Contents

1 Introduction

In almost all 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. Traditionally especially for amplicon metagenomics data set, OTUs(Operational Taxonomic Units) based on 16S rRNA genes are used as the basic units for diversity analysis. OTUs can be good replacement 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.

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 with large size because of the high diversity. For instance, there may be 4 petabase pairs of DNA in a gram of soilZarraonaindia:2013aa. 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.

Here we propose a novel concept - IGS (informative genomic segment) and use IGS as a replacement of OTUs to be the cornerstone for diversity analysis of whole shotgun metagenomics data sets. IGSs represent the unique information in a metagenomics data set and the abundance of IGSs in different samples can be retrieved by the reads coverage through an efficient k-mer counting method. This samples-by-IGS abundance data matrix is a promising replacement of samples-by-OTU data matrix used in 16S rRNA based analysis and all existing statistical methods can be borrowed to work on the samples-by-IGS data matrix to investigate the diversity. We applied the IGS-based method to several simulated data sets and a real data set - Global Ocean Sampling Expedition (GOS) to do beta-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 The concept of IGS(informational genomic segment)

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.

IGS may be a good concept in whole genome shotgun metagenomic diversity analysis, especially while facing large amount of "dark matters", unknown species. We don't care about species, we only care about how much information there is in the sample.

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 used to estimate total number of IGSs in the sample. Rarefaction curve based on number of IGSs can also be generated.

For beta diversity, here we will generate a samples-by-IGS data matrix, as a replacement of samples-by-OTU data matrix in 16s 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/disimilarity/distance between samples and do further analysis like clustering and ordination. QIIME and Mother can do this kind of jobs pretty well.

Using median k-mer frequency can decrease the influence of sequencing error, but can not eliminate the influence of errors. This can cause some problems in the following analysis, which will be discussed in details.

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

Median k-mer abundance can represent sequencing depth of a read(cite diginorm). 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.

The Figure 1a shows the abundance distribution of reads from 4 simulated sequencing data sets with different sequencing depth - 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 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 the figure in upper position to show the number of IGSs and their respective reads coverage, as shown in figure in lower position. We sum up the numbers of IGSs with different reads coverage for each data set and get the result as shown in below. 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%) 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 below. 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,

which is pretty close to observed value. See Table 1

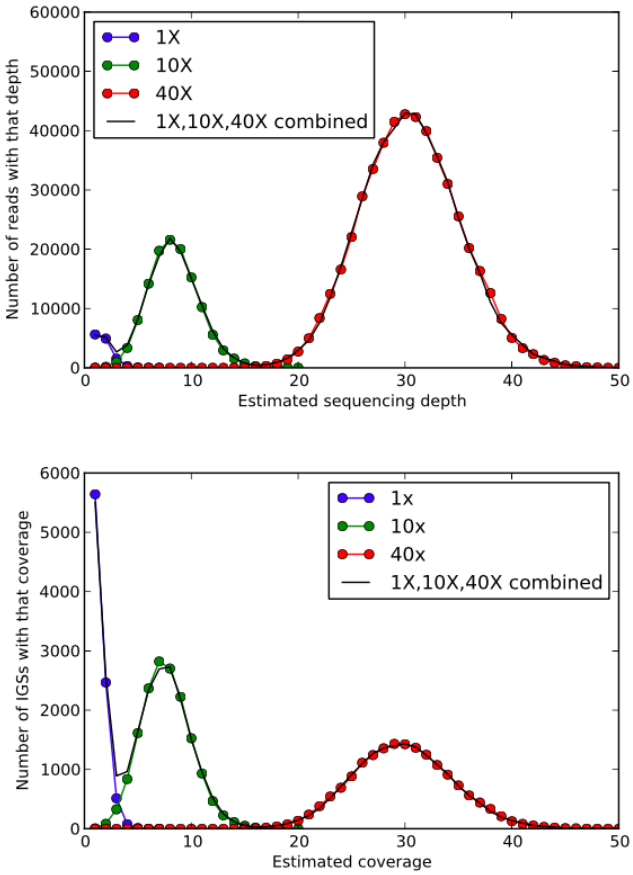


Figure 1: **from reads to IGS**

4 IGS can be used to do alpha diversity analysis

Basically the abundance distribution of IGSs with different coverage in a sample data set is acquired using the method shown above, like:

3 23

4 24

5 25

6 25

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

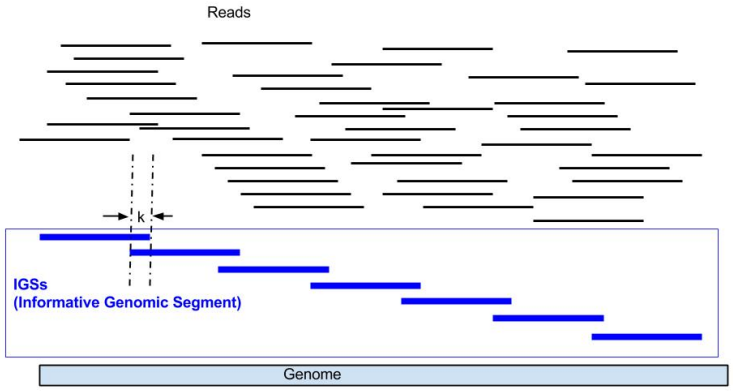


Figure 2: **the concept of IGS**

...

Here 23 IGSs with coverage as 3, this number is calculated from dividing the total number of reads with coverage as 3, which is 69, by the coverage 3: 69/3. Similarly there are $96/4 = 24$ IGSs with coverage as 4.

If we draw an analogy between IGSs and OTUs, this is like there are 23 different OTUs with 3 reads mapped to, and 24 different OTUs with 4 reads mapped to.

Then list all the different IGSs and the corresponding count,and we can get a long list with each IGS and the corresponding coverage.The coverage of an IGS can be considered as the abundance of such IGS in a sample. The list looks like:

IGS_ID abundance
1 3
2 3
3 3
...
23 3

24 4
25 4
...
47 4
48 5
...
...

This list is the counterpart of an OTU table in OTU based diversity analysis.

With such table at hand, numerous existing statistical methods and software packages can be used to investigate the alpha diversity.

5 IGS can be used to do beta diversity analysis

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

A typical OTU table across different samples is like this, which is also called samples-by-OTU data matrix.
OTU_ID Sample1_ID Sample2_ID Sample3_ID
OTU1 3 4 2
OTU2 2 5 0
OTU3 3 1 4

Like a OTU table, we hope to have the IGS table for the IGSs:
IGS_ID SampleA SampleB SampleC SampleD
IGS1 5 1 2 1
IGS2 5 1 2 1

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

Firstly, as how we get the coverage of a read from a sample dataset in this sample dataset, we can get the coverage of a read from a sample A dataset in another sample B dataset. We can still use the median k-mer count to represent the coverage. The basic idea is the same.

Because a read must derive from a segment in the genome of some species in a sample, if a read R from sample A with a coverage C_A in sample A has a coverage as C_B in sample B, that means that segment of genome in sample A from which read R derive also exists in sample B. That genomic segment has a coverage as C_A in sample A and has a coverage as C_B in sample B. Roughly there should be about C_A reads (read R should be one of them) in sampleA

covering that genomic segment and C_B reads in sampleB covering that genomic segment. Meanwhile, the C_A reads in sampleA should all have a coverage as C_B in sampleB, just like read R as one of them. Similarly, the C_B reads in sampleB should all have a coverage as C_A in sampleA.

Ok, now let's make an example.

Suppose there are 6 reads in sample A, all have a coverage as 3 in sampleA, and have a coverage as 2 in sampleB.

According to the discussion about IGS in previous section, the 6 reads cover 2 IGSs with a coverage as 3 for each IGSs. There should be 4 reads in sampleB covering the exact same 2 IGSs, with a coverage as 2 in sampleB.

So now we have 2 distinct IGSs with redundancy as 3 and 2 in the two samples respectively.

Note: small number is used in the analysis above as example, but it should be emphasized that the analysis is based on large number statistically.

Let's expand this example from 2 samples to 4 samples(A,B,C,D), as shown in figure above.

Let's say we find 10 reads in sampleA, with coverage as 5-1-2-1 in samples A-B-C-D respectively. (We call "5-1-2-1" "coverage spectrum" across samples.) So there should be **about** 2 reads in sampleB, 4 reads in sampleC, 2 reads in sampleD, all of which have a "coverage spectrum" as "5-1-2-1". Basically these 18 reads altogether cover 2 distinct IGSs, which apparently exist in all the 4 samples. The 2 distinct IGSs has a redundancy as 5,1,2,1 in the 4 samples respectively.

If we draw an analogy between IGSs and OTUs, this is like there are 2 OTUs, both with 5,1,2,1 reads mapped to in sample A,B,C,D respectively.

Like a OTU table, here we can have the IGS table for the two IGSs:
IGS_ID SampleA SampleB SampleC SampleD
IGS1 5 1 2 1
IGS2 5 1 2 1

6 An experiment using a simple simulated data sets

For this experiment, firstly we create 6 synthetic samples (Sample 1-6) based on 9 synthetic 10K genomes (genome A-I), with different composition of species and diversity.

The species composition for each synthetic sample is as below:
sample1: AAAB
sample2: AABC
sample3: ABCD
sample4: ABCE
sample5: AFGH

sample6: IFGH

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. So the species abundance in each sample is as below:

sample1: genomeA - 30, genomeB - 10

sample2: genomeA - 20, genomeB - 10, genomeC - 10

sample3: genomeA - 10, genomeB - 10, genomeC - 10, genomeD - 10

sample4: genomeA - 10, genomeB - 10, genomeC - 10, genomeE - 10

sample5: genomeA - 10, genomeF - 10, genomeG - 10, genomeH - 10

sample6: genomeI - 10, genomeF - 10, genomeG - 10, genomeH - 10

An a simple experiment, there is no sequencing errors introduced in the synthetic reads data sets.

Figure 3 and Figure 4 show that IGS method can yield the information about the difference of samples correctly. Sample 5 and sample 6 and very close to each other on the figure, which is true if we check the species composition of the two samples shown above.

Figure 5 shows the method can yield the richness information correctly. From the figure, samples with 4 different species have the richness almost twice as large as the sample with 2 different species.

From these results, we show the IGS method can work well to a simplest scenario, with high sequencing depth (10X) and no sequencing error. Next we will check the influence to 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.

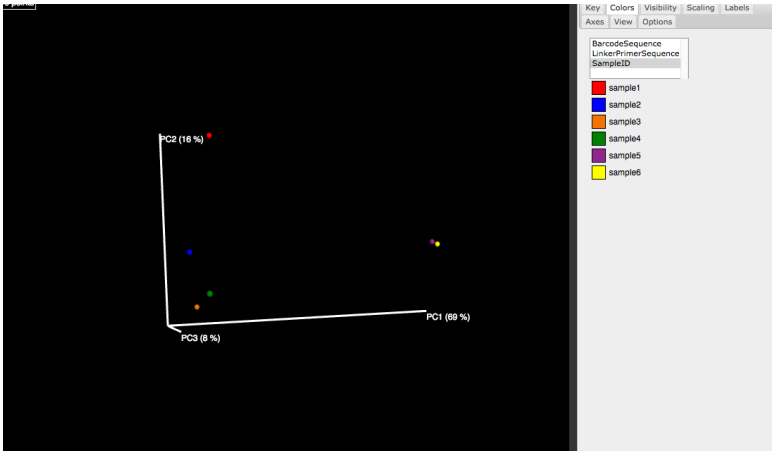


Figure 3: **Ordination of the 6 synthetic samples using IGS method**

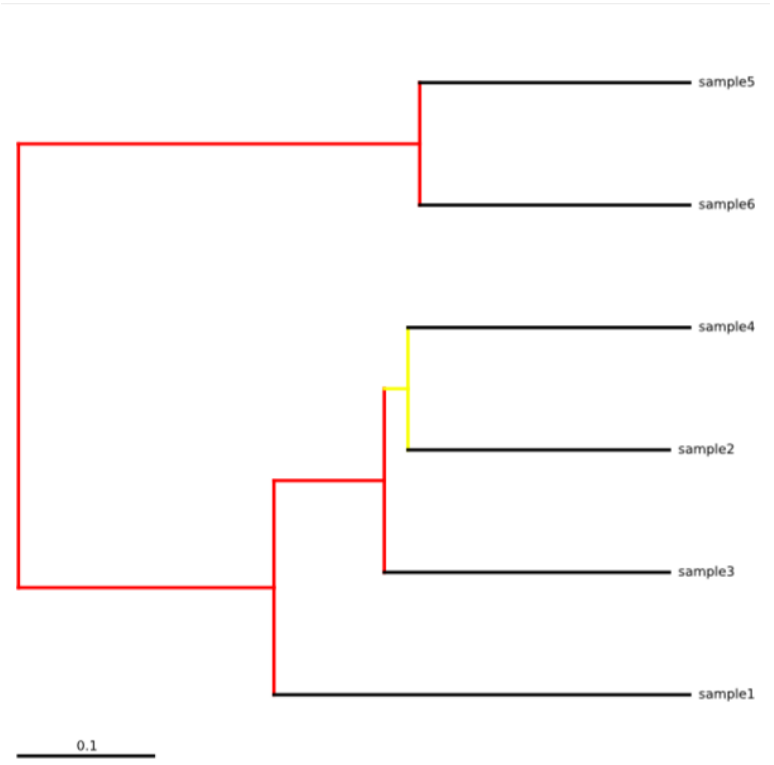


Figure 4: **Clustering of the 6 synthetic samples using IGS method**

7 Evaluating the influence of variable sequencing depth and sequencing error

Previously we have shown the IGS method generally works to a simple simulated data sets, with high sequencing depth and no sequencing error. Since in many situations we deal with metagenomic data sets with low sequencing depth, like soil or sea water samples, we want to know if we can still get reliable insights from the low coverage sequencing data. Also it is a fact that all sequencing technology will generate some errors. As discussed in the background section, one of the reasons we develop the IGS method is that based on the abundance counting of reads rather than k-mers, it is expected that the IGS method is more robust to sequencing error. We also wonder how error correction will help improving the results of IGS method.

6 synthetic samples are generated with the same species composition as in last section. For each sample, sequencing reads data sets with different sequencing depth are simulated. A series of sequencing reads data sets are also generated with sequencing error as 3%, besides a series of error-free data sets.

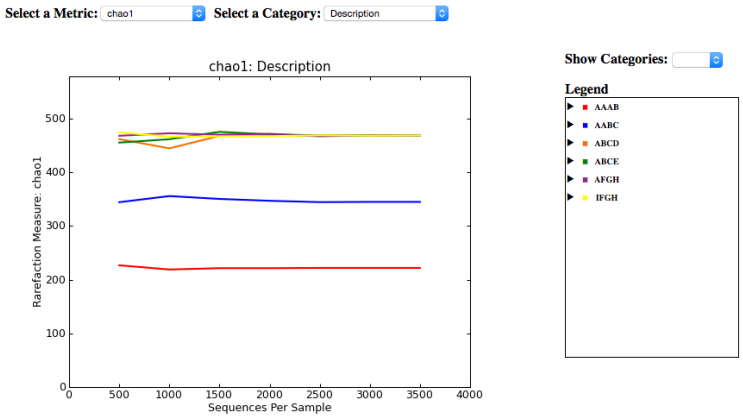


Figure 5: **Richness estimation of the 6 synthetic samples using IGS method**

Here we plan to use these simulated data sets to check the effectiveness of the IGS based method(generally and the effectiveness of error correction/filtering) and the influence of variable sequencing depth.

7.1 comparison of dissimilarity matrix from beta diversity analysis

To evaluate the effectiveness of beta diversity analysis using IGS based method, we compare the dissimilarity matrix generated by IGS based method with that generated from another metagenomics comparison tool - Commet(Compareads), and the true matrix,since we know exactly the species composition of the simulated data sets.

The clustering and ordination are all from the dissimilarity matrix. We think comparing matrix directly makes more sense than comparing the clustering and ordination plot. So we will not show the clustering and ordination figure in this evaluation. If the matrix can reflect the real relationship between samples reliably, the clustering and ordination will only be routine job.

The true dissimilarity matrix of the 6 simulated samples using bray-cutis metric from species composition directly is shown in Figure 6. For a simulated data set with 10x coverage and no error introduced (which will tell us the optimal performance of IGS method), the dissimilarity matrix can be calculated by using IGS method, as shown in Figure 7. We can see the absolute values in the matrix are not very close to that in the real matrix.But the relative values correspond to that in the real matrix well to show the relative distance between each pair of samples. To get a objective metric, we use Mantel (citing)test to calculate the correlation value between the two matrixes. The correlation is 0.9714, which means a very positive correlation between the two matrices.

We are very confident that the matrix from IGS method can reflect the true relationship between samples pretty well.

Next we test how well the matrix calculated by various methods can reflect the real relationship between samples. The simulated data sets with sequencing depth as 1 and 10, with sequencing error as 3% and without sequencing error are used in this experiment. For the data sets with sequencing error, we use a HMM based error correction tool to preprocess the reads to check the effectiveness of error correction. We also compare the performance of IGS based method and another metagenome comparing tool - Comet.

As shown in Figure 8, firstly, for all data sets, the matrix from IGS method has a higher correlation to golden standard than that from Comet. As expected, the matrix from data sets with sequencing error has a lower correlation than that from error-free data sets. Also Comet is more sensitive to sequencing error rate, compared to IGS method. However, error correction can increase the correlation significantly. Also higher coverage will yield more accurate matrix, which is not surprising.

Figure 9 shows how well the matrix calculated from data set with variable coverage can reflect the real relationship between samples. It is as expected that higher coverage data will yield better/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 how reliable the result will be if we only use a small proportion of data from a large metagenomic data set.

	sample1	**sample2**	**sample3**	**sample4**	**sample5**	**sample6**
sample1	0.00	0.25	0.50	0.50	0.75	1.00
sample2	0.25	0.00	0.25	0.25	0.75	1.00
sample3	0.50	0.25	0.00	0.25	0.75	1.00
sample4	0.50	0.25	0.25	0.00	0.75	1.00
sample5	0.75	0.75	0.75	0.75	0.00	0.25
sample6	1.00	1.00	1.00	1.00	0.25	0.00

Figure 6: **Dissimilarity matrix between synthetic samples using Bray-cutis from species composition directly**

7.2 Evaluate alpha diversity analysis by estimating size of metagenome

We can use statistic 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 x (reads_length - k-size +1)

	sample1	sample2	sample3	sample4	sample5	sample6
sample1	0.000000	0.354200	0.60460	0.660600	0.803225	1.000000
sample2	0.354200	0.000000	0.42205	0.508875	0.836375	1.000000
sample3	0.604600	0.422050	0.00000	0.564600	0.893100	1.000000
sample4	0.660600	0.508875	0.56460	0.000000	0.892150	1.000000
sample5	0.803225	0.836375	0.89310	0.892150	0.000000	0.422075
sample6	1.000000	1.000000	1.00000	1.000000	0.422075	0.000000

Figure 7: **Dissimilarity matrix between synthetic samples using Bray-cutis from sequencing reads using IGS method**

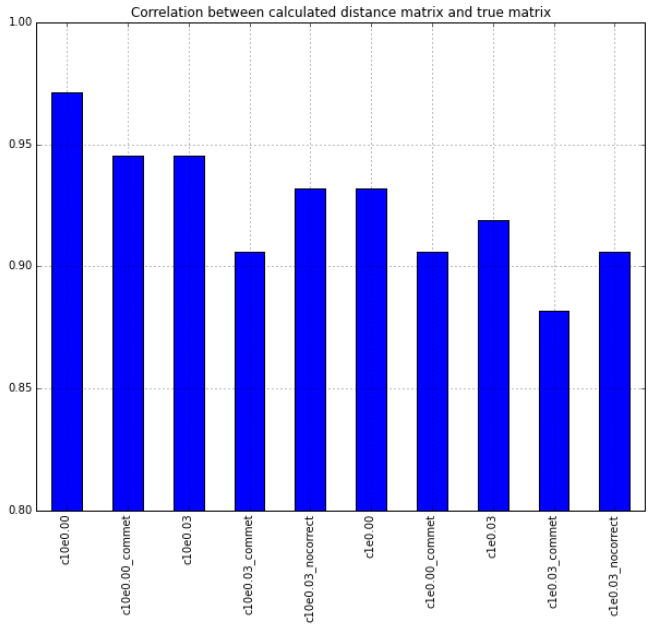


Figure 8: **Correlation between calculated distance matrix and true matrix from different data sets and using different methods**

Here we check how accurate the estimated size of genome and coverage is using different data sets with variable coverage/sequencing depth.

The genome size of the 6 samples should be:
sample1: AAAB 2x100K = 200K bp
sample2: AABC 3x100K = 300K bp

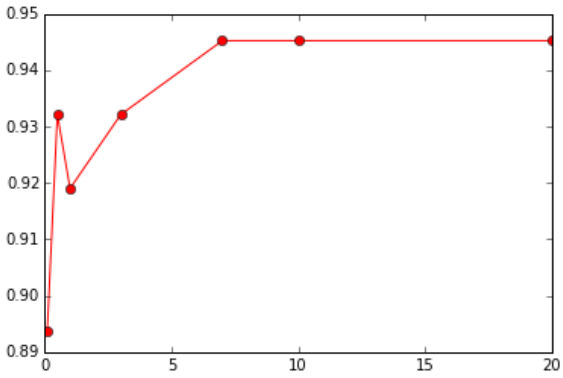


Figure 9: **Correlation between calculated distance matrix and true matrix from different dat sets with different sequencing depth**

sample3: ABCD 4x100K = 400K bp
sample4: ABCE 4x100K = 400K bp
sample5: AFGH 4x100K = 400K bp
sample6: IFGH 4x100K = 400K bp

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

Table 10 shows the estimated genome size of samples using error-free simulated sequencing data sets is close to true size shown above. If the sequencing error is introduced, the estimated genome size is inflated dramatically as shown in Table 11, which is not surprising. However after applying error correction, we can still get good estimation of genome size, as shown in Table 13. This proves again error correction can improve the effectiveness of IGS method.

Figure 13 shows the estimated genome size from data sets with variable coverage. The estimated genome size keeps increasing as we use more reads, with higher coverage,which is higher than actual genome size. It's interesting that the estimated genome size is very high with low coverage, (probably due to more unique k-mers, which makes error correction more difficult. Then the estimated genome size drops a little bit as coverage from 0.1X to 1-3X, then starts to climb again as coverage increases, probably because there are always more erroneous k-mers that cannot be corrected. The good thing is that the climbing rate is not that high, this is believed to be due to the effectiveness of our error correction algorithms.

It is important to point that even though the absolute value of estimated genome size may be overestimated. The relative relationship between samples are reliable, as shown in the figure. Sample3,4,5,6 all have 4 species, while sample 2 has 3 species, and sample 1 has 2 species. They can be separately pretty well.

	observed_IGS	ace	goods_coverage	simpson_evenness	estimated_genome_size
sample					
sample1	2335	2335	1	0.785113	189135
sample2	3494	3494	1	0.849064	283014
sample3	4618	4618	1	0.923235	374058
sample4	4623	4623	1	0.924541	374463
sample5	4611	4611	1	0.921446	373491
sample6	4632	4632	1	0.923992	375192

Figure 10: **Coverage = 10x, No error**

	observed_IGS	ace	goods_coverage	simpson_evenness	estimated_genome_size
sample					
sample1	17424	55634.121283	0.706557	0.361417	4506363.823955
sample2	19526	50195.556857	0.679834	0.501933	4065840.105448
sample3	21414	45097.526012	0.659845	0.624600	3652899.606962
sample4	21390	45350.043739	0.659646	0.621892	3673353.542853
sample5	21325	44507.263605	0.663239	0.624775	3605088.352024
sample6	21395	45419.187020	0.659144	0.622005	3678954.148641

Figure 11: **Coverage = 10x, error = 3%, no error correction**

	observed_IGS	ace	goods_coverage	simpson_evenness	estimated_genome_size
sample					
sample1	2442	2445.313100	0.999352	0.757776	198070.361092
sample2	3562	3564.924176	0.999398	0.816215	288758.858253
sample3	4613	4618.386655	0.998890	0.887475	374089.319053
sample4	4589	4595.167016	0.998726	0.884003	372208.528258
sample5	4615	4621.682515	0.998622	0.882599	374356.283742
sample6	4606	4611.082311	0.998953	0.884558	373497.667230

Figure 12: **Coverage = 10x, error = 3%, with error correction**

7.3 The IGS method can provide a whole framework to do alpha or beta 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.

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. Actually any metric can be applied to the IGS-by-samples table, abundance-based or incidence-baserd, richness or evenness.

Compareads(Commet) based on reads overlap between samples can get a

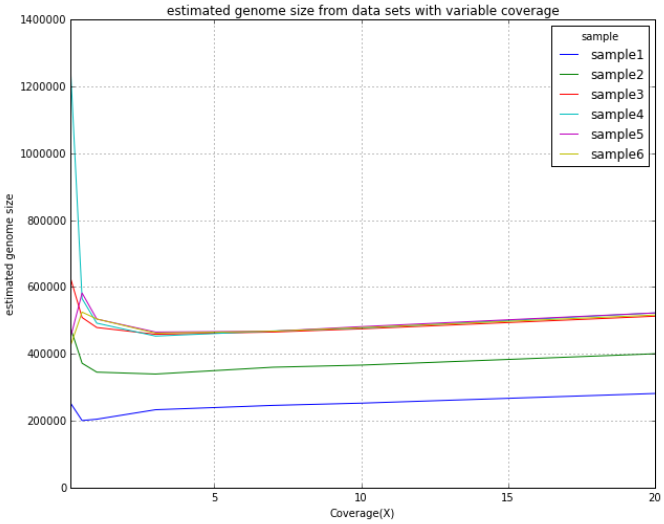


Figure 13: **estimated genome size from data sets with variable coverage**

matrix reflecting the real relationship between samples pretty good but it is stuck with one metric, which is based on the percentage of overlap reads between samples. This metric is like bray curtis , but not exactly the same.

8 GOS data sets: Sorcerer II Global Ocean Sampling Expedition

9 HMP metagenomics data set

10 GPGC soil sample - new and old

Using 1m and 2m randomly selected subsets can yield pretty good results.

11 iterated diversity analysis

As we load more reads, we will see the separation more clearly.
simulated data sets.

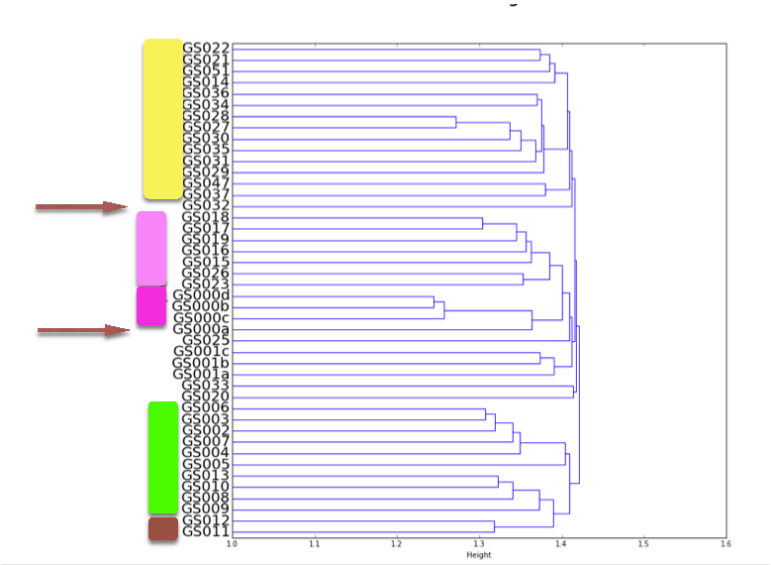


Figure 14: **cluster of GOS samples using IGS method**

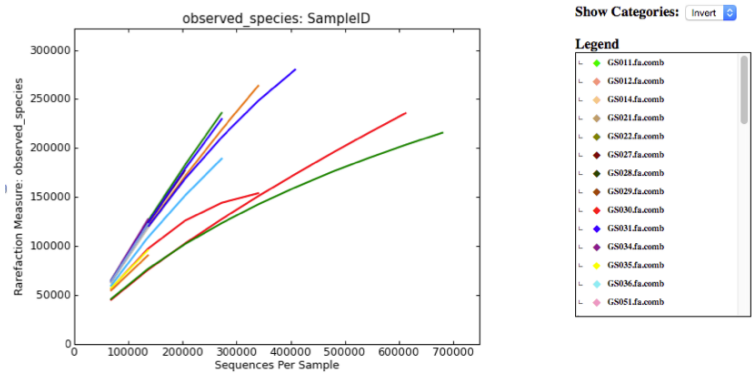


Figure 15: **cluster of GOS samples using IGS method**

12 IGS can be the foundation of a new framework to do microbial diversity analysis

13 IGS is promising to more problems

1. alpha-diversity analysis (1 sample) - richness/evenness - rarefaction curve - sequencing depth evaluation - genome size estimation - better choosing diginorm

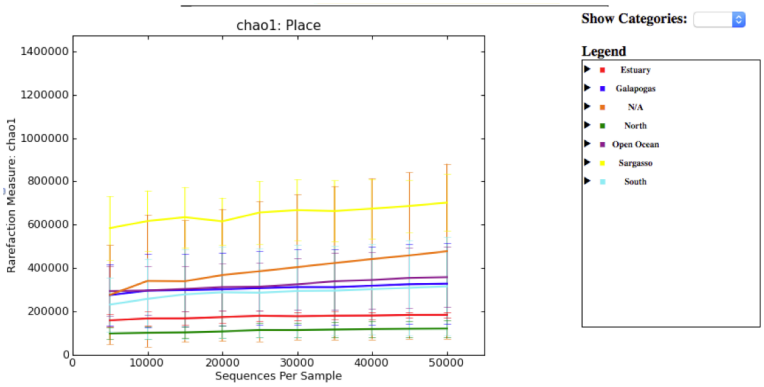


Figure 16: **cluster of GOS samples using IGS method**

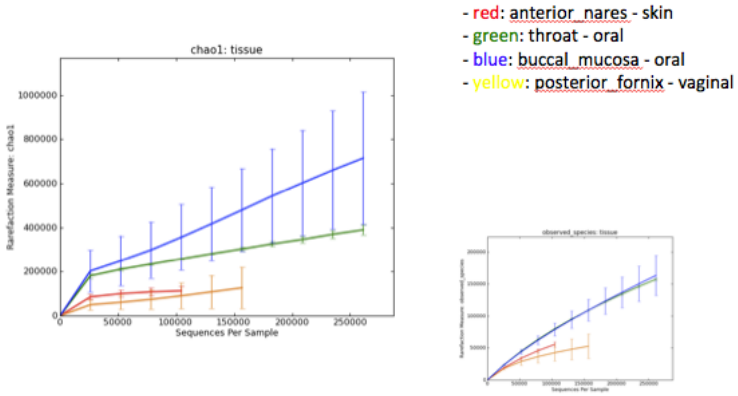
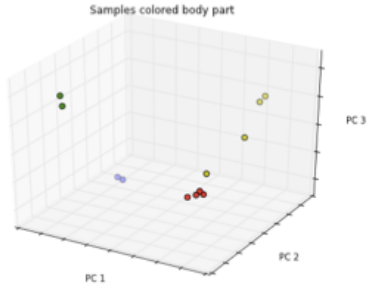


Figure 17: **rarefaction of HMP**

parameters(size of hashtables, etc.)

2. beta-diversity (multiple samples) - sample by sample comparison, clustering, ordination after getting segment-count table

3. other potential applications: - reads binning/classification (after clustering)(if number of samples is small, may not be effective) - extract IGS(reads) according different filters (shared by all samples, or some specific samples,) - co assembly (by extracting the reads with total coverage across samples ¿ 10, for example)



I picked the smallest 12 samples from ~700 HMP samples with 3 groups from different body parts. (skin, oral, vaginal).

- red: anterior_nares - skin
- green: throat - oral
- blue: buccal_mucosa - oral
- yellow: posterior_fornix - vaginal

Figure 18: **PCoA of HMP**

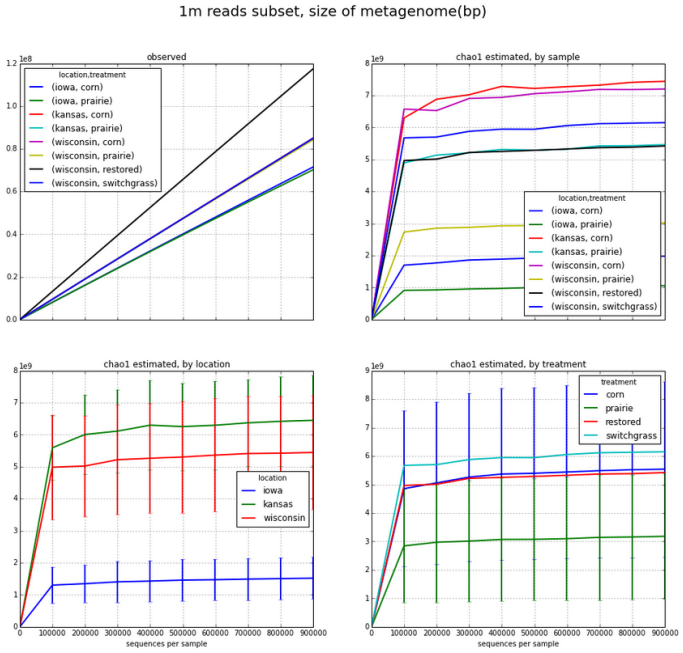


Figure 19: **alpha of GPGC**

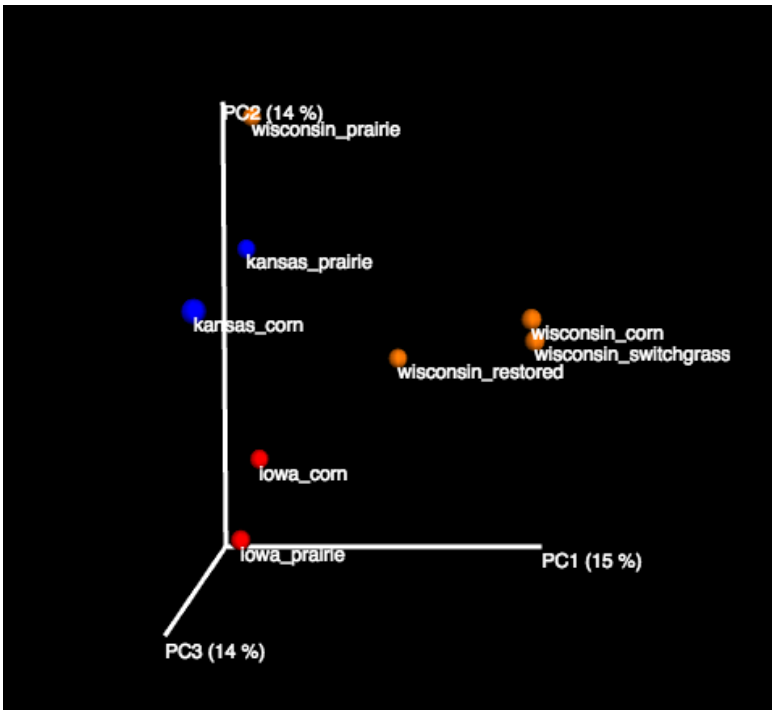


Figure 20: **beta of GPGC**

14 Conclusion

Advantage:

not only HMP, high depth, data but also low depth data, like soil metagenomics, which is impossible to use traditional method

15 Data

15.1 Four simulated reads data sets with different species abundance distribution

15.2 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 2

Table 2: **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 3: 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