



Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

Succinctus

A reference genome and methylome for the *Plasmodium knowlesi* A1-H.1 line

Ernest Diez Benavente^a, Paola Florez de Sessions^b, Robert W. Moon^a, Munira Grainger^c, Anthony A. Holder^c, Michael J. Blackman^{a,c}, Cally Roper^a, Christopher J. Drakeley^a, Arnab Pain^d, Colin J. Sutherland^a, Martin L. Hibberd^{a,b}, Susana Campino^a, Taane G. Clark^{a,e,*}

^a Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom^b Genomics Institute Singapore, Singapore^c The Francis Crick Institute, 1 Midland Road, London NW1 1AT, United Kingdom^d King Abdullah University of Science and Technology, Thuwal, Saudi Arabia^e Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

ARTICLE INFO

Article history:

Received 15 April 2017

Received in revised form 10 August 2017

Accepted 26 September 2017

Available online xxx

Keywords:

Plasmodium knowlesi

Reference genome

PkA1-H.1

Sequencing

Malaria

Methylation

ABSTRACT

Plasmodium knowlesi, a common parasite of macaques, is recognised as a significant cause of human malaria in Malaysia. The *P. knowlesi* A1H1 line has been adapted to continuous culture in human erythrocytes, successfully providing an in vitro model to study the parasite. We have assembled a reference genome for the PkA1-H.1 line using PacBio long read combined with Illumina short read sequence data. Compared with the H-strain reference, the new reference has improved genome coverage and a novel description of methylation sites. The PkA1-H.1 reference will enhance the capabilities of the in vitro model to improve the understanding of *P. knowlesi* infection in humans.

© 2017 The Authors. Published by Elsevier Ltd on behalf of Australian Society for Parasitology. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Plasmodium knowlesi, a common malaria parasite of long-tailed *Macaca fascicularis* and pig-tailed *Macaca nemestrina* macaques in southeastern Asia, is now recognised as a significant cause of human malaria. Clinical outcomes range from high parasitaemia to severe complications including death (Singh and Daneshvar, 2013). Early cases were largely misdiagnosed as *Plasmodium malariae*, a morphologically similar but distantly-related species (Lee et al., 2009). Although sporadic human cases had been described in the 1960s, the public health importance of *P. knowlesi* was first understood with the reporting of a substantial focus of human infections in Malaysian Borneo in 2004 (Singh et al., 2004). Molecular detection has confirmed that human cases of *P. knowlesi* infection are relatively common in eastern Malaysia and occur in other southeastern Asian countries including The Philippines, Indonesia, Vietnam and Myanmar (Kantele and Jokiranta, 2011). The geographical distribution of *P. knowlesi* appears to be constrained by the range of its natural macaque hosts and the Leucosphyrus mosquito group vector (Moyes et al., 2014). There is no evidence of sig-

nificant human-to-human transmission of *P. knowlesi* (Millar and Singh, 2015). Compared with other *Plasmodium* spp., the field of *P. knowlesi* genomics has been understudied, with only one reference genome available. The sequencing of the *P. knowlesi* reference “H-Pk1 (A+) clone” (PKNH, 14 chromosomes, 23.5 Mb, 5188 genes, 37.5% GC content) (Pain et al., 2008) has provided insights into novel genomic features, including highly variable *kir* and *sicavar* protein families, but is incomplete. Genetic diversity among *P. knowlesi* isolates is high compared with other members of the genus. Genome variation in this species exhibits dimorphism (Pinheiro et al., 2015), and there is evidence this may be driven by partitioning between the two distinct macaque hosts, *M. fascicularis* and *M. nemestrina* (Assefa et al., 2015).

Recently, the *P. knowlesi* A1.H1 (PkA1-H.1) line was the first to be successfully adapted to continuous culture in human erythrocytes, providing an in vitro model suitable for genetic modification (Moon et al., 2013). To support in vitro studies with this human-adapted clonal line we have assembled a new reference genome for the PkA1-H.1 line using PacBio (Pacific Biosciences Inc., USA) RS-II long read and Illumina HiSeq short read sequence data. An advantage of PacBio RS-II SMRT cell sequencing is the potential

* Corresponding author at: Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, United Kingdom.

E-mail address: taane.clark@lshtm.ac.uk (T.G. Clark).

<https://doi.org/10.1016/j.ijpara.2017.09.008>

0020-7519/© 2017 The Authors. Published by Elsevier Ltd on behalf of Australian Society for Parasitology.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Please cite this article in press as: Benavente, E.D., et al. A reference genome and methylome for the *Plasmodium knowlesi* A1-H.1 line. Int. J. Parasitol. (2017), <https://doi.org/10.1016/j.ijpara.2017.09.008>

to identify modifications on individual nucleotide bases, and thereby provide insights into methylation sites (Morgan et al., 2016). Cytosine and adenine DNA methylation is an epigenetic mark in most eukaryotic cells that regulates numerous processes including gene expression and stress responses. Genome-wide analysis of DNA methylation in *Plasmodium falciparum* has mapped the positions of methylated cytosines. This work has identified a single functional DNA methyltransferase, *PfDNMT* (PF3D7_0727300), which may mediate these genomic modifications (Ponts et al., 2013) and is thus a potential target for anti-malarial drugs. Analyses have revealed that the malaria genome is asymmetrically methylated, in which only one DNA strand is methylated, and this could regulate virulence gene expression and transcription elongation (Ponts et al., 2013). Using PacBio RS-II data, we describe for the first known time, over 40 000 potential modified bases in the *P. knowlesi* genome, including ~5% that were specifically pinpointed as 6-methyladenine modifications, which recently have been shown to have a role in epigenetic regulation of gene expression in other eukaryotic organisms (Greer et al., 2016). Both the PkA1.H-1 reference genome and the associated methylation data are available to support future in vitro and in vivo studies of *P. knowlesi* parasites, and assist the continuing development of anti-malarial drugs and vaccines.

DNA of high quality and molecular weight was purified from phenol chloroform extraction of magnetic activated cell sorted column enriched PkA1-H.1 schizonts (Bates et al., 2010). The DNA (20 µg, A260/280:1.98) was used to prepare 20 kb insert libraries, which were sequenced using nine SMRT cells on the PacBio RS-II device (Pacific Biosciences Inc.) at the Genome Institute of Singapore. The sequencing yielded a total of 365,956 reads with a mean length of 9645 bp. These data were complemented by raw sequences from the Illumina HiSeq2500 platform (500 bp fragment, 150 bp paired end reads), performed at King Abdullah University of Science and Technology, Kingdom of Saudi Arabia, and yielded in excess of 5 million reads. PacBio data were available for the A1-H.1 clone, but were insufficient in themselves to lead to a complete genome i.e. single contig chromosomes (Moon et al., 2016). The de novo assembly of the reads using the HGAP3 software pipeline within the SMRT Portal (Koren et al., 2012) yielded a total of 111 contigs with a theoretical ~145-fold coverage. These contigs were then used as a reference to map with the *bwa-mem* aligner (Li and Durbin, 2009) almost 5 million PkA1-H.1 Illumina 150 bp reads, leading to 7120 single nucleotide polymorphisms (SNPs) and insertion and deletion (indel) corrections. The high quality corrected contigs were ordered with Abacas software (<http://abacas.sourceforge.net>) using the current *Plasmodium knowlesi* H strain (PKNH) reference (version 2.0, www.genedb.org), and manually checked to remove possible errors. This led to 14 complete chromosomes (number of contigs: range 1–5) and mitochondrial (two contigs, copy number seven-fold), and apicoplast (one contig, 1.8 copies) genomes. The PkA1-H.1 genome was then annotated using the Companion webserver (<https://companion.sanger.ac.uk>).

The resulting PkA1-H.1 reference covers 98.2% of the PKNH (v2) reference and 100% of a smaller draft assembly for this line (Moon et al., 2016), and contains only 42 gaps. Compared with the PKNH (v2), PkA1-H.1 has 3993 SNP and 19,936 indel differences. The new genome improved characterisation of *kir* and *sicavar* genes, leading to a total genome length of 24.4 Mb (chromosome size range: 726,979–3,301,832). To assess the performance of the PkA1-H.1 reference, we aligned whole genome sequence data from 60 published human *P. knowlesi* isolates (Assefa et al., 2015; Pinheiro et al., 2015). Raw Illumina sequence data were downloaded for 60 *P. knowlesi* isolates (Assefa et al., 2015; Pinheiro et al., 2015) (Supplementary Table S1). The samples were aligned against the PkA1-H.1 reference using *bwa-mem* (Li and Durbin, 2009) and SNPs

were called using the *Samtools* software suite (Li, 2011), from which a set of high quality SNPs was filtered using previously described methods (Samad et al., 2015; Campino et al., 2016). The alignment process yielded an average coverage of ~143-fold across 99% of the reference genome, and 1,632,024 high quality SNPs (one every 15 bp) were characterised (Fig. 1). These coverage statistics were marginally superior to mapping to PKNH, which yielded average coverage of ~140-fold across 96% of the reference genome (Supplementary Table S1). The improvement includes an increase in coverage in the *kir* and *sicavar* genes, and the closing of several gaps in these genes that exist in the PKNH reference.

The 38-fold sequence coverage obtained during the new sequencing was in excess of the recommended 25-fold for the 6-methyladenine (m6A) and 4-methylcytosine (m4C) modification detection (Morgan et al., 2016). The *RS_Modification_and_Motif_analysis* program within the SMRT Analysis Portal was used to perform a methylation analysis (Morgan et al., 2016). Overall, we have identified 41,508 potential modified bases (*P* values <0.01) in the *P. knowlesi* genome (by chromosome: mean 2965; range 1089–5823) (Supplementary Table S2). Of the total, 7231 modifications (~17%) were identified from an independent unanalysed dataset (Moon et al., 2016). Approximately five percent (2218) of the modifications were specifically pinpointed as m6A modifications, a type that has been recently confirmed to play a role in epigenetic regulation of eukaryotic organisms such as *Caenorhabditis elegans* (Greer et al., 2016). Furthermore, 3646 (~9%) modified bases were classified as m4C methylation events. The proportion of the total adenosine and cytosine bases that are modified is 0.11 and 0.25, respectively. These fractions are lower than those reported in a study of *P. falciparum* for 5-methylcytosine (m5C), which estimated that two-thirds of the cytosine bases were methylated (Ponts et al., 2013). Analysis of the distribution of the methylation sites across the genome revealed that 45.3% of the m6A sites are located within gene boundaries. However, the m4C modifications were distributed more evenly, with 50.1% being intragenic (m4C versus m6A, *P* < 0.0004). We calculated the number of modified bases over a 10 kb sliding window, revealing a stable distribution over the different chromosomes (modified bases/kb: mean 1.6; range 1.38–1.81) (Fig. 1). In order to identify genomic regions (islands) with an accumulation of modified bases, the fold change of modified bases per window was compared with the chromosome average. This analysis revealed 85 10 kb regions that present at least an increase of two-fold change over the chromosome average in either of the two PacBio datasets analysed (Table 1, Supplementary Table S3). These regions included four methyltransferase genes, and several loci involved in the ribosome constitution and with DNA/RNA manipulation functionality. Further, the *reticulocyte binding protein NBPXa* gene (PKNH_1472300) was also identified, which has been demonstrated to be essential for the infection of human red blood cells in this *P. knowlesi* line (Moon et al., 2016).

The dominant cause of malaria in Malaysia is now *P. knowlesi*. To assist with disease control, a deeper understanding of the biology of this neglected parasite is required, and genomics has the potential to provide useful insights. The use of state-of-the-art technology such as PacBio RS-II SMRT cell sequencing has assisted with the biological understanding of constantly changing parasite populations (Ahmed et al., 2016). Using sequence data from PacBio RS-II and Illumina HiSeq2500 technologies, we have constructed a reference genome for the PkA1-H.1 line, which is the first known *P. knowlesi* line to be successfully adapted to continuous culture in human erythrocytes. The PkA1-H.1 reference is complete to a chromosome level, base-corrected, fully annotated, and spans over 98% of the PKNH “H strain” reference, including the closure of some of its gaps in highly variable *kir* and *sicavar* genes. The completeness of the PkA1-H.1 genome improved alignment of sequences and variant detection from published *P. knowlesi* clinical strains, when

Download English Version:

<https://daneshyari.com/en/article/8500133>

Download Persian Version:

<https://daneshyari.com/article/8500133>

[Daneshyari.com](https://daneshyari.com)