International Journal for Parasitology xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

International Journal for Parasitology

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

Comparative genomic analysis of the IId subtype family of *Cryptosporidium parvum* ☆

Yaoyu Feng [a,b,*], Na Li [a,c], Dawn M. Roellig [c], Alyssa Kelley [c], Guangyuan Liu [b], Said Amer [c,d], Kevin Tang [e], Longxian Zhang [f], Lihua Xiao [c,*]

[a] College of Veterinary Medicine, South China Agricultural University, Guangzhou 510642, China
[b] State Key Laboratory of Veterinary Etiological Biology, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou 730046, China
[c] Division of Foodborne, Waterborne and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA
[d] Department of Zoology, Faculty of Science, Kafr El sheikh University, Kafr El Sheikh 33516, Egypt
[e] Division of Scientific Resources, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA
[f] College of Animal Science and Veterinary Medicine, Henan Agricultural University, Zhengzhou 450002, China

ARTICLE INFO

Article history:
Received 19 May 2016
Received in revised form 19 December 2016
Accepted 22 December 2016
Available online xxxx

Keywords:
Cryptosporidium parvum
Genomics
Whole genome sequencing
Host adaptation
Transmission

ABSTRACT

Host adaptation is known to occur in *Cryptosporidium parvum*, with IIa and IId subtype families preferentially infecting calves and lambs, respectively. To improve our understanding of the genetic basis of host adaptation in *Cryptosporidium parvum*, we sequenced the genomes of two IId specimens and one IIa specimen from China and Egypt using the Illumina technique and compared them with the published IIa IOWA genome. Sequence data were obtained for >99.3% of the expected genome. Comparative genomic analysis identified differences in numbers of three subtelomeric gene families between sequenced genomes and the reference genome, including those encoding SKSR secretory proteins, the MEDLE family of secretory proteins, and insulinase-like proteases. These gene gains and losses compared with the reference genome were confirmed by PCR analysis. Altogether, 5,191–5,766 single nucleotide variants were seen between genomes sequenced in this study and the reference genome, with most SNVs occurring in subtelomeric regions of chromosomes 1, 4, and 6. The most highly polymorphic genes between IIa and IId encode mainly invasion-associated and immunodominant mucin proteins, and other families of secretory proteins. Further studies are needed to verify the biological significance of these genomic differences.
Published by Elsevier Ltd on behalf of Australian Society for Parasitology.

1. Introduction

Cryptosporidium parvum is the *Cryptosporidium* sp. responsible for watery diarrhoea in pre-weaned ruminants (Santin, 2013). It is also the most important zoonotic *Cryptosporidium* sp. in humans (Ryan et al., 2014). Previous sequence characterizations of the 60 kDa glycoprotein (gp60) gene had shown the existence of host adaption in *C. parvum*, with the occurrence of the IIa subtype family mostly in cattle, IIc subtype family mostly in humans, and IId subtype family mostly in sheep and goats, although all three subtype families are human pathogens and IId subtypes have been found in calves in some areas (Xiao, 2010). More recent sequence characterizations at other genetic loci have confirmed the existence of host adapted *C. parvum* subtype families (Widmer and Lee, 2010).

The genetic basis for host adaptation in *C. parvum* is not clear. The genome of one *C. parvum* IIa isolate (IOWA) from a calf in the United States, propagated through calf passages, was among the first two *Cryptosporidium* isolates sequenced (Abrahamsen et al., 2004). More recently, the genome of one IIc isolate from a child in Uganda and propagated in immunosuppressed mice was sequenced (Widmer et al., 2012). Due to the existence of significant sequence differences between the two isolates across the entire genomes, more comparative genomic analysis of other host-adapted *C. parvum* subtypes, especially different subtypes from the same area, is needed in order to better understand the genetic determinants for host adaptation in *C. parvum*.

In this study, we sequenced the genomes of two IId specimens of *C. parvum* from China and Egypt. As a control, we also sequenced the genome of one IIa specimen. The comparative genomic analysis

☆ *Note:* Nucleotide sequence data reported in this paper, including all Sequence Read Achieve (SRA) data and assembled contigs, are available in GenBank under the BioProject accession number PRJNA320419 and BioSample accession numbers SAMN04938568, SAMN04938569 and SAMN04938570.

* Corresponding authors at: College of Veterinary Medicine, South China Agricultural University, Guangzhou 510642, China. Fax: +86 21 6425 0664 (Y. Feng). Fax: +1 404 718 4197 (L. Xiao).
E-mail addresses: yyfeng@scau.edu.cn (Y. Feng), lxiao@cdc.gov (L. Xiao).

http://dx.doi.org/10.1016/j.ijpara.2016.12.002
0020-7519/Published by Elsevier Ltd on behalf of Australian Society for Parasitology.

revealed some differences in the number of several subtelomeric gene families between specimens sequenced in this study and the reference *C. parvum* IOWA, and in the sequences of the invasion-associated and immunodominant mucin-type secretory glycoproteins between IIa and IId subtype families.

2. Materials and methods

2.1. Cryptosporidium specimens

The genomes of three *C. parvum* specimens were sequenced in the study: specimen 31727 of the IIdA19G1 subtype, 34902 of the IIdA20G1 subtype, and 35090 of the IIaA15G1R1 subtype. Specimen 31727 was collected from a 1 month old dairy calf with diarrhoea in Zhengzhou, Henan Province, China in November 2008 and maintained through animal passages in gerbils. Specimen 34902 was collected in January 2011 from a 3 week old buffalo calf with diarrhoea in Sakha, Kafr El Sheikh Province, Egypt. Specimen 35090 was collected in October 2011 from a 5 week old dairy calf with diarrhoea in Al Nubaria, El Beheira Province, Egypt. The two IId subtypes from China and Egypt were targeted for sequencing in this project because they are commonly found in calves in both countries. The three subtypes chosen in this study represented the most common *C. parvum* subtypes in calves and humans in China and Egypt. For each specimen, faecal material was stored in 2.5% potassium dichromate at 4 °C for less than 6 months before use in *Cryptosporidium* oocyst isolation. *Cryptosporidium* species and subtypes were determined by PCR-RFLP analysis of the ssrRNA and sequence analysis of the gp60 genes, respectively (Xiao et al., 2009). The collection of faecal specimens used in the study was approved by the Institutional Committee of the Post-graduate Studies and Research at Kafr El Sheikh University, Egypt, and the Research Ethics Committee of Henan Agricultural University, Zhengzhou, China.

2.2. Oocyst isolation and whole genome sequencing

Cryptosporidium oocysts were isolated from stool specimens by sucrose and cesium chloride gradient centrifugation, and immunomagnetic separation as previously described (Guo et al., 2015a). They were subjected to treatment with 10% commercial bleach on ice for 10 min and five freezing-and-thawing cycles. DNA was extracted using the Qiagen DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA), and amplified using the REPLI-g Midi Kit (Qiagen). The amplified DNA was sequenced on an Illumina Genome Analyzer IIx (Illumina, San Diego, CA, USA). For specimen 31727, a 100 bp paired-end technique was used whereas for specimens 34902 and 35090, a 100 bp single-end technique was used. The sequence reads were analysed for sequencing quality using CLC Genomic Workbench 8.5 (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench). They were trimmed off by 10 nucleotides at the 5′ end and Phred score < 25 at both ends, using the error rate limit setting of 0.02, minimum read length of 65 nucleotides, and ambiguous trim setting of 2. Trimmed reads were assembled de novo using CLC Genomics Workbench with word size 50, bubble size of 400, mismatch cost of 2, insertion and deletion costs of 3, minimum contig length of 500 bp, and updating contigs after read mapping. The word size 50 was selected based on the outcome of de novo assemblies of several *Cryptosporidium* genomes using word sizes of 22 (default), 30, 40, 50 and 60, and assessment of assemblies using QUAST (http://bioinf.spbau.ru/quast). Raw sequence reads were also trimmed using BBDuk from the BBMap package (https://sourceforge.net/projects/bbmap/). Reads were trimmed at both ends for Phred score < 30, phix adapters from BBMap package resources, and 10 bp from the 5′ end with paired-end reads trimmed to equal length and reads shorter than 65 bp removed. Sequence reads were analysed for sequence quality before and after the cleaning using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed sequence reads were also de novo assembled using SPAdes (http://cab.spbu.ru/software/spades/), with word sizes 31, 41, 51 and 61.

2.3. Genome comparison and identification of gene insertions and deletions (indels)

For assessment of gene gains and losses among *C. parvum* isolates, contigs from each genome assembly were aligned with the eight assembled chromosome sequences of the *C. parvum* IOWA isolate of the IIaA15G2R1 subtype (version AAEE00000000.1) by using the progressive alignment algorithm of Mauve 2.3.1 (http://asap.genetics.wisc.edu/software/mauve/) with default options. This reference was also used in other analyses described below, and all analyses were conducted prior to the recent re-annotation of the IOWA genome, which increased the number of annotated genes from 3,805 to 3,865 (Isaza et al., 2015). Major insertions and deletions (indels) in genomic fragments were identified by manual inspection of the genome alignment. Potential genes in major insertions were identified using FGENESH (http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind) or geneid (http://genome.crg.es/software/geneid/geneid.html) webservers. The identities of the genes were established by blastp analysis of the deduced amino acid sequences against the National Center for Biotechnology Information (NCBI), USA, non-redundant protein sequence database. Contigs from contaminants were removed from the assembly through BLAST analysis against the NCBI non-redundant nucleotide database, which produced the final genome. Contigs obtained from the study were further aligned with the reference IOWA genome using NUCmer within the MUMmer 3.0 package (http://mummer.sourceforge.net/), with a minimum cluster length of 100 bp. Indels in the alignments were detected by using the MUMmer show-snps utility and in-house scripts.

2.4. Variant analysis and identification of highly polymorphic genes

To identify highly polymorphic genes, sequence reads of each genome were mapped to the reference *C. parvum* IOWA genome using CLC Genomics Workbench 8.5, with a mismatch cost of 2, insertion and deletion costs of 3, length fraction of 0.5 and similarity fraction of 0.8. The outcome of the read mapping was analysed using the Basic Variant Detection tool in the software, with minimum coverage of 10, variant probability of 90, required variant count of 2. The outcome of the variant detection was exported into Excel and the number of single nucleotide variants (SNVs) present per 1,000 bp along the 9.1 Mb *C. parvum* genome was plotted by using the Pivot Table function of Excel. Due to the likely presence of mismapping of sequence reads to multicopy genes and sequence heterozygosity at some loci (especially cgd2_1370, cgd1_3290, cgd3_4230, cgd5_4510, cgd5_4520, cgd6-1000, cgd6_5510), only homozygous SNVs were considered in the identification of highly polymorphic genes. The number of SNVs present among genomes sequenced in this study was also compared using the same approach.

Alternatively, sequence reads were also mapped to the *C. parvum* IOWA genome using Burrows Wheeler Alignment (BWA) (https://www.msi.umn.edu/sw/bwa) with default parameters. All BAM files were passed to SAMtools mpileup (http://www.htslib.org/) with parameters -g for computing genotype likelihoods, -C50 to handle excessive read depth that could cause errors, -P

Please cite this article in press as: Feng, Y., et al. Comparative genomic analysis of the IId subtype family of *Cryptosporidium parvum*. Int. J. Parasitol. (2017), http://dx.doi.org/10.1016/j.ijpara.2016.12.002