# Investigate the diversity of extremely complex metagenomic sample

Qingpeng Zhang

November 12, 2013

## Contents

# 1 Background and Significance

## 1.1 Introduction and Overview

Species diversity is an important measurement of ecological communities. Scientists believe that there is relationship between species diversity and ecosystem processes (Loreau et al., 2001). Evaluating the species diversity in a community is a central research topic in classic ecology. Many methods have been developed since decades ago, which make the question like "how many species of birds in this habitat" easier to answer. Nevertheless, scientists have not started to think seriously about other old questions like "How many species are there on earth?" (May, 1988) or "How many species are there in the ocean?" (Mora, Tittensor, Adl, Simpson, & Worm, 2011) until not long time ago. Why? The answer is simple: Microorganisms represet the vast majority of the Earth's biodiversity and the assessment of microbial diversity is simply hard. Actually it is believed that microbial diversity is the outermost frontier of the exploration of diversity (Magurran & McGill, 2011). Microorganisms are ubiquitous. They were the first forms of life on the Earth. There are more bacterial cells inhabiting our body than our own cells (Savage, 1977). They are essential to all life, too. There are several reasons why assessment of microbial diversity is such a challenge. The concept of species is ambiguous. Morphological examination is impossible. 95% or more of the microbial diversity in the biosphere can not be cultivated with standard culturing techniques (Curtis, Sloan, & Scannell, 2002). To tackle these obstacles, metagenomics emerges, with the boost given by the progress of next generation sequencing technology. Lots of metagenomics projects have been performed on samples from acid mine drainage channels to human gut. For complex samples like soil, the resulting data set will be huge. There are approximately a billion microbial cells, with about 4 petabase pairs of DNA($4 * 10^{12}$ bp). Since we have limited sequencing power and other financial strain, the resulting metagenomics data set from high diversity sample like soil only corresponds to a tiny fraction of the actual genomic content in the sample. The large size of data set and the low coverage make the assessment of microbial diversity of high diversity sample even harder. Novel method is highly needed.

### 1.1.1 Next-generation sequencing

Sequencing technology is changing really fast. Over the past decade, next-generation sequencing(NGS) has been the overwhelming technology and almost replaced classic Sanger sequencing technology. Illumina and Roche 454 are the two most popular platforms. Illumina can generate reads with shorter length, typically 100 bases for HiSeq and 150 bases for MiSeq (Qin et al., 2010; Mason et al., 2012), but with lower cost compared to Roche 454 sequencing technology, which generates reads 500 bases to 1K bases longer. In fact a recent study comparing Illumina versus Roche 454 for metagenomics shows that both platforms agreed on over 90% of the assembled contigs and 89% of the unassembled reads(Luo, Tsementzi, Kyrpides, Read, & Konstantinidis, 2012). Because of the advantage of low cost, there is an obvious trend that Illumina is dominating the sequencing market, which means while designing any tool for metagenomics, the developer should take the relatively short length of Illumina reads into account.

### 1.1.2 Metagenomics

It is believed that the word "metagenomics" was coined in 1998 (Handelsman, Rondon, Brady, Clardy, & Goodman, 1998), which can be translated as 'beyond the genome' (Gilbert & Dupont, 2011). At that time, basically it is the technique of cloning environmental DNA randomly and screening for interesting genes, especially 16S rRNA genes. This technique was firstly applied in practice by Schmidt et al. in 1991 (Schmidt, DeLong, & Pace, 1991). This was a crucial step in the investigation to microbial world. Before that it was standard protocol to culture and isolate microbes from cells or living organisms of any other species and then do analysis. This resulted seriously narrow picture of the diversity of an ecosystem as only a small portion of the microbial species (5% or less) in the biosphere can be cultured with standard culturing techniques (Sogin et al., 2006). Metagenomics with the concept of cloning DNA directly from the environment without cultivation brings the researchers the ability to explore the genomic DNA from all the genomes of all the organisms in an environmental community, culturable or unculturable.

The improvement of next generation sequencing technology with high throughput and low cost has been accelerating the metagenomics research recently. The number of microbial species in some ecological community is huge. In soil it is estimated that there exist millions of species with most of them in low abundance (Gans, Wolinsky, & Dunbar, 2005). Only using high throughput next-generation sequencing strategy can it be possible to sample the contents of those populations deeply enough to cover the rare species.

Currently there are two approaches in metagenomics. One is amplicon metagenomics, to amplify gene of interest like 16S rRNA genes as taxonomic markers and sequence the libraries (Sogin et al., 2006), which is the traditional way dating back to 1991 experiment by Schmidt et al. Many classic methods to access microbial diversity rely on this approach. The other one is whole genome shotgun metagenomics, to sequence the libraries of randomly isolated DNA fragments without screening. Since the whole genomes of organisms in the sample are available rather than the limited genes of interest like 16S/18S rRNA, this whole genome shotgun sequencing approach can provide better taxonomic resolution and more information benefiting other investigation (Tyson et al., 2004) (Qin et al., 2010). Now there have been thousands of metagenomic genomes available in online database like MG-RAST (Glass, Wilkening, Wilke, Antonopoulos, & Meyer, 2010).

There have been many metagenomics projects focusing on the microbial samples of different kinds of habitat, from extreme environment such as acid mine drainage channels with low complexity (Tyson et al., 2004), and medium complexity samples like human gut (Qin et al., 2010) and cow rumen (Hess et al., 2011), to high complexity samples like seawater (Venter et al., 2004) and soil (Gilbert et al., 2010).

Metagenomics studies have expanded our knowledge of microbial world in different habitats. Some of them shed light on the explanation of some serious human deseases. Studies have shown the associations between human gut metagenomes and type II diabetes (Qin et al., 2012), obesity (Turnbaugh et al., 2009; Kau, Ahern, Griffin, Goodman, & Gordon, 2011) or crohn's disease (Morgan et al., 2012).

In almost all of these metagenomics projects, diversity analysis plays an important role to supply information about the richness of species, the species abundance distribution in a sample or the similarity and difference between different samples, all of which are crucial to draw insightful and reliable conclusion. Next the challenges in the assessment of microbial diversity will be elaborated.

## 1.2 Challenges in measuring diversity of metagenomics

### 1.2.1 Concept of Diversity

When we try to characterize an ecological community, diversity measurement is often the first step. It is always desirable to know how many species there are in a sample, which is the concept of richness and how abundant each species is relative to others in the same sample, which is the concept of evenness. They are straightforward conceptually. But in practice, there are a large number of quantities that was suggested to measure species diversity to adapt the different scenarios of sampling individuals.

In a higher level, three diversity indices are well established and used in ecology, $\alpha$-diversity, $\beta$-diversity, and $\gamma$-diversity. $\alpha$-diversity is the diversity in one defined habitat or sample. $\beta$-diversity compares species diversity between habitats or samples.$\gamma$-diversity is the total diversity over a large region containing multiple ecosystems.

The concept of diversity has two aspects, richness and evenness. Richness is the total number of species identified in a sample, which is the simplest descriptors of community structure. Evenness is a measure of how different the abundance of a species is compared to other species in a community. If all the species in a community has the same abundance, the community has a higher evenness diversity. However all natural communities are highly uneven, which means a large number of species has rare abundance in the community. This raises a question about the effectiveness of using the measurement of richness to represent diversity. Is a community with 1 dominant species and 10 rare species more diverse than a community with 3 dominant species and 2 rare species? So a new metrics taking both richness and evenness into account is suggested. Two popular indice are Shannon diversity (Shannon, 2001), which is based on information theory and shows the information in a community as an estimate of diversity, and Simpson diversity index (Simpson, 1949), which basically shows the probability that two individuals picked randomly from a community belong to the same species.

Besides these two, Hill (Hill, 1973) proposed a new diversity index to use a weighted counts of species to measure diversity, based on the species abundance distribution. This can be considered as a generalized diversity index, since both Shannon and Simpson index and richness can be seen as special cases of Hill diversity index.

It is necessary to note that we can not tell if any index is better than other. It all depends on the characteristics of the community and the process of sampling and other factors. More often an index is used just because it has been used by many others before in the same scenario, it may be because it is more feasible in this scenario but it is not necessarily because it is better or even provide useful information.

In microbial ecology, richness is simply the most popular index to mesure microbial diversity, partially because of the challenge raised by the different characteristics of microbial community. Lots of methods to estimate richness in classic ecology were borrowed to tackle the problem of estimate microbial diversity, which will be discussed in the next section.

### 1.2.2 Diversity measurement in Microbial Ecology

There have been numerous mature methods and tools to measure diversity of macroorganisms in decades of development of classic ecology. One would think that we just need to borrow those methods to use in microbial field. Nevertheless in reality it is not that straightforward. The microbial communities are so different from macroorganisms like plant or animal communities, with the number of species many order of magnitude larger(Whitman, Coleman, & Wiebe, 1998). This fact raises serious sampling problems. It is really difficult to cover enough fraction of the microbial community even with impressively large sample size thanks to modern metagenomic approaches (Roesch et al., 2007). In a word, diversity measurement is a rather big challenge for microbial communities and novel and effective methods are highly demanded (Schloss & Handelsman, 2005).

**Concept of Species and OTU Identification using sequence markers** To borrow the methods of diversity measurement from classic ecology on the use of evaluating microbial diversity, the first problem is that in microbial world, there is no unambiguous way to define "species" (Stackebrandt et al., 2002). It is impossible to identify a microbial individual as a specific species morphologically. In fact in metagenomics the concept of "species" has been replaced by OTUs(Operational Taxonomic Units). An OTUs are those microbial individuals within a certain evolutional distance. Practically we mainly use 16S rRNA genes as the evolutional marker genes, because 16S rRNA genes exist universally among different microbial species and their sequences change at a rate corresponding with the evolutionary distance. So we can describe microbial individuals with higher than a certain percent(like 97%) 16S rRNA sequence similarity as one OTU, or belonging to one species (Schloss & Handelsman, 2005).

**Binning of Metagenomic Reads into OTUs** In classic ecology dealing with samples from macroorganisms communities, before we can use any statistical method to measure diversity, it is standard procedure to identify the species of each individual in the sample. It is the same for diversity measurement of microbial communities. Difference is that here we need to place the sequences(individuals) into respective "bin" or OTUs(species). There are two strategies to do such binning - Composition-based or intrinsic binning approach and similarity-based or extrinsic binning approach.

**Composition-based approach** Lots of efforts have been put to get a comprehensive category of reference microbial genome sequences (Human Microbiome Jumpstart Reference Strains Consortium et al., 2010; D. Wu et al., 2009). Currently there are a large number of finished or high-quality reference sequences of thousands of microbial species available in different databases and this number is still increasing quickly (Markowitz et al., 2012; Glass et al., 2010; Wang, Garrity, Tiedje, & Cole, 2007). So the first intrinsic composition-based approach is to use those reference genomes to train a taxonomic classifier and use that classifier to classify the metagenomics reads into bins. Different statistical approaches like Support Vector Machines (Patil, Roune, & McHardy, 2012), interpolated Markov models(Brady & Salzberg, 2011),naive Bayesian classifiers, and Growing Self Organizing Maps (Rosen, Reichenberger, & Rosenfeld, 2011) were used to train the classifier. Without using any reference sequences for the training, it is possible to use signatures like k-mers or codon-usage to develop reference-independent approach. The assumption is that the frequencies distribution of the signatures are similar of the sequences from the same species. TETRA is such a reference-independent tools using Markov models based on k-mer frequencies (Teeling, Waldmann, Lombardot, Bauer, & Glöckner, 2004). There is another tool using both TETRA and codon usage statistics to classify reads (Tzahor et al., 2009).

**Similarity-based approach** The similarity-based extrinsic approach is to find similarity between the reads sequences and reference sequences and a tree can be built using the similarity distance information. MEGAN (Huson, Auch, Qi, & Schuster, 2007) is a typical tool using this method, which reads a BLAST file output. Other sequence alignment tools can also be used here like BowTie2 or BWA. Recently, an alternative strategy was developed, which only uses the reference sequences with the most information rather than all the reference sequences to do alignment. Those reference sequences include 16S rRNA genes or some other specific marker genes. The benefit is obvious, it is more time-efficient since there are fewer reference sequences to align to. Also, it can provide better resolution and binning accuracy since the marker genes can be selected carefully with the best distinguishing power. AMPHORA2 (M. Wu & Scott, 2012) and MetaPhlAn (Segata et al., 2012) are two typical tools using this strategy.

**Statistics for Diversity Estimation** After the binning of sequences into OTU, we need statistics to help us estimate the diversity. Many statistical methods have been developed and widely used in classic ecology to macroorganisms. However the first difference between diversity measurement of macroorganisms and microbial community is that generally the microbial community diversity is much larger than observed sample diversity, thanks to the high diverse characteristics of microbial community and the limit of metagenomics sampling and sequencing. The first approach which is also considered as classic is rarefaction. Rarefaction curve can be used to compare observed richness among different samples that have been sampled unequally, which is basically the plot of the number of observed species as a function of the sampled individuals. It is worth noting that rarefaction curve shows the observed diversity, not the total diversity. We should never forget those unseen microbial species, which is pretty common for microbial community sampling.

To estimate the total diversity from observed diversity, different estimators are required.

The first one is extrapolation from accumulation curve. The asymptote of this curve is the total diversity, which means the number of species will not increase any more with sampling more individuals. To get the value of that asymptote point, from observed accumulation curve, a function needs to be assumed to fit the curve. Several proposals have been made to use this extrapolation method (R. K. Colwell, Mao, & Chang, 2004; Gotelli & Colwell, 2001). The problem is that if the sampling effort only covers a small fraction of the total sample, which means the accumulation curve just starts, it is difficult to find an optimal function to fit the curve. Different functions can fit the curve equally well but will deduct dramatically different asymptote value. So this curve extrapolation method should be used cautiously.

Another one is parametric estimator, which assumes that the relative abundance follows a particular distribution. Then the number of unobserved species in the community can be estimated by fitting observed sample data to such abundance distribution then the total number of species in the community can be estimated. Lognormal abundance distribution is mostly used in different project since most communities of macroorganisms has a lognormal abundance distribution and it is believed that it is also typical for some microbial communities (Curtis et al., 2002; Schloss & Handelsman, 2006; Quince, Curtis, & Sloan, 2008). It is understandable that there is always controversy as to which models fit the communities best since in an ideal world the abundance distribution should be inferred from the data,not be assumed unverifiably. The problem is that we can only infer the abundance distribution accurately when the sample size is large enough. There has been some attempts on this direction recently (Gans et al., 2005) and more robust methods are still needed.

If the species abundance distribution can not be inferred, we can still use nonparametric estimators to estimate the total diversity without assuming that abundance distribution arbitrarily. These estimators are related to MRR(mark-release-recapture) statistics, which compare the number of species observed more than once and the number of species observed only once. If current sampling only covers a small fraction of a diverse community, the probability that a species is observed more than once will be low and most species will be observed only once. If current sampling is enough to cover most species in the community, the opposite will be the case. A series of estimators invented by Chao are the representative estimators in this category, including Chao1 (Chao, 1984), Chao2 (Chao, 1987), ACE (Chao & Yang, 1993) and ICE (Lee & Chao, 1994). For example, Chao1 formular is:

$$S_{Chao1} = S_{obs} + \frac{{n_1}^2}{2n_2}$$

where $S_{obs}$ is the number of species observed, $n_1$ the number of species observed once(singletons, with only one individule), and $n_2$ the number of species observed twice(doubletons, with exactly two individuals) in the sample. The ACE uses data from all species rather than just singletons and doubletons. Its formular is:

$$S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{F_1}{C_{ACE}}{\gamma_{ACE}}^2$$

where $S_{rare}$ is the number of rare species (with few than 10 individuals observed) and $S_{abund}$ is the number of abundant species (with more than 10 individuals).

In past years there are several software packages that have been developed for biodiversity analysis. Out of them, EstimateS (R. Colwell, n.d.) is a software that can be used for general purpose diversity analysis, which implement a rich set of diversity analysis algorithms.

However it is not designed specifically for microbial diversity analysis. So microbial diversity data should be preprocessed to general population data to be fed into EstimateS. Two other softwares - MOTHUR (Schloss et al., 2009) and QIIME (Caporaso et al., 2010) are designed for microbial diversity. So they are more popular in microbial diversity analysis. CatchAll (Bunge, 2011) is a relatively newer package, which can estimate the diversity using both nonparametric and parametric estimators including many variants and return the results using different estimators and the respective credibility of the results.

### 1.2.3 Novel methods required for diversity measurement of large metagenomics sample

As reviewed in previous sections, the topic of microbial diversity measurement has been investigated for a long time with many methods and software packages developed. However there is still lots of room for more work to do.

Basically the mainstream methods to measure microbial diversity are still focusing on the use of 16S rRNA amplicon metagenomics data. Many of the software packages mentioned above are also supposed to accept 16S rRNA data as input. This is understandable that after all the concept of OTU is from the similarity of 16S rRNA sequences. Using 16S rRNA data to measure diversity is popular but is not without problems. 16S rRNAs may not be that reliable to be OTU markers. The reliability is sensitive to potential horizontal gene transfer and the variance of gene copy in bacterium. There is suggestion that alternative marker genes should be used, like single copy housekeeping genes.

There are many projects generating whole genome shotgun metagenomics data sets. However they are mainly used for assembly and annotation purpose. Less attention was paid to diversity measurement using these whole genome metagenomics data sets. One possible reason is that the whole genome metagenomics data sets are often with low depth given the high diversity of metagenomics samples compared to 16S rRNA ampicon metagenomics data set. Assembly and annotation are always challenging with the low depth and lack of reference sequences. It is also true for diversity measurement. On the other hand, although with low depth, some whole genome metagenomics data sets are with large size because of the high diversity. For instance, there may be 4 petabase pairs of DNA in a gram of soil(Zarraonaindia, Smith, & Gilbert, 2013). Many of those methods for sequence binning or diversity estimation do not scale well and will not work for large metagenomics data sets. For instance, many composition-based binning approach involves k-mer/signature frequency distribution calculation, which is rather computationally expensive. Even basic sequence alignment will be impossible for large metagenomics data set. Many of those statistical software packages to estimate diversity using various estimators are not prepared for the large scale of whole genome metagenomics data.

With the development of next generation sequencing technology, the cost of sequencing is dropping rapidly. Whole genome metagenomics sequencing is more popular and large amount of metagenomics data is being generated with increasing speed, which can not be even met by the increase of computational capacity. Novel methods that can scale well are extremely needed to deal with the increasingly large metagenomics data set.

# 2 Research Proposal

## 2.1 Introduction

The goal of this project is to develop a method to measure the diversity of extremely complex sample by using extremely large metagenomics data set and it will be reference-free, assembly-free, annotation-free and binning-free. We will provide data structures and algorithms to support this method, use this method on synthetic data set and small scale real metagenomics data set to prove its effectiveness and efficiency and finally utilize this method on an extremely large soil metagenomics data set to acquire better scientific understanding of soil microbial communities. Relative topics like comparison of metagenomic samples will also be discussed.

As reviewed in previous section in this proposal, measuring diversity of microbial community heavily relies on 16S rRNA sequencing data set. In the context of microbial ecology, the concept of ”species” is replaced by OTUs, which are defined by the similarity of 16S rRNA sequences. However in the context of whole genome shotgun metagenomics data set, the concept of OTUs loses its practicability. We have some methods to classify whole genome shotgun metagenomics reads into OTU bins. The binning can also benefit other task like functional annotation. The binning does provide some insights to the diversity of community. But it is limited heavily by its intense computational requirement and the inaccuracy due to the lack of reference sequences many binning methods rely on and background knowledge. Now the consequence is that facing a extremely large metagenomics data set from complex microbial community, either the binning of the large amount of reads will last forever because of the computationally expensive algorithm or only a small fraction of the reads can be classified accurately because of the lack of information of large number of rare species in the community. Besides binning, annotation and assembly are both facing great challenge while facing extremely large metagenomics dataset from extremely complex microbial community. So it is of great importance to have a method that can measure the microbial diversity without the need of binning, assembly, annotation and reference sequences.

We have been working on this method for several years, firstly based on the concept of k-mers, later on the concept of ”genomic segment”. Large volumes of sequence data started to flood within the last 5 years. In answer to that trend, we also developed an efficient k-mer counting approach based on sketch data structure when existing software can not meet the requirement of the increasing size of sequence data set any more.

Another relative problem is to compare different metagenomic samples including the comparison on diversity and on genomic contents. The methods developed for diversity estimation can easily be migrated to solve the comparison problem and scale well to extremely large data sets.

## 2.2 Specific Aim 1: Using a novel concept of IGS(informative genomic segment) for microbial diversity measurement

In classic ecology dealing with macroorganisms, diversity measurement is based on the concept of ”species”. For 16S rRNA amplicon metagenomics data set, it is based on the concept of "OTUs". When the concept of OTUs does not work for large shotgun metagenomics data set, in the beginning we proposed that the concept of k-mers(a DNA segment with the leng of k) can be used to measure diversity. K-mers can be considered as the atom of information in DNA sequences. One of the composition-based approaches to binning is to use the k-mer as the signature. Suppose the sizes of microbial genomes are similar and the difference between genomic content of microbial genomes is similar, the number of distinct k-mers in the sequence data set is related to the number of species in a sample. However, because of sequencing error, which is unavoidable due to the limit of sequencing technology, this k-mer based analysis doe not work well. One sequencing error on a read will generate at most k erroneous k-mers. In metagenomics data set with low coverage, most of the distinct observed k-mers are from sequencing errors.

Next we turned our gaze to the upper level - reads. A novel method termed as "digital normalization" was developed to remove abundant reads before assembly. However it also supplies a novel way to distil information from reads by decreasing the bad influence of sequencing errors so that we can use those informative reads to measure the microbial diversity. We term those informative reads as IGS(informative genomic segment), which can be considered as a segment of DNA on a microbial genome. Those IGSs should be different enough to represent the abstract information a genome contains. Suppose microbial genomes contain similar number of those IGSs, as they contain similar number of distinct k-mers, the number of IGSs will be related to the species richness in a sample, and the abundance distribution of IGSs will be related to species evenness in a sample. Many classic diversity estimation methods based on OTUs level described in sections above can be borrowed to estimate the diversity of IGSs and the diversity of actual species subsequently.

### 2.2.1 Preliminary results: an approach to count k-mer efficiently

K-mer counting plays a key role in our initial investigation on using distinct k-mers to measure microbial diversity. We needed to count how many distinct k-mers in different metagenomics data sets and get the abundance distribution. In the beginning we used an existing k-mer counting tool - Tallymer (Kurtz, Narechania, Stein, & Ware, 2008). However as we started to deal with larger metagenomic data, where we routinely encounter data sets that contain $300 \times 10^9$ bases of DNA and over 50 billion distinct k-mers (Howe et al., 2012), it was not efficient enough and for some data set it can not handle at all. So to tackle the k-mer counting problem to measure microbial diversity was the original motivation for exploring more efficient k-mer counting approach, which leads the development of khmer package (Zhang, Pell, Canino-Koning, Howe, & Brown, 2013), although it has been used for other purpose afterwards, like metagenome assembly and error correction.

This simple probabilistic data structure for k-mer counting is based on a CountMin Sketch, a generalized probabilistic data structure for storing the frequency distributions of distinct elements (Cormode & Muthukrishnan, 2005). Our implementation extends an earlier implementation of a Bloom filter, which has been previously used for both k-mer counting and de Bruijn graph storage and traversal (Bloom, 1970; Broder & Mitzenmacher, 2003; Melsted & Pritchard, 2011; Pell et al., 2012; Rizk, Lavenier, & Chikhi, 2013; Jones, Ruzzo, Peng, & Katze, 2012)

The two basic operations supported by khmer are `c = increment_count(kmer)` and `c = get_count(kmer).` Both operate on the data structure in memory, such that neither incrementing a count nor retrieving a count involves disk access.

The implementation details are similar to those of the Bloom filter in (Pell et al., 2012), but with the use of 8 bit counters instead of 1 bit counters. Briefly, Z hash tables are allocated, each with a different size of approximately H bytes; the sum of these hash table sizes must fit within available main memory. To increment the count for a particular k-mer, a single hash is computed for the k-mer, and the modulus of that hash with each hash table's size H gives the location for each hash table; the associated count in each hash table is then incremented by 1. To retrieve the count for a k-mer, the same hash is computed and the minimum count across all hash tables is computed. While different in implementation detail from the standard CountMin Sketch, which uses a single hash table with many hash functions, the performance details are identical (Pell et al., 2012).

An additional benefit of the CountMin Sketch is that it is extremely easy to implement correctly, needing only about 3 dozen lines of C++ code for a simple threadsafe implementation.

To determine the expected error rate — the average frequency with which a given k-mer count will be incorrect when retrieved — we can look at the hash table load. Suppose N unique k-mers have been counted using Z hash tables, each with size H. The probability that no collisions happened in a specific entry in one hash table is $(1-1/H)^N$, which can be approximated as $e^{-N/H}$. The individual collision rate in one hash table is $1-e^{-N/H}$. The total collision rate, which is the probability that a collision occurred in each entry where a k-mer maps to in all Z hash tables, is $(1-e^{-N/H})^Z$.

While the error rate can easily be calculated from the hash table load, the average *miscount* — the degree to which the measured count differs from the true count — depends on the k-mer frequency distribution. Because next generation sequencing data sets are dominated by low abundance k-mers from sequencing errors, these miscount values are generally low.

The khmer software implementation offers good performance, a robust and well-tested Python API, and a number of useful and well-documented scripts. While khmer does not always perform better than other k-mer counting software, it is competitive. In memory-limited situations with poor I/O performance, khmer is particularly useful, because it will not break an imposed memory bound and does not require disk access to store or retrieve k-mer counts. Also, because it provides a Python API for online counting, it is flexible and easy to expand. This is why it has been used in different tasks from microbial diversity estimation to metagenome assembly. It will still be a substantial tool for most of the tasks proposed below.

### 2.2.2 Preliminary results: median k-mer frequency of a read can represent sequencing depth

A novel method termed as "digital normalization" was developed to remove abundant reads from very large metagenomic data sets before assembly(Brown, Howe, Zhang, Pyrkosz, & Brom, 2012). Abundant reads are identified by estimating the sequencing coverage of reads.

Traditionally sequencing coverage of reads are estimated by mapping reads to an assembly or reference genome, which is impossible for metagenomic reads since we have no assembly yet.

However a deduction from this method becomes a substantial starting point for IGSs based microbial diversity estimation - we can use median k-mer frequency to represent the sequencing coverage of a read.

Firstly each read comes from a specific segment of DNA of some species. The more times this segment of DNA is sequenced, the higher the abundance of k-mers from that read would be after we count all the k-mers in the sequencing data set. If there is no sequencing errors, we can estimate the sequencing coverage of a read - a segment of DNA of a species precisely. However the sequencing errors do exist and can cause many erroneous low abundance k-mers. For instance, one substitution error will introduce $k$ low abundance k-mers.(see Figure 1)So the average k-mer abundance of a read would be lower than actual sequencing coverage. However if we choose *median* k-mer abundance rather than *average* k-mer abundance, the bad influence of that *single* sequencing error can be eliminated, if the $k$ is relatively small compared to reads length $L$. (when $L > 3k - 1$,precisely). When multiple sequencing errors happen in a single read, the median k-mer abundance will be affected. However considering the sequencing error rate and reads length, the probability multiple sequencing errors happen in a single read is much lower than single error happens.

Using simulated and real genomic data sets, with the help of khmer to get median k-mer abundance, we find that median k-mer abundance correlates well with mapping-based coverage. see Figure 2

### 2.2.3 Preliminary results: IGS(informative genomic segment) can represent the novel information of a genome

From discussion above, median k-mer abundance can represent sequencing depth of a read. For a sequencing reads data set with multiple species, the sequencing depth of a read is related to the abundance of species where the read originates. Figure 3a shows the abundance distribution with different sequencing depth of reads from 4 simulated sequencing data sets - 3 sequencing data sets generated with different sequencing coverage from 3 simulated random genomes respectively and 1 combined data set with all the previously mentioned data sets. No error is introduced in these simulated data sets. Obviously the reads from the three data sets can be separated by estimated sequencing depth. The combined data set can be considered as a sequencing data set with three species with different abundance.

Each point on the curve shows that there are $Y$ reads with a sequencing depth of $X$. In other word, for each of those Y reads, there are $X - 1$ other reads that cover the same DNA segment in a genome that single read originates. So we can estimate that there are $Y/X$ distinct DNA segments with reads coverage as $X$. We term these distinct DNA segments in species genome as IGS(informative genomic segment). We can transform Figure 3a to show the number of IGSs and their respective reads coverage, as shown in Figure 3b. We sum up the numbers of IGSs with different reads coverage for each data set and get the result as shown in Table 1. The sum numbers of IGSs here essentially are the areas below each curve in Figure 3b.

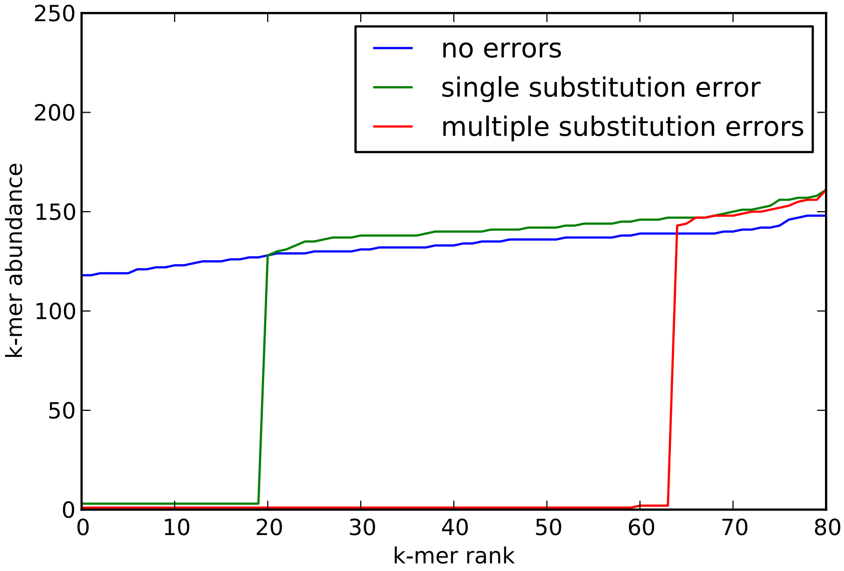

Figure 1: **Representative rank-abundance distributions for 20-mers from 100-base reads with no errors, a read with a single substitution error, and a read with multiple substitution errors.**
(Brown et al., 2012)

Table 1: **Total number of IGSs in different simulated reads data sets.**

| Data set | total number of IGSs |
|---|---|
| 1X depth | 8714 |
| 10X depth | 16321 |
| 40X depth | 16794 |
| 1X,10X,40X combined | 41742 |

Even though the datasets have different sequencing depth like 10X and 40X, they have similar numbers of IGSs. Dataset with 1X sequencing depth has fewer IGSs because the depth is not enough to cover all the content of the genome(63.2% (Lander & Waterman, 1988)) Essentially it is the maximum number of segments with length L on a genome out of which no two segments share any single k-mer. See Figure 4. Assume the species genome is totally random, which is the case in the simulated data set, the number of IGSs($N$) in a species genome is related to the size of genome($G$), read length($L$) and k size($k$), which can be denoted as

$$N = G/(L - k + 1)$$

For the simulated genome with size of 1M bps, read length as 80bps, k-mer size as 22bps, expected number of IGSs is $1000000/(80 - 22 + 1) = 16949$, pretty close to observed value.

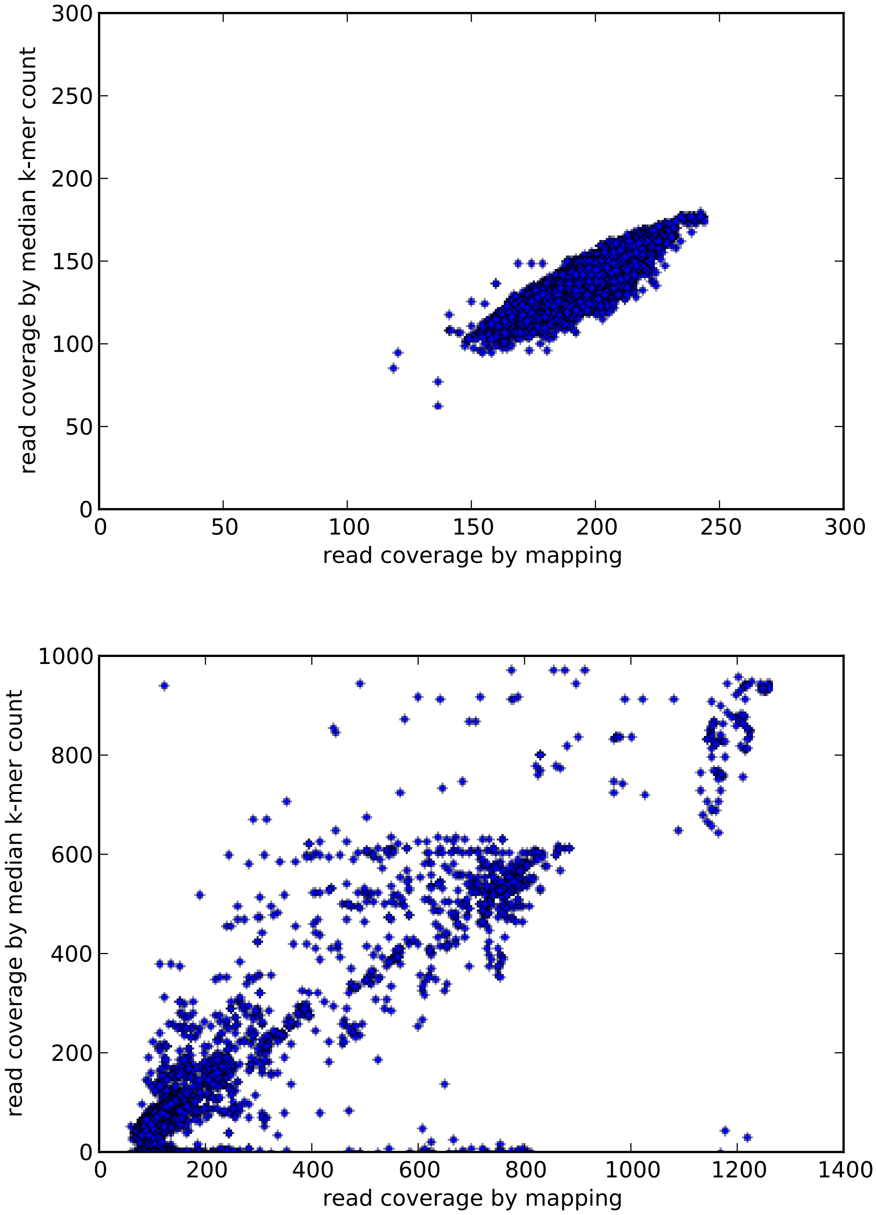

Figure 2: **Mapping and k-mer coverage measures correlate for simulated genome data and a real *E. coli* data set (5m reads). Simulated data $r^2 = 0.79$; *E. coli* $r^2 = 0.80$.**
(Brown et al., 2012)

Next we test the analysis using simulated sequencing reads data sets with sequencing error introduced according to Illumina error profile, see Figure 5 The patterns are similar to the experiment without sequencing error. However lots of reads with sequencing depth as

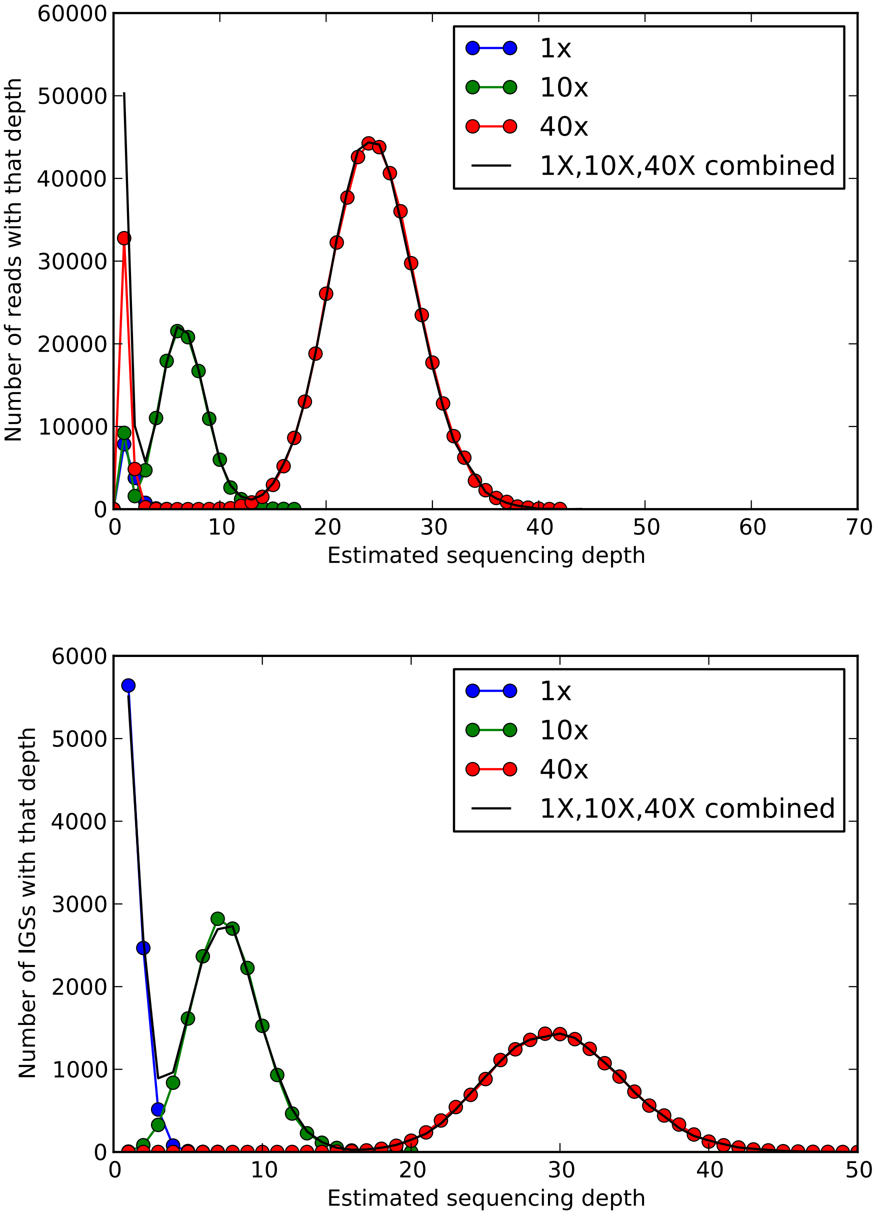

Figure 3: **From reads to IGSs, simulated data set, without sequencing error**

1 emerge. For high sequencing depth data set, most of them are generated by sequencing errors. So using IGSs based on using median k-mer abundance to represent sequencing depth can decrease the influence of sequencing errors, but can not eliminate it.

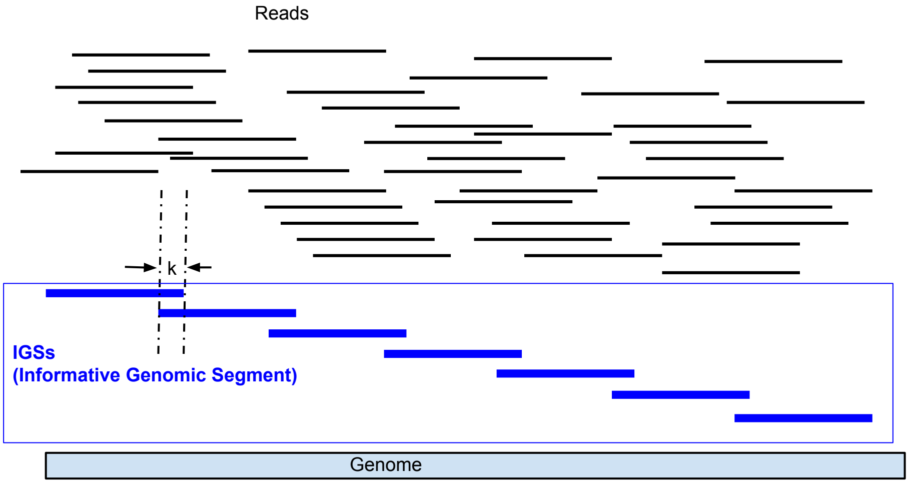

Figure 4: **IGSs are the segments with specific length on genome that do not share any k-mers with each other.**

### 2.2.4 Subaim 1a: refine the concept of IGSs

The concept of IGSs is promising to be used as the cornerstone for further investigation of microbial diversity. We still need to have better understanding of this novel concept.

Firstly it is necessary to do a survey of IGSs in real microbial genomes, like the abundance or distribution. Also it is required to investigate the relationship between the effectiveness of using IGSs and the size of microbial genomes and the repeat distribution in microbial genomes, since the similar genome size and random genome contents are the premise of using IGSs to estimate species diversity. If it can not be satisfied, which will be true in real microbial genome, what degree of influence it will have on the effectiveness should be investigated.

The influence of sequencing errors is still haunting. More efforts should be made to look into the statistics behind of the influence of sequencing errors. According to the statistics of using median k-mer abundance to represent sequencing depth, the influence of sequencing errors may not be as serious as observed. Certain modification to the algorithms of median k-mer abundance may be needed to mitigate the influence of sequencing errors. Other approaches also need to be investigated, like preprocessing data or do error correction using other tools beforehand.

### 2.2.5 Subaim 1b: "non-redundant reads" - another concept based on median k-mer abundance in read

Besides of the concept of IGSs, another method based on median k-mer abundance in read is also used to estimate diversity and compare metagenomics samples. In digital normalization

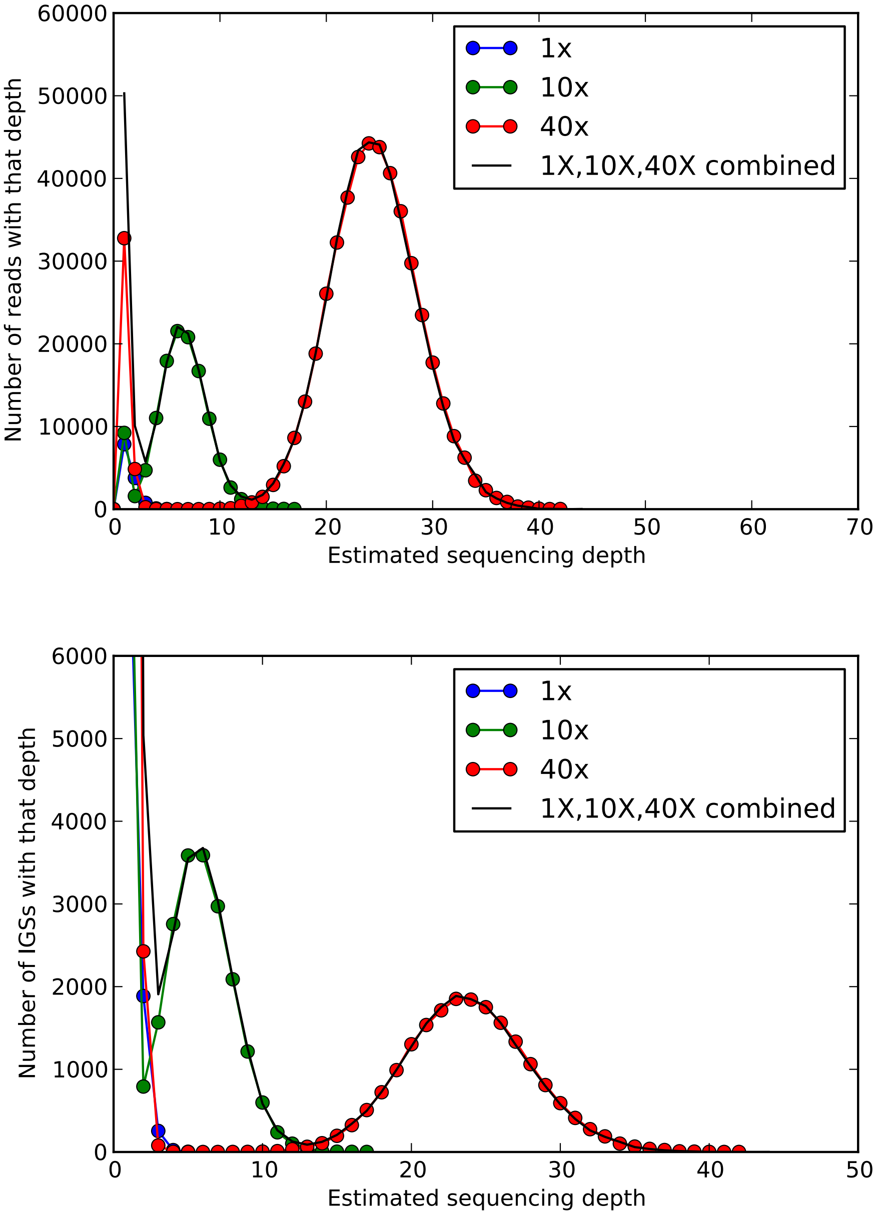

Figure 5: **From reads to IGSs, simulated data set, with sequencing error**

protocol, ”non-redundant reads” is defined as the reads with coverage lower than a specific number C. All ”redundant reads” will be discarded in digital normalization. This process can be shown as following pseudocode:

```
for read in dataset:
    if estimated_coverage(read) < C:
        accept(read)
    else:
        discard(read)
```

Firstly we can draw accumulation curve based on these "non-redundant reads".Without the influence of sequencing errors, the number of non-redundant reads will increase with the increase of sampled reads. However it will saturate after the sampled reads have covered all the genome with certain coverage C. No more new reads will have an estimated coverage lower than C.

The "non-redundant reads" concept is related to IGSs, especially if we use C as 1. In this case, only the reads that have current estimated coverage as 0 will be considered as "non-redundant reads". Basically these "non-redundant reads" are all the possible reads in a genome that do not share more than half of the k-mers in the read with each other, while IGSs are all the possible reads in a genome that do not share any k-mer. So the number of "non-redundant reads" is also related to the underlying species genome and can also be used as markers to estimate diversity.

Figure 6 shows the accumulation curves of "non-redundant reads" for those simulated sequencing data sets we used above. Obviously the accumulation curves show that the richness of the combined data set is higher than each data set with single species. The curves of the two data sets with single species are close to each other even they have different sequencing depth(10X vs 40X). We set the y-axis scale as same as x-axis scale on purpose to show the trend of saturation. Because of the sequencing errors, all three accumulation curves will never stop climbing. However the slop of the increase of curves based on "non-redundant reads" is much milder than the curves based on "distinct k-mers" that we tested before. This reiterate the conclusion that using median k-mer abundance can decrease the influence of sequencing error, although can not eliminate it.

This method of identifying "non-redundant reads" introduced in this section will be used for other purpose as discussed below. The relationship between the concept of "non-redundant reads" and IGSs is worth further investigation. Integration of the two concepts are also possible.

### 2.2.6 Subaim 1c: Borrowing statistical methods from OTU based diversity analysis

Many of the statistical methods used for OTU based diversity analysis can be used based on IGSs or "non-redundant reads".

Accumulation curves are shown in previous section. Some parametric estimators will benefit from our novel concept of IGSs or "non-redundant reads" since their abundance distribution gives some hints about the underlying species abundance distribution,which is important for parametric estimators.

Non-parametric estimators like Chao1, ACE, ICE can also be tested since we can get the abundance of IGSs or "non-redundant reads", which is basically the number of times that specific segment on genome are seen. Here the influence of sequencing error should be considered seriously, since the errors will generate many erroneous singleton - with abundance as 1. This will ruin the effectiveness of Chao1 especially, because this estimator relies heavily

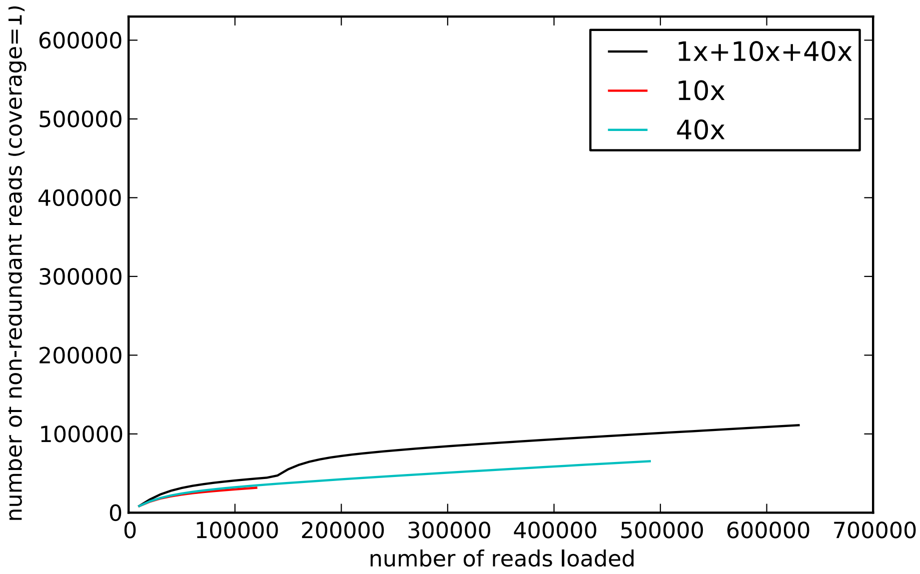

Figure 6: **Accumulation curve of ”non-redundant reads” of three simulated sequencing data sets**

on low abundance units like singleton and doubletons.

Other methods to estimate diversity and evenness should also be considered. A recent publication(Haegeman et al., 2013) argued the estimation of species richness is intrinsically so problematic that 1) species richness cannot be estimated from sample data alone 2) Relative species richness cannot be estimated from sample data alone, because of the existence of large number of rare species in microbial community(Mao & Colwell, 2005; Bent & Forney, 2008). Diversity indexes taking both richness and evenness into account like Shannon and Simpson are recommended, especially the generalized Hill index. These indexes are discussed in section Concept of Diversity above.

### 2.2.7 Subaim 1d: Sequencing depth of metagenomics dataset

Median k-mer abundance in a read correlates to sequencing depth as discussed above. Further investigation can be done based on this and question like ”how much more metagenomic sequencing is needed to achieve a specific sequencing depth?” needs to be answered.

## 2.3 Specific Aim 2: Comparing multiple metagenomic samples

In previous section, concepts and methods are discussed to estimate the diversity of one metagenomic sample, which is $\alpha$-diversity. Next we will investigate the $\beta$-diversity - the comparison of metagenomic samples.

### 2.3.1 Specific Aim 2a: Compare the diversity of metagenomics samples

Basically the ultimate solution to this problem is get the absolute/quantitative measurement of diversity of samples using methods to work on each sample independently then we can compare those measurement directly. We have discussed the direct methods to evaluate microbial diversity of a sample in section above.

However we can still have some idea about the difference of diversity between samples by comparing the characteristics directly without knowing the absolute or quantitative measurement explicitly.For richness comparison, we may have difficulty to get a quantitative measurement of richness for sample, but we can compare the rarefaction curves to have some idea, although theoretically this has serious flaws especially if the rarefaction curve is far from saturation. See Figure 6 for an example.

The shape of rarefaction curve is related to the abundance distribution. It is worth trying to do the comparison qualitatively of evenness in this way.

### 2.3.2 Specific Aim 2b: Compare the contents of multiple metagenomics samples

This is a different question from previous one. This comparison will give the information about how similar the contents of two samples are. Here similarity means the common elements shared by two or more samples. That samples have similar richness only means there are similar number of species in these samples. This does not necessarily mean they share many common species. After we get the common reads we can do annotation to those reads to have some biological insight of the similarity of samples, like what kind of species the samples share. This information can also facilitate the process of metagenome assembly of large diverse samples. If we find out two samples share large amount of contents, we can combine the sequencing reads and do assembly taking advantage of the larger number of reads to have better coverage of rare species.

There have been some work on content-based similarity measures and similar metagenomic data retrieval(Mitra, Klar, & Huson, 2009; Su, Xu, & Ning, 2012; Liu, Hsiao, Cantarel, Drábek, & Fraser-Liggett, 2011), most of which are based on a relatively small number of features, often coming from existing annotation information. Recently there is an paper on arXiv (Seth, Välimäki, Kaski, & Honkela, 2013) demonstrating a method using counting k-mers to estimate the similarity of metagenomic data sets. Rather than using all the k-mers, a subset of "informative" k-mers are selected to decrease the computation complexity. The performance of their method on extremely large data set is not clear.

Here we use the counting of general common elements in two metagenomic data sets to evaluate the similarity. Previously we also tried to use distinct k-mers as the common elements. Because of the serious influence of sequencing error, the inflation of low abundance erroneous k-mers makes that k-mer based approach infeasible.

Now we turn to reads as the common elements. Basic idea is similar to the identification of "non-redundant reads". The difference is that here we use the coverage of a sampleA read $R_A$in sampleB to test if the DNA segment in sampleA that read $R_A$ covers also exists in sampleB.

We simulated several test data sets to see how this method works. Firstly we simulated four sequencing data sets with 100 species but different types of species respectively and with sequence errors introduced by Illumina error profile. SampleA and Sample B share 80 common species. SampleC and sampleD share 60 common species and both of them share 20 common species with either sampleA or sampleB. All the 100 species have even abundance distribution for simplicity. Table 2 shows the percentage of reads in each sample that have coverage in another sample. It fits the underlying similarity of species pretty well. For example sampleA and sampleB share 80 common species. Now 74.3% of reads in sampleA have coverage in sampleB. The difference between 74.3% and 80% is probably due to sequencing error. But the influence is not that much actually.

Table 3 shows the result for four simulated datasets with same composition of species but with lower sequencing depth.(1X compared to 20X as in Table 2) This shows the observed percent of reads in one sample covered by another sample is related to the sequencing depth. This is easy to explain, since lower sequencing depth will not cover all the contents in the sample. But as long as the sequencing depth is high enough to cover all the contents ( like >10X), the percent reflects the actual similarity of genomic contents in different samples.

The first two series of simulated data sets assume even species distribution, which is not true for real microbial sample. For the next experiment we use more complicated data sets, with uneven species abundance distribution.

SampleA has 10 species with ID (1,2,3,4,5,6,7,8,9,10) and uneven relative abundance (20:18:16 :4:3:2:2:2:2:2).

SampleB has 10 species with ID (1,2,3,14,15,16,17,18,19,20) and uneven relative abundance (20:18:16 :4:3:2:2:2:2:2).

SampleB has 10 species with ID (21,22,3,4,5,6,7,8,9,10) and uneven relative abundance (2:2:2 :2:2:3:4:16:18:20).

In summary, A and B have high similarity on individual level, low similarity on species level. A and C have high similarity on species level, low similarity on individual level. B and C low similarity on species level and low similarity on individual level.

Table 4 shows that this method actually estimate the similarity on genomic contents, or individual level. SampleA and sampleB only shares 3 common species, but all high abundance species, so from Table 4, 70% of reads in sampleA have coverage in sampleB.

However we can use digital normalization to normalize the species abundance by discarding high abundance reads. We normalized the abundance of the 10 species in each sample to have an even abundance distribution, with sequencing depth of 5X for all of the species. Then Table 5 shows using this method to normalized data sets can get the actual similarity on species level, as 35% of reads in sampleA have coverage in sampleB.

Further work need to be done to integrate standard dissimilarity metrics like Jaccard distance to have better understanding of the question and result.

### 2.3.3 Specific Aim 2c: Tackle the extremely complex metagenomics samples

Based on an efficient k-mer counting approach that scales well, we have the advantage that our methods can be used to deal with the extremely complex metagenomics sample with extremely large sequencing data set being generated.

We will firstly apply the methods on some smaller real datasets like from Human Gut project. Finally we will tackle the most diverse microbial community on earth - soil by

Table 2: **20X, with error, 100 species, even abundance distribution**

| | sampleA | sampleB | sampleC | sampleD |
|---|---|---|---|---|
| sampleA | | 74.3% | 18.6% | 18.6% |
| sampleB | 74.3% | | 18.6% | 18.6% |
| sampleC | 18.6% | 18.6% | | 55.7% |
| sampleD | 18.5% | 18.5% | 55.8% | |

Table 3: **1X, with error,100 species, even abundance distribution**

| | sampleA | sampleB | sampleC | sampleD |
|---|---|---|---|---|
| sampleA | | 30.2% | 7.5% | 7.5% |
| sampleB | 30.2% | | 7.6% | 7.6% |
| sampleC | 7.2% | 7.3% | | 22.7% |
| sampleD | 7.2% | 7.3% | 22.7% | |

Table 4: **6X, with error, 10 species,uneven abundance distribution**

| | sampleA | sampleB | sampleC |
|---|---|---|---|
| sampleA | | 71.3% | 38.4% |
| sampleB | 71.3% | | 2.6% |
| sampleC | 21.8% | 8.5% | |

Table 5: **with error, 10 species, digital normalized to 5X coverage**

| | sampleA | sampleB | sampleC |
|---|---|---|---|
| sampleA | | 34.8% | 68.2% |
| sampleB | 35.3% | | 10.9% |
| sampleC | 58.1% | 11.0% | |

exploiting the terabytes of sequences. We will estimate the sequencing coverage to determine how much more effort we need to satisfy our research need. We have metagenomic data sets of soil samples gathered from different locations with different historical treatment. We hope we can get some scientific insights from estimating the diversity of each samples and the comparison between samples. We are still facing serious challenge in assembling the soil metagenome since the coverage is still low and the community is so diverse. By comparing the contents of the metagenomic data sets from different samples, we will know if they are similar enough to be combined for a co-assembling effort to have better contigs.

Although flaws and disadvantages still exist in our methods, trying to decode these extremely diverse microbial communities with numerous unknowns and ”dark matter” will be worthy effort and any insights from our work will be valuable.

## 2.4 Additional idea

Firstly the methods described in the proposal will be integrated into khmer package. Also it is worth trying to use both shotgun whole genome metagenomics data and 16S rRNA amplicon metagenomic data for the same sample. It seems that people are busy working on either aspect only but not many try to integrate the two types of metagenomic data. Hopefully this can give some insightful result.

# References

Bent, S. J., & Forney, L. J. (2008, Jul). The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *ISME J*, *2*(7), 689-95.

Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, *13*(7), 422-426.

Brady, A., & Salzberg, S. (2011, May). Phymmbl expanded: confidence scores, custom databases, parallelization and more. *Nat Methods*, *8*(5), 367.

Broder, A. Z., & Mitzenmacher, M. (2003). Survey: Network applications of bloom filters: A survey. *Internet Mathematics*, *1*(4), 485-509.

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., & Brom, T. H. (2012, 03). A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv preprint*.

Bunge, J. (2011). Estimating the number of species with catchall. *Pac Symp Biocomput*, 121-30.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010, May). Qiime allows analysis of high-throughput community sequencing data. *Nat Methods*, *7*(5), 335-6.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, 265–270.

Chao, A. (1987, Dec). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, *43*(4), 783-91.

Chao, A., & Yang, M. C. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, *80*(1), 193–201.

Colwell, R. (n.d.). Estimates: statistical estimation of species richness and shared species from samples. 2004. *Consultado en: http://viceroy. eeb. uconn. edu/estimates*.

Colwell, R. K., Mao, C. X., & Chang, J. (2004). Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, *85*(10), 2717–2727.

Cormode, G., & Muthukrishnan, S. (2005, April). An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, *55*(1), 58–75.

Curtis, T. P., Sloan, W. T., & Scannell, J. W. (2002, Aug). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci U S A*, *99*(16), 10494-9.

Gans, J., Wolinsky, M., & Dunbar, J. (2005, Aug). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, *309*(5739), 1387-90.

Gilbert, J. A., & Dupont, C. L. (2011). Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci*, *3*, 347-71.

Gilbert, J. A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C. T., Brown, C. T., et al. (2010). Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project. *Stand Genomic Sci*, *3*(3), 243-8.

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., & Meyer, F. (2010, Jan). Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc*, *2010*(1), pdb.prot5368.

Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology letters*, *4*(4), 379–391.

Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., & Weitz, J. S. (2013, Jun). Robust estimation of microbial diversity in theory and in practice. *ISME J*, *7*(6), 1092-101.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998, Oct). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*, *5*(10), R245-9.

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., et al. (2011, Jan). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, *331*(6016), 463-7.

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, *54*(2), 427–432.

Howe, A. C., Jansson, J., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., & Brown, C. T. (2012, 12). Assembling large, complex environmental metagenomes. *arXiv preprint*.

Human Microbiome Jumpstart Reference Strains Consortium, Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., et al. (2010, May). A catalog of reference genomes from the human microbiome. *Science*, *328*(5981), 994-9.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007, Mar). Megan analysis of metagenomic data. *Genome Res*, *17*(3), 377-86.

Jones, D. C., Ruzzo, W. L., Peng, X., & Katze, M. G. (2012, 07). Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *arXiv preprint*.

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011, Jun). Human nutrition, the gut microbiome and the immune system. *Nature*, *474*(7351), 327-36.

Kurtz, S., Narechania, A., Stein, J. C., & Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, *9*(1), 517.

Lander, E. S., & Waterman, M. S. (1988, Apr). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, *2*(3), 231-9.

Lee, S.-M., & Chao, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics*, 88–97.

Liu, Z., Hsiao, W., Cantarel, B. L., Drábek, E. F., & Fraser-Liggett, C. (2011, Dec). Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics*, *27*(23), 3242-9.

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J. P., Hector, A., et al. (2001, Oct). Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science*, *294*(5543), 804-8.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., & Konstantinidis, K. T. (2012). Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community dna sample. *PLoS One*, *7*(2), e30087.

Magurran, A. E., & McGill, B. J. (2011). *Biological diversity: frontiers in measurement and assessment* (Vol. 12). Oxford University Press Oxford.

Mao, C. X., & Colwell, R. K. (2005). Estimation of species richness: mixture models, the role of rare species, and inferential challenges. *Ecology*, *86*(5), 1143–1153.

Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., et al. (2012, Jan). Img/m: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res*, *40*(Database issue), D123-9.

Mason, O. U., Hazen, T. C., Borglin, S., Chain, P. S. G., Dubinsky, E. A., Fortney, J. L., et al. (2012, Sep). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to deepwater horizon oil spill. *ISME J*, *6*(9), 1715-27.

May, R. M. (1988, Sep). How many species are there on earth? *Science*, *241*(4872), 1441-9.

Melsted, P., & Pritchard, J. K. (2011, January). Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC bioinformatics*, *12*, 333.

Mitra, S., Klar, B., & Huson, D. H. (2009, Aug). Visual and statistical comparison of metagenomes. *Bioinformatics*, *25*(15), 1849-55.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011, Aug). How many species are there on earth and in the ocean? *PLoS Biol*, *9*(8), e1001127.

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*, *13*(9), R79.

Patil, K. R., Roune, L., & McHardy, A. C. (2012). The phylopythias web server for taxonomic assignment of metagenome sequences. *PLoS One*, *7*(6), e38581.

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M., & Brown, C. T. (2012, Aug). Scaling metagenome sequence assembly with probabilistic de bruijn graphs. *Proc Natl Acad Sci U S A*, *109*(33), 13272-7.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010, Mar). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59-65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012, Oct). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, *490*(7418), 55-60.

Quince, C., Curtis, T. P., & Sloan, W. T. (2008, Oct). The rational exploration of microbial diversity. *ISME J*, *2*(10), 997-1006.

Rizk, G., Lavenier, D., & Chikhi, R. (2013, Mar). Dsk: k-mer counting with very low memory usage. *Bioinformatics*, *29*(5), 652-3.

Roesch, L. F. W., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K. M., Kent, A. D., et al. (2007, Aug). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*, *1*(4), 283-90.

Rosen, G. L., Reichenberger, E. R., & Rosenfeld, A. M. (2011, Jan). Nbc: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, *27*(1), 127-9.

Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol*, *31*, 107-33.

Schloss, P. D., & Handelsman, J. (2005, Mar). Introducing dotur, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol*, *71*(3), 1501-6.

Schloss, P. D., & Handelsman, J. (2006, Jul). Toward a census of bacteria in soil. *PLoS Comput Biol*, *2*(7), e92.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009, Dec). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*, *75*(23), 7537-41.

Schmidt, T. M., DeLong, E. F., & Pace, N. R. (1991, Jul). Analysis of a marine picoplankton community by 16s rrna gene cloning and sequencing. *J Bacteriol*, *173*(14), 4371-8.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012, Aug). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*, *9*(8), 811-4.

Seth, S., Välimäki, N., Kaski, S., & Honkela, A. (2013). Exploration and retrieval of whole-metagenome sequencing samples. *arXiv preprint arXiv:1308.6074*.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3–55.

Simpson, E. H. (1949). Measurement of diversity. *Nature*.

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., et al. (2006, Aug). Microbial diversity in the deep sea and the underexplored ”rare biosphere”. *Proc Natl Acad Sci U S A*, *103*(32), 12115-20.

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A. D., Kämpfer, P., Maiden, M. C. J., et al. (2002, May). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol*, *52*(Pt 3), 1043-7.

Su, X., Xu, J., & Ning, K. (2012, Oct). Meta-storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. *Bioinformatics*, *28*(19), 2493-501.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., & Glöckner, F. O. (2004, Oct). Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, *5*, 163.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009, Jan). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480-4.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., et al. (2004, Mar). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37-43.

Tzahor, S., Man-Aharonovich, D., Kirkup, B. C., Yogev, T., Berman-Frank, I., Polz, M. F., et al. (2009). A supervised learning approach for taxonomic classification of core-photosystem-ii genes and transcripts in the marine environment. *BMC Genomics*, *10*, 229.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004, Apr). Environmental genome shotgun sequencing of the sargasso sea. *Science*, *304*(5667), 66-74.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007, Aug). Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Appl Environ Microbiol*, *73*(16), 5261-7.

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998, Jun). Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A*, *95*(12), 6578-83.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., et al. (2009, Dec). A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*, *462*(7276), 1056-60.

Wu, M., & Scott, A. J. (2012, Apr). Phylogenomic analysis of bacterial and archaeal sequences with amphora2. *Bioinformatics*, *28*(7), 1033-4.

Zarraonaindia, I., Smith, D. P., & Gilbert, J. A. (2013, Mar). Beyond the genome: community-level analysis of the microbial world. *Biol Philos*, *28*(2), 261-282.

Zhang, Q., Pell, J., Canino-Koning, R., Howe, A. C., & Brown, C. T. (2013). These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *arXiv preprint arXiv:1309.2975*.