Computational and Structural Biotechnology Journal 12 (2014) 26–33

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Mini Review

Evolutionary genomics and population structure of *Entamoeba histolytica*

Koushik Das, Sandipan Ganguly *

Division of Parasitology, National Institute of Cholera and Enteric Diseases, P-33, CIT Road, Scheme XM, Beliaghata, Kolkata 700010, India

ARTICLE INFO

Article history:
Received 11 July 2014
Received in revised form 2 October 2014
Accepted 3 October 2014
Available online 31 October 2014

Keywords:
Genetic polymorphism
Disease outcome
Genetic recombination
Genotyping
Short tandem repeat loci
Single nucleotide polymorphism

ABSTRACT

Amoebiasis caused by the gastrointestinal parasite *Entamoeba histolytica* has diverse disease outcomes. Study of genome and evolution of this fascinating parasite will help us to understand the basis of its virulence and explain why, when and how it causes diseases. In this review, we have summarized current knowledge regarding evolutionary genomics of *E. histolytica* and discussed their association with parasite phenotypes and its differential pathogenic behavior. How genetic diversity reveals parasite population structure has also been discussed. Queries concerning their evolution and population structure which were required to be addressed have also been highlighted. This significantly large amount of genomic data will improve our knowledge about this pathogenic species of *Entamoeba*.

Contents

1. Introduction

Amoebiasis 4caused by the gastrointestinal parasite *Entamoeba histolytica* is one of the major parasitic diseases after malaria and is responsible for approximately 100,000 human deaths per annum [1]. The parasite has an interchangeable two stage life cycle consisting of an infective cyst form and a motile pathogenic trophozoite form. Infection is endemic in many developing countries where poor sanitation and malnutrition are common. Infection can also be restricted to a certain population in some developed countries (among male homosexual population in Japan) [2,3]. The global prevalence of infection (estimated in 1986) suggested that 10% of the world population was infected by this parasite [4]. *E. histolytica* infection develops variable disease outcomes. 90% of infected individuals remain asymptomatic, while only 10% develops symptoms of invasive amoebiasis [5,6]. However, the global prevalence was estimated prior to the differentiation of *E. histolytica* from its non-pathogenic sibling *Entamoeba dispar* in 1993 [7]. Regardless of this epidemiological modification, invasive amoebiasis is still relatively a rare outcome of *E. histolytica* infection. Specific determinants for the diverse outcomes of this infection still remain obscure. However, host genetics and parasite genotype could be two possible factors [8,9]. Exploring the hidden genetic trait of parasite, directly linked to its virulence or associated with disease outcome, motivates a substantial area of *Entamoeba* research. Intra and inter-specific genomic comparisons have been conducted to identify the parasites' genetic factor linked to its virulence or associated with differential disease causing abilities [10–13]. These studies also provide some interesting and

* Corresponding author. Tel.: +91 33 2363 3855; fax: +91 33 2370 5066.
E-mail addresses: koushikdas55@gmail.com (K. Das), sandipanganguly@gmail.com (S. Ganguly).

http://dx.doi.org/10.1016/j.csbj.2014.10.001

valuable information concerning the evolution and population structure of this parasite. Recent information concerning evolutionary genomics of *E. histolytica* and their association with parasite phenotype and its virulence have been discussed. How parasite population structure is revealed by genetic diversity has also been discussed. Questions related to their evolution and population structure have also been emphasized in this review.

2. Whole-genome sequences of *Entamoeba* species

Several species of *Entamoeba* infects a wide range of hosts [14]. The simplest morphological characteristic like the number of nuclei per cyst has been exploited to distinguish between species [15]. However, morphological variations do not always reflect species-level differences and significant genetic diversity exists among morphologically indistinguishable organisms [15]. Some species like the oral parasite *Entamoeba gingivalis* do not produce cysts [14]. Phylogenetic relationships among *SSU rRNA* gene sequences of *Entamoeba* species suggested that *E. dispar*, *Entamoeba nuttalli* and *Entamoeba moshkovskii* are closely related to *E. histolytica*, while *Entamoeba invadens* and *Entamoeba coli* are distantly related [15]. *E. dispar*, morphologically identical with *E. histolytica* is usually considered as an avirulent commensal of human gut [14]. However, a recent study suggested that a certain strain of *E. dispar* (ICB-ADO), isolated from a Brazilian patient can cause amoebic liver abscess (ALA) in hamsters [16]. *E. moshkovskii* is microscopically indistinguishable from *E. histolytica* and *E. dispar* in its cyst and trophozoite form. It was initially thought to be a free living protozoan species [17] but a recent study suggested that *E. moshkovskii* infects humans and causes diarrhea and colitis in infants [17]. *E. dispar* infection is, in general much more common than *E. histolytica* worldwide [18]. Since, worldwide prevalence of *E. histolytica* infection [4] was estimated prior to the genetic discrimination of *E. histolytica* from *E. dispar*, the prevalence value can be completely erroneous and *E. dispar* could be a potential contributor to the prevalence figures in endemic areas [19]. *E. moshkovskii* can be found more frequently in regions where amoebiasis shows high prevalence [19,20]. *Entamoeba bangladeshi*, recently discovered from Bangladesh was clearly grouped with the clade of *Entamoeba* infecting humans, including *E. histolytica* [21]. *E. invadens* is a reptilian parasite and is an important model for encystation process. *E. invadens* can be induced to encyst in axenic laboratory culture, while encystation has not yet been achieved in axenically grown *E. histolytica* trophozoites [14].

The genome sequence of *E. histolytica* strain HM1:IMSS was published and analyzed in 2005 [22–24]. The genome assembly contains 20, 800, 560 bp of DNA in 1496 scaffolds. The genome has a high AT content (approximately 75%). Approximately half of the assembled sequence is predicted to be coding, with 8333 annotated genes [14]. Genome assembly of *E. dispar* strain SAW760 is of a similar size to that of *E. histolytica* strain HM1:IMSS. It consists of 22,955,291 bp of DNA in 3312 scaffolds. AT content is also quite similar to that of *E. histolytica* strain HM1:IMSS (approximately 76.5%). 50% of the assembled sequence is predicted to be coding, with 8749 annotated genes [14]. Genome assembly of *E. invadens* strain IP1 appears to be larger than that of *E. histolytica* strain HM1:IMSS and *E. dispar* strain SAW760. It contains 40,888,805 bp of DNA in 1149 scaffolds. AT content is comparatively less (approximately 70%). Approximately 38% of the assembled sequence is predicted to be coding, with 11,549 annotated genes. As per AmoebaDB database version 4.1 [25, www.amoebadb.org], genome assembly of *E. moshkovskii* strain Laredo consists of 25, 250,000 bp of DNA in 1147 scaffolds. AT content is approximately 64%. A total of 12,518 annotated genes are present. According to AmoebaDB database version 4.1 [25, www.amoebadb.org], genome assembly of *E. nuttalli* strain P19 consists of 14, 399,953 bp of DNA in 5233 scaffolds. AT content is approximately 75%. A total of 6187 annotated genes are present.

3. Structure and organization of genome

Structure of *E. histolytica* genome has been extensively reviewed by Clark et al. [24]. Many interesting evolutionary features of *E. histolytica* genome have been highlighted. *E. histolytica* have gained a significant number of metabolic genes (at least 68) through horizontal gene transfer from bacteria [14,22,24]. Orthologues of these genes found in both *E. histolytica* and its evolutionary distant species *E. invadens* [15] indicate that gene transfer is ancient [14].

The haploid genome of *E. histolytica* strain, HK9 is 3×10^7 bp in size, based on renaturation kinetics experiments [26]. Hybridization of gene marker to pulse field gels identified 14 linkage groups with 1–4 chromosomes per linkage group per nucleus [27]. Tetra-nucleated *E. histolytica* cyst must contain at least one to two genome copies (1n–2n) in each of the nuclei [28]. However, karyotype analysis of *E. histolytica* trophozoite revealed the presence of at least 4 functional copies of many structural genes and therefore probably a ploidy that is a multiple of four [28]. Ploidy can vary even within a cell lineage under different growth conditions [28]. However, this phenomenon was only studied in-vitro and whether this occurs in nature is not known. The rRNA gene occurs in circular DNA molecules that exist in multiple copies per nucleus [29]. These circular structures could be important for determining parasite phenotypes. The rDNA episome varies in size from 15 kb to 25 kb depending on *E. histolytica* strains. The rDNA episome in *E. histolytica* virulent strain HM1:IMSS has two rDNA units per circle, while *E. histolytica* avirulent strain Rahman has only a single rDNA unit in its episome [30]. Moreover, Jasson et al. reported that structural genes for hemolysins were present within the ribosomal RNA repeat on extra-chromosomal DNA element of *E. histolytica* [31].

Initial characterization of *E. histolytica* genome revealed some unusual features of its organization. *E. histolytica* genome is highly repetitive (about 40% of the sequences are assigned to repetitive elements). Among them, tRNA genes are exceptionally abundant; with an estimated 4500 copies (about 10 times of human genome) were present. Moreover, most of these tRNA genes are clustered and organized into 25 distinct arrays. The tRNA arrays are composed of tandemly repeated units encoding between 1 and 5 tRNA acceptor types [32]. The intergenic regions of these tRNA genes comprises of short tandemly repeated sequences (STRs) which resembles the micro/mini satellites of eukaryotic genomes. The only difference is that unlike randomly dispersed micro/mini satellites, STRs form a part of a larger unit which is itself tandemly arrayed [32]. tRNA genes are thought to be "hotspots" for recombination and mutation due to their unique structural organizations [32]. The arrangement of tRNA gene showed inter-specific variation. *E. histolytica* has 2 versions of tRNA array containing Asn^{GTT} and Lys^{CTT} genes [i.e. (*N-K1*) and (*N-K2*)], while *E. dispar* genome contains only 1 type of [*N-K*] array. *E. moshkovskii* array units are significantly smaller than their homolog in *E. histolytica* and *E. dispar* and their intergenic regions do not contain any STRs [32]. STR regions between these tRNA array units showed high degree of intra-specific variation in their repeat number, type and arrangement patterns [13]. These particular features make them very useful as population genetic markers for quantification of evolutionary divergence of this fascinating parasite. The only proposed function of this tRNA array unit is nuclear matrix binding [33]. Moreover, circumstantial evidence also suggests that they may be located either at subtelomeric or at chromosomal ends and could be functional replacements of traditional telomere repeats [32].

4. Genomic rearrangements and transposable elements

Unlike *Plasmodium* which has a stable genomic organization even among distantly related species, *Entamoeba* exhibit high degree of genomic plasticity and instability [14]. Genome rearrangement associated with tissue invasion and organ tropism has been reported as one possible explanation for the different tRNA STR genotypes identified in liver