Capturing variation in metagenomic assembly graphs with MetaCortex

Samuel Martin[1,#], Martin Ayling[1,#], Livia Patrono[2], Mario Caccamo[3], Pablo Murcia[4] and Richard M. Leggett[1,*]

[1] Earlham Institute, Norwich, NR4 7UZ, UK
[2] Robert Koch Institute, Nordufer 20, 13353 Berlin, Germany
[3] NIAB EMR, New Road, East Malling, Kent ME19 6BJ, UK
[4] MRC-University of Glasgow Centre for Virus Research, Glasgow, G61 1QH, UK

* To whom correspondence should be addressed. Email: richard.leggett@earlham.ac.uk
[#] Authors contributed equally.

ABSTRACT

The assembly of contiguous sequence from metagenomic samples presents a particular challenge, due to the presence of multiple species, often closely related, at varying levels of abundance. Capturing diversity within species, for example viral haplotypes, or bacterial strain-level diversity, is even more challenging. We present MetaCortex, a metagenome assembler based on data structures from the Cortex *de novo* assembler. MetaCortex captures intra-species diversity by searching for signatures of local variation along assembled sequences in the underlying assembly graph and outputting these sequences in sequence graph format. MetaCortex also implements a novel assembly algorithm for representing intra-species diversity in standard linear format. We show that MetaCortex produces accurate assemblies with higher genome coverage and contiguity than other popular metagenomic assemblers on mock viral communities with high levels of strain level diversity, and on simulated communities containing simulated strains. We also show that accuracy can be increased further by using the sequence graph produced by MetaCortex to create highly accurate single contig sequences.

INTRODUCTION

The well documented increase in yield and reduction in cost of DNA sequencing technologies has led to a rapid increase in the use of shotgun approaches for studying metagenomic samples [1]. A first analysis step is often taxonomic classification of reads by comparison with reference databases. However, deeper analysis is enabled by assembling sequence data to form longer contiguous sequence (contigs). Such assembly may facilitate improved classification, clustering of sequences (particularly where reference genomes are unavailable), or analysis at the scale of genes.

There are two fundamental approaches for the assembly of sequencing data: Overlap-layout-consensus (OLC) assembly, and de Bruijn graph assembly. In OLC assembly, each read is compared to every other read, and reads that overlap well are merged together to form contigs. The de Bruijn graph technique utilises directed graphs to represent the k-mers (short sequences of length k) present in a set of reads, and this representation of the read set turns the assembly problem into a graph traversal problem. Traversing a graph has a lower order of time complexity than OLC, and so the computational time to perform an assembly can be significantly reduced. Furthermore, the amount

of memory required to build the graph is proportional to the total k-mers present in the sample, rather than the total number of reads. This is particularly important as new generations of sequencing technologies are producing ever greater amounts of data.

Metagenomic assembly presents several challenges beyond those of de novo genomic assembly [2]. The main additional difficulties are unknown diversity and unknown species abundance. Diverse species may not share k-mers, particularly if k is large, and in this case it can be inferred that subgraphs representing distinct species are mostly disconnected and existing assembly techniques from de novo genome assembly may be used on each disconnected subgraph. However, it is often not the case that a metagenomic sample consists solely of diverse species, and separating closely related species, or many strains of a single species, is a difficult task, particularly at low abundance.

The problem of capturing below-species level diversity in an assembly presents an even bigger challenge. The genomes of different strains from a single species can differ by single nucleotide polymorphisms (SNPs), large-scale structural variation, and anything in between. In many ways, the challenges faced here are similar to those faced when separating haplotypes in de novo assembly, where much of the same sequence is shared between haplotypes. In this case, one can use the ploidy of the organism and the k-mer coverages to guide the assembly. However, in the metagenomic case we do not necessarily know what the expected coverage of each strain is, and in low abundance cases it will be difficult to distinguish a SNP from a sequencing error. The challenge is at its most stark when studying viral metagenomes. Due to their short replication times, large population sizes and lack of proofreading mechanisms (Coronaviruses are a notable exception [4]), viruses can evolve extremely rapidly. As such, viruses are often referred to as a quasispecies consisting of a set of related strains; and the genome sequence of a strain is sometimes referred to as a haplotype [5].

Due to the throughput of current generation sequencing technologies, it is possible to have a metagenomic data set consisting of several terabytes of reads. Many assemblers are incapable of assembling such large datasets within realistic time and memory constraints. A common strategy to make the problem tractable is to randomly subsample the read set to obtain a much smaller one which can be assembled. This process will usually not affect the assemblies of high abundance species in the sample, but there is evidence that the assemblies of the low abundance species will be incomplete and of poorer quality [3].

Many current short-read metagenomic assemblers utilise the de Bruijn graph paradigm. Popular examples include Ray Meta [6], MEGAHIT [7], MetaVelvet [8], and metaSPAdes [9]. A key part of the implementation of all these tools is to collapse an assembly graph into linear sequences (usually in the form of a FASTA file). This facilitates easy downstream analysis, but the act of converting the assembly graph into linear sequence has the effect of removing the understanding of sequence diversity that is implicit in the graph. Several assemblers have been created specifically for the assembly of viral quasispecies, such as SAVAGE [10] and VICUNA [11]. These tools are able to

assemble the haplotypes present in an isolated sample but may be less effective at assembling quasispecies from metagenomic samples. Within the (single-) genome assembly world, there is a growing awareness of the importance of capturing the genome graph in the output from assembly tools. This has resulted in the development of the FASTG format [12] and, more recently and with wider adoption, the graphical fragment assembly (GFA) format for assembly graph files [13]. These have been implemented in a number of tools including recent versions of ABySS [14], SPAdes [15] and SDG [16]. However, as yet, specialist metagenomic assembly tools lack the ability to capture variation through GFA output.

Here we introduce MetaCortex, a de Bruijn graph metagenomic assembler that combines the generation of assembly graph files that preserve intra-species variation (e.g. viral haplotypes) with new graph traversal algorithms to output variant contig sequences. MetaCortex is built upon data structures developed for the Cortex assembler [17]. Whilst MetaCortex can be used to assemble any metagenomic dataset, we have developed features to specifically target metagenomic datasets with high levels of strain diversity (e.g. viruses), and to represent this diversity in the resulting assembly. MetaCortex captures variation by looking for signatures of polymorphisms in the de Bruijn graph constructed from the reads and represents this in sequence graph format (both FASTG and GFA v2), and the usual FASTA format. The sequence graph provides information on local variation, such as SNPs and indels, along each contig identified by MetaCortex. By using the efficient data structures from Cortex, MetaCortex is capable of utilizing all k-mers from large metagenomic datasets (given that enough memory is provided), and able to perform assemblies from these datasets on a single CPU in just several hours.

We show that MetaCortex is able to produce highly contiguous assemblies capturing almost all genome level diversity, and with a low level of misassemblies. By outputting sequence graph files, we were able to capture strain level diversity that is not present in the contigs and use this to manually assemble contigs that were specific to individual strains in a sample.

MATERIAL AND METHODS

MetaCortex uses Cortex's hash table structure to store k-mer information and to encapsulate the de Bruijn graph structure. For reasons of memory efficiency, the maximum k-mer size must be specified when building MetaCortex. The default maximum value is 31, with 63, 95,127,160 or 192 also possible. The size of the hash table is user-defined and should be sufficient to contain the totality of the dataset being assembled (if when loading the reads into the hash table, it becomes full, the user is warned, but execution continues). Two parameters determine the hash table size - the width n, and the height b. The size of the hash table (the number of k-mers that can be stored) is then given by $s = b \times 2^n$. Multiplying these by the size of a hash table entry gives the total memory utilisation of the hash table, which will be reported by the software.

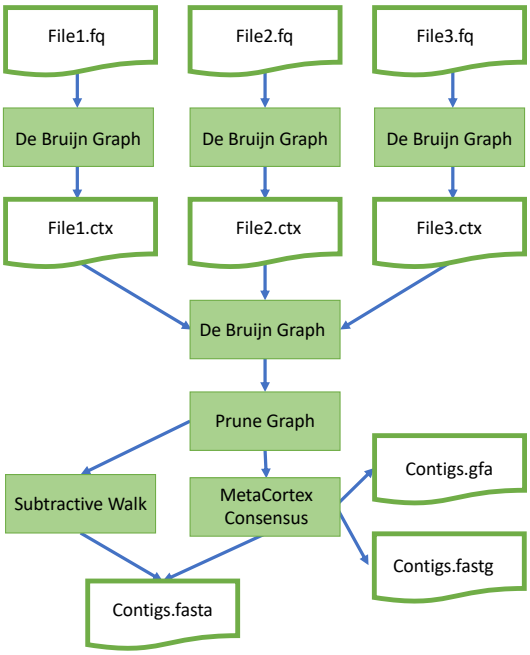

Figure 1: Flow chart depicting typical MetaCortex workflow.

After the construction of the hash table from the read files, a binary representation of the de Bruijn graph can be written to disc and this can then be used as the input to later assemblies. This can be used to speed up assembly time for subsequent assemblies of the same dataset, or to parallelise the reading of FASTA or FASTQ files. In the latter case, the individual CTX files can be merged to construct a de Bruijn graph from the whole read set. Figure 1 depicts a typical workflow.

The main innovation that MetaCortex introduces is two new graph traversal algorithms for metagenomic datasets. The core algorithm, MetaCortex Consensus (MC) is able to produce FASTA, GFA, and FASTG outputs. This is the option that should be selected to obtain sequence graphs. The Subtractive Walk algorithm (SW) produces only FASTA output, but performs much faster than MC, and attempts to write a single sequence for each variant that is detected. The user can also choose to only write unitigs (the maximal paths where each inner node has degree 2) as the FASTA output, by specifying the (existing) PerfectPath algorithm.

The MetaCortex approach

First, reads are decomposed into k-mers, and a de Bruijn graph is constructed; this graph (which will likely consist of several disconnected subgraphs) can be initially pruned to remove nodes which form short 'tips', or which fail to meet a minimum level of coverage. Tips up to a length of 100 are pruned by default, and the default minimum coverage is 2 (these values can be modified with command line options). Following construction of the de Bruijn graph, one of the following metagenomic traversal algorithms is executed. Formal descriptions of the algorithms can be found in the Supplementary Materials.

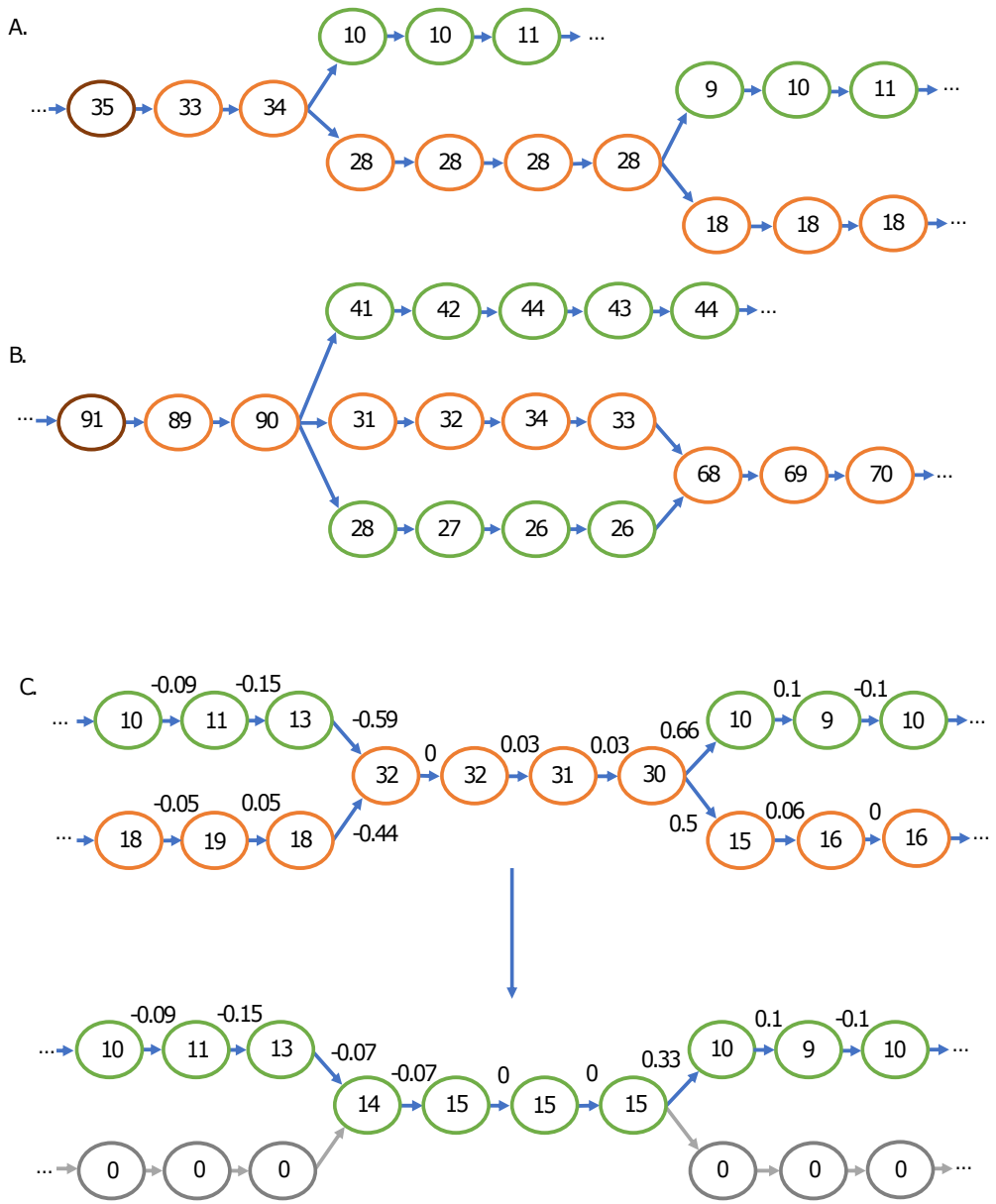

Figure 2: Depictions on de Bruijn graphs, with the coverage for each node represented. **A.** The path chosen by MetaCortex Consensus, with the highest coverage node highlighted in dark orange, and the chosen path in orange. **B.** The path chosen when two of the branches form a bubble. Because the two bubble branches, added together, represent a combined higher coverage than the top branch, a route through the bubble is selected for the path. **C.** The progression of the Subtractive Walk algorithm. The numbers above the edges are the normalised coverage difference δ. The first graph is before coverage subtraction, with the path chosen highlighted in orange, and the minimum coverage node highlighted in dark orange. The second graph is after coverage subtraction. Nodes belonging to exactly one path are removed (shown in grey), whilst nodes that are shared between paths remain with reduced coverage.

MetaCortex Consensus algorithm

This algorithm seeks to find consensus paths through each connected component of the overall graph structure. For each node in the graph, the connected component containing this node is explored to find the highest coverage node and to determine the size of the component. If it is sufficiently large, it is traversed starting at the highest coverage node. At each branching point in the graph, the branch with the highest coverage is favoured if it meets a minimum coverage threshold (Fig. 2a). However, should the paths from two other branches later join together (so that they form a bubble) and have higher coverage collectively, then the highest coverage of these two branches is chosen (Fig. 2b). Traversal continues until a tip is reached, there are no unvisited nodes adjacent to the current node, or there are no branches of sufficiently high coverage.

Once a path has been identified, the sequence it represents is written out to a FASTA file. Coverage statistics for the path are included in the header line. If the user has selected to have GFA2/FASTG output, the path is traversed to identify polymorphisms. At each node along the path, any branches that meet a minimum coverage threshold (and are not part of the original path) are explored, depth-first, with the highest coverage node taken at any subsequence branches. If at any point we return to a node on the original path at a position after the original branch, then this new path is written as an alternative path in the GFA file. The path traversal then continues from where the alternative path joins the original path.

Next, each node from the connected component is removed from the graph, and a new connected component is explored. Thus, we obtain one contig for each sufficiently large connected component. However, if the '-M' flag is specified, only those nodes in the path are removed, the remaining nodes in the current component will be reconsidered, and we may obtain multiple contigs for each connected component.

Subtractive Walk algorithm

In the SW algorithm, each node is examined, and for any that meet a minimum coverage requirement, the connected component they are contained in is explored to find the node with locally maximal coverage. (Note that unlike in MC, the whole component may not be explored, in order to speed up the process.) From this node, the highest coverage path is obtained and written to the FASTA file, as in MC, except that at branches, the highest coverage branch is always taken (i.e. bubbles are disregarded).

Next, MetaCortex estimates the number of variants covering each node in the path. First, the lowest coverage node along the path is found and this is assumed to have one variant covering it. There may be more than one node with the same minimal coverage, in which case, the node closest to the end of the path is chosen. Then the path in each direction of the node is walked, and at each step, the quantity δ is calculated as

$$\delta = \frac{c_{\text{current}} - c_{\text{previous}}}{\max(c_{\text{current}},\ c_{\text{previous}})}$$

where c_{current} is the coverage of the node at the current step, and c_{previous} is the coverage of the node at the previous step. This results in a value between -1 and 1. If this value is less than $-\Delta_{\text{SW}}$ (a value determined by the user with option -W, and set to 0.8 by default) then the number of variants covering this node is incremented; if it is greater than Δ_{SW}, the number of variants covering this node is decremented or it remains at 1.

After assigning a value to each node estimating the number of variants covering that node in the path, the coverage of nodes is adjusted in the following way. Nodes with 1 variant covering them are reduced to 0, and are essentially removed from the graph. Nodes with more than 1 variant covering them are reduced by an amount which is linearly interpolated from the nearest 1-variant nodes before and after this node in the path (Fig. 2c). This process is repeated until all nodes have been examined.

This algorithm is based on the assumption that, for two adjacent nodes in the graph that represent k-mers that only appear consecutively in the metagenome (i.e. the first node has outdegree one, and the second node has indegree one), the change in coverage between them will be small compared to their coverage values. On the other hand, for adjacent nodes in the graph that represent k-mers that appear consecutively in the metagenome, but at least one of which also appears elsewhere (i.e. at least one of the outdegree of the first node and indegree of the second node is greater than one), the change in coverage between them may be significant compared to the coverage values. The relative change in coverage across the path is what the value δ measures. This assumption, however, is only true for samples that have been sequenced with shotgun metagenome sequencing, and for regions of low sequence complexity (for example, repeat regions) this may not be the case.

RESULTS

To test MetaCortex, we assembled read sets from two mock communities of twelve viruses, at varying abundance; a lab mix of five strains of HIV in equal abundance; and two simulated communities with high levels of strain variation. In each case we set the k-mer length equal to 63, and used MetaCortex v0.4.

To compare MetaCortex's performance with that of existing de-novo metagenome assemblers, we also assembled the same datasets using MEGAHIT (v1.1.1), Ray Meta (v2.3.1) with k-mer length equal to 63, MetaVelvet (v1.2.02) with k-mer length equal to 63, and metaSPAdes (v3.14). For the real sequence data, we used trim-galore v0.5.0 [18] (a wrapper script around Cutadapt [19]), to trim low quality bases and adapter sequence from the reads, with the flags --paired and --retain_unpaired where appropriate. Since this can break the pairing of paired end reads and result in single-ended reads, we assembled these sets using both "paired + single" and "single only" modes, for the assemblers where these were available.

Statistics on the assemblies were obtained using MetaQUAST [20], which by default considers only those contigs greater than 500bp in length. A position on a contig is considered a misassembly by MetaQUAST if either the left flanking sequence aligns over 1kbp away from the right flanking sequence on a reference genome; flanking sequences overlap on more than 1kbp; or flanking sequences align to different strands, chromosomes, or genomes. We set the flag “ambiguity-usage” equal to one, so that only the best alignment from each contig is used when calculating certain statistics. For each dataset, the exact commands used to produce each assembly and the full MetaQUAST report of all assemblies are available in the Supplementary Materials.

Assembly of sequenced mock viral communities

Two communities of twelve viruses, each containing two ssDNA viruses and ten dsDNA viruses, were assembled from mock viral community read sets made available for benchmarking purposes [21]. The first community, Mock A, consisted of the dsDNA viruses each (theoretically) at 9.82% abundance, and the ssDNA viruses at 0.92%. The read set consisted of 2x250bp paired end reads, with a total read count of 96m. Using MetaCortex (SW), we were able to assemble 99.95% of the viral genomes with only two misassemblies. Mismatch and indel rates were also very low, at 7.21 per 100kbp and 3.61 per 100kbp respectively. Surprisingly, both the ssDNA viruses at low abundance had 100% coverage, while the others ranged from 100% to 98.744%. Eight virus genomes were each assembled in a single contig, while the assembly of all other genomes ranged from 2 to 19 contigs.

Table 1 shows that MetaCortex recovers the highest genome fraction, in the fewest contigs, and with a high average NG50 (this was calculated as the average of the individual NG50’s for each species), whilst maintaining a very low error rate. Only MEGAHIT recovered a similar genome fraction across all species in the mock (Figure 3.B), but the assembly was less contiguous and had a much higher mismatch rate and misassembly rate. Notably, both assemblies by Ray Meta failed to assemble any of the genomes of the ssDNA phage alpha3, and the dsDNA phages PSA-HS1 and PSA-HS2. Figure 3.A shows the genome fraction recovered for each genome and each assembler.

The total aligned length of the assembly by MetaCortex was 836kbp, out of a total assembly length of 6.2Mbp. We used BLAST to query the longest unaligned contigs (greater than 100,000bp) against the nt database. This revealed that each had alignments to one of Pseudoalteromonas or *Cellulophaga baltica*. These were host species that the mock community was grown on, so this represents legitimate assembly of DNA present in the sample. Including these species in the reference list for MetaQUAST, we had a total aligned length of 5.1Mbp, of which 4.3Mbp aligned to *C. baltica* over 41 contigs (covering 82.35% of the *C. baltica* genome).

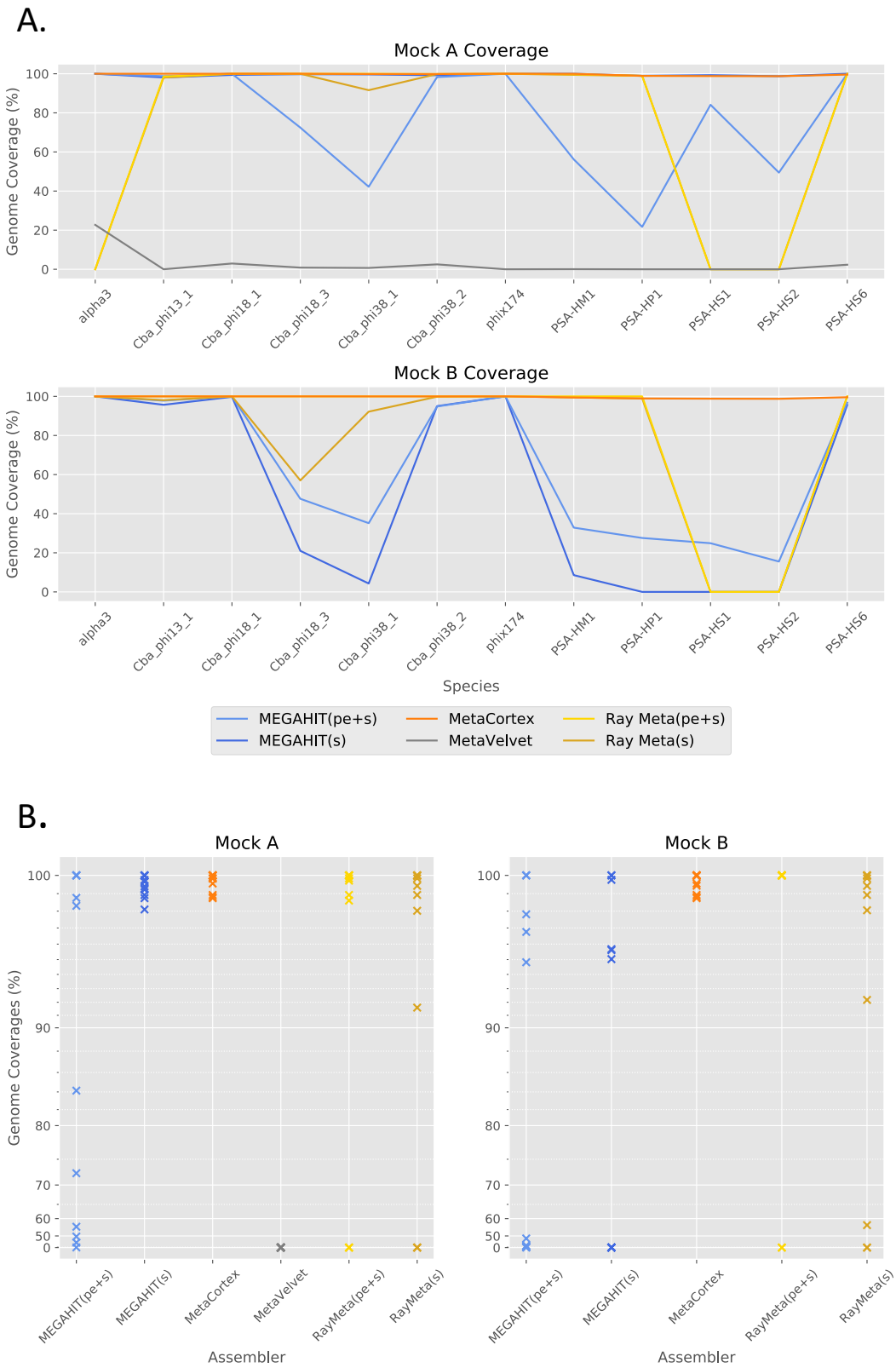

Figure 3: A. Scatterplot of genome coverages for Mock A and Mock B by assembler. **B.** Plot of genome coverage by viral species.

	Genome Fraction	Genomes Covered >50%	Contigs	Total Aligned Length	Average NG50	Misassemblies	Mismatch rate	Indel rate
MetaCortex	99.952%	12/12	437	835,722	47,119	2	7.21	3.61
MEGAHIT (s)	99.542%	12/12	5,479	823,716	34,161	7	32.25	3.78
MEGAHIT (pe+s)	70.582%	9/12	4,555	696,467	10,729	13	321.39	6.27
MetaVelvet	0.965%	0/12	1,139	15,815	1,099	5	1613.18	849.04
Ray Meta (s)	85.267%	9/12	1,792	522,876	36,444	0	3.46	2.88
Ray Meta (pe+s)	86.401%	9/12	1,521	605,259	49,370	0	3.60	3.03

Table 1: Summary statistics for assemblies of Mock A. The contigs field describes the number of contigs in the assembly of length greater than or equal to 500bp. Mismatch and indel rates are the number of occurrences per 100kbp.

The second community, Mock B, consisted of the dsDNA viruses at 3.51% abundance, and the two ssDNA viruses each at 32.47% abundance. Using MetaCortex (SW), we assembled 99.85% of the community, with 2 misassemblies (Table 2). As before, both ssDNA viruses had 100% genome coverage, while the others ranged from 100% to 98.744%. Seven virus genomes were assembled in a single contig, and the assembly of all other genomes varied from between two and eight contigs. As was the case in Mock A, approximately 4Mbp of unaligned contigs were found to align to *C. baltica*, covering 82.431% of its genome.

As with Mock A, we find that for Mock B, MetaCortex gives the largest genome fraction (Figure 3.B) in the fewest number of contigs, with a high average NG50. In this mock, Ray Meta produced a higher NG50, but with a much smaller genome fraction and more misassemblies. MEGAHIT struggled to assemble this set, after performing well on Mock A; in both cases the assemblies by MEGAHIT recovered a smaller genome fraction and contained more misassemblies than the others. MetaVelvet had failed to complete an assembly after 20 days (and failed to write any output for five days). Figure 3.A shows the genome fraction recovered for each genome and each assembler.

For both MockA and MockB, metaSPAdes failed after running out of memory, with 3TB of RAM available. This is beyond the memory limits for many researchers, and the usual strategy at this point is to assemble a much smaller subsample of reads instead. However, recent studies suggest that subsampling can drastically reduce the length and proportion of conserved genes in the subsequent assembly, when compared to the assembly of the full dataset [3].

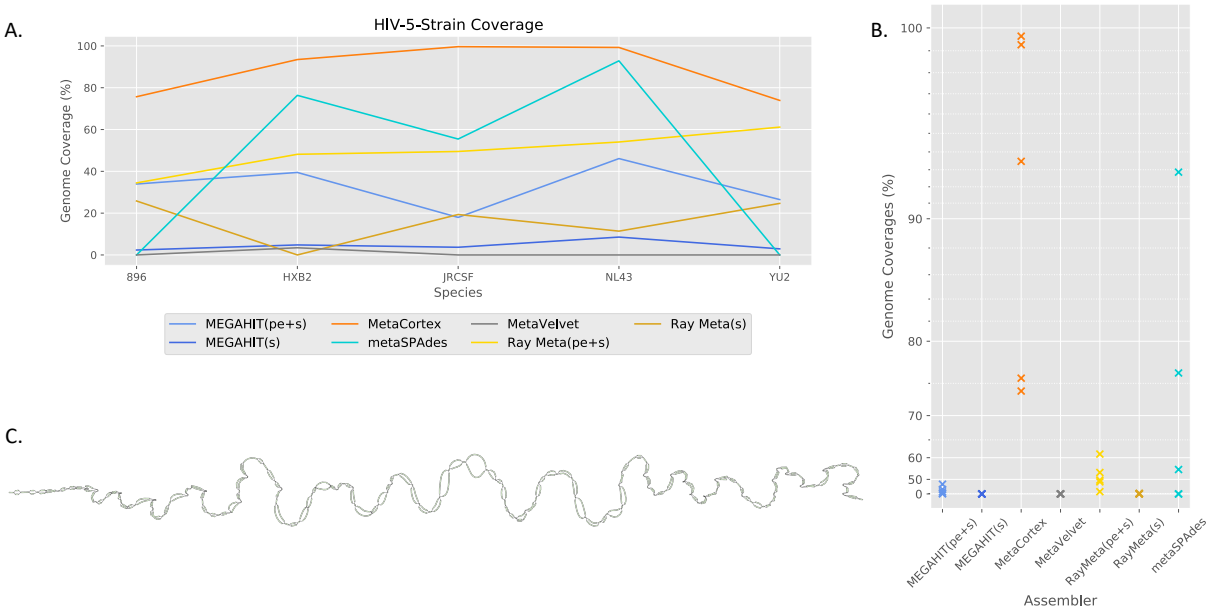

Figure 4. Assemblies of HIV-5-Strain dataset. **A.** Coverage for each viral species for each assembler. **B.** Scatterplot of genome coverages by assembler. **C.** Visualisation of GFA sequence graph output from MetaCortex. The sequence graph represents a single contig, with local variation shown by branching sequences.

	Genome Fraction	Genomes Covered >50%	Contigs	Total Aligned Length	Average NG50	Misassemblies	Mismatch Rate	Indel Rate
MetaCortex	99.854%	12/12	990	904,325	49,885	2	9.19	3.28
MEGAHIT (s)	39.058%	6/12	4,092	361,487	7,112	5	56.19	5.45
MEGAHIT (pe+s)	55.673%	6/12	3,682	589,734	14,432	27	677.54	7.35
Ray Meta (s)	81.307%	10/12	1,645	498,289	34,803	0	5.44	3.02
Ray Meta (pe+s)	87.647%	10/12	1,361	1,302,931	104,024	4	5.23	4.48

Table 2: Summary statistics for assemblies of Mock B.

Assembly of sequenced HIV lab mix

In order to evaluate the Subtractive Walk algorithm's ability to capture variants, we assembled an Illumina read set sequenced consisting of five well-studied strains of HIV-1 (89.6, HXB2, JR-CSF, NL4-3, YU2) in equal abundance. Reads and reference sequences for each strain were made available in [22]. Each strain was between 93% and 97% identical to the other strains, making this dataset particularly difficult to assemble. After adapter trimming, the read set contained 308Mbp of sequence, representing coverage of 6,367x.

So that sequence shared between strains would be included in the assembly of each strain, we set the SW delta parameter to 0.4 (from the default 0.8) and the min coverage parameter to 25 (from the default 10). With these values, MetaCortex was able to assemble 89.818% of the target genomes across 56 contigs (with a total length of 86,536bp) and no misassemblies.

We also assembled this dataset with MEGAHIT, Ray Meta, metaSPAdes, and MetaVelvet. Each of these tools produced assemblies with genome fractions far below that produced by MetaCortex (0.69% to 50.5%, Figure 4) and in some cases with higher misassembly rates (Table 3). Only Ray Meta produced assemblies with a significantly lower mismatch rate, but this represented only 16% of the genome fraction.

	Genome Fraction	Contigs (>= 500bp)	Total Aligned Length	Misassemblies	Mismatch Rate	Indel Rate
MetaCortex	89.818%	56	85,845	0	2464.62	110.46
MEGAHIT	4.465%	88	2,198	2	2453.70	46.30
MEGAHIT (pe+s)	32.848%	63	12,822	0	3511.20	62.92
metaSPAdes	48.422%	18	55,804	0	2817.26	123.79
MetaVelvet	0.690%	2	333	0	2395.21	299.40
Ray Meta	16.230%	10	7,855	0	12.74	12.74
Ray Meta (pe+s)	50.487%	31	27,560	0	970.28	98.26

Table 3: Summary statistics for assemblies of 5-strain HIV mix.

GFA output captures strain level variation and facilitates visualisation

We used MetaCortex to create a sequence graph of the assembly, as an alternative method to capture strain-level diversity. Using the MC algorithm, MetaCortex constructed a single contig 9,101bp long, and a sequence graph containing polymorphisms along the contig. In order to capture only true variation between the strains, we set the minimum coverage to 1,000x (slightly less than the total coverage divided by the number of strains). To capture as many individual SNPs as possible, we set $k = 31$. Figure 4.C shows the resulting sequence graph, as visualised by GfaViz [23].

To demonstrate that this sequence graph contains more information than the corresponding contig, we constructed one contig for each of the five strains present in the sample. First, we created a mapping of each sequence in the graph to a database of the reference genomes using BLAST. Then, we constructed contigs in the following way. For each strain, a walk is performed through the sequence graph, where at each branch, the branch whose sequence has the highest score for the strain is chosen. Scores are calculated as the mapping identity multiplied by the proportion of the

alignment covering the query (in ambiguous cases, the branch corresponding to the original contig is chosen). Each walk corresponds to a contiguous sequence, which forms the strain-level contig. A python script to parse the GFA file and BLAST mapping and construct the individual strain assemblies is available in the MetaCortex repository, under scripts/strain_assembly.py.

Next, we used dnadiff, part of MUMmer [24] to compare each strain-specific contig and the original contig to the strain's reference sequence. We found that the strain-specific contigs had alignments with a higher average identity to the corresponding reference strain and contained far fewer SNPs and indels (Table 4).

	NL4-3		**JR-CSF**		**HXB2**		**89.6**		**YU2**	
	Original	Strain specific	Original	Strain specific	Original	Strain specific	Original	Strain specific	Original	Strain specific
Total Length	9,101	9,089	9,101	9,106	9,101	9,089	9,101	9,094	9,101	9,089
Aligned Bases Ref	100%	100%	100%	100%	100%	100%	99.98%	99.98%	100%	100%
Aligned Bases Query	100%	100%	100%	100%	100%	100%	100.00%	100%	100%	100%
Avg Identity	96.75	98.00	96.46	97.95	96.26	97.43	94.46	97.05	94.99	95.77
Total SNPs	223	133	237	111	281	191	374	219	362	292
Total Indels	62	24	50	33	60	36	53	22	67	51

Table 4: Summary statistics comparing each strain-specific assembly to the base assembly.

Assembly of simulated viral and bacterial datasets

We tested MetaCortex's performance on two simulated communities, one viral and one bacterial, both with a high amount of strain level variation. For both communities we used the software CAMISIM [25] to simulate the community composition, strain level variants, and 15 Gb of Illumina 2 × 150 bp paired end reads, with a HiSeq 2500 error profile. Using the reference genomes (both real and simulated) we were able to compare the performance of MetaCortex (with parameters min coverage 5, SW delta 0.8), MEGAHIT, Ray Meta, metaSPAdes and MetaVelvet, on these datasets. Since we assembled the simulated reads directly (i.e., without adapter trimming or other pre-processing), we used the paired-end assembly mode for Ray Meta and MEGAHIT. For the bacterial community, Ray Meta failed after running out of memory (3TB available).

The viral community consisted of six species: Human mastadenovirus F, Human herpesvirus 5, Human respiratory syncytial virus, Influenza B virus, Reovirus 3 and Zika virus; and four simulated strains of Influenza B virus. The composition of the community is shown in Figure 5.C. For this dataset, MetaCortex recovered the highest overall genome fraction, with individual genome fractions

ranging from 95.38% to 100%, and had no misassemblies. MEGAHIT recovered a similar genome fraction but had significantly more misassemblies and a higher mismatch rate. Both Ray Meta and metaSPAdes recovered smaller genome fractions, although with lower error rates (Table 5), and Ray Meta had the highest average NG50. Individual genome coverage for each assembler is displayed in Figure 5.B.

	Genome Fraction	Genomes covered > 50%	Contigs (>= 500bp)	Total Aligned Length	Average NG50	Misassemblies	Mismatch Rate	Indel Rate
MetaCortex	99.703%	10/10	5,554	1,215,870	26,929	0	289.11	4.87
MEGAHIT	98.709%	10/10	1,017	746,262	25,943	17	550.91	1.55
metaSPAdes	92.550%	8/10	23	334,243	26,771	0	54.12	0.83
MetaVelvet	0.315%	0/10	73	489	0	0	2755.27	0
Ray Meta	77.241%	3/10	14	531,198	51,957	3	10.59	0

Table 5: Summary statistics for assemblies of simulated viral community.

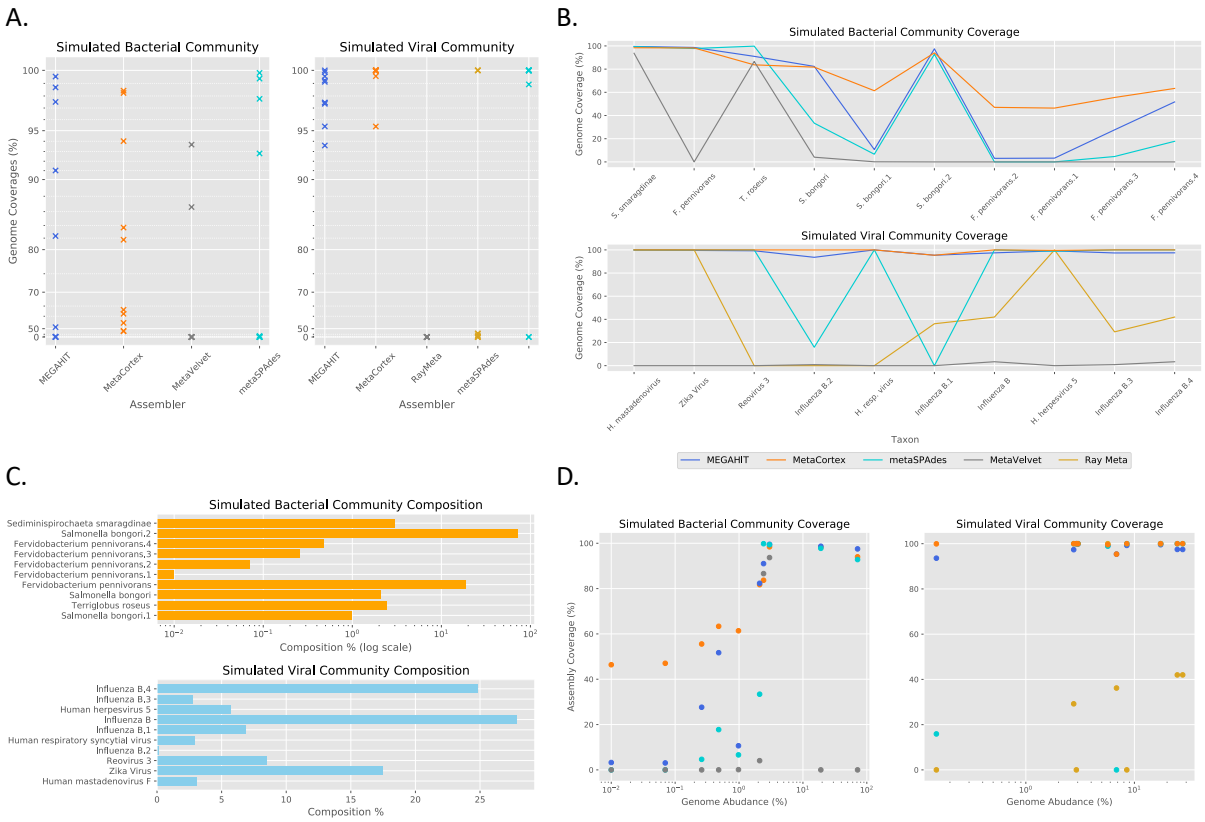

Figure 5: A. Scatterplot of genome coverages by assembler of simulated datasets. **B.** Coverage of each species in simulated datasets by assembler. **C.** Composition of simulated communities. **D.** Genome abundance against assembly coverage for assemblies of simulated datasets.

The bacterial community consisted of four species randomly chosen from the well-known MBARC-26 mock community [26]: *Terriglobus roseus*, *Salmonella bongori*, *Fervidobacterium pennivorans*, and *Sediminispirochaeta smaragdinae*; plus two strains simulated from *S. bongori*, and four strains simulated from *F. pennivorans* (Figure 5.C).

MetaCortex again recovers the highest genome fraction and produces highest NG50, but this time with the highest number of misassemblies. MetaCortex also has the highest minimum genome fraction for the individual strains, with the lowest being 46.4% (Figure 5.A). The next highest overall genome fraction is given by MEGAHIT, although in this case two of the simulated strains have a genome fraction of only 3%, and another with 10%. Full details can be seen in Figure 5.

The assembly by MetaVelvet has the lowest error rates, with no misassemblies, but fails to assemble any sequence for six of the genomes in the community, and for two of the genomes assembles less than 5%. The assembly by metaSPAdes also has low error rates but recovers significantly less of the genomes than MetaCortex and MEGAHIT, failing to assemble any of the least two abundant strains (F. pennivorans.1 and F. pennivorans.2) and less than 7% of both F. pennivorans.3 and S. bongori.1. The assembly by Ray Meta failed after running out of memory.

	Genome Fraction	Genomes covered > 50%	Contigs (>= 500bp)	Total Aligned Length	Average NG50	Misassemblies	Mismatch Rate	Indel Rate
MetaCortex	80.614%	8/10	3,764	25,367,807	127,616	139	588.90	6.90
MEGAHIT	64.187%	6/10	6,602	19,325,437	95,125	119	424.32	2.67
metaSPAdes	54.308%	4/10	273	16,668,491	125,853	1	110.58	2.13
MetaVelvet	26.885%	2/10	3,719	9,132,995	643	0	0.60	0.21

Table 6: Summary statistics for assemblies of simulated bacterial community.

DISCUSSION

Compared to the challenge of assembling a single isolate species, metagenomic assembly presents significant additional hurdles related to the presence of closely related species and their differing abundance in a sample. These challenges can be particularly obvious in viral communities, where a rapid evolutionary rate can make it especially difficult to distinguish between and within quasispecies. In recent years, a number of new assembly tools have attempted to tackle the challenges of metagenomic assembly using various heuristic approaches. To our knowledge, all have adopted approaches which reduce assembly graphs to sets of contigs, losing the variation captured by the underlying graph structures. We address this in MetaCortex, a new assembly tool that preserves strain interconnectedness by outputting sequence graph files as an alternative to contigs. In addition, a new Subtractive Walk algorithm enables MetaCortex to estimate the number of variants in a given

subgraph and to output representative contigs. This algorithm was able to recover a high proportion of five strains of HIV from a lab mixed community, and five strains (one real, four simulated) of influenza virus from a simulated viral community, without misassembly. The latter result is particularly encouraging, as several of these strains were at very low abundance in the sample.

We tested MetaCortex on a number of mock and simulated datasets which have known answers, allowing comparative evaluation against other common metagenomic assembly tools. We first evaluated MetaCortex's performance assembling published sequence data from two mock viral communities. We found that MetaCortex performs better than current popular metagenome assemblers on both communities, in terms of contiguity and genome coverage, whilst maintaining a low error rate that is comparable to or, in cases, significantly lower than the tools it was compared to. Our next experiment involved assembling sequence data from a mix of closely related HIV strains. Here we found that MetaCortex far outperforms the other assemblers in terms of genome coverage whilst maintaining a high contiguity and comparable error rate. Furthermore, by using the unique sequence graph output, we were able to construct a single contig for each strain which had a high identity mapping to the reference, and significantly fewer SNPs and indels when compared to the original assembly.

On simulated communities with high levels of strain variation, we found that MetaCortex consistently recovers a significantly higher genome fraction of the low abundance genomes compared to other assemblers (Figure 5.D). For the simulated viral community, MetaCortex recovered the highest overall genome fraction with the highest NG50 and had no misassemblies. For the bacterial community, MetaCortex again produced the highest NG50, but we observed a higher rate of misassembly compared to the other assemblers. This is not unexpected, as MetaCortex does not attempt to recognise regions of low sequence complexity, which are more common in bacteria than in viruses and can be a source of misassemblies. Future improvements to MetaCortex will look for signatures of low complexity regions in the de Bruijn graph, so that these can be traversed more carefully to avoid such misassemblies.

Overall, we found that MetaCortex consistently recovers a very high genome fraction when compared to other popular metagenome assemblers. In particular, our simulated datasets show that MetaCortex is especially effective at recovering the genomes of extremely low abundant species. For the assembly of viral communities, MetaCortex had low error rates comparable with the lowest of the other assemblers tested.

DATA AVAILABILITY

MetaCortex is open source and available to download from https://github.com/SR-Martin/metacortex. Full documentation, including instructions for installation, is available at https://metacortex.readthedocs.io/en/latest/. All assemblies, simulated reads, and simulated genomes

used in this paper have been deposited online on Zenodo, and can be found at DOI 10.5281/zenodo.4957491.

SUPPLEMENTARY DATA

Supplementary Data are available online.

ACKNOWLEDGEMENTS

We are grateful to Kirsten McLay for helping to enable some of the laboratory work that led to this publication. This research was supported in part by the NBI Computing Infrastructure for Science Group, which provides technical support and maintenance to Earlham Institute's high-performance computing cluster and storage systems.

FUNDING

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation, through Responsive Mode award BB/M004805/1, Core Capability Grant BB/CCG1720/1, Core Strategic Programme Grant BB/CSP1720/1.

CONFLICT OF INTEREST

None stated.

REFERENCES

1. Mitchell,A.L., Scheremetjew,M., Denise,H., et al. (2017) EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res., 46(D1), D726–35.
2. Ayling,M., Clark,M.D. and Leggett,R.M. (2020) New approaches for metagenome assembly with short reads. Briefings in Bioinformatics, 21(2), 584-594.
3. Cattonaro,F., Spadotto,A., Radovic,S. and Marroni,F. (2018) Do you cov me? Effect of coverage reduction on metagenome shotgun sequencing studies. F1000research 7:1767, 10.12688/f1000research.16804.4
4. Denison,M.R., Graham,R.L., Donaldson,E.F., Eckerle,L.D. and Baric,R.S. (2011) Coronaviruses. RNA Biology, 8:2, 270-279.
5. Gregori,G., Perales,C., Rodriguez-Frias,F., Esteban,J.I., Quer,J. and Domingo,E. (2016) Viral quasispecies complexity measures. Virology, 493, 227-237.
6. Boisvert,S., Raymond,F., Godzaridis,É. et al. (2012) Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol., 13, R122, 10.1186/gb-2012-13-12-r122
7. Li,D., Liu,C.M., Luo,R., Sadakane,K. and Lam,T.W.(2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 31(10), 1674-1676, 10.1093/bioinformatics/btv033
8. Namiki,T., Hachiya,T., Tanaka,H. and Sakakibara,Y. (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res., 40(20), e155, 10.1093/nar/gks678

9. Nurk,S., Meleshko,D., Korobeynikov,A. and Pevzner,P. (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Res., 27, 8824-834, 10.1101/gr.213959.116
10. Baaijens,J.A., Aabidine,A.Z.E., Rivals,E. and Schönhuth,A. (2017) De novo assembly of viral quasispecies using overlap graphs. Genome Res., 27(5), 835-848, 10.1101/gr.215038.116
11. Yang,X., Charlebois,P., Gnerre,S. et al. (2012) De novo assembly of highly diverse viral populations. BMC Genomics, 13, 475, 10.1186/1471-2164-13-475
12. FASTG - An expressive representation for genome assemblies. http://fastg.sourceforge.net/ Accessed 25 May 2021.
13. GFA: Graphical Fragment Assembly (GFA) Format Specification. https://github.com/GFA-spec/GFA-spec/ Accessed 25 May 2021.
14. Jackman,S.D., Vandervalk,B.P., Mohamadi,H., Chu,J., Yeo,S., Hammond,S.A., Jahesh,G., Khan,H., Coombe,L., Warren,R.L. and Birol,I. (2017) ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res., 27, 768-777.
15. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D., Pyshkin,A.V., Sirotkin,A.V., Vyahhi,N., Tesler,G., Alekseyev ,M.A. and Pevzner,.PA. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol., 19(5), 455-77.
16. Yanes,L., Accinelli,G.G., Wright,J., Ward,B.J. and Clavijo,B.J. (2019) A Sequence Distance Graph framework for genome assembly and analysis. F1000Research ,8,1490.
17. Iqbal,Z., Caccamo,M., Turner,I., Flicek,P. and McVean,G. (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat. Genet., 44, 226–232, 10.1038/ng.1028
18. Krueger,F. Trim Galore. Retrieved from https://github.com/FelixKrueger/TrimGalore, Nov 2020.
19. Martin,M., (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 17(1), 10-12, SSN 2226-6089, 10.14806/ej.17.1.200
20. Mikheenko,A., Saveliev,V. and Gurevich,A. (2016) MetaQUAST: evaluation of metagenome assemblies. Bioinformatics, 32(7), 1088-1090, 10.1093/bioinformatics/btv697
21. Roux,S., Solonenko,N.E., Dang,V.T., Poulos,B.T., Schwenck,S.M., Goldsmith,D.B., Coleman,M.L., Breitbart,M. and Sullivan,M.B. (2016) Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. PeerJ, 4, e2777.
22. Di Giallonardo,F., Töpfer,A., Rey,M., Prabhakaran,S., Duport,Y., Leemann,C., Schmutz,S., Campbell,N., Joos,B., Lecca,M., et al. (2014) Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res., 42, e115.
23. Gonnella,G., Niehus,N. and Kurtz,S. (2019) GfaViz: flexible and interactive visualization of GFA sequence graphs. Bioinformatics, 35(16), 2853–2855, 10.1093/bioinformatics/bty1046
24. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004). Versatile and open software for comparing large genomes. Genome Biology, 5, R12.
25. Fritz,A., Hofmann,P., Majda,S. et al. (2019) CAMISIM: simulating metagenomes and microbial communities. BMC Microbiome, 7, 17, 10.1186/s40168-019-0633-6

26. Singer,E., Andreopoulos,B., Bowers,R.M., Lee,J., Deshpande,S., Chiniquy,J., Ciobanu,D., Klenk,H.P., Zane,M., Daum,C., Clum,A., Cheng,J.F., Copeland,A. and Woyle,T. (2016) Next generation sequencing data of a defined microbial mock community. Sci. Data, 3:160081, 10.1038/sdata.2016.81