Reference-guided Assembly of Metagenomic Sequences

Victoria Cepeda[1,2,*], Bo Liu[1,2], Mathieu Almeida[2], Christopher M. Hill[1,2], Mihai Pop[1,2,*]

[1]Department of Computer Science, University of Maryland, College Park, Maryland, USA.

[2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA.

*vcepeda@cs.umd.edu, mpop@umd.edu

ABSTRACT

Metagenomic studies have primarily relied on *de novo* approaches for reconstructing genes and genomes from microbial mixtures. While database driven approaches have been employed in certain analyses, they have not been used in the assembly of metagenomic data. This is in part due to the small size and biased coverage of public genome databases, but also due to the inherent computational cost of mapping tens of millions of reads to thousands of full genome sequences.

Here we describe the first effective approach for reference-guided metagenomic assembly that can complement and improve upon *de novo* metagenomic assembly methods. Combined with *de novo* assembly approaches, we show that MetaCompass is able to generate significantly better results than can be obtained by either comparative or *de novo* assembly independently. Using this approach we report improved assemblies for 688 metagenomic samples from the Human Microbiome Project.

Introduction

Microorganisms comprise the majority of Earth's ecological diversity, and they play important functional roles in virtually all ecosystems. Particularly, human-associated microbial communities play a critical role in health and disease[1]. In many environments, however, more than 99% of the bacteria cannot be cultured by standard laboratory techniques[2]. Metagenomics involves the analysis of organismal DNA sequences obtained directly from an environmental sample, enabling studies of microorganisms that are not easily cultured in a laboratory. Metagenomic studies, pioneered in the early 2000s, have recently increased in number and scope due to the rapid advances of high-throughput sequencing technologies, which permit large amounts of DNA to be sequenced quickly and cheaply. For example the MetaHit consortium generated about 500 billion raw sequences from 124 human gut samples in its initial analysis[3], and the Human Microbiome Project (HMP) has generated over 1,128 reference microbial genomes, 9,811 16S sequence datasets, and 1,260 whole metagenome sequence datasets from healthy subjects[4]

The analysis of these vast amounts of data is complicated by the fact that reconstructing large genomic segments from metagenomic reads is a formidable computational challenge. Even for single organisms, the assembly of genome sequences from sequencing reads is a complex task, primarily due to ambiguities in the reconstruction that are caused by genomic repeats[5]. In metagenomic data, additional challenges arise from the non-uniform representation of genomes in a sample as well as from the genomic variants between the sequences of closely related organisms. Despite advances in metagenomic assembly algorithms over the past years[6–10], the computational difficulty of the assembly process remains high and the quality of the resulting data fairly low.

As a result, many analyses of metagenomic data are performed directly on unassembled reads[11–15], however the much shorter genomic context leads to lower accuracy. For example, PhymmBL reports an accuracy of just 59% on 100bp reads, as compared to 78% for 1000 bp segments[12]. The need for effective and efficient metagenomic assembly approaches remains high, particularly due to the fact that long read technologies (which partly mitigate the challenges posed by repeats) are not yet appropriate for metagenomic applications due to their high error rates and relatively high costs.

Reference-guided, comparative assembly approaches have previously been used to assist the assembly of short reads when a closely related reference genome was available [16–19] Comparative assembly works as follows: short sequencing reads are aligned to a reference genome of a closely related species, then their reconstruction into contigs is inferred from their relative locations in the reference genome [20]. This process overcomes, in part, the challenge posed by repeats as the entire read (not just the segment that overlaps adjacent reads) provides information about its location in the genome.

Currently, tens of thousands of bacterial genomes have been sequenced, and the number is expected to grow rapidly in the near future. These sequenced genomes provide a great resource for performing comparative assembly of metagenomic sequences, however they have yet to be used for this purpose in no small part due to the tremendous computational cost of aligning the reads from a metagenomic project to the entire reference collection of bacterial genomes.

In this paper we describe new algorithms and MetaCompass, a first assembly software package for the reference-assisted assembly of metagenomic data. We rely on an indexing strategy to quickly construct sample-specific reference collections, and show that this approach effectively complements *de novo* assembly methods. We also show that the combination of comparative and *de novo* assembly approaches can significantly boost the contiguity and completeness of metagenomic assembly, and use our new approach to provide an improved assembly of the data generated by the Human Microbiome Project [4].

Results

All assembly results were analyzed based on contiguity statistics, and also based on the number of complete genes found in the final assembly – a measure of how useful an assembly may be to downstream analyses. We distinguish between the total number of genes and the total number of phylogenetic marker genes – genes conserved across archaeal and bacterial organisms. The coverage of the set of marker genes has been used by the HMP and others[3,21,22] as measure of the completeness of an assembly.

Assembly of two artificial metagenomic samples

We first evaluated MetaCompass by assembling two synthetic (also known as mock) microbial communities (even and staggered) created during the Human Microbiome Project from the purified genomic DNA of 20 bacteria, one archea and one eukaryota for which finished genome sequences were available[4,23]. Since the true genome sequences are known, this dataset allows us to fully quantify the quality of the genomic reconstruction. After sequence quality trimming, 6.29 and 7.46 million Illumina reads with average length 61bp and 59bp were obtained for the even and staggered samples, respectively. We assembled these samples using MetaCompass with two different settings. First, since we knew all the genomes present in the samples, we did not perform reference genome selection (see Methods) before the assembly of these two samples. The assembly results

(MetaCompass* row in Table 1) can be considered as an upper bound on the performance of any assembly tool, because we know the exact genomes from which the metagenomic reads were obtained. Then we allowed MetaCompass to estimate the composition of the samples (MetaCompass row in Table 1). The taxonomic compositions of the sequenced metagenomic reads from these two samples were estimated using MetaPhyler[13] as shown in Supplementary Data 1, and the reference genomes were selected according to the depth of coverage estimated by MetaPhyler (see Methods).

We compare the performance of MetaCompass with that of five widely used *de novo* assemblers: IDBA-UD[8], MEGAHIT[24], SOAPdenovo2[10], SPAdes[25], and Velvet[7]. Compared with the other assemblers, MetaCompass produced significantly larger contigs and increased the number of predicted complete genes and marker genes. Note that here we are not trying to prove that MetaCompass is better than *de novo* assemblers, and actually in this particular setting, the comparison is not fair because our reference collection contains the exact genomes present in the samples. Rather, we are trying to show that the performance of MetaCompass can be excellent if the reference collection contains genomes highly similar to those in the metagenomic sample being assembled.

When dealing with large-scale data sets, run time is also a very important factor determining the applicability of a computational tool. Here, we evaluated the runtime performance of MetaCompass on an eight-core computer with 8 GB of memory for the even and staggered metagenomic samples mentioned above using single and multiple threads (see "MetaCompass 1 iteration" in Table 2). The comparative approach is slower than most *de novo* assemblers, however not substantially, and it has similar running time compared to one of the most effective metagenomic assemblers (IDBA-UD). Running multiple iterations of MetaCompass' consensus routine (see Methods) led to a modest increase in runtime.

Assembly of stool microbiome data

Mock communities are valuable for providing a baseline of performance but do not capture the true complexities of real datasets. To evaluate MetaCompass in a realistic setting, and to compare it to the results that can be obtained by *de novo* assembly, we analyzed two real stool metagenomic samples from the MetaHIT Project, obtained from healthy individuals.

The reference collection matters. To assess the effect of the reference collection used by MetaCompass, we also augmented our database with 241 genomes reconstructed directly from MetaHIT samples[26] (Supplementary Data 2). This addition significantly boosted the performance of MetaCompass, as seen in Table 3 (rows labeled MetaCompass* as compared to rows labeled MetaCompass).

Comparative and *de novo* approaches complement each other. We assembled the two stool samples with the *de novo* assembler SOAPdenovo2, a newer version of the assembler originally used to reconstruct these samples as part of the MetaHIT project. The contigs produced by the *de novo* assembly were smaller than those generated by MetaCompass, as evidenced both by the smaller maximum size, and smaller contigs needed to cover the most contiguous segments (the top 1-10Mbp) of the assembly. The comparison is even more striking when restricted to just those contigs from the *de novo* assembly that can be mapped to the reference genomes used by MetaCompass (rows labeled SOAPdenovo2* in Table 3). For these sequences, which are closely related to genomes available in the MetaCompass' reference database, MetaCompass can assemble larger contigs, more reads, and

cover more total DNA than SOAPdenovo2. At the same time, SOAPdenovo2 can assemble novel organisms, leading to more reads and a total amount of DNA in the final assembly.

To further explore the complementarity between the comparative and *de novo* approaches, we compared the number of reads included in the two assemblies. Across the two stool datasets a total of 109.1 million reads were assembled by either assembler. The majority (61%) was shared by the two assemblies, 7% were only assembled by MetaCompass, and 32% were only found in the SOAPdenovo 2 assembly, corresponding to the metagenomic sequences not found in the MetaCompass reference database. Combining the two approaches (see Methods) results in a final assembly that outperforms both by all metrics (total sequence covered, reads used, and contig sizes, Hybrid rows in Table 3).

Improvements from iterative assembly. Differences between the sequences being assembled and the reference genome used by MetaCompass can degrade the performance of the comparative assembly process. Iterative assembly (see Methods) can improve the quality of the reconstruction. Each iteration increases assembly contiguity and improves the number of reads that can be mapped to the final assembly, though the performance gains start plateauing after just 3 iterations (see Supplementary Data 3). The improved assembly quality comes at the cost of increases in runtime (Table 2).

MetaCompass can reconstruct complete bacterial genomes. As it is clear from the results shown above, MetaCompass can make effective use of reference genomes to effectively reconstruct related sequences from microbiome samples. In certain cases, the MetaCompass reconstruction can span entire microbial genomes. In the staggered mock community, the MetaCompass assembly includes a 1.9Mbp contig that almost perfectly matches the full length of the *Staphylococcus epidermidis* genome. In real metagenomic datasets - four retroauricular crease samples from the HMP project (NCBI accessions SRS024655, SRS024596, SRS013258, SRS046688) - MetaCompass reconstructed within each, a contig of length 2.56 Mbp closely related to the genome *Propionibacterium acnes*. Each contig is 99% identical to the reference genome *Propionibacterium acnes* KPA171202 (GenBank Accession: NC_006085), bacterium commonly found on skin surfaces, and which has a fairly small pan genome[27].

Reassembly of the data generated by the Human Microbiome Project

To further explore the benefits and limits of comparative approaches for metagenomic assembly, we re-analyzed with MetaCompass 688 metagenomic samples from the HMP Project. These samples cover 15 different body sites all coming from healthy individuals. As above, for each sample we generated both a comparative-only assembly, and a hybrid assembly that merges the comparative assembly with a *de novo* assembly of the same data. We ran MetaCompass using 3 iterations and 24 threads. Overall, MetaCompass effectively complemented the *de novo* assemblies, with the hybrid assembly outperforming the results obtained by SOAPdenovo2 - a newer version of the assembler used in the HMP project (Figure 1). The comparative assembly alone produced worse results than the original *de novo* assembly, an expected outcome given the fact that many host associated microbes have yet to be isolated or sequenced, and are thus not available in public databases. The comparative assembly performed better in terms of the number of marker genes completely reconstructed, likely due to the ubiquity and relatively high level of conservation of these genes.

The relative performance of comparative and *de novo* assembly approaches varied across body-sites due to the specific characteristics of the microbial communities being reconstructed. In stool (Figure

2), *de novo* assembly vastly outperformed MetaCompass across all metrics. Stool samples have very low human DNA contamination, leading to a much higher depth of coverage within the microbiome, factor that benefits *de novo* assembly. Furthermore, many bacteria, especially within the healthy gut microbiome, are anaerobic, hence difficult to culture and are underrepresented in sequence databases. For these bacteria the comparative approach is unsuitable. In the nares (Figure 1), the samples exhibit high levels of human DNA contamination, leading to much smaller sequence coverage of the microbiome. Here the *de novo* approach has limited effectiveness and MetaCompass generates better assemblies across all metrics. In vaginal samples (Figure 1), the human contamination is high, but the lower complexity of the normal microbiome allows for effective *de novo* assembly. The lower complexity of the community, and the fact that many members of the normal vaginal flora have been sequenced, leads to a smaller difference in performance between MetaCompass and the *de novo* assembly. In all situations, the hybrid, comparative + *de novo* assembly outperforms either approach, leading to a better assembly of the original data.

Discussion

We have described MetaCompass, a comparative metagenomics assembly method that relies on an indexing strategy to construct sample-specific reference collections. We show that comparative and *de novo* assemblies provide complementary strengths, and that combining both approaches effectively improves the overall assembly, providing a consistent increase in the quality of the assembly. The benefit of comparative assembly is highly dependent on the data available in the reference database as well as on the overall complexity of the microbial community being reconstructed. As the number of genomes in public databases is increasing, comparative approaches such as ours will be increasingly valuable for reconstructing near-complete genome sequences from metagenomic data. Already, using available genome sequences, we were able to improve upon the assembly of the data generated by the Human Microbiome Project, and the improved assemblies have been made available through the HMP DACC at http://hmpdacc.org/HMASM/.

Here we used the taxonomic profiling tool MetaPhyler as an index for the publicly available microbial genomes. Compared to whole-genome indices, the MetaPhyler index is based on just 31 phylogenetic marker genes commonly found in bacteria, thus providing a much more compact and efficient data-structure. Since MetaPhyler, and other similar tools[14,28] are designed for much broader use cases than that targeted here, it is likely that better performance in both memory and speed can be achieved by an indexing strategy designed specifically for the purpose of comparative metagenomic assembly, and we plan to explore such strategies in future work.

Also, we would like to note that comparative assembly provides new opportunities for the development of sequence alignment approaches that optimize the combined time of index creation and alignment. Most of the recent developments in sequence alignment have assumed index construction to be a one-time off-line operation, trading off a computationally intensive indexing approach for more efficient queries.

MetaCompass is released freely under an open source license at http://www.cbcb.umd.edu/software/metacompass .

Methods

Methods overview. MetaCompass operates in a two-step fashion. First we use MetaPhyler[13] to estimate the depth of coverage for all the genomes in the reference database that are closely related to

genomes in the metagenomic sample. Second, the genomes that are sufficiently well covered by reads (minimum 1% of abundance and coverage 0.5x) are selected as a sample-specific reference set to guide the assembly. These genomes are indexed, and then the metagenomic reads are aligned to them. The resulting read alignments are then used to construct contigs, and an iterative assembly step is used to refine these contigs. In a final, optional step, the comparative assembly is combined with a *de novo* assembly of the same dataset. The details of each analysis step are described below.

Selecting reference genomes. While comparative assembly approaches have already been described for single genomes[19,29] their use in metagenomic data is complicated by the fact that we do not know which reference genomes to use from among the tens of thousands of genomes now available in public databases. In principle one could simply index all available reference genomes and align to them the metagenomic reads. Building efficient indexes for large reference collections is computationally challenging for the most widely used short read aligners[30,31]. In addition, using an index comprised of all the genomes currently available requires a significant amount of memory during mapping, which may limit the usability of the tool in practice. For assembly, however we only need to use the genomes that are actually present in a sample. To identify these genomes, we rely on MetaPhyler, a taxonomic classification tool that indexes a collection of 31 bacterial core genes. The MetaPhyler index is much smaller than a whole-genome index, yet still allows us to identify which genomes have related sequences in the sample being assembled. The number of reads mapping to a particular gene within a genome can be used to estimate the depth of coverage of that genome in a sample. Only genomes estimated to be present at high enough abundance (minimum 1% of abundance) are retained for use as a reference during comparative assembly.

Aligning reads to reference sequences. The results presented in the paper are based on aligning the reads to the selected reference genomes with Bowtie 2[32] (parameters: --sam-nohead --sam-nosq --end-to-end --quiet -k 30 -p 12). However, our whole comparative assembly pipeline is designed and developed in a modular way such that any read mapping tool can be used.

Building contigs. In its simplest form, the comparative assembly approach involves mapping the reads to a genome and using their relative placement within this genome to guide the construction of contigs[19]. In the context of metagenomic data, however, this process is complicated by the fact that individual reads may map to multiple reference genomes, some of which are highly similar to each other. Adequately dealing with this ambiguity is critical for effective assembly. If all read mappings are retained, allowing a read to be associated with multiple reference genomes, the resulting assembly will be redundant, reconstructing multiple copies of the homologous genomic regions. If for each read a random placement is selected from among the multiple equivalent matches, none of the related genomes may recruit enough reads to allow assembly, thereby leading to a fragmented reconstruction. Assigning reads to genomes according to their estimated representation in the sample (determined, e.g., based on the number of reads uniquely mapped to each genome), may bias the reconstruction towards the more divergent reference genomes, which may lead to an overall poorer reconstruction of the genomic regions shared across related genomes. Here we propose a parsimony-driven approach – identifying the minimal set of reference genomes that explains all read alignments.

Formally, this problem can be framed as a set cover problem, an optimization problem which is NP-hard. To solve this problem, we use a greedy approximation algorithm, which iteratively picks the set of genomes that covers the greatest number of unused reads. It can be shown that this greedy algorithm is the best-possible polynomial time approximation algorithm for the set cover problem[33].

Given a set of reference genomes, selected as described above, a set of shotgun reads, and the alignment between each read and reference genome, the process of creating contigs is straightforward. For each nucleotide base of each reference genome, we look at the bases from the reads that are mapped to this locus, and pick the nucleotide with the highest depth of coverage as the consensus. In addition, to introduce an insertion, its depth of coverage should be higher than half of that of its neighbor nucleotides. Nucleotides from a reference sequence that do not match any base from the reads are discarded from the consensus sequence. Minimum depth of coverage and length for creating contigs can be specified through the program command-line options.

Improving assembly through iteration. Differences between the genome being assembled and the corresponding reference sequence may bias the reconstruction towards the reference sequence, leading to an imperfect assembly. Iterative assembly approaches have previously been effectively used to improve assembly quality[16,17] and we use this strategy here as well. Iterative assembly works as follows: (1) map shotgun reads to the original reference genomes; (2) create contigs based on the reads that are aligned; (3) use the newly created contigs and their surrounding sequences from the original genomes as new reference sequences, and iterate until the assemblies can not be improved further. By default, MetaCompass runs three rounds of iterative assembly.

Combining comparative and *de novo* assemblies. Comparative and *de novo* assembly approaches are complementary - comparative approaches are much more tolerant to repeats (the most significant challenge in genome assembly) and are effective at even low depths of coverage, while *de novo* approaches can assemble novel sequences not found in public databases. To leverage the complementary strengths of these tools, we use the light-weight assembler minimus2[34] to directly combine the contigs generated from the two approaches. The default parameters used by MetaAmos: (1) minimum overlap length is 100bp; (2) overlap similarity cutoff is 95%.

Gene prediction and marker gene detection. The genes were predicted in the contigs using MetaGeneMark [35](v2.7d) with the "MetaGeneMark_v1.mod" model parameter file and using the option "-n" to remove partial genes containing long strings of "N". The completion status of the genes (complete, lack 5', lack 3' and lack both) was defined by detecting all the common start codon ("ATG", "TTG", "GTG") and stop codon ("TAA", "TAG", "TGA") of prokaryotic genes.
The 40 universal single copy marker genes[36,37] were detected on the predicted genes using the standalone version of fetchMG (v1.0) http://www.bork.embl.de/software/mOTU/)[38].

Datasets used in our experiments. The Illumina reads of the even and staggered metagenomic samples of mock community from the Human Microbiome Project were downloaded from the NCBI SRA with BioProject ID 48475.

The Illumina reads of 688 metagenomic samples from the Human Microbiome Project were downloaded from the HMP Data Analysis and Coordination Center (www.hmpdacc.org).

The Illumina reads of two human gut microbial communities from Denmark (MH0012 female sample and MH0030 male sample) were downloaded from the MetaHIT project (http://www.metahit.eu/).

References

1. Hooper, L. V & Gordon, J. I. Commensal host-bacterial relationships in the gut. *Science* **292,** 1115–8 (2001).

2. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6,** 805–14 (2005).

3. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).

4. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486,** 215–21 (2012).

5. Kingsford, C., Schatz, M. C. & Pop, M. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics* **11,** 21 (2010).

6. Laserson, J., Jojic, V. & Koller, D. Genovo: de novo assembly for metagenomes. *J. Comput. Biol.* **18,** 429–43 (2011).

7. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–9 (2008).

8. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–8 (2012).

9. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–23 (2009).

10. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–72 (2010).

11. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38,** e191 (2010).

12. Brady, A. & Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* **6,** 673–6 (2009).

13. Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* **12 Suppl 2,** S4 (2011).

14. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9,** 811–4 (2012).

15. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8,** e1002358 (2012).

16. Dutilh, B. E., Huynen, M. A. & Strous, M. Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics* **25,** 2878–81 (2009).

17. Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* **11,** R41 (2010).

18. Husemann, P. & Stoye, J. r2cat: synteny plots and comparative assembly. *Bioinformatics* **26,** 570–1 (2010).

19. Pop, M. Comparative genome assembly. *Brief. Bioinform.* **5,** 237–248 (2004).

20. Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. Comparative genome assembly. *Brief. Bioinform.* **5,** 237–48 (2004).

21. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32,** 834–841 (2014).

22. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* gr.186072.114– (2015). doi:10.1101/gr.186072.114

23. Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* **21,** 494–504 (2011).

24. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* btv033– (2015). doi:10.1093/bioinformatics/btv033

25. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–77 (2012).

26. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32,** 822–828 (2014).

27. Tomida, S. *et al.* Pan-genome and comparative genome analyses of propionibacterium acnes reveal its genomic diversity in the healthy and diseased human skin microbiome. *MBio* **4,** e00003–13 (2013).

28. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).

29. Kolmogorov, M., Raney, B., Paten, B. & Pham, S. Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* **30,** i302–9 (2014).

30. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).

33. Feige, U. A threshold of ln n for approximating set cover. *J. ACM* **45,** 634–652 (1998).

34. Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8,** 64 (2007).

35. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).

36. Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311,** 1283–7 (2006).

37. Sorek, R. *et al.* Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318,** 1449–52 (2007).

38. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10,** 1196–9 (2013).

Acknowledgments
The authors were supported in part by the NIH, grants R01-HG-004885 and R01-AI-100947, and by the NSF, grants IIS-1117247 and IIS-0812111, all to MP. We would also like to thank Owen White and other members of the HMP DACC for helping us make public the revised assemblies of the HMP data.

Competing interests
The authors declare they have no competing interests.

Table 1. Assembly of the even and staggered mock metagenomic sample from Human Microbiome Project. The rows labeled "*" indicate that we directly provided MetaCompass the correct reference genomes, bypassing the automatic reference selection procedure. "Size to XXMbp" represents the size of the largest contig C such that the sum of all contigs larger than C exceeds XX Mbp. The percent of reads assembled ("Mapped reads (%)") is calculated by mapping reads back to the contigs using Bowtie 2. Missing values ("-") indicate that total assembly size is smaller than the corresponding cumulative assembly size.

Dataset	Tool	# Contigs	Total Size (Kbp)	Max Size (Kbp)	Size to 1Mbp (Kbp)	Size to 4Mbp (Kbp)	Size to 10Mbp (Kbp)	Mapped reads (%)	# Complete genes	# Complete Marker genes
	IDBA-UD	15,948	21,233.9	41.9	22.2	8.8	2.1	72.9	11,124	150
	MetaCompass	26,002	36,292.6	996.5	624.8	187.9	16.63	86.2	21,475	298
	MetaCompass*	**30,186**	**38,481.4**	**910.8**	**391.2**	**157.13**	**16.84**	**88.13**	**21,914**	**305**
Even	MEGAHIT	21,070	14,689.6	22.7	4.1	1.5	0.5	62.7	4,670	56
	SOAPdenovo2	6,917	3,264.0	8.3	0.5	-	-	7.9	640	3
	SPAdes	23,075	24,515.9	61.2	31.6	12.0	2.4	76.5	12,383	171
	Velvet	641	351.6	6.5	-	-	-	7.1	125	-
	IDBA-UD	8,320	15,575.7	132.6	54.1	21.9	1.8	79.4	10,265	142
	MetaCompass	4,170	18,309.4	785.3	473.7	139.4	39.8	84.1	15,093	213
	MetaCompass*	**3,542**	**18,866.4**	**785.3**	**473.7**	**155.0**	**40.7**	**84.1**	**15,962**	**215**
Staggered	MEGAHIT	6,910	10,386.5	107.1	53.4	16.6	0.34	60.8	6,736	98
	SOAPdenovo2	5,580	6,529.9	27.0	5.7	1.5	-	46.9	3,253	34
	SPAdes	13,219	16,403.6	46.3	20.4	6.7	1.3	80.8	8,784	126
	Velvet	4,684	2,123.2	8.4	0.4	-	-	21.1	345	1

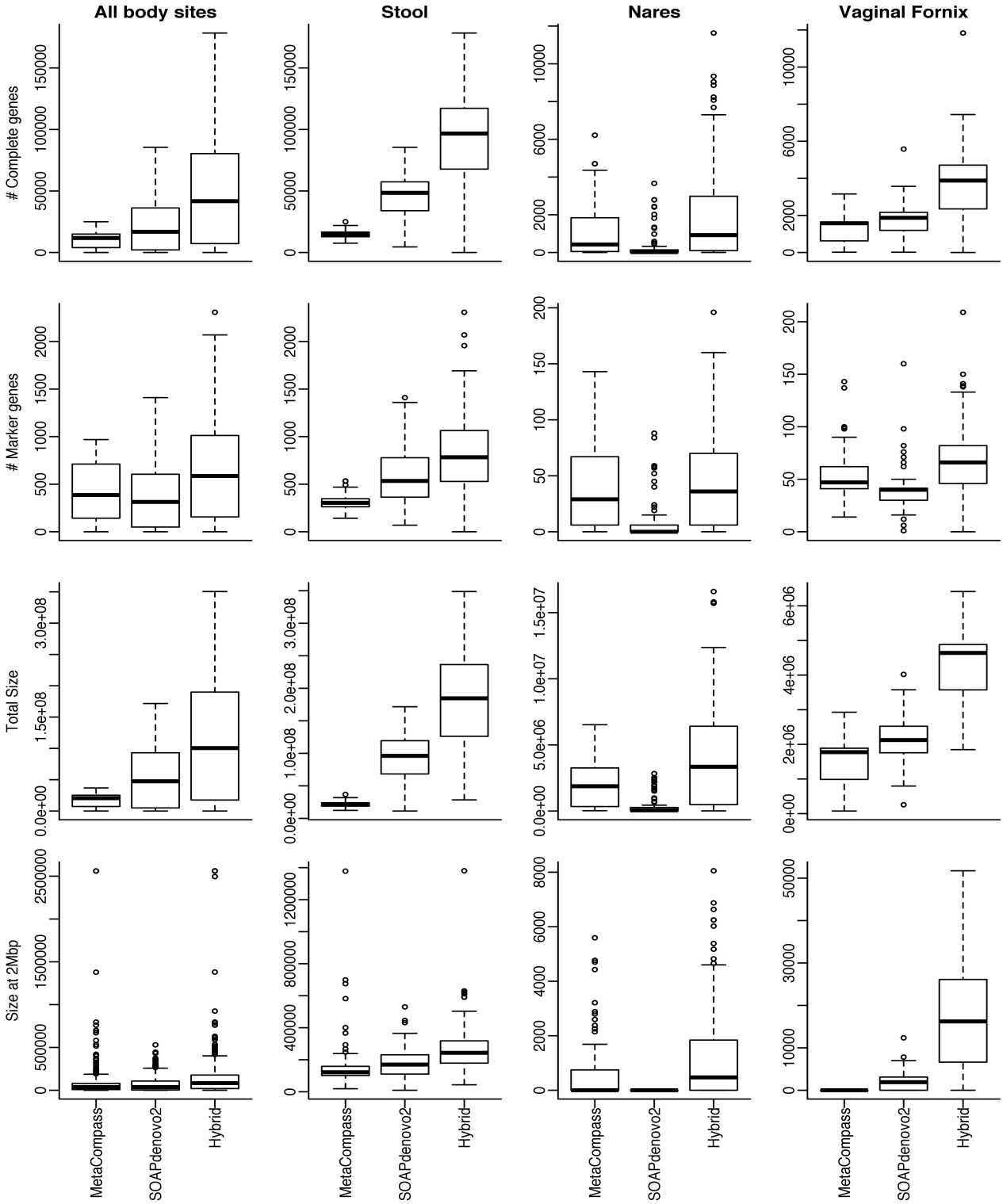
Table 2. Run time for assemblers on mock metagenomic samples.

Assembler	Even Run time with # threads (mm:ss)				Staggered Run time with # threads (mm:ss)			
	1	**2**	**4**	**8**	**1**	**2**	**4**	**8**
MetaCompass 1 iteration	35:08	28:39	20:42	16:52	31:34	27:00	17:34	16:34
MetaCompass 2 iterations	45:34	32:56	23:58	19:34	41:33	31:12	21:09	19:11
MetaCompass 3 iterations	51:52	34:51	25:55	21:36	56:00	36:19	24:25	20:30
IDBA-UD	39:32	27:28	19:32	16:00	42:29	27:17	18:32	16:34
MEGAHIT	-	17:50	06:02	03:22	-	26:51	07:57	04:47
SPAdes	09:22	06:02	03:53	02:59	36:22	25:56	23:36	23:17
SOAPdenovo2	02:42	02:14	01:47	01:28	02:46	02:08	01:48	01:41
Velvet	02:13	01:43	01:16	01:00	30:10	02:12	02:08	01:08

Table 3. Assembly statistics of comparative assembly (MetaCompass), de novo assembly (SOAPdenovo2) and a hybrid approach on 2 samples from the MetaHIT project. The rows labeled MetaCompass contain results obtained with reference genomes from the RefSeq database alone. The rows labeled MetaCompass* indicates that comparative assembly used, in addition, reference genomes reconstructed from stool samples in the MetaHIT project. The rows labeled SOAPdenovo2* show the statistics for only those contigs from the SOAPdenovo2 assembly that can be mapped to reference genomes. "Size to XXMbp" represents the size of the largest contig C such that the sum of all contigs larger than C exceeds XX Mbp. Missing values (-) indicate that total assembly size is smaller than the corresponding cumulative assembly size.

Dataset	Assembly tool	# Contigs	Total Size (Kbp)	Max Size (Kbp)	Size to 1Mbp (Kbp)	Size to 4Mbp (Kbp)	Size to 10Mbp (Kbp)	Mapped reads (%)	# Complete genes	# Complete marker genes
	MetaCompass	16,662	32,389.5	117.7	65.8	34.5	12.2	5.8	19,369	379
	MetaCompass*	44,374	113,725.6	810.4	620.0	270.4	141.4	41.4	79,791	1707
MH0012	SOAPdenovo2	11,3291	125,428.0	101.1	58.5	30.9	17.9	68.9	50,476	756
	SOAPdenovo2*	35,691	55,200.7	81.0	34.2	18.6	10.8	35.8	25,530	506
	Hybrid	121,008	211,924.7	810.4	620.0	270.4	147.9	74.9	117,283	1980
	MetaCompass	5,004	3,000.6	33.2	0.9	-	-	2.2	988	40
	MetaCompass*	34,619	47,300.8	602.4	302.3	88.8	29.6	39.9	28,644	749
MH0030	SOAPdenovo2	14,580	12,058.9	63.3	5.8	1.7	0.5	30.9	3,822	55
	SOAPdenovo2*	6,476	6,419.3	18.3	4.6	1.0	-	17.4	2,225	55
	Hybrid	42,686	55,474.5	602.4	302.3	88.8	30.5	53.1	30,996	750

Figure 1. Comparative assembly of metagenomic samples from the HMP Project. The boxplots include 688 metagenomic samples from all body sites, 137 stool samples, 87 nares samples and 51 posterior fornix samples.



Supplementary Data 1. Taxonomic compositions of the even and staggered mock samples estimated by MetaPhyler based on the Illumina shotgun reads. Coverage indicates the sum of the average depths of coverage of the genomes within a genus.

	Even		Staggered	
Genus level	**%Abundance**	**Coverage**	**%Abundance**	**Coverage**
Acinetobacter	10.5	3061	0.4	147
Bacillus	0.55	159	0.23	85
Bacteroides	7.55	2202	0.02	6
Clostridium	2.35	685	0.84	310
Deinococcus	32.99	9617	0.2	73
Enterococcus	1.57	457	0	1
Escherichia	0.7	203	3.87	1425
Helicobacter	6.36	1853	0.22	82
Listeria	2.58	751	0.05	20
Methanobrevibacter	1.21	352	6.39	2349
Neisseria	4.58	1335	0.19	69
Propionibacterium	5.84	1702	0.31	115
Pseudomonas	0.52	151	1.45	532
Rhodobacter	2.45	713	16.31	5999
Shigella	0.29	84	1.3	479
Staphylococcus	9.59	2796	54.1	19904
Streptococcus	8.52	2485	12.08	4445

Supplementary Data 2. MetaHIT genomes included to augment MetaCompass database of reference genomes.

MGS	Sample	Taxon ID (NCB	Scientific Name
MGS:2	MH0020	1262912	Parabacteroides sp. CAG:2
MGS:4	MH0014	1263075	Escherichia coli CAG:4
MGS:9	V1.CD35-1	1262967	Ruminococcus sp. CAG:9
MGS:12	V1.UC16-0	1263078	Eubacterium hallii CAG:12
MGS:13	MH0072	1263104	Roseburia intestinalis CAG:13
MGS:14	MH0012	1263090	Odoribacter splanchnicus CAG:14
MGS:15	MH0067	1263105	Roseburia inulinivorans CAG:15
MGS:18	O2.UC11-0	1262941	Roseburia sp. CAG:18
MGS:19	V1.UC25-0	1263070	Coprococcus comes CAG:19
MGS:20	MH0061	1262738	Bacteroides sp. CAG:20
MGS:24	MH0006	1263012	Firmicutes bacterium CAG:24
MGS:25	O2.UC9-0	1262984	Lachnospiraceae bacterium CAG:25
MGS:27	V1.UC11-5	1263068	Clostridium leptum CAG:27
MGS:29	MH0002	1262694	Alistipes sp. CAG:29
MGS:36	MH0002	1263079	Eubacterium rectale CAG:36
MGS:37	O2.UC49-2	1262757	Blautia sp. CAG:37
MGS:38	V1.CD41-0	1262889	Eubacterium sp. CAG:38
MGS:41	V1.UC4-5	1263021	Firmicutes bacterium CAG:41
MGS:42	V1.UC25-0	1263074	Dorea longicatena CAG:42
MGS:45	MH0115	1262947	Roseburia sp. CAG:45
MGS:50	O2.UC17-2	1262949	Roseburia sp. CAG:50
MGS:51	O2.UC37-0	1262979	Tannerella sp. CAG:51
MGS:52	V1.UC54-0	1262758	Blautia sp. CAG:52
MGS:57	V1.UC55-0	1262962	Ruminococcus sp. CAG:57
MGS:62	MH0003	1262828	Clostridium sp. CAG:62
MGS:64	V1.UC14-1	1262981	Erysipelotrichaceae bacterium CAG:64
MGS:65	MH0030	1.26E+12	Firmicutes bacterium CAG:65
MGS:67	MH0012	1263036	Alistipes putredinis CAG:67
MGS:68	MH0006	1263035	Alistipes finegoldii CAG:68
MGS:69	V1.CD36-0	1263059	Bifidobacterium longum CAG:69
MGS:72	O2.UC27-2	1263077	Eubacterium eligens CAG:72
MGS:74	MH0012	1262897	Faecalibacterium sp. CAG:74
MGS:80	MH0002	1263080	Eubacterium siraeum CAG:80
MGS:81	V1.CD18-3	1262842	Clostridium sp. CAG:81
MGS:83	O2.UC23-0	1262992	Firmicutes bacterium CAG:83
MGS:86	V1.CD4-0	1262895	Eubacterium sp. CAG:86
MGS:89	V1.UC15-3	-	-
MGS:95	MH0006	1262988	Firmicutes bacterium CAG:95
MGS:100	MH0038	1262940	Roseburia sp. CAG:100
MGS:102	O2.UC1-2	1262998	Firmicutes bacterium CAG:102
MGS:105	V1.CD15-3	1262872	Dorea sp. CAG:105
MGS:108	MH0012	1262950	Ruminococcus sp. CAG:108
MGS:114	V1.UC55-0	1263001	Firmicutes bacterium CAG:114
MGS:116	MH0006	1263095	Paraprevotella clara CAG:116
MGS:118	V1.CD15-3	1262978	Tannerella sp. CAG:118
MGS:122	V1.UC23-0	1262773	Clostridium sp. CAG:122
MGS:126	MH0014	1263106	Ruminococcus gnavus CAG:126
MGS:127	MH0012	1262774	Clostridium sp. CAG:127
MGS:129	V1.UC26-4	1263003	Firmicutes bacterium CAG:129
MGS:131	MH0137	1262862	Coprococcus sp. CAG:131
MGS:132	V1.CD35-1	1263065	Clostridium clostridioforme CAG:132
MGS:138	MH0003	1262775	Clostridium sp. CAG:138
MGS:139	V1.UC48-0	1262986	Proteobacteria bacterium CAG:139
MGS:145	V1.CD31-0	1263005	Firmicutes bacterium CAG:145
MGS:146	V1.UC30-0	1262879	Eubacterium sp. CAG:146
MGS:147	V1.UC31-0	1263061	Blautia hydrogenotrophica CAG:147
MGS:149	V1.UC11-5	1262776	Clostridium sp. CAG:149
MGS:154	MH0012	1263034	Akkermansia muciniphila CAG:154
MGS:156	O2.UC35-0	1262880	Eubacterium sp. CAG:156

MGS:157	V1.UC11-5	1262692	Alistipes sp. CAG:157
MGS:161	O2.UC32-2	1262881	Eubacterium sp. CAG:161
MGS:164	V1.UC56-0	1263102	Prevotella copri CAG:164
MGS:167	V1.UC21-0	1262777	Clostridium sp. CAG:167
MGS:177	V1.UC26-0	1262952	Ruminococcus sp. CAG:177
MGS:180	V1.UC5-3	1262882	Eubacterium sp. CAG:180
MGS:194	MH0006	1263008	Firmicutes bacterium CAG:194
MGS:196	MH0038	1262690	Acinetobacter sp. CAG:196
MGS:202	O2.UC52-2	1262884	Eubacterium sp. CAG:202
MGS:207	MH0147	1262914	Phascolarctobacterium sp. CAG:207
MGS:212	V1.CD20-4	1263009	Firmicutes bacterium CAG:212
MGS:217	O2.UC49-2	1262779	Clostridium sp. CAG:217
MGS:218	V1.CD18-0	1263072	Dialister invisus CAG:218
MGS:224	V1.UC14-1	1263067	Clostridium hathewayi CAG:224
MGS:227	V1.UC2-4	1263010	Firmicutes bacterium CAG:227
MGS:230	V1.UC55-4	1262782	Clostridium sp. CAG:230
MGS:234	V1.CD36-0	1263058	Bifidobacterium bifidum CAG:234
MGS:235	V1.CD17-0	1262854	Coprobacillus sp. CAG:235
MGS:236	O2.UC37-2	1263110	Streptococcus thermophilus CAG:236
MGS:238	MH0012	1263011	Firmicutes bacterium CAG:238
MGS:239	V1.UC25-1	1262705	Azospirillum sp. CAG:239
MGS:241	MH0062	1262911	Oscillibacter sp. CAG:241
MGS:242	O2.UC48-0	1262783	Clostridium sp. CAG:242
MGS:245	O2.UC47-0	1262784	Clostridium sp. CAG:245
MGS:251	O2.UC2-0	1262886	Eubacterium sp. CAG:251
MGS:253	MH0122	1262785	Clostridium sp. CAG:253
MGS:255	MH0011	1262923	Prevotella sp. CAG:255
MGS:257	V1.CD21-0	1262756	Blautia sp. CAG:257
MGS:259	MH0075	1263062	Butyrivibrio crossotus CAG:259
MGS:260	V1.CD15-3	1262706	Azospirillum sp. CAG:260
MGS:264	O2.UC30-0	1262786	Clostridium sp. CAG:264
MGS:265	V1.CD41-0	1262787	Clostridium sp. CAG:265
MGS:267	MH0142	1262684	Acetobacter sp. CAG:267
MGS:268	MH0054	1262693	Alistipes sp. CAG:268
MGS:269	V1.CD24-0	1262788	Clostridium sp. CAG:269
MGS:274	V1.CD29-0	1262888	Eubacterium sp. CAG:274
MGS:276	V1.UC14-1	1262699	Anaerostipes sp. CAG:276
MGS:277	MH0035	1262790	Clostridium sp. CAG:277
MGS:279	MH0020	1262924	Prevotella sp. CAG:279
MGS:287	MH0006	1263101	Phascolarctobacterium succinatutens CAG:287
MGS:288	MH0120	1262791	Clostridium sp. CAG:288
MGS:289	V1.UC38-0	1262851	Collinsella sp. CAG:289
MGS:290	MH0098	1262767	Catenibacterium sp. CAG:290
MGS:298	V1.UC54-0	1262876	Eggerthella sp. CAG:298
MGS:302	V1.UC12-0	1262793	Clostridium sp. CAG:302
MGS:306	V1.CD28-0	1262794	Clostridium sp. CAG:306
MGS:307	MH0175	1262795	Clostridium sp. CAG:307
MGS:308	O2.UC47-0	1263016	Firmicutes bacterium CAG:308
MGS:313	MH0157	1263017	Firmicutes bacterium CAG:313
MGS:314	MH0012	1262970	Subdoligranulum sp. CAG:314
MGS:317	V1.CD21-0	1262873	Dorea sp. CAG:317
MGS:318	MH0011	1262761	Butyrivibrio sp. CAG:318
MGS:321	MH0014	1263018	Firmicutes bacterium CAG:321
MGS:324	MH0097	1262969	Staphylococcus sp. CAG:324
MGS:325	MH0083	1263033	Acidaminococcus intestini CAG:325
MGS:338	V1.UC30-0	1262868	Cryptobacterium sp. CAG:338
MGS:341	V1.UC12-0	1263019	Firmicutes bacterium CAG:341
MGS:343	MH0099	1262796	Clostridium sp. CAG:343
MGS:344	V1.CD7-4	1262691	Akkermansia sp. CAG:344
MGS:345	MH0038	1263020	Firmicutes bacterium CAG:345
MGS:349	O2.UC41-2	1262797	Clostridium sp. CAG:349
MGS:352	MH0148	1262798	Clostridium sp. CAG:352
MGS:353	O2.UC58-0	1262955	Ruminococcus sp. CAG:353
MGS:354	MH0025	1262799	Clostridium sp. CAG:354

MGS:356	MH0111		1262800	Clostridium sp. CAG:356
MGS:357	V1.UC39-0		1262869	Dialister sp. CAG:357
MGS:364	V1.CD7-0		1262983	Lachnospiraceae bacterium CAG:364
MGS:368	O2.UC11-2		1262877	Eggerthella sp. CAG:368
MGS:373	O2.UC26-0	-	-	
MGS:377	V1.UC4-5		1263086	Megamonas funiformis CAG:377
MGS:397	MH0012		1262976	Sutterella sp. CAG:397
MGS:398	V1.UC13-0		1262852	Collinsella sp. CAG:398
MGS:403	O2.UC50-2		1262958	Ruminococcus sp. CAG:403
MGS:411	MH0173		1262802	Clostridium sp. CAG:411
MGS:413	MH0053		1262803	Clostridium sp. CAG:413
MGS:417	O2.UC3-0		1262804	Clostridium sp. CAG:417
MGS:433	MH0096		1262806	Clostridium sp. CAG:433
MGS:435	MH0012		1262695	Alistipes sp. CAG:435
MGS:437	V1.UC10-2		1263051	Bacteroides pectinophilus CAG:437
MGS:439	MH0043		1262899	Fusobacterium sp. CAG:439
MGS:440	MH0145		1262807	Clostridium sp. CAG:440
MGS:451	MH0100		1262809	Clostridium sp. CAG:451
MGS:452	MH0151		1262810	Clostridium sp. CAG:452
MGS:460	MH0157		1263024	Firmicutes bacterium CAG:460
MGS:462	V1.CD38-0		1262740	Bacteroides sp. CAG:462
MGS:465	V1.UC19-0		1262811	Clostridium sp. CAG:465
MGS:466	V1.CD27-0		1263025	Firmicutes bacterium CAG:466
MGS:470	O2.UC37-2		1262812	Clostridium sp. CAG:470
MGS:471	V1.CD15-3		1262948	Roseburia sp. CAG:471
MGS:472	MH0030		1262904	Mycoplasma sp. CAG:472
MGS:474	MH0006		1262926	Prevotella sp. CAG:474
MGS:475	MH0048		1263026	Firmicutes bacterium CAG:475
MGS:484	V1.UC40-0		1262759	Brachyspira sp. CAG:484
MGS:485	O2.UC60-0		1262927	Prevotella sp. CAG:485
MGS:488	V1.UC26-0		1262959	Ruminococcus sp. CAG:488
MGS:495	MH0035		1262987	Proteobacteria bacterium CAG:495
MGS:508	V1.UC9-0		1262815	Clostridium sp. CAG:508
MGS:510	MH0012		1262816	Clostridium sp. CAG:510
MGS:514	MH0009		1262696	Alistipes sp. CAG:514
MGS:520	MH0012		1262929	Prevotella sp. CAG:520
MGS:521	V1.UC2-4		1262977	Sutterella sp. CAG:521
MGS:524	MH0097		1262817	Clostridium sp. CAG:524
MGS:528	V1.UC6-0		1262700	Anaerotruncus sp. CAG:528
MGS:534	MH0143		1263027	Firmicutes bacterium CAG:534
MGS:536	O2.UC44-2		1263028	Firmicutes bacterium CAG:536
MGS:542	V1.UC13-3		1262687	Acidaminococcus sp. CAG:542
MGS:545	MH0009		1262742	Bacteroides sp. CAG:545
MGS:552	MH0143		1263029	Firmicutes bacterium CAG:552
MGS:555	MH0009		1263030	Firmicutes bacterium CAG:555
MGS:561	V1.UC49-1		1263089	Odoribacter laneus CAG:561
MGS:563	MH0053		1262961	Ruminococcus sp. CAG:563
MGS:567	MH0104		1262820	Clostridium sp. CAG:567
MGS:568	MH0004		1262821	Clostridium sp. CAG:568
MGS:571	MH0107		1262822	Clostridium sp. CAG:571
MGS:582	MH0115		1262997	Firmicutes bacterium CAG:582
MGS:590	MH0077		1262825	Clostridium sp. CAG:590
MGS:592	MH0168		1262931	Prevotella sp. CAG:592
MGS:594	MH0137		1262826	Clostridium sp. CAG:594
MGS:603	MH0035		1262891	Eubacterium sp. CAG:603
MGS:605	MH0099		1262855	Coprobacillus sp. CAG:605
MGS:617	MH0046		1262933	Prevotella sp. CAG:617
MGS:621	V1.UC26-4		1263100	Peptostreptococcus anaerobius CAG:621
MGS:628	MH0099		1262829	Clostridium sp. CAG:628
MGS:631	O2.UC52-0		1262996	Firmicutes bacterium CAG:631
MGS:632	MH0065		1262830	Clostridium sp. CAG:632
MGS:633	MH0143		1262744	Bacteroides sp. CAG:633
MGS:634	O2.UC38-0		1263083	Klebsiella variicola CAG:634
MGS:646	V1.CD21-0		1262995	Firmicutes bacterium CAG:646

MGS:649	V1.CD35-0	1262900	Fusobacterium sp. CAG:649
MGS:665	MH0115	1263071	Coprococcus eutactus CAG:665
MGS:678	V1.UC55-4	1262831	Clostridium sp. CAG:678
MGS:698	MH0126	1262856	Coprobacillus sp. CAG:698
MGS:702	MH0135	1262747	Bacteroides sp. CAG:702
MGS:709	MH0158	1262748	Bacteroides sp. CAG:709
MGS:710	MH0013	1262833	Clostridium sp. CAG:710
MGS:715	MH0183	1262834	Clostridium sp. CAG:715
MGS:719	O2.UC34-0	1263084	Lactobacillus amylovorus CAG:719
MGS:729	MH0021	1262835	Clostridium sp. CAG:729
MGS:755	V1.CD19-0	1262935	Prevotella sp. CAG:755
MGS:762	V1.UC36-0	1262837	Clostridium sp. CAG:762
MGS:768	V1.CD19-0	1262838	Clostridium sp. CAG:768
MGS:770	MH0006	1262751	Bacteroides sp. CAG:770
MGS:776	MH0146	1262906	Mycoplasma sp. CAG:776
MGS:777	MH0124	1262974	Succinatimonas sp. CAG:777
MGS:780	O2.UC1-2	1262839	Clostridium sp. CAG:780
MGS:782	MH0090	1262863	Coprococcus sp. CAG:782
MGS:788	V1.UC49-1	1262909	Odoribacter sp. CAG:788
MGS:791	MH0012	1262993	Firmicutes bacterium CAG:791
MGS:793	MH0102	1262840	Clostridium sp. CAG:793
MGS:798	MH0102	1262841	Clostridium sp. CAG:798
MGS:815	MH0137	1262901	Fusobacterium sp. CAG:815
MGS:822	V1.UC55-0	1263032	Firmicutes bacterium CAG:822
MGS:826	MH0124	1262857	Coprobacillus sp. CAG:826
MGS:831	MH0143	1262698	Alistipes sp. CAG:831
MGS:841	O2.UC24-2	1262894	Eubacterium sp. CAG:841
MGS:873	O2.UC60-0	1262936	Prevotella sp. CAG:873
MGS:877	MH0090	1262907	Mycoplasma sp. CAG:877
MGS:878	MH0118	1262686	Acholeplasma sp. CAG:878
MGS:884	V1.UC36-0	1262990	Firmicutes bacterium CAG:884
MGS:891	MH0057	1262937	Prevotella sp. CAG:891
MGS:914	MH0143	1262846	Clostridium sp. CAG:914
MGS:917	MH0143	1262688	Acidaminococcus sp. CAG:917
MGS:924	MH0069	1262938	Prevotella sp. CAG:924
MGS:927	O2.UC40-2	1262753	Bacteroides sp. CAG:927
MGS:933	O2.UC37-2	1262980	Veillonella sp. CAG:933
MGS:956	MH0144	1262908	Mycoplasma sp. CAG:956
MGS:964	V1.CD6-4	1262848	Clostridium sp. CAG:964
MGS:967	MH0067	1262849	Clostridium sp. CAG:967
MGS:977	MH0143	1262685	Acetobacter sp. CAG:977
MGS:988	MH0174	1262708	Bacillus sp. CAG:988
MGS:1000	MH0096	1262768	Clostridium sp. CAG:1000
MGS:1013	V1.UC11-5	1262769	Clostridium sp. CAG:1013
MGS:1031	V1.CD20-4	1262917	Prevotella sp. CAG:1031
MGS:1058	V1.CD19-0	1262918	Prevotella sp. CAG:1058
MGS:1060	MH0044	1262734	Bacteroides sp. CAG:1060
MGS:1138	MH0038	1262896	Faecalibacterium sp. CAG:1138
MGS:1185	MH0107	1262921	Prevotella sp. CAG:1185
MGS:1320	MH0057	1262922	Prevotella sp. CAG:1320
MGS:1329	V1.CD10-0	1263063	Clostridium bartlettii CAG:1329
MGS:1427	O2.UC20-2	1262874	Eggerthella sp. CAG:1427
MGS:1435	MH0124	1262867	Corallococcus sp. CAG:1435
MGS:5226	O2.UC43-0	1262930	Prevotella sp. CAG:5226

Supplementary Data 3. Assembly statistics after improving comparative assembly through iterative mapping on mock metagenomic samples and 2 samples from the MetaHIT project.

Dataset	# Iterations	# Contigs	Total Size(Kbp)	Max Size(Kbp)	Size at 1Mbp (Kbp)	Size at 2Mbp (Kbp)	Size at 4Mbp (Kbp)	Size at 10Mbp (Kbp)	Mapped reads (%)	# Complete genes	# Complete core genes
mock even	1	26002	36292.6	996.5	624.9	391.2	188.0	16.63	86.2	21475	298
	2	25822	37011.9	1960.7	1960.7	688.1	266.5	21.01	86.39	22119	335
	3	25786	37052.5	1960.7	1960.7	688.1	266.5	21.01	86.37	22160	336
mock even*	1	30186	38481.4	910.8	391.2	347.2	157.1	16.84	88.13	21914	305
	2	29631	39125.8	1960.7	1960.7	688.1	266.5	20.82	88.29	22554	340
	3	29598	39158.3	1960.7	1960.7	688.1	266.5	20.82	88.26	22584	342
mock stg	1	4170	18309.4	785.3	473.7	250.2	139.4	39.8	84.08	15093	213
	2	7567	20831.7	1160.0	1160.0	915.0	243.4	58.95	84.69	16151	228
	3	7620	21003.7	1160.0	1160.0	915.0	257.0	62.27	84.72	16260	228
mock stg*	1	3542	18866.4	785.3	473.7	250.2	155.0	40.69	84.12	15962	215
	2	3725	19289.2	1160.0	1160.0	870.9	233.4	59.63	84.39	16185	226
	3	3750	19320.9	1160.0	1160.0	870.9	257.0	62.52	84.42	16205	226
MH0012	1	41228	106812.1	810.4	348.7	283.5	219.9	111.5	40.34	25770	642
	2	43826	112434.5	810.4	620.0	342.3	270.4	135.09	41.02	28076	718
	3	44374	113725.6	810.4	620.0	342.3	270.4	141.42	41.38	28644	749
MH0030	1	33670	43988.0	566.9	287.2	159.9	72.4	19.27	38.88	73914	1463
	2	34396	46683.4	602.4	302.3	240.4	75.9	27.86	39.71	78704	1660
	3	34619	47300.8	602.4	302.3	240.4	88.8	29.59	39.95	79791	1707

The rows labeled "*" indicate that we directly provided MetaCompass the correct reference genomes, bypassing the automatic reference selection procedure.