Draft genome of the *Leptospira interrogans* strains, Acegua, RCA, Prea, and Capivara, obtained from wildlife maintenance hosts and infected domestic animals

Frederico S Kremer[1], Marcus R Eslabão[1], Sérgio Jorge[1], Natasha R Oliveira[1], Julia Labonde[1], Monize NP Santos[1], Leonardo G Monte[1], André A Grassmann[1], Carlos EP Cunha[1], Karine M Forster[1], Luísa Z Moreno[2], Andrea M Moreno[2], Vinicius F Campos[1], Alan JA McBride[1], Luciano S Pinto[1], Odir A Dellagostin[1]/+

[1]Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico, Núcleo de Biotecnologia, Pelotas, RS, Brasil
[2]Universidade de São Paulo, São Paulo, SP, Brasil

In the present paper, we announce new draft genomes of four Leptospira interrogans *strains named Acegua, RCA, Prea, and Capivara. These strains were isolated in the state of Rio Grande do Sul, Brazil, from cattle, dog, Brazilian guinea pig, and capybara, respectively.*

Key words: *Leptospirosis* - genomics - neglected diseases - bioinformatics

The *Leptospira* genus comprises at least 22 different species, some of which, like *Leptospira interrogans*, *Leptospira borgpetersenii*, *Leptospira santarosai*, *Leptospira noguchii*, and *Leptospira kirschneri*, are pathogenic and may cause leptospirosis (Boonsilp et al. 2013, Bourhy et al 2014). This neglected zoonosis is globally distributed and has become a reemerging public health problem in many countries, with stronger impact in tropical regions (Evangelista & Coburn 2010, Guerra 2013). Commonly found in rodents, leptospires may also infect and be hosted by different domestic and wildlife animals (Bharti et al. 2003). This wide variety of reservoirs may play a key role in the maintenance and transmission of the disease (Levett 2001). Therefore, genome sequencing of isolates from different hosts potentially provides a starting point to towards understanding the ability of *Leptospira* spp to adapt to specific host and the basis of pathogen-host interaction.

In the present study, whole-genome sequencing was performed for the strains Acegua, isolated from a stillborn bovine foetus (Monte et al. 2015), RCA, isolated from a domestic dog with clinical leptospirosis, Prea, isolated from Brazilian guinea pig (*Cavia aperea*) (Monte et al. 2013), and Capivara, isolated from capybara (*Hydrochoerus hydrochaeris*) (Jorge et al. 2012).

The isolates were cultured in Ellinghausen-McCullough-Johnson-Harris (EMJH) medium supplemented with 10% *Leptospira* enrichment EMJH (Difco, USA), 200 µg/mL 5-fluorouracil, and 5% foetal calf serum in an incubator at 30ºC without agitation. DNA extraction was performed using the commercial Illustra Bacteria GenomicPrep Mini Spin kit (GE Healthcare, USA), following the manufacturer instructions.

The whole genome sequences were obtained using an Illumina MiSeq paired-end library for Acegua, an Illumina MiSeq paired-end library and an Ion Torrent PGM fragment library for RCA and Prea, and an Ion Torrent PGM fragment library for Capivara. The raw reads were filtered by quality using Fastx-Toolkit (hannonlab.cshl.edu/fastx_toolkit/) and the paired-end reads were trimmed using Trimmomatic (Bolger et al. 2014).

De novo assembly was performed using A5 (Tritt et al. 2012), SGA (Simpson & Durbin 2012), and Ray (Boisvert et al. 2010) for Acegua, A5, SGA, Ray, MIRA (chevreux.org/), Newbler (roche.com/), and SPAdes (Bankevich et al. 2012) for RCA and Prea, and MIRA, Newbler, and SPAdes for Capivara. For each strain the *de novo* assemblies were merged using CISA (Lin & Liao 2013) and evaluated using QUAST (Gurevich et al. 2013). Genome annotation was performed as previous described (Kremer et al. 2015) using Prodigal (Hyatt et al. 2010), NCBI-BLAST+ (Altschul et al. 1990, Camacho et al. 2009), Uniprot (Apweiler et al. 2004), HMMER (Eddy 2011), AntiFam (Eberhardt et al. 2012), tRNAscan-SE (Lowe & Eddy 1997), RNAmmer (Lagesen et al. 2007), INFERNAL (Nawrocki et al. 2009), Aragorn (Laslett 2004), and Rfam (Griffiths-Jones et al. 2003), and manually reviewed using Artemis (Rutherford et al. 2000).

In silico multilocus sequence typing (MLST) was performed using BLASTn from the NCBI-BLAST+ and allele data from the *Leptospira* MLST scheme 1 (Boonsilp et al. 2013), obtained from PubMLST repository (pubmlst.org/).

The results of the *de novo* assemblies are presented in Table I. The isolates were initially sequenced using only the Illumina platform, but the high fragmentation in the resulting assembly for Prea and RCA isolates (data

doi: 10.1590/0074-02760160010
Financial support: CNPq, CAPES, FAPERGS, FAPESP (2011/18290-0, 2013/17136-2)
+ Corresponding author: odir@ufpel.edu.br
Received 12 January 2016
Accepted 3 March 2016

not showed) motivated the use of a second next-generation sequencing technology to improve the original draft sequences. Although usually not required, the combination of data of two or more platforms in the sequencing of a given genome may result in a more accurate assembly, considering that each sequencing technology has it owns bias. The most common errors associated with Illumina data occurs on CG-poor and CG-rich regions, while IonTorrent, duo to its chemistry, has a high error-rate in homopolymeric regions. In fact, both characteristics are found in *Leptospira* genomes.

During genome annotation (Table II), by using our pipeline, in addition to the coding DNA sequences, we were also able to identify many noncoding feature in all four genomes, including not only transfer RNAs and ribosomal RNAs, but also transfer-messenger RNAs (tmRNAs), RNase P *loci*, and riboswitches. There is an increasing interest in the analysis of gene expression in *Leptospira*, especially during infection (Matsui et al. 2012, Lehmann et al. 2013, Caimano et al. 2014, Eshghi et al. 2014). Recent studies have already performed whole-transcriptome sequencing of *L. interrogans* and many noncoding features associated with gene expression regulation and transcriptional/translational processing were identified, including RNase P, tmRNAs, riboswitches, as well other families of noncoding RNA. Therefore, the identification of noncoding features in the annotation of newly sequenced genomes may allow a more accurate description of the resulting transcriptome.

The *in silico* MLST sequence types (ST) for the four isolates are presented in Table III. Previously identified by variable-number tandem-repeat as *L. interrogans* serogroup Australis serovar Muenchen (Monte et al. 2015), the Acegua isolate was a match for ST24 that contains two *L. interrogans* serogroup Australis isolates, while the Capivara isolate was identified as ST17 that includes nine *L. interrogans* serogroup Icterohaemorrhagiae isolates (5 belonging to serovar Copenhageni and 2 to serovar Icterohaemorrhagiae). Preliminary analysis revealed that the *pfkB locus* was absent in the draft assemblies of RCA and Prea. To investigate this fact, the raw reads from these isolates were aligned using BLASTn against a reference set of *pfkB* alleles obtained from the PubMLST repository. The BLAST XML output

TABLE I

Summary of the assembly results

Isolate	Scaffolds[a] (n)	Assembly length (Mb)	N50 (bp)	CG (%)
Acegua	158	4.6	63,489	35.07
RCA	89	4.43	55,782	35.06
Prea	106	4.44	46,508	35.17
Capivara	160	4.51	45,526	34.98

a: all assembled sequences joined (or not) by linkage information.

TABLE II

Summary of the annotation results

Isolate	CDS	tRNAs	rRNAs	Other ncRNAs[a]	Riboswitches
Acegua	3734	37	4	2	3
RCA	3591	34	3	10	5
Prea	3616	33	5	10	3
Capivara	4146	37	3	7	3

a: includes the noncoding RNAs identified by Rfam and Aragorn; CDS: coding DNA sequence; ncRNAs: noncoding RNA; rRNAs: ribosomal RNAs; tRNAs: transfer RNAs.

TABLE III

Sequence types (ST) profiles of the Acegua, RCA, Prea, and Capivara strains based on the *Leptospira* multilocus sequence typing scheme 1

Isolate	*glmU*	*pntA*	*sucA*	*tpiA*	*pfkB*	*mreA*	*caiB*	ST
Acegua	1	4	2	1	5	3	4	24
RCA	1	1	2	2	10[a]	4	8	17
Prea	1	1	2	2	10[a]	4	8	17
Capivara	1	1	2	2	10	4	8	17

a: hit not found in the BLAST against the draft genome assembly.

was analysed by a Python script to identify reads that correspond to this *locus* using an identity threshold of 95%. The selected reads were saved in FASTQ format and filtered by quality using a minimum Phred score of 20 in at least 95% of the bases. After filtering, 83 reads remained in the Prea set, and 90 in the RCA set, corresponding to mean coverages of about 18 and 20-fold, respectively. Therefore, the absence of this *locus* in both draft genomes was a result of an assembly artifact. For each genome, the reads that aligned to the *pfkB* database were assembled using CAP3 (Huang & Madan 1999) and the resulting contigs were aligned against the same database to identify the corresponding alleles in the MLST scheme 1, that are showed in Table III.

The *Leptospira* genus comprises more than 300 serovars and pathogenic species were already reported in a wide variety of animal hosts. However, from the 233 genome sequences indexed in BioProject database and available at GenBank with host information, the major part (166) was obtained from human samples (ncbi.nlm.nih.gov/bioproject/). The sequencing of isolates obtained from wildlife animals, like *C. aperea* and *H. hydrochaeris*, both rodents and natural reservoirs, provide data for future pangenome and pathogenome analysis intending to understand the factors that guide the pathogen-host interactions. Additionally, the isolate Acegua, obtained from a bovine stillborn, also represents an interesting source of information about these interactions, since abortion induced by leptospirosis in cattle is usually associated to the serovar Hardjo of the species *L. interrogans* and *L. borgpetersenii*, not to Muenchen, although this serovar has been associated to abortions in pigs (Ellis et al. 1986).

Finally, the analysis of these isolates also provide new insights into the serogroups circulating in the south of Brazil, suggesting that while *L. interrogans* serogroup Icterohaemorrhagiae serovars Icterohaemorrhagiae and Copenhageni are present, they are not the only ones. Based on the MLST profiles, serovars belonging to serogroup Australis are also circulating among wild and domestic animals, and the comparative analysis of genomic data may be applied to trace their distribution and evolution. Furthermore, the availability of these new genome sequences from four *L. interrogans* strains, isolated from diverse hosts, will provide useful data towards understanding the molecular diversity and pathogenesis of these new strains.

Nucleotide sequence accessions - These Whole Genome Shotgun projects have been deposited at DDBJ/EMBL/GenBank under the accessions LCZF00000000 for Acegua, LJBP00000000 for RCA, LJBO00000000 for Prea, and LJBQ00000000 for Capivara. The versions described in this paper are LCZF01000000, LJBP01000000, LJBO01000000, and LJBQ01000000, respectively.

REFERENCES

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

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh L-SL 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res 32*: D115-D119.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin A V, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol 19*: 455-477.

Bharti AR, Nally JE, Ricaldi JN, Matthias MA, Diaz MM, Lovett MA, Levett PN, Gilman RH, Willig MR, Gotuzzo E, Vinetz JM 2003. Leptospirosis: a zoonotic disease of global importance. *Lancet Infect Dis 3*: 757-771.

Boisvert S, Laviolette F, Corbeil J 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol 17*: 1519-1533.

Bolger AM, Lohse M, Usadel B 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics 30*: 2114-2120.

Boonsilp S, Thaipadungpanit J, Amornchai P, Wuthiekanun V, Bailey MS, Holden MTG, Zhang C, Jiang X, Koizumi N, Taylor K, Galloway R, Hoffmaster AR, Craig S, Smythe LD, Hartskeerl RA, Day NP, Chantratita N, Feil EJ, Aanensen DM, Spratt BG, Peacock SJ 2013. A single multilocus sequence typing (MLST) scheme for seven pathogenic *Leptospira* species. *PLoS Negl Trop Dis 7*: e1954.

Bourhy P, Collet L, Brisse S, Picardeau M 2014. *Leptospira mayottensis* sp. nov., a pathogenic species of the genus *Leptospira* isolated from humans. *Int J Syst Evol Microbiol 64*: 4061-4067.

Caimano MJ, Sivasankaran SK, Allard A, Hurley D, Hokamp K, Grassmann AA, Hinton JCD, Nally JE 2014. A model system for studying the transcriptomic and physiological changes associated with mammalian host-adaptation by *Leptospira interrogans* serovar Copenhageni. *PLoS Pathog 10*: e1004004.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL 2009. BLAST+: architecture and applications. *BMC Bioinformatics 10*: 421.

Eberhardt RY, Haft DH, Punta M, Martin M, O'Donovan C, Bateman A 2012. AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database (Oxford) 2012*: bas003.

Eddy SR 2011. Accelerated profile HMM searches. *PLoS Comput Biol 7*: e1002195.

Ellis WA, McParland PJ, Bryson DG, Thiermann AB, Montgomery J 1986. Isolation of leptospires from the genital tract and kidneys of aborted sows. *Vet Rec 118*: 294-295.

Eshghi A, Becam J, Lambert A, Sismeiro O, Dillies M-A, Jagla B, Wunder EA, Ko AI, Coppee J-Y, Goarant C, Picardeau M 2014. A putative regulatory genetic *locus* modulates virulence in the pathogen *Leptospira interrogans*. *Infect Immun 82*: 2542-2552.

Evangelista KV, Coburn J 2010. *Leptospira* as an emerging pathogen: a review of its biology, pathogenesis, and host immune responses. *Future Microbiol 5*: 1413-1425.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR 2003. Rfam: an RNA family database. *Nucleic Acids Res 31*: 439-441.

Guerra MA 2013. Leptospirosis: public health perspectives. *Biologicals 41*: 295-297.

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

Huang X, Madan A 1999. CAP3: A DNA sequence assembly program. *Genome Res 9*: 868-877.

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics 11*: 119.

Jorge S, Monte LG, Coimbra MA, Albano AP, Hartwig DD, Lucas C, Seixas FK, Dellagostin OA, Hartleben CP 2012. Detection of virulence factors and molecular typing of pathogenic *Leptospira* from capybara (*Hydrochaeris hydrochaeris*). *Curr Microbiol 65*: 461-464.

Kremer FS, Eslabão MR, Provisor M, Woloski RDS, Ramires OV, Moreno LZ, Moreno AM, Hamond C, Lilenbaum W, Dellagostin OA 2015. Draft genome sequences of *Leptospira santarosai* strains U160, U164, and U233, isolated from asymptomatic cattle. *Genome Announc 3*: e00910-e00915.

Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res 35*: 3100-3108.

Laslett D 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res 32*: 11-16.

Lehmann JS, Fouts DE, Haft DH, Cannella AP, Ricaldi JN, Brinkac L, Harkins D, Durkin S, Sanka R, Sutton G, Moreno A, Vinetz JM, Matthias MA 2013. Pathogenomic inference of virulence-associated genes in *Leptospira interrogans*. *PLoS Negl Trop Dis 7*: e2468.

Levett PN 2001. Leptospirosis. *Clin Microbiol Rev 14*: 296-326.

Lin S-H, Liao Y-C 2013. CISA: contig integrator for sequence assembly of bacterial genomes. *PLoS ONE 8*: e60843.

Lowe TM, Eddy SR 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res 25*: 955-964.

Matsui M, Soupé M-E, Becam J, Goarant C 2012. Differential in vivo gene expression of major *Leptospira* proteins in resistant or susceptible animal models. *Appl Environ Microbiol 78*: 6372-6376.

Monte LG, Jorge S, Xavier MA, Leal FMA, Amaral MG, Seixas FK, Dellagostin OA, Hartleben CP 2013. Molecular characterization of virulent *Leptospira interrogans* serogroup icterohaemorrhagiae isolated from *Cavia aperea*. *Acta Trop 126*: 164-166.

Monte LG, Ridieri KF, Jorge S, Oliveira NR, Hartwig DD, Amaral MG, Hartleben CP, Dellagostin OA 2015. Immunological and molecular characterization of *Leptospira interrogans* isolated from a bovine foetus. *Comp Immunol Microbiol Infect Dis 40*: 41-45.

Nawrocki EP, Kolbe DL, Eddy SR 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics 25*: 1335-1337.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, Barrell B 2000. Artemis: sequence visualization and annotation. *Bioinformatics 16*: 944-945.

Simpson JT, Durbin R 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res 22*: 549-556.

Tritt A, Eisen JA, Facciotti MT, Darling AE 2012. An integrated pipeline for *de novo* assembly of microbial genomes. *PLoS ONE 7*: e42304.