GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Optimization of dengue virus genome assembling using GSFLX 454 pyrosequencing data: evaluation of assembling strategies

S.M.M. Casseb[1], J.F. Cardoso[2], R. Ramos[3], A. Carneiro[3], M. Nunes[2], P.F.C. Vasconcelos[1] and A. Silva[3]

[1]Departamento de Arbovirologia e Febres Hemorrágicas,
Instituto Evandro Chagas, Ananindeua, PA, Brasil
[2]Centro de Inovação Tecnológica, Instituto Evandro Chagas,
Ananindeua, PA, Brasil
[3]Laboratório de Polimorfismo do DNA, Instituto de Ciências Biológicas,
Universidade Federal do Pará, Rede Paraense de Genômica e Proteômica,
Belém, PA, Brasil

Corresponding author: S.M.M. Casseb
E-mail: samir.casseb@gmail.com

Genet. Mol. Res. 11 (4): 3688-3695 (2012)
Received October 21, 2011
Accepted March 13, 2012
Published August 17, 2012
DOI http://dx.doi.org/10.4238/2012.August.17.6

ABSTRACT. Currently assembling genomes without reference is one of the most important challenges for bioinformaticists all over the world in an attempt to characterize new organisms. The current study has used two dengue virus type 4 (DENV-4) strains recently isolated in Brazil, which have its genomes sequenced using the GSFLX 454 sequencer (Roche, Life Science) by the pyrosequencing method. The GSFLX 454 data were used for testing different genome assembling strategies. We described a pipeline that was able to recover more than 96% of the sequenced genome in a single run and could be helpful for further assembly attempts of other DENV genomes, as well as other RNA virus-like genomes.

Key words: Dengue; Pyrosequencing; Virus; Assembly; Pipeline; Genome

INTRODUCTION

Dengue virus (DENV) is one of the most important arthropod-borne viruses that affect humans worldwide (Carvalho et al., 2010). According to the World Health Organization, approximately 1 million cases of dengue are annually reported over 100 countries worldwide, with DENV being responsible for more than 250,000 cases of hemorrhagic dengue fever and 25,000 deaths (Carvalho et al., 2010).

Due to its important health impact, several studies have been conducted to assess the traits of DENV genotypes through the sequencing of its entire genome. In the past few decades, several strains of the 4 DENV serotypes (DENV-1 to 4) have been partially or completely sequenced using the Sanger strategy (Sanger et al., 1977).

The development of new technologies in molecular biology has drastically improved microbial research in various aspects, one of which is descriptive genomics. The rapid advances in new methods of DNA sequencing (next-generation sequencing, NGS), which are characterized by high-throughput data, enables the generation of hundreds of thousands of entire genomes in a short period of time (Ronaghi, 2001; Schuster, 2008).

The most widespread NGS platforms are Genome Sequencer FLX 454 (GSFLX 454; Roche Life Science, Germany), Solexa (Illumina, USA), and SOLID (Applied Biosystem, USA). The NGS platforms provide high genome coverage; on the other hand, the reads generated by NGS platforms are comparatively much smaller than those obtained by Sanger-based sequencers (Sanger et al., 1977). Despite the high-throughput capacity of next-generation sequencers, there is a need for developing new strategies for genome assembly (Schuster, 2008). These new platforms impose challenges for the genome assembly process, which is based on the analysis of similarity between 2 reads that generate a given scaffold sequence.

The development of new strategies to obtain complete genomes using NGS platforms as complement approaches requires the use of distinct assembling algorithms, as well as the use of data generated by different technologies (e.g., data generated by Sanger, GSFLX 454, SOLID, Illumina, etc.) (Birney, 2011). These approaches use a large number of contigs, and some platforms usually require a large amount of computational effort and human resources for processing and data mining (Hernandez et al., 2008; Nijkamp et al., 2010).

The GSFLX 454 (Roche) pyrosequencer can sequence approximately 100 million bases in a single time frame of 7.5 h and generate reads ranging from 250 to 400 bases (Hernandez et al., 2008).

NGS assembly approaches are basically characterized by the use of a reference genome for mapping the generated reads (reference assembly), and the *ab initio* approaches overlap-layout-consensus, Bruijn Graph, and Greedy Graph (Miller et al., 2010) perform the alignment on the basis of similarity between their own reads, with the size of overlapping regions (defined in k-mer) allowing for sequence extension and production of a continuous sequence (Miller et al., 2010).

An outbreak of DENV-4 occurred in Roraima State, Brazil, in July 2010 (Temporão et al., 2011); during this outbreak, full-length genomes were obtained for the assessment of a pipeline suitable for genome assembly by using mapping and/or *ab initio* methods. The use of either or both methodologies may be helpful in developing an efficient pipeline for assembling genomes of DENVs and other DENV-like viruses, such as ssRNA viruses.

MATERIAL AND METHODS

Sequences

Two DENV-4 sequences (GenBank accession Nos. JN559740 and JN559741) were used in the study and were kindly provided by the DENV sequencing group established at the Department of Arbovirology and Hemorrhagic Fevers of Instituto Evandro Chagas. DENV-4 sequences were obtained by the pyrosequencing method for *de novo* assembly using the GSFLX 454 (Roche, Life Science). Briefly, the extracted RNA was fragmented using $ZnCl_2$. Fragments in the range of 400 bp were selected by using the Bionanalyzer (Agilent, USA) and then ligated to specific adaptors (A and B) for sequencing in the GSFLX 454.

Analysis using pre-processing (Mothur)

The generated reads were first evaluated for the presence of possible homopolymers, as well as for preprocess sequences, to remove primers, generate group files, and screen for quality (Phred value) by using the Mothur v. 1.20 software (Schloss et al., 2009). Phred values were tested in a range of 10 to 30.

Assembly

Sequences input were assembled using 2 different approaches, namely, the *de novo* and mapping reference methods, by using 3 distinct computer programs: Newbler v. 2.5.3 (Data Processing Software Manual 454 Life Science, http://www.454.com/), Mira v. 3.2.1.15 (Chevreux et al., 2004), and the Geneious pro^{TM} 5.4 free-trial software (Biomatter's Geneious Software, http://www.geneious.com/).

For the *de novo* assembly method, when Mira or Newbler (gsAssembler) softwares were used, 3 different parameters for read size were set (50, 200, and 300 bp) in order to assemble the contigs. Parameters used for Newbler were as follows: input, 20 bp; isotig threshold, 100; isogroup threshold, 250; minimum overlap length, 40; minimum overlap identity, 70%; k-mer, 12 (seed step), and k-mer, 14 (seed length). The same parameters were set for the Mira software, and the Mauve software (Darling et al., 2010) was used to rearrange the contigs generated by *de novo* assembly against a DENV-4 reference genome available at the local GenBank database.

For the mapping reference assembly, gsMapper (Newbler Software) was used to rearrange the reads against a given reference sequence using the following parameters: input, 20 bp; all contig threshold, 100; large contig threshold, 200; minimum overlap length, 40; minimum overlap identity, 70%; k-mer, 12 (seed step), and k-mer, 16 (seed length). The same parameters were set for the Mira software.

Analysis of contaminant sequences

For removing possible contaminants from the targeted reads, the program Blast-N and Blast-X (BLAST database and the stand-alone sequence comparison software v. 2.2.25) (Altschul et al., 1990) were used. The results generated for each read were evaluated in terms of queries without the DENV genome on the basis of the parameters of coverage, E value, and max-identity score for other organisms. Alternatively, the Mauve software (Darling et al., 2010) was used to graphically identify the reads related to the reference genome.

RESULTS

Data quality

For assembling the sequences generated by the GSFLX 454, the reads were first inspected for quality, and no significant difference was observed in the final genome for Phred values higher than 20.

Contigs generation, removing contaminant sequences, and scaffold

The GSFLX 454 reads were assembled by using 2 distinct methods: *de novo* and mapping reference, as well as different softwares (Mira, Newbler, and Geneious Pro). Contigs were generated according to the method and software used for assembling the reads in large sequences. Irrespective of the method used, contaminant sequences were found in the entire data by using the BlastX/BlastN algorithm (http://www.ncbi.nlm.nih.gov) and removed before assembling the contigs. Most of the contaminant contigs corresponded to partial genomes of *Aedes albopictus* or *Aedes*-related genomes with high E values (1.0) and p-scores (100%) (Figure 1).

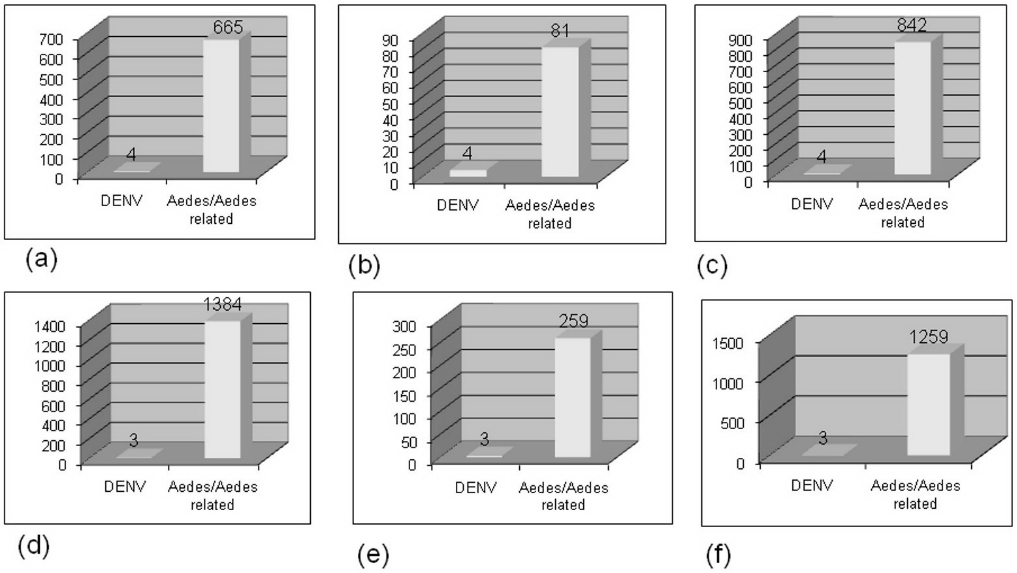

Figure 1. Analysis of contaminant (*Aedes albopictus* and *Aedes*-related sequences) and dengue virus (DENV) contigs after the GSFLX 454 sequencing data assembly. Strain ROR 1982 assembled by Newbler (**a**), Mira (**b**) and Geneious (**c**). Strain ROR 2010 assembled by Newbler (**d**), Mira (**e**), and Geneious (**f**).

The genomes generated using the different strategies and computer softwares showed different sizes, ranging from 9.477 to 10.268. In comparison to reference sequences available at the GenBank database (GU2899913 and EU854299), the used methods could recover 89-96% of the entire DENV-4 genomes, with 69 internal gaps in the open reading frame (ORF; 53 nucleotides at genome position 8.870-8.920 and 16 nucleotides, at genome position 9.231-9.246) in the strain ROR 1982 and an additional 12 nucleotides of the ROR 2010 ORF. Lack of short-terminal 5'- and 3'-ends were noted in both strains. Internal gaps in the ORF were eliminated when a consensus sequence was generated using data obtained from Newbler, Mira, and Geneious pro. Terminal regions were not recovered, and genome coverage also varied with the method and software used. The results obtained from the methods used are summarized in Table 1.

Table 1. Results for *de novo* and mapping reference strategies for DENV-4 strains according to N50, maximum (Max.) and minimum (Min.) number of reads, number of contigs, number of bases, genome coverage, recovered genome (size).

Method	Software	Strain	N50	Max.	Min.	No. of contigs	No. of bases	Genome coverage (x)	Genome size (%)
De novo	Newbler	ROR 1982	315	6.072	40	669	192.562	287	10.023 nt (94)
		ROR 2010/2854	325	5.261	38	1.387	389.316	280	10.212 nt (96)
	Mira	ROR 1982	987	6.072	31	85	37.678	145	9.477 nt (89)
		ROR 2010/2854	590	5.252	31	262	115.869	291	10.185 nt (96)
	Geneious	ROR 1982	404	6.224	36	846	310.708	367	10.206 nt (96)
		ROR 2010/2854	305	5.352	32	1.262	282.761	224	10.189 nt (96)
	Newbler/Mira/ Geneious	ROR 1982	NA	NA	NA	NA	NA	NA	10.268 nt (96)
		ROR 2010/2854	NA	NA	NA	NA	NA	NA	10.222 nt (96)
Mapping reference	Newbler	ROR 1982	8.825	8.825	331	3	10.194	3.398	10.194 nt (96)
		ROR 2010/2854	10.218	10.218	10.218	1	10.218	10.218	10.225 nt (96)
	Mira	ROR 1982	10.150	10.150	10.150	5	10.150	2.030	10.150 nt (95)
		ROR 2010/2854	10.227	10.277	10.227	1	10.227	10.227	10.227 nt (96)
	Geneious	ROR 1982	10.225	10.225	10.225	2	10.225	10.225	10.207 nt (96)
		ROR 2010/2854	10.237	10.237	10.237	1	10.237	10.237	10.237 nt (96)
	Newbler/Mira/ Geneious	ROR 1982	NA	NA	NA	NA	NA	NA	10.288 nt (96.6)
		ROR 2010/2854	NA	NA	NA	NA	NA	NA	10.239 nt (96.5)
De novo/ mapping reference	All	ROR 1982	NA	NA	NA	NA	NA	NA	10.288 nt (96.6)
		ROR 2010/2854	NA	NA	NA	NA	NA	NA	10.239 nt (96.5)

NA = not available.

Pipeline

A pipeline was developed for assembling the DENV-4 genomes by using the *de novo* and/or mapping reference strategies. Briefly, the reads were trimmed for adaptors using the Mothur software and used for contig generation; contigs were then generated by the *de novo* strategy using 3 distinct softwares (Newbler, Mira and Geneious). Subsequently, possible contaminant sequences were removed from the whole data by using a combination of BlastN and BlastX approaches, followed by assembling contigs related only to virus genomes. The assembled contigs were rearranged and validated against a reference sequence (chose by BlastN/BlastX analysis; high E value and p-score) by using the Mauve software algorithm. Figure 2 summarizes the pipeline strategy used for the assembly and developing the full-length genomes of the 2 DENV-4 strains, and Figure 3 demonstrates the efficiency of the 3 algorithms for assembling the DENV genomes.

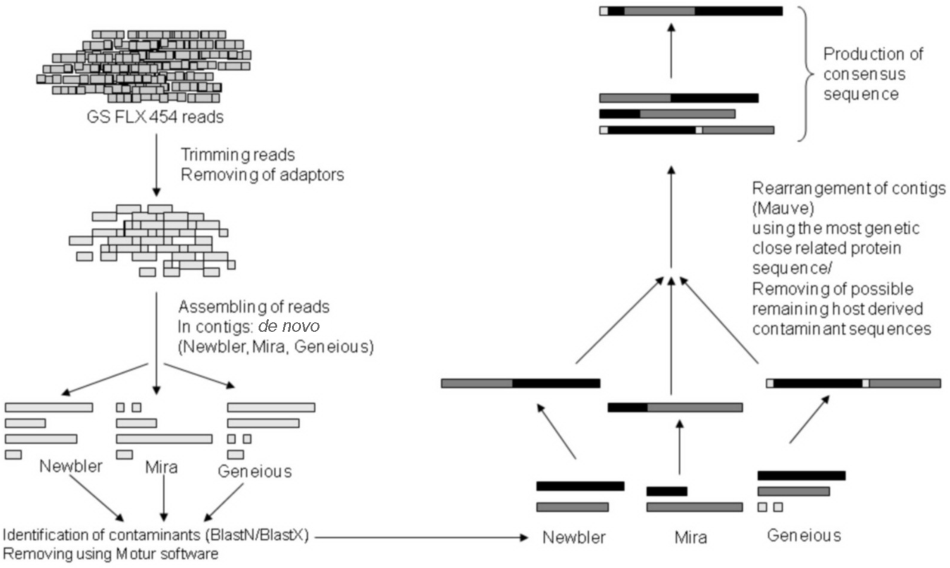

Figure 2. Proposed pipeline for assembly dengue virus type 4 genomes and related viruses using *de novo* strategy.

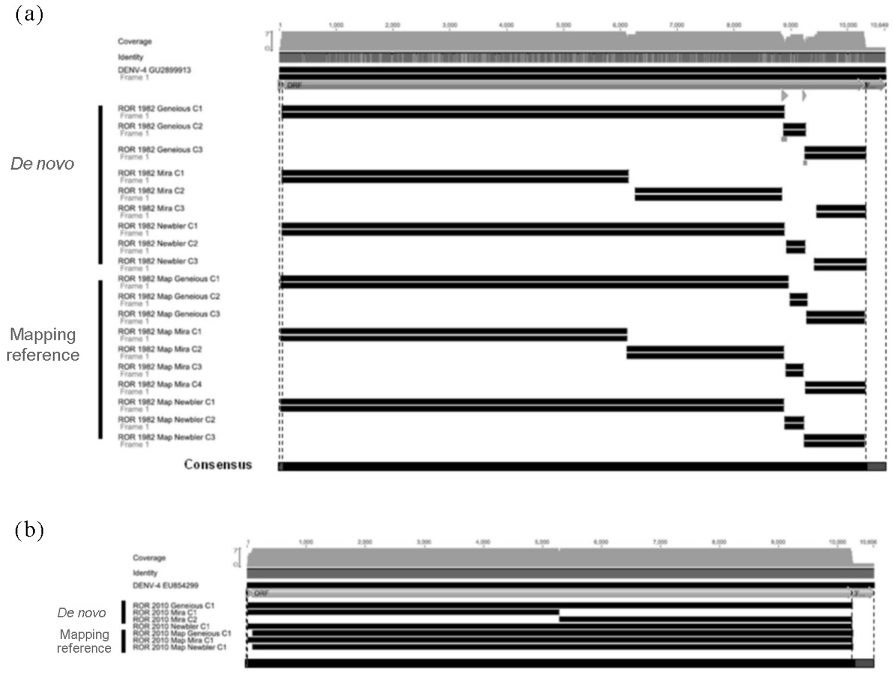

Figure 3. Graphic representation showing the association of three distinct software algorithms (Geneious, Mira and Newbler) and methods (*de novo* and mapping reference) for obtaining a consensus sequencing with high genome recovering percentage. **a.** Assembly for the strain ROR 1982. **b.** Assembly for the strain ROR 2010. Genome sizes are indicated over each reference genome.

DISCUSSION

Genome assembly by NGSs without the use of reference sequences is a challenging prospect in bioinformatics worldwide (de Magalhaes et al., 2010). In the past, efforts have been made to use NGS-based approaches, in particular pyrosequencing, for obtaining genetic data from different viral pathogens, such as HIV; avian leukosis virus; lyssaviruses; enteric viruses, and few arboviruses, including DENV-1 and phleboviruses (Victoria et al., 2009; Bimber et al., 2010; Bishop-Lilly et al., 2010; Day et al., 2010; Hedskog et al., 2010; De Wolf et al., 2011; Schinazi et al., 2011; Abbate et al., 2011; De Benedictis et al., 2011; Palacios et al., 2011).

The 2 DENV-4 strains used in the current study were chosen for testing a pipeline for assembling genomes generated by GS-FLX 454. In this pipeline for analyzing the DENV-4 genome, we employed a combined approach of using different softwares and strategies to evaluate the best parameters and process for rescuing the maximum length of the targeted genome in a single sequencing run.

A comparison between strategies (*de novo* and mapping) as well as softwares revealed that the proposed pipeline for the *de novo* method (see Figure 2) was accurate and highly efficient in comparison to the mapping reference strategy, recovering more than 90% of the proposed genome and with no errors. Few gaps (N = 69) were found within the ORF, which were easily closed when mapping reference and *de novo* sequencing were applied together. Notably, the contigs generated by one method were supplementary to those generated by the other, correcting and/or confirming the previously annotated position of a given nucleotide and validating the sequences available at the GenBank database (GU289913 and EU854299).

In most cases, the absence of a reference sequence to guide the reads for assembling the genomes is a major concern among bioinformaticists. The setting of the appropriate output parameter depending on the type of sequencer and software used for assembling the genomes is crucial for correctly assembling a sequence (Harismendy et al., 2009). In the case of GSFLX 454 (Roche) and Solid (Applied Biosystems), which are examples of sequencers for long reads and short reads with output parameter defaults of 200 and 50-75 bp, respectively (Harismendy et al., 2009). Reads with sizes greater or lesser than the determined threshold value are refused or ignored during the assembly process. In case of the DENV-4 genome, the parameters were increased from 50 to 350 bp, and in consequence, many reads with a mean size of 75 and 300 bp, which were previously ignored, were incorporated in the contigs and assembled as DENV-4 sequences. This suggests that the read size must be evaluated, and changes in default parameters may be necessary for different organisms.

In the current analysis, the prior removal of contaminant sequences and the identification of target-related contigs (in this case, virus-related contigs) by using a combination of BlastN, BlastX, and Mauve algorithms were extremely useful for sequence cleanup and to restrict the contigs to only viral genome sequences. Restriction of contigs to only viral-related genomes eliminated the computational effort and possible errors during the assembly processes. A combination of the algorithms of different softwares (Newbler, Mira, and Geneious) enabled better genome recovery in comparison to the reference genomes, GU2899913 and EU854299.

In conclusion, the proposed pipeline can be useful for the assembly of other DENV genomes of different viral serotypes (DENV-1, DENV-2, DENV-3, and DENV-4), as well as for other single-stranded, positive-sense RNA virus genomes.

ACKNOWLEDGMENTS

Research supported by CNPq/CAPES/FAPESPA (#573739/2008-0), University of Pará State and IEC/SVS from the Brazilian Government.

REFERENCES

Abbate I, Vlassi C, Rozera G, Bruselles A, et al. (2011). Detection of quasispecies variants predicted to use CXCR4 by ultra-deep pyrosequencing during early HIV infection. *AIDS* 25: 611-617.

Altschul SF, Gish W, Miller W, Myers EW, et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

Bimber BN, Dudley DM, Lauck M, Becker EA, et al. (2010). Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. *J. Virol.* 84: 12087-12092.

Birney E (2011). Assemblies: the good, the bad, the ugly. *Nat. Methods* 8: 59-60.

Bishop-Lilly KA, Turell MJ, Willner KM, Butani A, et al. (2010). Arbovirus detection in insect vectors by rapid, high-throughput pyrosequencing. *PLoS Negl. Trop. Dis.* 4: e878.

Carvalho SE, Martin DP, Oliveira LM, Ribeiro BM, et al. (2010). Comparative analysis of American Dengue virus type 1 full-genome sequences. *Virus Genes* 40: 60-66.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, et al. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14: 1147-1159.

Darling AE, Mau B and Perna NT (2010). Progressive mauve: multiple genome alignment with gene gain, loss, and rearrangement. *PLoS One* 5: e11147.

Day JM, Ballard LL, Duke MV, Scheffler BE, et al. (2010). Metagenomic analysis of the turkey gut RNA virus community. *Virol. J.* 7: 313.

de Magalhaes JP, Finch CE and Janssens G (2010). Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing Res. Rev.* 9: 315-323.

De Benedictis P, De Battisti C, Dacheux L, Marciano S, et al. (2011). Lyssavirus detection and typing using pyrosequencing. *J. Clin. Microbiol.* 49: 1932-1938.

De Wolf H, Van Marck H, Mostmans W, Thys K, et al. (2011). HIV-1 nucleotide mixture detection in the Virco®TYPE HIV-1 genotyping assay: a comparison between Sanger sequencing and 454 pyrosequencing. *J. Virol. Methods* 175: 129-132.

Harismendy O, Ng PC, Strausberg RL, Wang X, et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 10: R32.

Hedskog C, Mild M, Jernberg J, Sherwood E, et al. (2010). Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PLoS One* 5: e11345.

Hernandez D, Francois P, Farinelli L, Osteras M, et al. (2008). *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.* 18: 802-809.

Miller JR, Koren S and Sutton G (2010). Assembly algorithms for next-generation sequencing data. *Genomics* 95: 315-327.

Nijkamp J, Winterbach W, van den Broek M, Daran J-M, et al. (2010). Integrating genome assemblies with MAIA. *Bioinformatics* 26: 433-439.

Palacios G, Tesh R, Travassos da RA, Savji N, et al. (2011). Characterization of the Candiru antigenic complex (Bunyaviridae: Phlebovirus), a highly diverse and reassorting group of viruses affecting humans in tropical America. *J. Virol.* 85: 3811-3820.

Ronaghi M (2001). Pyrosequencing sheds light on DNA sequencing. *Genomes Res.* 11: 3-11.

Sanger F, Nicklen S and Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74: 5463-5467.

Schinazi RF, Massud I, Rapp KL, Cristiano M, et al. (2011). Selection and characterization of HIV-1 with a novel S68 deletion in reverse transcriptase. *Antimicrob. Agents Chemother.* 55: 2054-2060.

Schloss PD, Westcott SL, Ryabin T, Hall JR, et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol*. 75: 7537-7541.

Schuster SC (2008). Next-generation sequencing transforms today's biology. *Nat. Methods* 5: 16-18.

Temporão JG, Penna GO, Carmo EH, Coelho GE, et al. (2011). Dengue virus serotype 4, Roraima State, Brazil. *Emerg. Infect. Dis.* 17: 938-940.

Victoria JG, Kapoor A, Li L, Blinkova O, et al. (2009). Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J. Virol.* 83: 4642-4651.