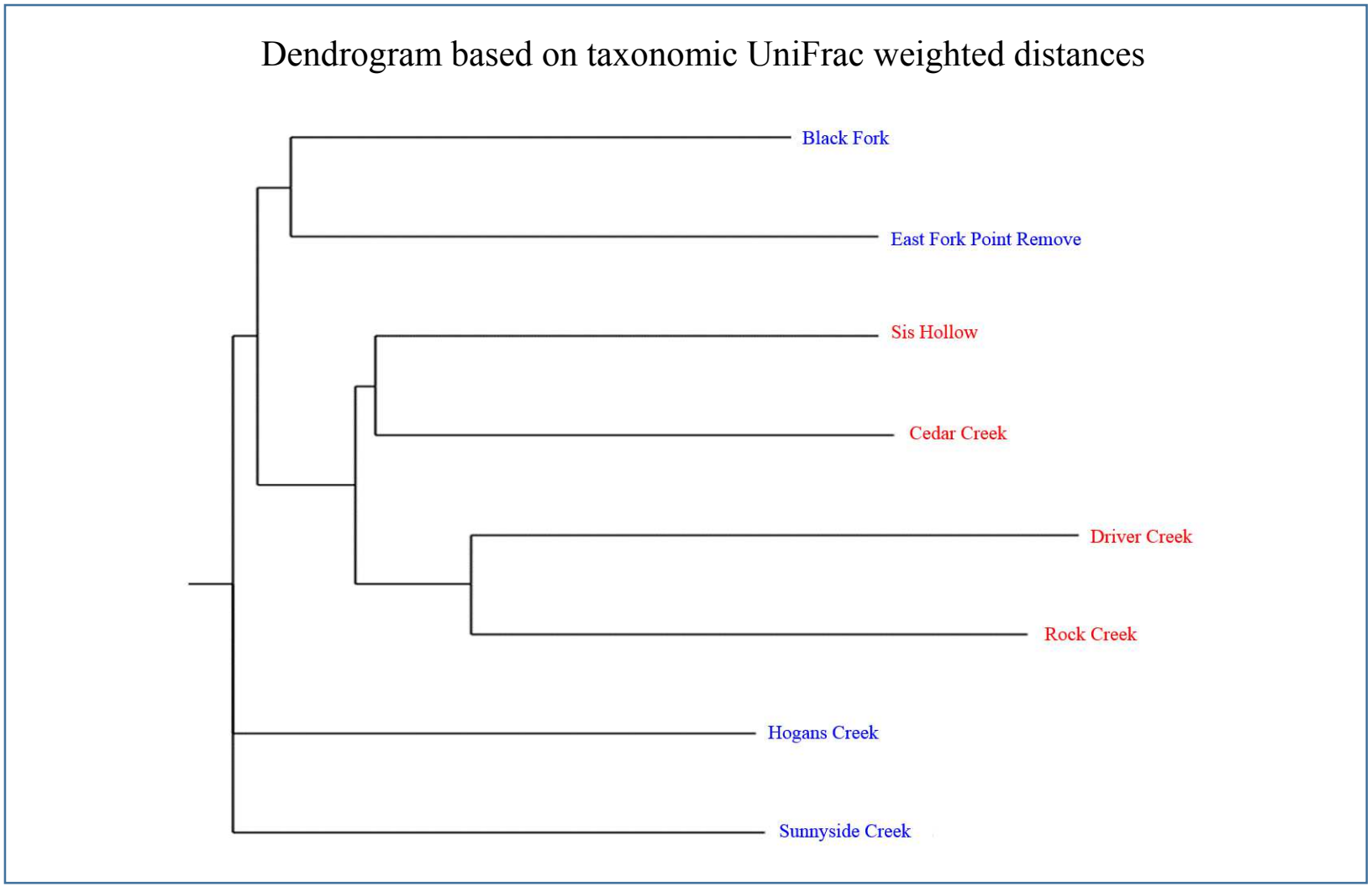

Figure 12: Neighbor-joining tree based on UniFrac weighted phylogenetic distance matrix. All four MICZ sites (in red) cluster together in a clade, with Sis Hollow and Cedar Creek being most similar. Two PICZ sites (in blue), Black Fork and East Fork Point Remove, are sister to the MICZ sites, with the remaining PICZ sites (Hogans Creek and Sunnyside Creek) sister to the previous PICZ sites.

Appendix 1 – supplemental

Final optimized Biofilm extraction protocol from Nasco Whirl-Pak Speci-Sponges™

The final biofilm protocol was as follows: Remove each Nasco Whirl-Pak Speci-Sponges™ sample from the -80°C freezer and thaw. Prepare 500ml of standard PBS buffer (400ml dH_2O, 0.1g KCl, 0.89g $NaHPO_4 \cdot H_2O$, 0.135g KH_2PO_4, thoroughly mixed then brought up to 500ml) and vacuum filter in a 1000ml capacity 0.2µ pore. Pour 35ml of PBS buffer into Whirl-Pak with sponge, reseal top opening and place flat on counter and apply alternating pressure by hand for five minutes, inducing a forced swirling of PBS into and out of the sponge to solubilize the biofilm material. Pipette resultant elution into two sterile 15ml tubes and centrifuge @ 5000 x G for five minutes. Remove supernatant, resuspend each pellet in a minimal amount of supernatant and place into a 1.5ml Eppendorf tube, previously weighed so as to allow for determination of pellet mass, and centrifuge @ 1000 x G for one minute. Remove supernatant and weight pellet. The process is repeated until ~ 0.2g of pellet is obtained with no remaining supernatant

The MOBIO Kit protocol used to extract the 16 samples

Sample pellets were suspended in 350µl of BF1 and transferred to PowerBiofilm® Bead Tube, then 100µl of BF2 was added, briefly vortexed, and incubated at 65°C for five minutes, with the bead tube labeled on top and side for clarity.

We used a BioSpec® Mini-Beadbeater 16 (30 seconds @ 3500) to lyse cells. The sample was then centrifuged @ 13000 X G for one minute and the supernatant was transferred to a fresh 1.5ml tube.

After bead beating, the supernatant remained pigmented, so an additional 100µl of solution BF3 was added (where BF3 is designed to remove humic and PS co-precipitates). This was vortexed then centrifuged @ 13000 X G for one minute and the ~400µl of supernatant was transferred to a fresh 1.5ml tube. Solution BF4 (900µl) was then added and briefly vortexed. Solution BF4 was kept in a water bath at 37°C to prevent precipitation prior to use, and 650µl of this solution was added to a spin-filter column then centrifuged @ 13000 X G for 1 minute. The flow-through was discarded and this process was repeated once to completely load all supernatant onto the filter. The filter was placed into a clean collection tube and solution BF5 was vortexed, with 650µl loaded onto the filter and centrifuged @ 13000 X G for one minute. The flow-through was discarded and 650µl of BF6 was loaded on the filter and centrifuged @ 13000 X G for one minute. The flow-through was discarded. The spin filter was then centrifuged @ 13000 X G for five minutes to completely dry the membrane. The filter was then placed in a clean collection tube and 50µl of BF7 (elution buffer), which was heated to 42°C, was carefully added to the center of the filter and allowed to incubate at room temp for two minutes, then centrifuged @ 13000 X G for 1 minute to elute. Each sample DNA extraction was quantified using Life Technologies Qubit® 2.0 Fluorometer.

Primers used

Four sets of primers used for testing and optimizing extractions: (1) universal bacterial – 27F (5'-AGAGTTTGATCMTGGCTCAG-3') and 1492R (5'-TACGGYTACCTTGTTACGACTT-3') which amplifies 97.3% of the full length (1541 nucleotide) bacterial 16S rRNA gene; 27F (spans positions 8 to 27 in *Escherichia coli* rRNA coordinates) and 1492R (spans positions 1492 to 1507) (Weisburg et al. 1991); (2) broad range

fungal – 5.8sF (5'-GTGAATCATCGARTCTTTGAA-3') and ITS1fR (5'-TCCGTAGGTGAACCTGCGG-3') which amplifies basidiomycete ITS sequences from mycorrhiza samples commonly used for molecular systematics at the species level (Gardes and Bruns 1993); (3) universal eubacterial – Eub338F (5'-ACTCCTACGGGAGGCAGCAG-3') and Eub518R (5'-ATTACCGCGGCTGCTGG-3') which amplifies a partial universal subset of the bacterial 16S rRNA gene; Eub338F (spans positions 320 to 338 in *Escherichia coli* rRNA coordinates) and Eub518R (spans positions 518 to 537) (Fierer et al. 2005); and (4) *Firmicutes* Lgc353F (5'-GCAGTAGGGAATCTTCCG-3') and Eub518R (5'-ATTACCGCGGCTGCTGG-3') which amplifies a partial subset of the bacterial 16S rRNA gene of *Firmicutes*; Lgc353F (spans positions 334 to 353) (Guo et al. 2008).

Appendix 2 – custom Perl script *convert_cls.pl*

```
#!/usr/bin/perl
# convert comma separated file to tab delimited
# author Wil Johnson
# simple script to do the conversion

use strict;
use warnings;

system ("clear");
system ("ls *.csv");
print "\n";
print 'What is the comma separated file you need to convert? ';
chomp (my $csv_file = <> );
print 'What shall I name the new file? ';
chomp (my $tabd_file = <> );
system ("< $csv_file tr \",\" \"\t\" > $tabd_file");
exit;
```

Figure A1. Perl script designed to elicit user response to modify a comma separated txt file into a tab delimited file regardless of the size or complexity of the file.

Appendix 2 – custom Perl script *headerMod.pl*

```
#!/usr/bin/perl

use strict;
use warnings;
use Bio::SeqIO;
=head1 Name
headerMod.pl
=head1 Usage
headerMod.pl  <fastaFile>
=head1 Synopsis
This scripts takes an Illumina generated fasta file with a QIIME incompatible header and converts it to work with QIIME.
Generally your header will be in some format such as:
#>M02146:10:000000000-A51MH:1:1101:13422:1525
A new fasta file will be written out containing the sequence(s) with new headers of the form:
#>'SampleID_1' 'uniqueSeqIdentifier' orig_bc='AGTCGTGCCTCC'
new_bc='AGTCGTGCCTCC' bc_diffs=0
like so
#>up.Rock_6 1101:12437:2258 orig_bc=AGTCGTGCCTCC new_bc=AGTCGTGCCTCC
bc_diffs=0
note the 3 numbers separated by ":" are still intact (these are not homogenous and will serve as the uniqueSeqIdentifier
The script substitutes the arbitrary unchanging initial header (M02146:10:000000000-A51MH:1:) with your matching ID (must be identical to mapfile ID) and adds the sequence number (QIIME required), retains the uniqueSeqIdentifier, and inserts the remaining QIIME requirements-the supplied barcode will be used for both "orig" and "new" so bc_diffs=0 will also be set
```

USING THIS SCRIPT:

1 - run script with FASTA file as single argument (headerMod.pl <fastaFile>)

2 - choose new modified FASTA file name (prompted)

3 - provide matching mapfile (prompted) > mapfile will be diplayed as will the original header from target Fasta file

4 - copy arbitrary header to be replaced and paste it accordingly (leave '>' character out of match to replace)

5 - choose SampleID to replace arbitrary run header (prompted)

6 - copy and paste correct barcode for specific sampleID

7 - inspect results for accuracy

You can use 'head' and 'tail' unix commands on your fasta file (if more than one sequence/header is within) to ensure all headers were matched and replaced

Another check: If your Illumina header is shorter than your QIIME compatible (likely) your new file should be larger, otherwise your pasted match did not match throughout your FASTA file

Check and rerun to be sure you only remove what doesn't change throughout your multiple fasta sequences. ENJOY!

=head1 Author

Wil H. Johnson, UofA

=cut

```
unless (@ARGV ==1){ die "Usage:  headerMod.pl  fastaFileName";}
my $oFile = shift;
system ("clear");
print "When running a 'white space' warning will appear-ignore this as white space is required for QIIME compatibility\n";
print "\nYour original Illumina fasta file name: $oFile\n";
print 'New QIIME compatible   fasta file name: ';
chomp (my $newFile = <> );
my $seq_in  = Bio::SeqIO->new( -format => 'fasta', -file => $oFile);
my $seq;
my $seq_out = Bio::SeqIO->new('-file' => ">$newFile", '-format' => 'fasta');
```

```
system ("clear");
system ("ls *.txt");
print "\nWhat is the mapfile associated with this sample? ";
chomp (my $mapfile = <> );
open(DATA, $mapfile) or die "Couldn't open file $mapfile, $!";
print "\nPartial mapfile $mapfile  with barcodes shown\n";
while(<DATA>){ print substr($_, 0, 48);
   print "\n";
}
print "\nIllumina header of original FASTA file     ";
system ("head -n1 $oFile");
print 'Copy & Paste characters to match & replace: ';
chomp (my $seq_char = <> );
print "working file: $oFile\nWhat is the new label (SampleID) to attach? ";
chomp (my $label = <> );
print 'What is the original barcode? ';
chomp (my $b_code = <> );
my $seqnumber=1;
open (STDERR, '>>', "log_$newFile");
while( $seq = $seq_in->next_seq() )
{
        my $seqName = $seq->id;
        $seqName =~ s/$seq_char/$label\_$seqnumber /g; #replace arbitrary with new label and
seqeunce count number globally
      $seqName =~ s/(gi\.\w*)\..*/$1/;
        $seqName=$seqName . " orig_bc=$b_code new_bc=$b_code bc_diffs=0"; #add in
remaining QIIME dependencies
      $seq->id($seqName);
        $seq_out->write_seq($seq);
        $seqnumber=$seqnumber+1;
}
```

```
system ("clear");;
print "\nYour sequences have been renamed and are in the file $newFile\n\n";
system ("ls -lF -1 $oFile $newFile");
print "\n\nOriginal header\n";
system ("head -n1 $oFile");
print "\n\nNew header\n";
system ("head -n1 $newFile");
my $headernumber=$seqnumber-1;
print "\nTotal number of modified headers= $headernumber\n\n";
my $filename = "log_$newFile";
my $filesize = -s $filename;
my $nfilesize = -s $newFile;
my $errornum = substr($filesize/$nfilesize,0,3);
if ($errornum = 1.3) {
        print "The error log only contains 'white space warnings' and will be deleted";
        system ("rm log_$newFile");
} else {
        print "There were errors check head and tail of log_$newFile!";
}
print "\n\n";
```