# Using the concept of informative genomic segment to investigate microbial diversity of metagenomics sample

TBD

January 14, 2016

## Contents

**Abstract**

# 1 Introduction

Species diversity is an important measurement of ecological communities. Scientists believe that there is a relationship between species diversity and ecosystem processes [**?**]. Evaluating the species diversity in a community is a central research topic in macroorganism ecology. Many methods have been developed over the last few decades, aimed at answering questions such as "how many species of birds are in this habitat". Nevertheless, until recently scientists had not started to think seriously about larger-scale questions such as "How many species are there on earth?" [**?**] or "How many species are there in the ocean?" [**?**] until recently. Why? The answer is straightforward: Microorganisms represent the vast majority of the Earth's biodiversity and the assessment of microbial diversity is quite difficult.

It is believed that microbial diversity is the outermost frontier of the exploration of diversity [**?**]. Microorganisms are ubiquitous. There are more bacterial cells in our body than human cells [**?**]. There are several reasons why assessment of microbial diversity is such a challenge. First, the concept of species is ambiguous. Morphological examination is impossible: fewer than 1% of microorganisms in the biosphere cannot be cultivated by traditional cultivation techniques[**?**]. To overcome this obstacle, metagenomics has emerged, driven by the progress of next-generation sequencing (NGS) technology. Lots of metagenomics projects have been performed on samples ranging from acid mine drainage channels to human gut.

In almost all metagenomics projects, diversity analysis plays an important role in supplying knowledge about the richness of species, the abundance distribution of species in a sample, and the similarity and difference between different samples. The topic of microbial diversity measurement has been investigated for a long time with many methods and software packages developed. However, there still remains lots of room for more work.

Traditionally used for amplicon metagenomics data set, OTUs(Operational Taxonomic Units) based on 16S rRNA genes are used as the basic units for diversity analysis on shotgun metagenomic data. OTUs can be good replacements

of the concept of "species" in metagenomics. Basically contigs are assembled from reads and are "binned" into OTUs using composition-based or similarity-based approaches. Then the diversity can be estimated by using the abundance information of the OTUs. The mainstream methods to measure microbial diversity are still focusing on the use of 16S rRNA amplicon metagenomics data. Many of the popular microbial diversity analysis software packages generally accept 16S rRNA data as input. This is understandable because 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. The 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 bacteria. There have been suggestions that alternative marker genes should be used, such as single copy housekeeping genes. Thus, measuring diversity beyond using 16S rRNA data is worth investigating. Recently there are many more 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 of large size because of the high diversity. For instance, there may be 4 petabase pairs of DNA in a gram of soil [**?**]. 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 approaches involve k-mer/signature frequency distribution calculation, which is rather computationally expensive. Even basic sequence alignment will be impossible for large metagenomics data sets. 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 NGS technology, the cost of sequencing is dropping rapidly. Whole genome metagenomics sequencing is more popular and a large amount of metagenomics data is being generated with increasing speed, which cannot be even met by the increase of computational capacity. For complex environmental samples such as soil, the resulting data sets can 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, the resulting metagenomics data sets from highly complex samples (e.g. soil) only correspond to a tiny fraction of the actual genomic content in the sample. The large size of data sets and the low sequencing coverage make the assessment of microbial diversity of high diversity sample even harder.

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. While the concept of OTUs can be used to analyze large shotgun metagenomics data set, normally assembly,

binning and annotation are required before doing diversity analysis. However for many metagenomics projects these are difficult tasks, lacking necessary reference genome or requiring expensive computation. So we are interested in finding an approach to bypass difficult tasks like assembly, binning, annotation and use the raw reads to make the diversity analysis of large shotgun whole genome metagenomic data possible.

We began such efforts by proposing that the concept of k-mer (a DNA segment with the length of k) could be used as the basic unit to measure the 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-mers as the source of signatures[**?**, **?**]. 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 correlates 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 does not work well on unassembled data. One sequencing error on a read will generate up to k erroneous k-mers. In metagenomics data sets, especially with high coverage, most of the distinct observed k-mers are from sequencing errors.

Next we shifted the focus from k-mers onto a higher level - reads. In previous chapter, we have discussed a novel approach to use median k-mer count in a read to estimate sequencing depth without a reference assembly, based on which the framework for streaming analysis of short DNA sequencing reads was developed. It also offers a novel way to distill information from reads by reducing the bad effect of sequencing errors so that we can use those informative reads to measure the microbial diversity. We term those informative reads as IGSs (informative genomic segments), which can be considered segments 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; then the number of IGSs will correlate with the species richness in a sample, and the abundance distribution of IGSs will be related to species evenness in a sample. Furthermore, we can get the abundance of the IGSs across different samples. Many classic diversity estimation methods based on OTUs described in the literature review chapter can be applied to estimate the diversity of IGSs and the diversity of actual species subsequently.

For alpha diversity, we can generate a list of IGSs and the respective abundance in a sample. Then existing estimators like Chao's can be applied to estimate the total number of IGSs in the sample. Rarefaction curves based on the number of IGSs can also be generated.

For beta diversity, we can generate a samples-by-IGS data matrix from the abundance of IGSs across samples, as a replacement of samples-by-OTU data matrix in OTU-based analysis and samples-by-species data matrix in traditional ecology. From that samples-by-IGS data matrix, we can use existing methods to calculate similarity/dissimilarity/distance between samples and do further analysis like clustering and ordination.

We applied the IGS-based method to several simulated data sets and sev-

eral real data sets - Global Ocean Sampling Expedition (GOS) to do diversity analysis and the samples were clustered more accurately than existing alignment-based method. We also tried this novel method to Great Prairie Soil Metagenome Grand Challenge data sets. Furthermore we will show some preliminary results using the IGS-based method for alpha-diversity analysis. Since this method is totally binning-free, assembly-free, annotation-free, reference-free, it is specifically promising to deal with the highly diverse samples, while we are facing large amount of dark matters in it, like soil.

# 2 Results

## 2.1 IGS(informative genomic segment) can represent the novel information of a genome

Median k-mer abundance can represent sequencing depth of a read, as discussed in last chapter (published as [**?**]). 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 from.

The upper plot in Figure 1 shows the abundance distribution of reads from 4 simulated sequencing data sets with different sequencing depths - 3 sequencing data sets generated with different sequencing coverage(1x, 10x, 40x) from 3 simulated random genomes respectively and 1 combined data set with all the aforementioned 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 IGSs (informative genomic segments). We can transform the upper plot in Figure 1 to show the number of IGSs and their respective reads coverage, as shown in lower plot. 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 the figure.

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%). The IGSs can be seen as the genomic segments on a genome with the length of reads.(Figure 2 Assume the composition of species genome is totally random, which is the case in the simulated data sets, 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

$$N = \frac{G}{L}$$

which is the number of reads that can have a 1X coverage of the genome. For the simulated genome with size of 1M bps, read length as 80bps, expected number of IGSs is

$$1000000/80 = 12500$$

,

which is close to the observed value (Table 1).

| Data set | total number of IGSs |
|---|---|
| 1X depth | 6419 |
| 10X depth | 12022 |
| 40X depth | 12371 |
| 1X,10X,40X combined | 30748 |

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

## 2.2 Using a simple simulated data set to evaluate the IGS method

For this experiment, firstly we create 6 synthetic samples (Sample 1-6) based on 9 synthetic 100K genomes (genome A-I), with different composition of species and diversity (Table 2). For sample1, there are two species - A and B, with abundance distribution as 3:1. The sequencing depth of all the synthetic data sets is 10X. As a simple experiment to demonstrate the effectiveness of the IGS based method, there is no sequencing error introduced in the synthetic reads data sets.

| sample ID | species composition | sequencing depth | abundance of species | size of metagenome (bp) |
|---|---|---|---|---|
| sample1 | AAAB | 10 | A:30 B:10 | 200K |
| sample2 | AABC | 10 | A:20 B:10 C:10 | 300K |
| sample3 | ABCD | 10 | A:10 B:10 C:10 D:10 | 400K |
| sample4 | ABCE | 10 | A:10 B:10 C:10 E:10 | 400K |
| sample5 | AFGH | 10 | A:10 F:10 G:10 H:10 | 400K |
| sample6 | IFGH | 10 | I:10 F:10 G:10 H:10 | 400K |

Table 2: **Six synthetic simple metagenomes**

To evaluate the effectiveness of alpha diversity analysis using IGS based method, we can use a metric to estimate the total number of IGSs in a sample, which can be used to calculate the estimated genome size of a sample using the formula below: size of genome = number of IGS $*$ reads_length

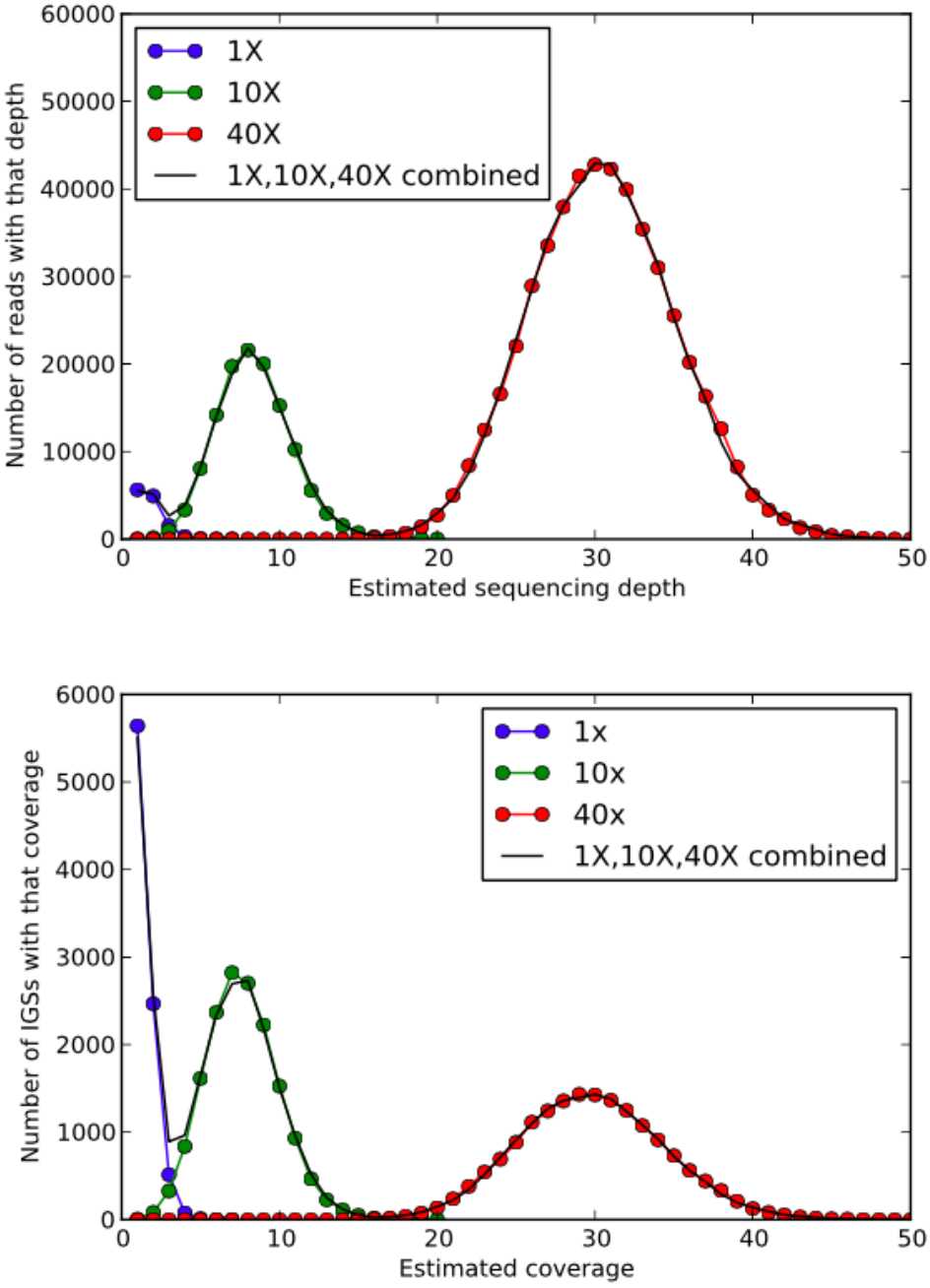


Figure 1: **Transforming reads into IGSs.**

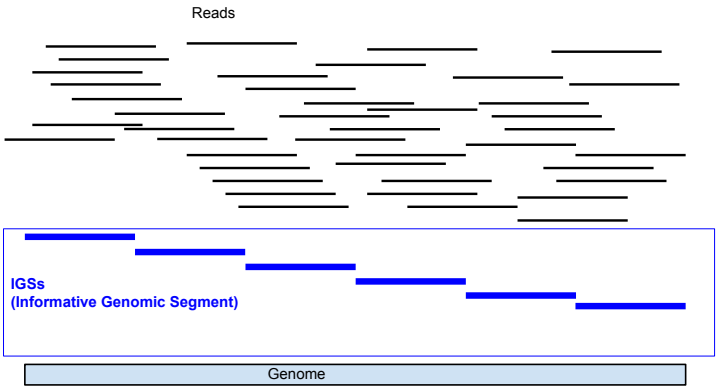


Figure 2: **IGS(informative genomic segment) can represent the novel information of a genome.**

| | observed IGS | ACE | simpson evenness | estimated genome size (Kbp) | real genome size (Kbp) |
|---|---|---|---|---|---|
| sample1 | 2002 | 2002.0 | 0.76 | 200.2 | 200 |
| sample2 | 3038 | 3038.0 | 0.83 | 303.8 | 300 |
| sample3 | 4076 | 4076.0 | 0.91 | 407.6 | 400 |
| sample4 | 4078 | 4078.0 | 0.91 | 407.8 | 400 |
| sample5 | 4069 | 4069.0 | 0.91 | 406.9 | 400 |
| sample6 | 4087 | 4087.0 | 0.91 | 408.7 | 400 |

Table 3: **Alpha diversity analysis result of the simple simulated data using IGS method.**

In this experiment, we use ACE metric since we find it is more accurate than Chao1, since it uses more abundance information.

Table 3 shows the alpha diversity analysis result of the simple simulated data using IGS method. The estimated genome sizes of the samples are close to real size. This is not surprising since for this simple experiment, there is no error introduced and the coverage is high (10x) to cover most of the genetic materials in the samples. Also the Simpson evenness shows the relative evenness of the samples correctly. Sample 1 is the least even with composed of two species with abundance ratio as 1:3. This shows that in this simple example, the IGS method cannot only analyze the richness of samples but also the evenness.

To evaluate the effectiveness of beta diversity analysis using IGS based method, we compared the dissimilarity matrix generated by IGS based method with the true matrix, since we know exactly the species composition of the simulated data set.

The true dissimilarity matrix of the 6 simulated samples using Bray-Curtis

| | sample 1 | sample2 | sample 3 | sample 4 | sample 5 | sample 6 |
|---|---|---|---|---|---|---|
| sample 1 | 0.00 | 0.25 | 0.50 | 0.50 | 0.75 | 1.00 |
| sample 2 | 0.25 | 0.00 | 0.25 | 0.25 | 0.75 | 1.00 |
| sample 3 | 0.50 | 0.25 | 0.00 | 0.25 | 0.75 | 1.00 |
| sample 4 | 0.50 | 0.25 | 0.25 | 0.00 | 0.75 | 1.00 |
| sample 5 | 0.75 | 0.75 | 0.75 | 0.75 | 0.00 | 0.25 |
| sample 6 | 1.00 | 1.00 | 1.00 | 1.00 | 0.25 | 0.00 |

Table 4: **Dissimilarity matrix between synthetic samples using Bray-curtis from species composition directly.**

| | sample 1 | sample2 | sample 3 | sample 4 | sample 5 | sample 6 |
|---|---|---|---|---|---|---|
| sample 1 | 0.00 | 0.35 | 0.60 | 0.66 | 0.80 | 1.00 |
| sample 2 | 0.35 | 0.00 | 0.42 | 0.51 | 0.84 | 1.00 |
| sample 3 | 0.60 | 0.42 | 0.00 | 0.56 | 0.89 | 1.00 |
| sample 4 | 0.66 | 0.51 | 0.56 | 0.00 | 0.89 | 1.00 |
| sample 5 | 0.80 | 0.84 | 0.89 | 0.89 | 0.00 | 0.42 |
| sample 6 | 1.00 | 1.00 | 1.00 | 1.00 | 0.25 | 0.00 |

Table 5: **Dissimilarity matrix between synthetic samples using Bray-Curtis from sequencing reads using IGS method.**

metric from species composition directly is shown in Table 4. For a simulated data set with 10x coverage and no error introduced (which again will tell us the optimal performance of IGS method), the dissimilarity matrix can be calculated by using the IGS method, as shown in Table 5. We can see the absolute values in the matrix are not very close to those in the real matrix. However, the relative values correspond to those in the real matrix well enough to show the relative distance between each pair of samples. To get a objective metric, we use the Mantel [**?**] test to calculate the correlation value between the two matrixes. The correlation is 0.9714, which means a very strong correlation between the two matrices. Thus the dissimilarity matrix from the IGS method reflects the true relationship between samples effectively.

If the matrix can reflect the real relationship between samples reliably, the clustering and ordination will only be routine tasks.

Figure 3 and Figure 4 show that IGS method can yield similarity between samples correctly. Sample 5 and sample 6 are very close to each other on the figure, which matches their species composition.

The clustering and ordination are all from the dissimilarity matrix. We think comparing matrices directly makes more sense than comparing the clustering and ordination plots. So we will not show the clustering and ordination figure in other evaluations in this section. Mantel correlation will be used to measure the accuracy of beta diversity analysis.

These results show that the IGS method can work well on a simple scenario, with high sequencing depth (10X) and no sequencing error. Next we will

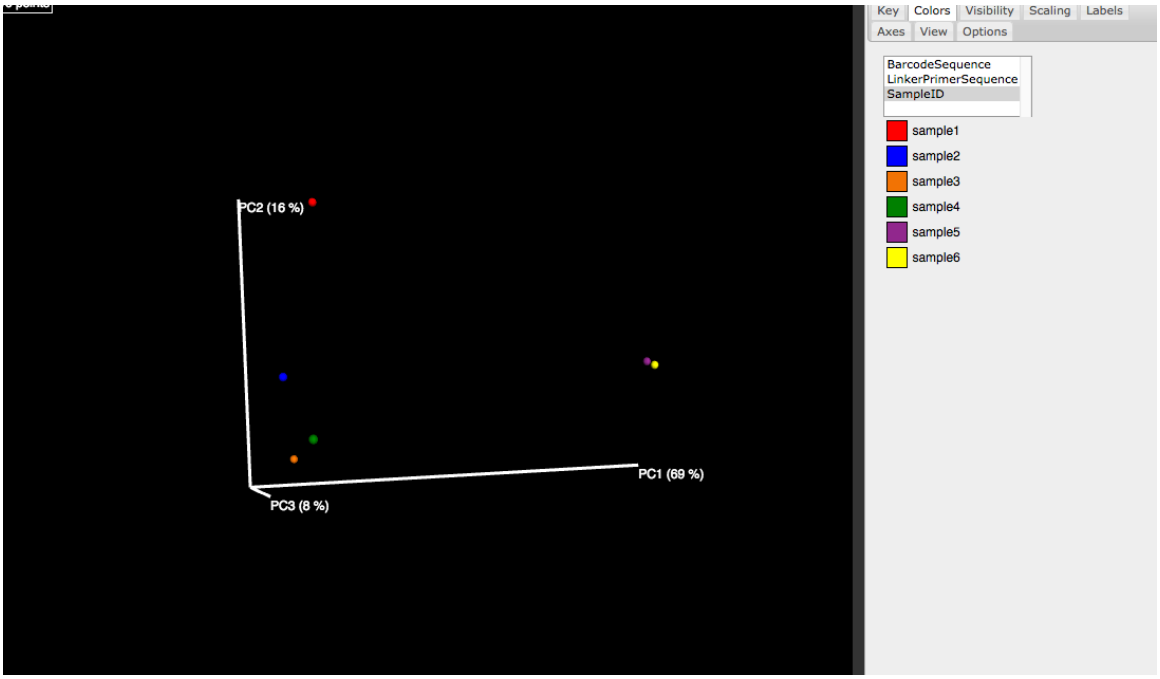


Figure 3: **Ordination of the 6 synthetic samples using IGS method.**

check the influence on the analysis accuracy of variable sequencing depth and sequencing error, and introduce new ways to preprocess the data to decrease the influence of sequencing error.

## 2.3 Improving the accuracy of this method in real world analysis

Previously we have shown the IGS method generally works on a simple simulated data set, with high sequencing depth and no sequencing error. In the real world, in many situations we have to deal with metagenomic data sets with relatively low sequencing depth, like soil or sea water samples. Also it is a fact that all sequencing technology generates some errors. As discussed in the introduction chapter, one of the reasons we developed the IGS method is that we expect the IGS method to be less prone to sequencing error based on the abundance counting of reads rather than k-mers. However the effect of those factors on the accuracy should still be observable.

In this section, we will analyze the effect of these factors on the accuracy of the IGS method and investigate ways we can reduce the effect in order to increase the accuracy of analysis.

As in last section, six synthetic samples were generated with the same species composition with same coverage as 10X but with different sequencing error rate

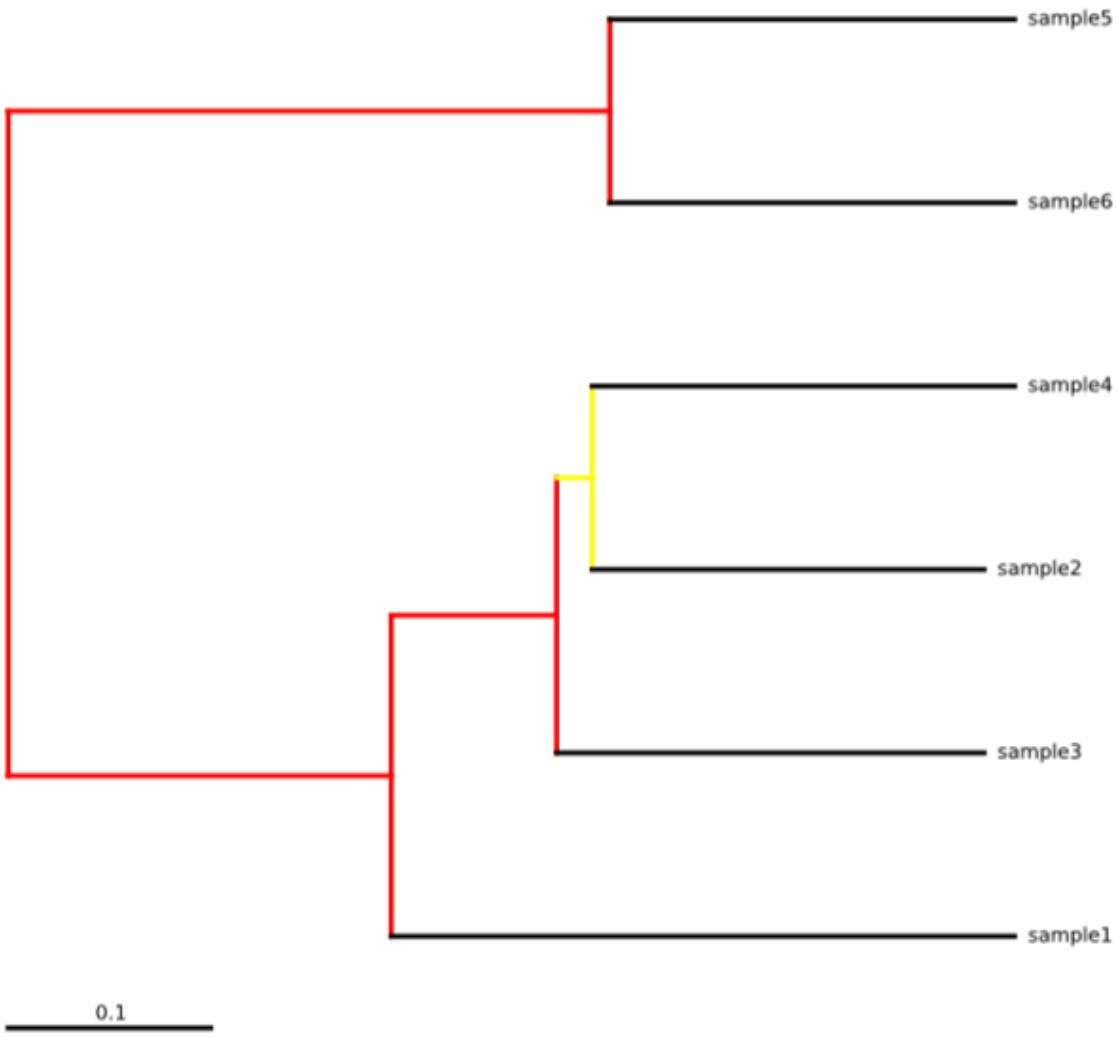


Figure 4: **Clustering of the 6 synthetic samples using IGS method**

(0.5%, 1.0%, 1.5%, and 0% - no error at all).

To show the influence of sequencing error on accuracy of the analysis, we compared the richness estimation using reads with different sequencing error rate, as shown in Figure 5. For data set without error (error rate = 0), the estimated size of the metagenome matches the real size perfectly. With increasing error rate, the size of metagenome is increasingly over-estimated. This is due

to several factors,which will be discussed below.

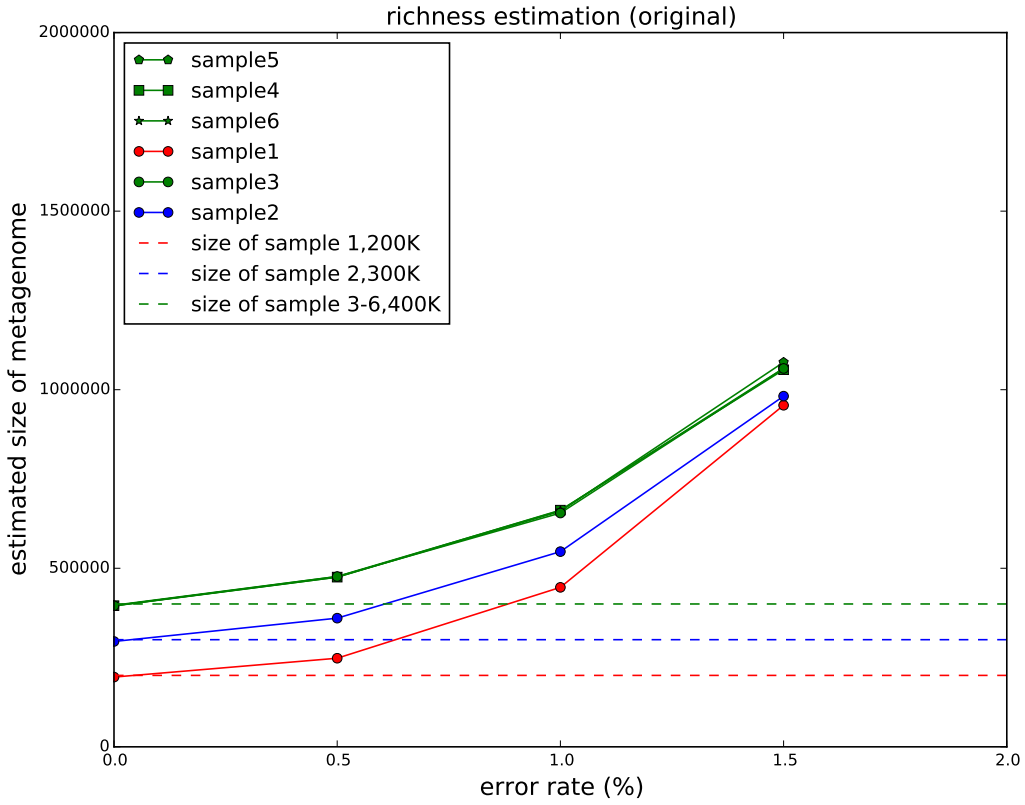


Figure 5: **Richness estimation using IGS method without adjustment.**

We also check the beta diversity analysis with different error rate and notice that the beta diversity is less prone to increasing sequencing error rate 6. We will therefore focus on alpha diversity in the discussion below.

### 2.3.1 the effect of sequencing error to the accuracy of analysis

The first factor to take into account is sequencing error. One sequencing error will generate up to k erroneous k-mers. This is the reason why it is difficult to use k-mer counting only to do diversity analysis, as a large proportion of k-mers in a reads data set are erroneous, especially for low coverage reads data. As discussed in the section about digital normalization, using median k-mer count to retrieve the coverage of a read is less prone to sequencing error, because this

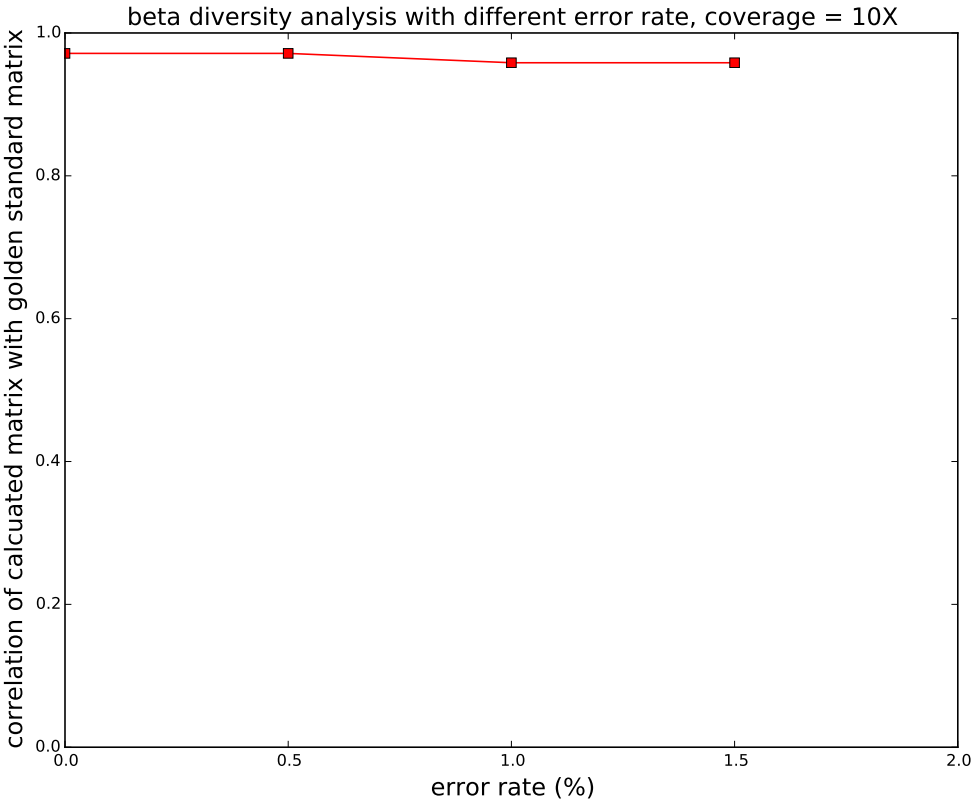


Figure 6: **Beta diversity analysis using IGS method without adjustment.**

will not always affect the median k-mer count.

Take the experiment we did previously as an example, for read length of 100bp and k as 19, one sequencing error will affect the count of 19 k-mers at most, and two sequencing errors will affect the count of at most 38 k-mers. The count of these k-mers will generally be much lower than the true count. So out of the 82 k-mers in the 100bp read, at most 38 k-mers will have incorrectly low counts. However, this will not affect the median k-mer count, which is the count of the 41th k-mer if ranked by count. However, if there are three or more errors in the read, the situation is more complicated. For 3 errors in a read, 3 to 57 k-mers will be affected by the errors to have an incorrect count as 1. The distribution of the probability about the number of affected k-mers can be acquired by a model similar to Lander-Waterman model used in genome

sequencing theory. Here we got the distribution using simulation, as shown in Figure 7. From this probability distribution, we can get the probability that 3 errors will affect more than 40 k-mers is 0.43. In this case, 3 errors will affect the median k-mer count of a read. We can also get such probability for 4 errors or more. Combining to the probability that a certain number of errors occur in a read with a specific sequencing error rate, which is easy to derive from binomial distribution, we can get the probability that the coverage of a read is incorrectly assessed as 1. Still for the example here, this probability is the probability that 3 errors occur in a read multiplied by the probability that 3 errors will affect median k-mer count, plus the probability that 4 errors occur in a read multiplied by the probability that 4 errors will affect median k-mer count,and so on.

Generally, let $P\_error(n, e, L)$ is the probability that $n$ errors occur in a read with length as $L$, with error rate as $e$ and $P\_effect(n, k, L)$ is the probability that $n$ errors in a read with length of $L$ affect median k-mer count. The probability that the coverage of a read is incorrectly assessed as 1 is

$$\sum_{n=3}^{\infty} P\_error(n, e, L) \times P\_effect(n, k, L)$$

, and by binomial distribution,

$$P\_error(n, e, L) = f(n; L, e) = Pr(X = n) = \binom{L}{n} e^n (1 - e)^{L-n}$$

Practically, when $n > 5$ and $e < 0.015$, $P\_error(n, e, L)$ is very small, we only consider number of errors in a read as 3, 4 and 5.

From the discussion above, the sequencing errors reduces the estimated coverage of some reads incorrectly to 1 and the probability this occurs to a read can be estimated. So to reduce the effect of sequencing error on this aspect, we can calculate the expected number of reads that are affected and remove those reads from the set of reads with coverage of 1 before generating list of IGS from the reads abundance distribution.

Also, we want to make sure 2 errors in a read will not affect median k-mer count, since it is more common to have 2 errors in a read practically. In this case,

$$2 \times k < \lfloor \frac{L - k + 1}{2} \rfloor,$$

we can get $k < L/5$,basically. For $L$ as 100, $k$ will be 19, which is what we choose in the testing. However, the $k$ should not be too small, or the k-mers cannot handle the diversity of information of a large data set.

Taking the sequencing error into account, we used the methods introduced above to adjust the estimation of metagenome size of the 6 synthetic samples. The estimation after adjustment is closer to real number, as shown in Figure 8..

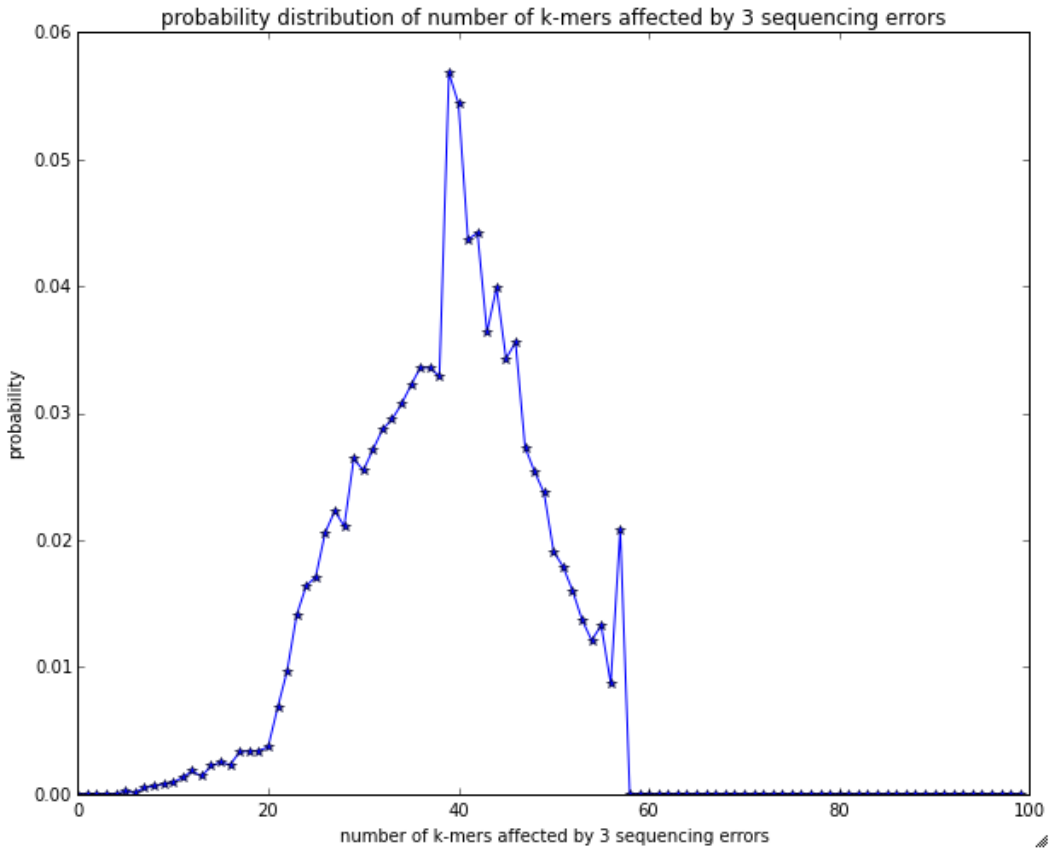


Figure 7: **Richness estimation using IGS method without adjustment.**

### 2.3.2 the effect of Bloom filter size on the accuracy of analysis

As discussed in the chapter about k-mer counting, the collision in bloom filter which we use for efficient k-mer counting will result in counting error. If the false positive rate for a specific bloom filter we use for k-mer counting is 0.1, 10% of the k-mers will have incorrect counts. When we use median k-mer count to get read coverage, such incorrect count has the effect on two aspects. On one hand, some k-mers in a read will have incorrect higher count. However, if the false positive rate is low, this will not affect median k-mer count. This shows the method of using median k-mer count to get read coverage is not only less prone to sequencing error, but also less prone to the inaccuracy characteristics of underlying data structure. One the other hand, this inaccurate count also affects the counts of those erroneous k-mers generated by sequencing error. For example, 3 errors in a read affect the count of 43 k-mers, the counts for these 43 k-mers are supposed to be 1. However, because of the collision in the Bloom

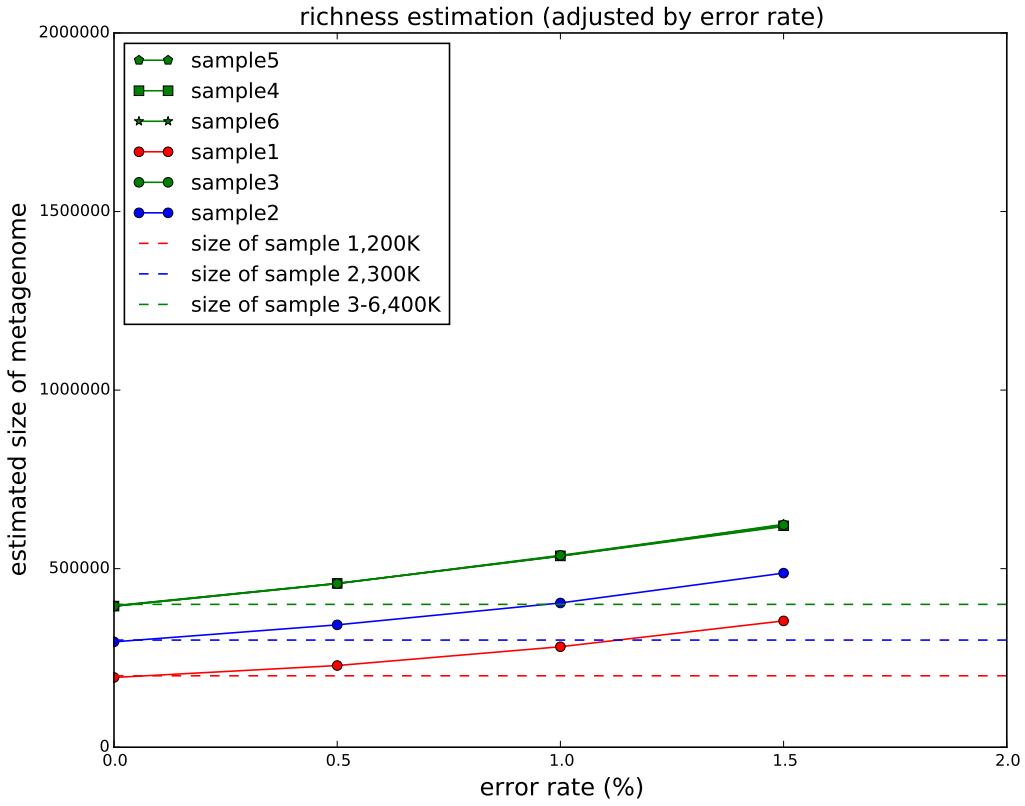


Figure 8: **Richness estimation using IGS method adjusted by sequencing error rate.**

filter and the resulting incorrect k-mer counting, if the false positive rate is 0.1, about 4 out of the 43 k-mers will have inflated count, mostly as 2. So the combined effect of sequencing error and collision in bloom filter is that some reads will have incorrect coverage as 1 and some reads will have incorrect as 2. We can get the percentage of total reads that will have such incorrect coverage, using statistical model similar to that discussed in last section. Using same example, 3 errors occur in a read, if the 3 errors affect 41-45 k-mers(with a chance of 0.20), the median k-mer count will be 2, due to the collision in bloom filter, while if he 3 errors affect more than 45 k-mers(with a chance of 0.24), the median k-mer count will be 1, purely due to sequencing errors.

We did the same experiment but also adjusted the estimation according to the false positive rate of bloom filter and got better estimation, as shown in

Figure 9.

With adjustment to estimation taking sequencing error and collision in bloom filter into account, as shown in Figure 9, the estimated genome size is closer to real number. With an error rate of 1%, a false positive rate of 0.1, and with 10X coverage data, the estimated genome size is about 20-25% more than real number. However the estimation is still increasing with higher error rate. This means there are still other factors influencing this accuracy of the estimation that we failed to take into account.

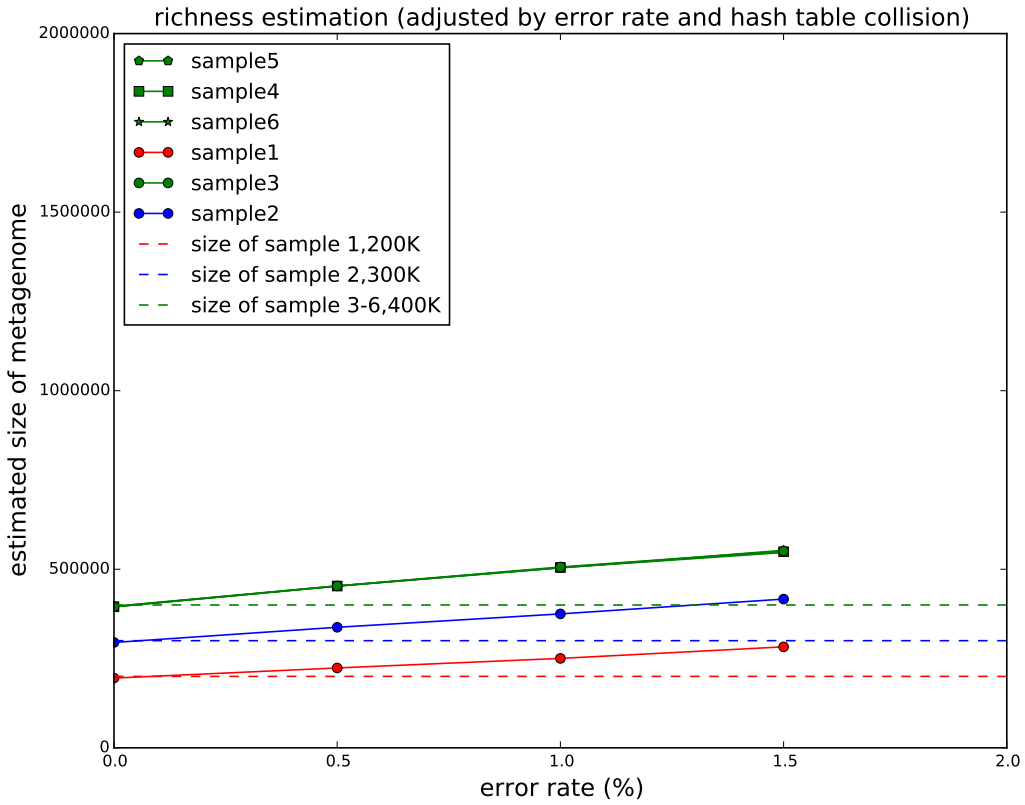


Figure 9: **Richness estimation using IGS method adjusted by sequencing error rate and false positive rate of bloom filter.**

## 2.4 the effect of sequencing depth to the accuracy of IGS method

We have shown that the IGS method can generate good result from relatively high coverage data (like 10X). It is expected that the higher the coverage of data is, the more accurate analysis we can conduct. However for many metagenomics project,especially environmental samples, it is difficult to yield high enough sequencing depth. We investigated the effect of sequencing depth on the accuracy of the IGS method.

Figure 10 shows how well the matrix calculated from a data set with variable coverage reflects the real relationship between samples. It is as expected that higher coverage data will yield a more accurate distance matrix. Note even with a coverage as low as 0.1, the correlation is 0.89. This can give us the hint about how reliable the result will be if we only use a small proportion of data from a large metagenomic data set. So the beta diversity analysis using IGS method not only is less prone to sequencing error, but also less prone to sequencing depth.

Figure 11 12 13 shows the estimated genome size from data sets with variable coverage with different error rates. It's interesting that the estimated genome size is very high with extremely low coverage. This is probably due to the limits of the statistical model in estimating the total size of information with limited observed information. After all, only a small proportion of the genomes in the sample are covered by reads.

We can see the pattern again here that higher error rate will influence the accuracy of genome size estimation, especially when the coverage is low. However for error rates from 0% to 1%, as long as the coverage is higher than 1X, the estimation of genome size starts to be stable. It is important to point that even though the absolute value of estimated genome size may be overestimated, the relationship between samples is reliable, as shown in the figures. Sample 3,4,5,6 all have 4 species, while sample 2 has 3 species, and sample 1 has 2 species. They can be separately effectively.

The estimation of genome size does not increase much with increasing coverage, even for the data set with error rate as 1%. This proves that the adjustment method discussed previously does eliminate most of the bad effect of sequencing errors. That being said, it is still beneficial to do some preprocessing to the data to reduce the error rate. If the error rate can be reduced from 1% to 0.5%, the estimated size of genome will be more accurate. This again demonstrates the importance of the streaming method doing error profile analysis discussed in chapter 4 above.

## 2.5 Compare IGS method to COMMET in beta diversity analysis

Next we test how well the matrix calculated by various methods can reflect the real relationship between samples. COMMET[**?**], the successor of Compareads [**?**], is one of few software packages for comparing metagenomes. It is based on

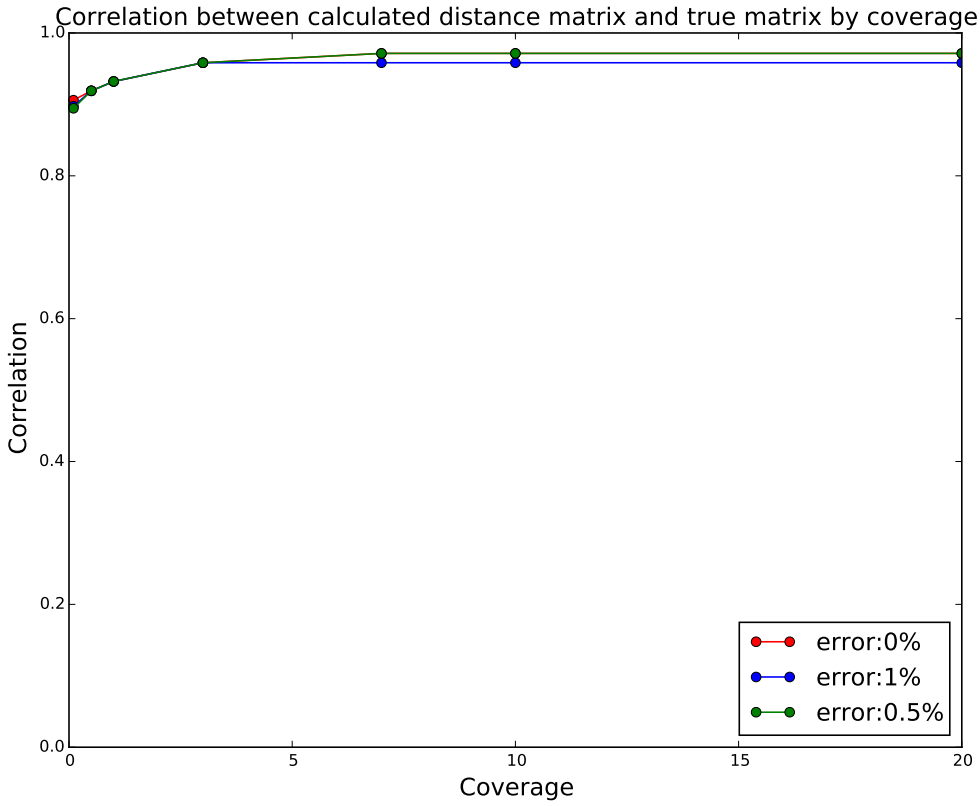


Figure 10: **Correlation between calculated distance matrix and true matrix from different data sets with different sequencing depth.**

the method of count shared reads between metagenomes. The higher percentage of reads shared by two metagenomes, the more similar the two metagenomes are inferred to be. So basically this is a straightforward abundance-based method to evaluate the similarity.

We have the simulated data set with sequencing depth as 0.1X and 10X, with sequencing error as 1% and without sequencing error was used in this experiment. This data set has the same species composition as that used in other experiment previously.

As shown in Figure 14, firstly, for all data sets, the matrix from IGS method has a higher correlation to golden standard than that from COMMET. As expected, the matrix from data sets with sequencing error has a lower correlation than that from error-free data sets. COMMET is more prone to sequencing er-

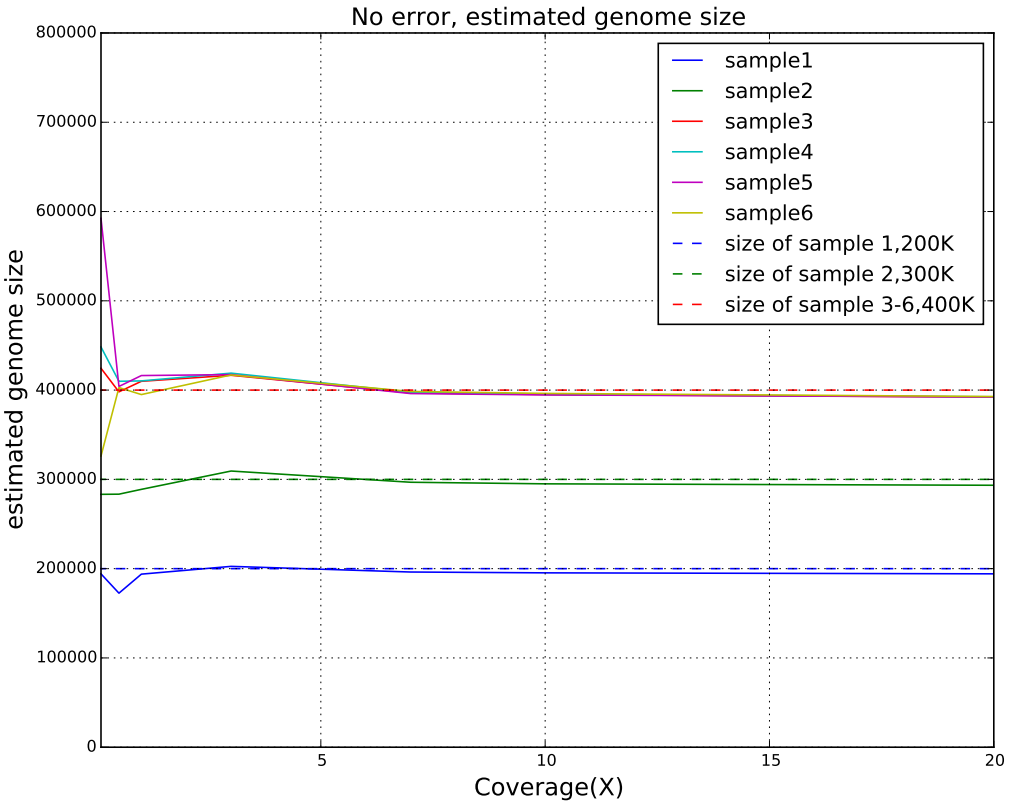


Figure 11: **Estimated genome size from data sets with variable coverage, without error.**

ror rate, compared to IGS method, for high coverage data or low coverage data. Also higher coverage will yield more accurate matrix, which is not surprising.

In the experiment below with real metagenomic dataset, we will see more evidence that the IGS method has better performance than some other metagenome comparison methods.

## 2.6 The IGS method can provide a whole framework to do alpha or Tbeta diversity, with good versatility.

From the testing using simulated data sets shown here, we are confident that our IGS method works well and can give reliable results from data sets with error and low sequencing depth.

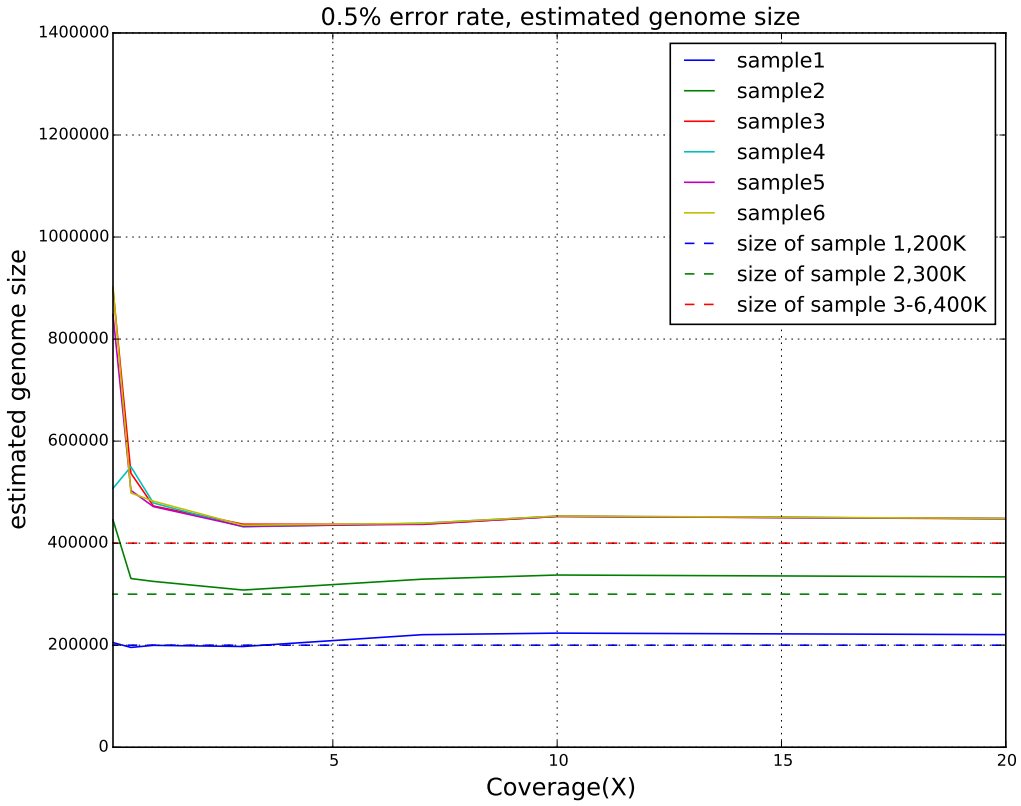


Figure 12: **Estimated genome size from data sets with variable coverage, with error rate as 0.5%.**

The IGS method can provide a whole framework to do alpha or beta diversity. Here we tested beta diversity using only Bray-Curtis metric and alpha diversity on richness only. In fact, any standard metric can be applied to the IGS-by-samples table.

The other software package to do metagenome comparison - Compareads/COMMET - based on reads overlap between samples can get a matrix reflecting the real relationship between samples. However, it is stuck with one metric, which is based on the percentage of overlapping reads between samples. This metric is like Bray-Curtis, but not exactly the same.

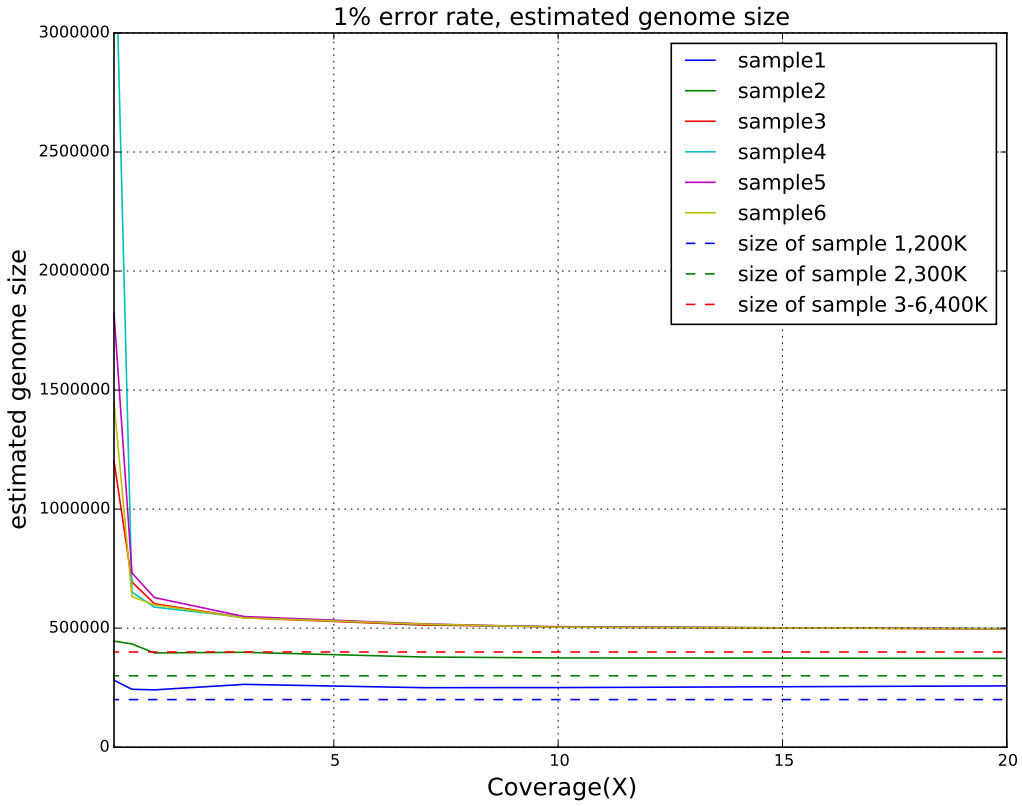


Figure 13: **Estimated genome size from data sets with variable coverage, with error rate as 1.0%.**

## 2.7 Applying IGS method to real metagenome data sets

Having shown that the IGS method delivered good results about microbial diversity from simulated synthetic data sets, we will now evaluate the novel method on several published metagenomic datasets, including samples from ocean, human microbiome and soil. For the ocean sample and human microbiome data sets, we will compare the result from IGS method with that from the original publication. For soil sample, since there is no other diversity analysis that has been conducted to these data sets, we will show the result we got from IGS and try to interpret the ecological meaning.

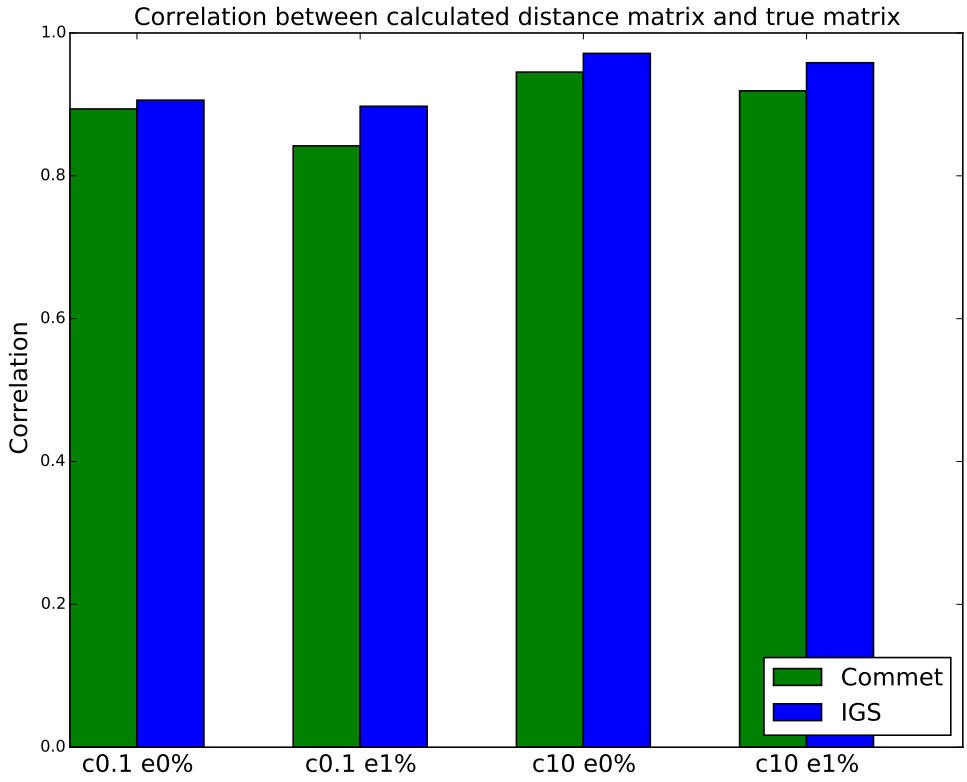


Figure 14: **Correlation between calculated distance matrix and true distance matrix from different data sets and using different methods.**

### 2.7.1 GOS data sets: Sorcerer II Global Ocean Sampling Expedition

We tested the IGS method on a well known public dataset from the Sorcerer II Global Ocean Sampling expedition. During the expedition, 44 water samples were collected from different locations across the Atlantic and Pacific Oceans and were sequenced using Sanger technology. The whole dataset is composed of 7m reads from of 44 samples. A whole metagenomic comparison of the samples was done using a sequence alignment method in the original research.

The IGS method took only several hours on MSU HPC to generate the dissimilarity matrix of the samples thanks to the scalability and distributability of the IGS based method. After clustering, Figure 15 shows that, consistently with the original study, the samples are clustered according to their geographical

origin. The group with yellow color contains samples from Tropical- Galapogas. The group with light purple color contains samples from Tropical -Open Ocean. The group with dark purple color contains samples from Sargasso. The group with green color contains samples from Temperate.

If we compare the cluster we got from IGS method with the cluster in the original study, we can see the IGS method yields a cluster more consonant with the sample origin than the method used in the original study. For example, in the original study, sample 14,21 and 22 from Tropical - Galapogas are separated from other Tropical- Galapagos samples, while in Figure 15 they are grouped together. Also, samples 00a,00b,00c,00d, all from the same location, are grouped together in our result,while in original research, sample 00a is separated from the other three samples.

Compared with the clustering generated using Compareads, our method is comparable, with some distinct differences. For example, sample 16 is clustered together with 15,17, 18, and 19 in our result. However, in the result by Compareads, sample 16 is clustered with 23 and 26, in contradiction to the geographical origin of the samples.

Next we used IGS method to analyze the alpha diversity. Figure 16 shows the rarefaction curve of IGSs of the samples. As expected, we cannot see the saturation,which means the sequencing data set is still far from deeply covered. Because the data sets for different samples have dramatically different sizes, we estimated the total number of IGSs using the Chao1 estimator with a limited number of reads in each sample (50000) to make sure the smallest data set has enough reads for comparison, as shown in Figure 17 .

We see that the richness of samples is related to the geographical origin. The samples from tropical areas have a higher richness than the samples from more northern areas. The relationship between samples is consistent with the clustering in beta diversity analysis shown above. As discussed in the section above about alpha diversity analysis to synthetic data, such number of total IGSs may be over-estimated but the relative relationship between samples on richness should be reliable. (This is not discussed in original research work on the GOS samples.)

### 2.7.2 Human Microbiome Project(HMP) metagenomics data set

We tested the IGS method on 12 HMP (Human Microbiome Project) samples from different body parts including skin, oral and vaginal. Principal component analysis (Figure 18) shows the samples are separated well by the body parts where they are collected.

Rarefaction curve and estimated number of IGSs show that the richness of samples is related to the body part where they are collected. The oral samples have higher richness than skin or vaginal samples, which is consistent with other research. [**?**]

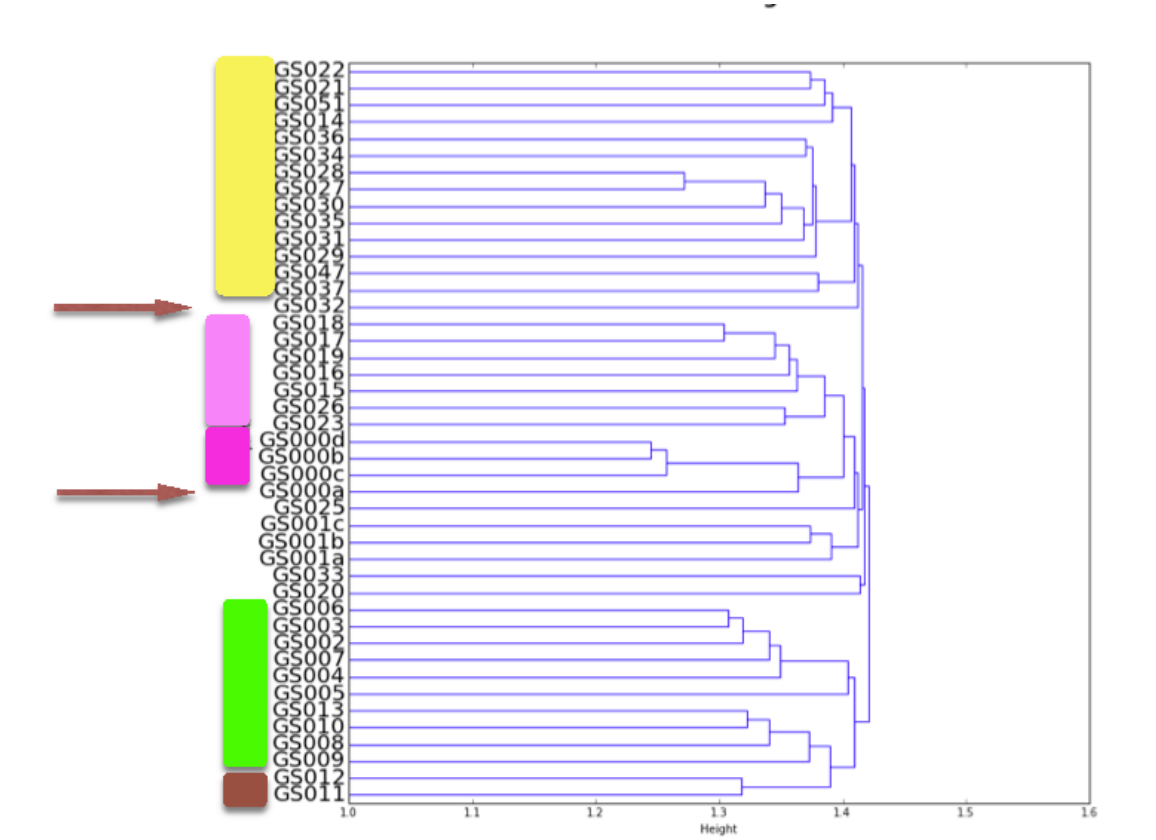


Figure 15: **Clustering of Global Ocean Sampling Expedition samples using IGS method.**

### 2.7.3 GPGC - Great Prairie Soil Metagenome Grand Challenge

Having tested the IGS method on two relatively smaller metagenomic data sets, we will now use it to analyze a larger data set from 8 soil samples collected from fields with different treatments and different locations across the great prairie region in the US. (Table 6).

As discussed above using simulated data sets, read data sets with lower sequencing coverage will reduce the accuracy of the analysis. However, as shown in Figure 10, with sequencing depth as 0.1x, the calculated distance matrix using IGS method still has a reasonably high correlation with golden standard distance matrix. So we can use subset of a large data set to acquire the diversity information, with the trade-off of lower accuracy.

For the GPGC datasets, we made a subset with 2 million reads from each sample and conducted the diversity analysis using the IGS method.

Principal coordinates analysis (Figure 20) shows the samples are separated

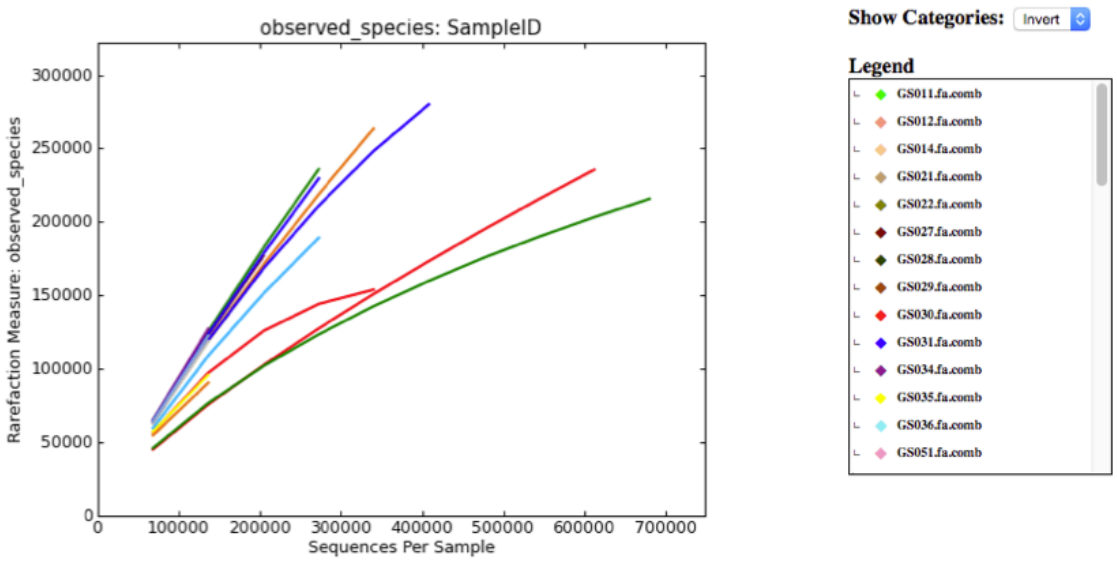


Figure 16: **Rarefaction curve of IGSs of Global Ocean Sampling Expedition samples.**

well by location where they are collected.This proves that the geographical origin plays a more important part in determining the similarity of genomic composition of samples, compared to different treatments.

Figure 21 shows the rarefaction curve and estimated number of IGSs of the samples. Basically the "corn" and "switchgrass" samples have higher richness than "restored" and "prairie" samples. This observation that cultivation increases the richness of soil is consistent with the intermediate disturbance hypothesis [**?**]. The disturbance from treatment like cultivation opens more niches and the stable communities in prairie eliminate some populations by the principle of competitive exclusion.

Its harder to explain the rank by state. The Kansas site experiences more drought stress and higher tempseratures. The Iowa and Wisconsin sites experience more cold, especially freezing conditions arresting their biology for 3-4 months. However, the freeze-thaw cycles also kill off some each cycle, which is similar to intermediate disturbance. With new growth each spring, this new growth would be the fast growers with less diverse. Why Iowa is the least diverse is still difficult to explain for now.

From the alpha diversity, we also have a rough estimation of the total size of the metagenome in Iowa soil, which is about 540G base pairs. This proves the high complexity of soil sample and we still need considerably more sequencing effort to achieve a reasonable high coverage.

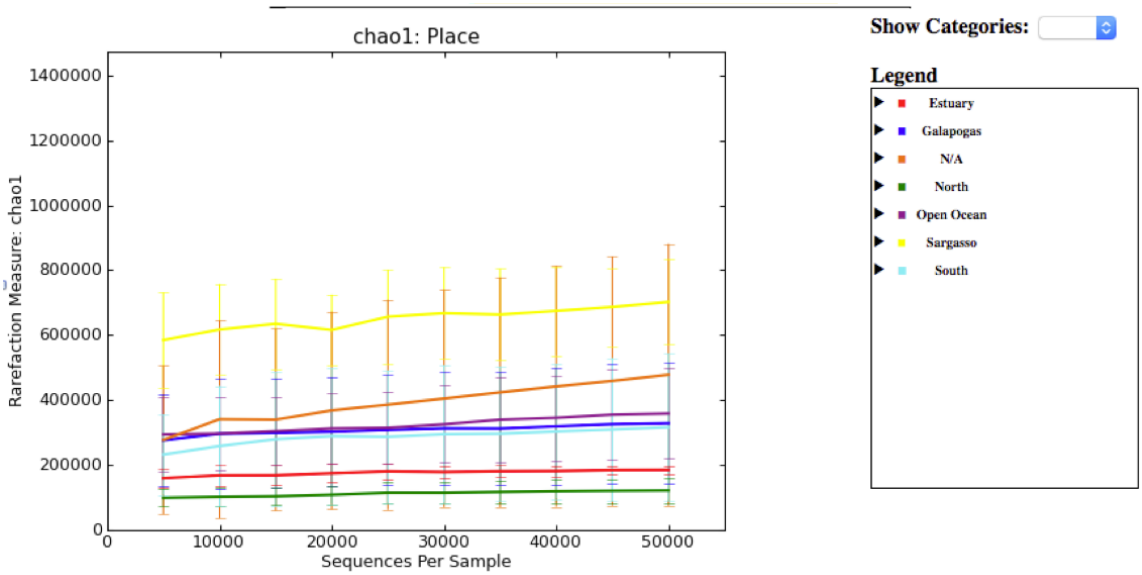


Figure 17: **Estimated number of IGSs of Global Ocean Sampling Expedition samples.**

### 2.7.4 More soil metagenomic samples

Additionally we test the IGS method on two other unpublished data sets. One is a series of soil samples collected from KBS with different treatment. Figure 22 shows the IGS method can separate the samples by treatment well.

The other data set is a series of soil samples from Amazon rainforest. The samples are separated well by the treatment. (Figure 23 ) It is also obvious that samples from forest have lower richness than prairie. (Figure 24)

# 3 Discussion

# 4 Future work

Though this manuscript demonstrated the performance of the new approach to analyze microbial diversity using whole genome shotgun sequencing data without the requirement of assembly, binning, or annotation, there is still plenty of room for improvement.

Primary questions in the process of developing the IGS based method are how many species there are in a sample or how similar the samples are with each other, mostly focusing on the quantitative aspect. Admittedly these are important questions to the microbial ecologists, but they are also curious about the qualitative aspect, such as what drives differences between samples and

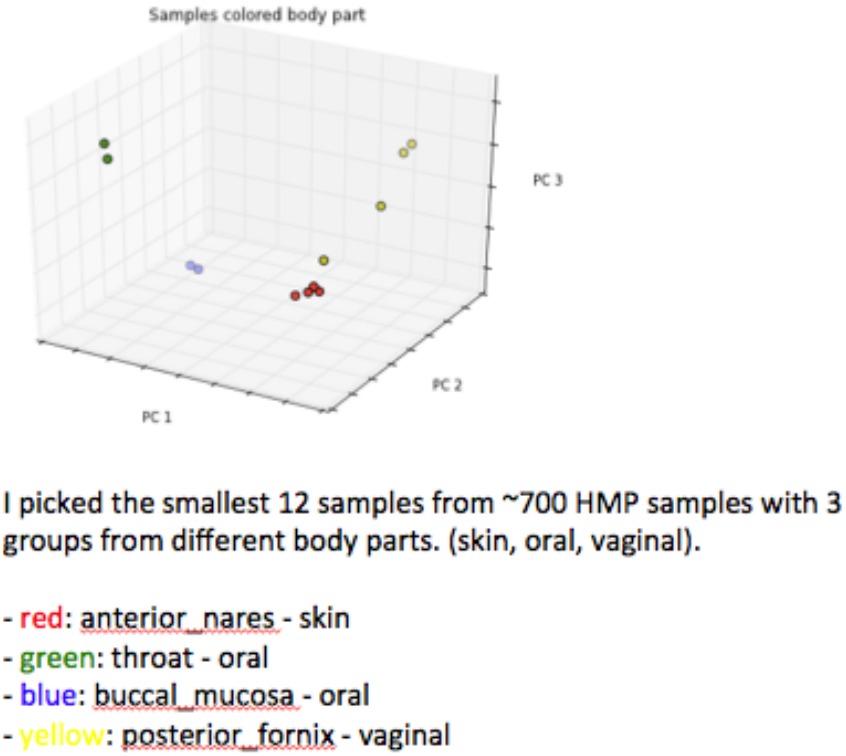


Figure 18: **Principal coordinates analysis of 12 Human Microbiome Project samples, red: anterior nares- skin, green: throat -oral, blue: buccal mucosa -oral, orange: posterior fornix -vaginal.**

eventually its functional potential[**?**]. Thus, a natural expansion of the IGS based framework will focus on answering questions mentioned above.

Now we have an efficient and scalable approach to obtain the coverage of a read in a sample, it is straightforward to extract the reads according to its coverage profile across samples so we can get a subset of reads that have specific properties, like the reads that are common in all samples. In this way we may collect these "common" reads across the samples and try to co-assemble them since now they should have higher coverage. Or we can get a subset of reads that are common in a group of samples but do not exist in another group of samples, like the samples from patients and healthy persons. These "signature" reads may offer important insights to understand what happens to the microbial community while the environment changes. Admittedly these kinds of "extraction" can be implemented using other methods like reads alignment method. However, they may not be as efficient and scalable as the IGS based method,

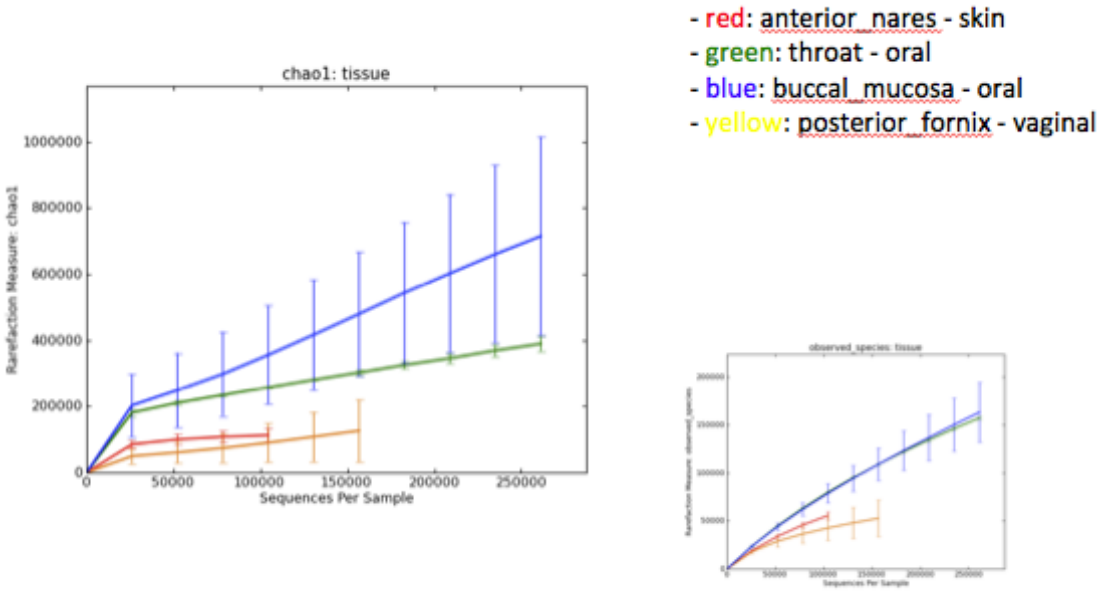


Figure 19: **Alpha diversity of 12 Human Microbiome Project samples: estimation of metagenome size of HMP samples, red: anterior nares-skin, green: throat -oral, blue: buccal mucosa -oral, orange: posterior fornix -vaginal.**

especially for extremely large metagenomic data.

One advantage of the IGS based method is that binning is not required in this procedure. Firstly, traditionally binning is used to classify contigs after read assembly effort. The similarity based binning method relies on sequence alignment, which is inefficient, even infeasible for large metagenomic data. Secondly, reference sequences are normally required for similarity based binning approach. The composition-based approach relies on the frequency profile of sequence signatures and machine learning approach on that profile, which is computationally expensive. The third approach based on coverage profile across samples was developed recently[**?**, **?**, **?**, **?**, **?**].Mostly, the coverage profile is used with the companion of composition frequency profile to classify contigs. The assumption on which the coverage profile based binning approaches are based on, that contigs with similar coverage profile across samples are more likely to be from the same microbial species, is actually similar to the assumption on which using IGS to do beta diversity is based, that the IGSs with similar coverage profile across samples are likely to be from the same microbial organism. Thus, it is promising to classify the IGSs by the coverage profile across samples. We

| sample | # of reads | size of .gz file | # of bps | ave. length |
|---|---|---|---|---|
| iowa corn | 1,514,290,825 | 46G | 144,202,427,079 | 95.2 |
| iowa prairie | 2,597,093,273 | 74G | 226,815,059,143 | 87.3 |
| kansas_corn | 2,029,883,371 | 66G | 206,933,829,048 | 101.9 |
| kansas_prairie | 4,987,358,734 | 145G | 499,387,223,498 | 100.3 |
| wisconsin_corn | 1,616,440,116 | 51G | 162,257,698,471 | 100.4 |
| wisconsin_prairie | 1,653,557,590 | 53G | 166,467,901,724 | 100.7 |
| wisconsin_restored | 226,830,595 | 11G | 34,241,520,930 | 151.0 |
| wisconsin_switchgrass | 310,966,735 | 13G | 40,259,619,921 | 129.5 |

Table 6: **GPGC data sets**

have already overcome the challenge of retrieving the coverage profile efficiently based on the probabilistic data structure, while in those coverage profile based binning approaches the coverage profile is normally retrieved by assembly of contigs and mapping reads back to contigs, which both require higher coverage reads to do assembly and are computationally expensive.

There are two obstacles to overcome in this coverage profile based IGS binning approach. First, with relatively small number of samples, the resolution will be limited, since the total number of different coverage profiles will be limited. This is probably the reason why most of those coverage profile based contig binning methods have to integrate composition profile information also. Second, on the other hand, if there are a large number of samples, there will be too many different coverage profiles. We can use more sophisticated approach to classify the coverage profiles to reduce the number of bins, as in those coverage profile based contigs binning methods. Another approach worthy of note is that there is a method termed partitioning developed in our group as a divide and conquer approach to scale metagenome assembly. It can be considered as a binning approach also, where the reads in the same partition are more likely to originate from the same microbial organism. We can try to integrate the partitioning and IGS coverage profile to improve the accuracy of the binning. In summary, this will be one of the first attempts to do reads binning. After the IGS/reads binning, we expect to do better assembly and annotation and gain more knowledge about the function and phylogenetic information.

We have shown that after adjustment according to sequencing error and collision rate of the bloom filter, the estimated size of metagenome is close to real number for synthetic data sets. Howevere, the difference between the estimation and real number is still increasing with higher error rate, which means there are other factors that affect the accuracy of estimation. This is worthy of further investigation. The size estimation of metagenome is extremely important in metagenomic data analysis and it is closely related to the estimation of sequencing depth or how much more effort is required to gain enough sequencing depth. As shown in the results, we are confident that the relative relationship between the richness of different samples is reliable from the IGS based alpha diversity analysis. How accurate the absolute value of the richness or the size of

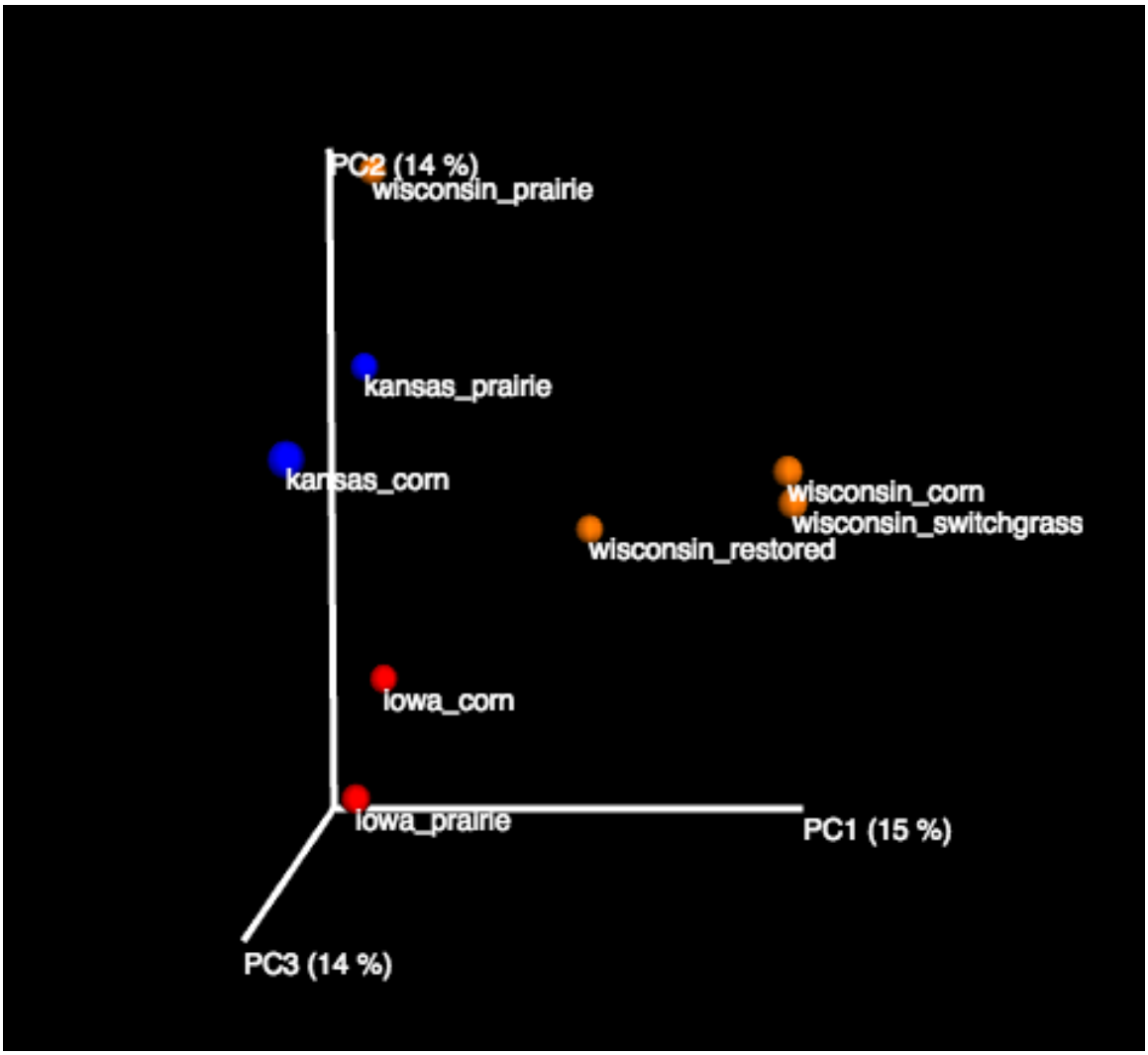


Figure 20: **Principal coordinates analysis of 8 Great Prairie Soil Metagenome Grand Challenge (GPGC) samples.**

metagenome in a real data is requires further investigation and new statistical model may be needed to adapt to the abundance distribution of IGSs. Furthermore, any information about the richness of a sample is beneficial to the optimal choice of parameter for digital normalization.

Our efforts to examine the effect of sequencing depth on the accuracy of beta diversity reveals that using a relatively small subset of the whole data set may get reasonably good result showing the separation of samples after clustering or ordination. However, how good the separation is seems to be

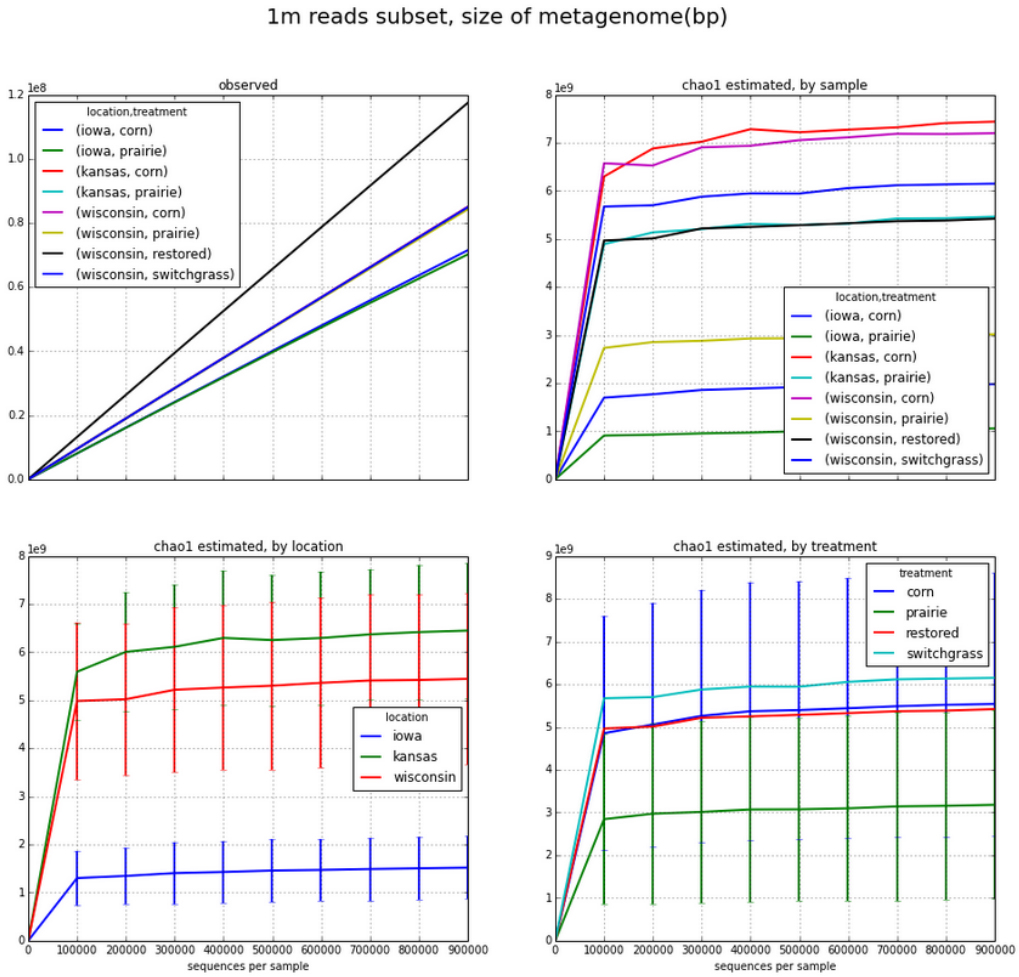


Figure 21: **Alpha diversity analysis of 8 GPGC samples. Upper left, rarefaction curve of IGSs. Upper right, estimated number of IGSs in different samples. Lower left, estimated number of IGSs in samples grouped by location (Iowa, Kansas and wisconsin). Lower right, estimated number of IGSs in samples grouped by treatment (corn, prairie, restored, switchgrass).**

related to the characteristics of samples and cannot be determined easily before starting the analysis. Thus, a potential approach should estimate beta analysis in an iterative way. We already know that using more data will benefit more accurate analysis or better separation for the purpose of comparing metagenomic samples. In such iteration procedure, pattern of separation can be monitored as more reads are loaded into the analysis and the procedure can be stopped

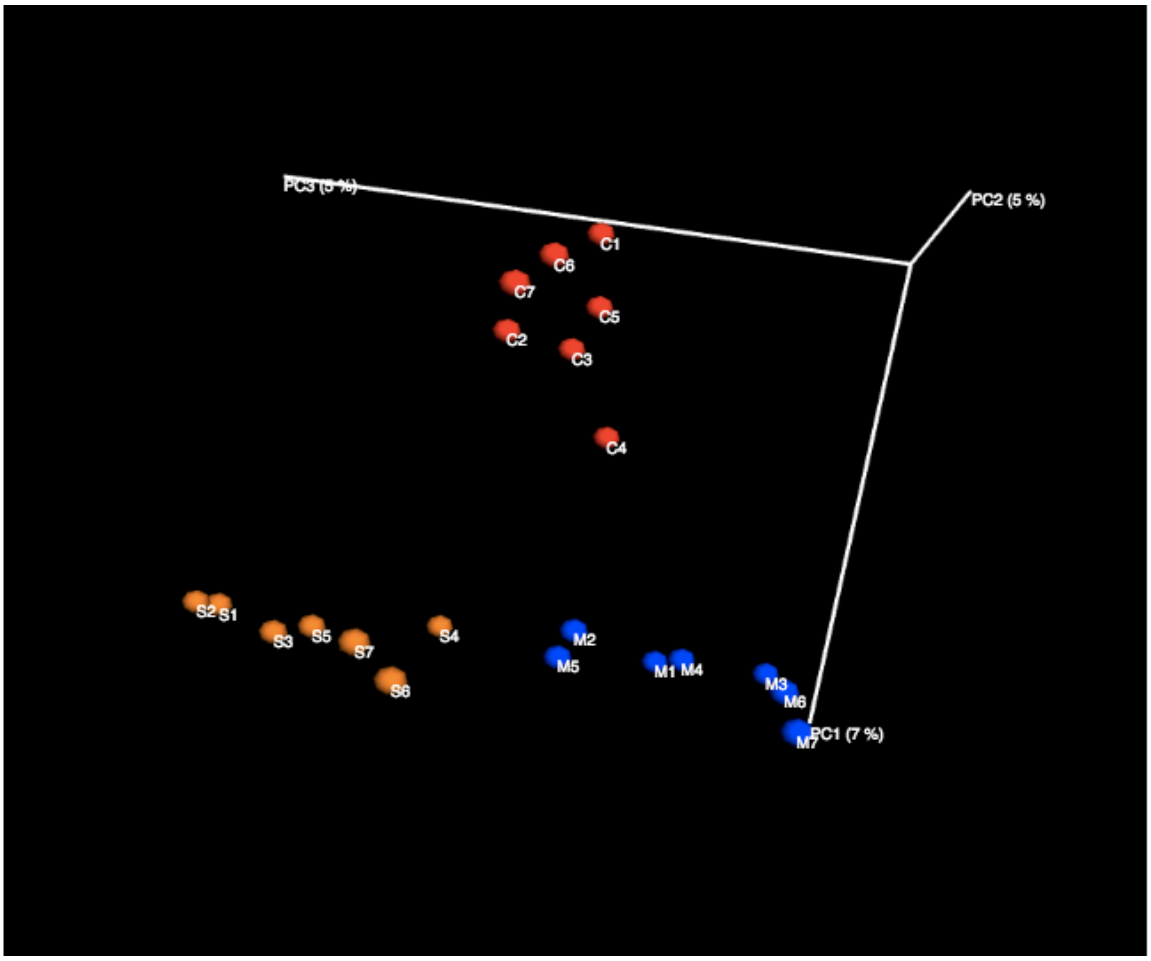


Figure 22: **Principal coordinates analysis of soil samples with different treatments collected from Kellogg Biological Station(KBS). Red, corn. Blue, miscanthus. Brown, switchgrass.**

as long as the pattern of the separation of samples is significant enough. This way, we may save lots of computational cost and still have enough information about the relationship between samples.

## 4.1 Concluding thoughts

Diversity analysis is a key part of the microbial ecology research, like of macroorganism ecology. However due to the obscure definition of the term "species" in microbial ecology, we can virtually never measure the diversity of species directly, rather we use other taxonomic concepts like operational taxonomic unit

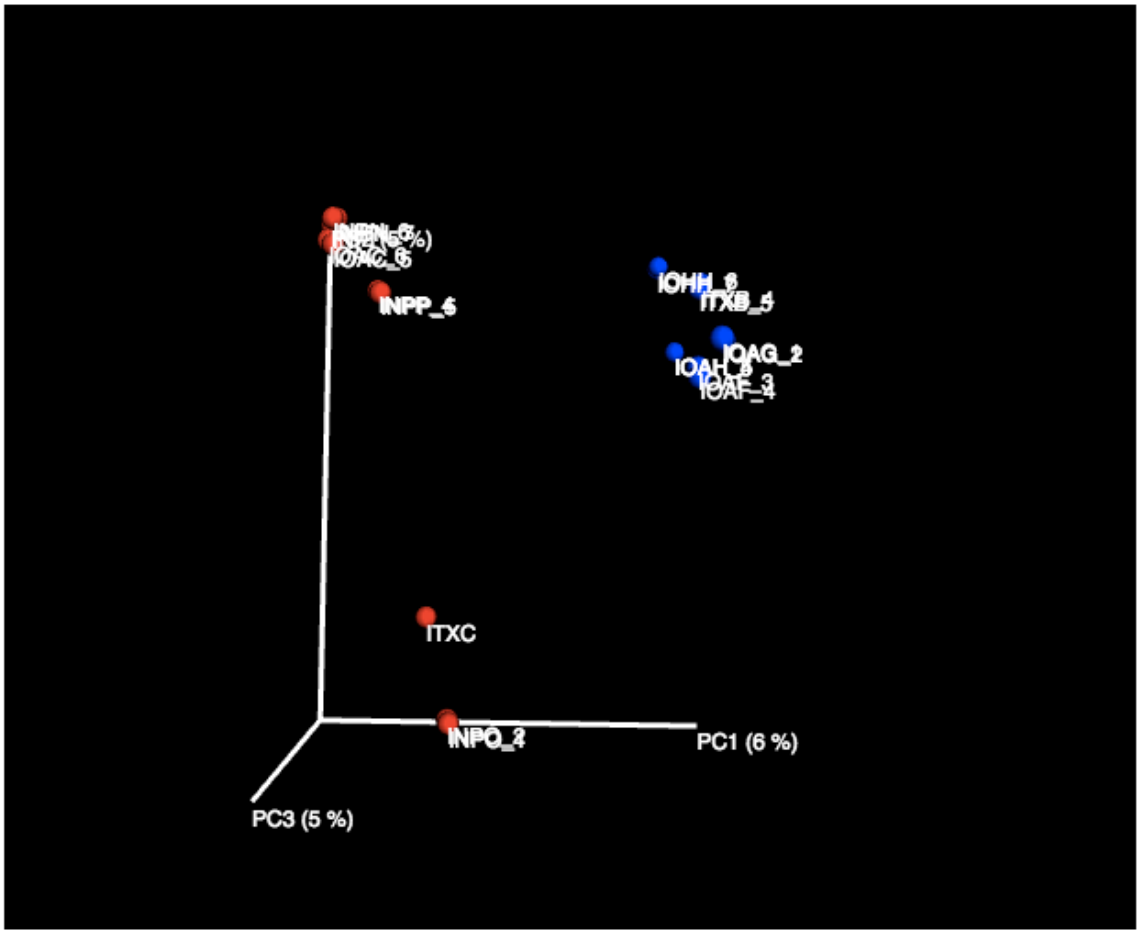


Figure 23: **Principal coordinates analysis of soil samples collected from Amazon rainforest. Red, forest samples. Blue, prairie samples.**

(OTU) to evaluate the diversity of microbial community, instead of species. 16S rRNA sequencing reads may be classified into different OTUs. Shotgun whole genome sequencing reads can also be classified into OTUs. But most, if not all the existing methods based on the concept of OTU rely heavily on preprocessing of original reads data in some way like assembly or external information like reference sequences for annotation. Both of the prerequisites are not satisfied for many metagenomic projects. For metagenomic data set with low sequencing coverage, the assembly process skews the analysis by including primarily the most abundant organisms. Sequences that are rare are not assembled into contigs and are therefore not included in the contig analysis It is common that only a small proportion of reads can be used in assembly especially in a complex environmental sample. [**?**] The reference sequence database is far from completion

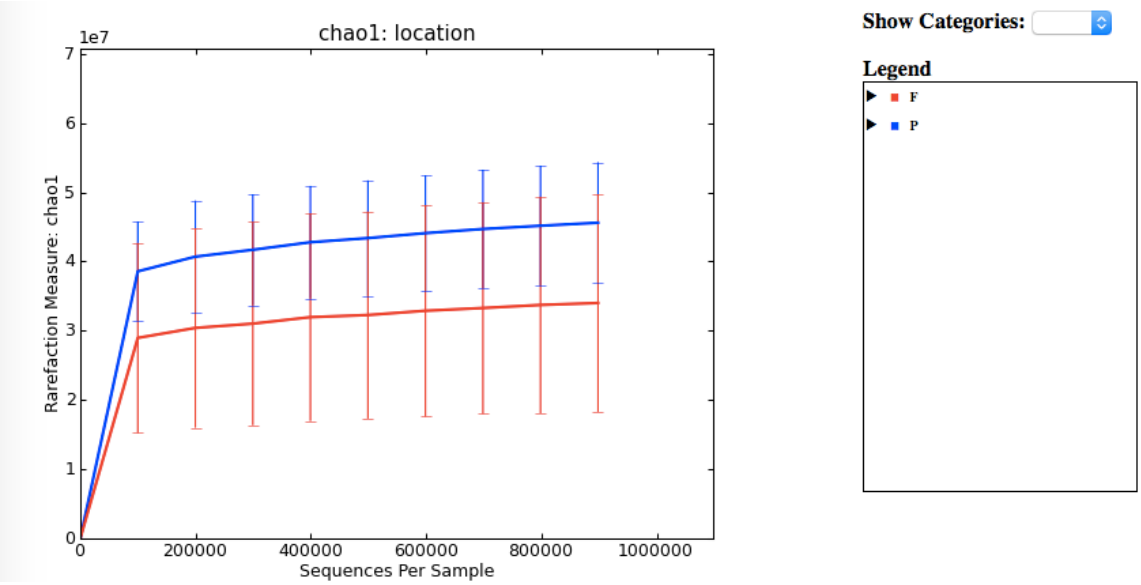


Figure 24: **Estimated number of IGSs in metagenomic data from soil samples collected from Amazon rainforest. Grouped by treatment. Red, forest samples. Blue, prairie samples.**

especially for microbes in environmental samples from soil or sea water. The paucity of reference databases affects the ability to identify functional capacity of microbes, when traits cannot be identified. Thus, applying these methods can only obtain an incomplete diversity of the microbial community in the metagenomic sequencing data set.

The IGS based method discussed in this dissertation offers a novel framework that overcomes the limitations of assembly, binning, or annotation without the requirement of reference sequence database. It can take advantage of all the information in the metagenomic reads data and gain a full picture of the diversity of the microbial community. Importantly, this is a new framework with the concept of IGS instead of OTU as the taxonomic unit to analyze microbial diversity. Thus, this framework can be used to perform all possible diversity analysis that OTU-based framework can do. Moreover, this is a more thorough approach than many other methods developed to solve only specific problems in the field of diversity analysis. For example, there are several methods developed to estimate the species richness in a metagenomic sample [**?**]. Our IGS based framework cannot only estimate the species richness or size of metagenome as shown in the section above, but also it can estimate the evenness or species abundance distribution of a metagenomic sample, which is also an important aspect of alpha diversity analysis. For beta diversity or compositional similarity analysis between metagenomic samples, there are several methods developed to compare metagenomic samples based on reads mapping or counting shared reads

[?]. However they only estimate abundance-based similarity, similar to the Bray-Curtis indices used in the experiment discussed in the section above. It should be noted that the IGS-based framework can also be used to estimate incidence-based similarity, which cannot be estimated using other existing approaches.

Besides the potential for a broad application of the IGS based framework, it is also efficient and highly scalable to handle extremely large metagenomic sequencing data sets. We have discussed the efficiency of the novel k-mer counting method and the following method of digital normalization, with the ability to retrieve the coverage of a read accurately and efficiently. We also performed a thorough analysis to examine the effect of the size of used data structure to the accuracy. We can take advantage of the probabilistic characteristics of the data structure to make a trade-off between expected analysis accuracy and expected usage of computational power. In this way, we make the analysis highly scalable to keep pace with the increasing size of metagenomic sequencing data.

In addition, we examined the effect of sequencing depth to the accuracy of estimating microbial diversity. It was expected that using more number of reads, that is, a data set with higher sequencing depth increases the accuracy of diversity estimation. For similarity analysis between samples (beta diversity), a data set with relatively low sequencing depth can still get decent results, as shown in the experiment with synthetic data and real soil data sets. However, for alpha diversity such as richness estimation, use of a data set with lower sequencing depth results in the diversity estimation more distant from the real number. Although the absolute value of such species richness of a sample is not accurate, the relative comparison of species richness between samples is less prone to smaller reads data with lower sequencing depth. These results suggested that for a specific purpose, only a subset of the large metagenomic reads data can be enough to achieve reasonably satisfying result. Under certain circumstance, this feature is quite helpful and can reduce the computational expense dramatically.

# 5 Conclusion

We established a series of approaches to enable scalable and effective investigation of microbial diversity using whole-genome shotgun metagenomic data. Firstly a k-mer counting package - khmer was developed to enable fast and memory efficient k-mer-based analysis of sequencing data sets[?, ?]. Khmer relies on Count-Min Sketch, a probabilistic data structure used to store the frequency of distinct elements efficiently. Unlike other data structures used for k-mer counting, such as hash tables, suffix arrays, and trie structure, the Count-Min Sketch has significantly low memory usage for sparse data sets with trade-off with counting false positive. We conducted extensive analysis on the performance of the counting algorithm and benchmark to compare the performance of the khmer to other k-mer counting packages. The initial motivation of developing khmer was to count the k-mers in metagenomes for diversity analysis. Now khmer has been widely used for many other purposes, from enabling large scale

de novo metagenome assembly to sequencing error detection and correction.

Based on the efficient k-mer counting package khmer, especially with the ability to do online counting and retrieval entirely in memory, we developed digital normalization[**?**], "a single-pass computational algorithm that systematizes coverage in shotgun sequencing data sets, thereby decreasing sampling variation, discarding redundant data, and removing the majority of errors." Digital normalization can reduce the computational expense of downstream analysis such as assembly dramatically because after the normalization of sampling variance, redundant reads are discarded as well as the errors in them. The algorithm of digital normalization has been used by many research groups to facilitate their analysis and has been implemented in different tools like Trinity and Illumina's TruSeq pipeline. Like digital normalization, based on the same approach to estimate sequencing depth without a reference assembly, a streaming approach to analyze and trim sequencing errors in short reads datasets was developed[**?**]. The approach offers a general framework for streaming sequence analysis and could be used for error correction and variant calling. Moreover, the approach can be applied generically to data sets with variable sequencing coverage, such as metagenomes especially.

Further more, by integrating efficient k-mer counting and a novel de Bruijn graph mapping method based on digital normalization we developed a novel approach to allow for scalable diversity analysis of large, complex metagenomes. A novel concept - IGS (informative genomic segment) is proposed to represent the unique information in a metagenomics data set. The IGSs can be used as a complement of OTUs to be the cornerstone for diversity analysis of whole shotgun metagenomics data sets. The abundance of IGSs in different samples can be retrieved by mapping the reads to de Bruijn graphs. In this procedure, not like many other microbial diversity analysis methods, assembly or binning is not required any more. This method was evaluated on multiple metagenomes from a variety of environments (e.g., human body part, seawater, soil). Given the velocity in growth of sequencing data, this method is promising for analyzing highly diverse samples with relatively low computational requirements. Further, as the method does not depend on reference genomes, it also provides opportunities to tackle the large amounts of unknown "dark matter" we find in metagenomic datasets.

# 6 Methods

## 6.1 Code availability

The algorithms of the IGS based diversity analysis are implemented in the khmer software package, written in C++ and Python, available at github.com/ged-lab/khmer/. khmer also relies on the screed package for loading sequences, available at
github.com/ged-lab/screed/. khmer and screed are Copyright (c) 2010 Michigan State University, and are free software available for distribution, modification,

and redistribution under the BSD license.

The code and detailed instruction used to generate all the results in this chapter is available at http://github.com/ged-lab/2013-diversity/.

## 6.2 Simulated data sets

### 6.2.1 Four simulated reads data sets with different species abundance distribution

### 6.2.2 Simulated sequencing reads of e.coli

Here we simulated 4 sequencing reads data sets with read length as 100bp of e.coli with different sequencing depth(50x and 150x) and different sequencing error rate(1%,2% and 0%). Table 7

Table 7: **Simulated sequencing reads data sets of e.coli**

| sample | coverage | error rate |
|---|---|---|
| A | 150 | 0.01 |
| B | 50 | 0.01 |
| C | 50 | 0.01 |
| D | 50 | 0.02 |

Table 8: GPGC Data sets

| sample | # of reads | size of .gz file | # of bps | ave. length |
|---|---|---|---|---|
| iowa corn | 1514290825 | 46G | 144202427079 | 95.2 |
| iowa prairie | 2597093273 | 74G | 226815059143 | 87.3 |
| kansas_corn | 2029883371 | 66G | 206933829048 | 101.9 |
| kansas_prairie | 0 | 145G | 0 | 0 |
| wisconsin_corn | 1616440116 | 51G | 162257698471 | 100.4 |
| wisconsin_prairie | 1653557590 | 53G | 166467901724 | 100.7 |
| wisconsin_restored | 226830595 | 11G | 34241520930 | 151.0 |
| wisconsin_switchgrass | 310966735 | 13G | 40259619921 | 129.5 |

## 6.3 Using IGS to analyze alpha diversity

Basically the abundance distribution of IGSs with different coverage in a sample data set can be acquired using the method shown above.

Suppose from a reads data set, the coverage distribution of reads is as shown in Table 9.

There are 69 reads with coverage as 3, 96 reads with coverage as 4, and so on. We transform this coverage distribution of reads into abundance distribution of IGSs, as shown in Table 10.

For example, there are 23 IGSs with abundance 3. This is calculated by dividing total number of reads with coverage as 3, which is 69, by the coverage

| coverage | number of reads |
|---|---|
| 3 | 69 |
| 4 | 96 |
| 5 | 125 |
| 6 | 150 |
| ... | ... |

Table 9: **Reads coverage distribution.**

| abundance | number of IGS |
|---|---|
| 3 | 23 |
| 4 | 24 |
| 5 | 25 |
| 6 | 25 |
| ... | ... |

Table 10: **IGS abundance distribution.**

3. Similarly there are 24 IGSs with abundance 4. If we draw an analogy between IGSs and OTUs, this is like saying there are 23 different OTUs with 3 reads mapped to each OTU, and 24 different OTUs with 4 reads mapped to each OTU, and so on.

Next all the different IGSs and the corresponding abundances can be listed, as shown in Table 11.

| IGS ID | abundance |
|---|---|
| 1 | 3 |
| 1 | 3 |
| 1 | 3 |
| ... | ... |
| 23 | 3 |
| 24 | 4 |
| 25 | 4 |
| ... | ... |
| 47 | 4 |
| 48 | 5 |
| ... | ... |

Table 11: **Listing IGSs with the corresponding abundances.**

This list is the counterpart of an OTU table in OTU based diversity analysis. With such table at hand, existing statistical methods and software packages can be directly used to investigate the alpha diversity.

## 6.4 Using IGS to analyze beta diversity

As with alpha diversity analysis, the OTU table is also a foundation for beta diversity analysis. As long as we get a reliable OTU table, there are existing pipelines to perform the beta diversity analysis.

A typical OTU table across different samples is like this, which is also called a samples-by-OTU data matrix, as shown in Table 12.

| OTU ID | Sample A | Sample B | Sample C |
|---|---|---|---|
| OTU1 | 3 | 4 | 2 |
| OTU2 | 2 | 5 | 0 |
| OUT3 | 3 | 1 | 4 |
| ... | ... | ... | ... |

Table 12: **Samples-by-OTU matrix.**

Like a OTU table, we hope to have the IGS table for the IGSs, as Table 13.

| IGS ID | Sample A | Sample B | Sample C |
|---|---|---|---|
| IGS1 | 2 | 3 | 2 |
| IGS2 | 2 | 4 | 0 |
| IGS3 | 3 | 3 | 4 |
| ... | ... | ... | ... |

Table 13: **Samples-by-IGS matrix.**

So now the problem is how we can generate a samples-by-IGS matrix as the counterpart of samples-by-OTU matrix so many of the existing tools/methods used for OTU-based diversity analysis can be borrowed for this kind of IGS-based analysis, just as what is shown above for alpha diversity analysis.

Firstly, using the same approach to get the coverage of a read in the sample data set where it is from (Figure 25-A), we can get the coverage of a read from sample A dataset in sample B dataset (Figure 25-B). We still use the median k-mer count to represent the coverage of a read.

For a data set with several samples to analyze, firstly we can get the coverage of a read across different samples and get a read coverage profile like "4:6:2", as shown in Figure 26(A). For all the reads in the samples we can get such read coverage profiles, as shown in Figure 26(B). We have already known that "contigs with similar coverage profiles are likely to have originated from the same microbial population"[**?**]. Several new binning methods based on coverage profiles have been developed based on such assumption[**?**, **?**, **?**, **?**, **?**]. Thus, here we can assume that reads with similar coverage profiles are likely to have originated from the same genomic region. Actually it is safer to say reads with different coverage profiles are not likely to have originated from the same genomic region. The next step is to group those reads with same coverage

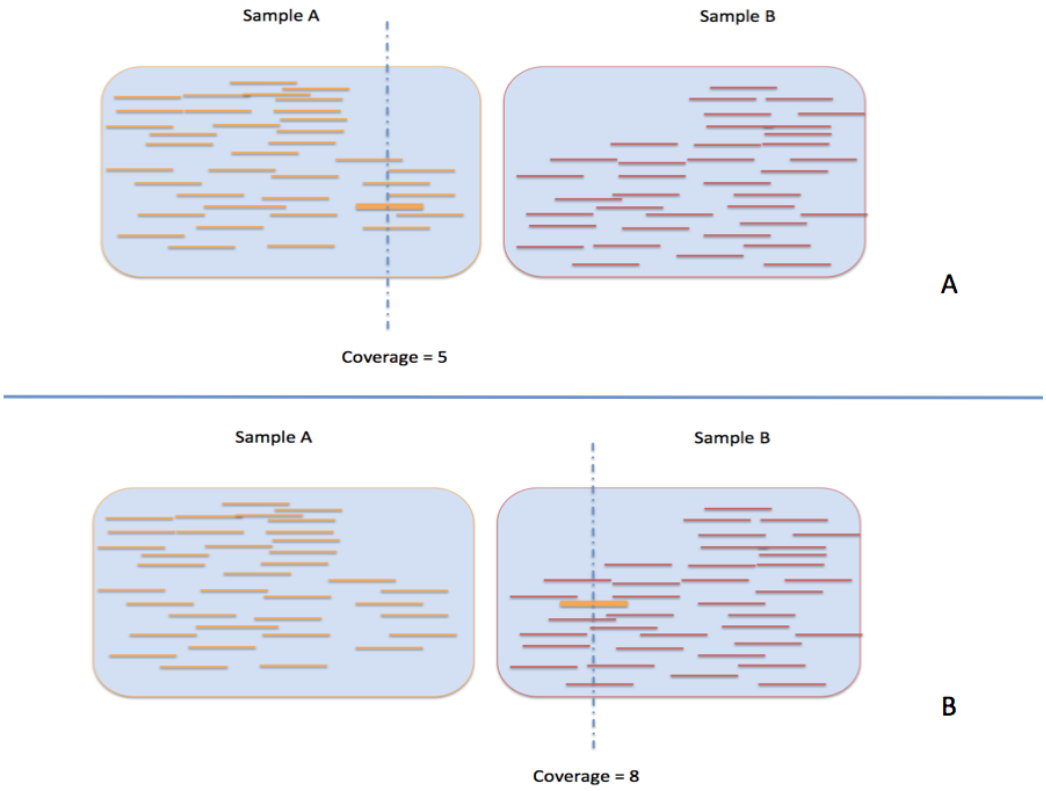


Figure 25: **Get the coverage of a read in samples. A read in sample A has the coverage of 5 in sample A, has the coverage of 8 in sample B.**

profiles together into "super bin" in different samples, as shown in Figure 26(C). The reads in each "super bin" may not be from the same species, however, they should be from the species that have same abundance profile across samples. In the example shown in the figure, the 6 reads from sample A, the 4 reads from sample B and the 4 reads from sample C all have the same coverage profile as "3:2:2". Actually, the numbers of reads from different samples with same coverage profile have similar ratio to the numbers in the profile, like "6:4:4" versus "3:2:2" in the example above. (Experiments using simulated data are not shown here.)

Next we can use the approach similar to the one used for alpha diversity analysis to estimate the size of the genomic region each "super bin" covers, represented by the number of IGSs. Still for the example in Figure 26(C), 6 reads from sample A grouped into the first "super bin" have a coverage of 3 in

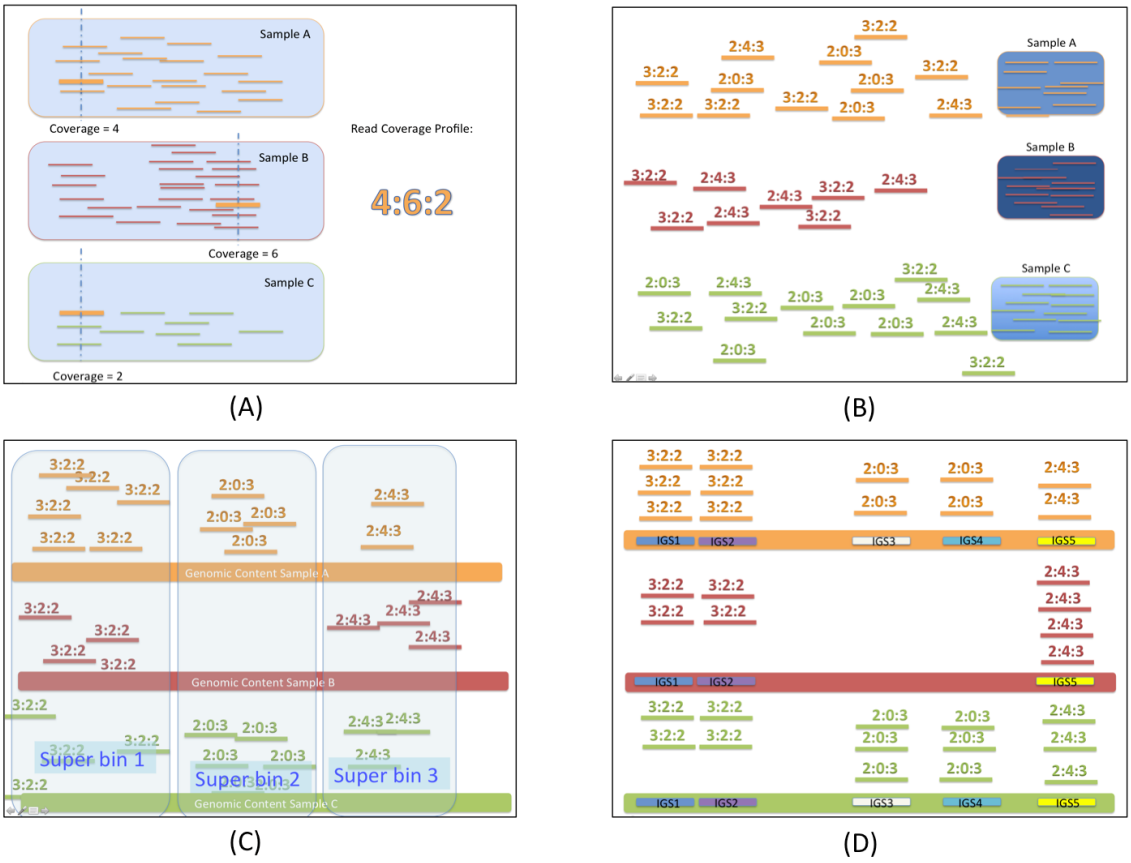


Figure 26: **From read coverage profile to IGS. (A): Get the coverage profile of one read. (B): Get the coverage profiles of all the reads in 3 samples. (C): Group the reads with same coverage profiles into "super bin". (D): Calculate the number of IGSs in each "super bin".**

sample A, where they originate from. The number of IGSs those 6 reads cover can be calculated as 6/3, which is 2. Thus 2 IGSs have an abundance profile as "3:2:2" across the samples. Similarly there are 2 IGSs with abundance profile as "2:0:3" and 1 IGS with abundance profile as "2:4:3", as shown in Figure 26(D).

List those IGSs and the corresponding abundance profiles across samples, we can have the samples-by-IGS matrix as shown in Table 14.

With such samples-by-IGS matrix, similarity matrix between samples can be calculated using different similarity indices, like Bray-Curtis. Next clustering and ordination methods can be applied to better interpret the relationship between samples.

| IGS | Sample A | Sample B | Sample C |
|---|---|---|---|
| IGS1 | 3 | 2 | 2 |
| IGS2 | 3 | 2 | 2 |
| IGS3 | 2 | 0 | 3 |
| IGS4 | 2 | 0 | 3 |
| IGS5 | 2 | 4 | 3 |

Table 14: **Samples-by-IGS matrix from example.**

# 7 Acknowledgments