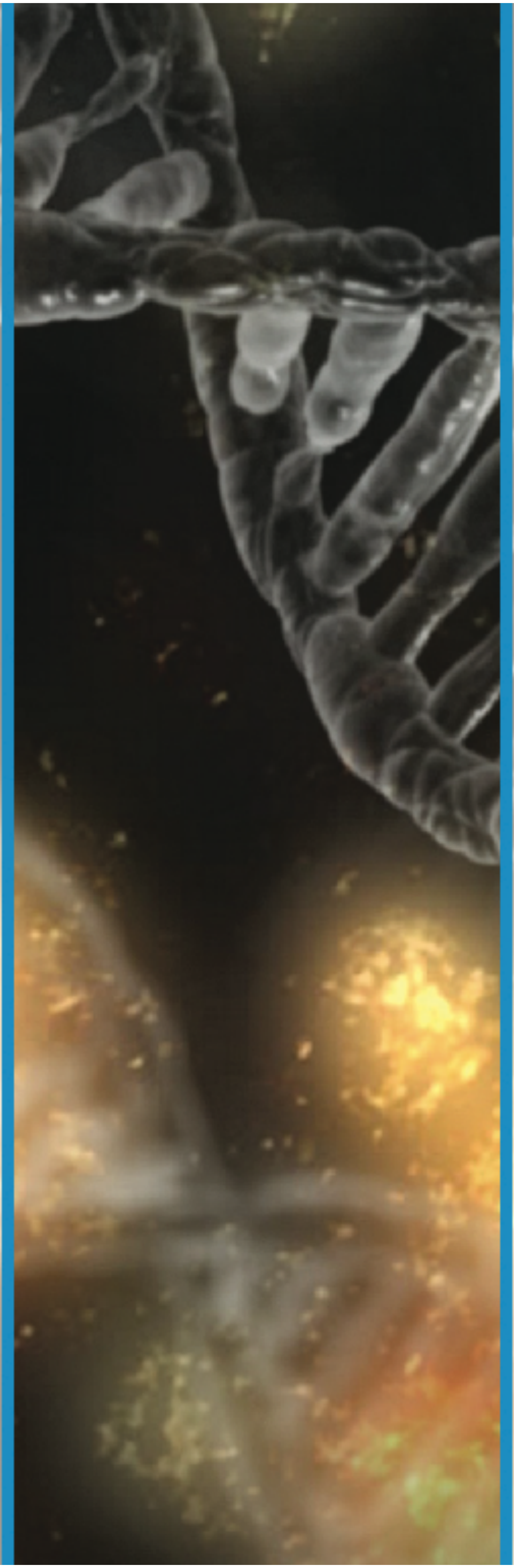


# Next-Generation DNA Sequencing (NGS)

*Description, Sample Collection, Data Interpretation & Case Studies*

## Table of Contents

# Next-Generation Sequencing (NGS)

## 1.0 Introduction

Biodegradation of groundwater contaminants, particularly emerging contaminants, is a complex process driven by the particular composition, dynamics, and metabolic functionality of subsurface microbial communities characterizing an impacted site. The development of effective strategies for the management and remediation of such sites requires an in-depth knowledge of the composition and metabolic potential of microbial communities.

Next-generation DNA sequencing (NGS), or high-throughput sequencing, is a collection of advanced technologies for ascertaining the precise order of bases within a DNA molecule. In addition to its unprecedented throughput, NGS offers the advantages of scalability and speed in determining DNA sequences much less expensively than previous sequencing methods. With NGS, one can survey in a cost-effective manner the genomes of entire communities or microbiomes, including those of unculturable constituents.

NGS provides identification of microorganisms present in a field or well sample down to the taxonomic level of genus with no prior knowledge of the microbial community composition. Each sequenced segment of DNA is indicative of a specific microorganism. Although metabolic activity cannot always be predicted from phylogeny, comprehensive identification of the microorganisms present in an environment offers deep insight into the potential microbial processes impacting bioremediation. No other microbial analysis provides more comprehensive characterization of the microbial community in a field sample or better answers the question: *What microorganisms are present?*

## 2.0 Next-Generation Sequencing

### 2.1 How Does NGS Work?

The various NGS platforms all provide massively parallel sequencing which allows millions of nucleic acid fragments to be sequenced simultaneously and rapidly.[1] While each NGS platform is unique (*e.g.*, Illumina MiSeq, Ion Torrent PGM), the overall steps and the underlying concepts of Next-Generation Sequencing are similar (Figure 1). The general methodology involves template or library preparation, nucleic acid sequencing, and data analysis. First, community genomic DNA (cgDNA) is extracted from an environmental sample and fragmented into a library of small nucleic acid segments. The ends of these DNA fragments are then ligated with a chemically synthesized adaptor molecule, which is a DNA molecule of known sequence. Second, the library is amplified and subsequently sequenced in millions of parallel reactions.

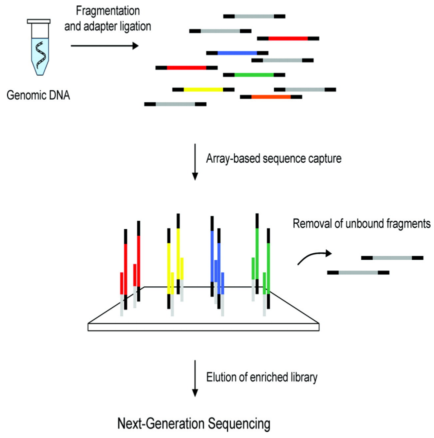

Fragmentation and adapter ligation
Genomic DNA
Array-based sequence capture
Removal of unbound fragments
Elution of enriched library
Next-Generation Sequencing

*Figure 1: The NGS Approach – DNA Library Construction, Amplification, and Massively Parallel Sequencing*

The sequencing step is similar to previous methods: the bases of each DNA fragment are sequentially identified from light signals emitted as the complement to each fragment strand is resynthesized. The net result is a set of newly identified 'strings' of nucleotides called 'reads' that represent specific members of the microbial community present in the original sample. Comparisons of next-generation sequencing results between samples can reveal important differences or shifts in the microbial community by location, over time, or in response to site activities.

## 2.2 NGS Data Analysis and Interpretation

NGS is not quantitative like quantitative polymerase chain reaction (qPCR). Sequencing results obtained from NGS technology are reported as relative abundances with units of "percent of hits"—the percent of total sequences that have been identified as belonging to a particular microbial genus. Because NGS generates massive sequencing datasets, it is necessary to apply a suite of bioinformatic tools to extract meaningful biological information and to make valid inferences and predictions.  These analytic and statistical techniques are described in more detail as follows.

**2.2.1 Diversity Indices.** The Shannon diversity index is a quantitative measurement that characterizes how many different genera are present in the sample and takes into account the distribution of the number of organisms classified to each genus present in the sample (commonly referred to as species evenness).[2,3] Shannon's diversity index increases in value as the number of genera increases and as the number of organisms present per genera becomes even. Simpson's index measures the probability that two individuals selected randomly from the sample would belong to different genera: the greater the value, the greater the sample diversity. The Chao1 index is an excellent indicator of species richness and is based on the number of reads when one (singleton) or two (doubleton) operational taxonomic units (OTUs) are observed. This value is the predicted number of genera based on the number of singletons and doubletons. The total genera observed is presented here, but does not include reads unclassified at genus species.

**2.2.2 Principal Coordinate Analysis.** Principal coordinate analysis (PCoA) is used to visualize differences in microbial communities between samples.[4] Unlike more traditional methods such as principal component analysis (PCA), PCoA calculates complex functions for the axes rather than dimensional scaling used in PCA. Therefore, PCoA is able to better demonstrate dissimilarities that may be nuanced in PCA tests. PCoA accomplishes this by using a dissimilarity matrix to assign each sample a location in dimensional space, then changes the coordinate system to display the data in two dimensions. This analysis allows us to visualize multidimensional data in two dimensions. The scatterplot in Fig. 2 shows a PCoA of the normalized relative abundance of all samples at the genus-level classifications.  Increasing distance between sample points on this plot indicates increasing dissimilarity between bacterial populations in the samples. From the opposite perspective, the microbial community compositions of samples that group near each other in the PCoA plot are more similar. For example, the bacterial community of MW2 is highly similar to that of MW5 (Fig.2, upper left corner). Conversely, the microbial community of MW1 is not particularly similar to those of any other sample collected.

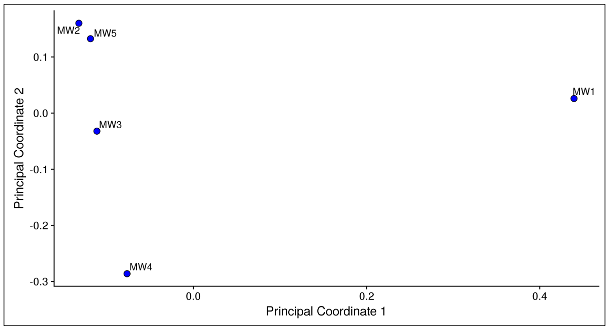


*Figure 2: Principal Coordinate Analysis.*

**2.2.3 Hierarchical Clustering Dendrogram.** Hierarchical clustering is accomplished by comparing dissimilarities between the samples using complete agglomeration of the Bray-Curtis dissimilarity. This groups together samples which are the least dissimilar. The length of the branches indicate the amount of dissimilarity between samples. Therefore, shorter branches are more similar. An example of a Hierarchical Clustering Dendrogram is shown in Fig. 3. The bar chart beneath each sample shows the relative abundance of the top 8 of genus-level classifications, along with all other classified and unclassified genera. Notice that samples MW2 and MW5 cluster together in Figure 3 while MW1 is an outlying branch.

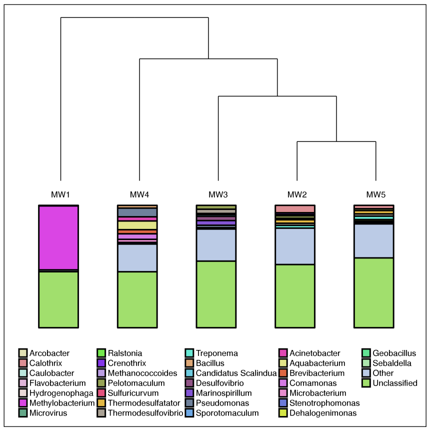
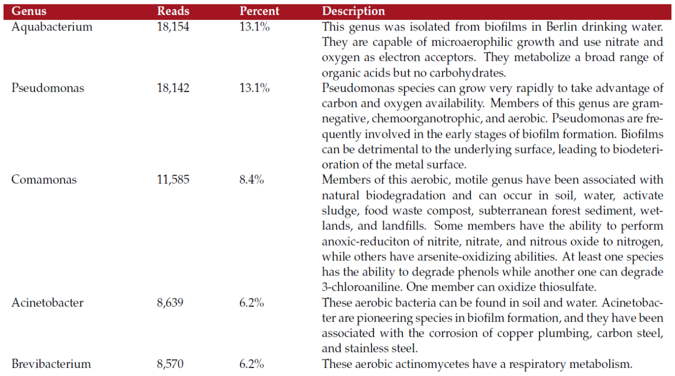


*Figure 3: Hierarchical Clustering Dendrogram.*

NGS is most appropriate for identifying members of the microbial community present in a sample when little is known about the process in question. NGS data are presented graphically using pie charts showing the relative proportion of the top phylum classification results (see Fig. 4 below) and top genus classification results. The top genus classification results are further elaborated in tables providing the specific genus, the corresponding number of reads and percent total reads, and a brief description of the primary metabolic activities exhibited by members comprising the particular genus. A partial example of top genus classification results is shown in Table 1.



| Genus | Reads | Percent | Description |
|---|---|---|---|
| Aquabacterium | 18,154 | 13.1% | This genus was isolated from biofilms in Berlin drinking water. They are capable of microaerophilic growth and use nitrate and oxygen as electron acceptors. They metabolize a broad range of organic acids but no carbohydrates. |
| Pseudomonas | 18,142 | 13.1% | Pseudomonas species can grow very rapidly to take advantage of carbon and oxygen availability. Members of this genus are gram-negative, chemoorganotrophic, and aerobic. Pseudomonas are frequently involved in the early stages of biofilm formation. Biofilms can be detrimental to the underlying surface, leading to biodeterioration of the metal surface. |
| Comamonas | 11,585 | 8.4% | Members of this aerobic, motile genus have been associated with natural biodegradation and can occur in soil, water, activate sludge, food waste compost, subterranean forest sediment, wetlands, and landfills. Some members have the ability to perform anoxic-reduciton of nitrite, nitrate, and nitrous oxide to nitrogen, while others have arsenite-oxidizing abilities. At least one species has the ability to degrade phenols while another one can degrade 3-chloroaniline. One member can oxidize thiosulfate. |
| Acinetobacter | 8,639 | 6.2% | These aerobic bacteria can be found in soil and water. Acinetobacter are pioneering species in biofilm formation, and they have been associated with the corrosion of copper plumbing, carbon steel, and stainless steel. |
| Brevibacterium | 8,570 | 6.2% | These aerobic actinomycetes have a respiratory metabolism. |

*Table 1: Top genera classification results.*

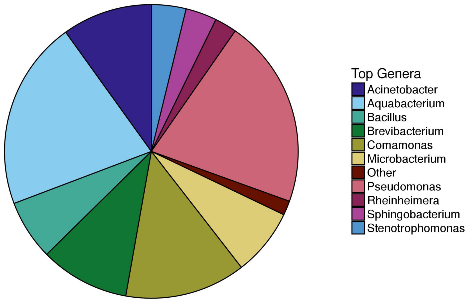


*Figure 4: Pie Chart Displaying Top Genus Classifications.*

In summary, the analysis of NGS data can provide broad insights into microbial community dynamics, such as differences in microbial diversity between a background well and a contaminated well, or overall temporal shifts in microbial community composition. Information on potential microbial activities occurring at a contaminated site can then be used to develop tools or select quantitative PCR targets for routine monitoring and ultimately controlling a complex microbial process.

# 3.0 Selecting Sampling Locations and Sample Collection Procedures

## 4.1 Sample Collection and Preservation

Collecting samples for NGS analysis is no more difficult than collecting groundwater or soil samples for common chemical analyses and can be readily incorporated into a routine sampling event. Below are guidelines to follow when collecting samples for any DNA-based analysis.

1. Use clean latex (or similar) gloves when collecting and handling samples.
2. Keep samples cold (~4°C) to minimize changes in the microbial community.
    a. Place samples on ice or freezer packs in a cooler immediately after collection.
    b. As soon as possible (preferably overnight), ship samples to the laboratory.
    c. Include enough ice/freezer packs to ensure that samples remain cold during shipment.

Microbial Insights (MI) has been receiving field samples for DNA-based analyses for over 25 years and has performed extensive in-house testing of sample preservation and shipping requirements. Overnight shipment at 4°C combined with immediate DNA extraction upon sample receipt at the laboratory minimizes changes to the microbial community.

NGS analysis can be performed on nearly any sample type including groundwater, soil, sediments, and Bio-Traps®. Groundwater samples can be submitted using 1 L poly bottles or using Bio-Flo filters (Figure 3). Bio-Flo filters can be readily attached to ¼ inch tubing and are compatible with low-flow purging/sampling pumps. For more detailed information on sample collection, complete protocols are available on the sampling page of the MI website (http://www.microbe.com/sampling-census/).

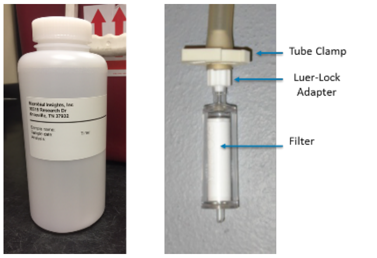


*Figure 5: Groundwater samples can be collected in 1L poly bottles or using Bio-Flo filters*

## 4.2 Selecting Sampling Locations

The number and locations of selected samples for NGS analysis depend upon the size of the impacted area and the variability in subsurface conditions across the site. For NGS, MI recommends collecting samples from monitoring wells representing distinct areas of the site including:

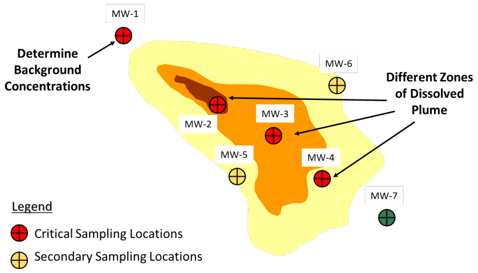

MW-1
Determine Background Concentrations
MW-6
Different Zones of Dissolved Plume
MW-2
MW-3
MW-5
MW-4
MW-7
Legend
⊕ Critical Sampling Locations
⊕ Secondary Sampling Locations

*Figure 6: Selecting Sampling Locations*

- A non-impacted, upgradient monitoring well to determine background microbial populations (Figure 6, MW-1);
- At least one source area monitoring well (MW-2); and
- Monitoring wells within the dissolved plume with substantially different contaminant concentrations or geochemical conditions (MW-3 and MW-4).

Analysis of a sample from an upgradient, non-impacted monitoring well is particularly important. As detailed in the following case studies below, NGS data interpretation emphasizes comparisons between results for background and impacted samples to determine whether contaminant-degrading microorganisms are enriched and growing in the dissolved plume.

# 4.0 Case Study – MNA Assessment at a Petroleum-Impacted Site

## 4.1 Site Background

This site was a pipeline release in an agricultural area. Impacted soils near the release were excavated and the site was undergoing MNA at the time of the study. Total BTEX was observed at concentrations of 4 to 7 mg/L. BTEX concentrations appeared to be stable or decreasing. Geochemical monitoring indicated notably lower sulfate concentrations in the impacted areas, higher dissolved methane concentrations than background, and higher alkalinity in the impacted wells. While indirect, the consumption of electron acceptors and increase in alkalinity provide a supporting line of evidence of microbial activity within the dissolved plume. The consumption of electron acceptors within the plume was a good general indicator of microbial activity, but that may not translate to biodegradation of the contaminants of concern, particularly benzene. There was high variability in MNA parameters potentially due to agricultural activities from neighboring farms as well as the variable conditions that are commonplace in an area that could be described as a wetland environment.

## 4.2 Site Management Questions and Study Design

To better examine the feasibility and performance of monitored natural attenuation (MNA), the site managers decided to examine the microbiology as a third, and more direct, line of evidence. The sampling strategy occurred over two years and included both background and monitoring wells. Specifically, quantitative PCR (qPCR) was used to quantify genes involved in anaerobic BTEX biodegradation, and NGS was used to evaluate overall differences and changes in the microbial community. Both molecular tools can be used to address a variety of distinct site-specific questions.

**Using qPCR to address site-specific questions:**

- Are BTEX degraders present at substantial concentrations under existing conditions?
- What are the concentrations of BTEX degraders in impacted areas?
- Is BTEX biodegradation a likely component of MNA at this site?

**Using NGS to address site-specific questions:**

- Are there differences in microbial community composition in the background versus plume?
- Did community composition and structure change over time?
- Were there specific microbial genera that were dominant in the community samples?
- Did one genus outperform another genus at an impacted area?

To answer the above questions and to assess the feasibility of MNA, qPCR and NGS analyses were performed on groundwater samples obtained from two background wells, MW-6 & MW-7, and three impacted monitoring wells, MW-8, MW-9, & MW-10. Groundwater samples were obtained from each monitoring well approximately quarterly for a year and a half (6 sampling events).

## 4.3 Assessment of Microbial Community Differences in the Background Versus Plume

Principal Component Analysis (PCoA) is one of the most efficient ways to visualize NGS data and to make meaningful observations based on the plethora of sequencing data. Principal coordinate analysis was performed for the five wells over the six sampling events (Fig. 7). Based on that analysis, we observed a strong grouping between two wells: MW-6 and MW-7. The three other wells, MW-8, MW-9, and MW-10, however, displayed another grouping pattern. After looking at site data, we know that the wells on the right are unimpacted (background) wells, and the wells to the left are contaminated. Thus, the overall microbial communities of the impacted wells were different from the background microbial populations, suggesting that petroleum hydrocarbons exerted a selective pressure on the microbial community.

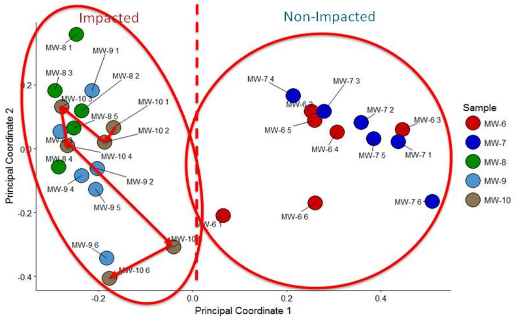


*Figure 7: Principal Coordinate Analysis of monitoring wells at all time points showed changes in microbial community structure.*

Using the PCoA analysis, changes in a single well over time can also be tracked. Figure 7 also shows temporal changes in the resident microbial community at well MW-10 from the first sampling event (Fig. 7, MW-10 1) to the 2nd, 3rd, 4th, 5th, and final sampling events (MW-10 6). Temporal changes for the other impacted wells were less dramatic but still evident. Overall, PCoA analysis of NGS data suggested that the presence of petroleum hydrocarbons was not the only factor influencing the microbial communities of the impacted wells. These communities were unstable and definitely changed over time, particularly during the last two sampling events, suggesting that those changing environmental conditions did affect the overall microbial community.

## 4.4 Assessment of Microbial Community Composition in Petroleum-Impacted Wells

Another way to visualize the shifts in microbial communities is through dissimilar hierarchical clustering. Figure 8 presents a dendrogram representing dissimilar hierarchical clustering of the NGS data obtained at this site. Well samples with more similar microbial communities are on the same branch, while samples with more dissimilar microbial communities are placed on different branches. Similar to PCoA, we see two distinct clusters—background wells (MW-6 and MW-7) on one large branch and contaminant-impacted wells on a separate large branch (MW-8, MW-9 and MW-10).

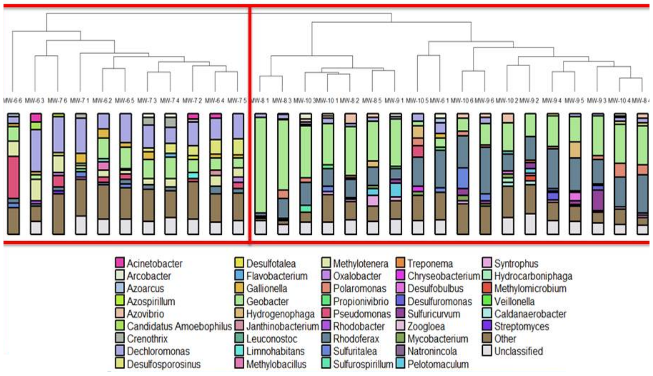


*Figure 8: Hierarchical Clustering demonstrates differences in microbial community composition between monitoring wells.*

In Figure 9, we can see the stacked bar charts representing the microbial community composition over time at impacted monitoring well MW-10. The first sampling event is at the top through the sixth sampling event at the bottom. *Geobacter* communities, indicated by the red arrows, were a major portion of the microbial community during the first four sampling events, but decreased substantially during the last two events. *Rhodoferax* populations, indicated by blue arrows, increased, and this trend proved true for the other impacted samples. *Geobacter* and some strains of *Rhodoferax* are iron reducing bacteria and are viewed as competitors to one another so this shift in the microbial population is interesting.

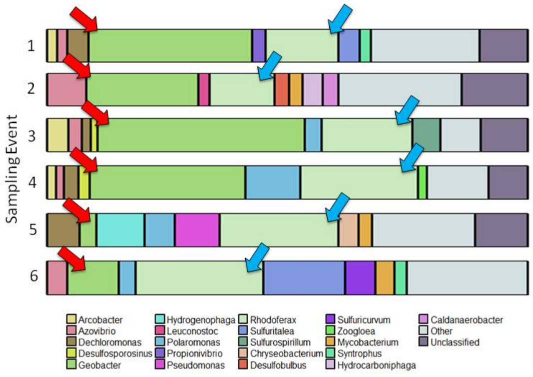


*Figure 9: Microbial community changes in well MW-10 over time.*

Looking at the total percent concentrations of *Rhodoferax* and *Geobacter* in all sampling wells over all sampling events, the top two wells are unimpacted and the bottom row of charts are the impacted wells. In the background wells, the populations of *Geobacter* and *Rhodoferax* were low and remained relatively constant. Conversely, iron-reducing *Geobacter* and *Rhodoferax* populations were more substantial in the impacted wells. The NGS data provided evidence of a trend in selection of iron reducers in impacted wells where hydrocarbon degradation had led to consumption of oxygen and therefore, the generation of anaerobic conditions. Looking at the bottom row of charts with the impacted wells, we see that MW-8 had the largest decline of *Geobacter* as *Rhodoferax* became more prevalent. In MW-9 and MW-10, *Geobacter* was eventually outcompeted by *Rhodoferax.* With nearby agricultural activities and a wet environment, fertilizer and other amendments may have caused spikes in available nutrients and electron acceptors. *Geobacter* are best known as iron reducers. Some species of *Rhodoferax*, which may be reclassified as *Albidoferax* in the near future, are also iron reducers, but others can use additional electron acceptors like oxygen and nitrate. If there was an influx of these more energetically favorable electron acceptors, that could have given *Rhodoferax* populations a competitive advantage over *Geobacter* species. Nitrogen availability could also have an impact. *Rhodoferax* cannot fix nitrogen, whereas *Geobacter* spp. generally can. If nitrogen availability was limited during the early sampling events, *Geobacter* would have had a competitive advantage. An influx of nutrients during the later sampling events would have eliminated that advantage. Variability in the subsurface conditions likely had a secondary impact on the overall microbial community at this site.

## 4.5 qPCR Quantification of Anaerobic BTEX Degraders

While NGS provided insight into changes in the microbial community, quantitative PCR (qPCR) was used to quantify specific functional genes involved in anaerobic BTEX biodegradation to assess the potential for anaerobic BTEX biodegradation under existing site conditions (Fig. 10). The benzylsuccinate synthase gene (BSS, blue bars) encodes the enzyme responsible for initiating anaerobic biodegradation of toluene and other alkyl-substituted benzenes. The ABC assay (red bars) targets the gene encoding anaerobic benzene

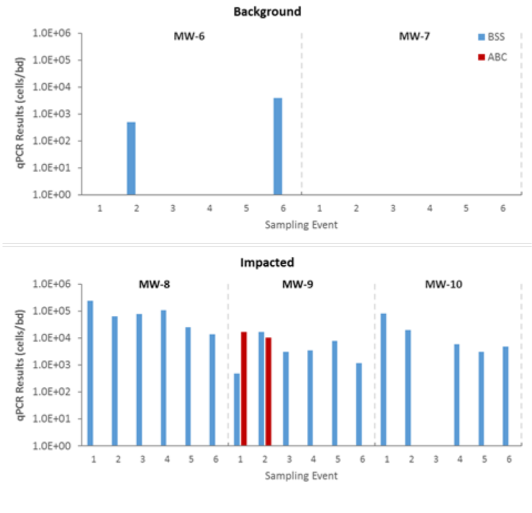


*Figure 10: qPCR Quantification of Functional Genes involved in Anaerobic BTEX Biodegradation.*

carboxylase, which catalyzes the first step in the only characterized pathway for anaerobic biodegradation of benzene. As shown in Figure 10, concentrations of BSS genes were typically below detection limits in groundwater samples from background wells MW-6 and MW-7. In impacted wells MW-8, MW-9, and MW-10, however, BSS genes were routinely detected at relatively high concentrations (~$10^4$ cells/mL).  While not detected in all impacted samples, ABC was detected at monitoring well MW-9 confirming the presence of bacteria capable of anaerobic benzene biodegradation. Thus, the qPCR results demonstrated growth of anaerobic BTEX degraders within the dissolved plume providing a strong line of evidence for the feasibility of MNA at the site.

### 4.6 Conclusions

In summary, NGS and qPCR analyses complement one another. NGS provided broad insights into microbial community dynamics at the impacted site, whereas qPCR analysis, being much more specific, provided another line of evidence for biodegradation which could not have been derived from the NGS results. NGS revealed that the microbial community composition of the samples from impacted wells were very different than the background populations, and overall microbial diversity was lower in the impacted wells. With the depletion of oxygen, there were higher proportions of anaerobes like *Geobacte*r in the wells. NGS also revealed population shifts over time as *Rhodoferax* appeared to outcompete *Geobacte*r during the later sampling events potentially due to the changing electron donor and nutrient availability. The qPCR results demonstrated growth of high concentrations of anaerobic BTEX degraders within the dissolved plume and provided a strong supporting line of evidence for MNA as a site management strategy.

**NGS and qPCR are highly complementary molecular biological tools (MBTs)**

**NGS Conclusions**

- The microbial communities of impacted wells were very different than background populations.
- With oxygen depletion, proportions of anaerobes like *Geobacter* were higher in impacted wells.
- Microbial populations shifted over time as *Rhodoferax* appeared to outcompete *Geobacter* potentially due to changes in electron donor and nutrient availability.

**qPCR Conclusions**

- Concentrations of BSS and ABC, functional genes involved in anaerobic BTEX biodegradation, were substantially higher in the impacted wells than in background wells.
- Enrichment and growth of anaerobic BTEX degraders within the dissolved plume under existing site conditions strongly suggested that biodegradation was a component of MNA.

## 5.0 Quality Assurance/Quality Control (QA/QC) Parameters

For more than 25 years, the primary mission at Microbial Insights (MI) has been to provide the most accurate and precise data in the industry to ensure that our clients can use our results as an integral part of site management decisions.

The accuracy of MI's data is attributed not only to the quality of our assays and continued investment in instrumentation but also the experience of our staff and rigorous QA/QC procedures that are second to none.

- **Date of Extraction:** DNA and RNA extractions are performed the day that the samples are received by MI to minimize the possibility of any changes to the microbial community prior to analysis.
- **Extraction Blanks:** An extraction blank (no sample added) is processed alongside each set of field samples from DNA extraction through analysis to ensure that cross contamination has not occurred.
- **Negative Controls:** A negative control (no DNA) is included to ensure that cross contamination has not occurred.

## 6.0 References

1. Shendure, J, & Ji, H. Next-generation DNA sequencing. *Nature Biotechnology* 26, 1135–1145 (2008).
2. Gotelli, N. J. & Colwell, R. K. Quantifying biodiversity: Procedures and pitfalls in the measurement and composition of species richness. *Ecology Letters* **4,** 379–391 (2001).
3. Hill, M. O. Diversity and evenness: A unifying notation and its consequences. *Ecology* **54,** 427–432 (1973).
4. Buttigieg, P. L. & Ramette, A. A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses. *FEMS Microbiology Ecology* **90,** 543–550 (2014).