Use of simulated data sets to evaluate the fidelity of metagenomic processing methods [AU: Title shortened (our limit is 13 words). OK as edited?]

Konstantinos Mavromatis[1], Natalia Ivanova[1], Kerrie Barry[1], Harris Shapiro[1], Eugene Goltsman[1], Alice C McHardy[2], Isidore Rigoutsos[2], Asaf Salamov[1], Frank Korzeniewski[1,3], Miriam Land[3], Alla Lapidus[1], Igor Grigoriev[1], Paul Richardson[1], Philip Hugenholtz[1] & Nikos C Kyrpides[1]

[1]Department of Energy Joint Genome Institute (DOE-JGI). 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [2]IBM T.J. Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, New York 10598, USA. [3]Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. Correspondence should be addressed to K.M. (KMavrommatis@lbl.gov).

RECEIVED 4 DECEMBER 2006; ACCEPTED XX XX 2007; PUBLISHED ONLINE XX XXXXXX 2007; DOI:10.1038/NMETHXXX

Metagenomics is a rapidly emerging field of research for studying microbial communities. To evaluate methods presently used to process metagenomic sequences, we constructed three simulated data sets of varying complexity by combining sequencing reads randomly selected from 113 isolate genomes. These data sets were designed to model real metagenomes in terms of complexity and phylogenetic composition. We assembled sampled reads using three commonly used genome assemblers (Phrap, Arachne and JAZZ), and predicted genes using two popular gene finding pipelines (fgenesb and CRITICA/GLIMMER). The phylogenetic origins of the assembled contigs were predicted using one sequence similarity–based (blast hit distribution) and two sequence composition–based (PhyloPythia, oligonucleotide frequencies) binning methods. We explored the effects of the simulated community structure and method combinations on the fidelity of each processing step by comparison to the corresponding isolate genomes. The simulated data sets are available online to facilitate standardized benchmarking of tools for metagenomic analysis.

Recent advances in sequencing technology are mediating a transition from organismal to community genomics (metagenomics), allowing us to directly examine the molecular blueprints of microbial communities. Metagenomic data processing follows essentially the same steps as processing of shotgun sequences generated for isolate genomes, starting with the assembly of sequence reads, followed by gene prediction and functional annotation. The genome sequence of an isolated microorganism, however, is typically derived from a clonal population, whereas metagenomic projects sample the genomes of multiple species and strains present in highly variable abundance in a microbial community. Quality control steps that detect assembly and gene-finding errors, such as finishing and gap closure, or manual curation of genes and functions, are mostly omitted. The final output of this process is similar to that

generated for draft isolate genomes and includes scaffolds, contigs and unassembled reads. An additional step of assigning scaffolds and contigs to phylogenetically related groups, called binning, is necessary because multiple species are present in the data set. This can range from coarse-level groupings such as domain (bacteria, archaea) down to fine-level groupings such as individual strains of a given species, depending on the binning method, community structure as well as sequencing quality and depth.

To our knowledge, no metagenome-specific assemblers are yet available, and methods developed for isolate genomes have been used with parameter modifications, such as JAZZ[1], Celera Assembler[2] and Phrap[3]. Owing to the nature of the samples and the algorithms used, assembly can produce chimeric contigs and scaffolds comprising reads from different organisms. Strain heterogeneity can add considerably to this problem as the probability of co-assembly increases with closely related genomes[4]. Furthermore, it is difficult to distinguish whether the chimerism is the result of assembly or is natural, owing to homologous recombination[1].

Accuracy of gene prediction methods, originally developed for finding genes in isolate microbial genomes, is impaired by shorter average sequence fragment length and a higher frequency of sequencing errors, which leads to inevitable fragmentation of genes. Additionally, the chimeric nature of assembled metagenomic sequences enhances the problem. Methods for *ab initio* gene prediction usually rely on identification of oligonucleotide composition and codon usage in the sequence fragments to predict coding regions[5,6]. Depending on the phylogenetic heterogeneity of the organisms in the environmental sample, however, the oligonucleotide composition and codon usage preference of the different contigs can be quite different. This makes it difficult for most gene prediction algorithms to produce accurate models of protein-coding regions. Evidence-based methods that rely on the similarity of new sequences to genes in the database may also fail to identify genes when there are no sequenced homologs, or when the predicted genes contain chimeric or shuffled sequences. Owing to these limitations, the quality of gene annotations is likely to be reduced.

Binning methods fall into two main categories: sequence composition–based and sequence similarity–based. Composition-based classifiers distinguish genomes from one another by intrinsic features of the sequence such as oligonucleotide frequencies caused by codon usage or restriction-site frequencies[7–9]. Short sequences are difficult to classify by this approach because of insufficient signal[9]. Two composition-based classifiers have been applied to metagenomic data sets, TETRA[9] and PhyloPythia[10]. Similarity-based methods assign metagenomic fragments to their closest phylogenetic neighbor based on coding-sequence identity[3]. This approach is entirely dependent on the availability of reference sequences that are related to the species present in the microbial community under study.

Unfortunately, owing to the cultivation bias, the phylogenetic representation of naturally occurring microbial species is very poor[11], and similarity-based binning has limited resolution.

A fundamental problem common to all methods applied to metagenomic data sets is the inability to quantify error rates since the 'correct' solution is not known. The rapid accumulation of metagenomic data sets[12] makes the need for benchmarking such methods even more pressing. To address this need, we created three data sets of varying complexity, and benchmarked several commonly used assembly, gene prediction and binning methods against them.

RESULTS

Simulated data sets

We constructed three simulated metagenomic data sets of varying complexity by combining sequencing reads randomly selected from 113 isolate genomes sequenced at the Department of Energy Joint Genome Institute (DOE-JGI) and available through IMG[13]. This approach allows the incorporation of real sequencing and sequence-dependent processing errors and does not rely on their simulation. We designed the first data set (simLC) to simulate low-complexity communities dominated by a single near-clonal population flanked by low-abundance ones. These types of data sets result in a near-complete draft assembly of the dominant population, as seen, for example, in bioreactor communities[3,14] (**Supplementary Table 1** online).

We designed the second data set (simMC) to resemble moderately complex communities with more than one dominant population, also flanked by low-abundance ones, as has been observed in an acid mine drainage biofilm[1] and *Olavius algarvensis* symbionts[15] (**Supplementary Table 1**). These types of communities usually result in substantial assembly of the dominant populations according to their clonality. The third data set (simHC) simulates high-complexity communities lacking dominant populations, such as agricultural soil[16], where no dominant strain is present, and typically results in minimal assembly.

Assembly

We assembled the simulated data sets using three commonly used programs at the DOE-JGI; Phrap v3.57 (see URL in Methods), Arachne v.2 (ref. 17) and JAZZ[18] (**Supplementary Methods** online).

First we investigated the degree of assembly (**Table 1**). Phrap incorporated more reads and produced the largest number of contigs in all three meta-assemblies. This trend is most obvious in the case of the simHC data set, where Phrap assembled 40% of the reads compared to only 2.4% and 2.3% by JAZZ and Arachne, respectively. Only a small percentage of the reads assembled by Phrap, however,

were incorporated into major contigs (≥10 reads), as expected for a community with no dominant populations. In contrast, the assemblers had comparable performance for the simLC data set, which most closely resembled an isolate genome. As data-set complexity increased, JAZZ exhibited a considerable drop in degree of assembly (**Table 1**).

We then assessed contig chimerism, caused by coassembly of reads from two or more isolate genomes, by reference to the known origin of individual reads (**Fig. 1 and Supplementary Figure 1**). We assigned contigs to the phylogenetic group that contributed the majority of constituent reads at each taxonomic level from strain to domain, and we defined the degree of contig chimerism at each taxonomic level as the percentage of reads not belonging to the major phylogenetic group.

Encouragingly, the majority of contigs comprised reads belonging to a single isolate genome for all assembly methods (**Fig. 1a**,**b**). Chimerism at progressively higher taxonomic ranks indicated that Arachne produced the highest proportion of accurate contigs, followed by JAZZ and Phrap (**Fig. 1a**,**b**). In absolute terms, Phrap produced the highest number of homogeneous contigs (**Table 1**), but the inability to distinguish these from chimeric contigs makes the other assemblies, with lower chimeric fractions, more reliable. We attribute the generation of chimeric contigs to the presence of ubiquitous sequences (for example, transposases) and low-quality sequencing mainly at the end of the reads (data not shown).

The percentage of chimeric contigs in the simLC assembly was small, and the taxonomic level at which read homogeneity was achieved varied (**Fig. 1c**). By contrast, most simMC chimeric contigs were the result of coassembly of strains belonging to the same species (**Fig. 1c**). We anticipated this as the simMC data set comprised closely related dominant populations, and this highlights the fact that none of the assemblers could effectively discriminate sequences belonging to strains of the same species (**Fig. 1a**). Chimerism was distributed randomly at all taxonomic levels among the simHC contigs, reflecting the absence of a large number of sequences from the same organism. For all assembly methods and all data sets, the degree of chimerism was most pronounced in contigs larger than 8 kb (**Fig. 1b**,**c**). Therefore, employing these assemblers for highly complex community data sets, such as simHC, has limited value since most contigs are below 8 kb and will provide misleading information.

Gene prediction

We applied the two gene prediction pipelines that have been previously used for the DOE-JGI metagenomic projects, to each of the assemblies resulting in 18 sets of identified genes. The first, fgenesb (see URL in Methods), was used to predict coding sequences both on the assembled contigs and the unassembled reads. The second pipeline used a combination of CRITICA and GLIMMER (CG), which is also used to predict genes in all the isolate genomes sequenced at the DOE-JGI[19].

To evaluate the accuracy of each pipeline, we compared the genes identified on the simulated data sets to the genes originally predicted on the corresponding reads of the isolate genomes (reference genes) using blastp[20], and categorized them into four groups. The first comprised genes common to both data sets (correctly identified genes). In this group we included only genes identified on the same sequence reads, with >80% amino acid identity over 50% of the shortest gene length. Genes falling below these thresholds formed the second group (inaccurately predicted genes).The third group contained genes predicted in the simulated data sets with no corresponding reference gene (newly predicted genes). Finally, reference genes without a corresponding predicted gene in the simulated data set formed the group of missed genes. We expected reference genes represented by <90 bp in the meta-assemblies to be missed by blastp owing to their length and were excluded from the comparisons. These comprised <7.5% of all identified genes in simLC and simMC and <10% in simHC.

Fgenesb correctly identified 10–30% more reference genes on the contigs than the CG pipeline in every data set (**Fig. 2a**). Both pipelines called 7–15% of the genes inaccurately (**Fig. 2a**), hence the difference in correctly called genes betweenthe two pipelines is due to CG missing a greater proportion of reference genes, mainly located on small contigs (data not shown). Additionally, 1–10% of the genes were newly predicted. The effect of assembly quality was striking in the gene-prediction results. In all cases, the higher quality Arachne assemblies resulted in more accurate gene predictions (higher correctly and lower inaccurately identified genes) than the other two assemblers (**Fig. 2a**).

We also evaluated the accuracyof the gene calls on unassembled reads (where low-abundance species are usually represented). Fgenesb correctly identified ~70% and missed ~20% of reference genes on unassembled reads in all data sets (**Fig. 2b**). The remaining 10% of reference genes were inaccurately called and another ~8% were newly predicted. Notably, the contribution of assembly to accurate gene prediction was not more than 20%, whereas its effect on the missed and inaccurately predicted genes were only slightly higher. The CG pipeline exhibited poor results (7% accurately predicted, 85% missed and 8% inaccurately predicted genes), and we did not use these data for the following steps of the analysis.

Although some of the inaccurate or new genes could be real (that is, they were either miscalled or missed in the analysis of the original isolate genomes), it is more likely they represent gene prediction errors that can be attributed to the following factors. In the case of fgenesb, the use of gene modeling parameters of one 'generic' microorganism cannot describe the diversity observed in communities, especially in complex communities such as simHC. Contig chimerism exacerbates this problem. For the CG pipeline, the low percentage of accurately predicted genes could be explained by CRITICA's low sensitivity, which uses only sequence similarity and di-codon frequencies as a measure of coding

probability. Therefore it would be strongly affected by the heterogeneity of the metagenomic sequence fragments and the absence of similar sequences in the database. Furthermore, both pipelines are affected by the presence of low-quality sequences (especially in singlets and contigs of low coverage) with errors (for example, frameshifts).

Based on the simulated data sets, we predict that approximately 10–20% of genes called in metagenomic data sets (that have not been manually curated) are inaccurate. This noise will contribute to the contamination of sequence databases with 'ghost' genes and may lead to the generation of inaccurate community metabolic models.

Gene function prediction

Gene-centric or environmental gene tag (EGT) analysis is a recently proposed method for comparing the metabolic potential of microbial communities[21,22]. This approach is especially useful for highly fragmented metagenomic data sets, and is dependent on the accuracy of gene calling and annotation. To assess the effect of annotation accuracy on gene-centric analyses, we compared the annotations of the simulated data sets to those of the reference genes (reference annotation). We also assessed the effect of excluding singlet annotations from the EGT profiles because they have been omitted in previous studies[1,3]. We chose the widely used Clusters of Orthologous Groups (COG) classification[23] as the basis for comparison. Functional prediction profiles for each data set using fgenesb and the CG pipeline were compared to each other using hierarchical clustering (**Fig. 3**).

The distance between the reference and fgenesb annotations reflects gene-calling errors caused by the inaccurate gene models and much shorter average sequence-fragment length. We estimate that these errors resulted in 5–20% of COGs having misrepresented frequencies (both over- and underrepresented; **Fig. 3**). We attribute the small distances between the fgenesb annotations for each data set to the fact that the majority of genes are called on singlets with a uniform error rate (**Fig. 2b**), resulting in profiles essentially assembly independent.

Excluding singlet annotations produced pronounced differences in EGT profiles of the same data sets, and in all cases profiles based on genes predicted on contigs and singlets were more similar to the reference set than to the genes predicted by any of the methods on contigs alone (**Fig. 3**). We attribute this to the fact that the majority of the genes are called on singlets for all data sets (**Fig. 2**), which will likely be the case for most real metagenomic data sets. Therefore, it is critical to include singlet annotations in EGT calculations.

Binning

We binned assembled sequences using three different methods previously used at the DOE-JGI. These include two sequence composition–based methods (PhyloPythia[10] and *k*mers), and a sequence similarity–based approach (BLAST distr). *k*mers and Blast distr were developed at the DOE-JGI (**Supplementary Methods**). All three methods assign contigs to phylogenetic groups by comparison to third-party data (isolate genomes), but the potential binning resolution differs. In specific, BLAST methods classified contigs only to the predefined rank of Class, *k*mer to Family and PhyloPythia to varying ranks from domain to genus. It has already been shown that the fidelity of composition-based binning declines with decreasing fragment length[9,10], and chimerism will very likely reduce binning fidelity. Indeed, binning of short sequences (<8 Kb in simLC and simMC, and the entire simHC) resulted in low-quality bins (**Supplementary Table 2** online), and thus we excluded them from subsequent analysis.

For benchmarking binning accuracy, we determined the reference identity of each bin as the lowest taxonomic rank to which all contigs belong. For example, if a bin comprised contigs assembled from the genomes of *Rhodopseudomonas palustris* (a member of the class Alphaproteobacteria), and *Xylella fastidiosa* (a member of the class Gammaproteobacteria) its reference identity would be the phylum proteobacteria because this is the lowest rank to which all contigs belong. If one or more of the contigs in a bin were chimeric, we used their phylogenetic identity, based on the majority read composition, in the reference identity calculation.

We used PhyloPythia both in a sample-specific and a generic mode (**Supplementary Methods**), and in both cases it performed better than the other two methods, as it typically exhibited higher specificity values (**Fig. 4** and **Supplementary Table 2**). Training PhyloPythia on contigs belonging to individual dominant community members provided higher resolution binning for those organisms than runs based on the generic model (**Supplementary Table 2**). However, it also resulted in slightly lower specificity values because by attempting to bin contigs to its training rank of genus, it included small amounts of sequence data from more distantly related organisms.

In most cases, both BLAST distr and *k*mer performed poorly, as evidenced by their low average specificity values and high s.d. (**Supplementary Table 2**). BLAST distr is dependent on the availability of closely related reference genomes, which are frequently absent[11]. Even when a closely related genome is available, the variation in genomic content between similar organisms may result in the absence of corresponding genes in the reference genome. Notably, *k*mer, although failing to assign bins to the correct taxonomic group, did produce phylogenetically coherent clusters at the rank of order and above.

To determine which combinations of tools best approximate the true population structure, we calculated the relative abundance of the two dominant populations in the binned data (that is, Alpha and

Gamma proteobacteria for both simLC and simMC) and compared it to the original data set (**Supplementary Fig. 2** online). Typically, dominant populations are overrepresented, and their ratio is distorted in the final binned data sets. This can be attributed to either insufficient assembly of minor populations or the assignment of contigs to very broad bins (for example, bacteria).

Notably, chimeric contigs did not have a noticeable effect on binning accuracy. This was probably because small grossly chimeric contigs had been excluded from the analysis and because chimericity largely occurs at taxonomic levels below binning resolution of each method (**Fig. 1c**). Contigs belonging to a given dominant population were assigned with variable accuracy and taxonomic resolution, and distributed across multiple bins, as shown by low sensitivity values (**Supplementary Table 2**). We attribute this to intrinsic limitations of the binning methods, as we observed it in many cases regardless of the assembly and binning method used. Ideally, researchers would like to see binning down to individual component populations. But even with the best binning method used in the present study only a fraction of large contigs were accurately assigned down to genus (~60% of contigs for the sample-specific model version of PhyloPythia).

DISCUSSION

Even though a large amount of metagenomic data has already being generated[12] methods to process these data are in their infancy, and objective measures of their efficacy are lacking. This study provides for the first time such a quantitative measure, through the design of simulated metagenomic data sets of varying complexity. We present a critical evaluation of various assembly, gene prediction and binning methods, previously used for analysis of metagenomic data sets at the DOE-JGI, by benchmarking them against the simulated data sets. Although this study does not test all methods presently available to analyze such data, it highlights the utility of the simulated data sets and illustrates some of the typical problems of existing methods to guide future improvements.

Although all metagenomics processing steps will greatly benefit from the availability of an adequate number of reference genomes from all branches of the tree of life, this study additionally demonstrates that there is considerable need for both the improvement of existing methods and the development of new ones. The iterative application of methods may also contribute to an increase in the quality of the metagenomic analysis as downstream steps often provide information about the quality of the previous ones.

The simulated data sets and comparative analysis of the methods presented here are available at the FAMES webserver (see URL in methods) which can also be used as a tool for the evaluation of new methods. Selected data sets are also available in the IMG/M[24] system, which facilitates their analysis and

the identification of errors. We anticipate that these simulated data sets will become a standard metric for comparison and improvement of methods used in metagenomic analysis.

METHODS

Additional methods. Descriptions of the data sets and the methods used are available in **Supplementary Methods**.

URL. Phrap: http://www.phrap.org; fgenesb: http://sun1.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb;FAMES: http://fames.jgi-psf.org

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank A. Lykidis and I. Anderson from the Genome Biology Program at DOE-JGI for their feedback and comments on this manuscript. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and the University of California, Lawrence Livermore National Laboratory under contract number W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract number DE-AC02-05CH11231 and Los Alamos National Laboratory under contract number W-7405-ENG-36.

AUTHOR CONTRIBUTIONS

K.M. and N. I. performed the analysis, K.B., H.S. and E.G. performed assemblies with Phrap, JAZZ and Arachne respectively, A.C.M. performed binning with PhyloPythia, A.S. performed gene predictions with fgenesb and developed and performed binning with BLAST distr, F.K. developed and performed binning with *k*mer, M.L. performed gene prediction with the GLIMMER/CRITICA pipeline, A.L., I.G., P.R. and I.R. supported the project, P.H. and N.C K. supported the project and contributed conceptually. K.M., P.H. and N.C.K. wrote the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Tyson, G.W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).

2. Venter, J.C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).

3. Garcia Martin, H. *et al.* Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.* **24**, 1263–1269 (2006).

4. Hallam, S.J. *et al.* Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc. Natl. Acad. Sci. USA* **103**, 18296–18301 (2006).

5. Delcher, A.L., Harmon, D., Kasif, S., White, O. & Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).

6. Lukashin, A.V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).

7. Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G. & Fertil, B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* **16**, 1391–1399 (1999).

8. Karlin, S. & Burge, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**, 283–290 (1995).

9. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glockner, F.O. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163 (2004).

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

11. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3**, 0003 (2002).

12. Liolios, K., Tavernarakis, N., Hugenholtz, P. & Kyrpides, N.C. The genomes on line database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.* **34**, D332–D334 (2006).

13. Markowitz, V.M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34**, D344–D348 (2006).

14. Strous, M. *et al.* Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**, 790–794 (2006).

15. Woyke, T. *et al.* Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**, 950–955 (2006).

16. Tringe, S.G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).

17. Jaffe, D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).

18. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).

19. Chain, P. *et al.* Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. *J. Bacteriol.* **185**, 2759–2773 (2003).

20. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

21. DeLong, E.F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496–503 (2006).

22. Tringe, S.G. & Rubin, E.M. Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6**, 805–814 (2005).

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

24. Markowitz, V.M. *et al.* An experimental metagenome data management and analysis system. *Bioinformatics* **22**, e359–e367 (2006).

Table 1 | Assembly summary

	simLC			simMC			simHC		
	Phrap	Arachne	JAZZ	Phrap	Arachne	JAZZ	Phrap	Arachne	JAZZ
Total reads	100,628	100,628	100,546	125,652	125,652	125,513	118,064	118,064	117,890
Contigs (>1 reads)	10,320	1,400	1,953	12,644	4,692	5,840	19,236	558	1,066
Homogeneous contigs	9,055	1,333	1,941	11,026	4,329	5,282	16,451	477	1,012
Reads in contigs	56,843	33,065	29,390	82,693	55,022	39,808	47,338	2,744	2,832
Degree of assembly in contigs (%)	56.49	32.86	29.23	65.81	43.79	31.71	40.10	2.32	2.40
Major contigs (≥ 10 reads)	482	367	503	1,980	1,372	876	86	20	11
Homogeneous contigs	287	330	502	1,380	1,133	605	5	18	9

Reads in major contigs	30,852	28,060	24,405	50,267	39,190	1,9719	1,521	290	192
Degree of assembly in major contigs (%)	30.66	27.88	24.28	40.00	31.19	15.71	1.29	0.25	0.16

The total number of reads each data set comprised is reported in the first line. The number of contigs, homogeneous contigs, reads included in these and the degree of assembly for each assembly method used are reported, both for all contigs and for contigs with at least 10 reads.

Figure 1 | Quality of assembly. (**a**) Percentage of assembled sequence that contains assembly errors at any phylogenetic level. Only contigs with at least 10 reads are included. (**b**) Degree of chimericity of contigs; color corresponds to the assembler used. (**c**) Degree of chimericity of contigs; color corresponds to the taxonomic level from which the contig is homogeneous. All contigs with at least 2 reads are included in **b** and **c**. Larger versions of panels b and c are included in **Supplementary Fig. 1**.

Figure 2 | Gene prediction in data sets. (**a**) Predicted genes on assembled sequences. (**b**) Predicted genes on unassembled reads. The combination of assembler/gene prediction method is shown on the *x* axis. The total number of original genes included in these sequences are shown on the top of the columns.

Figure 3 | Hierarchical clustering of genes assigned to COGs in the simulated data sets. COGs are on the horizontal axis. Red and green color represent over- and under-abundant functions in each data set, respectively.

Figure 4 | Specificity and sensitivity values for selected binning methods. Only contigs larger than 8 kb were used. Error bars indicate s.d. A complete list of all the specificity and sensitivity values are available in **Supplementary Table 2**.