Supplementary Methods

Description of the datasets. Random reads from 113 isolate microbial genomes present in the Integrated Microbial Genomes system v1.3 (IMG) [1] and sequenced at the DOE Joint Genome Institute were selected to form three distinct simulated metagenomic datasets (Supplementary Table 1). Reads corresponding to the ends of the same clone (pair reads) were selected where possible. The selected reads belonged to libraries of 3kb and 8kb.

Assembly. Phrap v3.57 (http://www.Phrap.org), Arachne [2] and JAZZ [3] genome assemblers were used. The parameters used for each tool were chosen based on prior experience and in order to make them more inclusive and less stringent. The parameters were as follows: for PHRAP: minmatch 30, maxmatch 55, minscore 55, max_subclone_size 50000, revise_greedy, vector_bound 20, for Arachne: recycle_bad_contigs=False, maxcliq1=500, maxcliq2=500, fast_draft_consensus=False, mc_min_overlap=30, and for JAZZ read depth penalties were turned off, as were the models assuming that the dataset contained a single primary constituent.

Gene calling. Genes were predicted with either fgenesb (www.softberry.com) or the combination of CRITICA [4] /GLIMMER [5] used by the Oak Ridge National Laboratory microbial pipeline [6]. Cluster of Orthologous Groups (COGs) [7] were calculated by comparison to the CDD database using reverse psi blast [8]. Hierarchical clustering was performed with the program CLUSTER [9] and visualization with TreeView [10].

Binning. Three methods for binning were used namely *k*mer, PhyloPythia, BLAST distribution. *k*mer calculated the oligonucleotide frequencies of all metagenomic sequences and compared to a reference set of finished genomes[11] from the Integrated Microbial Genomes system v1.3 (IMG)[1]. The metagenomic bin was assigned to the taxonomic family of the best matching isolate bin, using a chi-square measure. The comparison was performed on both strands of any sequence and the best match was chosen. The oligomers used were (a) of seven tandem nucleotides (NNNNNNN where N can be any nucleotide) and (b) of length of eight nucleotides following the pattern NNxNNxNN (where N can be any nucleotide and x is ignored).

PhyloPythia [12] was based on the comparison of sequence patterns between the metagenomic sequences and a reference set of genomes, resulting in assignment of sequence fragments to phylogenetic lades. Initially, sequences were binned using a classification based on a generic training set (untrained PhyloPythia) and bins were assigned to taxonomic levels ranging from domain to family. Subsequently, contigs containing single copy genes belonging to the most abundant species were used as training set (trained PhyloPythia), in this case bins were assigned to genera as well. In both occasions two groups of bins were created, the first with high p-value of 0.85 and the second with a lower p-value of 0.5.

BLAST distr was based on the classification of sequences based on the distribution of BLAST hits of predicted genes to taxonomic classes. Genes were predicted using fgenesb and compared to a database composed of protein sequences from 253 complete bacterial and archaeal genomes, downloaded from NCBI's ftp site. Homologs with E-value less than 1e-05 were assigned a normalized blast score (bit score of the blast hit divided by the bit score of the query against itself). A metagenomic sequence was assigned to the phylogenetic class with the highest total normalized score, if at least 50% of its predicted genes had hits in this class and an average normalized score per ORF was greater than 0.2. In all methods the most abundant genomes in the simulated datasets were excluded from the reference sets.

For every binning method the values of specificity (Sp) and sensitivity (Sn) were calculated as follows:

Sp = (True positive) / (True positive + False positive),

Sn = (True positive) / (True positive + False negative).

References

1. Markowitz, V.M. et al.. The integrated microbial genomes (IMG) system. *Nucleic Acids Res* **34**, D344-8 (2006).

2. Batzoglou, S. et al.. ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12**, 177-89 (2002).

3. Aparicio, S. et al.. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301-10 (2002).

4. Badger, J.H. & Olsen, G.J.. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* **16**, 512-24 (1999).

5. Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L.. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636-41 (1999).

6. Chain, P. et al.. Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph Nitrosomonas europaea. *J Bacteriol* **185**, 2759-73 (2003).

7. Tatusov, R.L. et al.. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

8. Altschul, S.F. et al.. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).

9. de Hoon, M.J.L., Imoto, S., Nolan, J. & Miyano, S.. Open source clustering software. *Bioinformatics* **20**, 1453-4 (2004).

10. Saldanha, A.J.. Java Treeview--extensible visualization of microarray data. *Bioinformatics* **20**, 3246-8 (2004).

11. Sandberg, R., Winberg, G., Branden, C.I., Kaske, A., Ernberg, I. & Coster, J.. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res* **11**, 1404-9 (2001).

12. McHardy, A., Martin, H., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I.. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**, 63-72 (2006).