Assembling metagenomes, one community at a time

Andries J. van der Walt[1,2†], Marc W. Van Goethem[1†], Jean-Baptiste Ramond[1], Thulani P. Makhalanyane[1], Oleg Reva[2] & Don A. Cowan[1]

1 Centre for Microbial Ecology and Genomics (CMEG), Department of Genetics, University of Pretoria, Pretoria, South Africa

2 Centre for Bioinformatics and Computational Biology, Department of Biochemistry, University of Pretoria, Pretoria, South Africa

† Contributed equally

† Corresponding authors: Andries van der Walt or Marc Van Goethem

Centre for Microbial Ecology and Genomics (CMEG)

Natural Sciences Building 2

Lynnwood Road

University of Pretoria

Pretoria 0028

South Africa

Email: andriesvanderwalt@gmail.com / mwvangoethem@gmail.com

Abstract

Metagenomics allows unprecedented access to uncultured environmental microorganisms. The assembly of metagenomic sequences facilitates gene prediction and annotation, and enables the assembly of draft genomes of dominant taxa, including uncultured members of a community. However, while numerous platforms have been developed for this critical step, there is currently no strict standard for the assembly of metagenomic sequence data. To assist with selection of an appropriate metagenome assembler we evaluated the capabilities of eight prominent assembly tools on nine publicly-available environmental metagenomes. We found that assembler choice ultimately hinges on the scientific question at hand, the resources available and the bioinformatic competence of the researcher. We provide a concise workflow for the selection of the best assembly tool.

Introduction

The 'science' of metagenomics has greatly accelerated the study of uncultured microorganisms in their natural environments, providing unparalleled insights into microbial community composition and putative functionality[1]. Even though shotgun metagenomic sequencing provides comprehensive access to microbial genomic material, many of the encoded functional genes are substantially longer (~1000 bp[2]) than the length of reads provided by the sequencing platforms[3] most commonly used for shotgun metagenomic studies (Illumina HiSeq 3000, 2 x 150 bp; http://www.illumina.com/). Thus, raw sequence data alone are typically not sufficient for an in-depth analysis of a communities functional gene repertoire. Moreover, unassembled metagenomic sequence data are fragmented, noisy, error prone and contain uneven sequencing depths[4].

To assist in the accurate and thorough analysis of metagenomes, sequence data can be assembled into larger contiguous segments (contigs)[5]. To this end, numerous metagenome assembly tools (assemblers) have been developed, the vast majority of which assemble sequences in *de novo* fashion. In short, metagenomic sequences are split into predefined segments (*k*-mers), which are overlapped into a network, and paths are traversed iteratively to create longer contigs[6]. *De novo* assembly is advantageous as it allows for more confident gene prediction than is attainable from unassembled data[7]. Furthermore, *de novo* assembled metagenomes facilitate the discovery and reconstruction of novel genomes and/or genomic elements[8].

Improvements to assembly quality have greatly expanded the scope of questions that can be answered using shotgun metagenome sequencing including, for example: determination of microbial community composition and functional capacity[9], microbial population properties[10], comparisons of microbial communities from various environments[11], extraction of full genomes from metagenomes[5] and genomics-informed microorganism isolation[12]. Each of these questions require researchers to emphasise specific features of the metagenome. Genome-centric questions[5,12] require long contigs/scaffolds, while gene-centric questions[9-11] require high confidence contigs and the assembly of a large proportion of the metagenomic dataset.

Considering the wealth of available assemblers, it is particularly important that researchers understand assembler performance, especially for investigators who lack appropriate bioinformatic expertise. Firstly, an assembler needs to produce a high proportion of long contigs (>1000 bp). Long contigs allow for more accurate interpretation of full genes within a genomic context and facilitate the reconstruction of single genomes. A good assembler should also utilize most of the raw sequence data to generate the largest assembly span possible. Furthermore, an assembler needs an intuitive and user-friendly interface to enable assembly with minimal effort and rapid processing of the metagenomic data. Finally, tools should be able to assemble metagenomes using the least computational resources possible. Metagenomic assemblers are consistently being developed, this requires regular benchmarking, as with other bioinformatic tools[13].

Here we benchmark seven prominent open-source metagenome assemblers (Velvet v1.2.10[14], MetaVelvet v1.2.02[15], SPAdes v3.9.0[16], metaSPAdes v3.9.0[17], Ray Meta v2.3.1[18], IDBA-UD v1.1.1[19], MEGAHIT v1.0.6[20] as well as the commercially-available CLC Genomics Workbench v8.5.1 (QIAGEN Bioinformatics; https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/; Supplementary Table 1). We compare each assemblers performance on nine complex metagenomes from three distinct environments (i.e., three publicly available metagenomes each from soil, aquatic and human gut niches). While most of the assemblers assessed here have been tested and reviewed extensively[21-25], in this article we provide an elegant reference framework which both experienced and inexperienced researchers can use to determine which assembler is best aligned with their project scope, resources and computational background.

Materials and Methods

Metagenomic datasets

In this study we contrast the assemblies of nine publicly available metagenomic datasets uploaded to the MG-RAST server (http://metagenomics.anl.gov/), the SRA (https://www.ncbi.nlm.nih.gov/sra) or the JGI IMG/M database (https://img.jgi.doe.gov/). The metagenomes are from three distinct environments, namely; soil (Iowa[8], Oklahoma[26], and Permafrost[27]); aquatic (Kolkata Lake (unpublished data), Arctic Frost Flower[28] and Tara Ocean[29]) and human gut niches (Scandinavian Gut[30], European Gut[31] and Infant Gut[32]; Table 1). Each dataset was unique in its complexity and sequencing was performed at different depths. All metagenomes were sequenced using Illumina short read technology producing paired-end reads ranging from 100 to 151 bp in length. Most datasets were sequenced on the Illumina HiSeq 2000 platform, except for the Permafrost metagenome which was sequenced using an Illumina Genome Analyzer II, and the Kolkata Lake metagenome which comprised sequences generated by an Illumina MiSeq. This allowed for comparisons of each assemblers' performance under different coverage and taxonomic diversity. We opted to exclusively evaluate metagenomes sequenced using Illumina platforms due to their popularity and applicability to metagenomic datasets[3].

Prior to assembling the short read metagenomes, we used Prinseq-lite v0.20.4[33] for read quality control. We removed all reads with mean quality scores of less than 20 [*-min_qual_mean* 20], and removed all sequences contains any ambiguous bases (*N*) [*-ns_max_n* 0].

After quality filtering, we assessed the level of coverage of each metagenome using Nonpareil, a statistical program that uses read redundancy to estimate sequence coverage[34].

Evaluation of the metagenome assemblers

Most assemblies were performed on a local server (48 Intel® Xeon® CPU E5-2680 v3 @ 2.50 GHz processors, 504 GB physical memory, 15 TB disk space) using 8 threads. However, SPAdes, metaSPAdes and IDBA-UD required more memory, and assembly was performed on the Lengau cluster of the Centre for High Performance Computing (CHPC) for the Iowa and Oklahoma soil datasets. SPAdes, metaSPAdes, IDBA-UD and MEGAHIT iteratively analyse *k*-mer lengths to find the optimal value, and these assemblers were allowed to optimise their own *k*-mer lengths. The other assemblers used *k*-mer values of 55 (Velvet: 51; MetaVelvet: 51; SPAdes: 33, 55, 71; metaSPAdes: 33, 55, 71; Ray Meta: 55; IDBA-UD: 20, 30, 40, 50, 60, 70, 71; MEGAHIT: 21, 41, 61, 81, 99; CLC Genomics Workbench: 55). To control for *k*-mer length bias, we compared each assembler's performance at *k*-mer lengths between 50 and 61. Quality of the generated assemblies were assessed using MetaQUAST. This tool calculates basic assembly statistics, including number of contigs above various lengths (500 bp, 1 kbp, 5 kbp and 50 kbp), assembly span above various lengths (500 bp, 1 kbp, 5 kbp and 50 kbp), *N*50 lengths and *L*50 lengths. To assess the amount of data that was used for each assembly, we mapped back the short fragment sequencing reads to the

111 constructed metagenomes. This was performed using Bowtie 2[35], using the sensitive setting. Time
112 and memory (RAM) taken to complete assembly was calculated using an in-house bash script.

113 The heatmap was generated using the heatmaply package[36] in R. Values were calculated as a mean
114 proportion from the average value obtained for each statistic. This provided ratios of over- or under-
115 performance relative to the average assembly statistic (-2 to +2). The radial plots were generated
116 using log-transformed data for each assembly statistic of relevance to ensure concise representation
117 of the data.

118 Data availability is provided in Supplementary Tables 1 and 2. A link to each to assembler
119 benchmarked is provided, as are the accession numbers for all nine metagenomes assessed.

120

Results

Metagenome data and dataset complexity

Using Nonpareil, we confirmed that the soil metagenomes were more complex (less redundant) than the aquatic and human guts metagenomes, which were the least complex (highly redundant; Figure 1)[37-39]. All the human gut metagenomes came close to sequencing saturation (with at least 75% of the diversity sequenced; Figure 1). The infant gut metagenome was sequenced to above 90% estimated average coverage (~94%). However, all the sequencing depths reached were insufficient to describe the complete spectrum of microbial members in the samples assessed. For example, the largest metagenome assessed here, the Iowa soil metagenome, only described 48.8% of the total microbial diversity despite the utilization of 47 Gbp of sequence data. We note the published predictions that 2-5 Gbp of sequence data would fully capture an entire natural microbial community[40].

Estimates of the number of microbial species per gram of soil still vary substantially, with values ranging from 2000[41] to more than 830000[37]. These estimates do not include eukaryotic microbes, which generally possess much larger genomes and are much more difficult to fully sequence[42]. We therefore propose that the sequencing depth required to provide comprehensive coverage of soil metagenomes should be increased by an order of magnitude, to ~100 Gbp. This is a function of the extreme taxonomic heterogeneity of soil microbial communities, and highlights the challenges associated with assembling low coverage metagenomes.

Strategy and approaches of the current research

We defined four criteria to assess the performance of each metagenomic assembler tested; (1) ease of use and assembler attributes, (2) quality of assemblies generated and computational requirements, (3) influence of sequencing depth and coverage and (4) suitability to different environments.

1. Ease of use and assembler attributes

Many researchers entering the field of metagenomics are inexperienced in the use of intricate bioinformatic tools, and may lack extensive computational resources. To assess the ease of use for inexperienced computational biologists we evaluated the availability of a web application or graphical user interface (GUI), ease of installation, availability and completeness of manuals, MPI compatibility and programming language.

Seven of the assemblers tested here use command-line interfaces (CLI) and are open-source freeware (Velvet, MetaVelvet, SPAdes, metaSPAdes, Ray Meta, IDBA-UD and MEGAHIT). Only the commercial software CLC Genomics Workbench (Qiagen) implements a GUI (Supplementary Table 1). CLC is easily installed on most Linux, Windows or MacOS computers, whereas all other assemblers are limited to Unix-based operating systems. The GUI is intuitive, and users can

assemble simply by using a point-and-click interface. CLC provides substantial support (via manuals and web based tutorials) and was the most user-friendly assembler tested here.

Unix-based assemblers are inherently more difficult to use and must be installed or compiled from source code using the CLI. All assemblers that are CLI-based can be downloaded from GitHub, while some tools (SPAdes, metaSPAdes, Ray Meta, Velvet and MetaVelvet) provide download links from their respective parent websites. All tools, barring SPAdes and metaSPAdes, provide Message Passing Interface (MPI) compatibility, allowing parallelization which reduces computational time. All tools assessed here provide manuals or 'readme' files either on their websites or GitHub repositories, although others, such as IDBA-UD, MetaVelvet, are not comprehensive and lack information on installation or implementation. Tools with more complete manuals (MEGAHIT and Ray Meta) feature extensive wiki pages and frequently asked questions. The number of citations, websites, programming languages and MPI compatibility of all the tools assessed are provided in Supplementary Table 1.

2. Benchmarking quality of assemblies generated and computational requirements

Evaluating metagenome assembly quality is challenging without the use of known reference genomes for diverse microbial communities. We compared assembly quality using many standard metrics, including the total number of contigs longer than 500 bp, 1 kbp (referred to as long contigs throughout) and 50 kbp (referred to as ultra-long contigs throughout), maximum contig length, *N50* length of the contigs (length of the median contig, representing the length of the smallest contig at which half of the assembly is represented), mapping rate and assembly span (total length assembled using contigs > 500 bp). We used MetaQUAST to evaluate these assembly quality statistics[22].

We selected the Tara Ocean metagenome[29] for a comparison of each assembler at *k*-mer lengths between 50 and 61. We selected this range as the assemblers which automatically optimize *k*-mer values generally set sizes within this range. We set the other non-optimizing assemblers to 55. Compared to the other metagenomes, the Tara Ocean metagenomic dataset is of intermediate complexity and sequencing depth (Figure 1, Supplementary Table 2). This metagenome was sequenced on an Illumina HiSeq instrument, which is currently the most widely used shotgun metagenome sequencing technology[3]. This 5.4 Gbp metagenome comprised more than 27 million high-quality read pairs with a mean read pair length of 200.3 bp (Supplementary Table 2).

SPAdes (1415), Ray Meta (1329), IDBA-UD (1166) and metaSPAdes (1124) provided assemblies with high *N50* values (> 1000 bp), while the assemblies generated using CLC, MEGAHIT, Velvet and MetaVelvet produced *N50* statistics below 1000 bp (Figure 2; Table 1; Supplementary Figure 1). Overall, the assembly spans varied considerably with SPAdes (275.9 Mbp), MEGAHIT (210.6 Mbp), metaSPAdes (202.8 Mbp) and IDBA-UD (179.7 Mbp) assembling the largest metagenomes. Assembly span was correlated with the number of reads mapping back to the assemblies (R^2 = 0.92; Supplementary Figure 2), with SPAdes and metaSPAdes having the highest values (Table 1). Both

IDBA-UD and MEGAHIT mapped back more than 50% of the sequence reads to the assemblies. SPAdes also produced the most contigs over 1 kbp (70711), while MEGAHIT, IDBA-UD and metaSPAdes created fewer contigs in that size range, but all were comparable to each other (between 48640 and 56243 contigs). The largest contig was assembled by SPAdes (197 kbp), followed by metaSPAdes (142 kbp) and IDBA-UD (101 kbp). These three assemblers also produced the most 'ultra-long' contigs (> 50 kbp); with 54, 37 and 2 contigs, respectively.

The computational requirements of an assembly tool should be a major consideration when selecting an assembler. We evaluated all assemblers in relation to the time taken to assemble the Tara Ocean metagenome (Supplementary Figure 3; Table 1) using the same number of threads (*n*=8; Supplementary Figure 3A). Velvet, MetaVelvet and CLC assembled the metagenome in less than an hour, while MEGAHIT and Ray Meta were substantially slower, assembling over multiple hours. IDBA-UD, SPAdes and metaSPAdes required considerably more time to complete assembly, taking approximately 24 hours, or more. In terms of memory requirements, SPAdes was the most 'memory expensive' (157 GB of RAM), followed by Velvet and MetaVelvet (both 109 GB), which is substantially more RAM than is available on an average desktop computer (16 GB). By contrast, MEGAHIT (11 GB) and CLC (16 GB) were the most memory efficient assemblers (Supplementary Figure 1B and 3; Table 1).

Overall, SPAdes, metaSPAdes, IDBA-UD and MEGAHIT displayed the best performances in assembling this metagenome of intermediate size and complexity, as they produced very high *N50* values, a high proportion of long contigs and the widest assembly spans. While SPAdes was the best assembler overall, MEGAHIT was the most memory efficient, as it produced an assembly comparable to the best performing assemblers while using only a fraction of computational resources.

3. Benchmarking influence of sequencing depth and coverage

Temperate soil communities are generally more diverse than extreme counterparts (e.g., permafrost; Supplementary Table 3, Figure 1)[11]. Subsequently, high levels of diversity within these biomes require much deeper sequencing effort. Differences in microorganism abundances and strain level heterogeneity introduce complications during metagenome assembly, resulting in increased memory requirements and longer computational run-times, which may challenge assemblers. The two temperate soil metagenomes assessed here have vastly different sequencing depths, thus providing us with the scope to assess the influence of sequencing depth on the performance of each assembler. The Oklahoma soil metagenome[26] had a low sequencing depth (9 Gbp) and estimated coverage (11%), c.f. the Iowa soil metagenome[8], which had a very high sequencing depth (47 Gbp) and 49% estimated coverage (Figure 1, Supplementary Table 2). We predicted that deeper sequencing effort would be correlated with an increase in metagenome coverage[34].

All assemblers successfully assembled the Oklahoma metagenome, although SPAdes required considerably more memory (up to 1 TB RAM, Supplementary Table 3). Nevertheless, SPAdes produced the best assembly statistics for most categories (9548 long contigs and an assembly span of 54.3 Mbp; Supplementary Table 3; Figure 2). IDBA-UD and MEGAHIT used less than 500 GB of RAM and were comparable in performance (3828 and 3416 long contigs, and assembly spans of 17.2 Mbp and 20.2 Mbp, respectively; Supplementary Table 3; Figure 2). It is noteworthy that while metaSPAdes was one of the best performing assemblers for the Tara Ocean metagenome (Figure 2), it performed poorly here (Supplementary Table 3; Figure 2), suggesting that metaSPAdes is ill-suited to assembling low coverage metagenomes.

The massive Iowa soil metagenome could not be assembled by either SPAdes or IDBA-UD using our available computing resources (1 TB of RAM). This is in agreement with the methodology described by the authors who generated this dataset, who digitally normalized and partitioned the Iowa metagenome to allow for assembly using Velvet[8]. Remarkably, MEGAHIT and CLC assembled the Iowa metagenome using less than 500 GB of RAM. MEGAHIT performed best across most categories tested (assembly span of 1036.5 Mbp, largest contig of 104841 bp, and 277623 long contigs; Figure 2), while CLC produced the third-best assembly (assembly span of 432.7 Mbp, largest contig of 70207 and 114196 long contigs), using less than 64GB of memory. MetaSPAdes performed comparably to MEGAHIT but had much higher computational resource requirements to assembly the Iowa soil metagenome, using up to 1TB of RAM (assembly span of 873.8 Mbp, largest contig of 188499 bp, and 225046 long contigs).

Overall, we found that sequencing depth greatly influenced the performance of the assemblers, although the most memory-efficient tools, MEGAHIT and CLC, performed well irrespective of sequencing coverage. SPAdes and IDBA-UD produced good assemblies for the Oklahoma soil metagenome, but were extremely expensive in terms of memory and failed to assemble the Iowa soil metagenome. We found that metaSPAdes produced a better assembly for the Iowa soil metagenome than the Oklahoma dataset. MetaSPAdes performed optimally for the assembly of the high-coverage metagenome, but was less efficient in the assembly of the low-coverage metagenome.

4. Benchmarking suitability to various environments

Environmental samples are widely dissimilar in microbial community complexity and have distinct taxonomic compositions. In this study, we assembled metagenomes from three environmental biomes of different phylotypic complexities. Overall, SPAdes, MEGAHIT, IDBA-UD and metaSPAdes assembled most of the metagenomes well, according to the parameters we evaluated (Supplementary Tables 3-5). SPAdes consistently provided the largest contigs and the widest assembly spans. MEGAHIT demanded far fewer computational resources, and yet produced similar assemblies to metaSPAdes and IDBA-UD. CLC provided assemblies of moderate to high quality, was the easiest to use and performed particularly well on large metagenomes. Together, these

results indicate that no single assembler performs best across all sequencing platforms and datasets.

How to select a metagenome assembler

Bioinformatics projects can be limited by memory (RAM) requirements. SPAdes, metaSPAdes, IDBA-UD, Velvet and MetaVelvet all have large memory requirements during the assembly of massive datasets. MEGAHIT and CLC are extremely memory efficient, as they required less than 500 GB and 64 GB of RAM, respectively, to assemble the massive Iowa soil metagenome. MEGAHIT, for example, generates succinct *de Bruijn* graphs to achieve efficient memory usage[20].

Our results indicate that although many assemblers perform comparably, their applicability is defined by the research question at hand. SPAdes, for example, generated good assemblies with multiple long and ultra-long contigs for most datasets. These are ideal characteristics for genome-centric studies, which require the binning of draft genomes from community sequence data[43]. By contrast, metaSPAdes considers read coverage during assembly, making it more applicable for microbial community profiling[17]. While SPAdes and metaSPAdes produced the best assemblies in general, MEGAHIT performed comparably and emerged as a rapid and memory efficient alternative assembler.

In conclusion, we argue that when selecting an assembler, the primary consideration should be the research question. Selecting an appropriate assembler is essential to make full use of metagenomic sequence dataset. The primary objectives of the project, whether gene- or genome-centric, for example, should dictate the choice of assembler. We suggest that a secondary consideration should be the computational resources available to the researcher. Some assemblers are very memory efficient, while others sacrifice computational resources for improved assembly quality. Finally, as most assemblers use a CLI (and are more flexible than those constrained by a GUI), the GUI-based CLC platform is an excellent alternative if bioinformatic skill level is a consideration.

Other analyses

In additional analyses (Figure 2), we compared the performance of each assembler on a low diversity soil metagenome (Supplementary Table 3), other aquatic metagenomes (Supplementary Table 4) and human gut microbiomes (Supplementary Table 5).

Discussion

Over the last decade, high throughput sequencing has revolutionised the field of microbial ecology[44]. Amplicon-based technologies have allowed for near-complete classification of whole microbial communities, including populations of bacteria, archaea and fungi[45]. The emergence of two key platforms for analysing amplicon sequencing data, mothur[46] and QIIME[47], has allowed for methodological standards to be set, which enables robust comparisons between studies[48].

While whole community shotgun metagenome sequencing has facilitated the in-depth description of microbial communities from diverse environments, such as the human gut[49] and acid mine drainage systems[50], no standards exist with regard to assembly platforms or their use. While numerous reviews on strategies to analyse metagenomic data have been published[51], there are currently no standard assembly procedures implemented to enable thorough comparative analyses between projects. Numerous pipelines for processing metagenomic sequence data are available. These typically integrate existing tools into a single workflow for rapid, standardized analysis (e.g., MG-RAST, MetAMOS, and IMG/M)[52-54]. However, few of these pipelines are as widely used as mothur or QIIME in barcoding studies. This is partly because integrated metagenome analysis tools, such as MetAMOS, do not achieve the flexibility afforded by using each tool individually (e.g., using separate tools for assembly, binning and taxonomic assignment).

Consequently, investigators can analyse unassembled reads[11], optimize their assembly parameters or even develop their own tools to assemble their data prior to further analysis[55]. However, within the scope of metagenome assembly, essential details are often omitted when describing methods[56]. This leads to methodological discrepancies, and severely limits the possibility of making routine, robust comparisons between studies. This issue was recently highlighted by Vollmers, et al. [21], who reported that the taxonomic diversity patterns of microbial communities differed substantially, depending on the assembler used.

In our comparative analyses of the most popular assembly platforms, SPAdes produced the most long contigs, independent of the metagenome origin. SPAdes is ideal for genome-centric research questions that require long and ultra-long contigs, such as those that aim to bin and reconstruct single genomes from shotgun metagenomes[16]. By contrast, MEGAHIT, IDBA-UD and metaSPAdes provided very large assembly spans and consider sequence coverage during assembly, reducing the number of misassemblies generated. These tools are thus more appropriate for research questions related to taxonomic profiling of natural microbial communities, for functionally annotating microbial communities, for the analysis of population scale dynamics or for comparison of microbial communities across biomes[17,20]. Overall, MEGAHIT produced some of the best assemblies throughout this study, while only using a fraction of the computational resources required by other assemblers. We strongly recommend MEGAHIT for researchers who do not have access to large computational resources. Finally, the CLC assembler is ideal for researchers who lack a depth of bioinformatic knowledge, or who prefer to use a GUI and are willing to invest in software which is

easier to use. CLC is easy to install, has an intuitive interface and provides a compromise in which assembly quality may be sacrificed for ease of use. Strikingly, the most widely cited assembler assessed here (Velvet cited 5621 times; Supplementary Table 1) did not perform well across most metagenomes, while scarcely cited platforms (MEGAHIT, metaSPAdes cited 82 and 8 times, respectively; Supplementary Table 1) performed well across most statistics assessed here.

No assembler tested here consistently provided superior assemblies across the different metagenomes. Consequently, we propose a viable methodology for the selection of an appropriate assembler, dictated by (1) the scientific research question posed, then by (2) the computational resources available, and (3) the bioinformatics skill level of the researcher (Figure 4). In light of the above proposed framework, we urge researchers to carefully consider the assembler used (as well as the entire bioinformatics pipeline followed) while specifically bearing in mind their research question and what feature of the dataset they want accentuated.

Acknowledgements

We thank the National Research Foundation under the grant numbers 102910 (AJvdW) and 97891 (MWVG) and the University of Pretoria Genomics Research Institute for financial assistance. We thank Dr Surendra Vikram for helping with the installation of assemblers, as well as Dane Kennedy from the CHPC, South Africa with his assistance in accessing the large memory nodes of the CHPC. We thank S. de Scally for support and helpful discussions.

Author contributions

A.J.v.d.W. contributed research, analysis, formulation of ideas and writing. M.W.V.G. contributed research, analysis, formulation of ideas and writing. J.B.R. contributed formulation of ideas, direction and writing. T.P.M. contributed formulation of ideas and writing. O.R. contributed direction and writing. D.A.C. contributed formulation of ideas, direction and writing.

Competing financial interests

The authors declare no competing financial interests.

References

1 Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**, 525-552 (2004).

2 Xu, L. *et al.* Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Molecular biology and evolution* **23**, 1107-1108 (2006).

3 Clooney, A. G. *et al.* Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PloS one* **11**, e0148028 (2016).

4 Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nature Reviews Genetics* **14**, 157-167 (2013).

5 Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature communications* **7**, 13219 (2016).

6 Compeau, P. E., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* **29**, 987-991 (2011).

7 Loman, N. J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104: H4. *Jama* **309**, 1502-1510 (2013).

8 Howe, A. C. *et al.* Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences* **111**, 4904-4909 (2014).

9 Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews* **68**, 669-685 (2004).

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

11 Fierer, N. *et al.* Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences* **109**, 21390-21395 (2012).

12 Wurch, L. *et al.* Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nature communications* **7** (2016).

13 Baruzzo, G. *et al.* Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods* (2016).

14 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829 (2008).

15 Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research* **40**, e155-e155 (2012).

16 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**, 455-477 (2012).

17 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. metaSPAdes: a new versatile de novo metagenomics assembler. *arXiv preprint arXiv:1604.03071* (2016).

18 Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome biology* **13**, 1 (2012).

19 Peng, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-1428 (2012).

20 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).

21 Vollmers, J., Wiegand, S. & Kaster, A.-K. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective-Not Only Size Matters! *PloS one* **12**, e0169662 (2017).

22 Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, btv697 (2015).

23 Vázquez-Castellanos, J. F., García-López, R., Pérez-Brocal, V., Pignatelli, M. & Moya, A. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC genomics* **15**, 37 (2014).

24 Pignatelli, M. & Moya, A. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PloS one* **6**, e19984 (2011).

25 Charuvaka, A. & Rangwala, H. Evaluation of short read metagenomic assembly. *BMC genomics* **12**, S8 (2011).

26 Luo, C. *et al.* Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. *Applied and environmental microbiology* **80**, 1777-1786 (2014).

27 Hultman, J. *et al.* Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* (2015).

28 Bowman, J. S., Berthiaume, C. T., Armbrust, E. V. & Deming, J. W. The genetic potential for key biogeochemical processes in Arctic frost flowers and young sea ice revealed by metagenomic analysis. *FEMS microbiology ecology* **89**, 376-387 (2014).

29 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

30 Tremaroli, V. *et al.* Roux-en-Y gastric bypass and vertical banded gastroplasty induce long-term changes on the human gut microbiome contributing to fat mass regulation. *Cell metabolism* **22**, 228-238 (2015).

31 Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99-103 (2013).

32 Bäckhed, F. *et al.* Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell host & microbe* **17**, 690-703 (2015).

33 Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864 (2011).

34 Rodriguez-R, L. M. & Konstantinidis, K. T. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30**, 629-635 (2014).

35 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).

36 Galili, T. heatmaply: interactive heat maps (with R). *Month* (2016).

37 Gans, J., Wolinsky, M. & Dunbar, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**, 1387-1390 (2005).

38 Torsvik, V., Øvreås, L. & Thingstad, T. F. Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science* **296**, 1064-1066 (2002).

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

40 Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *science* **304**, 66-74 (2004).

41 Schloss, P. D. & Handelsman, J. Toward a census of bacteria in soil. *PLoS computational biology* **2**, e92 (2006).

42 Lynch, M. & Conery, J. S. The origins of genome complexity. *science* **302**, 1401-1404 (2003).

43 Becraft, E. D. *et al.* Single-Cell-Genomics-Facilitated Read Binning of Candidate Phylum EM19 Genomes from Geothermal Spring Metagenomes. *Applied and environmental microbiology* **82**, 992-1003 (2016).

44 Shokralla, S., Spall, J. L., Gibson, J. F. & Hajibabaei, M. Next-generation sequencing technologies for environmental DNA research. *Molecular ecology* **21**, 1794-1805 (2012).

45 Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular ecology* **21**, 2045-2050 (2012).

46 Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75**, 7537-7541, doi:10.1128/AEM.01541-09 (2009).

47 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**, 335-336 (2010).

48 Plummer, E., Twin, J., Bulach, D. M., Garland, S. M. & Tabrizi, S. N. A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. *Journal of Proteomics & Bioinformatics* **8**, 283 (2015).

49 Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *science* **312**, 1355-1359 (2006).

50 Kantor, R. S. *et al.* Genome-resolved meta-omics ties microbial dynamics to process performance in biotechnology for thiocyanate degradation. *Environmental Science & Technology* (2017).

51 Scholz, M. B., Lo, C.-C. & Chain, P. S. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current opinion in biotechnology* **23**, 9-15 (2012).

52 Markowitz, V. M. *et al.* IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research* **42**, D568-D573 (2014).

53 Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D. & Meyer, F. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols* **2010**, pdb. prot5368 (2010).

54 Treangen, T. J. *et al.* MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol* **14**, R2 (2013).

55 Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59-64 (2014).

56 Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208-211 (2015).

478 **Figure and Table legends.**

479 **Table 1. Computational requirements for assembly of the Tara Oceans metagenome.** Time
480 required is given in seconds, minutes and hours for illustrative purposes and memory in GB of RAM
481 required.

482 **Figure 1. Nonpareil estimates of sequence coverage (redundancy) for the 9 metagenomes**
483 **studied**. Metagenomes are grouped according to their environmental niche.

484 **Figure 2. Radial plots showing the assembly statistics for two soil metagenomes.** A) Oklahoma
485 soil metagenome. B) Iowa soil metagenome

486 **Figure 3. Proposed workflow to select a metagenome assembler based on the research**
487 **question, the computational resources available and the bioinformatic expertise of the**
488 **researcher**

Table 1. Assembly statistics and computational requirements of the Tara Oceans metagenome. Time required is given in seconds, minutes and hours for illustrative purposes and memory is provided in GB of RAM required.

	Tara Ocean							
	CLC	IDBA-UD	MEGAHIT	metaSPAdes	MetaVelvet	Ray Meta	SPAdes	Velvet
Number of contigs (≥ 500 bp)	50,716	163,815	216,938	185,419	67,161	6,128	220,178	57,816
Total length	46,069,409	179,686,756	210,621,485	202,770,058	55,972,515	7,277,214	275920,632	45,425,460
No. of long contigs (≥ 1 kbp)	10,720	50,498	56,243	48,640	12,590	2,179	70,711	8,802
No. of ultra-long contigs (≥ 50 kbp)	0	2	1	37	0	0	54	0
Largest contig	39,748	101,400	62,649	141,519	30,177	41,443	197,381	21,980
N50	880	1,166	982	1124	805	1,329	1415	749
L50	14,113	38,236	58,246	39,033	21,544	1,345	39,617	19,631
Mapping rate (%)	38.98	52.24	55.92	64.03	41,17	8.25	64.46	48.19
Time (seconds)	3,527	69,782	10,455	125,862	2,527	16,419	80,039	2342
Time (minutes)	58.78	1,163.03	174.25	2,097.70	42.12	273.65	1,333.98	39.03
Time (hours)	0.98	19.38	2.90	34.96	0.70	4.56	22.23	0.65
Memory required (GB)	16.23	42.84	10.58	66.53	109.37	42	157.75	109.37

Figure 1. Nonpareil estimates of sequence coverage (redundancy) for the 9 metagenomes studied. Metagenomes are grouped according to their environmental niche.

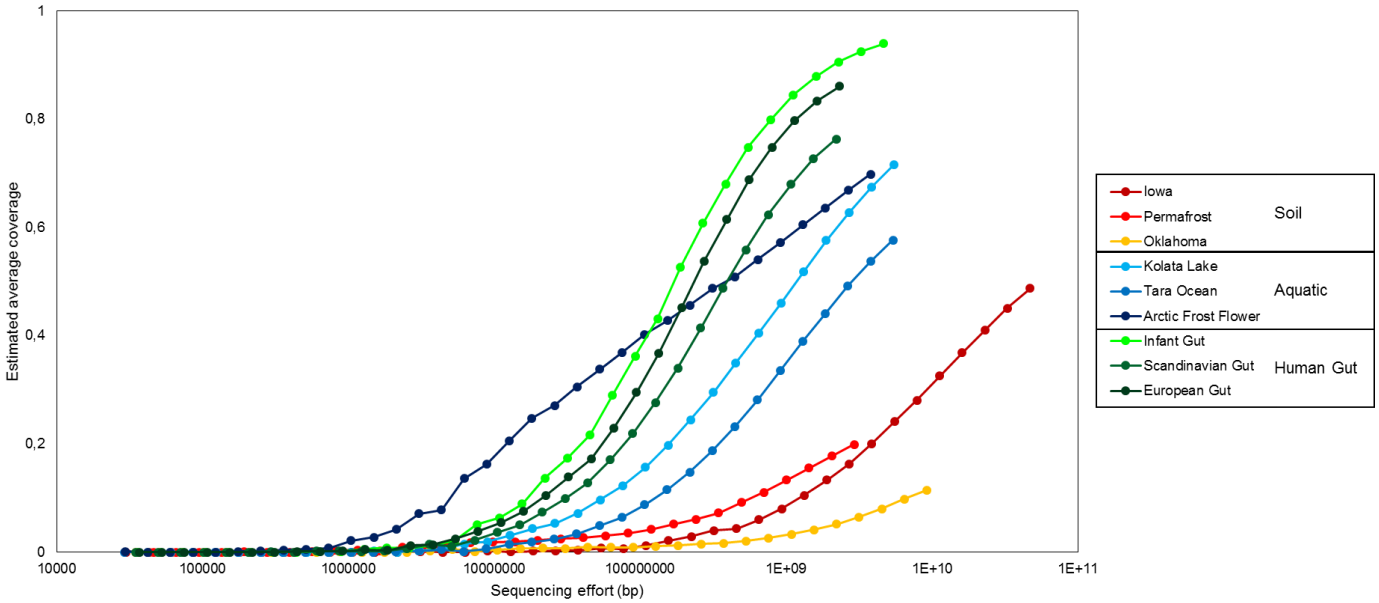

Figure 2. Radial plots showing scaled assembly statistics for the all the metagenomes assessed. Metagenomes are labelled above the respective radial plots.

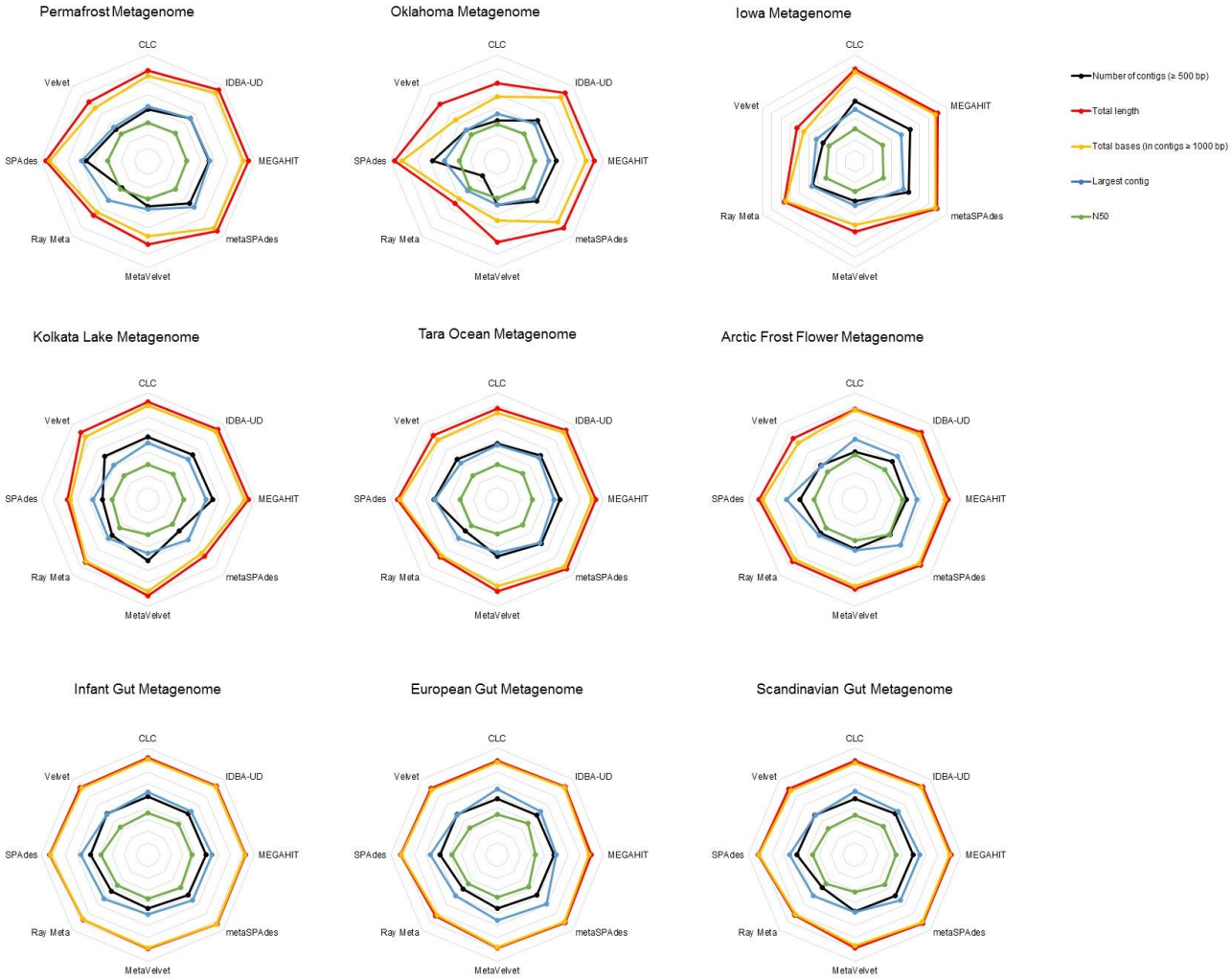

Figure 3. Proposed workflow to select a metagenome assembler based on the research question, the computational resources available and the bioinformatic expertise of the researcher.