A NEW APPROACH FOR SIMULTANEOUS DNA-BASED MONITORING OF THE POLLUTED ENVIRONMENTS

By Shahrokh Shekarriz,

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University
Master of Science (2016)
Hamilton, Ontario (Department of Biology)

TITLE: a new approach for simultaneous dna-based monitoring of the polluted environments
AUTHOR: Shahrokh Shekarriz (McMaster University)
SUPERVISOR: Dr. Brian Golding
NUMBER OF PAGES: x, 75

Abstract

Taxon composition and biodiversity analyses are known powerful parameters for environmental site status and environment diagnosis. Many ecological studies assess taxon composition through traditional species identification and use bioindicator species to evaluate environmental conditions. The recent breakthrough in bulk sample sequencing combined with DNA barcoding has created a new era for environmental monitoring. Metabarcoding approaches are more robust in studying alpha, and beta diversity compare to the DNA barcoding and the conventional method of species identification, particularly for rare and cryptic species. Here we built upon ecological studies of bioindicator species and transferred the traditionally named taxa to DNA-based approaches. We developed a small customized DNA database for biodiversity assessment and taxonomic identification of environmental DNA samples using high-throughput amplicon sequences. It contains macroinvertebrate species that are known as indicators of specific environmental conditions. By implementing this small database into the KRAKEN algorithm for the first time, we were able to assess environmental biodiversity compared to other popular methods of taxonomic classification, especially in polluted environments where the taxonomic composition globally change by the presence of anthropogenic drivers. Our method is incredibly faster, and it requires significantly less computational power in contrast to common homology-based techniques. To evaluate our approach, we have also studied the importance of database's size and the depth of sequencing in taxonomic classification of high-throughput DNA sequences.

Acknowledgements

I would like to begin by express sincere gratitude to my advisor Prof. Brian Golding for his patience, guidance, and immense knowledge. The door to Dr. Golding's office was always open when I had questions. He has taught me to think critically not only about the scientific studies but also on a personal level. He has given me opportunities that broadened my understanding of computational biology and science as a whole. I would also like to extend my gratitude to the member of my advisory committee – Dr. Jonathan Stone, for his guidance that widened my research and writing this thesis.

I would also like to thank the organizations that contributed to this project; Environment Canada, Park Canada, and McMaster University. I would like to extend my deepest thank to Dr. Mehrdad Hajibabaei and his lab members that provided us the invaluable data and support. And many thanks to my wonderful labmates who helped through my whole master project.

Finally, I can never repay the debt owned to my cousin, Mehrdad, for his support over my entire educational career and beyond. My family supported me from the beginning, and they have the bigger impact that they ever know.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Shahrokh SHEKARRIZ, declare that this thesis titled, “a new approach for simultaneous dna-based monitoring of the polluted environments” and the work presented in it are my own.

Chapter 1

Introduction

1.1 Biodiversity analyses

Human activities as a result of economic, cultural and intellectual goals are making significant global environment changes. A list of mechanisms includes; elevated CO2 and other greenhouse gasses, increased nutrient loading and excessive water consumption, different types of intense land usage and finally human-facilitated species invasions are all causing global biodiversity changes in our environment. Biodiversity alteration can lead to natural species invasion and cause massive shifts in the ecosystem. Changes in ecosystem services mediated by altered species traits, biodiversity and ecosystem processes influence the services that benefit humanity (Chapin III et al., 2000) (Fig. 1.1). Biodiversity as a measure that quickly respond to the global environment change (Fig. 1.1), has essential components that are helpful to study the global environment change mechanisms. These components include; species richness (number of species), species evenness (relative abundance), species composition (particular species), non-additive effects (interaction among species), and the temporal and spatial variation.

Species richness can not reflect the changes in ecosystem functioning as a global pattern. For example, increased microbial richness elevates the decomposition of organic matter (Salonius, 1981), while there seems to be no relationship between plant species richness and the decomposition rate(Wardle, Bonner, and Nicholson, 1997). In contrast to species richness, species evenness requires more attention as it does respond to human activities very quickly (Chapin III et al., 2000). Particular species in an ecosystem (species composition) actively regulate the ecosystem process by moderating energy and material fluxes or by adjusting abiotic condition (Tilman et al., 1997). For example, earthworms and termites, soil invertebrates, have an effect on the species composition of the aboveground vegetation and wildlife by modifying the turnover of organic matter and nutrition supply (Lavelle et al., 1997). Species interactions comprise; competition, trophic and mutualism can also impact the ecosystem processes through direct and indirect modification. It can directly alter pathways of energy and material flow (Ruiter, Neutel, and Moore, 1995) or indirectly refine the abundances of bioindicator species.(Power et al., 1996). Also, based on the global-diversity hypothesis, diversity can minimize the chance of global ecosystem alteration caused by environment shifts

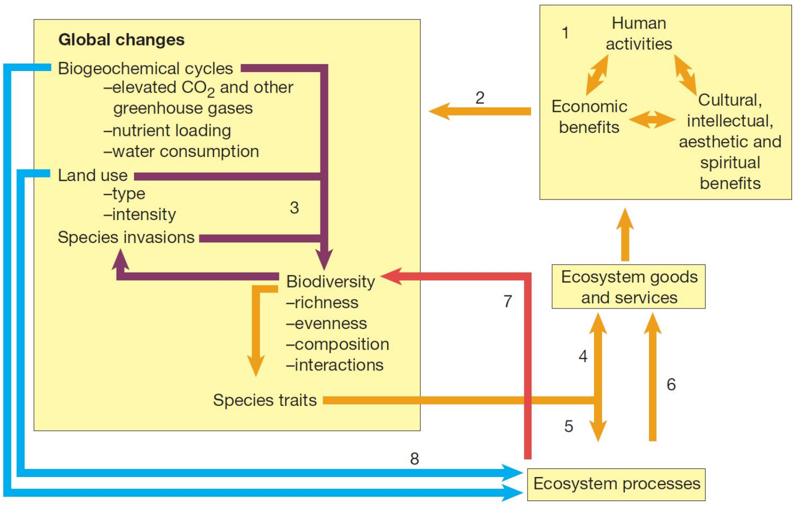


FIGURE 1.1: The role of biodiversity in global environment change from Chapin III et al. (2000)

(McNaughton, 1977). Diversity makes an ecosystem resilient to change as the probability of maintaining the current properties will increase by the presence of larger number of functionally similar species (Chapin III and Shaver, 1985). For instance, Microbial communities with higher species richness are less responsive to ecosystem processes (Naeem and Li, 1997).

1.2 Macroinvertebrates as indicators of environments

A bioindicator can be defined as “a species or group of species that readily reflects the abiotic or biotic state of an environment, represents the impact of environmental change on a habitat, community, or ecosystem, or is indicative of the diversity of a subset of taxa, or of the wholesale diversity, within an area” (McGEOCH, 1998). Macroinvertebrates are potential bioindicators for assessing environments because they have: abundant medium-sized bodies, medium growth rate, moderate population turnover (not as high as microorganisms and not as low as higher plants or animals), active and passive dispersal mechanism indicative of changes in ecosystem function(Hodkinson and Jackson, 2005). Changes in these organisms can indicate: a changing physical environment, a changed chemical environment, specifically with different sort of pollution, the comparable quality or conservation value of habitat, which refers to the importance of ecological communities and shifts in the ecological status of the habitat with respect to time and place (Hodkinson and Jackson, 2005).

Thus far, the role of many invertebrates as potential bioindicators has been well-documented through ecological studies (Paoletti, 2012). Although the indication role of invertebrates in terrestrial habitats is less developed than is their usage in aquatic environments (Bonada et al., 2006), terrestrial macro-invertebrates possess the same bioindication characteristics as marine invertebrates (Rosenberg, Danks, and Lehmkuhl, 1986). For example, benthic macroinvertebrate taxa can tolerate pollution differently, and their different responses were used for water assessment. Some of them, like larvae in the orders Plecoptera and Ephemeroptera, were shown to be pollution intolerant. While other species such as tubificid worms and chironomid midges survived under deoxygenated conditions related to anoxia (Rosenberg, Resh, et al., 1993). Invertebrate bioindicators demonstrate three levels of responses including individual animal level, species population-level, and community-level responses (Hodkinson and Jackson, 2005). Choosing a suitable level of responses based on the purpose of a study is an important aspect that needs to be adequately addressed. Individual animals can illustrate single environmental pollutants while the whole invertebrate community can highlight more general values like conservation or continuing forest degradation (McGEOCH, 1998).

Individual animal-level responses focus more on the physiology and behavior of a single animal to environmental stress, which are known as short-term bioindicators. For instance, the presence of pollutions like fertilizers and heavy metals have caused fluctuating asymmetry in some groups of invertebrates (Clarke, 1993; Tessier et al., 2000). Fluctuating asymmetry (FA) is a morphological phenomenon that causes higher levels of asymmetry in body shape and size (Parsons, 1992). The population responses at the species-level are the simplest way to interpret the role of invertebrates as bioindicators because this level of responses does not contain a diverse population of species (community-level) or variable factors that cause physiological reactions in an individual (individual animal level). For example, the mortality of soil invertebrates specifically *Lumbricus spp.*, collembola, snails and isopod species can be affected by the presence of soil contamination including heavy metals and pH (Cortet et al., 1999; Van Straalen, 1998). Similarly, marine macroinvertebrates such as amphipods (Duan et al., 2000), caddisflies, and mayflies (Benton and Guttman, 1990) exhibit significant species population level responses to heavy metal exposures.

The community-level responses are more complicated than the other levels, however they provide a valuable approach for biological monitoring as they integrate multiple species to evaluate multi-source environmental effects. The community level utilizes species richness, evenness, and taxonomic composition to assess environmental changes in a particular site. Although this level of responses provides realistic ecological responses, it suffers from poor accuracy in traditional taxonomic identification. For example, a group of benthic macroinvertebrates includes Ephemeroptera, Plecoptera, and Trichoptera is known as indicators of river pollutions and their richness evaluated as a valuable index for running water ecosystems (Rosenberg, Resh, et al., 1993). But, this subset of taxa does not include Oligochaete worms and chironomid midges, pollution-sensitive groups, due to their challenging taxonomic identification.(Hodkinson and Jackson, 2005)

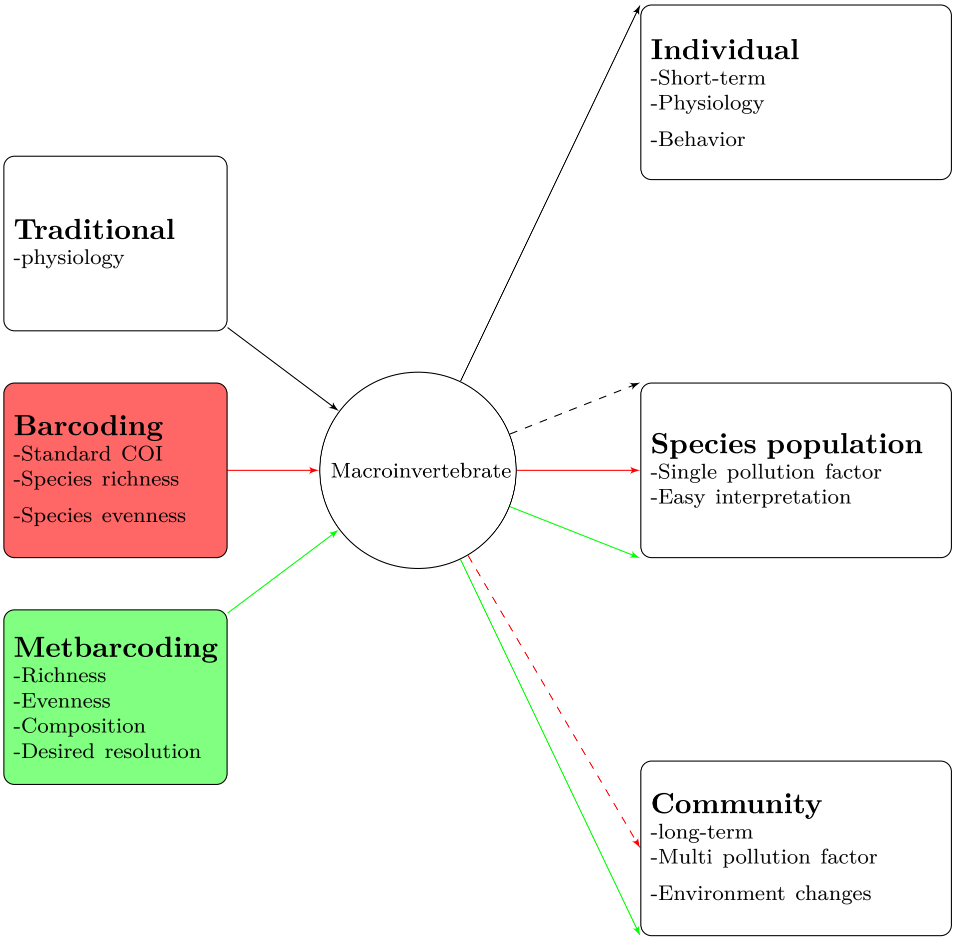


FIGURE 1.2: The three primary methods of assessing environmental eukaryotes and the level of ecological responses that they can produce using macroinvertebrate bioindicators. The solid lines represent the best level of response that a method generates and the dash lines are potential other usages of a method.

There are two main challenges with traditional environment monitoring. Taxonomic identification, the level of species responses, and the complexity of integrated responses of multiple species population in the community-level analysis. Using conventional monitoring methods requires a high degree of expertise and keys for species identification, as only a few of them are well-characterized. Identification errors are more frequent at the species-level rather than the family-level and some morphologically immature species, or cryptic species are impossible to identify based on standard keys. As a result, a higher level of taxonomy has been used for monitoring programs in the past (Jones, 2008; Sweeney et al., 2011). Although a higher taxonomic level might be useful for broad scale impacts in an environment, the minor or small effects, which are essential in many studies will be ignored (Hewlett, 2000). As different species or genera might have variable responses within a family, site diagnostics for impacting factors are more complex in higher level identification (Pettigrove and Hoffmann, 2005). Biodiversity assessments at the community-level using conventional methods are extremely labor-intensive and suffer regarding accuracy. Taxonomists use a subset of easy sampling bioindicators or their own specialist group due to the significant number of taxa involved at this level of study. Consequently, other taxonomically challenging groups of bioindicators that represent environment's effects are excluded from their analysis (Rosenberg, Resh, et al., 1993).

With the proposal of DNA barcoding (Hebert, Cywinska, Ball, et al., 2003) and the introduction of environmental barcoding (metabarcoding) (Hajibabaei et al., 2011), it's possible to choose different levels of study and many invertebrates responses to evaluate the environmental changes (Fig. 1.2). It's important to fully understand the limitation and strength of each method and choose the appropriate approach based on the ecological purposes of a project. Figure 1.2 Illustrate the three methods of environmental monitoring using macroinvertebrates and different level of responses that each method can potentially produce. The traditional way of assessing macroinvertebrates is suitable for studying individual responses, especially for physiological and behavior reactions of an individual to ecological effects (Merritt and Cummins, 1996). Although some authors used conventional monitoring approaches for population density evaluation at species-level (Hodkinson and Bird, 1998; Frati, Fanciulli, and Posthuma, 1992), this method suffers regarding accuracy and is extremely labour-intensive.

DNA barcoding (1.3) provides a reliable and fast approach for assessing the species population-level responses, but it's costly and sometimes not feasible to use this method for community level analysis. This method uses richness and evenness attributes to evaluate environmental condition at species-level. Metabarcoding (1.4) employs high-throughput sequences (HTS) to analyze bulk environmental eukaryotes samples. Metabarcoding utilizes richness, evenness, and composition at the different taxonomic level to predict long-term and multi-factor environmental changes at the community level. Also, it's practical to make use of this approach at species-population level using enormous available databases and proper markers. Taberlet et al. (2012) reviewed the benefits of DNA-based methods for selecting a different level of responses based on choosing the right marker and methodology.

1.3 DNA Barcoding

Since 2003 (Hebert, Ratnasingham, and Waard, 2003), a campaign was started to overcome the limitation of traditional taxonomic identification. Researchers used gene variants, mainly mitochondrial genes, to assign a barcode for each species. The main advancement was a gradual increase in including a broad range of organisms and developing barcodes for different organisms (Hebert and Gregory, 2005). Although some groups of organisms are still waiting to be added to the campaign, the taxonomic method of inquiry for a wide range of organisms was a significant advancement(Hajibabaei et al., 2007). The standardized DNA sequencing approach in taxonomic studies leads to a standard routine protocol for different organizations and agencies in environmental monitoring. The originally proposed DNA barcode, a 650bp fragment of mitochondrial cytochrome c oxidase one gene (COl), was shown to be successful for over 90 percent of species due to the presence of unique COl barcode sequences (Hajibabaei et al., 2007). Fragment sequences of two genes (rbcl and matK) have been chosen as the main plant DNA barcode (Hollingsworth et al., 2009). Internal transcribed spacers (ITS) and 18SrDNA have been designated as fungi and protist barcodes (Schoch et al., 2012). After assembling a reference DNA barcode library from well-characterised species, identification of newly obtained or preserved species by comparison to a reference database is a routine procedure.(Hajibabaei et al., 2007)

1.4 Metagenomics and metabarcoding

New advancements in high-throughput sequencing (HTS) increased the speed and volume of sample sequencing while simultaneously decreasing the cost of sequencing(Mardis, 2008; Margulies et al., 2005). Previously, traditional methods in studying taxonomic composition and environmental monitoring only focused on single specimens. With the advent of new technologies, it's possible to explore the taxonomic composition and phylogenetic relationships of bioindicator communities in the environment through comparative analysis of mixtures of DNA (Handelsman et al., 1998). The HTS applications accelerated the process of sample preparation, taxon identification and environmental monitoring.

Previously, metagenomics was known as a culture or clone independent approach for studying only prokaryote communities. The 16S ribosomal gene region has been utilized extensively for microbial diversity assessments (Lee et al., 2012), the prokaryote communities associated with human body (Caporaso et al., 2011), the evaluation of prokaryotic biodiversity in freshwater (Luo et al., 2012) and soil (Lauber et al., 2009). More recently, researchers used this approach for environmental and community biodiversity analysis of eukaryotes in bulk samples (Drummond et al., 2015; Gibson et al., 2014). DNA metabarcoding consists of high-throughput amplicon sequences of cytochrome oxidase subunit one were shown to be successful in biodiversity assessments of environmental samples that contain mixtures of eukaryotes DNA (Hajibabaei et al., 2011). The availability of a standard DNA library depends on the type of marker; for example, Cytochrome c oxidase 1 has a massive standard DNA library, which can be used for species-level identification of samples (Taberlet et al., 2012).

The amplicon metabarcoding primarily used 454 sequencing platform. The 454 sequencing (pyrosequencing) produces a long read length, however, has some disadvantages including fewer sequences compared to other platforms, a higher number of sequencing errors and, increased cost of reagents compared to other methods (Luo et al., 2012). Currently, the sequencing platform that researchers mostly use for amplicon-based HTS is Illumina MISEQ or HISEQ. Because the COI barcode region is too long for Illumina platforms, authors have developed modified primers covering a shorter region of COI barcode (Leray et al., 2013; Hajibabaei et al., 2011). There are two other sequencing techniques that don't use fragment amplification (PCR) to characterize species communities. These methods include shotgun sequencing (Simon and Daniel, 2011) and gene enrichment techniques (Dowle et al., 2015). But, both methods are currently too expensive for rapid identification of bulk environmental samples.

1.5 DNA-based biomonitoring

Traditional biomonitoring approaches focus on periodic local-scale sampling followed by an extended period of data processing, which often ends with unverified data in terms of taxonomic precision. Typical output from these type of studies is a binary result, which can only report impacted or not impacted environments (Baird and Hajibabaei, 2012). In monitoring studies, the roles and functions of species providing species level information should be considered, specifically in making conservation decisions. More recently, the advent of new tools in DNA sequencing and DNA-based approaches in taxon identification carved a new path towards assessment of complex multi-stressor scenarios and represented the future diagnosis for a given environmental condition (Hajibabaei et al., 2011; Carew et al., 2013). High-throughput data generated by amplicon sequencing of bulk samples linked to DNA barcode libraries create a better understanding of environment biodiversity. This new technique is promising to extract precious biodiversity data from bulk samples. Moreover, the data generated in such studies contain OTUs, which create new opportunities for data analysis and taxonomic assignments. "Biomonitoring 2.0" (Baird and Hajibabaei, 2012) has been proposed as an alternative scheme for ecosystem monitoring , especially in a multi-stressor environment. Although much work needs to be done to overcome the current challenges in this approach, it seems to have a potential for biomonitoring studies in future.

1.5.1 Challenges

While shifting from individual sampling to DNA-mixture sampling was an important step forward, it requires careful consideration of data interpretation specifically in terms of reliability and repeatability of taxonomic assignment. One possible problem to deal with is DNA that can persist beyond the lifespan of an individual (Dejean et al., 2011). Another challenge is to address mixed data sources comprising known, named taxa and OTUs in the development of diagnostic indices. Also, the establishment of a standard approach to collect, preserve and subsequently analyse field samples in a manner that would be compatible with current and future DNA analysis methods. Finally, there has been a move toward generating information on the relative abundance of the organism in biomonitoring samples, but it may cause some false results due to PCR amplification bias (Polz and Cavanaugh, 1998; Baird and Hajibabaei, 2012).

1.6 Taxonomic classification algorithms

There are three main supervised learning approaches (methods that compare query sequence to databases for taxon assignment): similarity search (use homology or alignment based methods; e.g., BLAST (Altschul et al., 1997)), composition methods (use k-mer counts or frequencies; e.g., LMAT (Ames et al., 2013), KRAKEN (Wood and Salzberg, 2014), RDP (Wang et al., 2007)) and phylogenetic methods (use evolutionary models and homology-based methods; e.g., FastTree (Price, Dehal, and Arkin, 2009), EPA (Berger, Krompass, and Stamatakis, 2011)) (Bazinet and Cummings, 2012).

Almost all taxonomic assignments in current metabarcoding studies employ a similarity search. They use a combination of BLAST (most commonly BLASTn) and different versions of the lowest common ancestor (LCA) algorithm, e.g., MEGAN (Huson et al., 2007). Although this approach has substantial accuracy even for short query sequences, it searches each query sequence against a gigantic database (mostly non-related organism) which can take a long time. Given the expansion of the GenBank database, each BLAST job will take more and more time in the future (Fig. 1.3). Furthermore, it suffers from poor assignment accuracy when a match for the query sequence is not in the database, specifically compared to other classification methods.

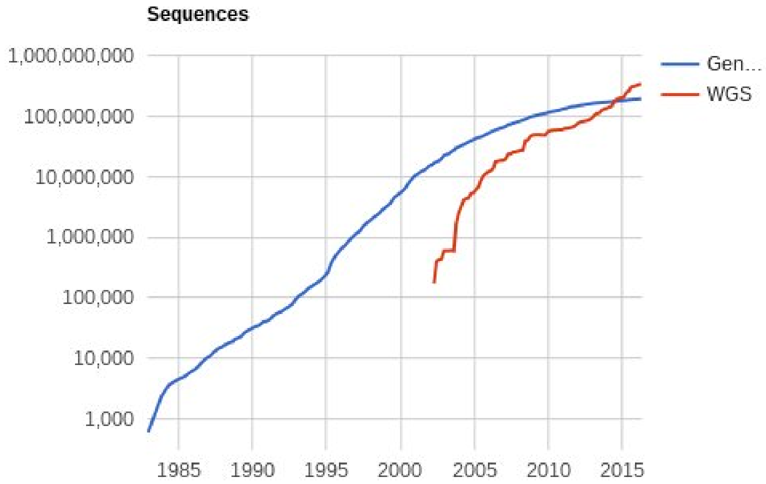


FIGURE 1.3: The GenBank nucleotide database is growing exponentially. This figure represents the growth of GenBank sequences (1985 - 2015) (NCBI, 2016)

The phylogenetic approach for taxonomic classification employs evolutionary models utilizing maximum likelihood, neighbor-joining, or Bayesian methods to calculate the suitable place of a query sequence on a phylogenetic tree (Bazinet and Cummings, 2012). These programs use the simple observation to find where an inserted branch divergent from a node that represents a species or higher rank. This approach requires enormous computational power as it contains multiple alignments, fixed topology (e.g., NCBI taxonomy), and the insertion of a query sequence into the reference alignment (e.g., GreenGenes database; DeSantis et al. (2006)).

The other popular approach is a compositional model. There are three types of algorithms in compositional models (Bazinet and Cummings, 2012) including the Naive Bayesian classifiers (Porter et al., 2014), interpolated Markov models (IMMs), and kmer/k-nearest-neighbor algorithms (Ames et al., 2013). The main advantage of this approach is faster query sequence classification compared to alignment-based methods. As an example, KRAKEN developers compared the speed of different classification algorithms and reported their program as 150 to 240 times faster than the closest competitor (Megablast)(Wood and Salzberg, 2014). KRAKEN processes data at a rate of over 1.5 million reads per minute (rpm) while Megablast classified at 7,143rpm for a Hiseq metagenome (an Illumina sequencing system designed for production-scale genome with maximum 1500 GB output). KRAKEN analyzed data over 1.3 million rpm however Megablast had a speed of 2,830 rpm for a Miseq metagenome (an Illumina sequencing machine designed to target small genome and amplicon with maximum 15 GB output) (Wood and Salzberg, 2014). But, the biggest problem with this approach is the computational power, as the pre-computed databases of these programs need to be downloaded locally before analyzing a particular dataset. This issue may be the case for users with limited computer disc space.

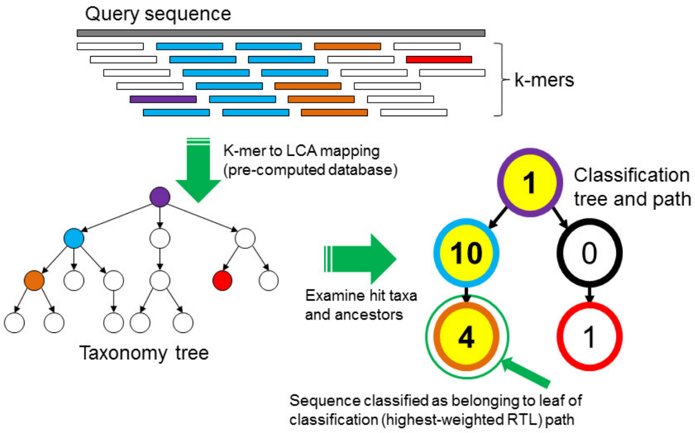


FIGURE 1.4: The KRAKEN sequence classification algorithm (Wood and Salzberg, 2014)

A sequence model of database for each group of target organisms in kmer-based method needs to be computed separately. For example, LMAT and KRAKEN create different sizes of standard databases for 16S ribosomal RNA sequences.They both have a database that contains k-mers and the LCA of all organisms that contain that k-mer. Query sequences will be classified by searching the database for each k-mer in a query sequence. Then, using the result of LCA taxa in the pre-computed database, they assign an appropriate label for the sequence. Sequences that have no kmers in the database are left unclassified (Fig. 1.4). LMAT uses a 17-20 value for "k" and each k-mer is mapped to the individual source genomes minus a non-redundant search of taxonomic identifiers associated with the k-mers, unlike alignment-based methods (Ames et al., 2013). KRAKEN builds the database with k = 31 by default, but users can change this value. KRAKEN's authors also performed a speed comparison against LMAT using one of the samples discussed in the LMAT published results. KRAKEN was 38.82 faster than LMAT and 7.55 times faster than a version of LMAT using a smaller database.

There is a trade-off between precision and sensitivity in taxonomic classification. The accuracy of an assignment can increase with some cost to sensitivity. For instance, NBC (Rosen et al., 2008) and PhymmBL (Brady and Salzberg, 2009) algorithms label all the reads as accurately as possible. Because there is not adequate information to label some of the reads, the algorithms that label all the reads cause higher false positive in classification. While KRAKEN (Wood and Salzberg, 2014) leaves a sequence unclassified if not enough evidence exists. Wood and Salzberg (2014) compared the performance of KRAKEN and three other traditional classification algorithms; NBC, PhymmBL and Megablast for identical query sequences and they reported that KRAKEN yields a higher precision at the cost of lower sensitivity. But, a user can customize the accuracy value in KRAKEN that helps to increase the efficiency of classification based on the objective of a study. KRAKEN's authors also concluded that their program classified the reads very similar to Megablast. These two programs were 97.5 % similar in terms of sensitivity and KRAKEN was slightly more accurate compared to Megablast (Wood and Salzberg, 2014).

1.7 Project purpose statement

The fundamental goal of this thesis is to introduce a new approach for DNA-based biomonitoring (1.5) of a polluted environment. This project evaluates the role of databases in the biodiversity assessment (1.1) of high-throughput amplicon sequences (HTS) (1.4) utilizing a small customize database. This study provides a remarkably faster approach for biodiversity estimation of eukaryotic communities (1.2) to address one of the main challenges (1.5.1) in DNA-based biomonitoring. We have aimed to accomplish three goals. First, transfer the past ecological knowledge of bioindicator species (1.2) to DNA-based approach by collecting and assembling a small library of unique species as the environmental indicators and compare the performance of such a small database with the currently popular methods for biodiversity evaluation. Second, increase the efficiency and accuracy of DNA-based biomonitoring method for individuals with limited resources in terms of time and computational power. We have employed a highly fast k-mer based algorithm (KRAKEN Wood and Salzberg (2014)) as a k-mer based method for taxonomic classification (1.6) of eukaryotic communities for the first time. And third, investigate the importance of read numbers (depth of sequencing) and the size of a database in calculating alpha diversity indices.

1.7.1 Applicability

"Biomonitoring 2.0" (Baird and Hajibabaei (2012); Fig. 1.5) uses bulk sequenced biological samples to run against a standard taxon library (which includes all known species) to assemble taxon composition. Afterward, all the other factors including taxon stressor-response library and water chemistry should be analyzed. Expectedly, the output is a complex multi-dimensional result which requires deep consideration and analysis to have a robust diagnosis for an environment. Probably there will be lots of non-related organisms in bulk samples, which are not beneficial for biomonitoring purposes, and including these taxa in the site composition only add noise to the data analysis. For instance, Gibson et al. (2015) identified all their sequences utilizing MegaBlast algorithm contain a database of all available COI sequences, then they retrieved only benthic metazoan phyla (i.e., Chordata, Mollusca, Arthropoda, Annelida) from their classified sequences. After that, they generated a subset of the matrix including representatives of Ephemeroptera, Trichoptera, and Odonata to calculate all their biodiversity metrics.

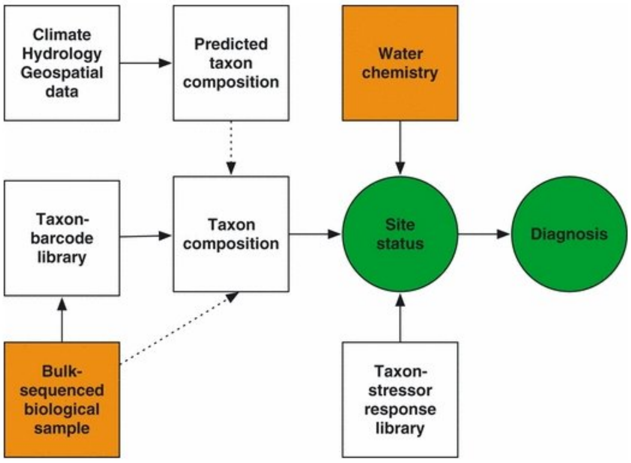

Climate Hydrology Geospatial data
Predicted taxon composition
Water chemistry
Taxon-barcode library
Taxon composition
Site status
Diagnosis
Bulk-sequenced biological sample
Taxon-stressor response library

FIGURE 1.5: Biomonitoring 2.0 schema for environment monitoring (Baird and Hajibabaei, 2012)

To avoid redundant analyses and increase the efficiency of Biomomitoring 2.0, we propose a real-time monitoring program. This method builds on a fully targeted taxon database of invertebrate bioindicator species. We are building upon past knowledge of indicator taxa that have been studied and characterized as indicators of different stressor responses through many years of ecological research. In this model, some post-analysis of bulk samples will be avoided. Furthermore, by choosing a robust assignment algorithm, the efficiency of our approach will be higher than the time-consuming methods like basic local alignment search tool (BLASTn)(Fig. 1.3).

1.7.2 Research questions

1. What is the relationship between the read numbers (depth of sequencing) and alpha diversity indices in a polluted environment?
2. How small can a database be, but still provide sufficient indication of environmental condition?
3. What are the most well-known aquatic and terrestrial macroinvertebrate bioindicators that ecological researchers confirmed their community responses to environment changes?
4. How does KRAKEN algorithm classify the HTS reads compared to other common taxonomic classification program utilizing a small custom database?
5. Is a small database of macroinvertebrate sequences able to distinguish polluted and non-polluted environments regarding taxonomic composition and alpha diversity indices?
6. Is a small library of macroinvertebrate sequences capable of estimating taxonomic composition and other biodiversity metrics in a contaminated environment compare to the gigantic database used by BLAST?

1.7.3 Hypothesis and predictions

1. There is a point in the taxonomic classification of HTS amplicons, where alpha diversity will not increase even by using a higher number of the reads. Presumably, there is a saturation level for taxonomic identification in amplicon-based HTS.
2. In polluted environments due to the low species diversity of environmental samples the size of a required database for taxonomic classification of HTS amplicons significantly decrease.
3. A small database of targeted macroinvertebrate bioindicators can sufficiently estimate biodiversity of a polluted environment compared to the other enormous eukaryotic databases.
4. The KRAKEN program, a kmer-based classifier, improves the DNA-based environmental monitoring by reducing the time and required computational power for the analysis of amplicon-based HTS.

Chapter 2

Essential factors in HTS taxonomic classification

2.1 introduction

High-throughput sequencing (HTS) has revolutionised biodiversity analysis of environmental communities. We have gained an in-depth knowledge of community dynamics, in particular for some morphologically cryptic organism such as viruses, bacteria, fungi and, macroinvertebrates (Angly et al., 2006; Buee et al., 2009; Shokralla et al., 2012). Nevertheless, the data generated by HTS machines are sensitive to the technical procedures used to produce the sequence reads. It is important to understand the structure and type of the generated output by each sequencing method or platform. With this understanding, time and money resources can be better allocated. For example, there are concerns about using HTS counts as an absolute abundance in ecological assessments due to the polymerase chain reactions (PCR) and sequencing biases (Amend, Seifert, and Bruns, 2010). However, bioinformatics tools such as UCLUST (or UPRASE) (Edgar, 2010) addressed this issue by introducing a technique to represent the clustered operational taxonomic unit (OTU) that are much closer to the number of species sampled.

In another study, Smith and Peay (2014) tested the correlation between PCR replicate and calculated indices in α and β diversity. They found no significant relationship between increased number of PCR replicates and the diversity measures. But, pseudo-β-diversity decreased with higher sampling depth. They also reported that the Illumina platform significantly recovered greater richness compared to the 454 sequencing platform. One other critical factor in HTS taxonomic classification is a database of known DNA sequences to label the unknown query DNA sequences correctly. The cytochrome oxidase one (COI) gene is a standard mitochondrial gene marker for the animal identification (Hebert, Ratnasingham, and Waard, 2003). There are relatively large libraries of this gene sequence available both on NCBI (Geer et al., 2009) and BOLD (Ratnasingham and Hebert, 2007) systems.

Although these databases (NCBI & BOLD) are not large enough yet to classify all the unknown sequences accurately, the main current challenge is how to utilize these gigantic databases as efficiently as possible. Given the expansion of these databases (Fig. 1.3) adequate identification of a query DNA sequence from HTS requires allocation of substantial time and computer power sources. One way to address this problem is to assemble a customized smaller library of a targeted group of organisms from these data sources. This is a primary goal of our project.

2.1.1 Objective

In this chapter, we evaluate the importance of read numbers (depth of sequencing) as an essential element in taxonomic classification of high-throughput sequences. We expect to observe a saturation level by increasing read numbers where diversity measures will not be affected even by higher read numbers. Also, building a custom database of invertebrate bioindicator species is one of the primary purposes of my master project. We hypothesise that a small database containing around 2000 bioindicator species can predict α diversity in environments contaminated with anthropogenic pollutants. To evaluate the performance of such a small database, we have decided to construct different size databases with three different taxonomic classification algorithms: KRAKEN (Wood and Salzberg, 2014), LMAT (Ames et al., 2013) and BLASTn (Altschul et al., 1997). To test the performance of these three algorithms independent of their default databases, the exact same genomic libraries were constructed for all of them.

We precede the analysis by retrieving the dataset from Gibbons et al. (2014), a spatiotemporal study of sediments in the Tongue River (Montana, USA), comprising six sites along 134 km of river sampled in both spring and fall for two years. All the samples were collected from the polluted sites with significant anthropogenic drivers including increased pathogenicity, antibiotic metabolism markers, and metabolic signature of coal and coalbed methane. We expect that in such a polluted environment the α diversity will not increase even with a higher number of reads. Also, in polluted environments due to the low species diversity of environmental samples the size of a required database for taxonomic classification of HTS amplicons significantly decreases.

2.1.2 Diversity indices

The diversity index is a quantitative method to measure the taxonomic diversity in a community and takes into account how evenly the taxa are distributed in an ecosystem (Hill, 1973; Jost, 2006). Traditionally the taxonomic diversity refers to the species diversity, but other categorize such as genera, family, haplotype, or operational taxonomic unit (OTU) can be used to calculate the diversity index. The two main indices for biodiversity measurements are The Shannon and Simpson index. The Shannon-Weaver or Shannon index (Shannon, 1949) assumes all the species are represented in a sample, and it gives the same weight to all the species. However, the Simpson index (Simpson,

1949) gives more weight to dominant species. Consequently, few rare species with few representative will not affect the diversity. For example, in a community containing 6 *Chironomus atrella*, 5 *Boyeria grafiana*, 1 *Hydropsyche tenuis*, 3 *Caenis youngi*, and 12 *Ablabesmyia monilis*. The Shannon index (H) is 1.37 and the Simpson index (D) is equal to 0.70. In the formulas below p_i is the proportion of individuals i, and S is the number of individuals ($\sum_{i=1}^{S} p_i = 1$), and b is the base of the logarithm. It is most common to use natural logarithms, but $b = 2$ has theoretical justification. (Oksanen et al., 2007).

$$H = -\sum_{i=1}^{S} p_i \log_b p_i \qquad \text{(Shannon)} \tag{2.1}$$

$$D = 1 - \sum_{i=1}^{S} p_i^2 \qquad \text{(Simpson)} \tag{2.2}$$

2.2 The importance of read numbers

2.2.1 Methods

Gibbons et al. (2014) sampled six locations in Tongue River (Fig. 2.1) each fall and spring over two years. The sites B, E, and S are all downstream of a small town (Ranchester, Birney, and Ashland respectively). All the locations, except site B, are downstream of Decker coal mine and Reservoir Dam. The C and BG sites are close to methane extraction wells. The site B is near Decker coal mine and downstream from irrigated farmland. The W site is also downstream of irrigated farmland. Except for sites BG and B, all duplicate samples in each six sampling sites and the four time points were used for amplicon sequencing. For all 44 samples, the V4 region of 16S rRNA gene was sequenced using the Illumina MiSeq platform.

The sequencing adapters were trimmed and their low quality reads were removed by Gibson et al. (2014) on the online available dataset, which is compatible with the fastqc (Andrews et al., 2010) results. Initially, we classified all the reads in 44 samples using local mega BLASTn (Altschul et al., 1997), which is the most popular homology-based algorithm, then the top hit reads were identified to both genus-level and species-level using MEGAN 5.3 (Huson et al., 2007). Afterward, we used two k-mer based algorithms to classify the taxonomic composition of samples, KRAKEN (Wood and Salzberg, 2014) and LMAT (Ames et al., 2013), at the genus-level and species-level.

One approach to comparing the performance of these three algorithms with different read numbers was to make various subsets of a sample and then re-classify all the subsets with these three algorithms. However, we decided to simulate the performance of these three algorithms for different size subsets of a sample because the first approach is time-consuming, it takes up a lot of disk space, and the result would be as same as the re-classification approach. All these taxonomic classification algorithms assign a taxon

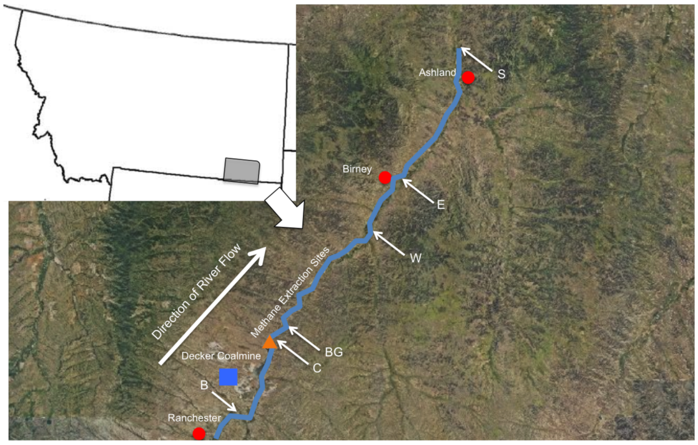


Figure 2.1: The location of Tongue River sampling site (Gibbons et al., 2014). Small towns are indicated with red points. The Decker coal mine and the Tongue River Reservoir Dam are illustrated with the blue square and the orange triangle respectively. The direction of river flow is south to north.

to each read, one read to one taxon. So, at a selected level of taxonomy, the number of identified taxa is equal to the read numbers. We have randomly sampled the $\log_{10}$ number of reads (identified taxa at each level) with substitution and the α diversity indices calculated for different subsets of whole real numbers. We have simulated the Simpson, and Shannon indices of different size read numbers in a sample at both genus-level and species-level using the vegan package in R (Oksanen et al., 2007). The result of all three programs was compared at different size subsets using the ggplot2 package in R (Wickham, 2009).

2.2.2 Result

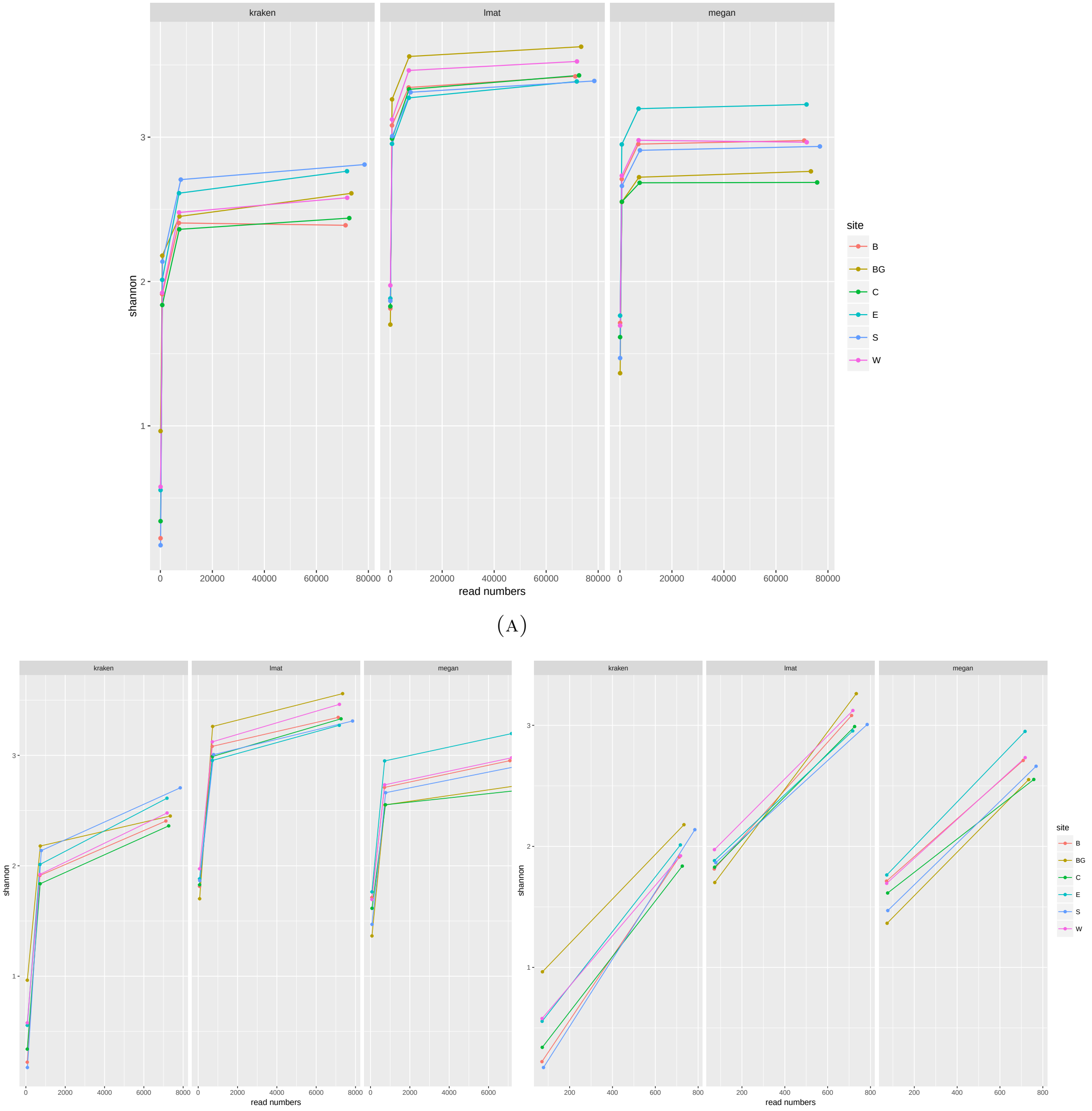
To have a better evaluation of the three algorithms (KRAKEN, LMAT, and BLASTn), we have decided to focus on the same taxonomic level and same diversity index for all the subsets and programs. We chose genus-level as the main taxonomic level and Shannon index as the information statistic index. Although some authors have chosen the suitable α diversity index based on their popularity (Gibson et al., 2015), the desired diversity index should be selected based on available data and the purposes of a project. Here as we are randomly sub-sampling the read numbers and calculating the α diversity for that particular number of sequences, we have focused on the Shannon index that assumes all the species represented in a sample and give the same weight to all the taxa.

We have used the Shannon index to compare the α diversity of various polluted sites from Gibson et al. (2014). Also, we compared the performance of different methods of classification in generating the Shannon index from these locations. Figure 2.2 illustrates how different programs (KRAKEN, LMAT, and BLASTn) calculated the Shannon index of six distinct sites with $\log_{10}$ sub-sampling of read numbers. The x-axes in Figure 2.2 show the read numbers and the y-axes is the Shannon index for each location. Each box in Figures 2.2a, 2.2b, 2.2c demonstrate the performance of different algorithms for the same range of read numbers. For a better demonstration of Shannon values at various sites, we have separated different range of read numbers. Figure 2.2a demonstrate the log10 sizes of read numbers, Figure 2.2b is the read numbers less than 10 percent of total reads and Figure 2.2c represents the read numbers less than 1 percent of full reads. We also compared the mean of Shannon values for all 44 samples together at different read numbers utilising three methods of taxonomic classification (Fig. 2.3). Again Figures 2.3a, 2.3b, and 2.3c illustrate various range of read numbers for better clarification.

Figure 2.2a illustrates that there is no change in Shannon index as read numbers are increased when read numbers are larger than 20,000. This result is consistent at different sites with various types of pollution. Figure 2.2a shows that there is a saturation level at around 20,000 read numbers where the Shannon values of all sites calculated by different algorithms plateau even with increasing read numbers. However Figures 2.2b and 2.2c demonstrate a linear relation between numbers of the reads and Shannon values at various sites. Figure 2.2b shows that by reducing the reads from around 8000 to 1000 the Shannon index at various sites calculated by three different methods significantly decreased. The relation between Shannon and depth of sequencing intensified from around 800 reads to 100 number of sequencing reads with a steeper slope (Fig. 2.3) in different sites calculated by various programs.

Figure 2.2 illustrates notably different Shannon values at different sites containing samples with the same read numbers. This difference highlights that community structure has been altered by anthropogenic drivers at each location, which is compatible with results from Gibbons et al. (2014). Various classification algorithms produce substantially different biodiversity assessments, Figure 2.3 demonstrates the performance of the three algorithms for all the samples as a function of different read numbers. The primary cause of this difference can be the structure of these algorithms or the size of their database. We have used the smallest database and all the default setting of these three programs that have been suggested by their authors for regular usage, but these databases were constructed differently and contained various number of taxa.

Figure 2.3a is comparing the mean of the Shannon index for all samples calculated by three classification methods (KRAKEN, LMAT, and BLASTn) at sub-sampling of read numbers. Our result indicates that there is no change in the Shannon index as read numbers are increased for subsets containing over 20,000 of read numbers. While an increasing relation between read numbers and α diversity for samples containing less than 8000 reads (Fig. 2.3b, 2.3c) is evident in all methods of classification.

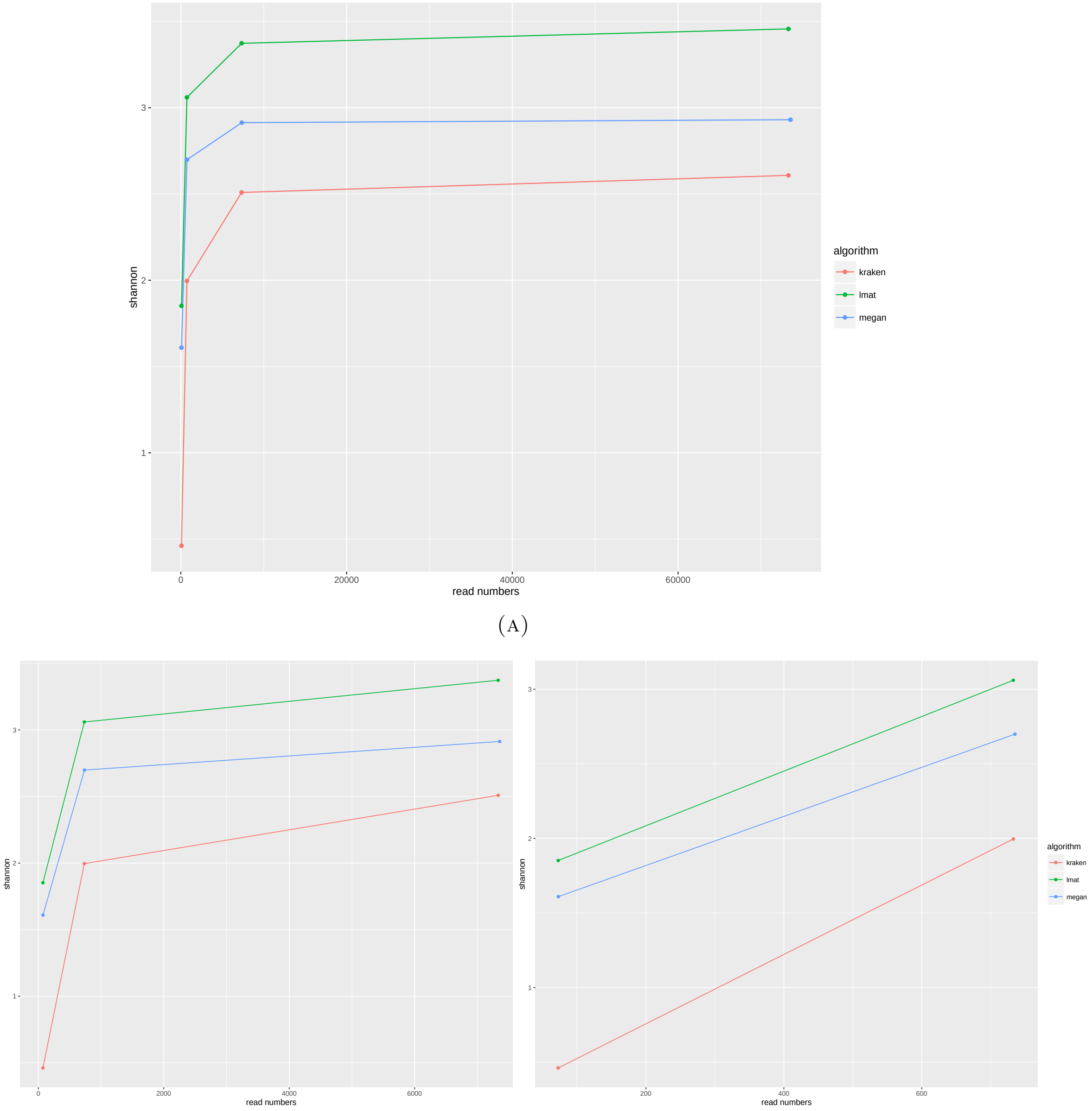


(A)

(B)

(C)

Figure 2.2: The relation between read numbers and Shannon index at various sites. The x-axis is the read numbers and the y-axis displays the Shannon index. Sites are colored differently and each dot here represents the mean of the Shannon index for all samples at a site. Figure 2.2a illustrate all sizes of sequence read numbers. Figures 2.2b and 2.2c demonstrate less than 10 % and 1 % of read numbers



(A)

(B)

(C)

FIGURE 2.3: The relation between read numbers and the Shannon index compared with various classification methods. The x-axis is the read numbers and the y-axis displays the Shannon index. Algorithms are colored differently and each dot here represents the mean of the Shannon index for all samples. Figure 2.2a illustrate all sizes of sequence read numbers. Figures 2.2b and 2.2c demonstrate less than 10 % and 1 % of read numbers

2.2.3 Discussion

The cost of HTS sequencing is rapidly decreasing by the introduction of new technologies. However, they have usually been considered highly expensive methods for large-scale biomonitoring (Sims et al., 2014). One possible approach to overcome the high cost of HTS is to increase the number of samples multiplexed (tagged) in an HTS run so that the efficiency of one sequencing run with fixed read number would be maximized. Simultaneously, increasing the number of samples multiplexed in a run will generate fewer read numbers for each sample. The purpose of DNA-based environmental monitoring is the taxonomic identification of bulk environmental samples and using diversity measurements (α and β diversity) to estimate the environmental conditions of an ecosystem. We investigated the required read numbers, the depth of sequencing, for taxonomic identification, and in calculating α diversity utilizing amplicon HTS method.

We have demonstrated the importance of sequence read numbers in DNA-based diversity measurements. Our result suggest that at around 20,000 sequences the alpha diversity index will plateau even with increasing the sequence reads. In other words, there is a saturation level for taxonomic identification in amplicon HTS sequencing at around 20,000 sequences (Fig. 2.2a). This result can be used to design future DNA-based monitoring projects in order to reduce the cost of sequencing.

The other challenge in environmental metabarcoding is the availability of sequencing platforms independent of the cost. Normally, there is only one sequencing platform available in one laboratory, and various labs have their preference of sequencing machine. These different platforms generate various outputs depending on their technologies and methods, but most of them are capable of producing more than 20,000 sequences per run (Reuter, Spacek, and Snyder, 2015). As we illustrated in our result (Fig. 2.3a), the low required numbers of sequences in taxonomic classification of amplicons HTS can be used for standardization of DNA-based environmental monitoring in various labs regardless of their sequencing platforms.

The results also indicate that HTS is necessarily for biodiversity community analyses of bulk environmental samples. As we have illustrated in Figures 2.2b, and 2.2c there is a relation between read numbers and alpha diversity. The Shannon index is reduced by lower number of sequences. This result suggests that using single read sequencing of samples (DNA barcoding) is not capable of generating α diversity as same as metabarcoding approach. This issue has been highlighted previously by other authors (Hajibabaei et al., 2011).

Figure 2.3 highlights the difference between these three methods of classifying HTS sequences. We have used all the default settings intentionally to focus more on the differences between various programs in normal usage. The main causes of variation in the Shannon index calculated by the three methods (Fig. 2.3a) are the algorithm structure and the default size of their databases. This issue shows on the importance of a standard pipeline for DNA-based environmental monitoring analyses. To generate a reproducible result in biomonitoring, researchers should accelerate their efforts in finding and utilizing the best post-sequencing pipeline.

In contrast to gene functional analyses, there is a saturation level in taxonomic identification using HTS methods. We would expect to see this saturation level at around 20,000 read numbers, but it can be different depend on the library preparation and sequencing methods. Unfortunately, Gibbons et al., 2014 did not list measurements of the level of contamination in their sampling locations, and their sampling did not contain clean locations. So we could not test the relation between sequence read numbers and the level of pollution.

2.3 The size of a database

Building a custom database of invertebrate bioindicator species is one of the primary purposes of my master project. I have constructed a small database containing 1927 bioindicator species. To evaluate the performance of such a small database, we decided to construct different size databases for three different taxonomic classification algorithms: KRAKEN (Wood and Salzberg, 2014), LMAT (Ames et al., 2013) and BLASTn (Altschul et al., 1997). To test the execution of these three algorithms independent of their database, the exact same genomic libraries were constructed for all three methods.

2.3.1 Methods

We continued the analysis with the same dataset from Gibbons et al. (2014), which is 16srRNA amplicons sequencing (see section 2.2.1). The bacterial reference genome from NCBI was used to construct different size databases for taxonomic classification of all samples (Pruitt, Tatusova, and Maglott, 2007). The available bacterial reference genomes contain 2786 unique taxa retrieved from NCBI (for each species there are a various number of accessions available in NCBI, but the total number of unique bacterial reference sequence at the date of downloading this data was 2786 species). Different size libraries assembled by random sampling of the bacterial reference genomes (RefSeq) using a custom python code. We constructed different size databases contain a different number of species to test the importance of database in calculating diversity indices. The $\log_{10}$ number of species randomly chose to assemble the databases. Table 2.1 shows the number of unique species in each database. The exact same libraries as shown in table 2.1 were used to build databases for three HTS classification algorithms.

Genome libraries	**Number of species**
DB1	2786
DB2	2500
DB3	2000
DB4	1000
DB5	100
DB6	10

TABLE 2.1: Constructed libraries of bacterial reference genome

We followed the KRAKEN manual to construct the custom databases. The NCBI taxonomy contains the GenInfo Identifier (GI number) as well as taxonomic information downloaded (Federhen, 2012). The previously constructed libraries of bacterial reference genomes were used to build the databases for KRAKEN algorithm (Tab. 2.1). Subsequently, the taxonomic information assigned to each genome using `-add-to-library` option in `kraken-build` package for all the six databases (Wood and Salzberg, 2014).

We have utilized the `-build` switch from the `kraken-build` package to construct the databases. To create all the databases, we used `K=31` and `M=31` in the `kraken-build` program. Here `K` is the fixed number of bases for each k-mer in a query sequence, in other words the k-mers are 31-mers (see section 1.6). `M` is the length of minimizer in base per (bp). Minimizer serves to keep the k-mers that are adjacent in a query sequence close to each other, which helps KRAKEN to use central processing unit (CPU) more efficiently. Any changes in M value can significantly affect the speed of the KRAKEN.

We identified the taxonomic composition of all the 44 samples utilizing six different KRAKEN databases. Unfortunately, we were not able to build a custom database for LMAT. We contacted the LMAT's authors to help with the problem, but they were not responsive. All the six libraries (Tab. 2.1) have been employed to create BLASTn databases, utilizing the `makeblastdb` application from BLAST tools (Altschul et al., 1997). All the 44 samples from Gibson et al. (2014) have been run against these databases separately, using the local megaBLASTn (Altschul et al., 1997) algorithm. Subsequently, the top hit reads with lowest common ancestor (LCA) were identified to genus level employing MEGAN 5.3 (Huson et al., 2007). The Shannon index at genus-level for both algorithms calculated for each sample utilizing the vegan package in R (Oksanen et al., 2007). We also used a custom code in R to bootstrap the Shannon values for 1000 replication in each site and calculated the mean of Shannon index for individual sites.

2.3.2 Result

We have used two different classification algorithms to calculate the Shannon index at the genus-level for all the samples from Gibson et al. (2014). We utilised six different size databases (Tab. 2.1) for both KRAKEN and BLASTn program to test the importance of the database in the ecosystems with significant anthropogenic pollutants (see section 2.2.1). Figure 2.4 illustrates the correlation between the size of a database and Shannon index for various sampling locations. Here the y-axis is the same Shannon scale for two programs; BLASTn in the first row and KRAKEN in the second row. The x-axis is five different size databases with 100, 1000, 2000, 2500, and 2786 number of species that constructed in the same manner for both programs. Each site is shown with different colours and the Shannon value for each sample within a site is calculated with two programs. Each point in figure 2.4 is the Shannon value for the single sample.

The comparison of the sites in figure 2.4 suggest that there are various community structure at different sampling locations, which is again compatible with the result of Gibbons et al. (2014). When the number of taxa in a database are reduced, the Shannon values change dependent on the composition of samples. The α diversity in all the sites classified by both programs shows a significant correlation between the size of database and Shannon index for databases containing less than 1000 number of species (Fig. 2.4). The Shannon values were reduced when the number of taxa in the databases was less than 1000.

The relationship between the size of databases containing more than 1000 species, and the Shannon index is not the same for samples in various sites or even within a site. The result calculated by KRAKEN in sites BG and W illustrate mostly lower Shannon values with fewer numbers of species in the database. However, the α diversity calculated by KRAKEN in sampling locations B, S, E and calculated by BLASTn in sites E, C, BG, W, S in most samples for databases containing more than 1000 taxa were not significantly changed by smaller size databases (Fig. 2.4). Figure 2.4 also shows that the Shannon index for some samples calculated by BLASTn in site E, S, C, B and calculated by KRAKEN in site C increased in databases contain more that 1000 number of species.

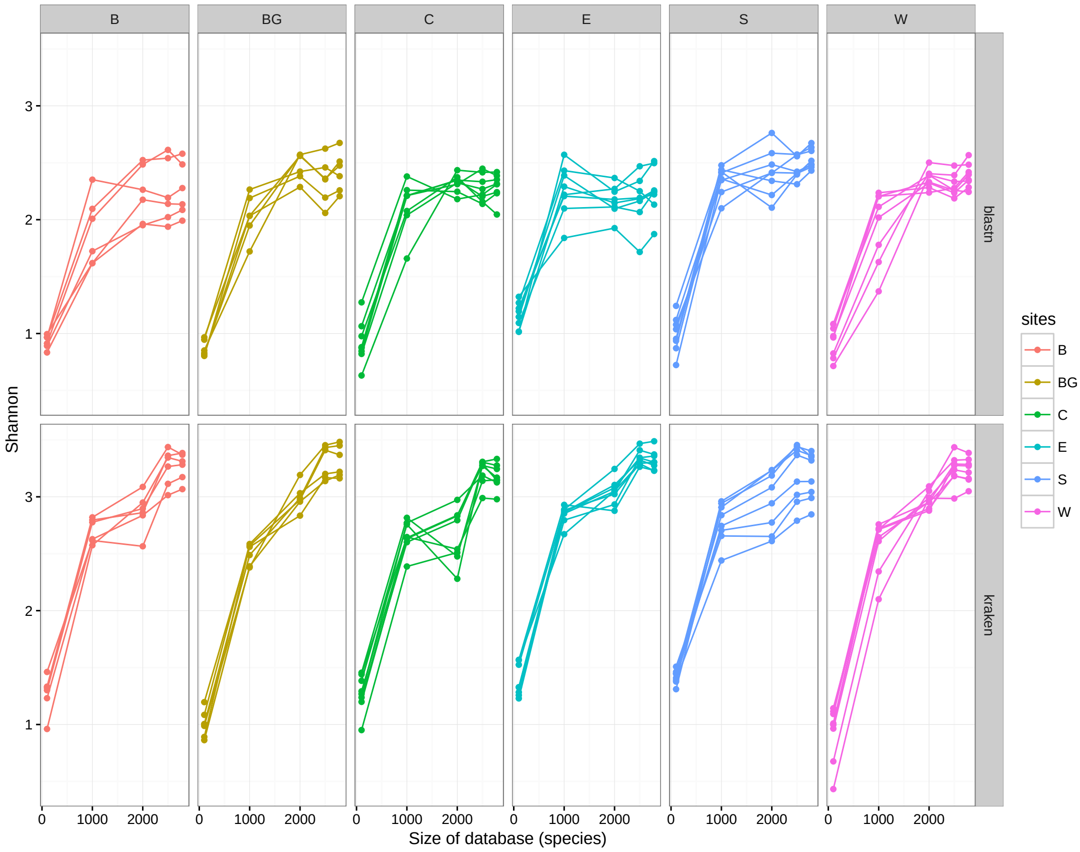


FIGURE 2.4: Correlation between the size of a database and α diversity (Shannon index) for individual samples from polluted sites (B - W). The x-axis is the number of species in a database (Tab. 2.1). The y-axis is Shannon index. Each dot is representing one sample and sites are separated with different colors. The result of KRAKEN and BLASTn algorithms compared for each sites.

Figure 2.5 compares the correlation between the size of the database and the Shannon index for the BLASTn algorithm in various sites with anthropogenic pollutants. Here the y-axis is the Shannon index, and the x-axis is five different sizes of database constructed for the BLASTn classification program (Fig. 2.5). Distinct colours separate sites, and each dot shows the bootstrapped mean of Shannon values for all samples within a site (Fig. 2.5). The mean for each Shannon index was calculated by bootstrapping with 1000 replications. Sites B and BG contained six samples and the other sites included eight samples.

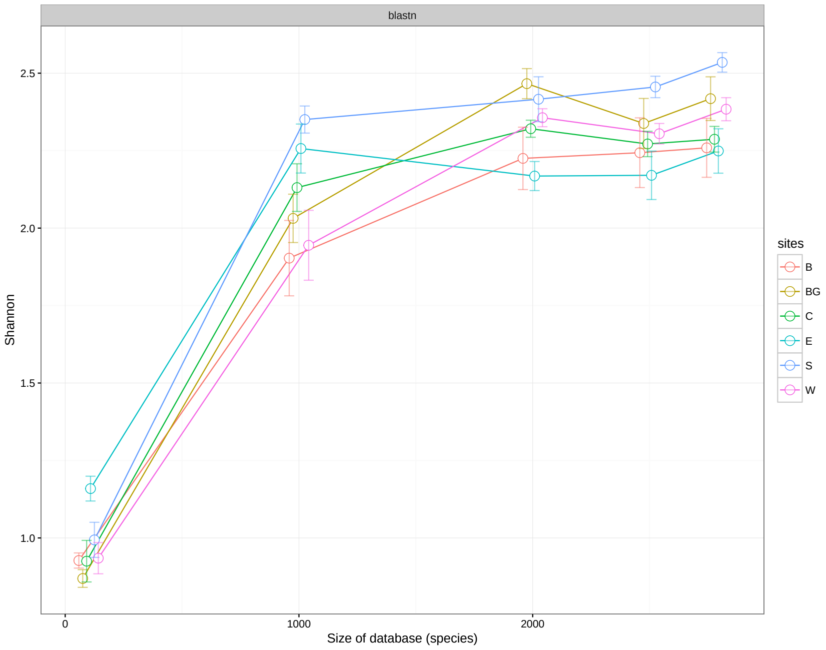


FIGURE 2.5: Correlation between Shannon index and the size of database in BLASTn program. The x-axis is different size databases and y-axis is Shannon index. Each site is shown with different colors. The points are dodged to prevent overlapping the error bars

Figure 2.5 shows that by reducing the number of taxa in the BLASTn program from 1000 species to 100 species the Shannon index significantly decreased. However, figure 2.5 illustrates that Shannon index is not much dependent to the size of the databases in BLASTn program for databases include more than 1000 taxa. There were not significant changes in Shannon for most of the sites when we reduced the database from 2500 to 1000 taxa. Also, figure 2.5 shows that by reducing the number of taxa in the database from 2000 to 1000 in site B and by decreasing the size from 2500 to 2000 in site BG the Shannon index even increased, which was presumably caused by changes in samples' evenness. Even with the major reduction of the database from 2786 to 1000 species, the Shannon changes are smaller than error bars in most of the sites.

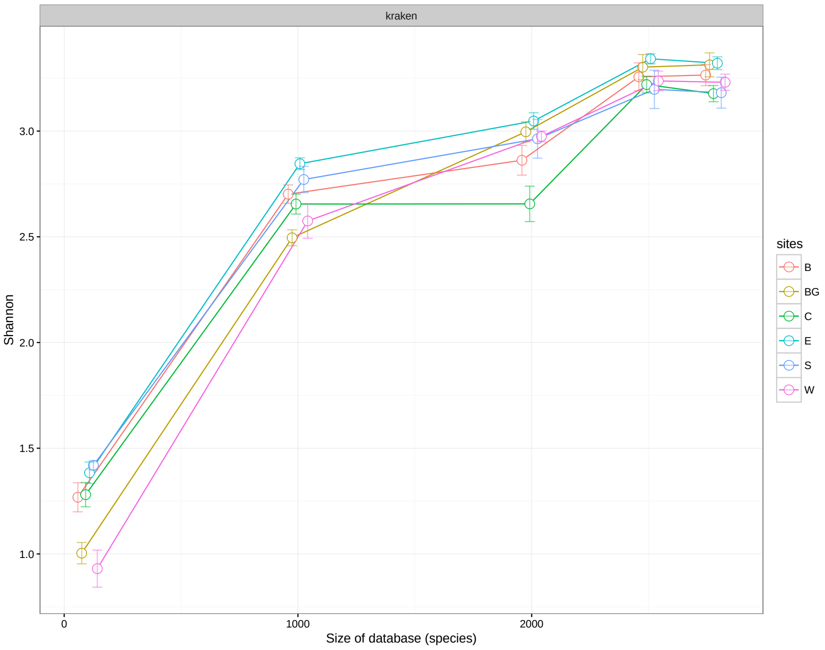


FIGURE 2.6: Correlation between Shannon index and the size of database in KRAKEN program. The x-axis is different size databases, and the y-axis is Shannon index. The points are dodged to prevent overlapping the error bars

We have also compared the relation between the Shannon and number of taxa in a database for the KRAKEN program at various sites that contain anthropogenic drivers. In the figure 2.6 the y-axis is the Shannon index, and the x-axis is five different size database constructed for BLASTn classification program. Distinct colours separate sites, and each dot shows the bootstrapped mean of Shannon values for all samples within a site (Fig. 2.5). The mean for each Shannon index was calculated by bootstrapping with 1000 replications.

Figure 2.6 displays that by reducing the number of taxa in the KRAKEN program from 1000 species to 100 species the Shannon index significantly decreased. However, this result illustrates that there is no significant relation between the size of the databases and the Shannon index in KRAKEN program for databases include more than 1000 taxa. There were not significant reduction in Shannon when we reduced the database substantially from 2000 to 1000 taxa. Also, figure 2.6 shows that by reducing the number of taxa in the database from 2000 to 1000 in site C the Shannon index significantly increased, which presumably caused by changes in samples' evenness. Although the KRAKEN program is more sensitive (lower minimum support, and lower accuracy) in taxonomic classification compared to the BLASTn algorithm, the Shannon changes are smaller than error bars in most of the sites when we reduced the size of a database from 2000 to 1000 taxa or from 2750 to 2500 species (Fig. 2.6).

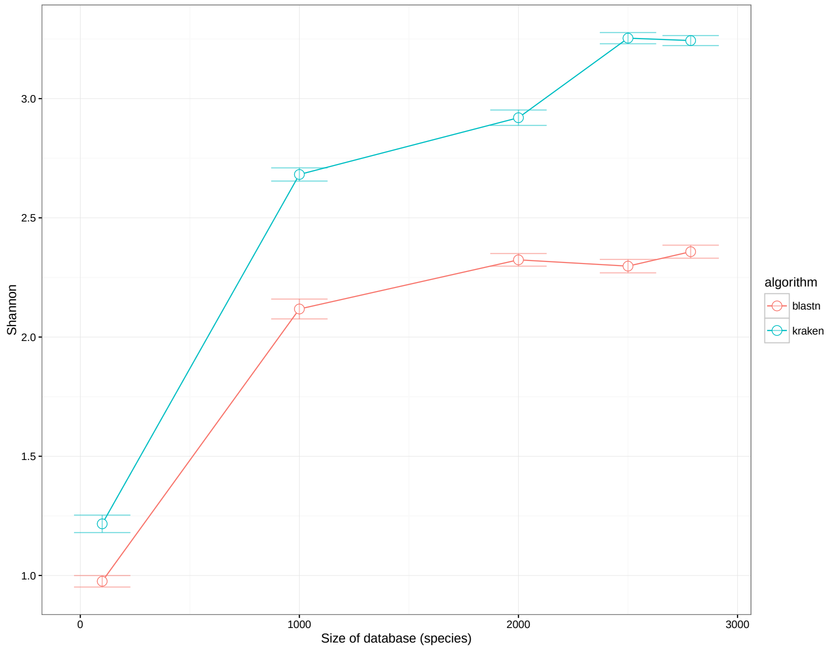


FIGURE 2.7: A comparison of KRAKEN and BLASTn algorithm in calculating alpha diversity with different size databases (Tab. 2.1).

Also, we focused on the performance of KRAKEN and BLASTn algorithms independent of the sites. Figure 2.7 demonstrates the correlation of two different taxonomic classifiers to the size of the database. Again the x-axis in figure 2.7 displays the same five databases for both algorithms. The databases include 100, 1000, 2000, 2500, and 2786 unique species randomly chosen from bacterial reference genomes available at NCBI. Here each dot shows the bootstrapped mean of Shannon values for all samples from Gibson et al. (2014) and based on 1000 replications.

In both classifiers, the Shannon index was significantly reduced by decreasing the number of taxa from 1000 to 100 in a database. Figure 2.7 illustrates the difference between taxonomic identifiers even with the same database implemented in both of them. When we reduced the number of species from 2786 to 2000 in a database, the BLASTN program did not change significantly. However, the KRAKEN algorithm significantly reduced for the same reduction of species in a database. Although the Shannon index of both programs was reduced when we substantially decreased the number of taxa in a database from 2000 to 1000, the Shannon index reduction is not comparable to the database reduction. It seems that the KRAKEN classifier as a more sensitive program, based on default settings, is more correlated to the size of the database compared to the BLASTN which is a more accurate program by default setting. The small error bars in figures 2.5, 2.6, and 2.7 indicates that the changes in Shannon values with a reduced number of taxa in the database are not because of the limited numbers of samples.

2.3.3 Discussion

Comparing the Shannon values of individual samples between and within sampling locations (Fig. 2.6) suggests that community structure is different among sampling locations. However, these variances between sites are not statistically significant. Figures 2.5 and 2.6 demonstrate different means of Shannon index among sites but these differences are smaller than the error bars in most of the points. Gibbons et al. (2014) discussed that the core taxonomic composition of sites is pretty similar, but the presence of distinct taxa vary with sites and anthropogenic drivers. This argument is reflected in our α diversity analysis, because the Shannon values are pretty similar between sites (Fig. 2.5, 2.6) while these values are different within a site (e.g. Sites S and B).

It seems that the number of taxa is not as crucial as the importance of taxa in a database for biodiversity analyses in polluted environments. In such ecosystems, identification of the core taxonomic composition is more critical than classifying all the present species for biodiversity measurements. When we reduced the number of taxa in a database from 2000 species to 1000 species, the Shannon index did not change significantly in both programs. This argument was also supported by Gibbons et al. (2014), as the core taxonomic composition is constant between the sites and changes in the composition of the other variable taxa do not affect the evenness and consequently the α diversity.

Exploring the performance of KRAKEN and BLASTn programs in calculating diversity indices suggest that they both reduced by decreasing the number of species from 1000 to any number lower. However, the relationship between the number of taxa and the α diversity is not consistent when we reduced the number of taxa in a database from 2786 to 1000. The BLASTn (Fig. 2.5)is less dependent on the number of taxa in a database compared to the KRAKEN algorithm (Fig. 2.6). When we reduce the number of species from 2786 to 1000, the Shannon values in BLASTn less changed compared to the KRAKEN. This dependency is due to the higher sensitivity of KRAKEN in taxonomic classification of samples with default settings compared to the megablast options used for BLASTn approach. It was possible to decrease the differences between these two programs by modifying the level of accuracy and minimum support number. But, we left the settings at their default to examine the execution of these algorithms for regular usage.

As mentioned earlier, the evenness is one of the main components of the biodiversity that quickly reflects the changes in environments, and it is being used as a tool to study human effects on environments 1.1. Unfortunately, we were not able to investigate the relationship between evenness (α diversity) and the level of anthropogenic pollutants in various sites. Because Gibbons et al. (2014) did not categorize their sampling locations based on the intensity of anthropogenic pollutants, they have calculated various β diversity measurements based on the correlation of particular variables comprising the distance between sites, temperature, salinity, and pH. They have found a remarkable association between community structure and the sites nearest in space.

Our result highlighted that changes in evenness caused significant alteration in the α diversity analysis. As previously mentioned, even by decreasing the size of the database the Shannon index increased in two particular sites. By reducing the number of taxa from 2000 to 1000 using the BLASTn algorithm in site E and with the KRAKEN program in site C, the Shannon values increased. This variation in α diversity can be explained only by changes in evenness because the sample's richness will be reduced by the presence of fewer taxa in the database and the only other factor that can influence the Shannon index is the evenness.

This result shows that the core taxonomic composition of samples has a profound effect on calculating the α diversity, particularly in polluted environments. Considering that there are not significant differences between sites in calculating alpha diversity using two programs(Fig. 2.5, 2.6). Also, figure 2.7 demonstrates that with reducing the taxonomic richness, decreasing the number of species from 2000 to 1000, the alpha diversity was not significantly reduced. In polluted environments the core taxonomic composition contains resilient species. Assembling these resilient taxa and construct a database that could identify these bioindicator species will produce reasonably the α diversity estimation similar to the result from enormous database of all known organism.

Although Gibbons et al. (2014) indicated that the community structure has been altered by anthropogenic drivers, we can not confirm this argument as they did not include samples from clean sites in their analyses. It is not clear if the differences in community structures between the sites are due to the presence of the anthropogenic drivers or because of the natural composition of the sampling locations. Also, we could not investigate the relationship between the level of pollutants and the required number of taxa in a database because Gibson et al. (2014) did not categorize their sampling locations based on the intensity of anthropogenic pollutants.

Human activities in the ecosystems could lead to lower taxonomic diversity and survival of the resilient species in the environments. The composition of these resistant species would be stable under a various level of pollutants and by identifying these flexible species in such contaminated environments the chance of accurate diversity measurements will increase. If these resilient taxa would be the target of identification using a database contains these species the chance of accurate biodiversity estimation would be as same as including all the known species into the database. Also, the accuracy and sensitivity of the classification algorithm are crucial for avoiding false identification that may cause inaccurate diversity measurements.

Chapter 3

A small customized DNA database

3.1 introduction

The main challenges with the current biomonitoring (metabarcoding) studies are the efficiency of a sequence classification in terms of time and computational power, the presence of a sufficiently targeted database for environmental monitoring in eukaryotes communities, and the standard approach to address mixed data sources (Dejean et al., 2011; Baird and Hajibabaei, 2012). Almost all of the biomonitoring studies use the homology-based algorithms (e.g. BLAST; Altschul et al., 1997) for taxonomic identification, which is a time-consuming method because all the query sequences should be compared to a gigantic eukaryotic database one by one. The homology-based method requires another algorithm (e.g. MEGAN; Huson et al., 2007) to trim the final matches with the database and find the lowest common ancestor for each match. Consequently, each query sequence that is compared to the whole database generates an output that takes substantial disk space.

These two issues, the time of HTS classification and the disk space requirement, have helped to prevent a standard approach for biomonitoring of eukaryotic communities. Different environmental agencies and research groups have their preferential pipelines to address these challenges given their available monetary and non-monetary resources. The different pipelines are based on each project's purposes and lead to incompatibility in HTS taxonomic identification. For the environmental agencies, the rapid identification and interpretation of an ecosystem is a fundamental element as they should diagnose and anticipate the environmental issues as quickly as possible. For the research groups the monetary issues, such as the cost of sequencing, and the computational power are the most significant difficulties.

The purpose of this project is to introduce an approach to overcome these challenges and facilitate the establishment of a standard method for biomonitoring using environmental DNA and eukaryotic communities. We show that a small customized database of bioindicators helps to reduce the time of classification and evaluation of the diversity of an ecosystem. This type of database would be sufficiently accurate for biodiversity assessments in polluted environments, which are the environments of interest for most of the environmental studies to investigate human impacts on the ecosystems.

We have also explored the performance of a k-mer based algorithm for taxonomic classification of eukaryotic communities for the first time. we hypothesize that k-mer based algorithms enormously reduce the time of classification for HTS reads, simultaneously it has high identification accuracy, and it takes much less computational disk space compared to homology-based algorithms. We have combined a small customized database of bioindicators and a suitable HTS classification algorithm to introduce a new pipeline for biodiversity assessments of polluted environments and compared the performance of our method to current popular biomonitoring post-sequencing pipelines.

3.1.1 Objectives

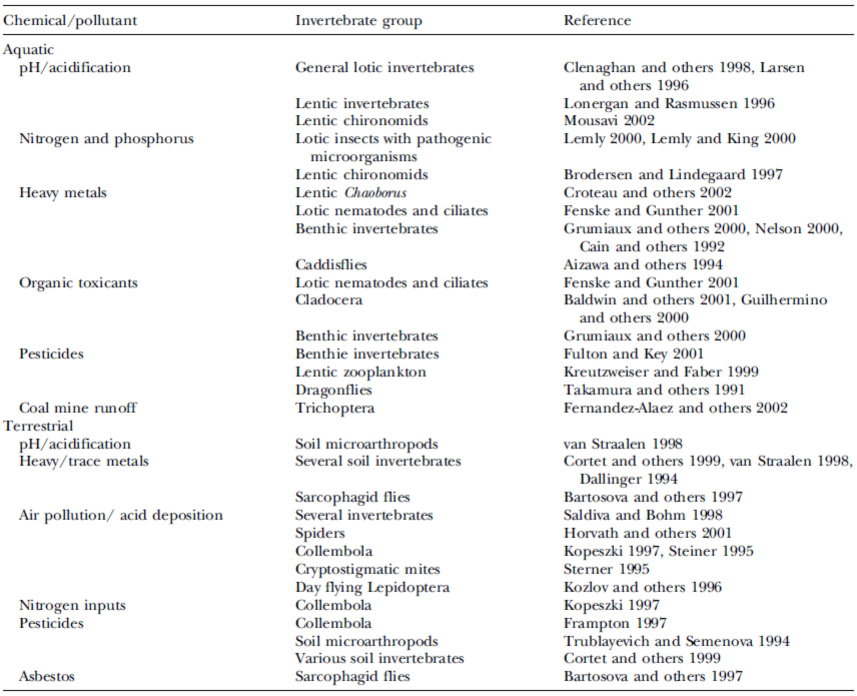
We explained the possibility of reducing the cost of sequencing, as one of the main challenges with rapid DNA-based biomonitoring, with generating less sequence read numbers in taxonomic identification and biodiversity assessments of amplicon HTS in chapter 2 of this thesis. We also concluded that the number of taxa in a database is not as important as the role of those taxa in constructing the community composition of a contaminated environment. These two conclusions from chapter 2 are compatible with our hypotheses for this chapter. A small database of targeted bioindicators can sufficiently estimate biodiversity of a polluted environment compared to the other enormous eukaryotic databases. The KRAKEN program, a k-mer based classifier, improves the DNA-based biomonitoring by reducing the time, cost, and required computational power for the analysis of amplicon-based HTS.

3.2 Library of macroinvertebrate bioindicators

There are criteria for the selection of bioindicator invertebrates based on conventional taxonomic identification. But, the major challenge is to select bioindicator species for environmental monitoring purposes and transfer this knowledge to DNA-based approaches. Bioindicator invertebrates based on conventional taxonomic identification are well-known and stable species that are easily identified and are known to respond to stress factors or changes in habitat, they are generally abundant organisms suitable as whole-community representatives (Hilty and Merenlender, 2000). Hodkinson and Jackson (2005) reviewed and critically evaluated the suitability of different terrestrial and aquatic invertebrates as a management tool for the monitoring of changes in an ecosystem. They classified a range of factors in environments and presented potential studies as three categories of bioindicators. The first category includes the range of chemical factors in aquatic and terrestrial environments that are potential for being monitored by invertebrates (table. 3.1). The Second group consists of indicator invertebrates in evaluating habitats for biodiversity, condition, and structure (table. 3.2). The third category contains indicators of habitat management, degradation, restoration and improvement (table. 3.3).

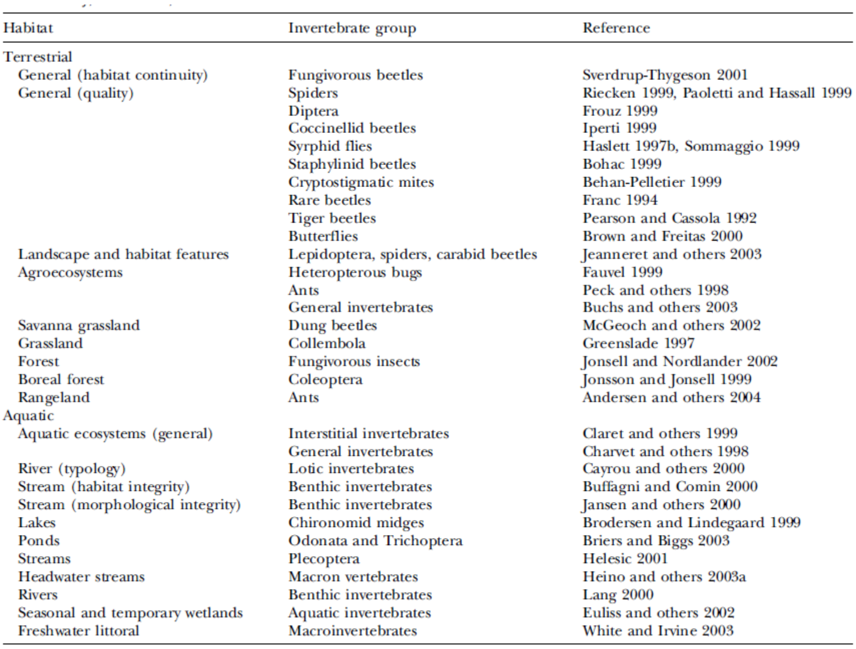
Some of the terrestrial and aquatic macroinvertebrates are potential bioindicators for a range of chemical changes in environments. These bioindicators are responsive to a single chemical parameter or a combination of chemical effects. These chemical parameters include pH, many plant nutrients such as nitrogen and phosphorus, or the presence of heavy metal compounds like cadmium (Brodersen and Anderson, 2002). For instance, *Hexagenia limbata*, a mayfly native to North America, has been used as the indicator of lethal anoxic conditions caused by eutrophication (Krieger et al., 1996), (for more examples see table 3.1).

Some groups of invertebrates can also indicate the habitat quality and conservation qualities of an environment. The bioindicators in this category are often ecologically important taxa that might be affected by human developments (Hodkinson and Jackson, 2005). The presence of particular taxa can be used to indicate long-term stability within a habitat over time (Nordén and Appelqvist, 2001). For example, land snails and certain beetles related to persistent forest fungi have been highlighted as indicators of long-term stability within woodland (Nordén and Appelqvist, 2001; Sverdrup-Thygeson, 2001), (for more examples see table 3.2).



Chemical/pollutant	Invertebrate group	Reference
Aquatic		
pH/acidification	General lotic invertebrates	Clenaghan and others 1998, Larsen and others 1996
	Lentic invertebrates	Lonergan and Rasmussen 1996
	Lentic chironomids	Mousavi 2002
Nitrogen and phosphorus	Lotic insects with pathogenic microorganisms	Lemly 2000, Lemly and King 2000
	Lentic chironomids	Brodersen and Lindegaard 1997
Heavy metals	Lentic *Chaoborus*	Croteau and others 2002
	Lotic nematodes and ciliates	Fenske and Gunther 2001
	Benthic invertebrates	Grumiaux and others 2000, Nelson 2000, Cain and others 1992
	Caddisflies	Aizawa and others 1994
Organic toxicants	Lotic nematodes and ciliates	Fenske and Gunther 2001
	Cladocera	Baldwin and others 2001, Guilhermino and others 2000
	Benthic invertebrates	Grumiaux and others 2000
Pesticides	Benthie invertebrates	Fulton and Key 2001
	Lentic zooplankton	Kreutzweiser and Faber 1999
	Dragonflies	Takamura and others 1991
Coal mine runoff	Trichoptera	Fernandez-Alaez and others 2002
Terrestrial		
pH/acidification	Soil microarthropods	van Straalen 1998
Heavy/trace metals	Several soil invertebrates	Cortet and others 1999, van Straalen 1998, Dallinger 1994
	Sarcophagid flies	Bartosova and others 1997
Air pollution/ acid deposition	Several invertebrates	Saldiva and Bohm 1998
	Spiders	Horvath and others 2001
	Collembola	Kopeszki 1997, Steiner 1995
	Cryptostigmatic mites	Sterner 1995
	Day flying Lepidoptera	Kozlov and others 1996
Nitrogen inputs	Collembola	Kopeszki 1997
Pesticides	Collembola	Frampton 1997
	Soil microarthropods	Trublayevich and Semenova 1994
	Various soil invertebrates	Cortet and others 1999
Asbestos	Sarcophagid flies	Bartosova and others 1997

TABLE 3.1: The range of chemical factors in aquatic and terrestrial environments that have a potential for being biomonitoried by invertebrates from Hodkinson and Jackson (2005).

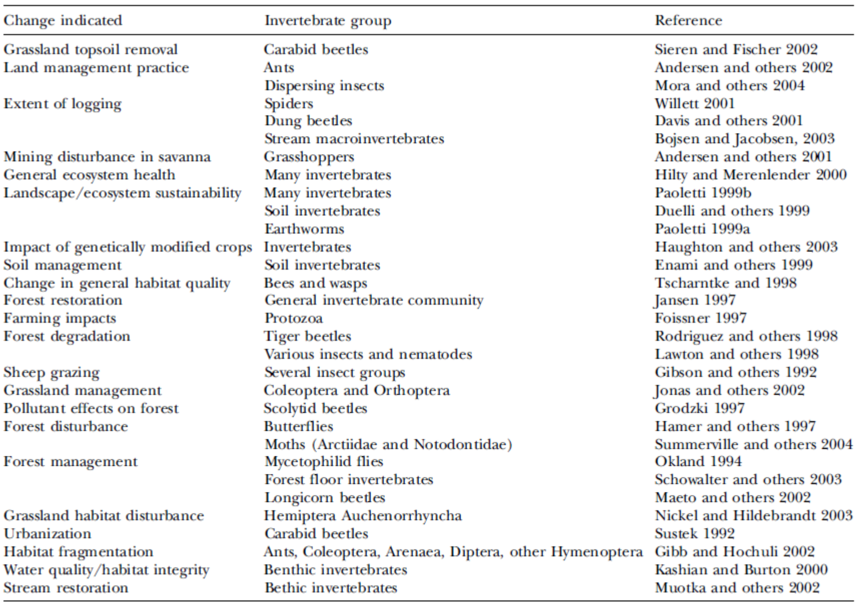


Habitat	Invertebrate group	Reference
Terrestrial		
General (habitat continuity)	Fungivorous beetles	Sverdrup-Thygeson 2001
General (quality)	Spiders	Riecken 1999, Paoletti and Hassall 1999
	Diptera	Frouz 1999
	Coccinellid beetles	Iperti 1999
	Syrphid flies	Haslett 1997b, Sommaggio 1999
	Staphylinid beetles	Bohac 1999
	Cryptostigmatic mites	Behan-Pelletier 1999
	Rare beetles	Franc 1994
	Tiger beetles	Pearson and Cassola 1992
	Butterflies	Brown and Freitas 2000
Landscape and habitat features	Lepidoptera, spiders, carabid beetles	Jeanneret and others 2003
Agroecosystems	Heteropterous bugs	Fauvel 1999
	Ants	Peck and others 1998
	General invertebrates	Buchs and others 2003
Savanna grassland	Dung beetles	McGeoch and others 2002
Grassland	Collembola	Greenslade 1997
Forest	Fungivorous insects	Jonsell and Nordlander 2002
Boreal forest	Coleoptera	Jonsson and Jonsell 1999
Rangeland	Ants	Andersen and others 2004
Aquatic		
Aquatic ecosystems (general)	Interstitial invertebrates	Claret and others 1999
	General invertebrates	Charvet and others 1998
River (typology)	Lotic invertebrates	Cayrou and others 2000
Stream (habitat integrity)	Benthic invertebrates	Buffagni and Comin 2000
Stream (morphological integrity)	Benthic invertebrates	Jansen and others 2000
Lakes	Chironomid midges	Brodersen and Lindegaard 1999
Ponds	Odonata and Trichoptera	Briers and Biggs 2003
Streams	Plecoptera	Helesic 2001
Headwater streams	Macron vertebrates	Heino and others 2003a
Rivers	Benthic invertebrates	Lang 2000
Seasonal and temporary wetlands	Aquatic invertebrates	Euliss and others 2002
Freshwater littoral	Macroinvertebrates	White and Irvine 2003

TABLE 3.2: The suggested use of bioindicators for evaluating habitats for biodiversity, condition, and structure from Hodkinson and Jackson (2005).

Bioindicator invertebrates are also suggested to indicate habitat change, degradation, and recovery. These indicators are being usually used as a part of biomonitoring programs to illustrate habitat degradation and the reduced biodiversity caused by negative human impacts such as agricultural practices, land use modification, and pollutant inputs. These biomonitor species sensitively respond to the change in the environment. Simultanesly, they are resilient enough that they will not be extinct from the environment. For example, as pollution increases, the macroinvertebrate community dominated by caddisflies, stoneflies, and mayflies shifts to a community dominated by chironomid and tubificid worms. When the stress is removed, the previous community composition will be restored again (Hodkinson and Jackson, 2005), (for more examples see table 3.3).

We have used the references in Hodkinson and Jackson (2005) to find potential bioindicators for each stressor-response. Although many of these studies failed to present species-level information for their indicator taxa, we used a combination of Genbank and The International Barcode of Life project (iBOL) to retrieve species level information. In some cases, we were not able to find any data for bioindicators in publicly available databases. In some studies, they couldn't identify a taxon to species-level, and, in some other cases, their identifications were not accurate. Furthermore, a bioindicator taxon may be merely the number of rare, local, or endangered species (Rosenberg, Danks, and Lehmkuhl, 1986), which explains why there are not any DNA records available from them.



Change indicated	Invertebrate group	Reference
Grassland topsoil removal	Carabid beetles	Sieren and Fischer 2002
Land management practice	Ants	Andersen and others 2002
	Dispersing insects	Mora and others 2004
Extent of logging	Spiders	Willett 2001
	Dung beetles	Davis and others 2001
	Stream macroinvertebrates	Bojsen and Jacobsen, 2003
Mining disturbance in savanna	Grasshoppers	Andersen and others 2001
General ecosystem health	Many invertebrates	Hilty and Merenlender 2000
Landscape/ecosystem sustainability	Many invertebrates	Paoletti 1999b
	Soil invertebrates	Duelli and others 1999
	Earthworms	Paoletti 1999a
Impact of genetically modified crops	Invertebrates	Haughton and others 2003
Soil management	Soil invertebrates	Enami and others 1999
Change in general habitat quality	Bees and wasps	Tscharntke and 1998
Forest restoration	General invertebrate community	Jansen 1997
Farming impacts	Protozoa	Foissner 1997
Forest degradation	Tiger beetles	Rodriguez and others 1998
	Various insects and nematodes	Lawton and others 1998
Sheep grazing	Several insect groups	Gibson and others 1992
Grassland management	Coleoptera and Orthoptera	Jonas and others 2002
Pollutant effects on forest	Scolytid beetles	Grodzki 1997
Forest disturbance	Butterflies	Hamer and others 1997
	Moths (Arctiidae and Notodontidae)	Summerville and others 2004
Forest management	Mycetophilid flies	Okland 1994
	Forest floor invertebrates	Schowalter and others 2003
	Longicorn beetles	Maeto and others 2002
Grassland habitat disturbance	Hemiptera Auchenorrhyncha	Nickel and Hildebrandt 2003
Urbanization	Carabid beetles	Sustek 1992
Habitat fragmentation	Ants, Coleoptera, Arenaea, Diptera, other Hymenoptera	Gibb and Hochuli 2002
Water quality/habitat integrity	Benthic invertebrates	Kashian and Burton 2000
Stream restoration	Bethic invertebrates	Muotka and others 2002

Table 3.3: The suggested use of bioindicators for habitat management, restoration, and improvement from Hodkinson and Jackson (2005).

We identified 1927 unique bioindicator species utilizing Hodkinson and Jackson (2005) references and criteria. For individual species, we searched for Cytochrome c oxidase subunit I (COI) and Cytochrome b (Cytb) markers in the GenBank nucleotide database (Benson et al., 2013) using the graphical interface. These two markers were shown to be successful for taxonomic identification of invertebrates specifically using bulk eukaryotic samples (Carew et al., 2013). The COI marker was used in preference to Cytb to download the sequences. We did not retrieve the Cytb sequences if we had the COI marker sequences of one species. We extracted the top 10 -15 accessions of each species depending on the total number of available accessions in the GenBank database. The list of bioindicator species and the library of sequences can be found at the GitHub page (https://github.com/shekas3/Bioindicators).

3.3 Implementation of the databases

Although some researchers use the whole library of COI markers, available at NCBI, for taxonomic classification of their amplicon HTS, we believe that this approach only adds enormous numbers of irrelevant taxa to the database. Consequently, the identification process takes a long time, and it requires massive disk space. For a library of bioindicators, we have constructed another local library of all available invertebrates in the GenBank as a reference sequence library to investigate the accuracy and efficiency of a bioindicator database. This library contains all the available sequences of the COI and Cytb markers with some manual specification within the Protostomia lineage in the nucleotide database of the GenBank (Benson et al., 2013).

We assembled the reference library of invertebrates utilizing a custom R program which loads `ape`, and `rentrez` packages (Paradis, Claude, and Strimmer, 2004). We specified a minimum of 100 and a maximum of 1000 bp `(100[SLEN]:1000[SLEN])` for the `COI` and `CYTB` genes in the Protostomia lineage from the nucleotide (`nuccore`) database in the Fasta format. Totally there are 1,147,778 sequences in the reference library of invertebrates in the multi-Fasta format. To calculate the total number of unique species available in our reference library, we developed a custom Python program to extract all the accession numbers from the downloaded sequences and mapped those accessions to a local taxonomy file. Totally there are 167,788 unique species available in our reference library of invertebrates.

Earlier in chapter 2, we have explored the performance of the KRAKEN algorithm for taxonomic classification of bacterial communities using the 16SrRNA marker. Our results showed that the KRAKEN (Wood and Salzberg, 2014) program is adequately efficient and accurate compared to the BLASTn (Altschul et al., 1997) algorithm for taxonomic classification of amplicon HTS. To address the biomonitoring challenges, we used the KRAKEN algorithm to construct a database of bioindicators and a reference database of invertebrates. To test the performance of the KRAKEN algorithm for taxonomic identification of eukaryotic communities, for the first time, we have assembled another database with the same reference invertebrates in the BLASTn algorithm.

After assembling the sequence library of bioindicators and the reference library of invertebrates, we implemented them into the different algorithms and constructed three databases for amplicon HTS classification.The ProK, ProM, and the SHAH were used to build databases for the KRAKEN and the BLASTn, and the KRAKEN algorithm respectively. For both approaches, ProK and ProM, we employed the reference library of invertebrate, and for the SHAH the library of bioindicators was used to construct the databases. Table 3.4 shows these three databases and compares the number of species, sequences as well as the disk space requirement for each method.

The SHAH database constructed using the KRAKEN algorithm, and the library of bioindicator species. We used a custom python code to convert the GenBank file format (gb) to a multi-Fasta format for the retrieved sequences of the bioindicators. The NCBI taxonomy includes the GI number as well as taxonomic information downloaded from the NCBI (Federhen, 2012). We have added all the sequences from the multi-Fasta file of bioindicators to the KRAKEN program's database employing the `-add-to-library` option from the `kraken-build` package and a Unix shell script to automate the process. We have utilized the `-build` switch from the `kraken-build` package to construct the database and assigned the taxonomy information to each sequence. To create the database, we used `K=31` and `M=31` in the `kraken-build` program.

We implemented the library of the reference invertebrates into the KRAKEN algorithm to assemble the ProK database. Again all the sequences available in the library were added to the database of the algorithm using `-add-to-library` and the taxonomy information was added to each sequence utilizing the `-build` switch from the `kraken-build` package. We also used the `K=31` and `M=31` in the `kraken-build` package to construct the ProK. The ProM database created using the reference library of invertebrates and the `makeblastdb` application from the BLAST tools (Altschul et al., 1997). The `makeblastdb` program requires Fasta file format and the taxonomy identifier file to build the database.

Parameter	**SHAH**	**ProK**	**ProM**
Number of species	1927	167,788	167,788
Number of sequences	20,235	1,147,778	1,147,778
Algorithm	KRAKEN	KRAKEN	BlastN
Method	K-mer based	k-mer based	Homology based
Database Size	8G	30G	405M

TABLE 3.4: Custom HTS classification databases for eukaryotes.

3.4 Samples

3.4.1 The Humber River watershed

The Humber River watershed was designated a Canadian Heritage River for its remarkable human heritage and recreational values in 1999 (CHRS, 2011). This river runs through a series of progressively more urbanized zones ranging from non-urbanized, natural landscapes to landscapes experiencing intense agricultural activity, to suburban towns to an urban metropolis. The presence of various conditions in the Humber River provides information regarding the individual and cumulative effect of urbanization and agricultural development on the river ecosystem. This river is maintained by the Toronto and Region Conservation Authority (TRCA). Sampling was conducted by Spall (2014) at the same locations that were sampled by TRCA in order to compare their results with the TRCA biomonitoring program. They used the same sampling method with some modifications to the TRCA sampling protocols.

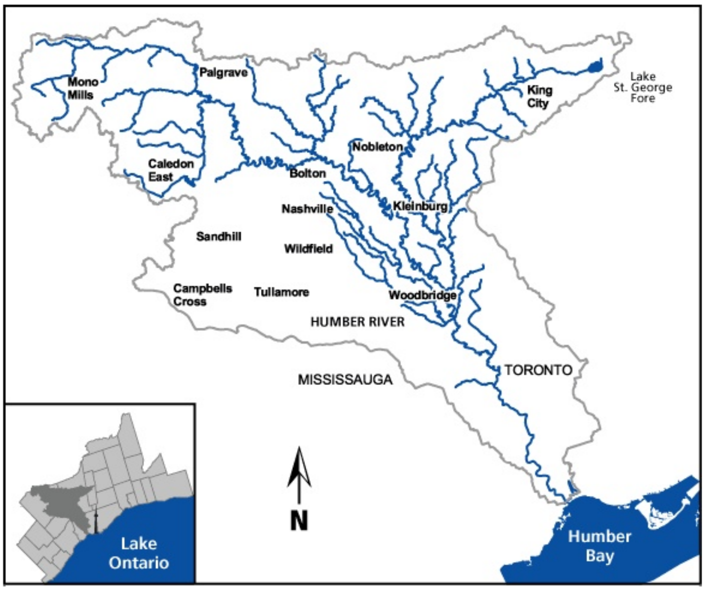


Figure 3.1: The location of Humber River in Toronto Ontario (CHRS, 2011).

The sampling conducted by the Hajibabaei lab contains 16 samples collected from 13 sites along the length of the river. Sampling was performed from August to September 2011. All the 16 samples were obtained in precisely the same manner. Benthic samples were obtained using standard D-frame kick-nets across ten transects of the river; with each transect placed at 10 m intervals. The contents were pooled into a sterilized container from which live sorting was performed in the field (Hajibabaei et al., 2011). Individual organisms were sampled using sterile forceps and placed into sterile sampling jars filled with 96 % ethanol as bulk samples (Spall, 2014).

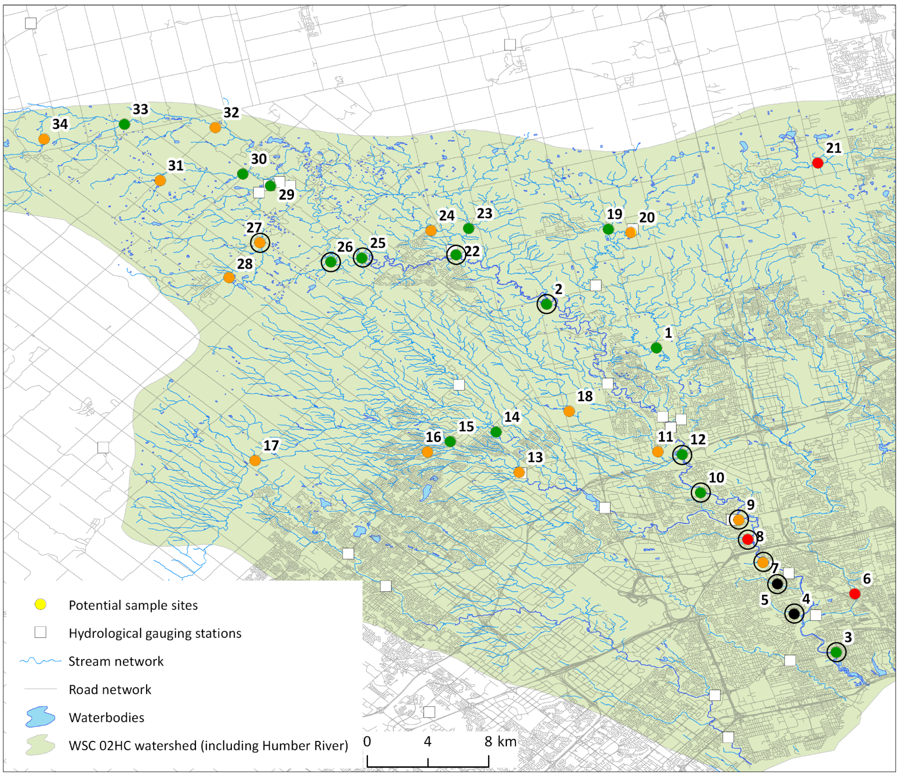


Figure 3.2: The location of Humber River sampling sites in Toronto Ontario (TRCA, 2011). The circled dots are those sites that were sampled by the Hajibabaei lab. The red, orange, green, and black dots are highly polluted, intermediate, non-polluted, and unknown ecosystems respectively.

Figure 3.2 shows all the 34 sites being monitored by TRCA annually. The sites are labeled with four distinct colors to indicate the site's condition. TRCA (2011) and Spall (2014) assigned these conditions using The Southern Ontario Land Resource Information System (SOLRIS). Those sites that were sampled by Spall (2014) are circled in black. The red points are those locations labeled as highly polluted sites. The orange points represent intermediate levels of pollution, and green dots are those from non-polluted ecosystems. The sites with unknown conditions are shown with the black points. Table A.1 illustrates the total number of samples and related sites from this study. The SOLRIS system provided the type of land cover and the other information that is found at each of the 34 locations. SOLRIS is a landscape-level inventory designed by The Ontario Ministry of Natural Resources to support planning and development projects in southern Ontario (for more information about the ecological data provided by SOLRIS see section A.1).

3.4.2 The Wood Buffalo Natinal Park

Gibson et al. (2015) studied the Peace-Athabasca delta to create baseline data to investigate the nature and importance of the biodiversity for early detection of potential anthropogenic effects from the nearby Alberta Oil Sands. A UNESCO World Heritage Site (designated 1982), the Peace-Athabasca delta is the largest inland freshwater delta complex in the world and is located in the Wood Buffalo National Park. The sampling in this study was conducted by Environment Canada and Parks Canada. These samples were collected in June 2012 from two adjacent deltas, the Peace and Athabasca deltas each containing four riverine wetland sampling locations. Three replicate samples of aquatic invertebrates, each located approximately 50 meters apart, were collected from each sampling sites. A standard Canadian Aquatic Biomonitoring Network (CABIN) pond net with a sterile 400 μM mesh, was used for collecting the samples (Gibson et al., 2015).

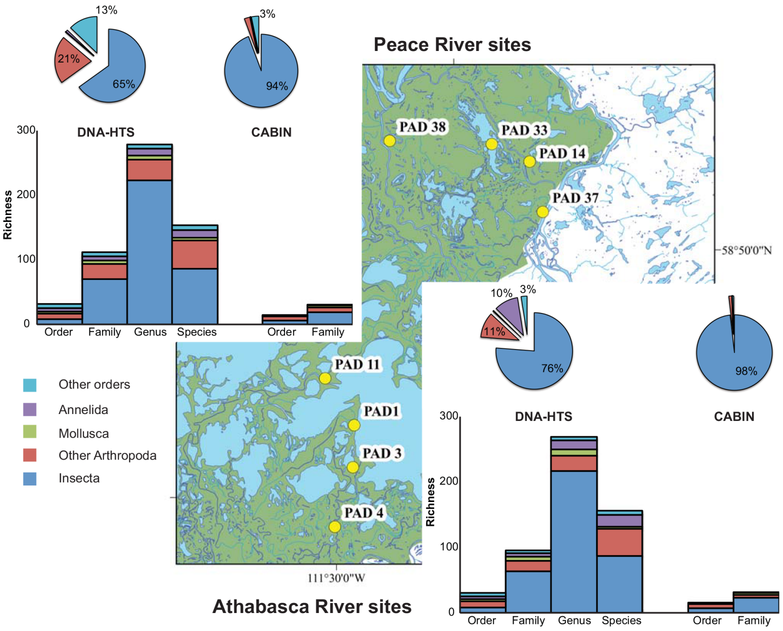


FIGURE 3.3: The location of the Peace and Athabasca River sites in the Wood Buffalo National Park from Gibson et al. (2015).

3.5 Methods

3.5.1 The Humber River samples

Library preparation

Figure 3.4 illustrates the library preparation and sequencing processes that were completed by the Hajibabaei lab at the University of Guelph (Spall, 2014). They homogenized the bulk samples in 95 % ethanol. Total DNA was extracted from the individual samples and the COI amplicon amplified using two primer pairs. The library of DNA templates were amplified from a single copy to tens of millions of copies and immobilized on beads using emulsion PCR (emPCR). Then the DNA template was sequenced utilizing the 454 Roche platform. All reagents, enzymes and primers required for the emPCR, as well as the remainder of the sequencing process are provided by Roche in the form of a sequencing kit (Spall, 2014).

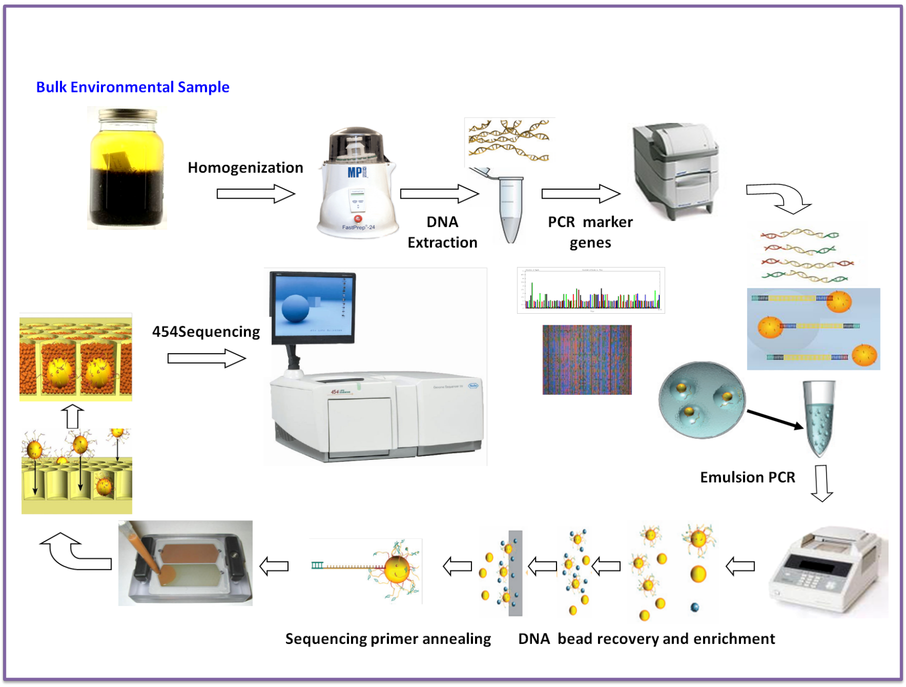

Bulk Environmental Sample
Homogenization
DNA Extraction
PCR marker genes
454Sequencing
Emulsion PCR
Sequencing primer annealing
DNA bead recovery and enrichment

FIGURE 3.4: library preparation and sequencing procedures (Spall, 2014).

Quality control and trimming

The Hajibabaei lab (Spall, 2014) provided us the raw sequencing output of 16 samples from the Humber River watershed. The 454 platform generates two files for each sequencing output; a FASTA file (`.fna`) that contains the DNA sequences and a quality control file (`.qual`) which quantifies the quality of each base in the DNA sequence. For each sample, the `.fna` and `.qual` files were combined and converted to the FASTQ format using a custom python code. The 454 machine removes the sequencing adapters by default before conversion of the `.sff` file to FASTA format in the final output. Low-quality sequences and those that were too short were removed from the dataset using the Qtrim software (Shrestha et al., 2014). Any of the sequences shorter than 120 base-pairs were trimmed as they are often of lower quality than the rest of the sequences. The reads with a mean quality score (Phred) of less than 20 (indicating a 99% accuracy per base) were also trimmed from the reads. The ambiguous bases (N_S) were removed from the middle, 3′, and 5′ end of the sequences. Those reads that satisfy all the above specifications were then converted to the FASTA format and labeled as "the clean data". Figure 3.5 shows all the post-sequencing procedures.

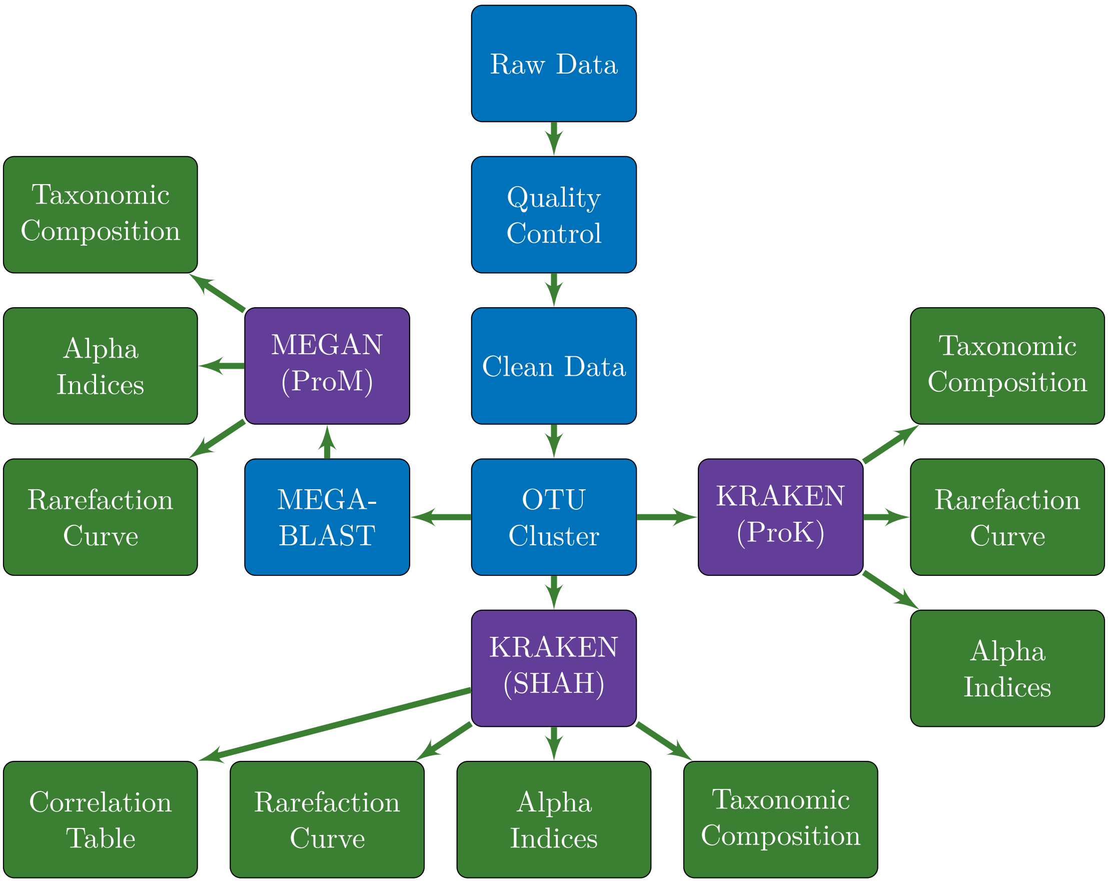


FIGURE 3.5: Schematic illustration of the steps of post-sequencing pipeline.

After filtering and trimming the low-quality reads, the UCLUST algorithm was used to dereplicate, sort, and cluster at 99 % sequencing similarity cutoff (Edgar, 2010). The `-derepfulllength` parameter was used to find exactly the same sequences and discard any exact duplicates. The reads were sorted by size employing the `-sortbysize` command in the UCLUST package. Chimera filtering was performed utilizing USEARCH with the de novo UCHIME (Edgar et al., 2011) algorithm. Amplicon sequencing may cause chimeric sequences that are formed by two or more biological sequences joined together. As the chimeric sequences may lead to false sequences, it's important to remove them in the analysis. The `-sortbylength` command in the USEARCH tool package was used to sort the reads by length, then the sequences clustered at 99 % similarity and the singletons retained using `-cluster-smallmem` command in the USEARCH program. All the filtered, dereplicated, sorted, non-chimeric, and clustered reads were recorded as OTUs.

Sequence classification

As shown in figure 3.5, we have employed three methods of classification with the previously constructed databases (Tab. 3.4) to identify the taxonomic composition of all the samples. First, we have used the MEGA-BLAST (Morgulis et al., 2008) algorithm against the local ProM database for the identification of samples. The ProM database is the custom database contains the COI gene of invertebrates that was constructed earlier. The MEGA-BLAST searches conducted with splitting the BLAST searches into small jobs and ran in parallel using the GNU Parallel (Tange et al., 2011). The BLAST hits less than or equal to the 1e-10 evalue retained, and the others discarded. The expected value (evalue) describes the number of expected hits by chance in a database of the particular size. The top hit matches with the lower common ancestor extracted using the MEGAN program (Huson et al., 2007).

When we uploaded the MEGA-BLAST output into MEGAN, we optimized the MEGAN parameters to the metagenomic setting as follows. A minimum support of 1 which refers to the minimum number of reads associated with a taxa told MEGAN that a species name only needed to appear one time in the top hits to be included in the analysis (The minimum support is set to 50 by default). The minimum score, the minimum required BLAST scores assigned to a taxa, was set to 100 (by default the minimum score in MEGAN is equal to 50). Finally, the top percent parameter was set to 10, meaning only the first 10 % of the hits were included when determining the higher-level taxonomy. The MEGAN algorithm assigned a taxonomy to each OTU using the GenBank information and these customized settings. All the identified taxa at the different levels of taxonomy were retained from the MEGAN results.

Second, we employed the KRAKEN algorithm (Wood and Salzberg, 2014) with the ProK database for the classification and identification of the Humber River samples. The ProK database is the custom database of invertebrates retrieved from GenBank earlier, and it has the same sequence content as the ProM database (Tab. 3.4). We have developed a shell script that includes `kraken`, `kraken-report`, and `kraken-mpa-report` commands from the KRAKEN tools package to classify and assign taxonomy to the OTUs. The `kraken` commands classified all the reads and assigned taxonomic IDs to each read. The `kraken-report` and `kraken-mpa-report` provide the taxonomic information and assign the number of reads to each taxa. The final output contains the number of reads covered by the clade, the number of reads directly assigned to taxa, a rank code, NCBI taxonomy, and the scientific name that will be used for the biodiversity analysis.

The last classification method as shown in figure 3.5 was to use the KRAKEN algorithm against the SHAH database. The SHAH database is the customized database of bioindicator invertebrates that was assembled earlier (see section 3.2). We have employed a shell script that contains the `kraken`, `kraken-report`, `kraken-mpa-report`, and `kraken-filter` commands from the KRKAEN package to classify and count the number of reads associated to each taxa. As mentioned earlier, the first three commands classify and assign taxonomic information to each read. The `kraken-filter` command provides a simple scoring scheme for the KRAKEN results. We have increased the accuracy of the KRAKEN algorithm in this method by setting the `-threshold` option to 0.10 in the `kraken-filter` command. The final output contains the number of reads covered by the clade, the number of reads directly assigned to taxa, a rank code, the NCBI taxonomy, and the scientific name that will be used for the biodiversity analysis.

The SHAH database is a very small database compared to the other databases that are typically used for taxonomic identification. The small size of the SHAH database may cause false identification of taxonomic composition. To prevent this possibility, we have increased the accuracy of the identification by defining a high confidence score. The `kraken-filter` command specifies a threshold to accept only reads with higher accuracy. If a taxa did not have a score exceeding the threshold, it will be labeled 'unclassified'. A sequence label score in this algorithm is the fraction of C/Q where C is the number of k-mers mapped to lowest common ancestor (LCA) value in the clade rooted at the label, and Q is the number of k-mers in the sequence that lack an ambiguous nucleotide (Wood and Salzberg, 2014). It's important to note that there is a trade-off between accuracy and sensitivity, by increasing the accuracy of classification, the sensitivity of identification will be decreased (see section 1.6 in the chapter one).

Biodiversity analysis

All the results from three different methods of classification contain four levels of taxonomy used to calculate diversity metrics. We have explored the taxonomic composition of samples from different conditions. Also, we investigated the performance of the SHAH database in estimating the composition of samples at species, genus, family, and order level of taxonomy. We have compared the rarefaction curves of the ProK and the SHAH database to test the importance of database in estimating taxonomic richness. The diversity indices, the Shannon and the Simpson index, were used to inspect the sites condition (polluted and non-polluted). The Pearson's product-moment correlation test with bootstrapping was used to determine the relationship between the SHAH database and the other databases in calculating alpha diversity (Shannon and Simpson diversity indices). For all the biodiversity analyses and exploration we used the vegan package (Oksanen et al., 2007) in R version 3.3.1 (Team et al., 2013).

3.5.2 The Wood Buffalo National Park

A dataset from the Wood Buffalo National Park (Gibson et al., 2015) was also retrieved to test the performance of the three methods of classification that we constructed earlier (Tab. 3.4). Although Gibson et al. (2015) did not conduct the sampling in polluted environments, it can be useful given there are two deltas with different biodiversity (for more information about the samples see section 3.4.2). Gibson et al. (2015) homogenized the benthic samples in 95 % ethanol, the total DNA was extracted, and the two fragments (F230, and BE) within the COI gene were amplified. The F230 fragment is 230bp in length and is found at the 5′ end of the standard barcode region. The BE fragment is 314bp and is found at 3′ end of the standard barcode region. The purified amplicons from the first PCR were used as a template for the pre-sequencing PCR using the Illumina-tailed primers. All the generated amplicons sequenced using the Illumina Miseq platform. Gibson et al. (2015) merged the forwards and reverse raw reads for all the 24 samples with a minimum 25 bp overlap. The PRINSEQ software (Schmieder and Edwards, 2011) was used to filter the paired-end reads with the minimum Phred score of 20 and minimum length of 100bp. The remaining reads dereplicated and clustered at 99 % sequence similarity using the USEARCH (Edgar, 2010) with the UCLUST algorithm. Also, Gibson et al. (2015) used the UCHIME algorithm (Edgar et al., 2011) for chimera filtering the sequence reads. We have retrieved all the filtered, dereplicated, non-chimeric and clustered reads for all the 24 samples of this study and followed our post-sequencing pipeline as figure 3.5 illustrates. All the reads were classified and identified using the three methods of classifications (Tab. 3.4). We have employed the MEGA-BLAST algorithm (Morgulis et al., 2008) against the ProM database and then imported the output into the MEGAN program (Huson et al., 2007) to find the top hits with the lowest common ancestor. The KRAKEN algorithm Wood and Salzberg, 2014 with the ProK database was used to classify and identify the reads. Finally, we have utilized the SHAH database to classify all the sequences as well. We have used the classification results for the biodiversity analysis and compared the performance of various methods of classification in estimating the α diversity. We have used the same parameters and algorithms to classify the WBNP as the Humber River samples (for more details see sequence classification in section 3.5.1).

3.6 Results

3.6.1 The performance of programs

The two main challenges in biomonitoring of eukaryotic communities are the time required and computational power for HTS analyses. To tackle these problems, we have constructed a small database of bioindicators using the KRAKEN algorithm for the first time in studying eukaryotic communities. The performance of our bioindicator database and the KRAKEN algorithm was investigated using three methods of classification to identify taxonomic composition and biodiversity analyses of two datasets (see section 3.4). Table 3.5 highlights the differences between three methods of taxonomic identification in classifying the same dataset. The comparison of the ProK with the ProM database is useful to investigate the performance of the KRAKEN algorithm. The ProK and the ProM are both using the same database, and the BLASTN method is known as a reference approach for HTS classification. Also, comparing the SHAH with the ProK database is helpful to explore the employment of a small database of bioindicator sequences in the environmental monitoring programs. The SHAH and the ProK both utilize the KRAKEN algorithm.

As shown in table 3.5 the SHAH approach contains 1,927 species and 20,235 sequences. For each species we extracted around ten accessions. For the species with unavailable COI gene records in the NCBI, we have retrieved the Cytb genes from NCBI. The ProM and the ProK include substantially larger databases compared to the SHAH with 1,147,778 sequences and 167,788 species. The SHAH and the ProK classified and identified the HTS dataset in 0.5 and 4.5 hours respectively. However, the ProM method classified the HTS reads in 32 hours using a cluster machine with 48 cores in parallel. The ProM method required 405 megabytes for the database and 1,296 gigabytes of disk space for the output files. The ProK needed 30 gigabytes and 713 megabytes of disk space for the database and output files respectively. And the SHAH generated a database of 8 gigabytes and 700 megabytes of the output file. All three approaches are very sensitive identifiers as we used a minimum support of close to one to assign the taxonomy. They are all fairly accurate methods with 98.25, 95.43, and 100 scores for the SHAH, ProK, and ProM respectively. These scores are not calculated consistently for the KRAKEN and the MEGA-BLAST algorithms, but they were suggested by the KRAKEN authors Wood and Salzberg (2014) as the thresholds levels. An Illumina amplicon dataset from The Wood Buffalo National Park (WBNP) with 24 samples that each contained 88,000 reads and 295 bp lengths in average was used to generate table 3.5.

Parameter	**SHAH**	**ProK**	**ProM**
Number of species	1927	167,788	167,788
Number of sequences	20,235	1,147,778	1,147,778
Algorithm	KRAKEN	KRAKEN	BlastN
Method	K-mer based	k-mer based	Homology based
Speed	0.5h	4.5h	32h
Database Size	8G	30G	405M
Output Size	700M	713M	1,296G
Min score	98.25	95.43	100
Min support	1read	1read	1read

TABLE 3.5: The performance of the three customized HTS taxonomic classification and identification approaches.

3.6.2 Taxonomic composition

The taxonomic composition of the Humber River samples was explored using the three different databases (Tab. 3.4). A summary of the taxonomic composition is given in figures 3.6 and 3.7 employing the ProK method that contains a custom reference database of invertebrates. The Humber River samples were separated and labeled into different columns according to their ecological conditions (for more information about the sampling sites and their conditions see section 3.4). The top abundant taxa were colored and labeled in figure 3.6, and the remaining low abundant taxa are shown in gray. Figure 3.6 illustrates that the diversity and the composition of samples are not particularly different in various sampling locations with diverse level of pollution. However, figure 3.7 shows that the species diversity is higher in non-polluted environments compared to polluted environments. The species diversity decreased with the increase of environmental pollutants. Comparing the composition of samples at the different levels of taxonomy indicates that a higher level of taxonomy is less responsive to environmental conditions (see figures A.1 and A.2 for other levels of taxonomy). Also, the comparison of the unknown environment condition with other ecological conditions in the Humber River samples, suggests that the composition of unknown samples are very similar to samples labeled as polluted. This result is also consistent at various level of taxonomy (Figs. 3.6, 3.7, A.1, A.2).

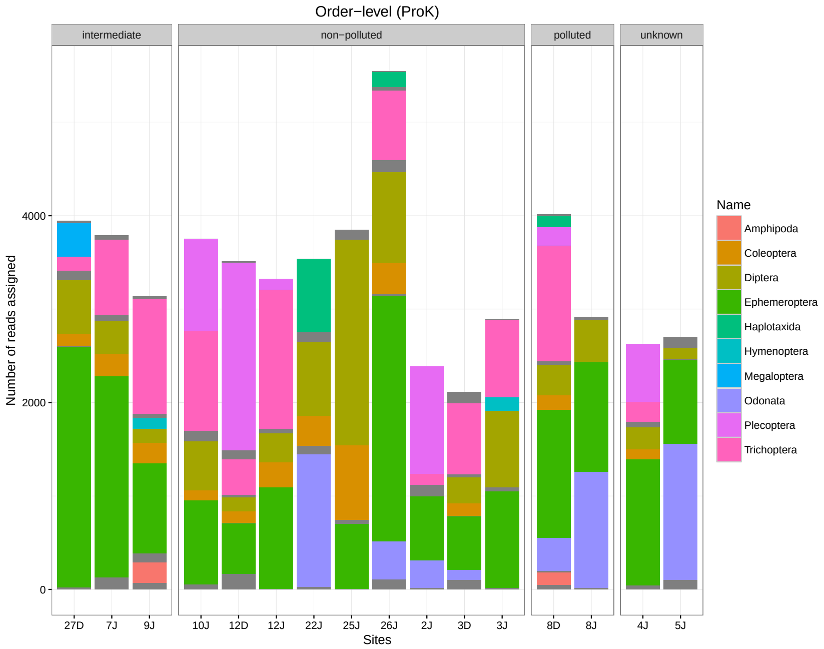


FIGURE 3.6: Taxonomic composition of the Humber River samples using the ProK method at the order-level.

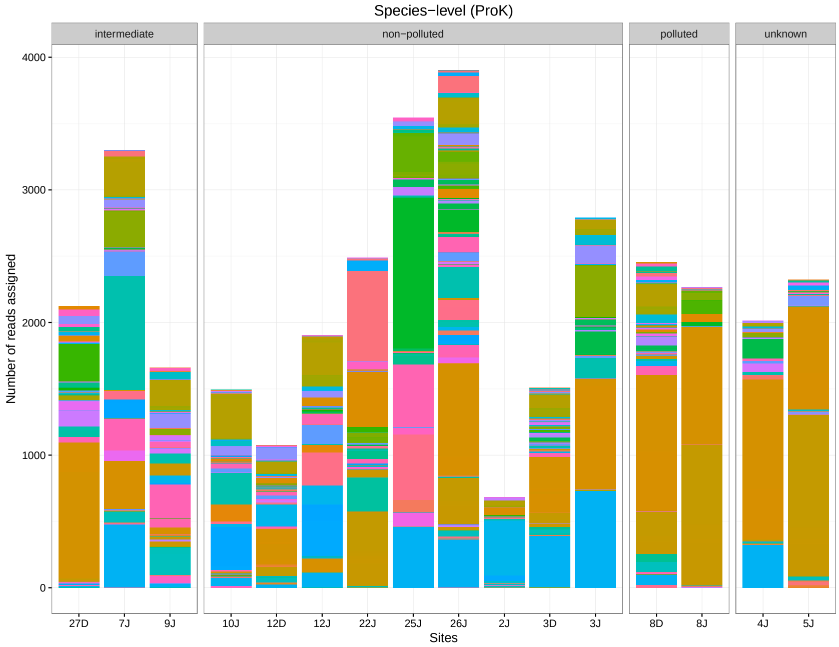


FIGURE 3.7: Taxonomic composition of the Humber River samples using the ProK method at the species-level.

We have compared the SHAH with the Prok classification approach in estimating the taxonomic composition of the Humber River samples. The SHAH employs a small database contains 1927 species while the ProK includes 167,788 species and 1,147,778 sequences. Figures 3.8 and 3.9 show a summary of the most abundant taxa at the order and family-level respectively. The performance of the SHAH and the ProK is compared at different ecological conditions. Each row in figures 3.8 and 3.9 compares the samples with the same ecological condition. The top abundant taxa were colored and labeled in figures 3.8 and 3.9, the remaining low abundant taxa are shown in gray. The x-axis shows all the samples from the Humber River that were identified using the ProK (on the left column) and the SHAH database (on the right). The y-axis illustrates the number of reads that assigned to a taxa (Figs. 3.8, 3.9). We have also compared these two classifier in estimating the taxonomic composition at the genus and species level (Figs. A.3, A.4).

Investigating the performance of the SHAH database compared to the ProK suggests that the SHAH identifier was able to estimate the compositions of the samples in polluted and unknown ecosystems. The SHAH database identified the most abundant taxa in sites 8D, 8J from polluted locations and sites 4J, 5J from unknown sampling locations accurately. This result is consistent at all four levels of taxonomy (Figs. 3.8, 3.9, A.3, A.4). However, the SHAH database was not able to estimate the taxonomic composition in non-polluted sampling sites. The composition of the most abundant taxa in sites 10J, 12D, 12J, 22J, 25J, 26J from non-polluted ecological condition is different between the ProK and the SHAH database. This difference is consistent among four level of taxonomy (Figs. 3.8, 3.9, A.3, A.4). The comparison of the SHAH and the ProK for the sites with intermediate level of pollution indicates that the SHAH result is not consistent with all the samples and taxonomic levels. The SHAH identified the most abundant taxa in the 27d site at the order, family, and genus levels while the composition of the 7J and 9J were not estimated accurately at various level of taxonomy. Also, we observed false taxonomic identification in sites 7J and 9J at family, genus and species levels using the SHAH approach.

3.6.3 Rarefaction curve

Rarefaction is an ecological technique to test if the difference in sample size causes the difference in richness. Obviously, species richness increases with sample size, but this increase should not affect the ecological interpretations. The rarefaction curve examines species numbers as a function of sample size. This technique calculates the expected number of species in a small collection of n individuals by randomly sampling from the large pool of N samples. The curves with steep slope illustrate that a large number of species diversity remains to be identified. However, the flatter curves suggest that an enough number of individuals or samples have been collected from the environment.

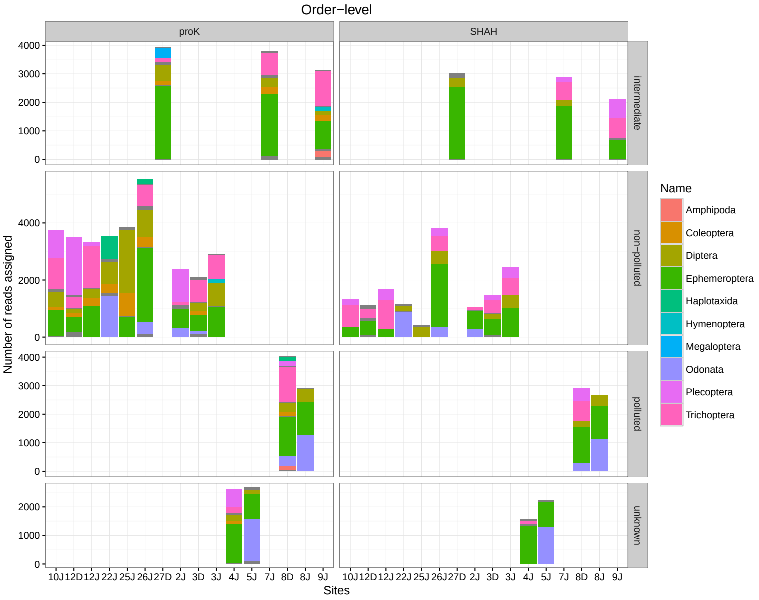


FIGURE 3.8: A comparison between the SHAH and the Prok approaches in estimating composition of the samples at the order-level. Each row compares the samples with the same ecological condition.

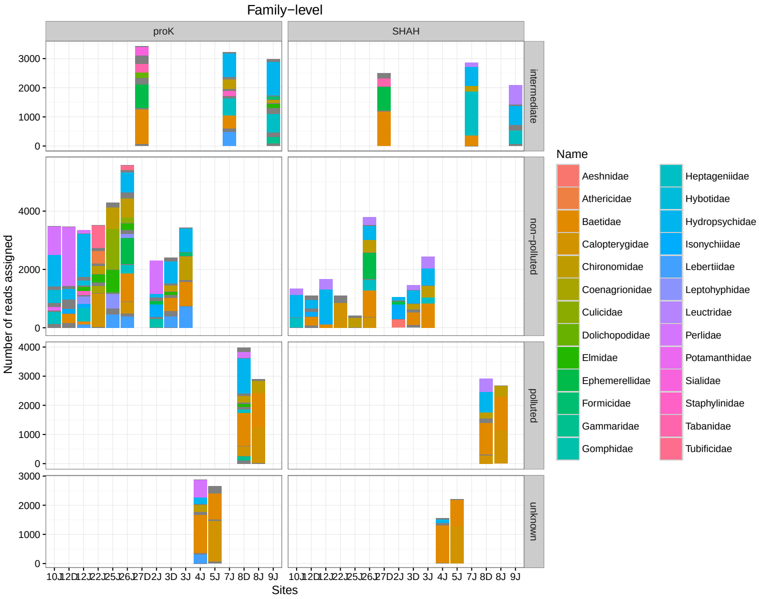


FIGURE 3.9: A comparison between the SHAH and the Prok approaches in estimating composition of the samples at the family-level. Each row compares the samples with the same ecological condition.

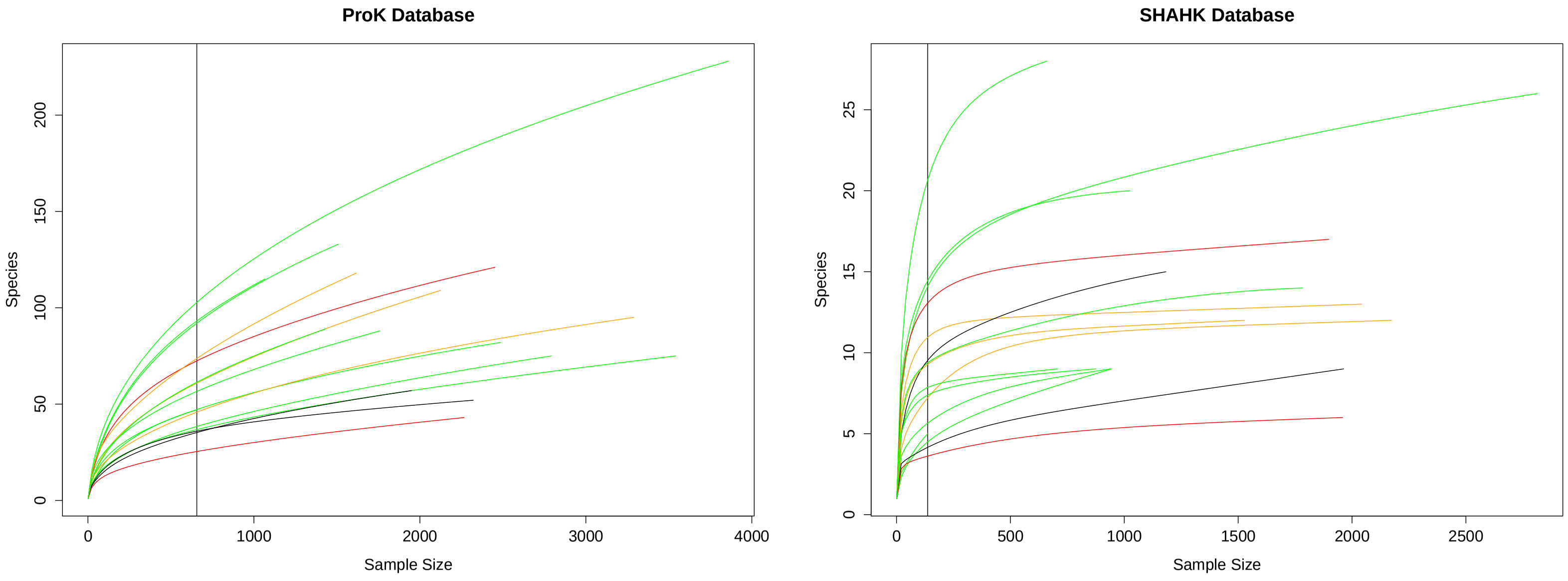
Figure 3.10 compares the rarefaction analyses constructed for the SHAH (Fig. 3.10b) and the ProK (Fig. 3.10a) databases. Each curve shows one site and curves are colored based on the ecological conditions of sampling sites. The non-polluted, polluted, intermediate, and unknown sampling sites are shown in green, red, orange, and black respectively. The x-axis shows the sample size (or the number of sequence reads), and the y-axis is the number of species (or OTUs that assigned to species). The vertical thresholds in figures 3.10a and 3.10b that cut the curves are the minimum sample count achieved over all the samples.

The comparison of the SHAH (Fig. 3.10b) and ProK (Fig. 3.10a) indicates that the SHAH approach requires smaller sample sizes compared to the ProK to estimate the taxonomic richness. As shown in figure 3.10 the rarefaction curves calculated for all the sites with the SHAH database plateau at around 140 number of sequences. However, the rarefaction curves constructed with the ProK method shows steeper slopes compared to the SHAH in various sampling locations. Also, this result suggests that the polluted sites (red curves) require smaller sample sizes to reach the saturation level compared to the non-polluted ecosystems (green cures) calculated by both methods. Although there are different expected numbers of species at the various sites for the same sample size, all the sampling locations plateau at around same number of sequences. The same saturation level for all the sites in one classification method indicates that the differences between the sites are not due to the size of the samples (Fig. 3.10).

3.6.4 Alpha diversity

TRCA (2011) and Spall (2014) labeled the Humber River sampling locations based on the environmental information provided by The Southern Ontario Land Resource Information System (for more information see sections 3.4 and A.1). We have calculated the biodiversity indices using three methods of identification (Tab. 3.5) and compared the results at various ecological conditions. Figure 3.11 shows the comparison of the SHAH, ProK, and ProM databases in calculating the Simpson index for various environmental conditions at the genus-level. The performance of different identification methods is separated in different boxes and various conditions are shown in distinct colors. For all the samples collected from the same condition and identified with the same method a box-plot is constructed (Fig. 3.11). Both Simpson and Shannon indicies are calculated at the four levels of taxonomy (Figs. 3.11, A.5, A.6, A.7, A.8).

Unfortunately, the number of samples collected from various environmental conditions in the Humber River sites are not equal. To tackle this problem we added type category to the collected samples based on the sampling locations and the ecological conditions. The samples collected from polluted, intermediate, unknown conditions are contaminated with anthropogenic drivers to a different extent. Those sites with polluted, intermediate, and unknown conditions were labeled as contaminated sites. As we mentioned earlier, the taxonomic composition of the Humber River sites showed that the composition of samples collected from the unknown conditions are pretty similar



(A) ProK

(B) SHAH

FIGURE 3.10: Comparison of rarefaction analyses between the ProK 3.10a and the SHAH 3.10b HTS classification methods using a reference sequence library of the invertebrates and a small library of bioindicators respectively. Each curve shows one site and curves are colored based on the ecological conditions of sampling sites. The non-polluted, polluted, intermediate, and unknown sampling sites are shown in green, red, orange, and black respectively.

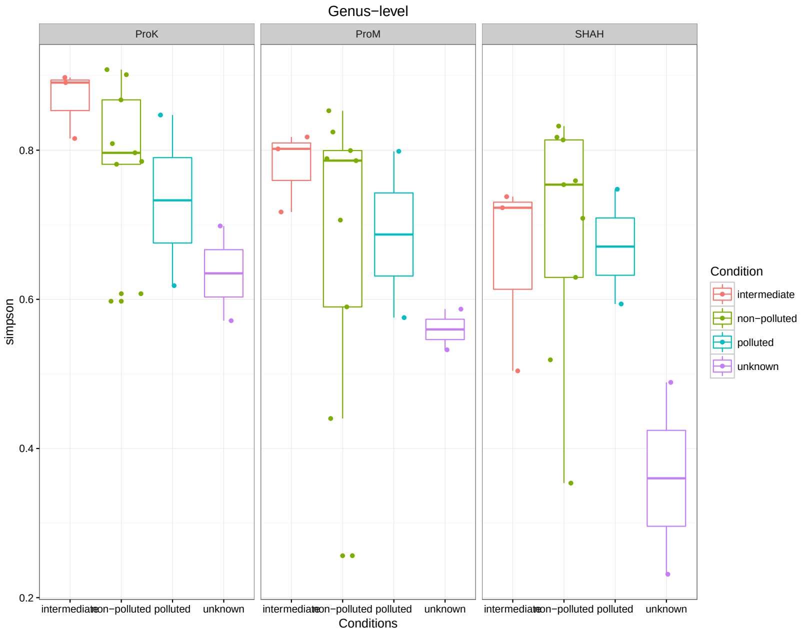


FIGURE 3.11: Comparison of the SHAH, Prok, and ProM in calculating the Simpson index for various environmental conditions at genus-level.

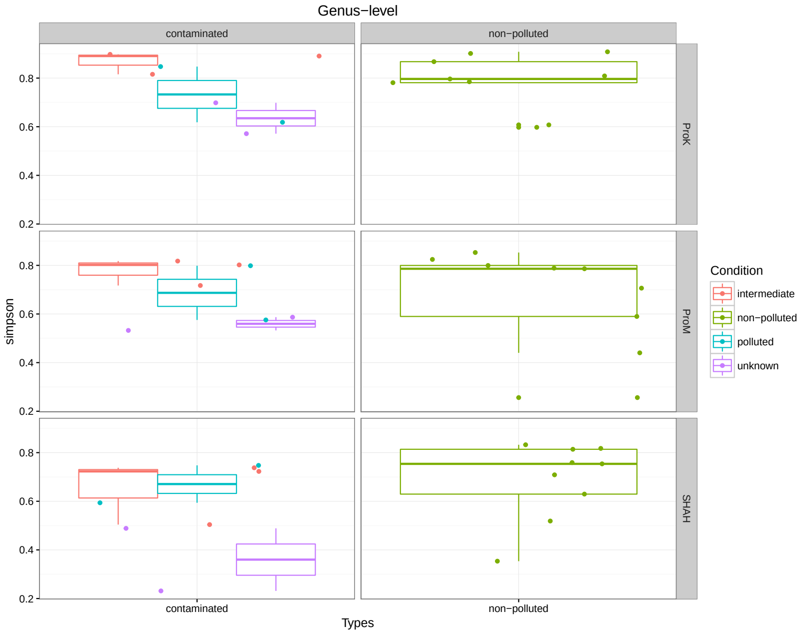


FIGURE 3.12: Comparison of the SHAH, Prok, and ProM in calculating the Simpson index for various environmental conditions and types at genus-level.

to polluted ecological conditions. Figure 3.12 illustrates the comparison of the SHAH, ProK, and ProM in estimating α diversity in various types of environments. The contaminated sites that include polluted, intermediate, and unknown conditions are shown in the left column and the non-polluted sites are located in the right column (Figs. 3.12). The performance of different methods are compared for various types and conditions and separated in the three rows in figure 3.12. Both Simpson and Shannon indicies are calculated to investigate the performance of multiple identification programs in various types of ecosystems at four level of taxonomy (Figs. 3.12, A.9, A.10).

Our results show that the SHAH was able to estimate the biodiversity accurately in the polluted environments compared to the ProK and ProM. Figure 3.11 indicates that the mean of Simpson calculated using the SHAH's result in polluted sites (the blue boxplots) is similar to the ProK and ProM that uses an enormous reference database. The similarity between the SHAH and other methods in polluted ecosystems are shown at different levels of taxonomy (Figs. 3.11, A.5, A.6, A.7, A.8). However, we were not able to identify a significant correlation between the SHAH and other methods in non-polluted environments. Unfortunately, the sampling in the Humber River sites was not conducted in the same fashion in various conditions. The absence of the same experimental design for all the sampling locations makes it difficult to interpret the correlation between conditions and the α diversity using multiple HTS classification methods. When we explored the performance of our programs within the type category the relations between the SHAH and other methods were more evident. As shown in figure 3.12 the SHAH estimated the biodiversity patterns in various conditions within the contaminated samples correctly. The similarity between the SHAH and other methods for the polluted, intermediate, and unknown sites are shown in contaminated section (left column) in figures 3.12, A.9, A.10. We used the Pearson-moment correlation by bootstrapping with 1000 replication to quantify the correlation between the identification methods in contaminated and non-polluted environments (Tab. 3.6).

We have analyzed the performance of three identification methods at various sites independently. Figures 3.13 and 3.14 compare the SHAH, ProM, ProK in calculating the Shannon and the Simpson index for each location. The x-axis is different sampling sites, and the y-axis is the Shannon index. The sites are separated based on their ecological conditions. As shown in figure 3.13 the individual Shannon values calculated by the SHAH is significantly correlated to the other methods in polluted, intermediate, and unknown conditions. The relation between the SHAH and other approaches are also shown in table 3.6. The Pearson-correlation table indicates that the Simpson and Shannon indices calculated by the SHAH method are significantly ($p<0.05$) correlated to the both ProK and ProM at four level of taxonomy in contaminated ecosystems. Although there are positive correlations in some of analyses between the SHAH and other methods in non-polluted environments, we were not able to identify a significant correlation between them.

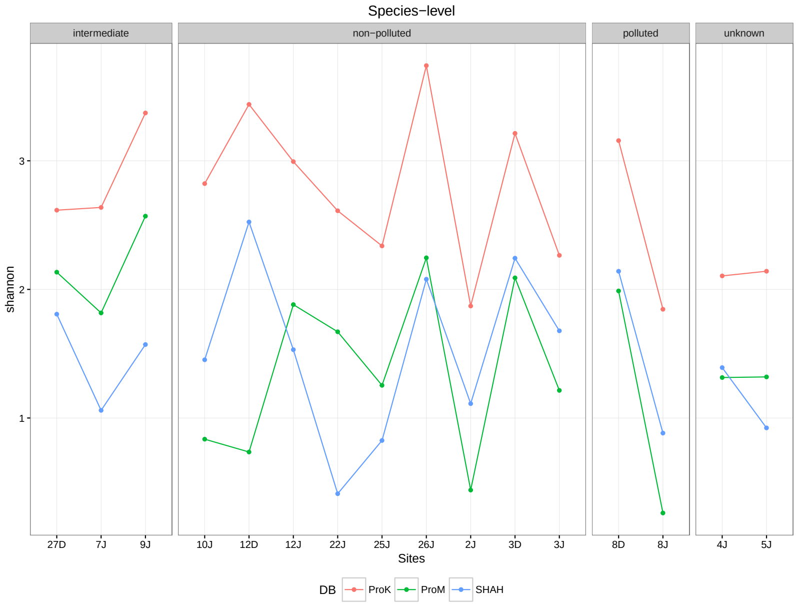


FIGURE 3.13: Correlation between the classification methods using the Shannon index at species-level. Each point show an individual site and the programs are illustrated with distinct colors.

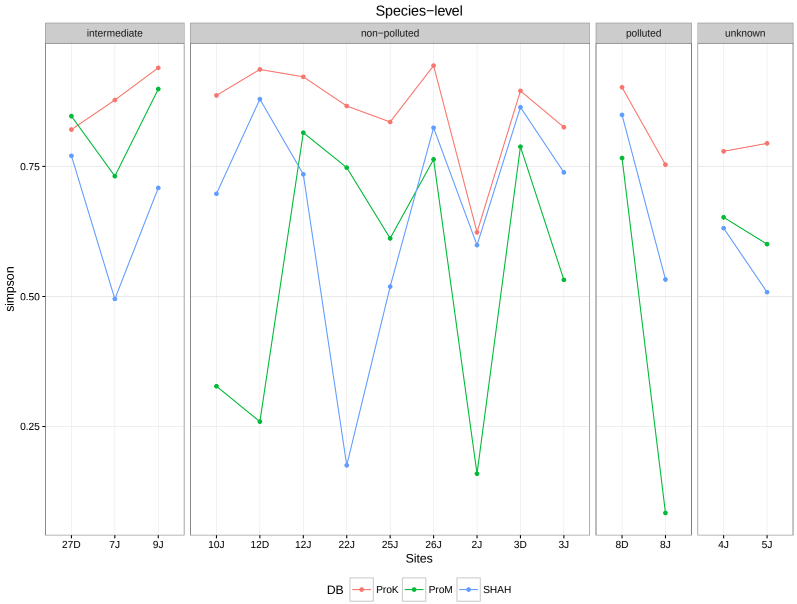


FIGURE 3.14: Correlation between the classification methods using the Simpson index methods at species-level. Each point show an individual site and the programs are illustrated with distinct colors.

	Simpson		Shannon	
	non-polluted	Contaminated	non-polluted	Contaminated
		Species		
(SHAH)VS(ProK)	0.323	0.481	**0.694**	**0.713**
(SHAH)VS(ProM)	-0.115	**0.541**	0.148	**0.667**
		Genus		
(SHAH)VS(ProK)	0.113	0.47	0.38	**0.764**
(SHAH)VS(ProM)	-0.08	**0.601**	0.069	**0.788**
		Family		
(SHAH)VS(ProK)	-0.114	**0.517**	0.25	**0.817**
(SHAH)VS(ProM)	-0.174	0.392	0.038	**0.742**
		Order		
(SHAH)VS(ProK)	0.4	**0.702**	0.577	**0.72**
(SHAH)VS(ProM)	-0.201	**0.698**	0.11	**0.775**

Table 3.6: Correlation between the SHAH and ProM or ProK in non-polluted and contaminated environments. The bold values indicate a positive significant ($p<0.05$) correlation.

Gibson et al. (2015) argued that the two deltas in the Wood Buffalo National Park have different biodiversity (see section 3.4 for more information about the dataset). We retrieved their raw data, analyzed with our post-sequencing pipeline (Fig. 3.5) and used this baseline data to test the performance of our HTS classification methods in such environments. Figures 3.15, 3.16 compare the biodiversity between the Athabasca and Peace Delta using the the three classification approaches. The x-axis shows all the sites within each delta and the y-axis illustrates the biodiversity index. Each point indicates the mean index for three samples and various identification methods are shown with different colors. The Shannon and the Simpson index were calculated for all the samples at four levels of taxonomy (Fig. 3.15, 3.16, A.11, A.12).

Our result for the WBNP is consistent with Gibson et al. (2015) conclusions; there are different biodiversities in the Athabasca River sites and the Peace River sites. All our classification methods (Tab. 3.5) estimated similar biodiversity in the Peace delta, for the sites 14, 33, 37, 38. As shown in figures 3.15, 3.16 the performance of the SHAH correlated with the ProK or the ProM. However, in the Athabasca Delta the ProM, ProK, and SHAH calculated different values for the same sites. The difference between the identification approaches in computing the diversity index highlights the importance of a standard approach in DNA-based environment monitoring. Our result illustrates the importance of the accuracy and sensitivity of the classification approaches as various methods may generate different estimations (Fig. 3.15).

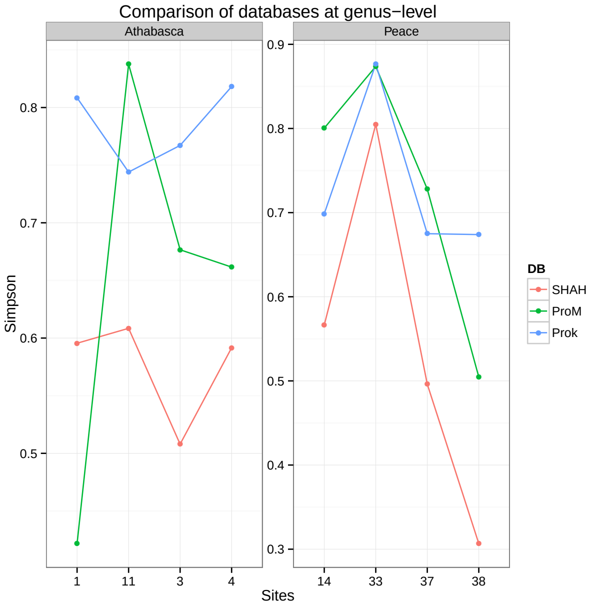


FIGURE 3.15: Correlation between the classification methods using the Simpson index at genus-level. Each point shows the mean of the Simpson index for each site and the programs are labeled with distinct colors.

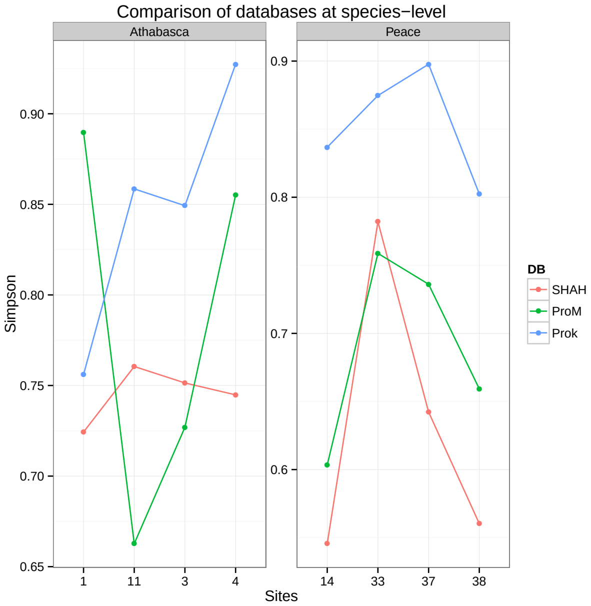


FIGURE 3.16: Correlation between the classification methods using the Simpson index at species-level. Each point shows the mean of the Simpson index for each site and the programs are labeled with distinct colors.

3.7 Discussion

The taxonomic composition of the Humber River samples using the ProK showed that the taxonomic diversity decreases with increasing pollutants in the ecosystem (Figs. 3.7, A.1, A.2. Nevertheless, the lower diversity in contaminated environments can not be seen at high-levels of taxonomy (Fig. 3.6). Previously, a higher level of taxonomy was used for monitoring programs because of the identification errors at lower levels (Jones, 2008; Sweeney et al., 2011). But, our result indicates that the biodiversity responses are less clear in the higher level of organization. This issue has been highlighted earlier by Hewlett (2000) and Pettigrove and Hoffmann (2005) in analyzing eukaryotic communities.

The SHAH, an HTS database that uses a small library of bioindicators sequences and the KRAKEN algorithm, was able to estimate the summary of taxonomic composition in the polluted ecosystems accurately (Figs. 3.8, 3.9). The capability of the SHAH can be explained through the lower diversity in such environments and presence of a targeted database at the core of the SHAH approach. The SHAH employs an extremely small DNA sequence library (20,235 sequences) compared to the GenBank database or even a custom reference database such as the ProM (1,147,778 sequences). The SHAH library was built upon the past knowledge of bioindicators, where the environmental ecologist predicted the presence of certain taxa in the particular type of ecosystems. The tiny database of the SHAH was enough to identify the composition of the samples derived from the polluted environment because the ecologist's predictions regarding the bioindicators were accurate.

As shown in figures (Figs. 3.8, 3.9), the SHAH was not able to estimate the composition of the samples in non-polluted environments because the SHAH did not contain enough sequences to accurately identify the taxonomic composition in these environments. Although the incapability of the SHAH in such ecosystem was anticipated, the false identification of the SHAH for the three samples (27D, 7J, 9J) with the intermediate level of pollution raised concerns about the accuracy of the SHAH (Figs. A.3, A.4). The KRAKEN includes a threshold tool that can be customized to obtain the higher level of accuracy. This tool should be used depending on the size of the database and the target environment because it can massively affect the sensitivity and accuracy of the identification using KRAKEN. The SHAH contains limited number of sequences, when its being used for the identification of samples with high diversity it requires a higher threshold value to avoid false identification in those cases where a direct match between the database and query sequence is not found.

We used the two datasets (see section 3.4 for more information) to investigate the performance of the three customized HST classification methods (Fig. 3.5). Unfortunately, the sampling in the Humber River watershed was not conducted consistently through various conditions and locations. The inconsistency of the sampling made our interpretation difficult for the α diversity analysis. As we did not have the same number of samples for each condition, the samples were categorized into non-polluted and contaminated ecosystems. We could not find a significant relationship between the SHAH and other methods in non-polluted samples because the SHAH was not able to identify the core composition of theses samples. It seems that the customized identification methods such as the SHAH can not be used for the regular biodiversity assessment of all environments. Customized databases should be constructed specifically for targeted environments. However, our result illustrated that the SHAH was able to calculate the biodiversity metrics at different levels of organization the same as ProK and ProM in the environments contaminated with anthropogenic pollutants. As shown in table 3.6 there was a significant correlation between SHAH and ProK or ProM in calculating the diversity indices. The relation between SHAH and ProK or ProM in the polluted ecosystems is due to the accurate identification of SHAH and the nature of the diversity index. As discussed in chapter two (see section 2.3.3) in α diversity analyses the size of the database is not as important as the ability of a database to identify the core taxonomic composition of the samples.

The capabilities of SHAH apply to any other datasets collected from the environment with lower diversity and known taxonomic composition. For instance, the SHAH estimated the biodiversity of the Peace Delta in the Wood Buffalo National Park the same as ProM, which is a popular, but time-consuming approach of identification (Fig. 3.15). But, the absence of a standard approach in taxonomic identification of the eukaryotic communities made it difficult to compare the methods and to determine the accuracy of each approach. The ProK and ProM with exactly the same database calculated different values for the same sites in the Athabasca Delta (Fig. 3.15). The difference is caused by the nature of their algorithms and the accuracy thresholds that we intentionally defined for them. Based on our settings ProM is the more accurate and less sensitive program compared to the ProK, and the SHAH stands between the ProK and the ProM regarding accuracy. For the compatibility and repeatability of the identification the presence of a standard classifier for the COI amplicons is necessary. For example, the Greengenes (DeSantis et al., 2006) database is known as a standard method for HTS classification of 16S rRNA amplicons.

We utilized the KRAKEN algorithm for the first time in classifying eukaryotic communities using the large size and extremely small databases. KRAKEN is a potential algorithm for simultaneous DNA-based monitoring of environments. This program can overcome some of the current challenges in the metabarcoding with substantial time and computer power reductions (Tab. 3.5). Researchers can simply classify and identify their large HTS datasets using their laptop in a short period. The use of the customized databases (e.g. the SHAH) can facilitate the process and the required disk space with the same level of accuracy.

The rarefaction analyses showed that the smaller size databases such as the SHAH require smaller sample sizes to reach the saturation level for taxonomic richness specifically in the polluted ecosystems. In other words, the customized databases such as the SHAH can help to reduce the cost of the biomonitoring programs as they require fewer read numbers. This conclusion is compatible with our result from chapter 2 where we showed the taxonomic identification of samples with anthropogenic pollutants requires fewer numbers of sequencing reads. One possible approach to overcome the high cost of HTS is to increase the number of samples multiplexed (tagged) in an HTS run so that the efficiency of one sequencing run with fixed read number would be maximized. Simultaneously, increasing the number of samples multiplexed in a run will generate fewer read numbers for each sample.

There are some concerns regarding the reliability, and the accuracy of the HTS counts as an absolute abundance in ecological assessments due to the polymerase chain reaction (PCR) amplifications and sequencing biases (Amend, Seifert, and Bruns, 2010). Although algorithms such as UCLUST finds the clustered operational taxonomic unit (OTU) that are much closer to the number of species, we can not argue that the OTUs are the absolute relative species abundance. There are other sequencing techniques such as shotgun sequencing (Simon and Daniel, 2011) and gene enrichment techniques (Dowle et al., 2015) that don't use fragment amplification (PCR) to characterize species communities. But, both methods are currently too expensive for rapid identification of bulk environmental samples.

In future, efforts should be made towards the establishment of a standard approach specifically for identification of the bulk eukaryotic communities. The optimum level of accuracy and sensitivity should be identified using various programs. This standard level can be used for the reliability and repeatability of DNA-based environmental monitoring. The SHAH database can be improved by adding more bioindicator species that depend on the target environment. The presence of various types of small customized DNA databases based on the target environment may lead to a decreasing the cost of biomonitoring programs and an increasing the efficiency. The KRAKEN algorithm is a potential algorithm for the regular HTS identification instead of homology-based programs. Currently, the KRAKEN package contains only prokaryotes database with the default settings, but the other databases for various groups of the organism should be constructed to be available for common usage.

Appendix A

Chapter 3 Supplements

A.1 Land Cover DATA

Sample	Date	Site	Condition	Type
1	June	2	non-polluted	non-contaminated
10	June	22	non-polluted	non-contaminatedd
11	June	25	non-polluted	non-contaminated
12	June	26	non-polluted	non-contaminated
13	Dec	3	non-polluted	non-contaminated
15	Dec	12	non-polluted	non-contaminated
2	June	3	non-polluted	non-contaminated
8	June	10	non-polluted	non-contaminated
9	June	12	non-polluted	non-contaminated
14	Dec	8	polluted	contaminated
6	June	8	polluted	contaminated
3	June	4	unknown	contaminated
4	June	5	unknown	contaminated
7	June	9	intermediate	contaminated
5	June	7	intermediate	contaminated
16	Dec	27	intermediate	contaminated

Table A.1: The available samples from the Humber River watershed. All the samples collected in 2011 and sequenced in June and December 2012.

Table A.2 shows the land cover data provided by SOLRIS for four of the sites in the Humber River sampling location. The data provided by The Southern Ontario Land Resource Information System (SOLRIS) which is a landscape-level inventory designed by The Ontario Ministry of Natural Resources to support planning and development projects in the south of Ontario Spall (2014).

Site	Forest	Mixed forest	Deciduous forest	Coniferous forest	Hedge rows	Plantations	Built-up pervious	Built-up impervious	Transportation	Open water	Swamp	Marsh	Bog	Fen	Extraction	Un-differentiated
3	0.98%	4.47%	5.53%	1.27%	0.83%	2.84%	3.03%	17.67%	6.81%	0.57%	4.55%	0.92%	0.02%	0.00%	0.14%	50.37%
8	0.54%	0.58%	3.47%	0.09%	0.74%	0.25%	2.23%	14.86%	5.52%	0.30%	2.89%	1.42%	0.01%	0.00%	0.01%	67.07%
12	1.31%	6.92%	7.92%	1.99%	0.99%	4.87%	1.97%	8.86%	4.51%	0.84%	6.33%	1.03%	0.02%	0.00%	0.25%	52.18%
27	0.77%	3.05%	4.57%	1.20%	0.62%	3.71%	0.00%	0.70%	2.04%	0.44%	4.47%	0.91%	0.03%	0.00%	0.01%	77.49%

Parameters	Description
Forest	Tree cover 60%; Upland tree species 75%; Canopy cover 2m in height
Mixed forest	Tree cover 60%; Upland conifer and deciduous tree species 25% each; Canopy cover 2m
Deciduous forest	Tree cover 60%; Upland deciduous tree species 75%; Canopy cover 2m
Coniferous forest	Tree cover 60%; Upland conifer tree species 75%; Canopy cover 2m
Hedge rows	Tree cover 60%; height 2m; linear arrangement; 10m < width 30m
Plantations (tree cultivated)	Tree cover 60%; height 2m; linear organization; uniform tree type
Built-up area (pervious)	Urban recreation areas (golf courses, playing fields etc)
Built-up area (impervious)	Residential, industrial, commercial and civic areas
Transportation	Highways; Roads
Open Water	No macrophyte vegetation, trees or shrub cover
Swamp	Open, shrub and treed communities; water table seasonally or permanently at, near or above substrate surface; tree or shrub cover 25%; dominated by hydrophytic shrub and tree species
Marsh	Open, shrub and treed communities; water table seasonally or permanently at, near or above substrate surface; tree or shrub cover < 25%; dominated by emergent hydrophytic macrophytes
Bog	Open, shrub and treed communities; water table seasonally or permanently at, near or above substrate surface; tree cover (trees 2m) < 25% sphagnum peat substrate
Fen	Open, shrub and treed communities; water table seasonally or permanently at, near or above substrate surface; tree cover (trees 2m) < 25%; sedges, grasses and low (<2m) shrubs dominate, sedge and brown moss peat substrate
Extraction	Pits, quarries
Undifferentiated	Data not available

TABLE A.2: Land cover data provided by SOLRIS

A.2 Taxonomic composition

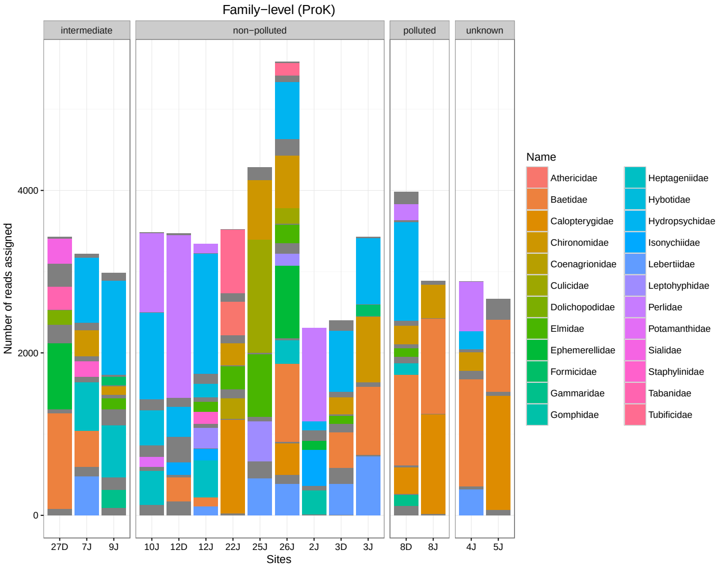


Figure A.1: Taxonomic composition of the Humber River samples using the ProK method at the family-level.

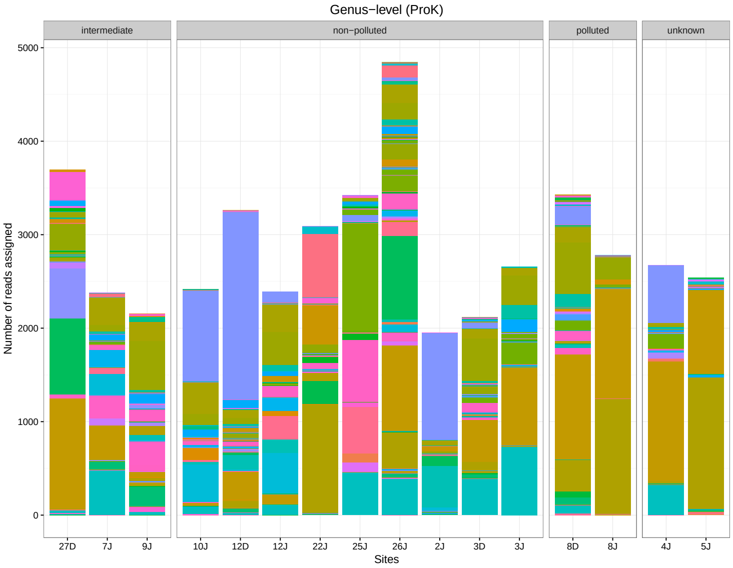


Figure A.2: Taxonomic composition of the Humber River samples using the ProK method at the genus-level.

A.3 Alpha diversity analyses

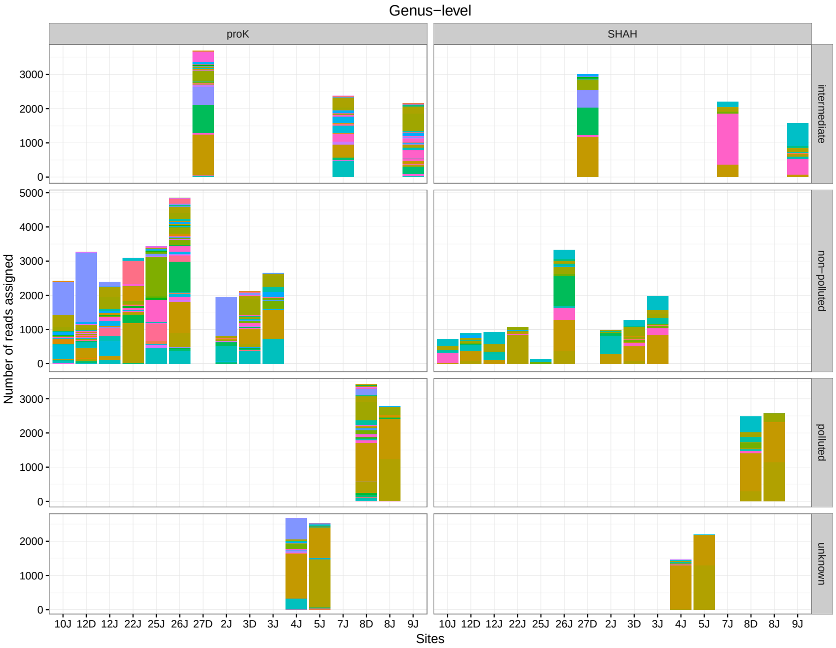


Figure A.3: A comparison between the SHAH and the Prok approaches in estimating composition of the samples at the genus-level of taxonomy. Each row compares the samples with the same ecological condition.

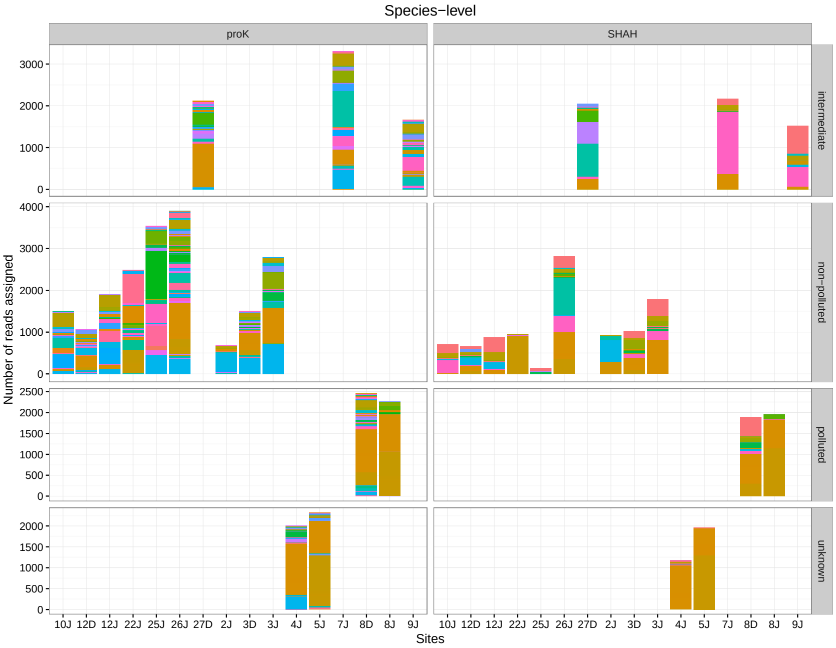


Figure A.4: A comparison between the SHAH and the Prok approaches in estimating composition of the samples at the species-level of taxonomy. Each row compares the samples with the same ecological condition.

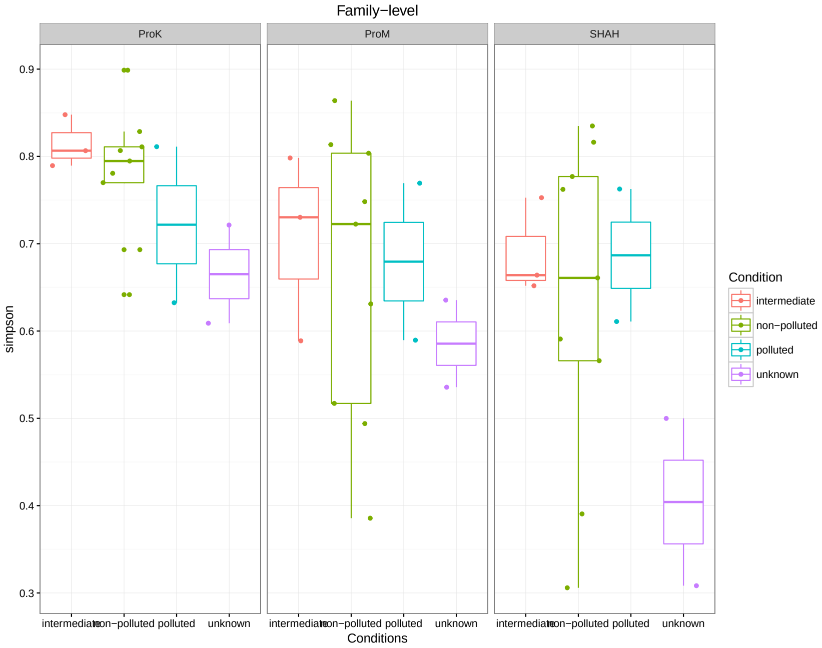


Figure A.5: Comparison of the SHAH, Prok, and ProM in calculating the Simpson index for various environmental conditions at family-level.

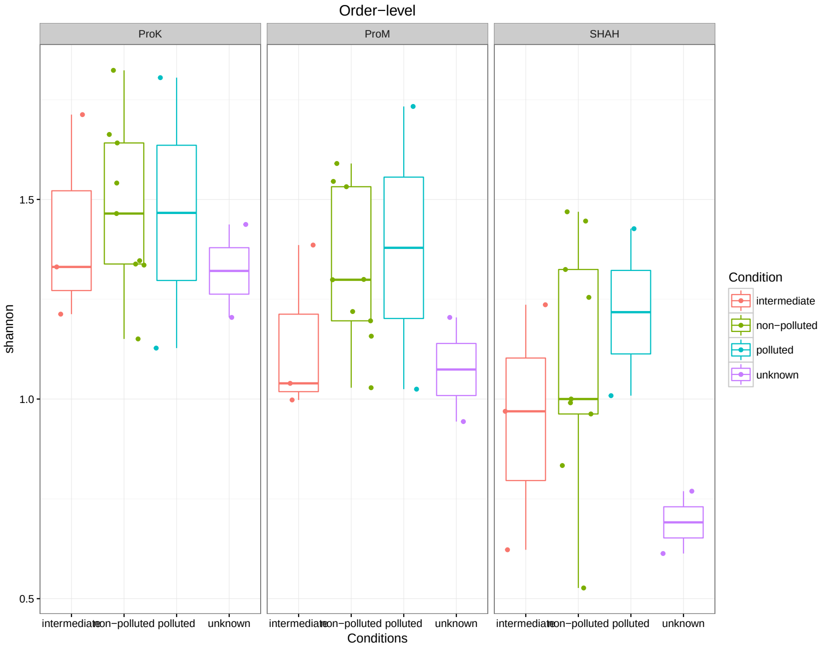


Figure A.6: Comparison of the SHAH, Prok, and ProM in calculating the Shannon index for various environmental conditions at order-level.

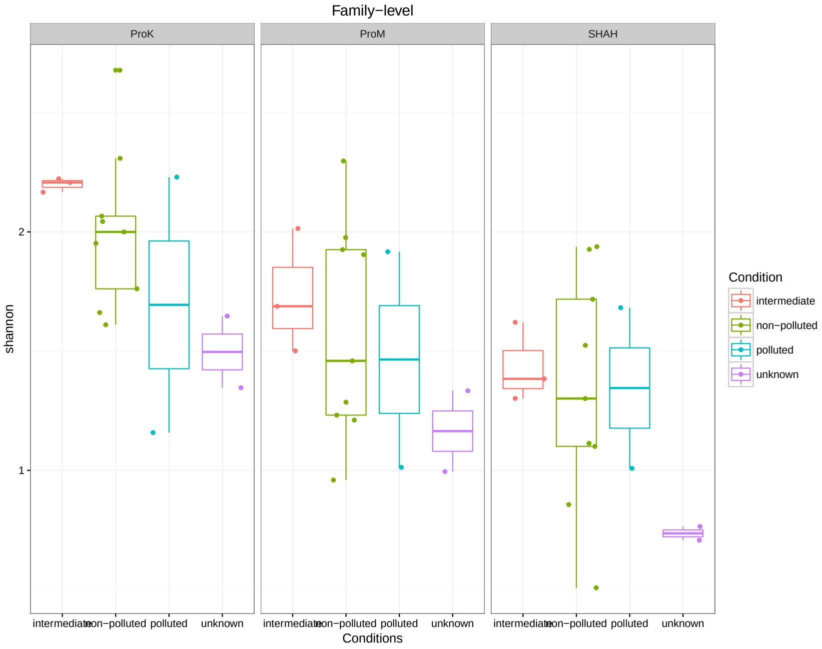


FIGURE A.7: Comparison of the SHAH, Prok, and ProM in calculating the Shannon index for various environmental conditions at family-level.

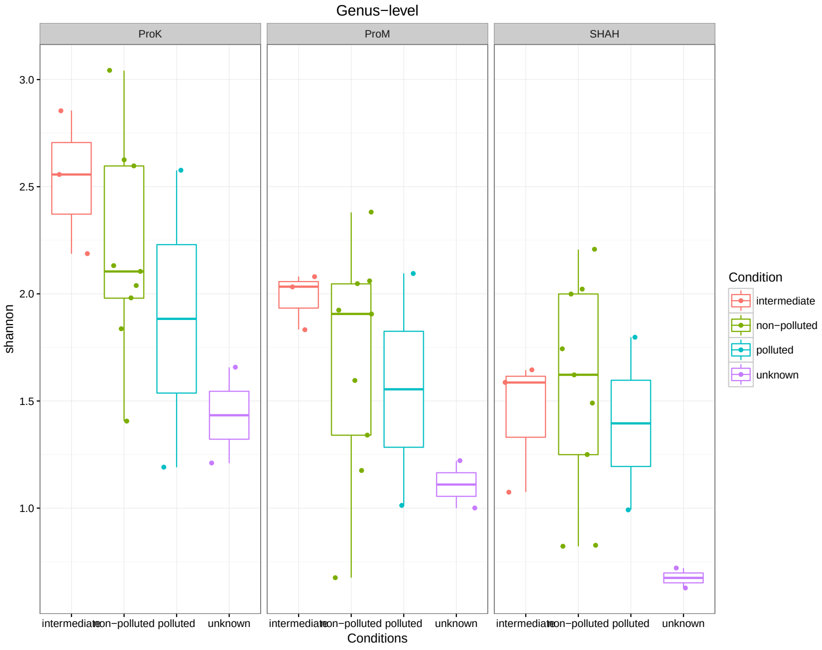


FIGURE A.8: Comparison of the SHAH, Prok, and ProM in calculating the Shannon index for various environmental conditions at genus-level.

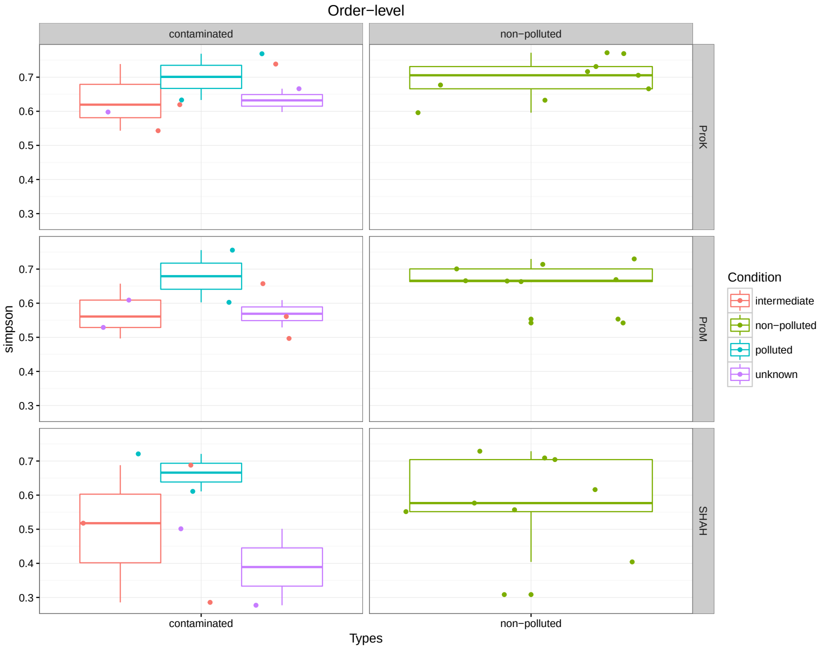


FIGURE A.9: Comparison of the SHAH, Prok, and ProM in calculating the Simpson index for various environmental types at order-level.

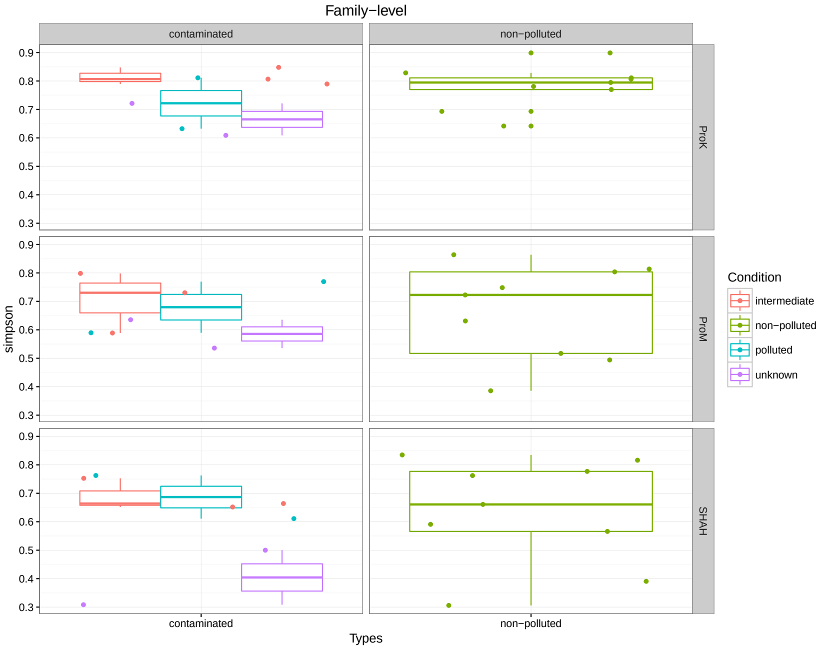


FIGURE A.10: Comparison of the SHAH, Prok, and ProM in calculating the Simpson index for various environmental types at family-level.

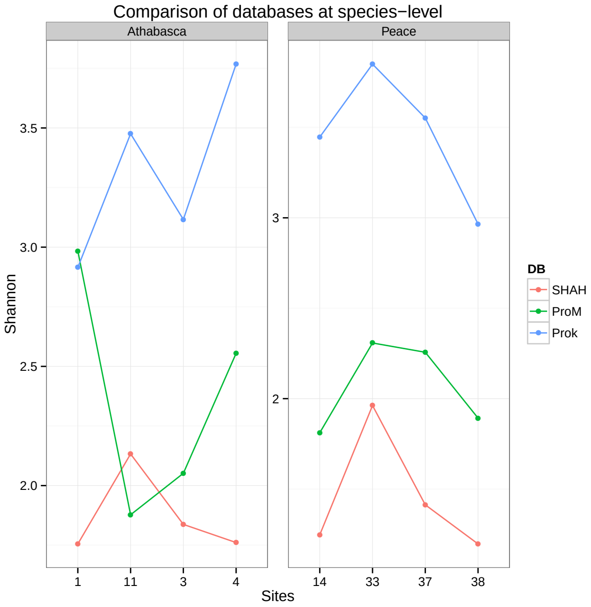


FIGURE A.11: Correlation between the classification methods using the Shannon index at species-level. Each point show the mean of the Shannon for each site and the programs are labeled with distinct colors.

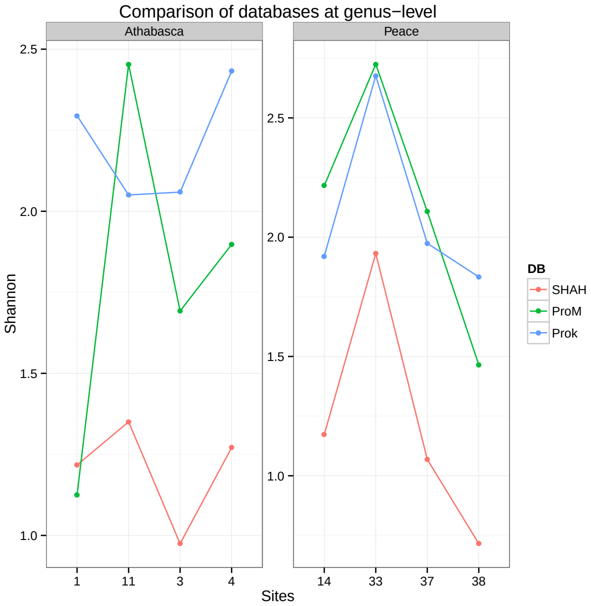


FIGURE A.12: Correlation between the classification methods using the Shannon index at genus-level. Each point show the mean of the Shannon for each site and the programs are labeled with distinct colors.

Bibliography

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997). “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. *Nucleic acids research* 25.17, pp. 3389–3402.

Amend, A. S., K. A. Seifert, and T. D. Bruns (2010). “Quantifying microbial communities with 454 pyrosequencing: does read abundance count?” *Molecular Ecology* 19.24, pp. 5555–5565.

Ames, S. K., D. A. Hysom, S. N. Gardner, G. S. Lloyd, M. B. Gokhale, and J. E. Allen (2013). “Scalable metagenomic taxonomy classification using a reference genome database”. *Bioinformatics* 29.18, pp. 2253–2260.

Andrews, S. et al. (2010). “FastQC: A quality control tool for high throughput sequence data”. *Reference Source.*

Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, et al. (2006). “The marine viromes of four oceanic regions”. *PLoS biol* 4.11, e368.

Baird, D. J. and M. Hajibabaei (2012). “Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing”. *Molecular Ecology* 21.8, pp. 2039–2044.

Bazinet, A. L. and M. P. Cummings (2012). “A comparative evaluation of sequence classification programs”. *BMC bioinformatics* 13.1, p. 92.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers (2013). “GenBank”. *Nucleic acids research* 41.D1, pp. D36–D42.

Benton, M. J. and S. I. Guttman (1990). “Relationship of allozyme genotype to survivorship of mayflies (Stenonema femoratum) exposed to copper”. *Journal of the North American Benthological Society*, pp. 271–276.

Berger, S. A., D. Krompass, and A. Stamatakis (2011). “Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood”. *Systematic biology*, syr010.

Bonada, N., N. Prat, V. H. Resh, and B. Statzner (2006). “Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches”. *Annu. Rev. Entomol.* 51, pp. 495–523.

Brady, A. and S. L. Salzberg (2009). “Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models”. *Nature methods* 6.9, pp. 673–676.

Brodersen, K. P. and N Anderson (2002). “Distribution of chironomids (Diptera) in low arctic West Greenland lakes: trophic conditions, temperature and environmental reconstruction”. *Freshwater Biology* 47.6, pp. 1137–1157.

Buee, M., M. Reich, C Murat, E Morin, R. H. Nilsson, S Uroz, and F. Martin (2009). "454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity". *New Phytologist* 184.2, pp. 449–456.

Caporaso, J. G., C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, et al. (2011). "Moving pictures of the human microbiome". *Genome Biol* 12.5, R50.

Carew, M. E., V. J. Pettigrove, L. Metzeling, and A. A. Hoffmann (2013). "Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species". *Front Zool* 10.1, p. 45.

Chapin III, F. S. and G. R. Shaver (1985). "Individualistic growth response of tundra plant species to environmental manipulations in the field". *Ecology* 66.2, pp. 564–576.

Chapin III, F. S., E. S. Zavaleta, V. T. Eviner, R. L. Naylor, P. M. Vitousek, H. L. Reynolds, D. U. Hooper, S. Lavorel, O. E. Sala, S. E. Hobbie, et al. (2000). "Consequences of changing biodiversity". *Nature* 405.6783, pp. 234–242.

CHRS (2011). *Canadian Heritage Rivers System.* URL: http://http://chrs.ca/the-rivers/humber/ (visited on 06/18/2016).

Clarke, G. M. (1993). "Fluctuating asymmetry of invertebrate populations as a biological indicator of environmental quality". *Environmental Pollution* 82.2, pp. 207–211.

Cortet, J., A. Gomot-De Vauflery, N. Poinsot-Balaguer, L. Gomot, C. Texier, and D. Cluzeau (1999). "The use of invertebrate soil fauna in monitoring pollutant effects". *European Journal of Soil Biology* 35.3, pp. 115–134.

Dejean, T., A. Valentini, A. Duparc, S. Pellier-Cuit, F. Pompanon, P. Taberlet, and C. Miaud (2011). "Persistence of environmental DNA in freshwater ecosystems". *PloS one* 6.8, e23398.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB". *Applied and environmental microbiology* 72.7, pp. 5069–5072.

Dowle, E. J., X. Pochon, J. C Banks, K. Shearer, and S. A. Wood (2015). "Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates". *Molecular Ecology Resources*.

Drummond, A. J., R. D. Newcomb, T. R. Buckley, D. Xie, A. Dopheide, B. C. Potter, J. Heled, H. A. Ross, L. Tooman, S. Grosser, et al. (2015). "Evaluating a multigene environmental DNA approach for biodiversity assessment". *GigaScience* 4.1, pp. 1–20.

Duan, Y., S. I. Guttman, J. T. Oris, and A. J. Bailer (2000). "Genotype and toxicity relationships among Hyalella azteca: I. Acute exposure to metals or low pH". *Environmental Toxicology and Chemistry* 19.5, pp. 1414–1421.

Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST". *Bioinformatics* 26.19, pp. 2460–2461.

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince, and R. Knight (2011). "UCHIME improves sensitivity and speed of chimera detection". *Bioinformatics* 27.16, pp. 2194–2200.

Federhen, S. (2012). “The NCBI taxonomy database”. *Nucleic acids research* 40.D1, pp. D136–D143.

Frati, F., P. P. Fanciulli, and L. Posthuma (1992). “Allozyme variation in reference and metal-exposed natural populations of Orchesella cincta (Insecta: Collembola)”. *Biochemical systematics and ecology* 20.4, pp. 297–310.

Geer, L. Y., A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, and S. H. Bryant (2009). “The NCBI biosystems database”. *Nucleic acids research*, gkp858.

Gibbons, S. M., E. Jones, A. Bearquiver, F. Blackwolf, W. Roundstone, N. Scott, J. Hooker, R. Madsen, M. L. Coleman, and J. A. Gilbert (2014). “Human and environmental impacts on river sediment microbial communities”. *PLoS One* 9.5, e97435.

Gibson, J., S. Shokralla, T. M. Porter, I. King, S. van Konynenburg, D. H. Janzen, W. Hallwachs, and M. Hajibabaei (2014). “Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics”. *Proceedings of the National Academy of Sciences* 111.22, pp. 8007–8012.

Gibson, J. F., S. Shokralla, C. Curry, D. J. Baird, W. A. Monk, I. King, and M. Hajibabaei (2015). “Large-Scale Biomonitoring of Remote and Threatened Ecosystems via High-Throughput Sequencing”. *PloS one* 10.10, e0138432.

Hajibabaei, M., G. A. Singer, P. D. Hebert, and D. A. Hickey (2007). “DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics”. *TRENDS in Genetics* 23.4, pp. 167–172.

Hajibabaei, M., S. Shokralla, X. Zhou, G. A. Singer, and D. J. Baird (2011). “Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos”. *PLoS one* 6.4, e17497.

Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman (1998). “Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products”. *Chemistry & biology* 5.10, R245–R249.

Hebert, P. D., A. Cywinska, S. L. Ball, et al. (2003). “Biological identifications through DNA barcodes”. *Proceedings of the Royal Society of London B: Biological Sciences* 270.1512, pp. 313–321.

Hebert, P. D. and T. R. Gregory (2005). “The promise of DNA barcoding for taxonomy”. *Systematic biology* 54.5, pp. 852–859.

Hebert, P. D., S. Ratnasingham, and J. R. de Waard (2003). “Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species”. *Proceedings of the Royal Society of London B: Biological Sciences* 270.Suppl 1, S96–S99.

Hewlett, R. (2000). “Implications of taxonomic resolution and sample habitat for stream classification at a broad geographic scale”. *Journal of the North American Benthological Society* 19.2, pp. 352–361.

Hill, M. O. (1973). “Diversity and evenness: a unifying notation and its consequences”. *Ecology* 54.2, pp. 427–432.

Hilty, J. and A. Merenlender (2000). “Faunal indicator taxa selection for monitoring ecosystem health”. *Biological conservation* 92.2, pp. 185–197.

Hodkinson, I. D. and J. Bird (1998). “Host-specific insect herbivores as sensors of climate change in arctic and Alpine environments”. *Arctic and Alpine Research*, pp. 78–83.

Hodkinson, I. D. and J. K. Jackson (2005). “Terrestrial and aquatic invertebrates as bioindicators for environmental monitoring, with particular reference to mountain ecosystems”. *Environmental Management* 35.5, pp. 649–666.

Hollingsworth, P. M., L. L. Forrest, J. L. Spouge, M. Hajibabaei, S. Ratnasingham, M. van der Bank, M. W. Chase, R. S. Cowan, D. L. Erickson, A. J. Fazekas, et al. (2009). “A DNA barcode for land plants”. *Proceedings of the National Academy of Sciences* 106.31, pp. 12794–12797.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster (2007). “MEGAN analysis of metagenomic data”. *Genome research* 17.3, pp. 377–386.

Jones, F. C. (2008). “Taxonomic sufficiency: the influence of taxonomic resolution on freshwater bioassessments using benthic macroinvertebrates”. *Environmental Reviews* 16.NA, pp. 45–69.

Jost, L. (2006). “Entropy and diversity”. *Oikos* 113.2, pp. 363–375.

Krieger, K. A., D. W. Schloesser, B. A. Manny, C. E. Trisler, S. E. Heady, J. J. Ciborowski, and K. M. Muth (1996). “Recovery of burrowing mayflies (Ephemeroptera: Ephemeridae: Hexagenia) in western Lake Erie”. *Journal of Great Lakes Research* 22.2, pp. 254–263.

Lauber, C. L., M. Hamady, R. Knight, and N. Fierer (2009). “Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale”. *Applied and environmental microbiology* 75.15, pp. 5111–5120.

Lavelle, P, D Bignell, M Lepage, W Wolters, P Roger, P. Ineson, O. Heal, and S Dhillion (1997). “Soil function in a changing world: the role of invertebrate ecosystem engineers”. *European Journal of soil biology* 33.4, pp. 159–193.

Lee, C. K., B. A. Barbier, E. M. Bottos, I. R. McDonald, and S. C. Cary (2012). “The inter-valley soil comparative survey: the ecology of Dry Valley edaphic microbial communities”. *The ISME journal* 6.5, pp. 1046–1057.

Leray, M., J. Y. Yang, C. P. Meyer, S. C. Mills, N. Agudelo, V. Ranwez, J. T. Boehm, and R. J. Machida (2013). “A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents”. *Frontiers in zoology* 10.1, p. 1.

Luo, C., D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis (2012). “Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample”. *PloS one* 7.2, e30087.

Mardis, E. R. (2008). “The impact of next-generation sequencing technology on genetics”. *Trends in genetics* 24.3, pp. 133–141.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, et al. (2005). “Genome sequencing in microfabricated high-density picolitre reactors”. *Nature* 437.7057, pp. 376–380.

McGEOCH, M. A. (1998). “The selection, testing and application of terrestrial insects as bioindicators”. *Biological Reviews of the Cambridge Philosophical Society* 73.02, pp. 181–201.

McNaughton, S. J. (1977). “Diversity and stability of ecological communities: a comment on the role of empiricism in ecology”. *American Naturalist*, pp. 515–525.

Merritt, R. W. and K. W. Cummins (1996). *An introduction to the aquatic insects of North America*. Kendall Hunt.

Morgulis, A., G. Coulouris, Y. Raytselis, T. L. Madden, R. Agarwala, and A. A. Schäffer (2008). “Database indexing for production MegaBLAST searches”. *Bioinformatics* 24.16, pp. 1757–1764.

Naeem, S. and S. Li (1997). “Biodiversity enhances ecosystem reliability”. *Nature* 390.6659, pp. 507–509.

NCBI (2016). *GenBank and WGS Statistics*. URL: http://www.ncbi.nlm.nih.gov/genbank/statistics/ (visited on 05/25/2016).

Nordén, B. and T. Appelqvist (2001). “Conceptual problems of ecological continuity and its bioindicators”. *Biodiversity & Conservation* 10.5, pp. 779–791.

Oksanen, J., R. Kindt, P. Legendre, B. O’Hara, M. H. H. Stevens, M. J. Oksanen, and M. Suggests (2007). “The vegan package”. *Community ecology package* 10.

Paoletti, M. G. (2012). *Invertebrate biodiversity as bioindicators of sustainable landscapes: Practical use of invertebrates to assess sustainable land use*. Elsevier.

Paradis, E., J. Claude, and K. Strimmer (2004). “APE: analyses of phylogenetics and evolution in R language”. *Bioinformatics* 20.2, pp. 289–290.

Parsons, P. (1992). “Fluctuating asymmetry: a biological monitor of environmental and genomic stress”. *Heredity* 68.4, pp. 361–364.

Pettigrove, V. and A. Hoffmann (2005). “A field-based microcosm method to assess the effects of polluted urban stream sediments on aquatic macroinvertebrates”. *Environmental Toxicology and Chemistry* 24.1, pp. 170–180.

Polz, M. F. and C. M. Cavanaugh (1998). “Bias in template-to-product ratios in multitemplate PCR”. *Applied and environmental Microbiology* 64.10, pp. 3724–3730.

Porter, T. M., J. F. Gibson, S. Shokralla, D. J. Baird, G. B. Golding, and M. Hajibabaei (2014). “Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier”. *Molecular ecology resources* 14.5, pp. 929–942.

Power, M. E., D. Tilman, J. A. Estes, B. A. Menge, W. J. Bond, L. S. Mills, G. Daily, J. C. Castilla, J. Lubchenco, and R. T. Paine (1996). “Challenges in the quest for keystones”. *BioScience* 46.8, pp. 609–620.

Price, M. N., P. S. Dehal, and A. P. Arkin (2009). “FastTree: computing large minimum evolution trees with profiles instead of a distance matrix”. *Molecular biology and evolution* 26.7, pp. 1641–1650.

Pruitt, K. D., T. Tatusova, and D. R. Maglott (2007). “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins”. *Nucleic acids research* 35.suppl 1, pp. D61–D65.

Ratnasingham, S. and P. D. Hebert (2007). “BOLD: The Barcode of Life Data System (http://www. barcodinglife. org)”. *Molecular ecology notes* 7.3, pp. 355–364.

Reuter, J. A., D. V. Spacek, and M. P. Snyder (2015). “High-throughput sequencing technologies”. *Molecular cell* 58.4, pp. 586–597.

Rosen, G., E. Garbarine, D. Caseiro, R. Polikar, and B. Sokhansanj (2008). “Metagenome Fragment Classification Using -Mer Frequency Profiles”. *Advances in bioinformatics* 2008.

Rosenberg, D. M., H. Danks, and D. M. Lehmkuhl (1986). "Importance of insects in environmental impact assessment". *Environmental Management* 10.6, pp. 773–783.

Rosenberg, D. M., V. H. Resh, et al. (1993). *Freshwater biomonitoring and benthic macroinvertebrates.* Chapman & Hall.

Ruiter, P. C. de, A.-M. Neutel, and J. C. Moore (1995). "Energetics, patterns of interaction strengths, and stability in real ecosystems". *Science* 269.5228, p. 1257.

Salonius, P. (1981). "Metabolic capabilities of forest soil microbial populations with reduced species diversity". *Soil Biology and Biochemistry* 13.1, pp. 1–10.

Schmieder, R. and R. Edwards (2011). "Quality control and preprocessing of metagenomic datasets". *Bioinformatics* 27.6, pp. 863–864.

Schoch, C. L., K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, E. Bolchacova, K. Voigt, P. W. Crous, et al. (2012). "Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi". *Proceedings of the National Academy of Sciences* 109.16, pp. 6241–6246.

Shannon, C. E. (1949). "Communication theory of secrecy systems". *Bell system technical journal* 28.4, pp. 656–715.

Shokralla, S., J. L. Spall, J. F. Gibson, and M. Hajibabaei (2012). "Next-generation sequencing technologies for environmental DNA research". *Molecular ecology* 21.8, pp. 1794–1805.

Shrestha, R. K., B. Lubinsky, V. B. Bansode, M. B. Moinz, G. P. McCormack, and S. A. Travers (2014). "QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform". *BMC bioinformatics* 15.1, p. 1.

Simon, C. and R. Daniel (2011). "Metagenomic analyses: past and future trends". *Applied and environmental microbiology* 77.4, pp. 1153–1161.

Simpson, E. (1949). "Measurement of Diversity". *Nature* 163, pp. 688–688.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting (2014). "Sequencing depth and coverage: key considerations in genomic analyses". *Nature Reviews Genetics* 15.2, pp. 121–132.

Smith, D. P. and K. G. Peay (2014). "Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing". *PLoS One* 9.2, e90234.

Spall, J. (2014). "Investigating the Utility of Next Generation Sequencing for Evaluating Biodiversity in Benthic Communities". PhD thesis. University of Guelph.

Sverdrup-Thygeson, A (2001). "Can'continuity indicator species' predict species richness or red-listed species of saproxylic beetles?" *Biodiversity & Conservation* 10.5, pp. 815–832.

Sweeney, B. W., J. M. Battle, J. K. Jackson, and T. Dapkey (2011). "Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality?" *Journal of the North American Benthological Society* 30.1, pp. 195–216.

Taberlet, P., E. Coissac, M. Hajibabaei, and L. H. Rieseberg (2012). "Environmental DNA". *Molecular Ecology* 21.8, pp. 1789–1793.

Tange, O. et al. (2011). "Gnu parallel-the command-line power tool". *The USENIX Magazine* 36.1, pp. 42–47.

Team, R. C. et al. (2013). "R: A language and environment for statistical computing".

Tessier, L, J. Boisvert, L.-M. Vought, and J. O. Lacoursière (2000). "Anomalies on capture nets of Hydropsyche slossonae larvae (Trichoptera; Hydropsychidae) following a sublethal chronic exposure to cadmium". *Environmental Pollution* 108.3, pp. 425–438.

Tilman, D., J. Knops, D. Wedin, P. Reich, M. Ritchie, and E. Siemann (1997). "The influence of functional diversity and composition on ecosystem processes". *Science* 277.5330, pp. 1300–1302.

TRCA (2011). *Toronto and Region Conservation Authority.* URL: `https://trca.ca/conservation/environmental-monitoring/` (visited on 09/30/2010).

Van Straalen, N. M. (1998). "Evaluation of bioindicator systems derived from soil arthropod communities". *Applied Soil Ecology* 9.1, pp. 429–437.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy". *Applied and environmental microbiology* 73.16, pp. 5261–5267.

Wardle, D., K. Bonner, and K. Nicholson (1997). "Biodiversity and plant litter: experimental evidence which does not support the view that enhanced species richness improves ecosystem function". *Oikos*, pp. 247–258.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* Springer Science & Business Media.

Wood, D. E. and S. L. Salzberg (2014). "Kraken: ultrafast metagenomic sequence classification using exact alignments". *Genome Biol* 15.3, R46.