Testing the fidelity of assembly, binning and gene calling in metagenomes

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Konstantinos Mavromatis[1], Natalia Ivanova[1], Kerrie Barry[1], Harris Shapiro[1], Eugene Goltsman[1], Alice McHardy[2], Isidore Rigoutsos[2], Asaf Salamov[1], Frank Korzeniewski[1], Miriam Land[3], Phil Hugenholtz[1], Nikos C Kyrpides[1].

1. Joint Genome Institute. Walnut Creek, CA 94598

2. IBM Thomas J.Watson Research Center, Yorktown Heights NY 10598

3. Oak Ridge National Laboratory, Oar Ridge, Tennessee 37831

Introduction

In an efford to evaluate methods used to analyse metagenomes, we constructed three synthetic metagenomic datasets of increased complexity by combining reads from a selection of 113 isolate genome sequencing projects available through the Joint Genome Institute.

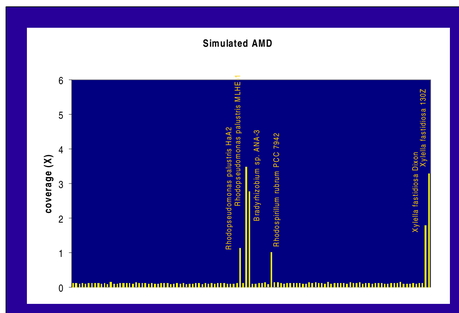
Isolate genomes were selected to represent populations in metagenomic datasets based on similar patterns of genome size, GC content and relative phylogenetic position. Reads were randomly sampled from the selected genomes to match the read depth of their corresponding populations in the metagenomic assemblies.

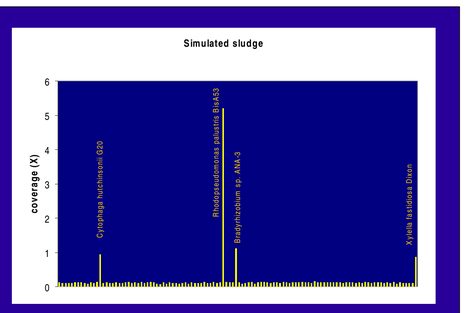
Sampled reads were then assembled using three programs used to assemble the real metagenomic data (Phrap, Arachne and JAZZ). Assembled contigs were binned using three different methods (oligonucleotide frequency, pattern discovery, best blast hit) and genes were called using two gene prediction pipelines (fgenesh, Critica/Glimmer). Due to the nature of the simulated metagenomes we were able to evaluate the quality of each step in the process. We explore the role of the population distribution as well as the algorithms used in the quality of the final metagenomic dataset used for further analysis.

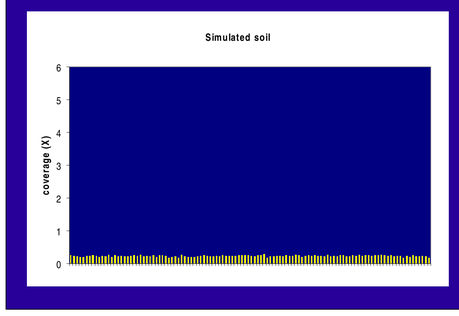
The results of the analysis show the dependency of each step on the dataset profile as well as the different limitations ot the methods.

The results of the analysis, as well as the sequences will be available to the public. New methods for assembly, binnning and gene calling can be tested and compared to each other using these datasets.

Simulated datasets

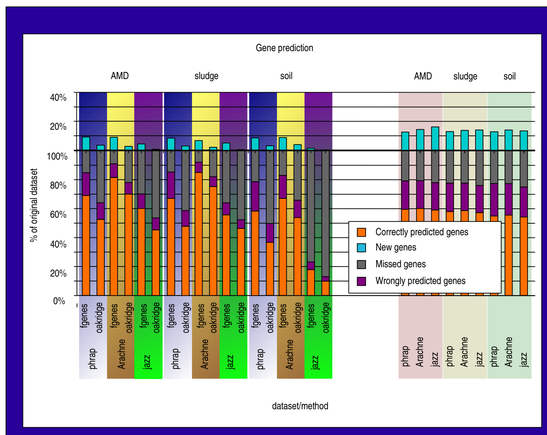






Each dataset approximately 100,000 reads were selected randomly to match the species diversity and abundance found in 3 types of metagenomic samples sequenced by the JGI, including (i) acid mine drainage (AMD) biofilm, representative of low complexity samples with few dominant organisms, (ii) biological phosphorus removing (EBPR) sludge community, which has higher diversity but only one highly abundant species, and (iii) Minnesota agricultural soil sample, representative of complex communities with very high species diversity and low abundance of each species

Gene calling



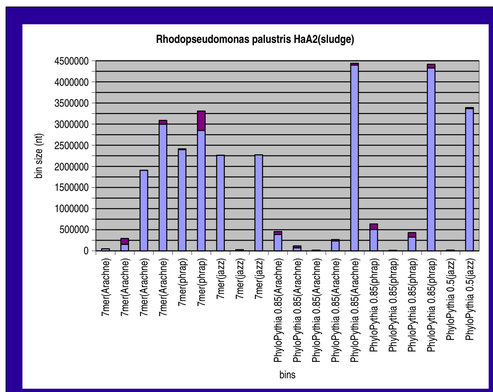
Contigs and scaffolds generated by 3 assemblers and unassembled reads were annotated by 2 *ab initio* gene finding algorithms, namely Fgenesb and Genome Analysis Pipeline – an automated gene prediction pipeline used to annotate isolate prokaryotic genomes at the DOE Joint Genome Institute.

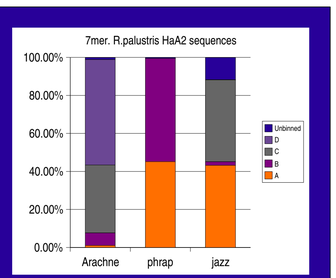
- Fgenesb models potential protein-coding regions by 5-th order Markov chain probabilities and non-coding upstream regions by 2-nd order positional Markov-chain probabilities. A "generic bacterial gene model" was employed. It was calculated based on a combined set of 10,000 CDSs from 100 complete bacterial genomes.
- Genome Analysis Pipeline uses a combination of CRITICA and Glimmer, where CRITICA finds potential protein-coding regions based on sequence similarity and dicodon bias, and Glimmer is used to fill the gaps.

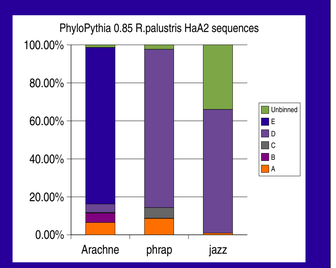
The number of predicted genes is compared with the set of genes that were predicted on the original, isolate genomes

Fgenes predicts a larger number of genes in all datasets. A group of predicted genes (~10%-20%) are predicted on different frame or strand (wronly predicted genes). A larger percent is not found on the assembled contigs (~20%-80%)

Binning example



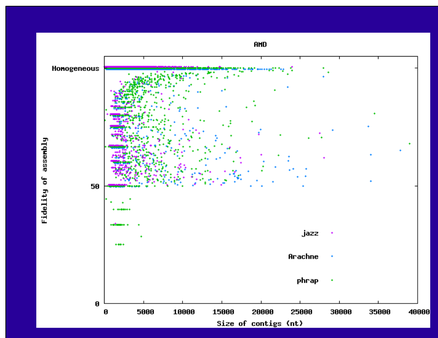


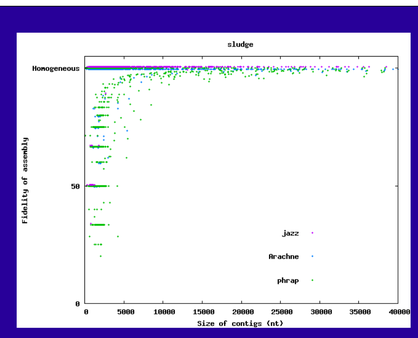


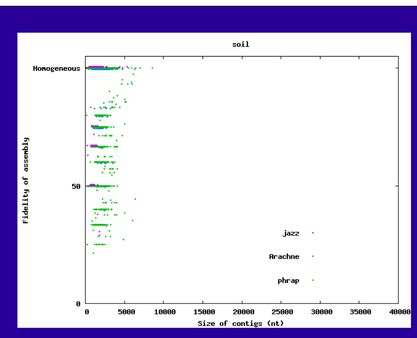
Examples of binning of *Rhodopseudomonas palustris* HaA2 in sludge

- The majority of the sequence is assigned to bins (Right).
- These bins are either homogeneous or containa small fraction of sequences from other organisms (Left).

Assembly



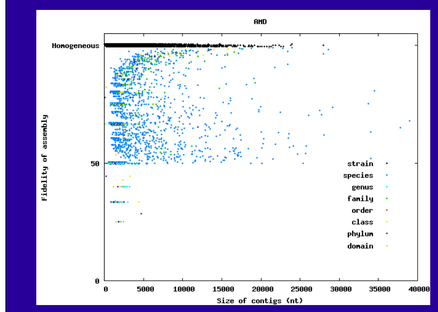


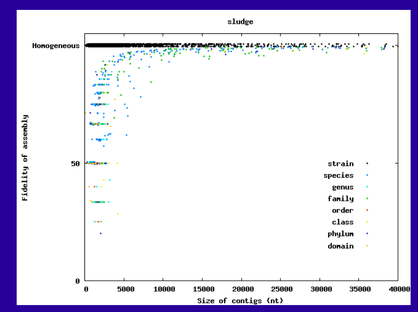


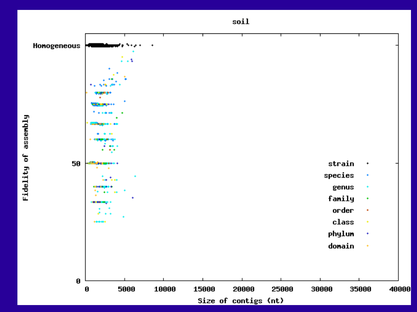
The reads from 3 simulated metagenomic datasets were assembled using 3 assembly tools, JAZZ , PHRAP and Arachne.

The following assembly parameters were used: for JAZZ read depth penalties were turned off, as were the models assuming that the dataset contained a single primary constituent. Parameters for Arachne were recommended by the developers to make it more inclusive and less stringent; parameter settings were as follows: recycle_bad_contigs=False, maxcliq1=500, maxcliq2=500, fast_draft_consensus=False, mc_min_overlap=30. Parameters for PHRAP were set as follows: minmatch 30, maxmatch 55, minscore 55, max_subclone_size 50000, revise_greedy, vector_bound 20.

	Sim sludge (Phrap)	Sim sludge (Arachne)	Sim sludge (JAZZ)
Reads	100628	100628	100546
Scaffolds			2753 / 33249
Contigs (>1 reads) / reads	10320 / 56843	1400 / 33065	3773 / 33247
Major contigs (≥ 10 reads) / reads	482 / 30852	367 / 28060	504 / 24417
	Sim soil (Phrap)	**Sim soil (Arachne)**	**Sim soil (JAZZ)**
Reads	118064	118064	117890
Scaffolds			3871 / 8984
Contigs (>1 reads) / reads	19236 / 47338	558 / 2744	4077 / 8984
Major contigs (≥ 10 reads) / reads	86 / 1521	20 / 290	11 / 192
	Sim AMD (Phrap)	**Sim AMD (Arachne)**	**Sim AMD (JAZZ)**
Reads	125652	125652	125513
Scaffolds			4352 / 44851
Contigs (>1 reads) / reads	12644 / 82693	4692 / 55022	8160 / 44846
Major contigs (≥ 10 reads) / reads	1980 / 50267	1372 / 39190	876 / 19719

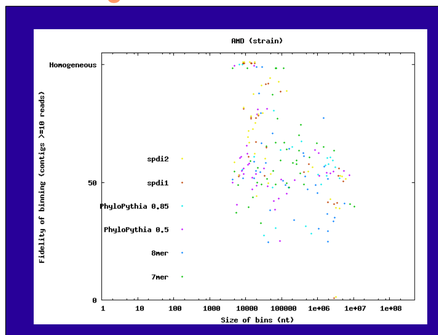


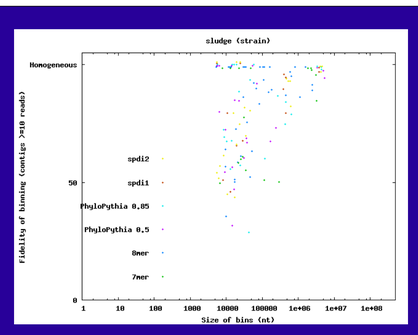


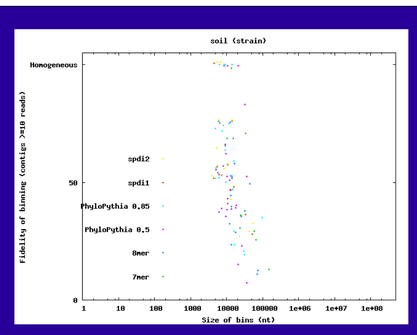


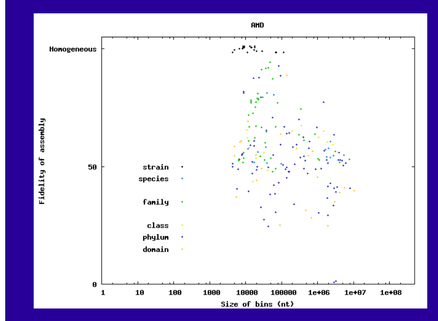
In AMD in which sequences from closely related organisms are in abundance all assemblers generated a large number of chimeric contigs that have sequences from different strains. In sludge the number of chimeric contigs was significantly lower. In soil the low sequence coverage resulted ot a low number of contigs in most cases missassembled. In all cases phrap appears to create the largest number of missassembled contigs, while Arachne is the most conservative.

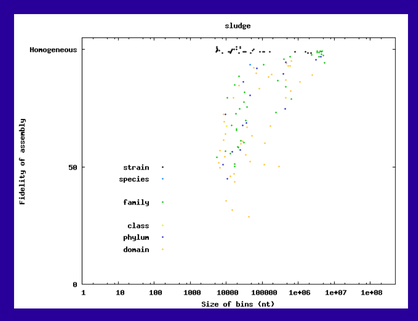
Binning

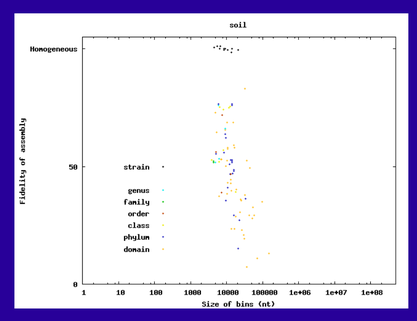








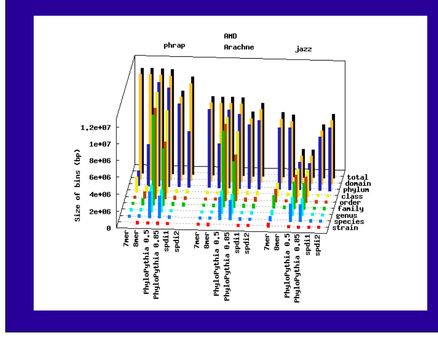


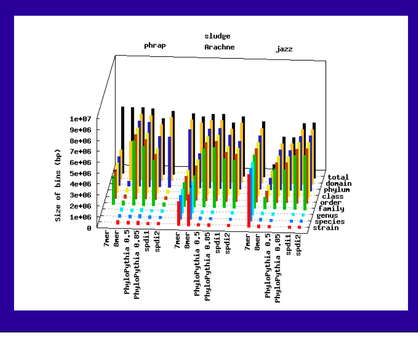


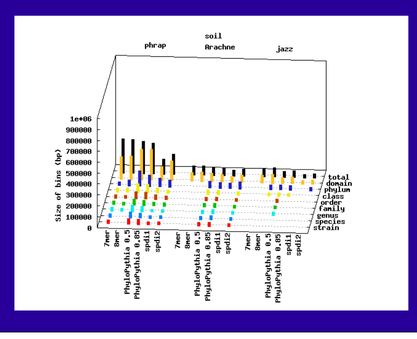
Binning of scaffolds and contigs was performed by three methods:

- PhyloPythia (Rigoutsos I. submitted for publication), a method based on the distribution of oligonucleotided patterns on contigs. We performed binning using two different cutoff values for the assignment (0,5, 0,85).
- Oligonucleotide frequency, irrespective of gene prediction. Two variants of an oligonucleotide frequency-based method were used; both of them were developed at the DOE Joint Institute using an algorithm similar to that implemented in TETRA [38]. One of the methods analyzes heptamer composition (NNNNNNN) of the contigs and another uses counts the frequencies of degenerate octamers (NN-NN-NN) (F. Korzeniewski, unpublished).
- SPDI was developed at the JGI (A. Salamov, unpublished). It is based on the analysis of phylogenetic distribution of the best BLAST hits of the predicted protein-coding sequences. First, all predicted proteins were compared to a database of all proteins encoded by 253 complete archaeal and bacterial genomes by BLASTp (the genomes of the dominant species in the simulated datasets were excluded from the comparison). Normalized BLAST score (bit score divided by self-score of a query sequence) has been calculated for all BLAST hits with e-value less than 1e-05 and a phylogenetic class with the highest total normalized score was assigned to a given sequence fragment, if (i) at least 50% of CDSs encoded by this fragment have hits in this class, and (ii) an average normalized BLAST score per CDS is greater than 0.2 (SPDI2) or greater than 0.1 (SPDI1).

Binning accuracy is generally lower in the AMD dataset, due to higher number missassembed contigs. PhyloPythia on average performs better above the level of family in the sludge and AMD datasets. Soil, due to the low number of contig and missassemblies is poorly binned.







This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.