Title
Annotation of metagenome short reads using proxygenes

Authors
Dalevi, Daniel
Ivanova, Natalia N.
Mavromatis, Konstantinos
et al.

Publication Date
2008-09-15

Annotation of Metagenome Short Reads Using Proxygenes

Daniel Dalevi[1], Natalia N. Ivanova[2], Konstantinos Mavromatis[2], Sean Hooper[2], Ernest Szeto[1], Philip Hugenholtz[3], Nikos C. Kyrpides[2], and Victor M. Markowitz[1,*]

[1]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[2]Genome Biology Program , DOE Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, CA 94598, USA
[3]Microbial Ecology Program, DOE Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, CA 94598, USA

ABSTRACT

Motivation: A typical metagenome dataset generated using a 454 pyrosequencing platform consists of short reads sampled from the collective genome of a microbial community. The amount of sequence in such datasets is usually insufficient for assembly, and traditional gene prediction cannot be applied to unassembled short reads. As a result, analysis of such datasets usually involves comparisons in terms of relative abundances of various protein families. The latter requires assignment of individual reads to protein families, which is hindered by the fact that short reads contain only a fragment, usually small, of a protein.

Results: We have considered the assignment of pyrosequencing reads to protein families directly using RPS-BLAST against COG and Pfam databases and indirectly via proxygenes that are identified using BLASTx searches against protein sequence databases. Using simulated metagenome datasets as benchmarks, we show that the proxygene method is more accurate than the direct assignment. We introduce a clustering method which significantly reduces the size of a metagenome dataset while maintaining a faithful representation of its functional and taxonomic content.

Contact: vmmarkowitz@lbl.gov

1 INTRODUCTION

The ultimate goal of metagenomic studies of a microbial community (microbiome) is to determine the systemic properties including genetics, metabolism, physiology and behavioral aspects of all community members, their interactions with various biotic and abiotic factors, transfer of energy and nutrients, and ecosystem dynamics. In practice, such comprehensive studies are seldom feasible and the scope of metagenomic analysis of most microbial communities is limited to genomic and metabolic reconstruction of the dominant population(s) including identification of key metabolic pathways likely to be present or absent in these populations. For most metagenome projects the amount of sequence data is insufficient for assembly and classification of sequences into different populations thus preventing even limited population-specific genomic and metabolic reconstruction. In these cases a gene-centric analysis using environmental gene tags (EGTs) is employed (Tringe et al. 2005). In this approach protein coding sequences (CDS) are identified in unassembled or partially assembled metagenomic sequences using an *ab initio* or evidence-based gene finder. These CDSs are further assigned to protein families, such as COGs (Tatusov et al. 1997), Pfams (Bateman et al. 2004) and TIGRfams (Selengut et al. 2007) and comparison of the relative abundance of protein families is performed. Proteins are assigned to families using reverse position-specific BLAST (RPS-BLAST) against position specific scoring matrices (PSSMs) of COGs in the CDD database (Marchler-Bauer et al. 2002) and enzyme-specific PSSMs in the PRIAM database (Claudel-Renard et al. 2003), and using hmmsearch against hidden Markov models (HMMs) in Pfam and TIGRfam databases. Alternatively, associations of proteins with functional subsystems can be achieved via BLAST searches against databases of annotated proteomes such as SEED (Overbeek et al. 2005).

The quality of annotations for metagenomic sequence data is lower than that of isolate microbial genomes due to higher rate of sequencing errors and data fragmentation. However identification of CDSs, their assignment to protein families and enumeration of representatives in metagenomes do not pose a problem even for completely unassembled reads generated by the Sanger sequencing platform, nor do they distort the functional or taxonomic profiles of the datasets. Such profiles may be distorted by the biases inherent to Sanger sequencing which involves cloning of metagenomic DNA into vectors, propagation of the vector within host bacteria and DNA amplification. The extent and the impact of such biases are largely unknown and therefore are difficult to account for in the downstream analysis. These problems and the relatively high cost of Sanger sequencing led to the increasing popularity of another variant of shotgun metagenome sequencing, which does not require cloning of environmental DNA and employs 454 Life Sciences pyrosequencing platform (Edwards et al. 2006). This type of sequencing raises another challenge to the downstream analysis: the depth of sequence generated by the pyrosequencing platform is usually insufficient for assembly, so the resulting metagenomes consist of individual unassembled reads. Furthermore, unlike Sanger sequencing which generates individual reads of 600-800 bp, each encoding a full-length protein or a significant portion thereof, pyrosequencing reads are 100 to 200 bp long and contain only a (usually small) fragment of a protein. As a result, traditional procedures for finding CDSs and assigning them to protein families cannot be applied to such sequences.

For protein family assignment of unassembled and/or short sequences, such as those generated by 454 platforms, two strategies can be envisioned: (1) ***direct*** assignment to protein families using translated read sequences for searches against family-specific PSSMs or HMMs; or (2) assignment via ***proxygene*** which we define as a full-length protein identified by a BLASTx search of read sequences against a protein sequence database and then used as a representative of a read or group of reads.

The perceived disadvantage of direct assignment of 454 reads to protein families is the low sensitivity of assignment in the case of

*To whom correspondence should be addressed.

RPS-BLAST, high computational demands in the case of hmmsearch and possible biases introduced by different degrees of sequence conservation within different protein families, which may explain why published metagenome studies followed a proxygene approach (Angly et al. 2006, Edwards et al. 2006, Thurnbaugh et al. 2006). These studies provided insufficient details about the methods employed for the selection of proxygenes (e.g., using best BLAST hit or multiple BLAST hits, resolution of functional annotation conflicts if more than one BLAST hit was used, etc.) or the reliability of the protein family assignment based on proxygenes.

In this paper we examine the reliability of direct and indirect assignment strategies using simulated metagenomic datasets created from pyrosequencing reads generated for isolate microbial genomes. We show that indirect assignment using proxygenes is more accurate than the direct method using RPS-BLAST. We also introduce a clustering method that reduces significantly the size of the derived datasets while maintaining the accuracy of functional and taxonomic assignments based on proxygenes. The reduction in size allows maintaining a compact yet comprehensive overview of the functional and taxonomic content of a metagenome.

2 METHODS

2.1 Simulated datasets

Reads from 22 genome projects, sequenced at the Joint Genome Institute (JGI) using the 454 GS20 pyrosequencing platform that produces ~100 bp reads, were selected and the genomes were split into three groups based on their phylogeny and the number of reads to ensure similar sizes for the simulated datasets. From each genome project, reads were sampled randomly at four different levels of coverage (0.1X, 1X, 2X and 4X per genome), resulting in a total of 12 simulated datasets (Table 1). The coverage is defined as the average number of times a nucleotide is sampled.

The position of each read on the assembled contigs was identified by BLASTn. Only the best hit of each read, with identity >95%, was kept and used to identify a position of the read with respect to the CDSs predicted on the assembled contigs using the JGI annotation and analysis pipeline. The nucleotide sequences of the genomes, the coordinates of the reference genes and their functional annotation were extracted from version 2.2 of the IMG database (http://img.jgi.doe.gov). At each level of coverage the CDSs overlapping the reads by more than 50nt comprised the reference gene set; the assignment of a read to a protein family was considered correct if it coincided with the family assignment of the gene from which the read has originated.

2.2 Assignment of 454 reads to protein families

We considered two ways of assigning reads to protein families: (1) direct assignment of the reads using RPS-BLAST against profiles of COGs and Pfams and (2) assignment via a proxygene. For direct assignment, translated RPS-BLAST search of reads against PSSMs in the CDD database was performed with an e-value cutoff in the range of 10^{-1} to 10^{-8} retaining the best hit only.

Proxygenes for 454 reads were found by BLASTx of the reads against the protein sequences in the IMG 2.2 database using e-value cutoffs in the range of 10^{-1} - 10^{-8}. Proxygenes were either assigned as the best BLASTx hit of a read (BH) or using a simple clustering method (see Figure 1). For the latter, the set of all reads $\{x_1, \ldots, x_N\}$ that have at least one hit below the cutoff have been clustered using the following algorithm:

1. Let $x=x_1$ and $i=1$.
2. Add x to group number G_i.
3. Extract the set of all proteins (A) that x has hits to, and add them to G_i.
4. For each protein p in A, extract all other reads $x_j, \ldots, x_M$ that have a best hit to p, and add them to G_i.
5. For each x in $\{ x_1, \ldots, x_M\}$ repeat step 2 until no more reads or proteins can be added to G_i.
6. Let x be the next unassigned read and let $i=i+1$, and repeat step 2.

This algorithm results in disjoint clusters (***proxy clusters***) in which no reads and no genes are members of more than one group. For each protein within the proxy cluster, the cumulative bit-score of its alignment to the reads within the same cluster is calculated. The protein with the highest cumulative bit-score is selected as a representative proxygene of this proxy cluster and is used for all further analyses, such as functional and taxonomic accuracy or determination of the overall functional profiles.

Most protein databases seem to be contaminated to some extent with rRNA sequences on which protein-coding genes have been predicted in different frames. Due to the high sequence conservation of rRNA genes, some of these "ghost" proteins are also conserved and even form "ghost" clusters which may contain proteins with no sequence similarity whatsoever and represent the same parts of rRNA sequences translated in different frames. Therefore before any protein family assignments of 454 reads were carried out, a filtering step has been introduced which involved BLASTn of the reads against an RNA database compiled of all rRNAs in IMG 2.2 in order to remove these reads.

3 RESULTS

454 reads can be associated with protein families by direct assignment or via proxygenes. Direct assignment compares the sequence of the read translated in 6 frames directly to the sequence profiles of protein families. Assignment via proxygenes is an indirect approach whereby BLASTx against a protein sequence database is used to identify a full-length protein ("proxygene") with high sequence similarity to the translated sequence of the 454 read. High sequence similarity between the read and the proxygene is considered as an indication that the read originated from a protein-coding gene which has high overall sequence similarity to the proxygene. Consequently, protein family membership and functional annotation of the read is considered to be the same as that of a full-length protein, which is used as a "proxy" of the 454 read in subsequent metagenome data analysis. Similarly, high sequence similarity between the read and the proxygene implies phylogenetic proximity of the organisms from which the read and the proxygene have originated, so that the full-length protein can be also used as a "proxy" of the read in assessing the taxonomic composition of the metagenome. However, the indirect approach may produce spurious hits to proteins that have little overall sequence similarity to the gene from which the read has originated. Accordingly, we have evaluated the accuracy of protein family assignments of the simulated datasets using direct and proxygene-based methods with several e-value cutoffs and assessed the accuracy of taxonomic assignments using both the BH and proxygene cluster approach.

It should be pointed out that although the simulated datasets used in this study faithfully reproduce some of the features of 454-sequenced metagenomes, such as the frequency and type of sequencing errors or variation (if any) of sequencing coverage, certain problems associated with processing of real metagenomes are hard to reproduce in a simulated environment. The main problem is the absence of a comprehensive collection of reference genomes; as a result only a small fraction of the genes in most metagenomic datasets generated to date are from organisms that have sequenced close relatives, thus limiting the detection of similarities between the short reads and reference genes. However, many of the genomes from which the 454 reads for the simulated datasets were selected belong to such over-sampled taxonomic groups as gamma- and betaproteobacteria (Table 1). In order to account for the potential errors resulting from a biased composition of reference databases and simulate the absence of close relatives of the sampled organisms, we followed the approach of (Mavromatis et al. 2007) and excluded all closely related genomes (either the same species or genus as the sampled genomes) from the reference database before carrying out BLASTx searches. The estimated sequence coverage is another unknown variable which may affect the results in the case of real metagenomes sequenced with 454 platform. For instance, the effect of resampling of a complex microbial community at very low sequence coverage is hard to estimate and it is possible that protein family composition and abundance will vary greatly from sample to sample. Similarly, comparison of completely unrelated microbial communities sampled at different coverage may result in virtually identical protein family abundance profiles. We attempted to address this problem by sampling the genomes included into each dataset at four different levels of coverage (0.1X, 1X, 2X, 4X) as described in the Methods section.

3.1 Evaluating accuracy of protein family assignment

In the first step we optimized the settings of RPS-BLASTx and BLASTx searches by using e-value cutoffs in the range of 10^{-1} to 10^{-8} and then estimated the accuracy of read assignments to COGs. The latter was calculated as the ratio of correct COG assignments (i.e. same COG assignment of the proxygene as that of the gene from which the read originated) over the total number of COG assignments. The results of this analysis (Figure 2) show that the direct assignment has invariably lower accuracy than the proxygene approach, with the exception of very low cutoffs for metagenome dataset M1 where direct approach performs as well as assignment via proxygenes (e.g. M1 at 4X in Figure 2). Most notably, the accuracy of COG assignment via proxygenes varied very little at different e-value cutoffs, with the percent of false assignments never exceeding 10% even at e-value of 10^{-1} (Figure 2). However, the percentage of reads assigned to COGs depends strongly on the cutoff and increases substantially at higher e-values (see Table 2). For example, about 39% of all reads in the dataset M3 were assigned to COGs at cutoff 10^{-1}, while at a more stringent cutoff of 10^{-5} used in previous studies (Angly et al. 2006, Edwards et al. 2006, Thurnbaugh et al. 2006) only 20% of the reads were assigned to COGs. This result is independent of the coverage, which is expected in the case of random sampling of reads. Since decreasing the e-value cutoff provides little reduction of the rate of false positive assignments while strongly affecting the overall number of reads assigned to COGs, the e-value cutoff 10^{-1} has been used in further analysis.

In addition to evaluating the e-value cutoffs, the effect of reference database composition was assessed by performing BLASTx searches against the reference database from which either the genomes of the same species or genomes of the same genus as the organisms used in the simulated datasets were removed. The effect of the reference database composition was most pronounced in the case of metagenome dataset M1 where removing all reference genomes belonging to the same genus as sampled organisms resulted in an error rate twice as high as that observed for the reference database with only same-species genomes removed. Conversely, removing all reference genomes of the same genus from the database had little if any effect on the accuracy of assignments for datasets M2 and M3. These results can be explained by the different taxonomic composition of metagenome datasets M2 and M3 as compared to dataset M1 (see Table 1): M1 has been sampled mostly from the representatives of Firmicutes, while M2 and M3 are composed almost exclusively of Proteobacteria, a phylum with more sequenced representatives than all other bacterial phyla combined. Even in the absence of the closest relatives, these genomes provide a comparative context rich enough for highly accurate assignment of reads.

The taxonomic composition of the simulated metagenome datasets also affected the percentage of reads assigned to COGs: at low e-values metagenome datasets M2 and M3 had twice as many reads assigned to COGs as dataset M1, although these differences were less prominent at higher e-values (Table 2). Furthermore, while the accuracy of direct assignments was essentially the same for all datasets at a given e-value, the accuracy of assignment via proxygenes was much higher for metagenome dataset M3 as compared to M1 at the same e-value, sequence coverage and reference database composition. Note that although reference databases contain significantly fewer genomes of Firmicutes than Proteobacteria, there are many phyla with even less sequenced representatives. It is expected that for the metagenomes composed of the members of such poorly sampled phyla the accuracy of proxygene-based assignment will be even lower. Thus proxygene-based comparisons of the metagenomes with vastly different taxonomic composition (e. g., those dominated by proteobacteria against those composed mostly of planctomycetes or chloroflexi) should be treated with caution, since several-fold differences in the accuracy of protein family assignments may result in gross errors in data interpretation. Our results emphasize the importance of a good reference database for the analysis of 454 data and indicate that although the availability of same-species reference sequences is highly desirable, it may be unnecessary as long as sequences of multiple and diverse representatives of the same phylum are present in the database.

3.2 Proxygene clustering

While the best BLAST hit (BH) is the simplest and most direct method of selecting a proxygene, this approach may result in a high level of redundancy: several reads may be associated with the same proxygene (see Figure 1.a), therefore there is no need to consider them as separate entities. Moreover, due to the presence of many closely related genomes in the reference databases, the read may have hits of nearly the same strength to several highly similar genes of which only one is chosen as a proxygene. Alternatively,

the reads originating from the same gene may become associated with different, but closely related proxygenes (see Figure 1.b). In terms of their protein family membership and functional annotation, all such proxygenes are equivalent and should be handled as one entity. Finally, treating each read-proxygene pair separately results in very large datasets, hardly amenable to any manual analysis by biologists and posing serious data management scalability problems.

In order to address these problems, we have developed a simple clustering algorithm for grouping the reads and proxygenes, as illustrated in Figure 1.c and described in the Methods section. This proxygene clustering provides a significant reduction in the size of the resulting datasets. Figure 3 shows a comparison between the number of proxygenes with and without proxygene clustering: the reduction is about 1.2 to 1.5 times at 0.1X coverage (BLAST evalue cutoff = 10^{-5}; removing same-genus genomes), whereas at 4X it is about 7 times for dataset M1, 10 times for M2 and 10 times for M3 (Table 2). This reduction is significant in light of the rapidly increasing number and size of metagenome datasets.

3.3 Taxonomic assignment of reads via proxygenes

In addition to assessing the functional content of various microbial communities, most metagenomic studies attempt to determine and compare the taxonomic composition of the samples. For metagenomic datasets generated with the pyrosequencing platform this question can be addressed by a proxygene-based approach, using the phylogenetic distribution of proxygenes as an estimate of the phylogenetic composition of a sample. Similarly, a proxygene cluster-based approach can be used to estimate the taxonomic composition of a sample, whereby the taxonomic identity of all reads assigned to a proxygene cluster is considered either the same as the representative proxygene (an approach used in this study) or as that of the lowest taxonomic group to which all proxygenes in the proxygene cluster belong.

Using the simulated data sets, we have examined the accuracy of the taxonomic assignment of reads using proxygene and proxygene cluster approaches. The accuracy of the assignment was measured as the fraction of true positives at different taxonomic levels (domain, phylum, class, order and family). As expected, the accuracy of assignment at the domain and phylum level is much higher as compared to the level of order and family with domain-level assignments being 100% accurate and the fraction of accurate family-level assignments varying from 20 to 60% for different metagenomes and different reference databases. The accuracy of taxonomic assignments at the phylum level reaches more than 90% for datasets M2 and M3, while the accuracy of assignments for M1 is only 60% at the same level. Similar to the accuracy of protein family assignments, this disparity appears to reflect the difference in taxonomic composition of the three simulated datasets, with M1 composed of representatives of less well-sampled phyla than M2 and M3.

At low sequence coverage (0.1X) the proxygenes and proxygene clusters are almost identical since most proxygene clusters contain only one or two reads. However, at higher (4X) coverage the clustering of the reads into proxygene clusters does not decrease the accuracy and in some cases it even improves the assignment, especially at higher taxonomic levels (phylum, class), which are most frequently used for estimation of the taxonomic composition of metagenomic samples. This result indicates that the reads are grouped into essentially consistent taxonomic clusters and selection of one proxygene as a representative of multiple reads effectively screens out some of the spurious hits that would adversely affect the accuracy of taxonomic assignments.

3.4 Hierarchical clustering against a reference

The relative frequencies of COGs and PFAMs are often used to compare metagenomic datasets obtained from different environments and detect functions that are over or under-represented (Tringe et al. 2005). Such analyses depend on an unbiased identification of COGs and PFAMs and comparable accuracy of their detection in different datasets irrespective of the taxonomic composition, population structure and variation in sequence coverage of microbial communities. Our results indicate that although the accuracy of protein family assignment is fairly high, it may vary greatly between metagenome datasets depending on their taxonomic composition and sequence coverage. Such variations may influence the results of gene-centric analysis performed on the individual protein families and even on their groupings, such as COG Pathways and Functional Categories. However, it is not clear whether these variations in accuracy could change the overall functional profiles of the metagenomes so severely as to affect the results of their profile-based clustering.

In order to address these questions, we performed hierarchical clustering of the datasets based on the relative frequencies of COGs produced by direct assignment, proxygenes and proxygene clusters (Figure 4). The placement of absolute (i.e. COG frequencies of all proteins from sampled isolate genomes) and sampled references (i.e. COG frequencies of the sampled genes) for all metagenome datasets shows that 454 sequencing indeed has little bias in terms of under- or over- sampling certain genomic regions. Furthermore, there is little difference between the profiles obtained by proxygene and proxygene cluster approaches indicating that there is little loss or gain of functional information with proxygene clustering. The error introduced by annotation in every case is nevertheless high since none of the metagenome datasets ends up in a cluster with the sample and absolute reference. As expected, 4X and 2X sampling references are closer to the absolute references than 1X and especially 0.1X. In addition, there is significant difference between the direct and proxygene-base profiles, since the profiles obtained by direct assignment mostly clustered together.

These results indicate that, although metagenome datasets generated by pyrosequencing platforms may indeed represent an unbiased sample of community DNA, significant biases can be introduced by subsequent processing of the data. These biases are mostly due to the skewed composition of reference databases and, depending on the taxonomic composition and population structure of the sample, they may be as difficult to account for as cloning biases of Sanger technology. While the accuracy of protein family assignments was sufficient to separate the three simulated metagenome datasets discussed in this paper, a similar separation may not be possible for real environmental samples that may be characterized by large disparity in sequence coverage due to different evenness and abundance of species distribution and considerable variation of the taxonomic composition.

4 DISCUSSION

We have compared methods for the annotation of short reads in metagenome datasets using benchmark datasets that model faithfully the main features of real datasets. While the proxygene based method is generally more accurate, its efficiency depends on the composition of the reference database. Thus, the metagenome datasets containing representatives of over-sampled phyla were annotated more efficiently. The accuracy of assignments did not increase significantly at lower e-value cutoffs, while selection of less stringent cutoffs (10^{-1}) allowed assignment of twice as many reads without increasing the rate of false positive assignments. We have also shown that the proxygene clustering has the important advantage of reducing substantially the size of metagenome datasets, while preserving faithfully their functional and taxonomic content.

Despite the increase of the average read length produced by newer generation of 454 sequencing platforms, such as GS FLX (~200 bp reads), it is expected that many metagenome datasets will remain unassembled due to prohibitively high amount of sequence data necessary to ensure even modest degree of assembly for all but the simplest microbial communities. Consequently, it is likely that gene-centric analysis will remain the method of choice for the analysis of many metagenomes and therefore the proxygene cluster based annotation presented in this paper has standing practical significance. We have applied this method to several metagenome datasets from ongoing metagenome studies (such as the PT3 and PT6 datasets listed in Table 2) that have been included into the IMG/M system (Markowitz et al. 2008). As soon as their analysis is completed and published, these datasets will be released as part of IMG/M's public version (http://img.jgi.doe.gov/m).

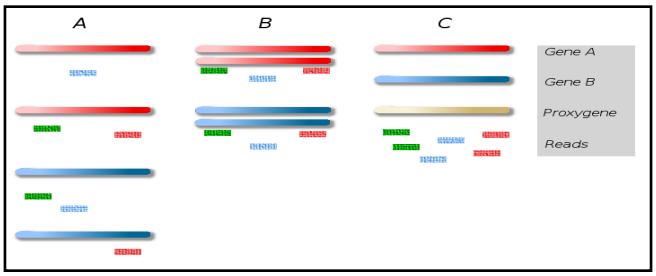

Fig. 1. Assigning reads to full-length (proxy) genes in a database: (a) each read is assigned to a separate proxygene by best BLAST hit: a read may be assigned to several identical proxygenes; (b) grouping identical proxygenes: a read may be assigned to several proxygenes; (c) proxygene-clustering: each read is assigned to a single proxygene.

FUNDING

The work presented in this paper was supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

Angly, F. et al. (2006) The marine viromes of four oceanic regions. *PLoS Biol* 4:11.

Bateman, A. et al. (2004) The Pfam protein families database. *Nucleic Acids Res* **32**, D138-41.

Claudel-Renard, C. et al. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Research* **31**, 6633-6639.

Edwards, R. A. et al. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, 57.

Markowitz, V. M. .et al. (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* **36**, D534-538.

Marchler-Bauer, A. et al. (2002) CDD: A Database of conserved domain alignments with links todomain three-dimensional structure. *Nucleic Acids Res* **30**, 281-283.

Mavromatis, K. et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods* **4**(6), 495-500.

Overbeek, R. et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**, 5691-702.

Selengut, J.D., Haft, et al. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes *Nucleic Acids Res* 35, D260-D264.

Tatusov, R. L. et al. (1997) A genomic perspective on protein families. *Science* **278**, 631-637.

Tringe, S. G. et al. (2005) Comparative metagenomics of microbial communities. *Science* **308**, 554-7.

Turnbaugh, P. J. et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027-31.

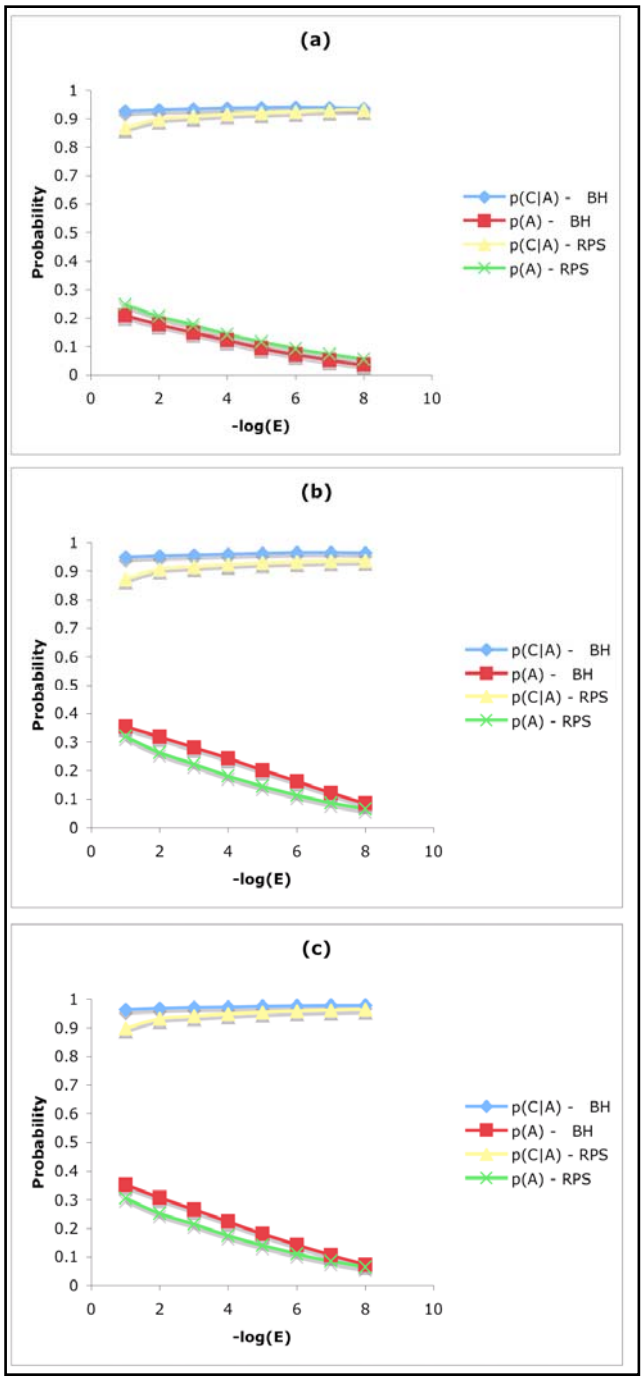

Fig. 2. Indirect annotation of COGs is compared to annotation using BH-proxygenes for three simulated datasets at 4X coverage (Table 1): (a) M1, (b) M2 and (c) M3. We removed all reference genomes that belong to the same species and/or genera as genomes used to create the simulated metagenomes before the BLASTx step. P(A) is the probability that a read is assigned to a COG and P(C|A) is the probability that an assigned COG is correct conditioned on the event that a COG has been assigned..

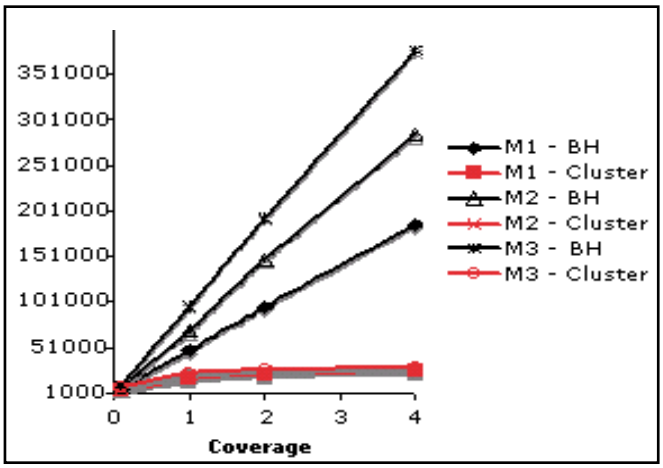

Fig. 3. The number of proxygenes is significantly reduced for all levels of coverage using the clustering approach. The number of proxygenes for the BH approach is shown in black for the three simulated datasets (M1, M2 and M3) at coverage 0.1X, 1X, 2X and 4X. The red lines show the number of proxygenes after clustering.

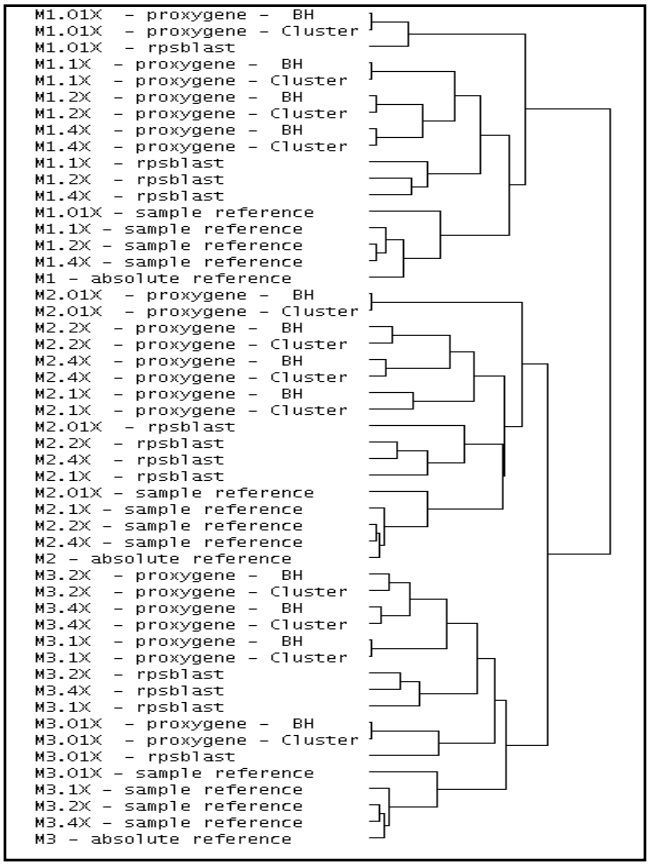

Fig. 4. Hierarchical clustering of relative COG frequencies of the three simulated metagenome datasets (M1, M2 and M3) at different levels of coverage. The absolute reference consists of the relative frequencies as they occur in the isolate genomes while the sample reference is the relative frequency of the reads that were sampled. The best BLAST hit proxygene (BH) and the proxygenes defined by the clustering approach (Cluster) are shown together with the direct annotation using RPS-BLAST (rpsblast).

Table 1. Genomes sampled for the simulated metagenome datasets. The size of each genome and total number or reads sampled for each dataset is shown. M1 = Metagenome dataset 1, M2 = Metagenome dataset 2 and M3 = Metagenome dataset 3.

		Genome	Reads sampled			
Dataset	Organism	size (bp)	0.1X	1X	2X	4X
M1	*Clostridium phytofermentans* ISDg	4533512	4638	46379	92756	185498
	Prochlorococcus marinus NATL2A	1842899	1866	18681	37360	74720
	Lactobacillus reuteri 100-23	2174299	2371	23710	47419	85352
	Caldicellulosiruptor saccharolyticus DSM 8903	2970275	2950	29496	58992	111422
	Clostridium sp. OhILAs	2997608	2934	29348	58697	117398
	Herpetosiphon aurantiacus ATCC 23779	6605151	6937	69387	138775	277553
	Bacillus weihenstephanensis KBAB4	5602503	4158	45463	91109	175869
	Halothermothrix orenii H168	2578146	2698	26980	53965	104554
	Clostridium cellulolyticum H10	3958683	3978	39802	79605	159206
M2	*Geobacter sp*. FRC-32	3982463	4225	42266	84525	158487
	Burkholderia multivorans ATCC 17616	6979389	7110	71074	142102	284221
	Delftia acidovorans SPH-1	6702581	7046	70448	140916	267735
	Comamonas testosteroni KF-1	5906374	6189	61895	123794	237264
	Geobacter lovleyi SZ	3871860	4300	43004	86009	153584
M3	*Shewanella putrefaciens* CN-32	4659220	4714	47151	94318	188633
	Shewanella loihica PV-4	4602594	4588	45882	91773	183536
	Halorhodospira halophila SL1	2678452	2690	26898	53796	110282
	Pseudomonas putida F1	5959964	6407	64080	128158	238005
	Shewanella baltica OS195	5310173	5378	53779	107548	215103
	Bifidobacterium longum bv. Infantis ATCC 15697	2832748	2898	28990	57981	112343
	Stenotrophomonas maltophilia R551-3	4544233	4685	46844	93699	179581
	Parvibaculum lavamentivorans DS-1	3854587	4501	39379	78764	157526

Table 2: Percentage of assigned reads together with the degree of reduction obtained using proxygene-clustering as opposed to BBH-proxygene. PT3 and PT6 are datasets from lean and obese mouse gut metagenome datasets generated using 454 (GS20) platform.

	Percentage assigned / Times reduction					
	BLAST e-value cutoff					
Dataset	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
PT3	17 / 4.0	13 / 3.9	10 / 3.8	7.7 / 3.5	5.6 / 3.3	3.9 / 3.0
PT6	19 /3.4	15 / 3.0	12 / 3.3	9.5 / 3.2	7.1 / 3.1	5.1 / 3.0
M1 (0.1X)	26 / 1.4	22 / 1.4	18 / 1.3	14 / 1.3	12 / 1.2	8.7 / 1.2
M1 (1X)	26 / 4.0	22 / 3.7	18 / 3.4	15 / 3.1	11 / 2.8	8.6 / 2.4
M1 (2X)	26 / 6.6	22 / 6.0	18 / 5.4	15 / 4.8	11 / 4.2	8.5 / 3.6
M1 (4X)	26 / 11	22 / 10	18 / 9.2	15 / 8.0	11 / 6.8	8.5 / 5.6
M2 (0.1X)	41 / 1.5	36 / 1.4	31 / 1.4	27 / 1.3	23 / 1.3	18 / 1.3
M2 (1X)	41 / 4.7	36 / 4.5	32 / 4.0	28 / 3.8	23 / 3.5	18 / 3.1
M2 (2X)	40 / 8.2	36 / 7.7	32 / 7.1	27 / 6.5	23 / 5.9	18 / 5.1
M2 (4X)	41 / 15	36 / 15	32 / 13	28 / 12	23 / 10	18 / 8.7
M3 (0.1X)	39 / 1.6	34 / 1.5	29 / 1.5	25 / 1.4	20 / 1.4	15 / 1.3
M3 (1X)	39 / 5.1	34 / 4.8	29 / 4.4	24 / 4.0	20 / 3.6	15 / 3.1
M3 (2X)	39 / 8.9	34 / 8.3	29 / 7.5	25 / 6.7	20 / 5.9	15 / 5.0
M3 (4X)	39 / 17	34 / 15	29 / 14	24 / 12	20 / 10	15 / 8.5