American Journal of Animal and Veterinary Sciences

Research Articles

Closing the Gaps in Rat Cytomegalovirus ALL-03 (Malaysian Strain) Genomic Scaffold

[1]Krishnan Nair Balakrishnan, [2]Ashwaq Ahmed Abdullah, [1]Yusuf Abba, [1]Jamilu Abubakar Bala, [3]Faez Firdaus Jesse Abdullah, [1]Farina Mustaffa Kamal, [1,2]Zeenathul Allaudin Nazariah, [3]Ideris Aini, [1,2]Noordin Mohamed Mustapha and [1,2]Mohd Azmi Mohd Lila

[1]*Department of Pathology and Microbiology,*
Faculty of Veterinary Medicine, Universiti Putra Malaysia, 43400, Serdang, Selangor D.E., Malaysia
[2]Institute of Bioscience, *University Putra Malaysia, 43400, Serdang, Selangor D.E., Malaysia*
[3]*Department of Veterinary Clinical Studies,*
Faculty of Veterinary Medicine, Universiti Putra Malaysia, 43400, Serdang, Selangor D.E., Malaysia

Article history
Received: 25-3-2015
Revised: 26-3-2015
Accepted: 28-5-2015

Corresponding Author:
Mr. Krishnan Nair Balakrishnan
Department of Pathology and Microbiology, Faculty of Veterinary Medicine,
Universiti Putra Malaysia,
43400, Serdang, Selangor D.E.,
Malaysia
Email: krishnan_ukm@yahoo.com

Abstract: Next generation sequencing technologies has revolutionized genomic research by producing a large volume of sequence data and lowest per base cost compared to the traditional sanger method. Although this technology offers many advantages, gap occurrences are commonly found in draft assemblies. The same problem was observed with Rat Cytomegalovirus (RCMV) ALL-03 (Malaysia strain), where a complete genome sequence could not produce the complete genome due to the presence of gaps in the draft genome. This restrains our ability to take full advantage of genome data. This study aimed to identify the sequence data present in the gap regions and close these gaps in order to produce a complete genome sequence for RCMV ALL-03. Twenty sets of specific primers were designed between two adjacent contigs and PCR was carried out to obtain the appropriate sequences in respective gap regions. Sanger sequencing was employed in the PCR product to get the gap sequences. Out of the five identified gaps in the RCMV ALL-03 genome sequence, only three were confirmed to be true gaps, while the other two were due to sequence repeats. In conclusion, all the gaps were closed successfully and complete genome sequence of RCMV ALL-03 can now be explored in further studies.

Keywords: Next Generation Sequencing, Cytomegalovirus, Sanger Sequencing, Genome, PCR

Introduction

Cytomegalovirus (CMV), an important pathogen belongs to the betaherpesviruses subfamily of herpesviruses, which infects many living organisms including humans (Weller, 1971). CMV causes acute, persistent and latent infections in human and animal population but remain asymptomatic in healthy individuals. In contrast, this virus can cause significant morbidity and mortality in immunocompromised and immunosuppresed patients, such as AIDS patients and organ transplant recipients, respectively (Livingston *et al*., 2001; Scalzo *et al*., 2009). Although primate CMV is closely related to Human CMV (HCMV), but these strains are not frequently used as a model for HCMV infection due to impracticalities and high expenses. These are the reasons why Murine CMV (MCMV) and Rat CMV (RCMV) became well known models for HCMV because of low cost, high reproductive rates and simplicity of handling (Mocarski *et al*., 2007). The major drawback of these strains is they do not cross the placental barrier and cause *in utero,* hence it is hard and complicated to use these models for congenital infections. To counter this, a new strain of RCMV strain ALL-03, acquired from the uterus and placenta of the *Rattus rattus diaardi* (house rat) (Loh *et al*., 2003), confirms the ability of this strains' vertical transmission in pregnant rats. This ability of the virus makes it an appropriate model of choice to study the congenital infection of CMV in humans. Hence, RCMV ALL-03 is a good model to study HCMV as it has the same pathogenicity and able to cross the placenta (Loh *et al*.,

2006). To further elucidate the pathogenesis of RCMV ALL-03, genome sequencing of this virus is much crucial.

CMV have the largest genome size of approximately 230-240 kbp of double stranded DNA with high Guanisine and Cytosine (G+C) content when compared to other herpesviruses (Mocarski *et al*., 2007). The human CMVs genome contains Unique Long (UL), Unique Short (US) and internal as well as terminal repeat regions. In contrast, the genome arrangements in animal CMV are linear without internal repeat regions but contain repeated sequences at genome termini (Christine Meyer, 2010). The core genes which are common to all herpesviruses are located at UL domain, while specific genes are located at US domains (Yu *et al*., 2003).

Like other herpesviruses, CMV has a large genomic size ranging from 195-240 kbp and it is also known as the largest among other herpesviruses. Currently, only two strains of RCMV have deposited the full genome sequence which are Maastricht strain (Vink *et al*., 2000) and English strain (Ettinger *et al*., 2012), which are 229,896 bp and 202,946 bp in length respectively. Recently, RCMV ALL-03 strain was sequenced using Next Generation Sequencing Illumina platform producing 198,895 bp arranged as single unique sequence flanked by 504 bp terminal direct repeats (unpublished data) (Yi, 2013). Unfortunately, the draft genome of RCMV ALL-03 is not complete because of the presence of gaps between different contigs. Eventhough sequencing technologies g improve and advance day by day, no sequencer produces adequate data to assemble a complete genome in a single experiment (Xing *et al*., 2011).

Sequencing reads will be assembled into a set of contiguous fragments known as contigs and arranged together to form a longer scaffolds. Hence, the draft assemblies have gaps and become incomplete assemblies (Xing *et al*., 2011).

In the public databases, more than a third of the genome sequences are in draft form and incomplete (Piro *et al*., 2014). Many of the draft assemblies in NCBI have gaps of different length and numbers depending on the size of the genome (Xing *et al*., 2011). These gaps make it difficult to study genetic variations, expression of RNA, chromosome conformation and interactions of protein-DNA from the incomplete draft assemblies (Shendure and Hanlee, 2008), thereby limiting comprehensive study of o f the genome e (Xing *et al*., 2011). The genomic data obtained from the RCMV ALL-03 sequence showed incomplete sequence due to gaps between contigs. This study was thus undertaken to evaluate the nature of the gaps and close them using specific designed primers from the partial genome data.

Materials and Methods

Virus Culture and Propagation

Monolayer cultures of Rat Embryonic Fibroblast (REF) cells from ATCC were cultured in 25cm^3 flask in an incubator at 37°C and 5% CO_2. Confluent cells were infected with RCMV ALL-03 and frequently observed for morphological changes known as Cytopathic Effect (CPE). When the CPE exhibits 90%, the virus harvested.

Virus Concentration and Purification

Viral supernatant concentrated using Polyethylene glycol 6000 (PEG 6000; Calbiochem, Darmstadt, Germany) followed by virus purification using 5 different concentrations of sucrose at 60%, 50%, 40%, 30% and 20% (w/v). Virus band as white opalescent detected and pulled out from gradient tubes. Purified virus, suspended in 1ml of PBS to further use.

DNA Extraction

The genomic viral DNA of RCMV ALL-03 was extracted as described by Lai *et al*. (1999).

Determination of DNA Concentration and Purity

Extracted DNA was subjected for concentration and purification using spectrophotometric by using a BioPhotometer™ plus (Eppendorf, Hamburg Germany) in accordance with the protocol given by manufacturer. The reading of the ratio 260 nm to 280 nm estimates the purity of DNA and pure DNA preparation has a ratio range of 1.8 to 2.0.

Primer Design

Primers used for PCR are listed in Table 1. All the primers were designed using a special software; CLC Genomic Workbench 4.7.2 at Institut Bio Sains, UPM and synthesized by the 1st Base Technologies company. Overall, 4 sets of primers were designed for each gap and designated as first, second and third trial. The primer positions are illustrated in Fig. 1. For the first trial, primers were carefully designed according to the available sequences at the end of different contigs flanking the gap region. For the second trial, result from the first trial was used as a template to design another 2 sets of primers. For the third trial, primers were designed further away from the gaps as illustrated below.

PCR &Gel Electrophoresis

PCR was carried out in a 25 µL reaction volume using a thermal cycler (Eppendorf, Hamburg, Germany) with the following cycling protocol; polymerase activation for 2 min at 95°C, denaturation for 20 sec at 95°C, annealing for 10 sec (depending on primer Tm) and extension for 10 sec at 70°C for 35 cycles. The final PCR product was examined against a 100bp DNA ladder (Vivantis, Lithuania) using a gel documentation system.

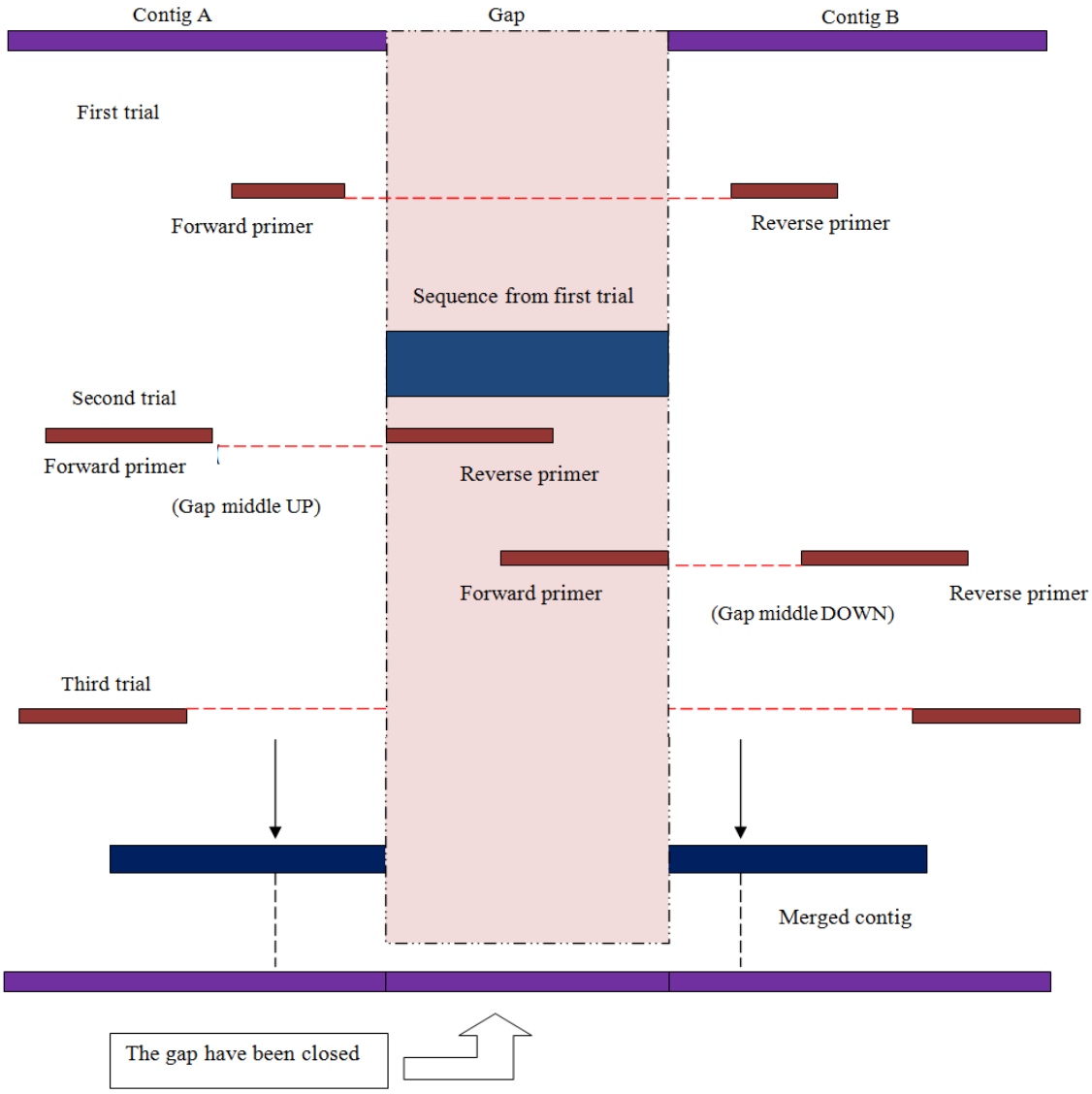

Fig. 1. Shows the position of 4 sets of primers which have been designed to close the 5 gaps in RCMV ALL-03 strain

Table 1. List of primers used for closing the gaps

			Second trial					
	First trial	Ta	Middle up	Ta	Middle down	Ta	Third trial	Ta
Gap 1	Forw 5' gacagaactaaccaaccc 3'	53°C	Forw 5'gacacacaactctaaaagca3'	42°C	Forw 5 cttctcgttctcgttctt 3'	51°C	forw 5' acggttttgctcttctgg 3'	48°C
	Rev 5' gaacgagaacgagaacga 3		Rev 5' cgagagcgagaacgagaa 3'		Rev 5' tcgcatatgtctctcgtc 3'		rev 5' ggttttttcacggtatctgg 3'	
Gap 2	Forw 5' cgcatacaaaaccaacct 3'	55°C	Forw 5'gatttacgtttgccggtg3'	47°C	Forw 5'ttcaattttcccgcacag3'	39°C	forw 5'cccacatttcaccgctatc 3'	54°C
	Rev 5' gacccccatctagatacaa 3'		Rev 5' gtcagttactttgaggatgg 3'		Rev 5' caatcaacacacccacac 3'		rev 5' cgcatacaaaaccaacct 3'	
Gap 3	Forw 5' agggctattgtcgaaaag 3'	48°C	Forw 5' cgggtgcattgtgtgata 3'	51°C	Forw 5' gtcttttcccgctttgtt 3'	50°C	forw 5' ggtaggtagtgagcgaaa3'	50°C
	Rev 5' gtcttttgtccttgagtg 3'		Rev 5' ggcaacaaagcggaaaaga 3'		Rev 5' cccccgtcatctttttct 3'		rev 5' caggttgatgtaggaagaa 3'	
Gap 4	Forw 5'ctctcgtagactgatttacct3'	49°C	Forw 5' agggggagggaaattttt 3'	53°C	Forw 5' cggggtctaaatatagttg3'	49°C	forw 5' gggtaaaggcaaagcatga3'	50°C
	Rev 5'ctccgctatttatgatctacc3'		Rev 5' ctgtctacatataacgcgat 3'		Rev 5' gaaccgaaaacaccagaa3'		rev 5' gtaaacgtgagaacatgga 3'	
Gap 5	Forw 5' aactaaccaacccccaac 3'	51°C	Forw 5' taatcaccctacctgaa 3'	47°C	Forw 5' gggaatcaccatcatagg3'	49°C	forw 5' cggggcgatcttcatatt 3'	48°C
	Rev 5' ggatgaagataggatggg 3'		Rev 5' ggaattgcacctatgatg 3'		Rev 5' agtcccggtttaataagc3'		rev 5' cctgctcttgtacggaaaa 3'	

Sequencing and Final Alignment

All the samples were sent for sanger sequencing using Applied Biosystems 3730XL Genetic Analyzer and results analyzed using Applied Biosystems DNA Sequencing Analysis Software v5.2 with KB Basecaller. Final alignment of the gaps with previous draft genome (unpublished data) (Yi, 2013) was carried out at Institute of Biosains UPM using software CLC Genomic Workbench.

Results

About 5 gaps have been identified and each gap is located at different position with different sizes in draft genome. Out of 5 gaps, only 3 are true gaps and remaining 2 were false gaps. The details of each gap were demonstrated with original draft and complete sequence result as below.

Gap 1 draft:

aaaaaacgagcgtatatatagacaaccgaaaacgttgttagatattattctgcttc
tcagtcaactcagacaccgctttatatagaacttgcatgtacccggaaaccttttct
cgatcaacgacagaactaaccaacccccaacacattaaccgcgtatacgcgtt
aatatggaaacgaatcgacccgagcatacatcgtcaccgatgaggcgcagtg
catcgatggatatacaacgatgtcgttctcnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnncc
gagcatacatcgtcaccgatgaggcgcagtgcatcgatggatatacaacgatg
tcgttctcgttctcgttcaccatcgcgatccagataccgtgaaaaaaccaggtac
actccgacaccaaggacgatggatcgggatagacactcacaca

Gap1 sequence result:

gacagaactaaccaacccccaacacattaaccgcgtatacgcgttaatatggaaa
cgaatcgacccgagcatacatcgtcaccgatgaggcgcagtgcatcgatggata
tacaacgatgtcgttctcgttctcgttcaccatcgcgatccagataccgtgaaaaaa
ccaggtacactccgacaccaaggacgatggatcgggatagacactcacaca

The first gap for this draft genome lies between 25748 and 25948. Initially the gap was 200 bp as indicated by small letter ‘n’. After closing the gap we get to know that only 60 bp are real missing sequences. The remaining gap was repetitive sequences.

Gaps derived from NGS can be formed due to occurrence of repeat sequences that are considerably longer than individual reads. In more evolved organisms, repeat regions tend to grow in number, size and complexity (Xing *et al*., 2011).

After fill up with nucleotides, the new sequences were blasted and compared with reference strains Maastricst and English. This gap region fall in gene coding region R32 and E32 in maastrischt strain and English strain respectively. This gene is responsible for encoding virion associated protein that elicits a strong humoral immune response (Vink *et al*., 2000).

Gap 2 draft:

ctaatgaataaattaaacctaatggaataaaccctagtaatataatatctagtaatt
ccagcaagtgtatatcccttcctagtgtaccctccctaaccccctccttccc
ccccaacattcccctacactgggggtccagtcttggnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnttcccggtccctttcccggtcccttccggtcccttcccggtcccttt
cccggtcccttccggtcccttcccggtccctcggctcctcggctcctccggtt
ccccggtcccgttccccggtcccgttccccggtcccttcccggccacccaattt
tggattttatttaggttggttttgtatg

Gap 2 sequence result:

Ccccccaacattcccctacactgggggtccagtcttggtcccctcggattct
ccggtccatccgttttcctccacggaaatgccggtttctcggctcctcaggtctcc
ccggttcttcggtaacctcggttcatcgacatctcggattcccggctctcatcggc
tccccggtccctttcccggtccctttcccggtccctttcccggtccctttcccggtc
cctttcccggtccctttcccggtccctttcccggtccctttcccggtccctttcccg
gtcccttcccggtccctttttcccggtccctttcccggtccctttcccggtccctttt
cccggtccctttcccggtccctttcccggtccctttcccggtcctcggctcctcg
gctcctccggttccccggtcccgttccccggtcccgttccccggtcccttcccg
gccacccaattttggattttatttaggttggttttgtatgcgaaaaattatcggacgc
tgagtttttacataaatgtatagacaaggttatccgtgctggtggcatgtttaaacg
tggctgaaaagatggcgccggattatggcgtgcgttcgcgccttttcaaatttcc
cgcttacggtgatgcacaagcagtacattacaaccacgaaaaaatatccc

Initially the gap was 200 base pair in original draft. After closing the gap, it was identified that the gap was 250 base pair with 50 additional nucleotides. The main reason for gap occurrence at this region was due to the repetitive sequence present after the gap. These repetitive sequences are true and these cause mis-assemblies when contigs are arranged resulting from short reads produced by Illumina assembler.

The blue color is the gap sequence and green colors are repetitive sequences.

Gap 3 draft:

tgcggccaagacggtgacggctttgtcgtacgacggacgtatgctgagtgggtc
gtacgtcgtttatactaaagagcaacttaaacgatccctctccacggacaagagg
gctattgtcgaaaagatcttaaaannnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnncgtcgtttatactaa
agagcaacttaaacgatccctctccacggacaagagggctattgtcgaaaagatc
ttaaaatttgtagataccctggcattctggatcataacaacgttagcgatccggag
accttgttatggttacttttttgtgggccccagagtctctgtcagaatccgacctgtttc
ggtcgtgatcgtgagtgcga

Gap 3 sequence result:

cgtggcttgtgaggcgtcttctccgcgtgatcggttcagattagttgagacggcc
ggtggaaacttttgttggtcacgaacgctttgccgaaggaaaggtctgagcag
actcaatgtggcgatacgagcttggtaggtagtgagcgaaacgagggtgtcttc
gacggtcttttgtccttgagtgatgatcgtgctagcggggccggtcttatcgcctc
cataccctcggtcccgggtatgcggccaagacggtgacggctttgtcgtacg
acggacgtatgctgagtgggtcgtacgtcgtttatactaaagagcaacttaaacg
atccctctccacggacaagagggctattgtcgaaaagatcttaaaatttgtagat

acccctggcattctggatcataacaacgttagcgatccggagaccttgttatggtt
actttttgtgggccccagagtctctgtcagaatccgacctgtttcggtcgtgatc
gtgagtgcgagttgtcttttcccgctttgttgccgccagttttctacgataccgtca
ccgattattcttcctacatcaacctggccgagttgtatgtgtatgtctggtataagg
attacgacttctcctcggagtcggcggggtgttacgatttaggtgaggtggccat
ggacagggtcaagaagacgttggcgtcggtttgtgataggtttggcgataaga
acgtacccgtttggccaatatcgtctcgaatatgcatattttgtgctttatataatca
aaacagggtatgtctggacttggcgaagaacgatattaatttcaccgcgtatagt
ccgataatcgtaaaagattgtcgtgacgctgcgac

The initial gap here was 200bp, but after several polymerase chain reaction and realignment we came to the conclusion that there is no any true gap existing. Repeats are usually assembled into single contigs and this can cause false gaps.

In this case, gap 3 has unknown regions in the draft genome (unpublished data) (Yi, 2013). But, after several PCR with specific primers, it showed that the gaps are not really existent because the sequencing result showed some repetitive sequence existing after the gap. The results we obtain were same as the sequence presented below the gap in the original draft. There is no surprise as herpesviruses especially Cytomegalovirus are rich with repetitive sequences. This causes some two contigs to create false gaps when arranged in a scaffold.

Gap 4 draft:

catagttagggtaaaggcaaagcatgacgcagtatctcctatgcaactgatgtct
ctcgtagactgatttacctgattaccgttgtcttttcggaagtatatttgggcccaac
atacgatgtagtagatgatcacgatggcgaagatgatttcggtgaaaatnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnngtctcttccgatctgtcttttcggaagtatatttgggc
ccaacatacgatgtagtagatgatcacgatggcgaagatgatttcggtgaaaat
cacgtagatcactagctgaacggggtctaaatatagttgcggcgtaaggtgatg
tatcgcgttatatgtagacaggttcatgttgccgaaatctataatttttggataatag
cacggaaaaccggcaccgggaaagtgtgccgct

Gap 4 sequence result:

catggtgctgattccgagcaaggcttggctcagattgaccaccgcggtccggta
tcgaacggtgccttgtagctttgggtggattttagcgagattgaaggcgctgcgt
tctgttgactcatagtgggtgacgaacgtgaccacaaaagccgttaagcatatg
aagtacatgcatttagcgaaggcgaccatggacgggaatctgaacgagaggc
tgaggacgtaaatctggaacgcgtccatagttagggtaaaggcaaagcatgac
gcagtatctcctatgcaactgatgtctctcgtagactgatttacctgattaccgttgt
cttttcggaagtatatttgggcccaacatacgatgtagtagatgatcacgatggc
gaagatgatttcggtgaaaatcacgtagatcactagctgaacggggtctaaatat
agttgcggcgtaaggtgatgtatcgcgttatatgtagacaggttcatgttgccga
aatctataatttttggataatagcacggaaaaccggcaccgggaaagtgtgccg
ctatggaaaatataacgatgtttacgaacgagattaacgataacactatagacat
agtccatgttctcacgtttacacgatctacgtgagagagcgtcattgtgtttatcgc
catcgtacctgctatcgcgtcgatggcatccgtctatggggtcaaatagtgaaaa
ctcagtatggcggcgtgtggtcttgtgacgtggtgtgttacatatctcattggtag
atcataaatagcggaggacggtgtctatgaagtatccactcgggttttgaaaacg
cgatggacagacagtggctgtccg

The initial gap here was 200bp but after several polymerase chain reactions and realignments, we come to conclusion that there is no gap existing. The repetitive sequence created the gap as seen previously in gap number 3.

Gap 5 draft:

tacactcacctcgtggcggtaaatctatgtgatcgatgtagtctccgacccgggg
atatccccgctttggggacatctccggatcgacggagccttcggagcggaggt
atctccgcatcggtgaagactccgcctcggagaggtctccgaagcgnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnngtctcggagcggtctccgtctcggagcggtctccgtct
cggagcggtctccgtctcggagcggtctccgtctcggagcggtctccgtctcg
gagcggtcgccgtatcggatccgcagtatccagcacgggaaaactcctgcgc
ggggaatcaccatcataggtgcaattcctcggagcggagtttt

Gap 5 second draft:

cccgctttggggacatctccggatcgacggagccttcggagcggaggtatctccgcatcg
gtgaagactccgcctcggagaggtctccgaagcgaggcaatcctcggatcggtacggtct
ccgtctcggagcggtctccgtctcggagcggtctccgtctcggagcggtctccgtctcgg
agcggtctccgtctcggagcggtctccgtctcggagcggtcgccgtatcggatccgcagt
atccagcacgggaaaactcctgcgcggggaatcaccatcataggtgcaattcctgtctcg
gagcggtctccgtctcggagcggtctccgtctcggagcggtctccgtctcggagcggtct
ccgtctcggagcggtctccgtctcggagcggtcgccgtatcggatccgcagtatccagca
cgggaaaactcctgcgcggggaatcaccatcataggtgcaattcctcggagcggagtttt
ttccgtacaagagcaggtctgaaactgcttcggtactctataccgcgccgttcttccttg
gtattcatcttttacggtcagtcacaccgtgtaatttgtttgttatttgtaatttttac
aggtctcggtgacgcccatctacctgtaacaggcttcccatcctatcttcatccttttta

Gap 5 final result:

cccgctttggggacatctccggatcgacggagccttcggagcggaggtatctccgcatcg
gtgaagactccgcctcggagaggtctccgaagcgaggcaatcctcggatcggtacggtct
ccgtctcggagcggtctccgtctcggagcggtctccgtctcggagcggtctccgtctcgg
agcggtctccgtctcggagcggtctccgtctcggagcggtcgccgtatcggatccgcagt
atccagcacgggaaaactcctgcgcggggaatcaccatcataggtgcaattcctcggagc
ggagtttttccgtacaagagcaggtctgaaactgcttcggtactctataccgcgccgtt
cttccttggtattcatcttttacggtcagtcacaccgtgtaatttgtttgttatttgtaa
ttttttacaggtctcgtgacgcccatctacctgtaacaggcttcccatcctatcttcatc
ctttttataggcggtggtatatgtgggcggaggtgaccttaactatgcaatccccacggt
ccactta

For this region, real gap did exist where the number of sequences in gap region is same as original drafts (unpublished data). The gap size was 200 base pair and, all the sequences in gap region were completely filled. However, after gap closure, we encountered repetitive sequences present in the original draft (unpublished data) but not in the new sequence result we obtained from

PCR as shown in Gap 5 second draft (Yi, 2013). To clarify, we realigned this region again using our data and we got to know that there were no repetitive sequences present. Hence, the false repetitive sequences were removed and the final result was produced.

Red color is gap, blue color is repetitive sequence where no gap in the area.. green color connected before and after the blue color.

Comparison of gap's sequences with RCMV English reference strain:

No. of gaps	Percentage match with reference strain
Gap 1	100%
Gap 2	99%
Gap 3	100%
Gap 4	100%
Gap 5	100%

For further verification, all the sequences in the gap region were compared with RCMV English strain which is a closely related strain with RCMV ALL-03 by using blast tool on NCBI website. Surprisingly, all the gaps showed nearly 100% match with reference strain. Therefore, it can be concluded that the nucleotides in the gap regions are very accurate and precise. According to Lapidus (2009), closely related reference strain can be used for comparative analysis and guide for contig mapping.

Discussion

Next Generation Sequencing of RCMV ALL-03 generated reads, which were constructed into contigs, oriented systematically to best represent their order in the true genome sequence by using RCMV English strain as a reference. Correct order and orientation of contigs resulted in an estimate of the complete genome size. Series of disconnected contigs were joint together into a complete continuous genome sequence known as scaffolding. When developing scaffolding, due to the virus complexity and larger size, final completed sequence could not produce successfully. This is because the size of inter-contig gaps represent as unknown region. This unknown region filled with character 'n' to produce a continuous draft of genomic sequence. It is not surprising to have 5 gaps or unknown region present in draft genome of RCMV all-03 (unpublished data) (Yi, 2013). There is no 100% error free for sequencing projects. Complications can be happen due to the physical limitations like chemical contents, technique of handling tools and gel electrophoresis or unidentified human mistakes (Lapidus, 2009). Draft genome of an organism can be generated in matter of weeks, but the complete sequence require many months or even years due to the additional, time, cost and experiments to finish the genome (Nagarajan *et al*., 2010).

Completing the genome scaffold can be carried out by wet laboratory methods, specific software's or combination of both. Although many advanced software's are available, we choose to close the gaps using conventional PCR method to perform additional sequencing (Nagarajan *et al*., 2010). Closing the gaps using PCR method is always considered as being slower, more tedious, labor intensive and time consuming, but it is the best reliable method to close the gaps (Lapidus, 2009). If the numbers of gaps are below 10, then it is recommended to use PCR method to close the unknown regions. For further validation, assembled contigs were also compared with complete reference genome (RCMV-E) to search and compare for similar sequences.

In our study we encountered three types of problems in closing the gaps. First the actual size of gap is not same as original draft (unpublished data) (Yi, 2013). Second, there are no true gap exist or occurrence of false gap. Third, the false sequence present after the gap.

The primary reason for gaps in assemblies from NGS sequencing data is the presence of repeat regions where the assembler may misassemble the sequence in a repetitive area (Tsai *et al*., 2010), which were seen in gap number 1, 2 and 5. In these gaps, repetitive sequences became the reason for gap occurrence. There are no any programs available to treat repeat regions and also to display the reads that link the contig. According to Mulyukov and Pevzner (2002), up to 80% of repetitive areas can be automatically done, while the remaining ones need manual laboratory experiments to finalize (Lapidus, 2009).

In some extreme cases, false gaps do happen where there are no real gaps existent in the draft genome (Tang *et al*., 2013). The same scenario happened for RCMV ALL-03 draft, where gap 3 and 4, which were previously identified as gaps turned out to be repeat regions or false gaps. These gaps were actually sequences identical with lower part of gap region. This data was also supported by the statement of Zerbino and Birney (2008), where gaps were observed due to repetitive sequences falling along the gap. Hence, the known region was mistakenly identified as a gap due to low coverage of reads. This confirms the appearance of gap 3 and 4. RCMV ALL -03 has a high GC content, which may be the reasons for the gap occurrence.

Other than running PCR, gaps can also be closed by different assemblers and parameters using the raw sequence data. Examples are IMAGE (Tsai *et al*., 2010), VELVET (Zerbino and Birney, 2008), Gapfiller (Boetzer and Pirovano, 2012), SOAPdenovo2 (Luo *et al*., 2012) and FGAP (Piro *et al*., 2014), all of which have their own strategy for assembly. Besides this, resequencing of the whole genome using different sequencing platforms can also be performed to avoid bias from previous results. However, this is labor intensive and costly

depending on the genome size. All the computer algorithms which have been developed to reconstruct entire genome from sequencing reads have never been perfect (Lapidus, 2009). Compared to sanger sequencing, Second Generation Sequencing (SGS) technologies like Illumina, produce shorter reads, high throughput with low cost. The drawbacks of SGS are shorter reads and coverage, which reduces the chances of connectivity. Furthermore, when a repetitive sequence is much longer than a read, then coverage alone cannot compensate and eventually all copies of the sequence will produce gaps in the scaffold (*Schatz et al*., 2010). Paired end sequencing offered by this SGS platform shows it is not as good as sanger sequencing. Hence, it still remains as a major question whether short reads are suitable for large genome projects. By choosing a good and better assembler, high quality draft genome can be produced more easier and faster to finish the process (Lapidus, 2009), while use of closely related reference strain can be used for comparative analysis and a guide for contig mapping, which was the reason why we used RCMV English strain as a reference strain.

There are numerous reasons for completing a draft genome. The draft genome of RCMV ALL-03 coverage is at least 90% of the genome and extra effort had to be made to exclude contaminating sequences, sequence errors and misassemblies. Missing sequences in gap 1 fall in the CDS region, which encodes for everyone associated protein. Hence, we cannot neglect any sequences in the gap region. The complete genome sequence is a high quality reference for comparison with other strains and very suitable for all types of detailed analysis of genomic, proteomic as well as in studying gene regulation.

Conclusion

The gaps in RCMV ALL-03 were closed successfully, hence producing a complete genome sequence which can be used for further exploration. It is very important to choose a best method either laboratory experiments or computer software's to close the gaps and this is based on the number of gaps present and the complexity of the genome.

Accession Number

The complete genome sequence of RCMV ALL-03 has been deposited at NCBI under the accession number: KP967684.

Acknowledgement

We thank the Institute of Bioscience, University Putra Malaysia, for allowing us to use the CLC Genomics Workbench software for genome assembly and annotations.

Funding Information

This work was supported by a Research University Grant Scheme (Rugs) under grant No. 9433909.

Author's Contributions

Krishnan Nair Balakrishnan: Designed and performed experiments, analyzed data, carried out final alignment and wrote the manuscript.

Ashwaq Ahmed Abdullah: Performed experiments, analyzed data and helped in final alignment.

Yusuf Abba: Analyzed and interpreted the data.

Jamilu Abubakar Bala: Interpreted data for the work.

Faez Firdaus Jesse Abdullah: Supervised the analysis and edited the manuscript.

Farina Mustaffa Kamal: Revised the article critically for important intellectual content.

Zeenathul Allaudin Nazariah: Designed the work and organized the study.

Ideris Aini: Revised the article critically for important intellectual content.

Noordin Mohamed Mustapha: Revised the article critically for important intellectual content.

Mohd Azmi Mohd Lila: Contributed substantially to the conception and design of the work. All authors have read and approved the final manuscript.

Ethics

No animals were used in this research.

References

Boetzer, M. and W. Pirovano, 2012. Toward almost closed genomes with GapFiller. Genome Biology, 13: R56. DOI: 10.1186/gb-2012-13-6-r56

Ettinger, J., H. Geyer, A. Nitsche, A. Zimmermann and W. Brune *et al*., 2012. Complete genome sequence of the English isolate of rat cytomegalovirus (Murid herpesvirus 8). J. Virol., 86: 13838. DOI: 10.1128/JVI.02614-12

Lai, K.Y., M.L. Mohd-Azmi, K.K. Khoo and A.R. Sheikh Omar, 1999. Random amplified polymorphic DNA (RAPD) for identification of new rat cytomegaloviruses isolated from rice-field rats. J. Vet. Malaysia, 11: 23-26.

Lapidus, A.L., 2009. Genome sequence databases (overview): Sequencing and assembly. University of California, Creek.

Livingston, R., A. Chatterjee and C.J. Harrison, 2001. Cytomegaloviruses, *eLS*: John Wiley & Sons, Ltd.

Loh, H.S., M.A. Mohd-Lila, S.O. Abdul-Rahman and L.J. Kiew, 2006. Pathogenesis and vertical transmission of a transplacental rat cytomegalovirus. Virol. J., 3: 42-42. DOI: 10.1186/1743-422X-3-42

Loh, H.S., M.L. Mohd-Azmi, K.Y. Lai, A.R. Sheikh-Omar and M. Zamri-Saad, 2003. Characterization of a novel rat cytomegalovirus (RCMV) infecting placenta-uterus of Rattus rattus diardii. Arch. Virol., 148: 2353-2367. DOI: 10.1007/s00705-003-0173-y

Luo, R., B. Liu, Y. Xie, Z. Li and W. Huang *et al*., 2012. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. Gigascience, 1: 18. DOI: 10.1186/2047-217X-1-18

Mulyukov, Z. and P.A. Pevzner, 2002. EULER-PCR: Finishing experiments for repeat resolution. Proceedings of the Pacific Symposium Kauai, Jan. 3-7, Hawaii, USA, pp: 199-210. DOI: 10.1142/9789812799623_0019

Mocarski, E., T. Shenk and R.F. Pass, 2007. Cytomegaloviruses. In: *Fields Virology*, Knipe, D.M., D.E. Howley, R.A. Lamb, M.A. Martin and B. Roizman *et al*. (Eds.), Lippincott Williams and Wilkins, Philadelphia, USA, pp: 2701-2772.

Nagarajan, N., C. Cook, M. Di Bonaventura, H. Ge and A. Richards *et al*., 2010. Finishing genomes with limited resources: Lessons from an ensemble of microbial genomes. BMC Genomics, 11: 242. DOI: 10.1186/1471-2164-11-242

Piro, V.C., H. Faoro, V.A. Weiss, M.B.R. Steffens and F.O. Pedrosa *et al*., 2014. FGAP: An automated gap closing tool. BMC Res. Notes, 7: 371. DOI: 10.1186/1756-0500-7-371

Scalzo, A.A., C.A. Forbes, L.M. Smith and L.C. Loh, 2009. Transcriptional analysis of human cytomegalovirus and rat cytomegalovirus homologues of the M73/M73.5 spliced gene family. Arch Virol., 154: 65-75. DOI: 10.1007/s00705-008-0274-8

Schatz, M.C., A.L. Delcher and S.L. Salzberg, 2010. Assembly of large genomes using second-generation sequencing. Genome Res., 20: 1165-1173. DOI: 10.1101/gr.101360.109

Shendure, J. and J. Hanlee, 2008. Next-generation DNA sequencing. Nat. Biotech., 26: 1135-1145. DOI: 10.1038/nbt1486

Tang, B., Q. Wang, M. Yang, F. Xie and Y. Zhu *et al*., 2013. ContigScape: A Cytoscape plugin facilitating microbial genome gap closing. BMC Genomics, 14: 289. DOI: 10.1186/1471-2164-14-289

Tsai, I.J., T.D. Otto and M. Berriman, 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biology, 11: R41. DOI: 10.1186/gb-2010-11-4-r41

Vink, C., E. Beuken and C.A. Bruggeman, 2000. Complete DNA sequence of the rat cytomegalovirus genome. J. Virol., 74: 7656-7665. DOI: 10.1128/JVI.74.16.7656-7665.2000

Weller, T.H., 1971. The cytomegaloviruses: Ubiquitous agents with protean clinical manifestations. N Engl. J. Med., 285: 267-274. DOI: 10.1056/nejm197107292850507

Xing, Y., D. Medvin, G. Narasimhan, D. Yoder-Himes and S. Lory, 2011. CloG: A pipeline for closing gaps in a draft assembly using short reads. Proceedings of the IEEE 1st International Computational Advances in Bio and Medical Sciences, Feb. 3-5, IEEE Xplore press, Orlando, pp: 202-207. DOI: 10.1109/ICCABS.2011.5729881

Yi, W.Q., 2013. Molecular characterization and of Glycoprotein B and Lower Matriz phosphoprotein of Rat cytomegalovirus. ALL-03.

Yu, D., M.C. Silva and T. Shenk, 2003. Functional map of human cytomegalovirus AD169 defined by global mutational analysis. Proce. Nat. Academy Sci., 100: 12396-12401.

Zerbino, D.R. and E. Birney, 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res., 18: 821-829. DOI: 10.1101/gr.074492.107