Summary of ARMO shotgun read assembly:

Author: Jiarong Guo

Tags: bioinformatics, assembly, metagenomics

Date: 2013-04-22

Slug: asseSum

Category: science

(with Adina Howe, James Tiedje, Titus Brown)

I have been working on the assembly of big shotgun metagenomic data from ARMO (Amazon Rain Forest Microbial Observatory) project. The biggest challenge is the huge data size, 2TB in fastq and more than 6 billions reads after read trimming. One lucky thing is that MSU provides High Performance Computer Clusters, so we have access to memory up to 1TB. Still, to my knowledge, no assemblers can handle this amount of data with 1Tb of memory. Well, some tools like SGA may be able to handle the data if they are very redundant, but this is not the case for soil metagenomic data. Digital normalization and partitioning, two data preprocessing methods designed by our lab (Dr. Titus Brown) are used to make the assembly doable.

Background and data

Amazon rain forest is very important for global ecosystem and microbial communities play key roles in the nutrient cycling. The goal of this study is to understand the effect of agriculture practice (conversion of forest to pasture) on microbial community. A survey based on 16s rRNA gene have shown that the land conversion cause higher local microbial diversity but more similar community across space. To further study the microbial community at functional level, shotgun metagenomics is used.

To evaluate the effects of spatial variation, samples are taken at corner of nested 0.01m, 0.1m, 1m, 10m, 100m square plot at forest site and pasture site. Each sample from a square plot are sequenced by two Illumina HighSeq lanes.

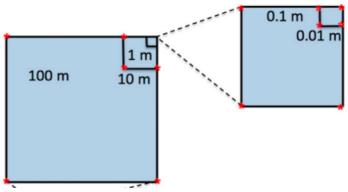

Figure 1. Experiment design. Samples are collected at corners are nested squares of 0.01, 0.1, 1, 10, 100 meters. Details are in this paper

Methods

Data trimming: the end of reads with quality score 2 (ASCII “B”) are trimmed and reads with length longer than 30bp are kept. Pair end reads with overlap are stitched with FLASH (version 1.0.3).

Digital normalization, sequencing artifact removing, and partitioning: This step is done by a bash script (https://github.com/jiarong/khmer/blob/master/ged-lab/armo-gjr/asse6.bash).

Assembly: VELVET assembler is used to assemble the group of sequence file with kmer size of 33, 37, 39, 49, 69. The one producing the most overall assembly length is picked. Another version of VELVET compiled with “BIGASSEMBLY” flag is used for assembly the biggest group of sequence file.

Results

Computational cost

Digital normalization: 150 cpu hours to finish 10 lanes of Illumina HighSeq data for each treatment (forest or pasture) with 1TB memory.

Artifact removal: 40 cpu hours to finish 10 lanes.

Partitioning: 1) 69 cpu hours for loading graph; 2) 346 cpu hours for partitioning graph (12 hours with 32 core on HPC); 3) 12 cpu hours for merging graph; 4) 42 cpu hours for annotating partition; 5) 13 cpu hours for extracting partitions. In total, the partitioning took one week to finish, because the most time consuming step, partitioning graph can be run with multi-threads.

Assembly: There is always one group file that is significantly larger than the others after partitioning because reads from different species could be connected by some conserved genes or artifacts. We call it the “lump”. The “lump” took about 20 cpu hours, while the other small group files took less than 0.5 cpu hours.

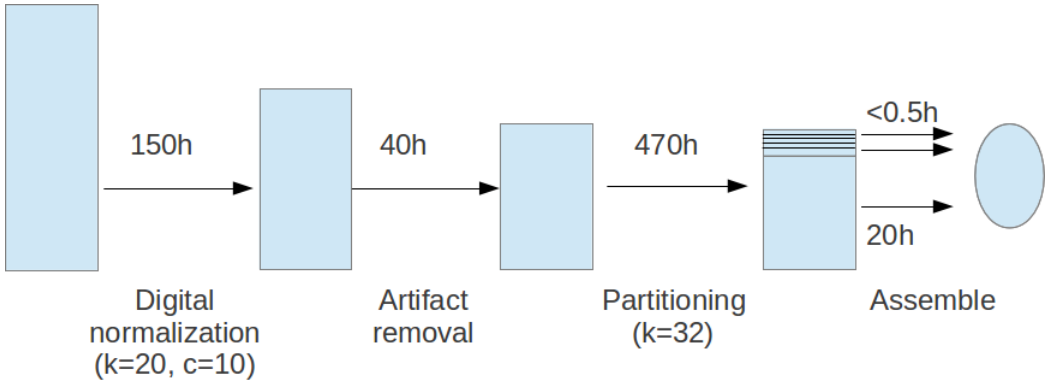

Figure 2: Flow chart of the assembling process.

Genomic saturation

Figure 3, 4 shows there is a fair amount of shared sequence among samples within the same treatment (forest), which also suggests the data within the treatment can be combined for better assemblies.

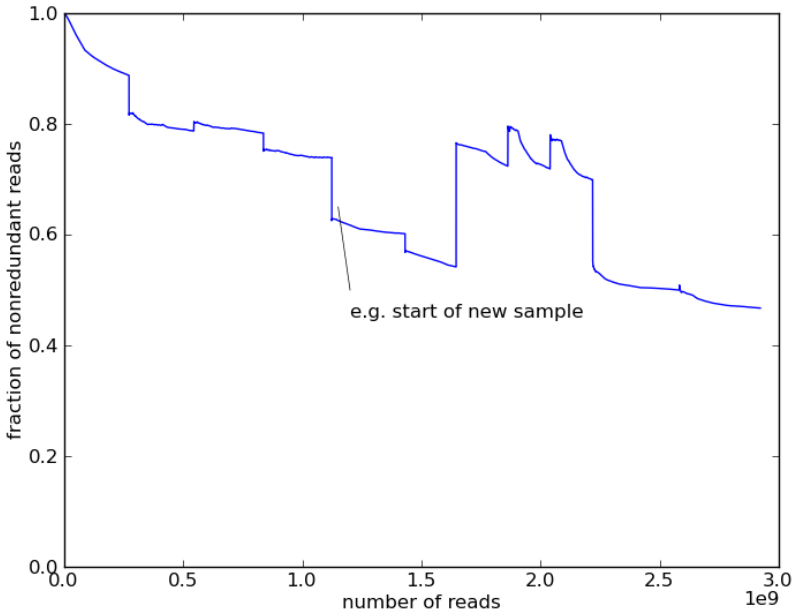

Figure 3. Genomic content saturation curve from digital normalization output from forest samples. The X axis the number of reads processed by digital normalization. The Y axis the fraction of reads kept after digital normalization in the current sample data.

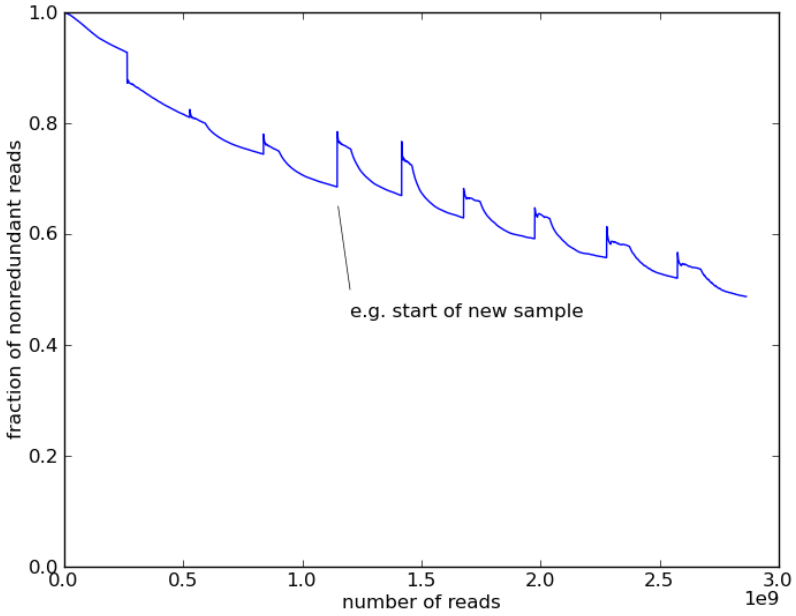

Figure 4. Genomic content saturation curve from digital normalization output from the pasture samples. The X axis the number of reads processed by digital normalization. The Y axis the fraction of reads kept after digital normalization in the current sample data.

Assembly statistics

Table 1: Assembly statistics. Minimum contig length of 800 bp is chosen. The low mapping percentage is due to the high minimum contig length cutoff.

Sample	>800bp contigs	total bp	max	mapping	mapping%	total reads
Forest	265073	271553890	9115	10732078	0.37%	2923068636
Pasture	497664	538669724	24861	23082813	0.81%	2863547487

As shown in Table 1, pasture data have more and longer contigs assembled (with 800bp cutoff). The DNA content in the two metagenome assemblies share little similarity (Table 2). Rank abundance curves (Figure 5) show the pasture assembly has better coverage than forest assembly.

Table 2 Similarity between two assemblies. Contigs covered is the fraction of total contigs covered by any contigs from the other sample. Total bp covered is the fraction of total basepairs covered by contigs from the other sample.

Sample	>800bp contigs	total bp	contigs covered	total bp covered
Forest	265073	271553890	9.30%	3.50%
Pasture	497664	538669724	5.20%	1.80%

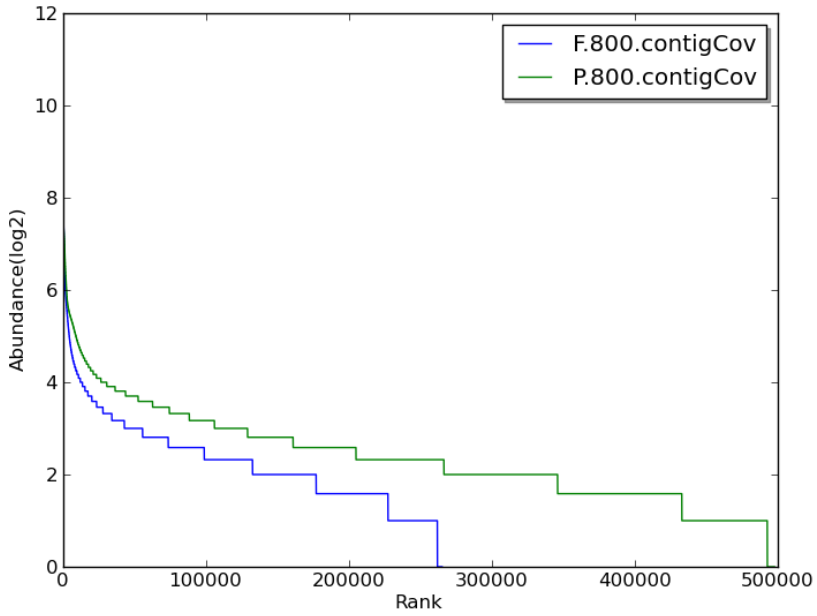

Figure 5. Rank abundance of forest and pasture samples. Bule curve (F.800.contigCov) is from the combined forest assembly with 800 minimum length and green curve (P.800.contigCov) is from the combined pasture assembly with 800 minimum length. The coverage is based on median of coverages on each base position in contigs after mapping.

Spatial variation

Clustering and heatmap of samples (Figure 6, 7) in each treatment based on the fold coverage of top 1000 most abundant contig shows that even though sharing a fair amount of genomic content, there are still spatial variation between samples. The closer the distance, the more similar the samples are.

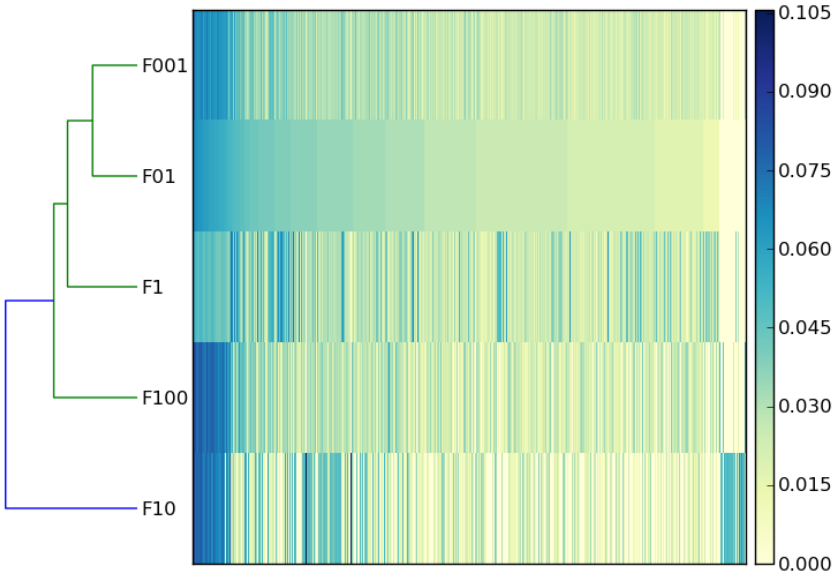

Figure 6. Heatmap of top 1000 most abundant contigs within forest samples. Samples are in the rows. "F" stands for forest. The number follow the "F" stands for the distance of sampling spot. For example, "F001" is 0.01m at forest. Contigs are the columns (name not shown due to large amount). The columns are sorted based on the abundance in F01 for A. The abundance matrix is transformed by hellinger transform prior to clustering.

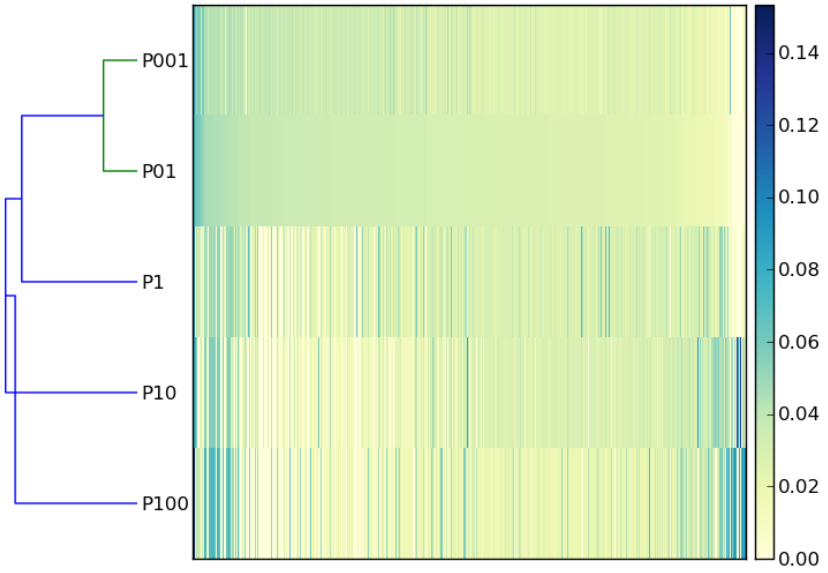

Figure 7. Heatmap of top 1000 most abundant contigs within pasture samples. Samples are in the rows. "P" stands for pasture. The number follow the "P" stands for the distance of sampling spot. For example, "P001" is 0.01m at pasture. Contigs are the columns (name not shown due to large amount). The columns are sorted based on the abundance in P01 for pasture. The abundance matrix is transformed by hellinger transform prior to clustering.

Conclusion

Digital normalization and partitioning are effective methods to assemble large metagenomic data. The assemblies from combined forest samples and from combined pasture samples share less than 5% similarity, which indicates the metagenomic content in forest and pasture are quite different. The microbial communities across space share a fair amount of genomic content, but they are still distinct from each other, especially when distance are large. The assembly data has been uploaded to MG-RAST for annotation. Next step will be comparing the communities with gene or functional category rather than just the contigs.