MetaCompass: Reference-guided Assembly of Metagenomes

Victoria Cepeda[1,2,+], Bo Liu[1,2,+], Mathieu Almeida[2], Christopher M. Hill[1,2], Sergey Koren[3], Todd J. Treangen[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.

[3]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA.

+ These authors contributed equally to this work

* mpop@umd.edu, to whom correspondence should be addressed

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 metagenomes. Here we describe the first effective approach for reference-guided metagenomic assembly of low-abundance bacterial genomes that can complement and improve upon *de novo* metagenomic assembly methods. When combined with *de novo* assembly approaches, we show that MetaCompass can generate more complete

assemblies than can be obtained by *de novo* assembly alone, and improve on assemblies from the Human Microbiome Project (over 2,000 samples).

Keywords: metagenome assembly, microbiome, low coverage assembly, comparative assembly

Background

Microorganisms play an important role in virtually all of the Earth's ecosystems, and are critical for the health of humans [1], plants, and animals. Most microbes, however, cannot be easily grown in a laboratory [2]. The analysis of organismal DNA sequences obtained directly from an environmental sample (a field termed metagenomics), enables the study of microorganisms that are not easily cultured. Metagenomic studies have exploded in recent years due to the increased availability of inexpensive high-throughput sequencing technologies. 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 hundreds of reference microbial genomes and thousands of 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 short reads is a complex task, primarily due to ambiguities in the reconstruction that are caused by genomic repeats [5]. In addition, metagenomic assemblers must be tolerant of non-uniform representation of genomes in a sample as well as of 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 assemblies requires improvement.

Consequently, many analyses of metagenomic data are performed directly on unassembled reads [11–15], however the much shorter genomic context leads to lower accuracy [16].The need for effective and efficient metagenomic assembly approaches remains high, particularly since long read technologies (which partly mitigate the challenges posed by repeats [17–19]) are not yet effective in metagenomic applications due to lower throughput, higher costs [20, 21], and higher required DNA quality and concentration.

Reference-guided, comparative assembly approaches have previously been used to assist the assembly of short reads when a closely related reference genome was available [22, 23]. 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 [23]. This process overcomes, in part, the challenge posed by repeats as the entire read (not just the segment that overlaps within adjacent reads) provides information about its location in the genome.

Currently, thousands of bacterial genomes have been sequenced and finished [24], and this number is expected to grow rapidly soon thanks to long read technologies. These sequenced genomes provide a great resource for performing comparative assembly of metagenomic sequences. Comparative approaches developed in the context of single genomes cannot, however, be directly used in a metagenomic setting. Simply mapping a set of reads to even hundreds of different genomes is

currently computationally prohibitive. Furthermore, genome databases comprise many variants of a same genome (e.g., the US FDAs GenomeTrackr project [25] alone has contributed over 60,000 different strains of *Salmonella*), and genome by genome analyses would result in redundant reconstructions of metagenomic sequences. We also note that some recent reference-guided strategies implemented in genomic analysis tools, such as the "--trusted-contigs" feature of the SPAdes assembler [26, 27] and StrainPhlan [28] ignore the fact that the data being reconstructed originates from genomes that are related but different from the genomes found in public databases. As a result, such approaches may actually mis-assemble the metagenomic data exactly within the genomic regions where novel biological signals may be located.

In this paper, we describe the first effective assembly software package for the reference-assisted assembly of metagenomic data. We rely on an indexing strategy to quickly construct sample-specific reference collections, thereby dramatically reducing the computational costs of mapping metagenomic reads to references. Furthermore, we eliminate redundancy in the assembly by disambiguating the mapping of reads against closely related genomes, and identify differences between the metagenomic data and the reference genomes in order to reduce the likelihood of mis-assembly.

We show that our approach effectively complements *de novo* assembly methods. We also show that the combination of comparative and *de novo* assembly approaches can boost the contiguity and completeness of metagenomic assembly, and provide an improved assembly of the entire whole-metagenome sequencing data generated by the Human Microbiome Project [4].

Our software is released freely under an open-source license at:

http://www.github.com/marbl/MetaCompass .

Results

All assemblies were compared based on contiguity statistics, number of errors, and based on the number of complete phylogenetic marker genes found in the final assembly – a measure of how useful an assembly may be to downstream analyses. The coverage of the set of marker genes has been used by the HMP and others [3, 29, 30] as measure of the completeness of an assembly.

Evaluation of performance on synthetic metagenomic dataset

We first evaluated MetaCompass by assembling a synthetic microbial community [31]. The synthetic sample was downloaded from the NCBI Short Read Archive (SRA) database, (SRR606249) and contains 54 bacteria and 10 archaea. Among these organisms, 55 had complete genome sequences in the NCBI RefSeq database (the database used by default by MetaCompass), and 9 were available only as a high-quality draft assembly at the time of publication. Since the true genome sequences are known, these data are ideal as they allow us to fully quantify the quality of the genomic reconstruction.

We set the minimum coverage in MetaCompass at 1-and 2-fold (see Methods), then performed reference genome selection (see Methods and Supplementary Table 1). The assembly results (Table 1, see MetaCompass 1X and 2X) can be considered an approximate upper bound on the performance of any assembly tool, as in this case 90% of the genomes recruited were exactly those from which the metagenomic reads were obtained. We compared the performance of MetaCompass with that of three widely used *de novo* assemblers: IDBA-UD (July 2016) [8], MEGAHIT (v1.0.6) [32], and

metaSPAdes (v3.9.0) [33]. Compared with these assemblers, MetaCompass achieved higher genome recovery (Table 1, Figure 1) and produced significantly larger and more accurate contigs (Table 1). When we decreased the MetaCompass minimum coverage threshold from 2-fold to 1-fold, we observed gains in maximum contig size and total aligned length, while retaining a similar error profile. However, we observe higher genome recovery at minimum coverage threshold 2 and 3. On the basis of the maximum contig size, total aligned length, error profile and genome recovery, we chose 2X as default setting for MetaCompass. Note that here we are not trying to prove that MetaCompass is better than *de novo* assemblers, and in this 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.

References removed from database

To provide a better idea of how MetaCompass would perform in a worst-case scenario, we removed from the database the genomes represented in the mock community (Supplementary Table 2), thereby forcing MetaCompass to recruit near-neighbor reference genomes, when available. (see 'MetaCompass.nr' row, Table 1). Median genome recovery for MetaCompass is just 1% less than that of *de novo* assemblers. The accuracy of the reconstruction, as measured by mismatch and indel rates, is higher than that of IDBA-UD and metaSPAdes (Table 1, MetaCompass.nr (2x)), while moderately lower than MEGAHIT.

The number of misassemblies and local misassemblies per 1 Mbp of assembled sequence (as reported by MetaQuast [34]) increased from 2.0 to 4.9 when reducing the coverage threshold to 1. To put this increase into context, we measured the total number of possible errors by evaluating the "accuracy" of the near-neighbor reference genomes recruited by MetaCompass with respect to the correct reference sequence (Figure 2, see hashed blue bar). This allows us to capture the real differences between the recruited reference genomes and the actual genome represented in the synthetic dataset [31], essentially providing an upper bound for the number of errors MetaCompass would make if it simply recapitulated the sequence of the selected reference genomes. As seen in Figure 2, MetaCompass is making five times fewer errors than would be expected, indicating our software is not unduly biased by the sequence and structure of the reference genome.

Evaluation of performance on downsampled synthetic datasets

To evaluate the ability of MetaCompass to assemble low-coverage genomes, we down-sampled the synthetic dataset to just 5 million paired-end reads, or 10% of the original data set. After down-sampling, the average coverage was reduced to approximately 3-fold (data not shown). The results (Table 2, Figure 3) highlight that MetaCompass can recover a median of 90% of each of the 64 genomes in the sample. While metaSPAdes comes in second place and is able to recover 80% (median recovery), it does so at the cost of a four times higher mis-assembly rate. The two remaining methods, MEGAHIT and IDBA-UD, leave a quarter to a half of the genomes unassembled (Table 3).

Computational performance

When dealing with large-scale data sets, the combination of total required memory and run time is an important factor in determining the applicability of a computational tool. We first evaluated the runtime performance of MetaCompass on a Linux 12-core server node with 80 GB of memory using the Shakya et al. synthetic dataset. The wall clock run time on this synthetic dataset for MetaCompass

is comparable to that of *de novo* assemblers, sometimes lower (Supplementary Table 3). MetaCompass (without PILON) and Megahit were the only approaches that required <16GB of RAM on a 100 million read dataset, highlighting the scalability of these methods to large datasets.

Reassembly of the data generated by the Human Microbiome Project (HMP2)

To further explore the benefits and limits of comparative approaches for metagenomic assembly, we re-analyzed with MetaCompass 2,077 metagenomic samples from the HMP Project (ftp://public-ftp.hmpdacc.org/Illumina/PHASEII/). These samples cover 15 body sites from four broad regions of the human body: oral, skin, stool, and vaginal. We compared the assemblies produced by MetaCompass with the official IDBA-UD assemblies reported by the HMP project [35]. Note that these assemblies were recently improved by Lloyd-Price et al. [36] but we did not include them in this study. Across all samples, on average, MetaCompass outperforms the HMP2 *de novo* approach, leading to an overall better assembly of the original data (Table 3, Figure 4). However, the relative performance of MetaCompass and the HMP2 assembly varied across body-sites due to the specific characteristics of the microbial communities being reconstructed. While MetaCompass generates more assembled sequence and complete marker genes across all body sites, the maximum contig size and size at 1 Mbp metrics vary per body site. In oral and stool samples (Figure 4), MetaCompass outperforms *de novo* assembly for all metrics. In skin and vaginal samples (Figure 4), the *de novo* approach has better contiguity statistics but MetaCompass assembles more complete marker genes. To gain further insight into these results we calculated the average nucleotide identity between the *de novo* assembled contigs and the recruited reference genomes for each body site. In all body sites, except for oral, the assembled contigs had 99% average nucleotide identity to the reference genomes. In the oral samples, the most distant reference genomes had only 97% identity to the assembled contigs, indicating that at least in part, the lower effectiveness of MetaCompass is due to the absence of a sufficiently closely related reference genome for some of the oral samples.

To further explore the drop in contiguity in skin and vaginal samples, we focused on just the contigs that mapped to bacterial genomes contained in the reference database, allowing for a direct comparison between MetaCompass and *de novo* contigs. The results shown in Table 4 indicate that for this set of contigs, MetaCompass outperforms the *de novo* approach for the vaginal samples. However, the *de novo* HMP2 assembly of the skin sample is still better in terms of complete genes recovered, but equivalent to MetaCompass with respect to complete marker genes recovered (a measure of assembly completeness).

Comparing reference-guided to *de novo* assembly on low-coverage HMP2 samples

To assess the ability of MetaCompass to assemble low-abundance organisms, we focused on all skin HMP2 samples. The skin samples had the second lowest average number of reads while still containing reasonable diversity and richness, as reported in Table 3. We removed the contigs assembled via *de novo* assembly from the MetaCompass output, collected the reference genomes that were used, mapped the HMP2 contigs to these reference genomes, and then evaluated the number of complete genes and complete marker genes in both. Compared to the HMP2 assembly, reference-guided assembly of these low coverage samples is able to reconstruct approximately 10% more marker genes (4,423 versus 3,915) than the *de novo* approach, roughly equating to 10 additional complete bacterial genomes in total.

We next searched for microbes that were present in the skin samples at relatively low coverage and explored the differences between the reconstructions generated by the HMP2 project and MetaCompass. Specifically, we identified a low coverage assembly of a *Propionibacterium acnes* genome reconstructed by both MetaCompass and the HMP in sample SRS057083**.** The HMP2

assembly covers less than 40% of the closest reference genome (NC_016516.1, *Propionibacterium acnes* TypeIA2 P.acn33), while the MetaCompass assembly covers more than 90% of the same genome.

Discussion

The benefit of comparative assembly is highly dependent on the reference genomes available in the database provided to MetaCompass. While MetaCompass can effectively use reference genomes that are distantly related to the genomes being assembled, the quality of the reconstruction is lower than can be achieved with closely related reference sequences. As the set of genome sequences available in public databases continues to increase, so will the effectiveness of reference-guided assembly approaches such as MetaCompass.

We have shown MetaCompass to be particularly effective in the assembly of low coverage or rare microbes, setting in which *de novo* assembly approaches simply cannot be used with good results. Improved assembly of low-abundance, rare microbes from existing datasets has the potential to provide valuable information in complex microbial communities or clinical samples where the host DNA comprises a large fraction of the data. Clinical applications are also a particularly relevant application domain for comparative approaches as the vast majority of publicly available genome sequences comprises human pathogens.

While MetaCompass provided an advantage over *de novo* approaches for most of the human-associated microbial communities sampled by the HMP project, in skin samples the performance of MetaCompass was on average lower than the assemblies produced by the HMP. This result could be

due to structural genome dynamics of bacterial defense systems commonly found in skin microbes [37–39], situation that introduces frequent structural variants between the reference genomes and the corresponding environmental isolates. We plan to further explore this hypothesis through graph-based analyses of *de novo* assemblies of the corresponding communities.

MetaCompass relies on the taxonomic profiling tool MetaPhyler as an efficient indexing strategy for identifying the reference genomes most closely related to the data being assembled. Compared to whole-genome indices, the MetaPhyler index is based on just 18 phylogenetic marker genes that are ubiquitous in bacteria, thus providing a compact and efficient data-structure. Using marker genes ensures that any genome present at a high enough coverage to allow assembly will be detected despite indexing just a small fraction of its genome. Since MetaPhyler, and other similar tools [40, 41] are designed for much broader use cases than those targeted here, it is likely that better performance in both memory and speed can be achieved by an indexing strategy designed specifically for comparative metagenomic assembly, and we plan to explore such strategies in future work. Furthermore, 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.

Conclusion

We have described MetaCompass, a computational pipeline for comparative metagenomic assembly. This novel method for metagenomic assembly leverages the increasing number of genome sequences available in public databases. We have shown 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. Even when distant reference genomes are recruited, MetaCompass is competitive with *de novo* genome assembly methods. These results are due to two critical steps. First, reference bias is avoided by constructing the consensus sequence from the reads within the sample, using the reference genome as just a guide, and by breaking the assembly where the reads indicate a structural disagreement with the reference. Second, unmapped reads are used in a *de novo* assembly process to reconstruct the sections of the metagenomic sample that are not similar to known reference genomes. In summary, we believe that reference-guided approaches such as MetaCompass, will increasingly replace the more computationally expensive and error-prone *de novo* assembly approaches as the collection of available reference genome sequences increases.

Methods

Methods overview. First, we use MetaPhyler [13] to identify reference genomes that are most closely related to the data represented in the input a sample. We use the NCBI RefSeq genome database (June 2016) as the standard reference collection for MetaCompass. We only retain for further consideration the genomes estimated by MetaPhyler to be represented at sufficient depth of coverage. These genomes are aligned using Bowtie2 [42] (v2.2.9). The resulting read alignments are then used to identify a minimal set of genomes that best explain all read alignments, then the read alignments are used to construct contigs. We developed the tool buildcontig to generate a consensus sequence for the contigs and then use Pilon [43] (v1.18) to correct the contigs in a way that reflects the genome being assembled and to avoid biasing the reconstruction towards the reference sequence. Contigs may be broken at this stage if the metagenomic sequence diverges from the reference

sequence. Finally the reads that were not included in the reference-guided process outlined above are assembled using MEGAHIT [32] (v1.0.6) to reconstruct the metagenomic segments not represented in the reference collection. The details of each analysis step are described below.

Selecting reference genomes. While comparative assembly approaches have already been described for single genomes [23, 44] their use in metagenomic data is complicated by the multiple unknown organisms and the thousands of genomes available in public databases. Building efficient indexes for large reference collections is computationally challenging for short read aligners [41], both in term of speed and memory consumption. For assembly, however we only need to use the genomes that are detected in a sample. To speed up the genome reference selection step, we reduce the 31 marker genes in MetaPhyler to 18 universally conserved marker genes in bacteria and archaea (intersection between the sets of genes used by FetchMG [45, 46] and MetaPhyler [13]). The MetaPhyler index is much smaller than a whole-genome index, yet still allows us to identify the closest reference genome detected in the sample being assembled. We further speed up the execution of MetaPhyler by restricting the analysis to just those reads that share at least one 28-mer with one of the marker gene sequences in the database. We rely on kmer-mask (http://kmer.sourceforge.net/wiki/index.php?Main_Page) to execute this filtering step. The selected reads are then aligned to the marker collection using BLASTN with the parameters '-word_size 28 -evalue 1e-10 -perc_identity 95 -max_target_seqs 100' and a minimum HSP alignment length of 35. Since closely related genomes can share the same marker genes, we retain all hits with a bit score equal to that of the top hit. Finally, we exclude from further consideration all the genomes with an estimated coverage below a user-selected coverage threshold (2-fold, by default).

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 [42] (parameters: --sam-nohead --sam-nosq --end-to-end --quiet --all -p 12). The alignments are then filtered to keep ties of lowest edit distance for

each reads, allowing a read to be aligned in multiple locations similar to the best-strata option of bowtie1.

Selecting a minimal reference set. 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 [23]. 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 [47].

Building contigs. 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 in each reference genome, we look at the bases from the reads that are mapped to each locus, and pick the variant (nucleotide or indel) with the highest depth of coverage as the consensus and report it. Minimum depth of coverage and length for creating contigs can be specified through the program command-line options.

Removing reference-bias with Pilon. Differences between the sequences being assembled and the reference genome used by MetaCompass can degrade the performance of the comparative assembly process. We employ Pilon [43] to "polish" the reference-guided assemblies, thereby changing the consensus sequence to resemble the data in the sample rather than the reference genome. During this process we also identify signatures of larger differences between the metagenomic sample and the reference sequence, and break the assembly at those locations.

Combining reference-guided and *de novo* assembly. We employ the *de novo* assembler MEGAHIT to assemble reads that were unable to be mapped back to the reference-guided assembly generated by MetaCompass. These reads represent microbes that are missing from our reference database and novel variants. This approach allows the final assembly to capture both reference and non-reference sequences. We chose MEGAHIT because it is currently the most efficient *de novo* assembler for metagenomics [48]. MEGAHIT is also the default assembly methods for the JGI metagenomic pipeline [49] and performed well in a recent review [50].

Gene prediction and marker gene detection. The genes were predicted in the contigs using MetaGeneMark [51](v3.26) 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

334 genes (complete, lack 5', lack 3' and lack both) was defined by detecting all the common start codon
335 ("ATG", "TTG", "GTG") and stop codon ("TAA", "TAG", "TGA") of prokaryotic genes.

336 The 40 universal single copy marker proteins [52, 53] were identified in predicted genes using the
337 standalone version of fetchMG (v1.0) http://www.bork.embl.de/software/mOTU/) [46].

338 **MetaQuast validation parameters.** The command used to run MetaQuast was: 'metaquast.py -R
339 ./shakya_references --fragmented --gene-finding'

340 **Synthetic metagenome assembly parameters.** IDBA-UD requires a single fasta file that was
341 generated using the IDBA 'fq2fa --merge --filter' command. MEGAHIT was run using the options '--
342 presets meta-sensitive --min-count 3 --min-contig-len 300 -t 12'. MetaSPAdes was run using the
343 options '--meta -t 12', then all contigs shorter than 300nt and with less than 3X coverage were
344 removed. IDBA-UD was run using the options '--min_count 3 --min_contig 300 --mink 20 --maxk
345 100 --num_threads 12'. MetaCompass was run using the options -m [1,2,3] -g 300 -t 16' on the
346 synthetic dataset and '-m 3 -g 300 -t 16' on the HMP2 samples.

347 **Data availability.** A list of all available HMP samples was obtained by combining those available
348 from the HMP Data Analysis and Coordination Center (DACC) (www.hmpdacc.org) and the HMP
349 SRA project PRJNA48479 on 11/16/2016. Any sample listed in the SRA and not in the DACC was
350 downloaded and processed by the HMP WGS Read Processing Protocol
351 (http://www.hmpdacc.org/doc/ReadProcessing_SOP.pdf). Three DACC samples were corrupt or
352 extracted to a duplicate SRS identifier (SRS023176, SRS043422, and SRS057182) and were
353 downloaded from SRA and processed as above. A total of 98 454 samples were excluded from the
354 downloaded set. This resulted in 2,713 samples. Some samples (504) had no references recruited and
355 were excluded from further analysis. This resulted in 2,209 MetaCompass assemblies. All HMP2
356 assemblies available at ftp://public-ftp.hmpdacc.org/HMASM/IDBA/ were downloaded (2,341 total

assemblies). A total of 2,077 samples (Supplementary Table 4) had both an HMP2 assembly and a MetaCompass assembly and were used for the analysis.

The set of known genomes for the synthetic dataset is available via the Supplementary Table 2 from Shakya *et al* [31].

Software availability. MetaCompass is available as an open-source package at: https://github.com/marbl/MetaCompass. The code is licensed under the Artistic License 2.0: https://opensource.org/licenses/Artistic-2.0

List of abbreviations

DACC – Data Analysis and Coordination Center

HMP – Human Microbiome Project

NCBI – National Center for Biotechnology

RefSeq - NCBI Reference Sequence Database

SRA – Short read archive

Funding

The authors were supported in part by the NIH, grants R01-HG-004885 and R01-AI-100947, by the NSF, grants IIS-1117247 and IIS-0812111, and the Office of Naval Research under cooperative agreement number N00173162C001, all to MP. SK was supported by the Intramural Research

Program of the National Human Genome Research Institute, National Institutes of Health. SK has received funding for travel and accommodation expenses in order to speak at Oxford Nanopore Technologies conferences.

Author's contributions

BL and MP designed the approach. VC, BL, TT, and MA implemented the algorithms, methods, and scripts described in the paper. VC, TT, and SK generated the assemblies presented in the paper. TT, VC, and MA validated the assembly results and performed the comparisons between different assemblers. VC, MP, BL, TT and wrote the paper. All authors were involved in reviewing and revising the manuscript. All authors read and approved the manuscript.

Acknowledgments

We would like to thank Owen White, Anup Mahurkar, and other members of the HMP DACC for helping us make public the revised assemblies of the HMP data. We thank C. Titus Brown and the other reviewers for their thoughtful reviews.

References

1. Hooper L V. Commensal Host-Bacterial Relationships in the Gut. Science (80-). 2001;292:1115–8. doi:10.1126/science.1058709.

2. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet. 2005;6:805–14. doi:10.1038/nrg1709.

398 3. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene
399 catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
400 doi:10.1038/nature08821.

401 4. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, et al. A framework for
402 human microbiome research. Nature. 2012;486:215–21. doi:10.1038/nature11209.

403 5. Kingsford C, Schatz MC, Pop M. Assembly complexity of prokaryotic genomes using short reads.
404 BMC Bioinformatics. 2010;11:21. doi:10.1186/1471-2105-11-21.

405 6. Laserson J, Jojic V, Koller D. Genovo: *De Novo* Assembly for Metagenomes. J Comput Biol.
406 2011;18:429–43. doi:10.1089/cmb.2010.0244.

407 7. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
408 Genome Res. 2008;18:821–9. doi:10.1101/gr.074492.107.

409 8. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and
410 metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.
411 doi:10.1093/bioinformatics/bts174.

412 9. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler
413 for short read sequence data. Genome Res. 2009;19:1117–23. doi:10.1101/gr.089532.108.

414 10. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with
415 massively parallel short read sequencing. Genome Res. 2010;20:265–72. doi:10.1101/gr.097261.109.

416 11. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic
417 Acids Res. 2010;38:e191. doi:10.1093/nar/gkq747.

418 12. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with
419 interpolated Markov models. Nat Methods. 2009;6:673–6. doi:10.1038/nmeth.1358.

13. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics. 2011;12 Suppl 2:S4. doi:10.1186/1471-2164-12-S2-S4.

14. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9:811–4. doi:10.1038/nmeth.2066.

15. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol. 2012;8:e1002358. doi:10.1371/journal.pcbi.1002358.

16. Menzel P, Ng KL, Krogh A, Marth G, Lipman D. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257. doi:10.1038/ncomms11257.

17. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10:563–9. doi:10.1038/nmeth.2474.

18. Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey SD, et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. 2013;14:R101. doi:10.1186/gb-2013-14-9-r101.

19. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2015;23:110–20. doi:10.1016/j.mib.2014.11.014.

20. Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VGH, McHardy AC, Nederbragt AJ, et al. Improved metagenome assemblies and taxonomic binning using long-read circular consensus

442 sequence data. Sci Rep. 2016;6:25373. doi:10.1038/srep25373.

443 21. Driscoll CB, Otten TG, Brown NM, Dreher TW. Towards long-read metagenomics: complete
444 assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a
445 freshwater lake co-culture. Stand Genomic Sci. 2017;12:9. doi:10.1186/s40793-017-0224-8.

446 22. Husemann P, Stoye J. r2cat: synteny plots and comparative assembly. Bioinformatics.
447 2010;26:570–1. doi:10.1093/bioinformatics/btp690.

448 23. Pop M, Phillippy A, Delcher AL, Salzberg SL. Comparative genome assembly. Brief Bioinform.
449 2004;5:237–48. doi:10.1093/bib/5.3.237.

450 24. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence
451 (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic
452 Acids Res. 2016;44:D733-45. doi:10.1093/nar/gkv1189.

453 25. Nutrition C for FS and A. Whole Genome Sequencing (WGS) Program - GenomeTrakr Network.
454 https://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/ucm3631
455 34.htm. Accessed 2 Oct 2017.

456 26. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New
457 Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol.
458 2012;19:455–77. doi:10.1089/cmb.2012.0021.

459 27. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al. Assembling
460 Genomes and Mini-metagenomes from Highly Chimeric Reads. Springer, Berlin, Heidelberg; 2013.
461 p. 158–70. doi:10.1007/978-3-642-37195-0_13.

462 28. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population
463 structure and genetic diversity from metagenomes. Genome Res. 2017;27:626–38.

464 doi:10.1101/gr.216242.116.

465 29. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes
466 in the human gut microbiome. Nat Biotechnol. 2014;32:834–41. doi:10.1038/nbt.2942.

467 30. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality
468 of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res.
469 2015;25:1043–55. doi:10.1101/gr.186072.114.

470 31. Shakya M, Quince C, Campbell JH, Yang ZK, Schadt CW, Podar M. Comparative metagenomic
471 and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities.
472 Environ Microbiol. 2013;15:1882–99. doi:10.1111/1462-2920.12086.

473 32. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for
474 large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics.
475 2015;31:1674–6. doi:10.1093/bioinformatics/btv033.

476 33. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic
477 assembler. Genome Res. 2017;:gr.213959.116. doi:10.1101/gr.213959.116.

478 34. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome
479 assemblies. Bioinformatics. 2013;29:1072–5. doi:10.1093/bioinformatics/btt086.

480 35. HMP2 assembly details. http://gembox.cbcb.umd.edu/metacompass.

481 36. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions
482 and dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61.
483 doi:10.1038/nature23889.

484 37. Puigbò P, Makarova KS, Kristensen DM, Wolf YI, Koonin E V. Reconstruction of the evolution
485 of microbial defense systems. BMC Evol Biol. 2017;17:94. doi:10.1186/s12862-017-0942-y.

38. Belkaid Y, Segre JA. Dialogue between skin microbiota and immunity. Science (80-). 2014;346:954–9. doi:10.1126/science.1260144.

39. Ambur OH, Davidsen T, Frye SA, Balasingham S V, Lagesen K, Rognes T, et al. Genome dynamics in major bacterial pathogens. FEMS Microbiol Rev. 2009;33:453–70. http://www.ncbi.nlm.nih.gov/pubmed/19396949. Accessed 6 May 2017.

40. Dutilh BE, Huynen MA, Strous M. Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. Bioinformatics. 2009;25:2878–81. doi:10.1093/bioinformatics/btp377.

41. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46. doi:10.1186/gb-2014-15-3-r46.

42. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. doi:10.1038/nmeth.1923.

43. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One. 2014;9:e112963. doi:10.1371/journal.pone.0112963.

44. Kolmogorov M, Raney B, Paten B, Pham S. Ragout-a reference-assisted assembly tool for bacterial genomes. Bioinformatics. 2014;30:i302-9. doi:10.1093/bioinformatics/btu280.

45. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, et al. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One. 2012;7:e47656. doi:10.1371/journal.pone.0047656.

46. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10:1196–9. doi:10.1038/nmeth.2693.

508 47. Feige U. A threshold of ln n for approximating set cover. J ACM. 1998;45:634–52.
509 doi:10.1145/285055.285059.

510 48. Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, et al. Metagenomic
511 assembly through the lens of validation: recent advances in assessing and improving the quality of
512 genomes assembled from metagenomes. Brief Bioinform. 2017. doi:10.1093/bib/bbx098.

513 49. Clum A. Genome Assembly at JGI. Genomic Technologies Workshop JGI User Meeting. 2015.
514 http://3q8i7m48ig9en9v121qx83t166i.wpengine.netdna-cdn.com/wp-
515 content/uploads/sites/2/2015/04/04_Alicia_2015.03.23.JGI-user-meeting-GT-workshop-Assembly-
516 talk-Clum-FINAL.pdf. Accessed 13 Oct 2017.

517 50. Greenwald WW, Klitgord N, Seguritan V, Yooseph S, Venter JC, Garner C, et al. Utilization of
518 defined microbial communities enables effective evaluation of meta-genomic assemblies. BMC
519 Genomics. 2017;18:296. doi:10.1186/s12864-017-3679-5.

520 51. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences.
521 Nucleic Acids Res. 2010;38:e132. doi:10.1093/nar/gkq275.

522 52. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic
523 reconstruction of a highly resolved tree of life. Science. 2006;311:1283–7.
524 doi:10.1126/science.1123061.

525 53. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. Genome-wide experimental
526 determination of barriers to horizontal gene transfer. Science. 2007;318:1449–52.
527 doi:10.1126/science.1147112.

528 54. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies.
529 Bioinformatics. 2016;32:1088–90. doi:10.1093/bioinformatics/btv697.

530 55. Li K, Bihan M, Yooseph S, Methé BA, Ludwig W. Analyses of the Microbial Diversity across

531 the Human Microbiome. PLoS One. 2012;7:e32118. doi:10.1371/journal.pone.0032118.

532

533

534

535 **Table 1. Evaluation of performance on synthetic dataset.** MetaCompass (X) indicates the
536 minimum coverage setting (1X or 2X), and MetaCompass.nr indicates all 64 reference genomes
537 comprising the Shakya et al. dataset were removed from the database. **Tool** indicates the assembler, **#**
538 **ctgs** the total number of assembled contigs reported by each assembler, **Max ctg** is the maximum
539 contig length (broken at errors) for all assembled contigs, **Median Genome Recovery** (**%**) is the
540 median percentage of each of the synthetic genomes that is recovered, **Complete Marker Genes**
541 **(median)** is the median number of fully reconstructed marker genes, **Total Aligned Length** is the
542 sum of the length of contigs aligned to the reference genomes, **Total Unaligned Length** is the sum of
543 the length of unaligned contigs, and **MetaQuast reported errors** are error statistics generated with
544 MetaQuast.

545

							MetaQuast reported errors				
Tool	**# ctgs**	**Max ctg**	**Median Genome Recovery (%)**	**Complete Marker Genes (median)**	**Total Aligned Length**	**Total Unaligned Length**	**Mismatches (/100kbp)**	**Indels (/100kbp)**	**Misasms (/1Mbp)**	**Local Misasms (/1Mbp)**	**Total Misasms (/1Mbp)**
MetaCompass (1X)	**18,766**	**7,057,109**	**100%**	**40**	**198,113,036**	**6,340,278**	**61.9**	1.9	**0.8**	1.1	**1.9**
MetaCompass (2X)	23,648	5,841,107	100%	**40**	195,836,655	6,198,040	63.1	**1.8**	0.9	1.1	2.0
MetaCompass.nr (2X)	42,852	1,151,857	98%	**40**	195,225,556	6,338,183	89.9	3.6	3.3	1.6	4.9
IDBA-UD	22,355	991,792	98%	39	186,777,879	6,186,424	98.6	3.5	5.3	**1.0**	6.3
MEGAHIT	35,351	1,151,857	99%	**40**	195,334,581	6,263,018	66.5	2.8	1.5	**1.0**	2.5
metaSPAdes	21,424	1,438,235	99%	**40**	192,795,050	6,208,276	97.1	3.7	1.3	**1.0**	2.3

546

Table 2. Evaluation of performance on down-sampled synthetic dataset. The synthetic dataset was down-sampled to only contain 10% of the total reads. **Tool** indicates the assembler, **# ctgs** the total number of assembled contigs reported by each assembler, **Max ctg** is the maximum contig length (broken at errors) of all assembled contigs, **Median Genome Recovery** (**%**) is the median percentage of each of the synthetic genomes that is recovered, **Complete Marker Genes (median)** is the median number of fully reconstructed marker genes, **Total Aligned Length** is the sum of the length of contigs aligned to the truth genomes, **Total Unaligned Length** is the sum of the length of unaligned contigs, and **MetaQuast reported errors** are error statistics generated with MetaQuast [54].

							MetaQuast reported errors			
Method	**# ctgs**	**Max Ctg**	**Median Genome Recovery (%)**	**Complete Marker Genes (median)**	**Total Aligned Length**	**Total Unaligned Length**	**Mismatches (/100kbp)**	**Indels (/100kbp)**	**Misasms (>1 kbp)**	**Misasms (<1 kbp)**
MetaCompass	71457	**962,929**	**90%**	**22**	**134,008,055**	3,009,931	**117.6**	**1.9**	**112**	**33**
IDBA-UD	43973	120159	45%	6	75,970,693	**1,564,008**	175.0	5.3	3447	93
MEGAHIT	62842	209,706	76%	15	105,665,678	2,774,432	128.0	4.1	772	122
metaSPAdes	67138	287,554	80%	16	111,636,826	3,154,199	133.0	4.3	470	115

Table 3. Re-assembly of 2,077 samples generated in the Human Microbiome Project. The results are aggregated by body site. **#** indicates the total reads per sample, **Avg cvg per sample (X)** is the mean estimate read coverage calculated based on the de novo assembly of each sample and body site, **Shannon Entropy (median)** is the Shannon diversity value per body site as reported in Li *et al* 2012 [55].The rows labeled MC contain results obtained with MetaCompass. The rows labeled HMP2 show the statistics for contigs from the production HMP2 assembly. **Total Size** (Mbp) is the total assembly size for each method, **Max ctg size (kbp)** is the size of the largest contig, **Median Size@1Mbp (kbp)** represents the mediad size of the largest contig C such that the sum of all contigs larger than C exceeds 1Mbp. **Median Complete Genes** represents the median number of complete genes per sample. **Median Marker Genes** indicates the median number of complete marker genes per sample.

HMP2 body site	#	Avg cvg per sample	Shannon Entropy (median) [55]	Asm	Total size (Mbp)	Max ctg size (kbp)	Median Size@ 1Mbp (kbp)	Median Complete Genes	Median Marker Genes
Oral	1107	20.0 ±8.1	2.4	HMP2	106,693	546.4	70.8	54,100	762
				MC	**135,586**	**892.3**	**95.8**	**63,144**	**915**
Skin	182	18.4 ±4.7	1.5	HMP2	2,944	890.7	**36.5**	4,654	78
				MC	**3,782**	**2,159.3**	15.1	**5,010**	**79**
Stool	427	17.4	**2.6**	HMP2	56,573	592.8	109.1	84,193	847

		±4.9		**MC**	**66,838**	**3,301.0**	**230.9**	**94,297**	**1,043**
Vaginal	159	7.8 ±4.5	0.2	HMP2	1,179	465.8	**28.7**	2,539	45
				MC	**1,458**	**558.0**	16.1	**2,934**	**60**
All (+NA)	2077 (202)	18.2 ± 5.6	1.9	HMP2	184,518	890.7	79.0	48,836	633
				MC	**232,161**	**3,301.0**	**114.6**	**57,639**	**764**

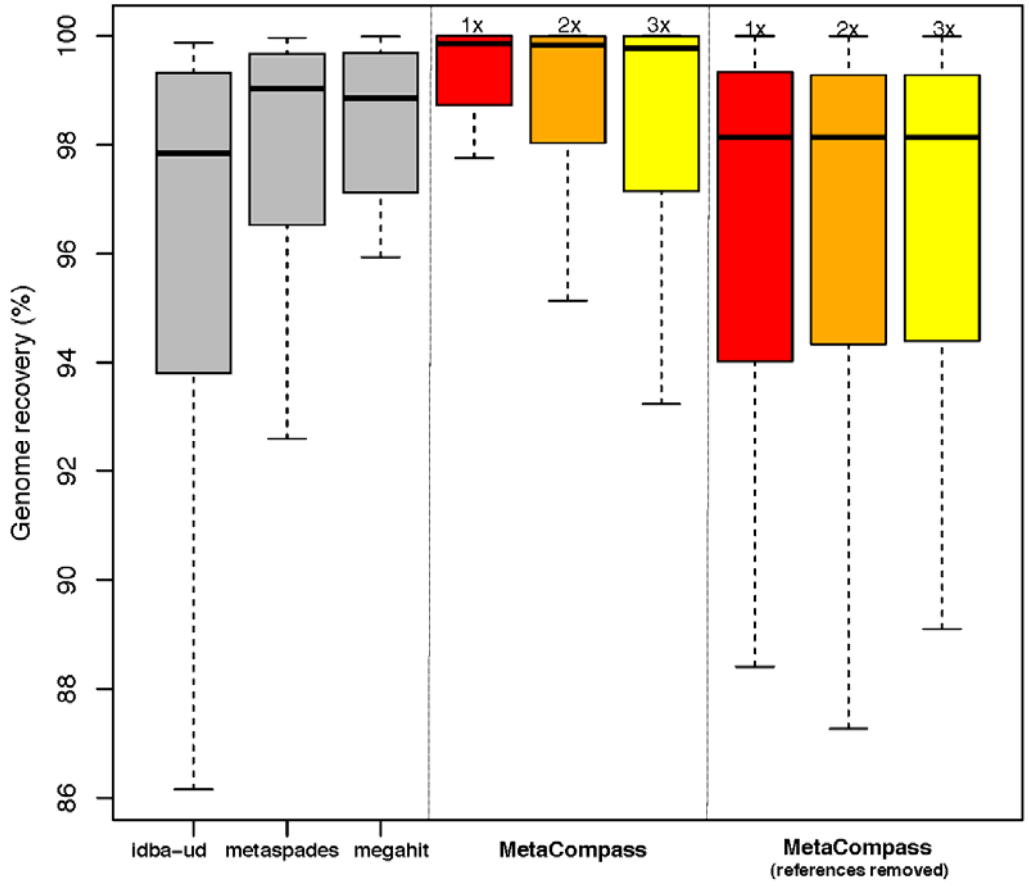
570

Table 4. Results of Human Microbiome Project analysis within the reference genomes. The results are aggregated by body site. **#** indicates the total reads per sample, **Avg cvg per sample (X)** is the mean estimate read coverage calculated based on the de novo assembly of each sample and body site, **Shannon Entropy (median)** is the Shannon diversity value per body site as reported in Li *et al* 2012 [55].The rows labeled MC contain results obtained with MetaCompass. The rows labeled HMP2 show the statistics for contigs from the production HMP2 assembly. **Total Size** (Mbp) is the total assembly size for each method, **Max ctg size (kbp)** is the size of the largest contig, **Median Size@1Mbp (kbp)** represents the median size of the largest contig C such that the sum of all contigs larger than C exceeds 1Mbp. **Median Complete Genes** represents the median number of complete genes per sample. **Median Marker Genes** indicates the median number of complete marker genes per sample.

HMP2 body site	#	Avg cvg per sample	Shannon Entropy (median) [55]	Asm	Total size (Mbp)	Max ctg size (kbp)	Median Size@ 1Mbp (kbp)	Median Complete Genes	Median Marker Genes
Oral	1107	20.0 ±8.1	2.4	HMP2	10,977	162.2	10.1	5411	176
				MC	**17,731**	**594.0**	**23.4**	**8899**	**265**
Skin	182	18.4 ±4.7	1.5	HMP2	615	716.9	**15.4**	**1973**	**35**
				MC	**618**	**2,159.3**	8.3	1652	**35**
Stool	427	17.4	2.6	HMP2	8,655	217.3	33.2	12512	142

		±4.9		**MC**	**8,665**	**3,301.0**	**104.3**	**12759**	**217**
	159	7.8	0.2	HMP2	310	172.8	7.8	1364	**37**
Vaginal		±4.5							
				MC	**330**	**558.0**	**8.0**	**1370**	**37**
All	2077	18.2	1.9	HMP2	23,126	716.9	12.7	5578	145
(+NA)	(202)	± 5.6							
				MC	**30,296**	**3,301.0**	**28.2**	**8201**	**202**

583



584

585 **Figure 1**. **Genome recovery percentages in synthetic metagenome** (MetaCompass versus *de novo*

586 assembly). Box plots represent distribution of genome recovery percentages (for the 64 genomes

present in the synthetic metagenome). x-axis indicates the assembly method, either IDBA-UD, metaSPAdes, MEGAHIT, or MetaCompass. MetaCompass was run both with the reference genomes present in the database (recruited as described in the methods) and without the truth reference genomes in the database (they were individually removed). y-axis indicates the genome recovery percentage, 0% indicates the genome was unassembled, whereas 100% indicates the genome was fully assembled.

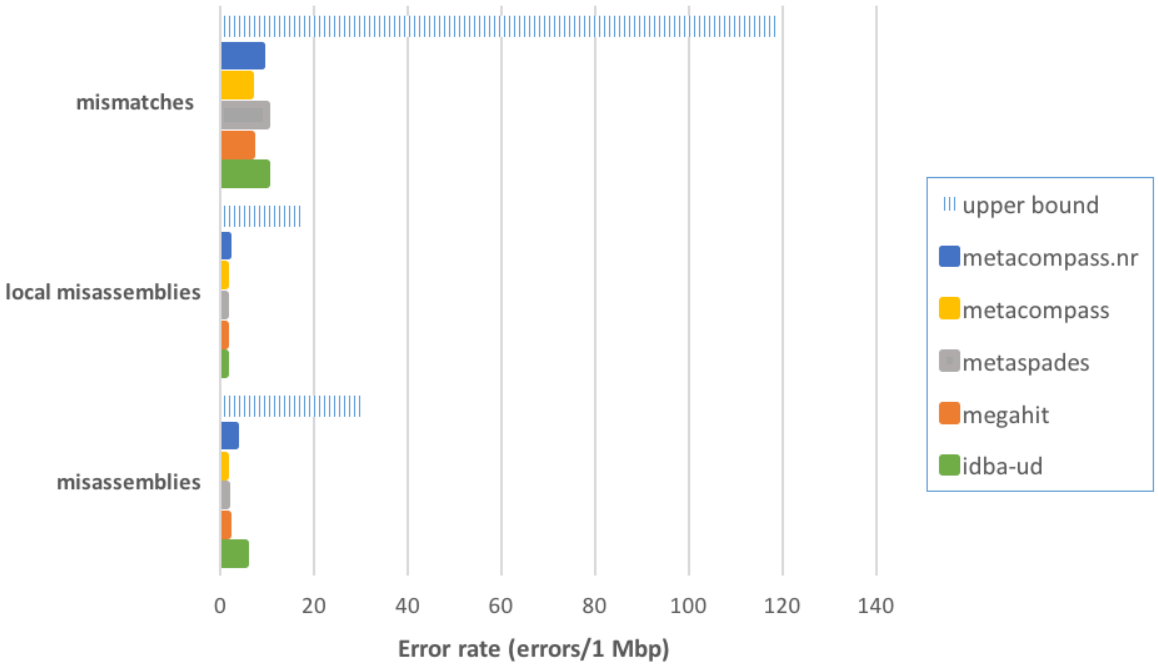
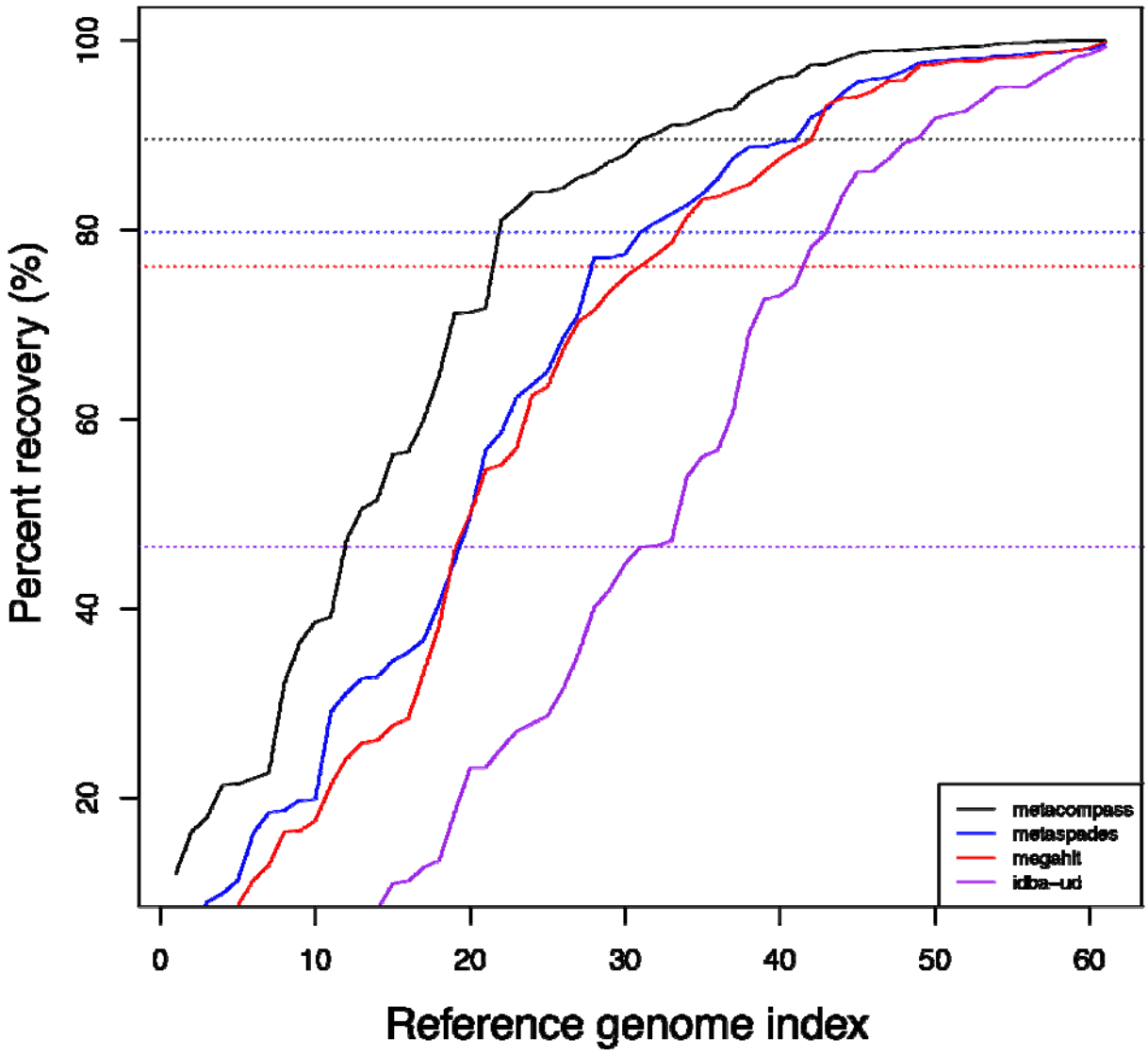


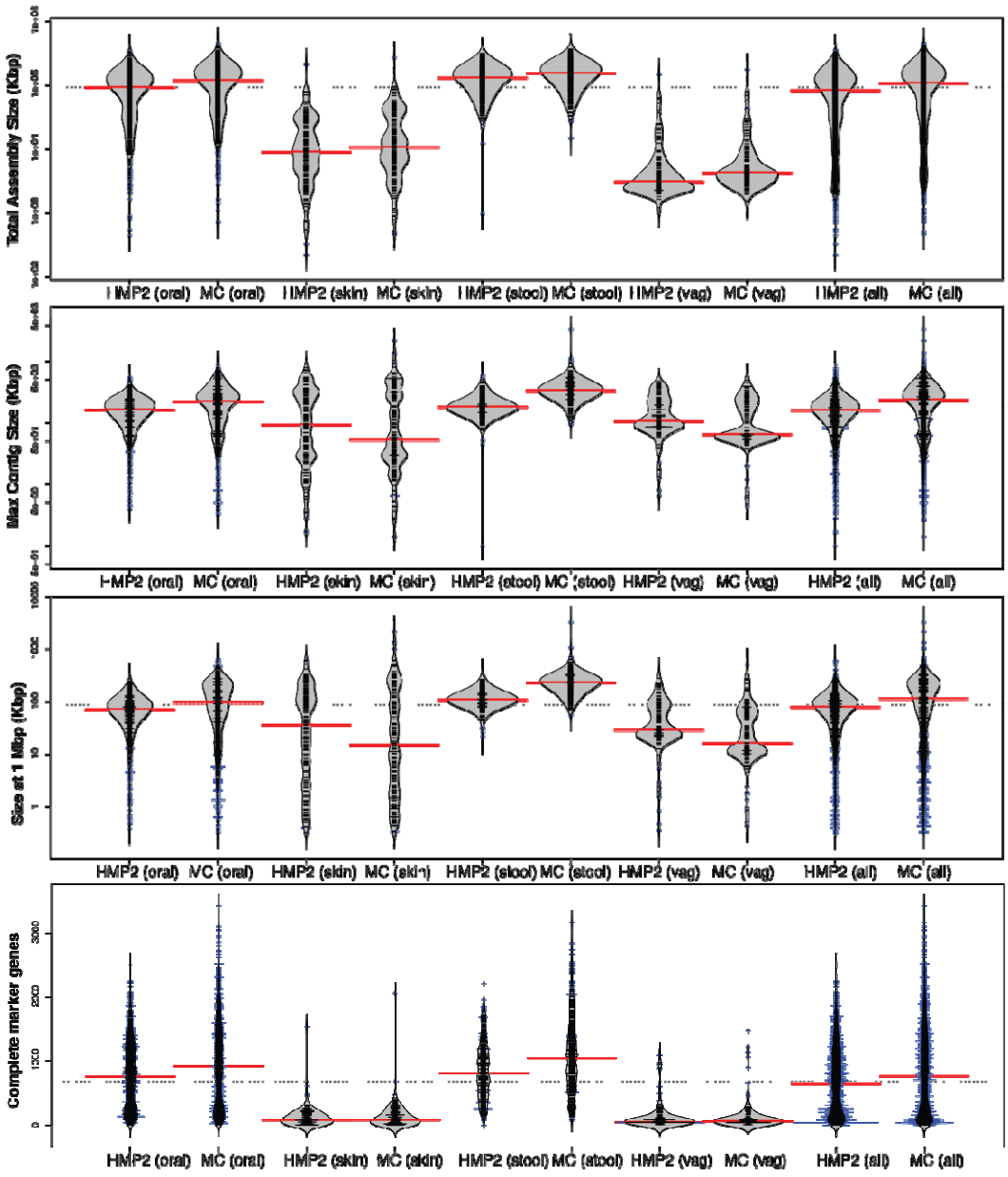
Figure 2. **Error profile on synthetic dataset**. The hashed blue bar represents the difference between the second-best reference genome (recruited by MetaCompass) and the true genome represented in the sample. This bar can be viewed as an upper bound on the errors metacompass.nr could make if it simply reconstructed the reference genome. **Mismatches** are the number of bases in a contig that differ from the reference genome. **Misassemblies** include large-scale (left flanking region aligns >1 kbp away from right flanking region) relocations, interspecies relocations, translocations, and inversions. **Local misassemblies** include small-scale (left flanking region aligns <=1 kbp away from

601 right flanking region) translocations and inversions. All errors are normalized to represent rates per 1
602 Mbp.



603

604 **Figure 3**. **MetaCompass performance on low coverage dataset**. Results obtained by down-
605 sampling the Shakya et al. synthetic genome to just 10% of the original set of reads. The 64 genomes
606 present in the sample are ordered per assembler by percent recovery, from lowest to highest. The y-
607 axis indicates how much of the n-th reference was covered by correctly assembled contigs (can range
608 from 0% to 100%). The colored dashed lines indicate the median percent recovery for each
609 assembler.



610

611 **Figure 4**. **Comparative assembly of 2,077 metagenomic samples from the HMP2 Project**. The
612 'bean plots' represent the distribution of assembly contiguity and completeness statistics across all
613 samples within the data. The x axis organizes the data by assembly and body site. The y-axis
614 indicates the statistic used to evaluate the assembly contiguity or completeness. The top panel shows
615 total assembly size (kbp), the second panel shows maximum contig size (kbp), the third panel shows
616 the size of the contig at 1 Mbp, and the bottom panel shows the complete marker genes assembled per
617 sample.