JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 24, Number 9, 2017
Mary Ann Liebert, Inc.
Pp. 874–881
DOI: 10.1089/cmb.2017.0008

Assessing the Impact of Assemblers on Virus Detection in a De Novo Metagenomic Analysis Pipeline

DANIEL J. WHITE,[1] JING WANG,[2] and RICHARD J. HALL[3]

ABSTRACT

Applying high-throughput sequencing to pathogen discovery is a relatively new field, the objective of which is to find disease-causing agents when little or no background information on disease is available. Key steps in the process are the generation of millions of sequence reads from an infected tissue sample, followed by assembly of these reads into longer, contiguous stretches of nucleotide sequences, and then identification of the contigs by matching them to known databases, such as those stored at GenBank or Ensembl. This technique, that is, de novo metagenomics, is particularly useful when the pathogen is viral and strong discriminatory power can be achieved. However, recently, we found that striking differences in results can be achieved when different assemblers were used. In this study, we test formally the impact of five popular assemblers (MIRA, VELVET, METAVELVET, SPADES, and OMEGA) on the detection of a novel virus and assembly of its whole genome in a data set for which we have confirmed the presence of the virus by empirical laboratory techniques, and compare the overall performance between assemblers. Our results show that if results from only one assembler are considered, biologically important reads can easily be overlooked. The impacts of these results on the field of pathogen discovery are considered.

Keywords: algorithms, assemblers, de novo metagenomics, pathogen discovery, test.

INTRODUCTION

Viral infection has a huge impact on human health, wildlife conservation, and agriculture. Finding novel viruses is notoriously difficult due to difficulty in culturing viruses (Fuhrman and Campbell, 1998) and the lack of an appropriate phylogenetic marker (Rohwer and Edwards, 2002). The metagenomic approach, which attempts to obtain sequence information from all genomes present in an environmental sample,

[1] Landcare Research, Auckland, New Zealand.

[2] Institute of Environmental Science and Research at the National Centre for Biosecurity and Infectious Disease, Upper Hutt, New Zealand.

[3] Animal Health Laboratory, Investigation and Diagnostic Centres and Response, Ministry for Primary Industries—Manatū Ahu Matua, Upper Hutt, New Zealand.

has gained in popularity due to detection and, in some cases, even the generation of complete genomes of novel uncultivated viruses. A typical metagenomic process includes the extraction of all nucleic acids from an environmental sample, the sequencing of all genomes present to create many random reads, and then followed by the identification of these reads by homology matching to nonredundant nucleotide or protein sequence databases using appropriate search algorithms such as BLASTN or BLASTX (Camacho et al., 2009) and RAPSEARCH2 (Zhao et al., 2012), respectively. The first study to use this technique was published in 2002 and, using cloning and Sanger sequencing, successfully revealed viral communities in sea water, including the identification of numerous novel viruses (Breitbart et al., 2002).

Since those early days, sequencing technology and bioinformatics capability have advanced significantly. With the advent of the current suite of sequencing-by-synthesis technologies, the number of randomly generated reads has moved from thousands to millions, which has led to a huge increase in power and sensitivity of de novo metagenomics, or the assembly of novel genomes with no reference genome (Mokili et al., 2012; Hall et al., 2014). This has helped to extend de novo metagenomics to the emerging field of pathogen discovery, which focuses on the discovery of novel viral etiological agents of diseases for which little or no background information is available (Cox-Foster et al., 2007; Smits et al., 2013; White et al., 2015).

The new sequencing technologies have also increased drastically the computing and bioinformatics resources required to handle such large data sets. For example, homology searching is an iterative and memory-exhaustive step, and if the researcher decides to match all the reads from a metagenomic sequencing run to the entire 185,019,352 sequences stored at GenBank (June, 2015), the computing resources required quickly becomes nontrivial. It is therefore often necessary to assemble the millions of sequence reads into longer contiguous stretches of DNA, or contigs. This has the benefits of reducing the number of homology searches that need to be done, thereby saving time and computing resources, and also increasing taxon assignment consistency and accuracy (Garcia-Etxebarria et al., 2014; Vazquez-Castellanos et al., 2014).

There is now a suite of assemblers that are used in, or have been designed for, metagenomic studies (Namiki et al., 2012; Haider et al., 2014; Roux et al., 2014; Guo et al., 2015). These use different assembly algorithms, such as the overlap–layout–consensus graph (OLC) and the de Bruijn graph (DBG) algorithms (Li et al., 2012). OLC adheres closely to the Lander–Waterman model of genomic mapping (Lander and Waterman, 1988) and forms a consensus contig from reads that overlap at a given threshold, for example, NEWBLER (Margulies et al., 2005), CELERA (Myers et al., 2000), and MINIMO (Treangen et al., 2011). The DBG algorithm, on the contrary, is a derivation of the Lander–Waterman model, which breaks reads into smaller kmers before overlapping to produce a final consensus contig, that is, the ''graph,'' for example, VELVET (Zerbino and Birney, 2008) and METAVELVET (Namiki et al., 2012). A third lesser used algorithm is the read probabilistic model, which uses a generative probabilistic model to construct reads, for example, GENOVO (Laserson et al., 2011). It is known that these algorithms behave differently dependent on the parameters of the sequence data set, such as sequence read depth, length, and the frequency of repeats (Li et al., 2012; Garcia-Etxebarria et al., 2014).

Despite the impact different assembly algorithms can have on taxa identification, and a known variation in performance of assemblers dependent on the characteristics of the data set; relatively little has been done to assess assembler performance on viral metagenomic data sets (Smits et al., 2014; Vazquez-Castellanos et al., 2014). One recent study, for example, has shown marked differences in assemblers to generate informative contigs, based on the length of contigs and the construction of problematic chimeric contigs (Vazquez-Castellanos et al., 2014). While these studies are informative and valuable, they have relied heavily on simulated data sets or have focused on the complete assembly and annotation of known viral genomes (Smits et al., 2014). The impact of assemblers on virus detection in an empirical setting, where viruses will be vastly outnumbered by other organisms present, remains uncertain. In a recent study of our own in which we used two different assemblers, MIRA and VELVET, we found that while the pipeline using VELVET assembled an entire virus genome from the metagenomic data set, later confirmed empirically in the laboratory, MIRA was only able to generate small contigs for which homology search results were buried among other results (White et al., 2016). For a technique that can rely on ''stand-out'' results, this has significant consequences.

In this study, we have tested the effectiveness of five assemblers, which are either well established in the field—VELVET (Zerbino and Birney, 2008), METAVELVET (Namiki et al., 2012), and MIRA (Chevreux et al., 1999)—or have been more recently developed—SPADES (Bankevich et al., 2012) and OMEGA (Haider et al., 2014)—to detect a viral genome in a metagenomic data set in which we know it exists. These assemblers use a strict DBG algorithm (VELVET, METAVELVET) or a DBG variation (SPADES), an OLC graph (OMEGA) or combined algorithms that use both OLC and greedy algorithms (MIRA). We tested their performance on a data set that we generated previously (White et al., 2016) as part of a study to discover viral pathogens in fecal matter of rowi kiwi (Apteryx rowi) that showed signs of minor verminous dermatitis (Gartrell et al., 2015). In that study, we revealed the presence of a novel virus, rowi kiwi circovirus-like virus (rowi kiwi CVLV), using a de novo metagenomic pipeline, which we subsequently confirmed using polymerase chain reaction and Sanger sequencing. In this study, overall performance of the assemblers was assessed using the rowi kiwi data set by investigating the number of reads that contribute to contigs, the number and length of contigs and, critically, the percentage of contigs that get taxonomically assigned. In particular, we have answered the question, can different assemblers be relied upon to detect a target viral genome in a metagenomic data set?

METHODS

The metagenomic data set used for this study originates from the fecal matter of eight rowi kiwi showing signs of disease and was generated as part of an earlier study (White et al., 2016). The raw metagenome data set consisted of 16,435,262 250 bp paired-end reads. The quality of reads was checked using FASTQC v0.10.1 (Babraham Institute, Cambridge) to ensure an average Phred score of 20 or greater, and sequences <100 bp were removed using the PRINSEQ-LITE v0.19.3 software package (Schmieder and Edwards, 2011). A threshold of 100 bp was chosen to (1) remove short reads, which tend to have an increased error rate; (2) decrease huge memory requirements imposed by homology searching numerous short, often uninformative, reads; and (3) improve the consistency of the data set between assemblers, as assemblers handle short reads differently. The cleaned data set contained 10,265,480 paired-end reads.

The assemblers chosen for this study are VELVET v1.2.10, METAVELVET v1.2.02, MIRA v4.0.2, SPADES v3.5.0, and OMEGA v1.0.2. A brief summary of the assemblers is given below and the parameters used for each assembler are summarized in Table 1.

MIRA

MIRA is a multipass DNA sequence assembler that uses the OLC and greedy algorithms to assemble prokaryote and small eukaryote genomes sequenced on most of the current sequencing technologies, within the same assembly if required. It can do de novo, hybrid, and true mapping assemblies (Chevreux et al., 1999). In this study, MIRA was run both excluding and including singletons (reads that pass quality filters but are not able to assemble into contigs).

VELVET

VELVET was one of the first assemblers to assemble short reads using the de bruin graph algorithm for de novo assembly in complex organisms, and implements error correction and repeat handling (Zerbino and Birney, 2008).

METAVELVET

METAVELVET is an extension of VELVET that attempts to control for highly abundant short reads that are common to metagenomes, which otherwise would be considered to be repeats. It deconstructs the de

Table 1. Assemblers Used and Their Parameters

Assembler	*Parameters*
MIRA	job = est,denovo,accurate parameters = COMMON_SETTINGS -NW:cnfs = no:cmrnl = no -GE:not = 12 SOLEXA_SETTINGS -CL:cpat = no
MIRA (singletons)	job = est,denovo,accurate parameters = COMMON_SETTINGS -NW:cnfs = no:cmrnl = no -GE:not = 12 SOLEXA_SETTINGS -AS:mrpc = 1-OUT:sssip = yes -CL:cpat = no
VELVET velveth:	kmer is 99 and used paired end reads
velvetg:	-cov_cutoff auto -exp_cov auto -unused_reads yes
METAVELVET velveth:	kmer is 99 and used paired-end reads
velvetg:	-cov_cutoff auto exp_cov auto -unused_reads yes
SPADES	-k 21,33,55,77,99,127—careful—pe1-1—pe1-2—cov-cutoff auto
OMEGA	-pe -l 100

bruin graphs made by VELVET and builds scaffolds on the subgraphs, with the aim of generating longer, more complete genomes (Namiki et al., 2012).

SPADES

SPADES is a de bruin graph assembler that uses multiple kmers in one run to deal with the problem of chimeric sequences being generated in the assembly of small prokaryotic genomes. Initially designed for single-cell bacterial genomes, it has recently been applied to virome analysis (Bankevich et al., 2012).

OMEGA

OMEGA is an OLC graph de novo assembler that uses a user-specified overlap length to assemble Illumina data into long reads (Haider et al., 2014).

Homology searching to the GenBank nonredundant nucleotide database (downloaded June 29, 2015) was done using BLASTN with search results restricted to five matches and a minimum E-value threshold of 1×10^{-3} (Camacho et al., 2009). Taxon assignment was visualized in MEGAN v5.10.5 (Huson et al., 2007) (software available from www-ab.informatik.uni-tuebingen.de/software/megan). All assembly and homology searches were performed on the Pan computer cluster at NeSI (National eScience Infrastructure), Auckland, New Zealand, using Intel E5-2680 core processing units, operating at 2.7 to 2.8 GHz on Ivy Bridge architecture with quad date rate (QDR) InfiniBand interconnect. One hundred eighty gigabytes RAM was assigned to all assemblies.

RESULTS

All results are summarized in Table 2. Assembly took as long as 1 day, 11 hours, and 6 minutes with SPADES, and as little as 14 minutes with VELVET. MIRA (singletons) generated the largest metagenome (consensus genome size 370,456,167 bp) and assembled the most contigs (1,616,750), METAVELVET generated the second largest on both counts, while OMEGA generated the smallest metagenome (8,196,640 bp) and assembled the least number of contigs (16,269). The largest contig was made by MIRA (with singleton option off) at 32,359 bp. while VELVETs largest contig was the smallest of all assemblers (2418 bp). The largest N50 was in OMEGAs metagenome, while the smallest was in MIRA (singletons).

The most number of contigs that were assigned a taxon was achieved using MIRA (singletons) with 1,353,011 contigs, while METAVELVET achieved the second largest amount (878,000) in accordance with the total number of contigs assembled. OMEGA achieved the least number of contigs with an assigned taxon (11,962) and also had the lowest proportion of contigs with assigned species (73.5%). The maximum number of virus species identified was found using VELVET (44), second was METAVELVET (39), and the least was with OMEGA (8). The maximum number of viruses identified that were not shared with another assembler was found with MIRA (singletons) (five—this number increases to six when both MIRA runs are considered together) and the second largest was with VELVET (3). Interestingly, 16 further viruses are specific to VELVET and METAVELVET when they are considered together. For all other assemblers, viruses were found in more than one instance.

As a means of assessing the efficacy of assemblers to construct contigs, we compared BLAST scores across pipelines for the highest scoring viral contig found across all assemblers. The highest blast score for a viral contig was 4744.0 achieved using SPADES, for a contig of length 2765 bp and with homology to Sewage-associated circular DNA virus 27 (GenBank access no. KM821762). All assemblers generated a contig that matched this virus. VELVET and METAVELVET assembled the shortest contig (2344 bp) and achieved the weakest (but still very strong) statistical support for a match to KM821762, while MIRA assembled the longest contig (2855 bp).

Finally, to determine how each assembler performed with sequence we know to be present (rowi kiwi CVLV), we blasted the complete genome of rowi kiwi CVLV (GenBank access no. KP202150) against each of the assemblies. Matches were found in all assemblies. The largest region of the rowi kiwi CVLV genome that matched a contig was found for the assembly generated by VELVET (100% homology over 1952 bp), while the smallest region was found for the assembly generated by METAVELVET (100% homology over 1580 bp). We also assessed how well contigs matched their closest homologue archived in GenBank at the time of study, Meles meles circovirus-like virus (GenBank access no. JQ085285). Only

TABLE 2. PERFORMANCE OF FIVE DIFFERENT ASSEMBLERS IN VIRUS DISCOVERY

Assembler	*MIRA*	*MIRA (singletons)*	*VELVET*	*META VELVET*	*SPADES*	*OMEGA*
General assembly						
Wall clock time	08:36:07	19:52:01	00:14:51	00:17:03	1d11:06:25	01:28:22
Size (megabytes)	221.1	418.5	384.7	384.7	38.2	9.8
No. of reads assembled[1] (%)	4,427,786 (43.1)	6,341,535 (61.8)	4,201,448 (40.9)	4,201,448 (40.9)	N/A	N/A
No. of contigs	512,750	1,616,750	1,002,664	1,002,675	87,340	16,269
N50	458	273	348	348	468	582
Maximum contig length (bp)	32,359	22,061	2418	2344	12,339	11,634
Consensus	202,611,745	370,456,167	350,400,905	350,424,276	34,622,064	8,196,640
Sensitivity						
No. of contigs assigned taxon id (%)	447,185 (87.2)	1,353,011 (83.7)	873,014 (87.1)	878,000 (87.6)	71,498 (81.9)	11,962 (73.5)
No. of "unknown" contigs (%)	62,664 (12.2)	257,368 (15.9)	124,513 (12.4)	119,625 (11.9)	15,018 (17.2)	4057 (24.9)
No. of viral taxa[2]	20	30	44	39	12	8
Number of unique viral taxa[3,4]	0	5	3	1	0	0
Power and accuracy						
Max bit score for a viral contig	4697	4697	4213	4213	4744	4473
Length of contig (in bp)	2885	2885	2344	2344	2765	2490
Taxon ID of highest scoring viral contig	Sewage-associated circular DNA virus 27	Sewage-associated circular DNA virus 27	Sewage-associated circular DNA virus 27	Sewage-associated circular DNA virus 27	Sewage-associated circular DNA virus 27	Sewage-associated circular DNA virus 27
Region of rowi kiwi CVLV with highest BLASTn hit[5]	1740/1740	1740/1740	1952/1952	1580/1580	1713/1713	1693/1694
Detection of rowi kiwi CVLV homologue[6] (No. of contigs assigned)	Yes (1)	Yes (1)	Yes (2)	yes (2)	No	No
Highest BLAST bit score to rowi kiwi CVLV homologue[6]	66.2 (54/66)	66.2 (54/66)	113 (149/207)	113 (149/207)	/	/

[1] Defined as the number of reads that contributed to contigs.

[2] Set at species level, with minimum number of contributing contigs before taxon assigned by MEGAN set at 2.

[3] Defined as the number of taxa not shared with another assembler.

[4] When VELVET and METAVELVET are combined, there are 19 viruses reported that are not found with the other assemblers.

[5] Identity across region expressed as a fraction of longest matching region, in base pairs.

[6] Meles meles circovirus-like virus (White et al., 2016).

CVLV, circovirus-like virus.

VELVET (2 contigs), METAVELVET (2 contigs), and MIRA (1 contig) assembled contigs that matched this virus, with the highest BLAST score (113) returned by the VELVET contigs. SPADES and OMEGA did not return any contigs with a statistical match to JQ085285.

DISCUSSION

Selecting an assembler to assemble metagenomic data remains one of the critical considerations in the metagenomic approach of viral pathogen discovery. In this study, we tested the sensitivity and accuracy of five popular metagenomic assemblers on a biomedical data set, already shown to contain a potentially pathogenic virus. The performance of assembly varied greatly. The most number of contigs was generated by MIRA (with singletons, 1,616,750) and the least by OMEGA (16,269), almost a 10-fold difference. The length of contigs also varied between assemblers with a maximum contig length of 32,359 for MIRA and as low as 2344 for METAVELVET. N50, a metric that explains the contig length at which 50% or more of contigs reached, and is less relevant for metagenomic data sets, was as high as 582 with OMEGA and as low as 273 for MIRA (with singletons). Overall, there was a spectrum of general assembly performance across algorithms, with VELVET (DBG) and MIRA (OLC with greedy) doing best. The one OLC-only assembler (OMEGA) performed the worst, although with a sample size of one for this algorithm it is hard to generalize across all OLC assemblers.

In terms of sensitivity, assembly using VELVET returned the greatest number of virus species, with a total of 44 virus species. VELVET combined with METAVELVET also returns the most number of virus species not seen with other assemblers, with a combined total of 19, where the next nearest was MIRA (singletons) with a substantially lower 5 unique species. Analysis with OMEGA, on the contrary, generated the least number of contigs that were assigned a taxon, the highest proportion of contigs that were not given an identifiable taxon (24.9%), and reported the least number of viral taxa. Taken together, VELVET appears to be the most sensitive for the detection of viral genomes; however, the apparent high virus return with this assembler may be overestimated. For example, VELVET produced a relatively large (over a million), but not very long, number of contigs. This is not surprising as VELVET was designed for single-genome assembly and, as such, it treats highly abundant, short reads as repetitive DNA and excludes such reads from assembly (Namiki et al., 2012). It is likely that many of the viruses unique to VELVET may actually result from short contigs giving quite low statistical support from homology searches and are therefore false positives.

Interestingly, all assemblers generated contigs that had very strong homology matches to Sewage-associated circular DNA virus 27. Bit scores were above 4000 for all assemblers, which show that all assemblers have generated accurate contigs for this virus. Interestingly, the relevant contig generated by VELVET is the shortest at 2344 bp, and consequently, the associated bit score is also the smallest at 4213. OMEGA, which performed relatively poorly on general assembly and sensitivity, returned a high bit score for this virus. Pipelines using each of the assemblers therefore are able to detect viruses with very strong signals.

Results from an assessment of the power of each assembler to detect virus sequence were revealing. Rowi kiwi CVLV sequence was recovered from all metagenomes when searched with the full genome, with a minimum sequence identity match with METAVELVET and maximum sequence identity match with VELVET. OMEGA was the only assembler to show a misalignment between its contigs and rowi kiwi CVLV genome, indicating an inferior assembly. On the contrary, neither SPADES nor OMEGA returned any significant matches to Meles circovirus-like virus, which was the most homologous organism to rowi kiwi CVLV archived in GenBank at the time of study, as shown in our earlier work (White et al., 2016). Taken together, it could be concluded that while each of the five assemblers tested here enables the detection of virus genomes with strong signals, sensitivity varies between assemblers, with higher sensitivity coming at a cost of higher rate of false negatives.

It is difficult to separate the best performing assemblers used here based on underlying assembly algorithms. According to the performance measurements we selected, which were chosen to assess ability to detect often rare and novel viruses in large, complex metagenomic data sets, VELVET, METAVELVET, and MIRA showed highest sensitivity and greatest power. MIRA uses quite different underlying algorithms to both VELVET and METAVELVET. A major advantage of VELVET over MIRA was the very low computing time required, which will keep time needed on computer clusters to a minimum, reducing overheads. However, the less time used to assemble reads may be somewhat mitigated by the additional time needed to screen through the large number of relatively less confident homology matches. Of interest, METAVELVET did not perform better than VELVET in this study. While the performance of these assemblers could vary with different kmer sizes, assembly optimization was not part of the scope of this study.

The importance of novel virus discovery using de novo metagenomics becomes apparent in disease-outbreak scenarios, where little information is known about the cause (Mokili et al., 2012). The development of sequencing technologies and software has facilitated a giant leap forward in capability, and will presumably continue to do so. New software is developed and often tested on data sets that suggest its superiority (Namiki et al., 2012; Haider et al., 2014; Guo et al., 2015). However, the structure of empirical data sets can differ significantly and software performance varies (Garcia-Etxebarria et al., 2014; Vazquez-Castellanos et al., 2014). We have shown here that software developed for metagenomic analyses does not perform equally well in the specific case of identifying viruses in complex metagenomic data sets, and that assemblers will be more or less desirable dependent on requirements. Overall, the VELVET software package had superior performance over most categories.

ACKNOWLEDGMENTS

The authors acknowledge those from the ecological genetics laboratory of Landcare Research, New Zealand, and the Richard Hall lab at the Institute of Environmental Science Research (ESR), New Zealand, who generated the metagenome from rowi kiwi fecal samples, and complete genome of rowi kiwi circovirus-like virus. They thank Daniel Tompkins for comments and suggestions. The authors also acknowledge the contribution of the New Zealand eScience Infrastructure (NeSI) high-performance computing facilities and the staff at the Centre for eResearch at the University of Auckland. New Zealand's national facilities are provided by the NeSI and funded jointly by NeSIs collaborator institutions and through the Ministry of Business, Innovation and Employment's Infrastructure programme (www.nesi.org.nz). This work was funded by the Ministry for Business, Innovation and Employment as part of the Crown Research Institute Capability Fund allocation to Landcare Research and ESR.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Bankevich, A., Nurk, S., Antipov, D., et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19, 455–477.

Breitbart, M., Salamon, P., Andresen, B., et al. 2002. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* 99, 14250–14255.

Camacho, C., Coulouris, G., Avagyan, V., et al. 2009. BLAST plus: Architecture and applications. *BMC Bioinformatics* 10, 421.

Chevreux, B., Wetter, T., and Suhai, S. 1999. Computer science and biology. *Proceedings of the German Conference on Bioinformatics, GCB'99.* Genome sequence assembly using trace signals and additional sequence information, 45–56.

Cox-Foster, D.L., Conlan, S., Holmes, E.C., et al. 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* 318, 283–287.

Fuhrman, J.A., and Campbell, L. 1998. Marine ecology—Microbial microdiversity. *Nature* 393, 410–411.

Garcia-Etxebarria, K., Garcia-Garcera, M., and Calafell, F. 2014. Consistency of metagenomic assignment programs in simulated and real data. *BMC Bioinformatics* 15, 90.

Gartrell, B.D., Argilla, L., Finlayson, S., et al. 2015. Ventral dermatitis in rowi (Apteryx rowi) due to cutaneous larval migrans. *Int J Parasitol Parasites Wildl* 4, 1–10.

Guo, X., Yu, N., Ding, X., et al. 2015. DIME: A novel framework for de novo metagenomic sequence assembly. *J Comput Biol* 22, 159–177.

Haider, B., Ahn, T.-H., Bushnell, B., et al. 2014. Omega: An overlap-graph de novo assembler for metagenomics. *Bioinformatics* 30, 2717–2722.

Hall, R.J., Wang, J., Peacey, M., et al. 2014. New Alphacoronavirus in Mystacina tuberculata Bats, New Zealand. *Emerging Infect Dis* 20, 697–700.

Huson, D.H., Auch, A.F., Qi, J., et al. 2007. MEGAN analysis of metagenomic data. *Genome Res* 17, 377–386.

Lander, E.S., and Waterman, M.S. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* 2, 231–239.

Laserson, J., Jojic, V., and Koller, D. 2011. Genovo: De novo assembly for metagenomes. *J Comput Biol* 18, 429–443.

Li, Z., Chen, Y., Mu, D., et al. 2012. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics* 11, 25–37.

Margulies, M., Egholm, M., Altman, W.E., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.

Mokili, J.L., Rohwer, F., and Dutilh, B.E. 2012. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* 2, 63–77.

Myers, E.W., Sutton, G.G., Delcher, A.L., et al. 2000. A whole-genome assembly of Drosophila. *Science* 287, 2196–2204.

Namiki, T., Hachiya, T., Tanaka, H., et al. 2012. MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* 40 (20):e155. DOI: 10.1093/nar/gKs678

Rohwer, F., and Edwards, R. 2002. The phage proteomic tree: A genome-based taxonomy for phage. *J Bacteriol* 184, 4529–4535.

Roux, S., Tournayre, J., Mahul, A., et al. 2014. Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* 15, 76.

Schmieder, R., and Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864.

Smits, S.L., Bodewes R., Ruiz-Gonzalez, A., et al. 2014. Assembly of viral genomes from metagenomes. *Front Microbiol* 5, 714.

Smits, S.L., Zijlstra, E.E., van Hellemond, J.J., et al. 2013. Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010–2011. *Emerging Infect Dis* 19, 1511–1513.

Treangen, T.J., Sommer, D.D., Angly, F.E., et al. 2011. Next generation sequence assembly with AMOS. *Curr Protoc Bioinformatics* 33, 11.8:11.8.1–11.8.18.

Vazquez-Castellanos, J.F., Garcia-Lopez, R., Perez-Brocal, V., et al. 2014. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* 15, 37.

White, D.J., Hall, R.J., Jakob-Hoff, R., et al. 2015. Exudative cloacitis in the kakapo (*Strigops habroptilus*) potentially linked to *Escherichia coli* infection. *N Z Vet J* 63, 167–170.

White, D.J., Hall, R.J., Wang, J., et al. 2016. Discovery and complete genome sequence of a novel circovirus-like virus in the endangered rowi kiwi, Apteryx rowi. *Virus Genes* 52, 727–731.

Zerbino, D.R., and Birney, E., 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* 18, 821–829.

Zhao, Y., Tang, H., and Ye, Y., 2012. RAPSearch2: A fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28, 125–126.

Address correspondence to:
Dr. Daniel J. White
Landcare Research
Private Bag 92170
Auckland Mail Centre
Auckland 1142
New Zealand

E-mail: whited@landcareresearch.co.nz