Journal of
Proteomics & Bioinformatics

Research Article Open Access

Efficiency of *Corynebacterium pseudotuberculosis* 31 Genome Assembly with the Hi-Q Enzyme on an Ion Torrent PGM Sequencing Platform

Adonney AO Veras[1], Pablo HCG de Sá[1], Kenny C Pinheiro[1], Diego Assis das Graças[1], Rafael Azevedo Baraúna[1], Maria Paula Cruz Schneider[1], Vasco Azevedo[2], Rommel TJ Ramos[1#*] and Artur Silva[1#*]

[1]Institute of Biological Sciences, Federal University Pará, Belém, Pará, Brazil
[2]Institute of Biological Sciences, Federal University Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[#]These authors contributed equally to this work

Abstract

Despite the high accuracy obtained through high throughput sequencing (HTS) platforms, including sequencing methods such as pyrosequencing (Roche 454), ligase (SOLiD System), and recently post-light sequencing using the Ion Torrent PGM that is able to detect ions released during sequencing, there are still many errors inherent in the chemistry used by these platforms, for example, INDEL (Insertion/Deletion) that are very common on platforms 454 of Roche and Ion Torrent PGM; Substitution abound in Illumina platforms and SOLiD. Thus, efforts to address these problems have been undertaken by the sequencing companies. To improve the accuracy of the Ion Torrent PGM reads, Life Technologies has developed an enzyme called Hi-Q.

This work aims to demonstrate the performance of the genome assembly of Corynebacterium pseudotuberculosis Cp31 using the enzyme Hi-Q. To evaluate the results, we used an Ion Torrent dataset obtained without the enzyme for the same strain.

The sequencing using the Hi-Q enzyme affected the accuracy of the reads, improving the assembly quality. As a result, a high number of complete genes related to the reference genome were obtained compared to the previous data (without Hi-Q). Furthermore, the use of Hi-Q reduced the number of contigs. After evaluated the GC bias in the genome produced through the Hi-Q, we identified a reduction of GC-Bias, what can reduce the amount and size of gaps generated in the genome assembly process. Furthermore, the comparison of the amount of pseudogenes observed in the genome annotation of *C. pseudotuberculosis* 31 available at NCBI and the genome sequenced by Ion Torrent PGM with a Hi-Q enzyme, show 3-fold less pseudogenes in the genome obtained through Hi-Q enzyme.

Thus, the high efficiency of Hi-Q was validated, which will be useful to whole-genome sequencing and RNA-Seq projects, due to its high accuracy, compared to the previous chemistry.

Keywords: NGS; Hi-Q enzyme; Genome assembly; Ion torrent; Bioinformatics

Introduction

NGS (Next-generation sequencing) platforms have brought about a revolution in the growth of biological knowledge [1]. These devices feature high data throughput at a reduced cost compared to platforms using the Sanger method [2,3].

The benchtop sequencing platforms have been launched in the past three years, for instance, Miseq by Illumina (www.illumina.com) and the Ion Torrent PGM (Personal genome machine) by Life Technologies [4].

The Miseq system performs sequencing by synthesis, similar to other Illumina sequencers [4]. However, its technology has dramatically reduced the time required per round compared to the Illumina HiSeq [5,6]. In 2011, Life Technologies released the Ion Torrent PGM, which uses sequencing based on semiconductor technology. This system detects hydrogen ions released during DNA sequencing, thereby inaugurating the post-light era [4,5].

Despite the increased accuracy of data produced by such sequencers, many errors remain (substitutions and indels), which are inherent to the chemistry used in these platforms, the method of library preparation and the type of sequencing used. These errors increase the complexity of data processing, and each of these devices introduces specific errors, such as the substitution errors that are prevalent on the Illumina and SOLiD platforms [7].

An error common to the Ion Torrent PGM and Proton platforms, which was observed in Roche 454, is related to the recognition of homopolymers due to the sequencing of these regions in a single cycle. Additionally, there is the issue of detecting the byproducts of the nucleotide incorporation reactions, i.e., pyrophosphates (454) and hydrogen ions (Ion Torrent and Proton). These are synthesis-based methods that measure reagent flow because the intensity of the flow of reactants is directly proportional to the amount of nucleotides incorporated. However, the relationship between the measured flow intensity and the number of nucleotides incorporated is nonlinear in

***Corresponding authors:** Rommel TJ Ramos, Institute of Biological Sciences, Federal University Pará, Belém, Pará, Brazil, Tel: (91) 32018426; E-mail: rommelthiago@gmail.com

Artur Silva, Institute of Biological Sciences, Federal University Pará, Belém, Pará, Brazil, E-mail: arturluizdasilva@gmail.com

Received October 21, 2014; **Accepted** December 14, 2014; **Published** December 21, 2014

Citation: Veras AAO, Sá PHCG, Pinheiro KC, Graças DA, Baraúna RA, et al. (2014) Efficiency of *Corynebacterium pseudotuberculosis* 31 Genome Assembly with the Hi-Q Enzyme on an Ion Torrent PGM Sequencing Platform. J Proteomics Bioinform 7: 374-378. doi:10.4172/jpb.1000342

homopolymeric regions, causing frequent errors in determining the length of such regions, which results in insertions and deletions [8].

The read mapping errors affect the *de novo* and reference assembly processes due to wrong insertions and deletions, which can cause frame-shifts that are identified during genome annotation and described, mainly, to Ion Torrent PGM [3,9,10].

To improve the quality of data produced on the Ion Torrent PGM sequencing platform, Life Technologies has developed a sequencing enzyme called Hi-Q with site-directed mutations in its molecular structure that reduce insertion and deletion error rates compared to the traditional chemistry.

This study aimed to demonstrate the performance during assembly and sequencing of a fragment library from the *Corynebacterium pseudotuberculosis* Cp31 genome using the traditionally marketed enzyme and Hi-Q, and compare the occurrence of pseudogenes between the genome available at NCBI and that assembled with Hi-Q enzyme.

Materials and Methods

Origin of the isolate, organism growth and DNA extraction

The model organism used in this study was C. *pseudotuberculosis* 31, which was isolated from a buffalo in Egypt [11]. The bacterium was grown and the DNA was extracted according to Ramos et al. [12].

Construction of libraries and sequencing

The steps to prepare the library included enzymatic DNA fragmentation into 400 bp fragments (Ion plus fragment library kit, PN#4471252 and Ion Shear Kit, PN#4471248), the binding of specific adapters, size selection performed on 2% E-Gel, PN#G661002, the amplification and dilution of the library, and, finally, emulsion in Ion Onetouch 2 with the Ion PGM Template OT2 400 kit, PN#4479882. The Ion OneTouch ES system was used in the enrichment step.

The Ion PGM Hi-Q Sequencing Solutions kit PN#A24569 was used for sequencing, along with the reagents available in the Ion PGM Hi-Q Sequencing Reagents Kit, PN#A24568. The sequencing was performed on an Ion 318 v2 chip, PN#4484354, following the manufacturer's recommended protocol for producing 400 bp reads.

Data analysis

The quality assessment for the raw data generated by sequencing using the two kits was performed using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/), followed by data processing using the Fastx-toolkit (http://hannonlab.cshl.edu/).

De novo assembly

The assembly process for both datasets was performed using two assembly tools, MIRA version 4.0.2 (http://www.chevreux.org/) and SPADES version 3.1.1 (http://bioinf.spbau.ru/).

The parameters used in the SPADES assembler were as follows: -iontorrent that enable a pipeline specific to data of the Ion Torrent PGM platform; additionally and -careful the error correcting step was performed before the assembly process. For Mira, the default parameters were used.

Assembly quality assessment

The assembly quality was assessed using the QUAST computational tool (http://quast.bioinf.spbau.ru/), which uses an annotated reference genome as the template to validate gene completeness, misassemblies, GC content, and other statistical metrics.

The software also allows multiple assemblies to be analyzed together, thus facilitating comparisons between different assemblies. Therefore, it is possible to evaluate the performance of the Hi-Q enzyme for genome assembly.

Genome annotation

The genome annotation was performed using RAST platform [13].

Results/Discussion

Bacterial growth, the extraction of chromosomal DNA and library construction for sequencing were performed using the same protocol for both samples. The raw data obtained from the two datasets, Cp31fragments and Cp31Hi-Q, were evaluated using FastQC (Figure 1). This analysis showed the higher quality (above Phred 20) of the Cp31Hi-Q data throughout the reads, so no quality filter was applied to the data sets.

The results for the assemblies obtained using the computational tools MIRA and SPADES are listed in Table 1. It was observed that even when using different assembly software, the Hi-Q library yielded better results than the library without Hi-Q. Furthermore, the SPADES tool yielded better values for N50, total contigs and total bases.

The assemblies were evaluated using the QUAST software [14], and the Hi-Q data were of higher quality, as shown in Figure 2. Figure 2A shows the gene completion of all four assemblies, and it is evident that the two assemblies using Hi-Q (blue and purple) featured better values for completeness relative to the reference (black dashes), which had fewer contigs than the assemblies without the enzyme (red and green); those without the enzyme had more contigs and lower gene completeness. These results demonstrate that the Hi-Q enzyme yields better performances in the assembly process.

Furthermore, the use of the Hi-Q enzyme was associated with the generation of longer contigs (Figure 2B), which is reflected by the blue and purple lines representing contigs >200 kb and >500 kb, respectively. In contrast, the data without Hi-Q (red and green) had contigs well below 100 kb. These results demonstrate once again the higher quality of the data produced with the Hi-Q enzyme.

These results demonstrate a significant improvement in the chemistry used in the sequencing kit, despite the improvements made for the generation of 400 bp reads. According to Loman et al. [4], there was lower accuracy (with a rate below 60% for homopolymers with a length equal to or greater than six bases) in the assembly of data from the Ion Torrent PGM platform compared to the results obtained on the 454 GS Junior and MiSeq platforms. This result is directly linked to difficulty in recognizing homopolymers.

After compare the genome annotation of *C. pseudotuberculosis* 31 present in the NCBI database and that sequenced with Hi-Q enzyme, 12 pseudogenes were shared between them and a reduction of 3-fold in amount of pseudogenes in the sequence produced by Hi-Q enzyme was observed (Figure 3). The list used to produce the Venn graph (Figure 3) is available in supplementary material. As an example, Figure 4 shows how the Hi-Q data improved the assembly in homopolymeric regions and fixed a frame-shift.

To evaluate how the Hi-Q improved the results, another analysis was performed using the data produced by the same sequencer (Ion Torrent PGM) using mate-pair and the traditional enzyme, and then

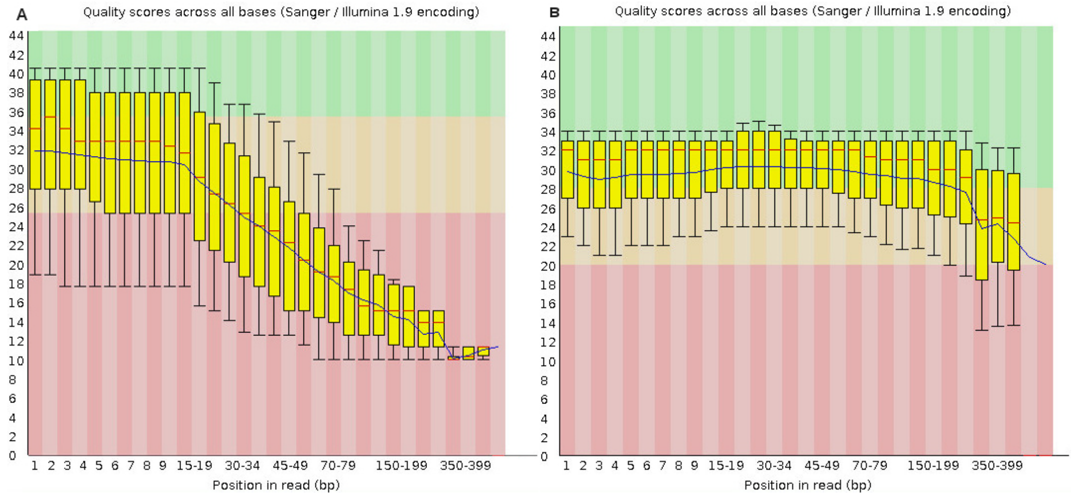


Figure 1: Quality assessment for the Cp31fragments (A) and Cp31Hi-Q (B) libraries.

Description	N50	Longest Contig	Shortest Contig	Total Contigs	Total Bases
Cp31 (Mira)	5,228	22,717	509	687	2,381,462
Cp31Hi-Q (Mira)	374,657	528,805	502	143	2,481,848
Cp31 (SPADES)	809	2,912	500	2,145	1,726,405
Cp31Hi-Q (SPADES)	345,051	655,218	1,582	15	2,387,472

Table 1: Assembly results using the MIRA and SPADES assemblers.

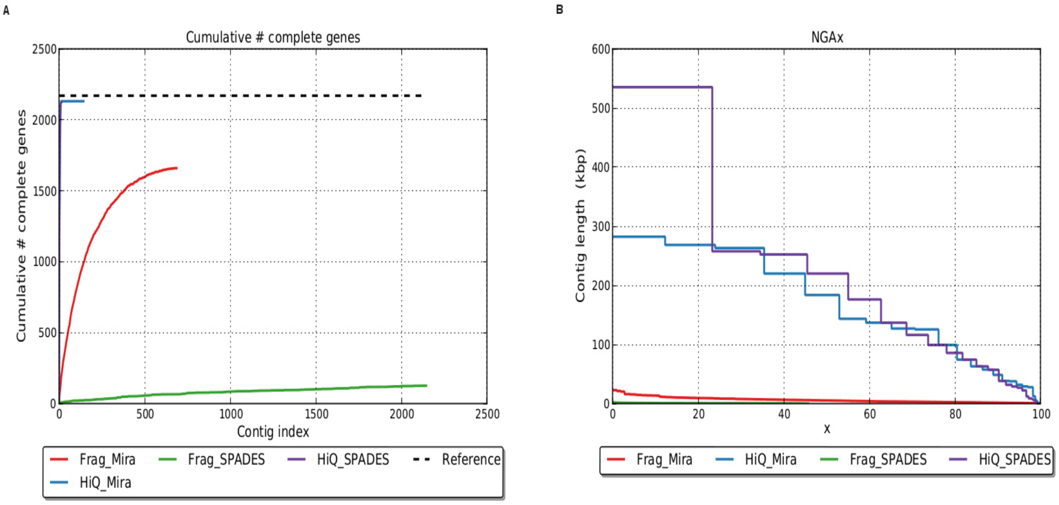


Figure 2: Assembly quality assessment. (A) Analysis of gene completeness and (B) Plot of contig length.

Description	N50	Longest Contig	Shortest Contig	Total Contigs	Total Bases
Cp31_mate_Mira	10,217	53,090	506	461	2,441,500
Cp31_mate_Spades	94,741	219,175	1,465	50	2,405,149
Cp31Hi-Q (Mira)	374,657	528,805	502	143	2,481,848
Cp31Hi-Q (SPADES)	345,051	655,218	1,582	15	2,387,472

Table 2: Comparasion between mate-pair and Hi-Q libraries.

we compared it to Hi-Q genome assembly. Despite using a mate-paired library (Cp31_mate), the final results show an inferior performance when compared to the assembly statistics of the Hi-Q enzyme data (Table 2), which presented the highest N50 and base pairs.

Conclusion

The efficiency of the Hi-Q enzyme was initially evident in the quality assessment of the raw data: better quality data were obtained from the Cp31Hi-Q library than data from the Cp31fragments library that was sequenced without the Hi-Q enzyme.

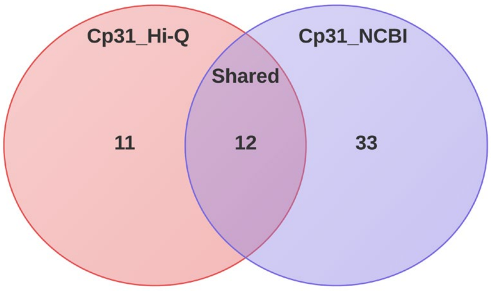
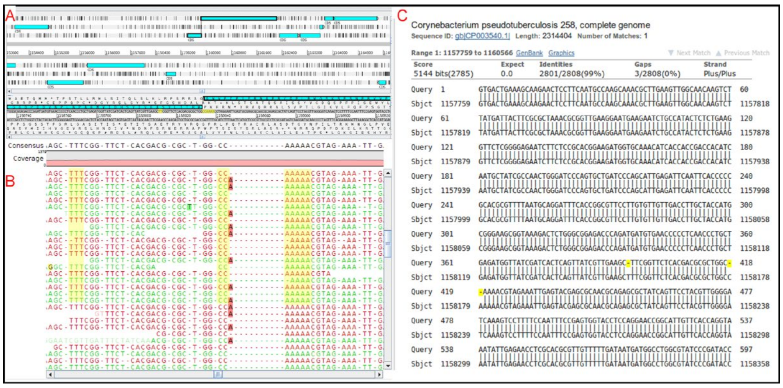


Figure 3: Venn graph of annotated pseudogenes for the C. pseudotubeculosis 31 (Cp31) sequences. Cp31_Hi-Q: the sequence produced with Hi-Q enzyme; Cp31_NCBI: the sequence present in NCBI database; Shared: pseudogenes shared by both genome sequences.



```
Corynebacterium pseudotuberculosis 258, complete genome
Sequence ID: gb|CP003540.1|  Length: 2314404  Number of Matches: 1

Range 1: 1157759 to 1160566 GenBank Graphics          ▼ Next Match  ▲ Previous Match
 Score             Expect   Identities          Gaps            Strand
 5144 bits(2785)   0.0      2801/2808(99%)      3/2808(0%)      Plus/Plus

Query  1        GTGACTGAAAGCAAGAACTCCTTCAATGCCAAGCAAACGCTTGAAGTTGGCAACAAGTCT  60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1157759  GTGACTGAAAGCAAGAACTCCTTCAATGCCAAGCAAACGCTTGAAGTTGGCAACAAGTCT  1157818

Query  61       TATGATTACTTCGCGCTAAACGCGGTTGAAGGAATGAAGAATCTGCCATACTCTCTGAAG  120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1157819  TATGATTACTTCGCGCTAAACGCGGTTGAAGGAATGAAGAATCTGCCATACTCTCTGAAG  1157878

Query  121      GTTCTCGGGGAGAATCTTCTCCGCACGGAAGATGGTGCAAACATCACCACCGACCACATC  180
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1157879  GTTCTCGGGGAGAATCTTCTCCGCACGGAAGATGGTGCAAACATCACCACCGACCACATC  1157938

Query  181      AATGCTATCGCCAACTGGGATCCCAGTGCTGATCCCAGCATTGAGATTCAATTCACCCCC  240
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1157939  AATGCTATCGCCAACTGGGATCCCAGTGCTGATCCCAGCATTGAGATTCAATTCACCCCC  1157998

Query  241      GCACGCGTTTTAATGCAGGATTTCACCGGCGTTCCTTGTGTTGTTGACCTTGCTACCATG  300
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1157999  GCACGCGTTTTAATGCAGGATTTCACCGGCGTTCCTTGTGTTGTTGACCTTGCTACCATG  1158058

Query  301      CGGGAAGCGGTAAAGACTCTGGGCGGAGACCCAGATGATGTGAACCCCCTCAACCCTGCT  360
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1158059  CGGGAAGCGGTAAAGACTCTGGGCGGAGACCCAGATGATGTGAACCCCCTCAACCCTGCT  1158118

Query  361      GAGATGGTTATCGATCACTCAGTTATCGTTGAAGC-TTCGGTTCTCACGACGCGCTGGC-  418
                ||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||| 
Sbjct  1158119  GAGATGGTTATCGATCACTCAGTTATCGTTGAAGCTTTCGGTTCTCACGACGCGCTGGCC  1158178

Query  419      -AAAACGTAGAAATTGAGTACGAGCGCAACGCAGAGCGCTATCAGTTCCTACGTTGGGGA  477
                 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1158179  AAAAACGTAGAAATTGAGTACGAGCGCAACGCAGAGCGCTATCAGTTCCTACGTTGGGGA  1158238

Query  478      TCAAAGTCCTTTTCCAATTTCCGAGTGGTACCTCCAGGAACCGGCATTGTTCACCAGGTA  537
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1158239  TCAAAGTCCTTTTCCAATTTCCGAGTGGTACCTCCAGGAACCGGCATTGTTCACCAGGTA  1158298

Query  538      AATATTGAGAACCTCGCACGCGTTGTTTTTGATAATGATGGCCTGGCGTATCCCGATACC  597
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1158299  AATATTGAGAACCTCGCACGCGTTGTTTTTGATAATGATGGCCTGGCGTATCCCGATACC  1158358
```

Figura 4: Frame-shift correction by Ion Torrent PGM, sequenced by Hi-Q enzyme. The Artemis interface present a frame-shift generated by 3 deletions events which are highlighted (A). After performed the Blast against Corynebacterium pseudotuberculosis 258, we confirmed the deletions were wrong (C). The same region was represented by Ion Torrent PGM (B), but the deletions were fixed, which shows an example of the efficiency of Hi-Q enzyme to address the problems to represent homopolymeric regions.

After the process of assembly an assessment of the completeness of the genes combined with the statistical parameters of the assembly is essential for assessing the accuracy of the assembly process. These analyses revealed better results for data produced with the Hi-Q enzyme on all metrics evaluated.

This result directly implies a reduction of frameshift errors, given that the number of broken genes is directly proportional to the number of such errors in the annotation.

Thus, the use of the Hi-Q enzyme is expected to reduce problems associated with sequencing errors and other experiments that rely on the Ion Torrent PGM sequencing platform.

Acknowledgment

This work was part of the Rede Paraense de Genômica e Proteômica supported by Fundação de Amparo a Pesquisa do Estado do Pará. AAOV, PHGS, VA, AS, RTJR were supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). KCP was supported by Fundação Amazônia Paraense de Amparo à pesquisa (FAPESPA).

References

1. Faino L, Thomma Bart PHJ (2014) Get your high-quality low-cost genome sequence. Trends in Plant Science 19: 288-291.

2. Sanger F, Nicklen S, Coulson A (1977) DNA sequencing with chain-terminating. Proc Natl Acad Sci USA 74: 5463-5467.

3. Wirawan A, Harris RS, Liu Y, Schmidt B, Schröder J (2014) HECTOR: a parallel multistage homopolymer spectrum based error corrector for 454 sequencing data. BMC Bioinformatics 15: 131.

4. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology 30: 434-439.

5. Henson J, Tischler G, Ning Z (2012) Next-generation sequencing and large genome assemblies. Pharmacogenomics 13: 901-915.

6. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

7. Scholz MB, Lo CC, Chain PSG (2012) Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Current Opinion in Biotechnology 23: 9-15.

8. Zeng F, Jiang R, Chen T (2013) PyroHMMsnp: an SNP caller for Ion Torrent and 454 sequencing data. Nucleic Acids Research 41: e136.

9. Oliveira LC, Saraiva TDL, Soares SC, Ramos RTJ, Sá PHCG, et al. (2014) Genome Sequence of Lactococcus lactis subsp. lactis NCDO 2118, a GABA-Producing Strain. Genome Announcements 2: 9-10.

10. Ramos RTJ, Carneiro AR, de Castro Soares S, Barbosa S, Varuzza L, et al. (2013) High efficiency application of a mate-paired library from next-generation sequencing to postlight sequencing: Corynebacterium pseudotuberculosis as a case study for microbial de novo genome assembly. Journal of Microbiological Methods 95: 441-447.

11. Silva A, Ramos RTJ, Ribeiro Carneiro A, Cybelle Pinto A, de Castro Soares S, et al. (2012) Complete genome sequence of Corynebacterium pseudotuberculosis Cp31, isolated from an Egyptian buffalo. Journal of Bacteriology 194: 6663–6664.

12. Ramos RTJ, Carneiro AR, Soares SDC, dos Santos AR, Almeida S, et al. (2013) Tips and tricks for the assembly of a Corynebacterium pseudotuberculosis genome using a semiconductor sequencer. Microbial Biotechnology 6: 150-156.

13. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: 75.

14. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072-1075.