Using the Concept of Informative Genomic Segment to Investigate Microbial Diversity of Metagenomics Sample

Qingpeng Zhang, C. Titus Brown

Department of Computer Science and Engineering
Michigan State University, East Lansing, MI, United States

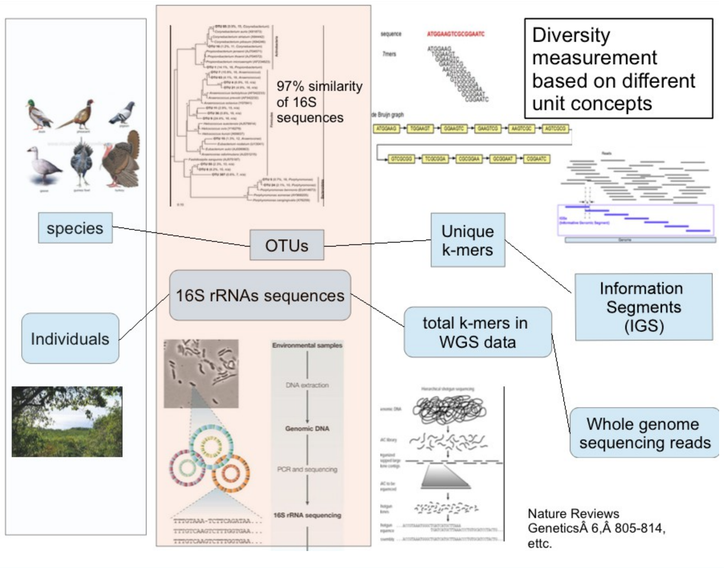
Department of Computer Science & Engineering

Introduction

In almost all the 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 OTUs(Operational Taxonomic Units) are used as the cornerstone for diversity analysis. 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 Global Ocean Sampling Expedition (GOS) dataset and the samples were clustered more accurately than existing alignment-based method. We also tried this novel method to MetaHIT data sets. 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.

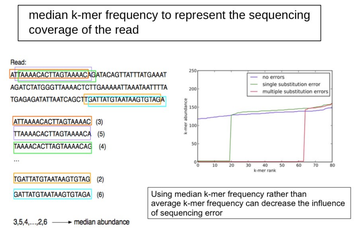
Background

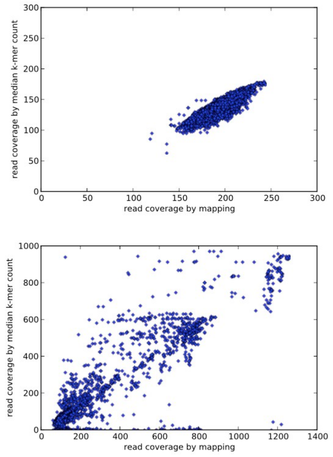
In traditional ecology, the concept of species is used to investigate diversity. In microbial ecology, the concept of OTU is used to investigate microbial diversity. OTU is mostly used for 16S data sets. And binning reads into OTU is typically required for OTU-based diversity analysis. Here we propose a new concept - IGS to replace the concept of OTU in 16S based diversity analysis and the concept of species in traditional ecology.



Nature Reviews GeneticsÂ 6,Â 805-814, ettc.

The concept of IGS(informative genomic segment)

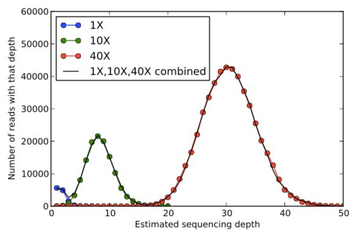




Mapping and k-mer coverage measures correlate for simulated genome data and a real E. coli data set (5m reads).
Simulated data r2 = 0.79;
E. coli r2 = 0.80.

(Brown et al., 2012)

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



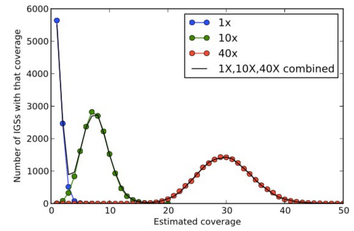


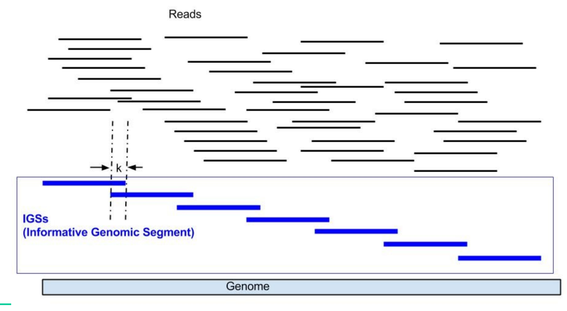
Table 1: Total number of IGSs in different simulated reads data sets

Data set	total number of IGSs
1X depth	8714
10X depth	16321
40X depth	16794

The abundance distribution with different sequencing depth of reads from 4 simulated sequencing data sets - 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.

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.

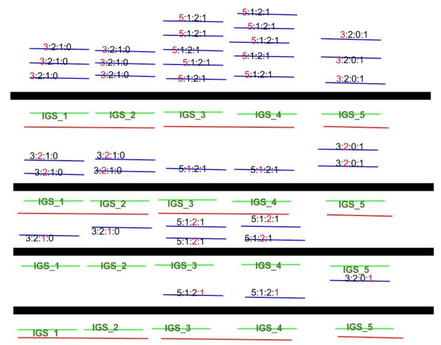
Each point on the curve shows that there are Y reads with a sequencing depth of X. In other word, for each of those Y reads, there are X-1 other reads that cover the same DNA segment in a genome that single read originates. So we can estimate that there are Y/X distinct DNA segments with reads coverage 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.



The concept of IGS for single genome

Using IGS to do diversity analysis

Build samples-by-IGS matrix to replace samples-by-OUT matrix



IGS_ID	SampleA	SampleB	SampleC	SampleD
IGS1	3	2	1	0
IGS2	3	2	1	0
IGS3	5	1	2	1
IGS4	5	1	2	1
IGS5	3	2	0	1

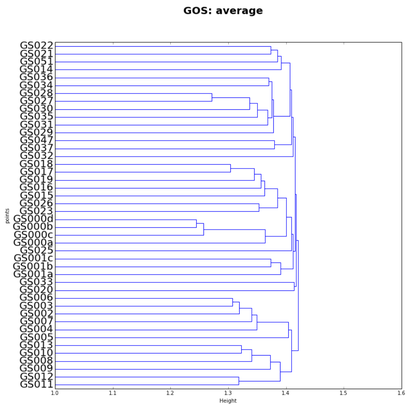
OTU_ID	SampleA	SampleB	SampleC	SampleD
OTU1	3	2	1	0
OTU2	3	2	1	0
OTU3	5	1	2	1
OTU4	5	1	2	1
OTU5	3	2	0	1

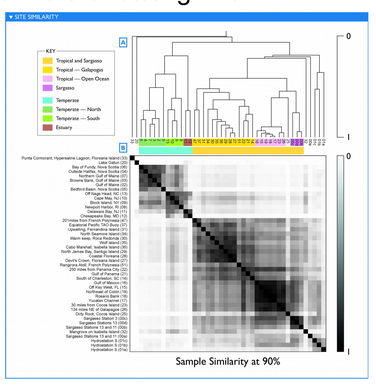
black: overall information in the genomes of all species in a sample
blue: reads, with the coverage spectrum across samples
green: IGSs
red: potential genomes of separate species

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.

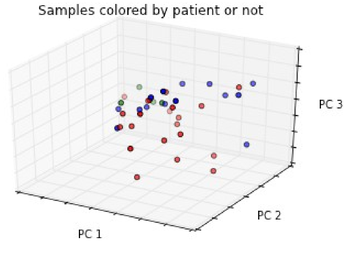
Apply IGS-based method on real data sets

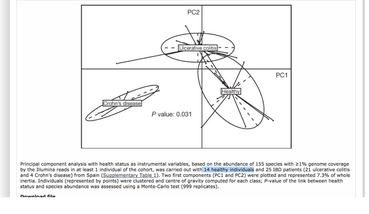
IGS-based method can get comparable if not better beta-diversity result than traditional methods based on reference/alignment.





Global Ocean Sampling Expedition (GOS) IGS-based vs. alignment-based (Rusch et al, 2007)





MetaHIT dataset (14 healthy individuals and 25 IBD patients) IGS-based vs. alignment-based (Qin et al, 2009)

This IGS-based method to do microbial diversity analysis is totally binning-free, assembly-free, annotation-free, reference-free and using this method to do alpha diversity analysis is under investigation.