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

**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*.

© 2014 Das and Ganguly. Published by Elsevier B.V. on behalf of the Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/3.0/).

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

**Contents**

1. Introduction .......................................................... 26
2. Whole-genome sequences of *Entamoeba* species .......... 27
3. Structure and organization of genome ......................... 27
4. Genomic rearrangements and transposable elements ...... 28
5. Large gene families and their diversities ..................... 28
6. Genetic diversity and population structure ................. 28
7. Conclusion .............................................................. 32
Acknowledgments ......................................................... 32
References ................................................................. 32

**1. Introduction**

Amoebiasis caused 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
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 of human gut [14]. However, a recent study suggested that a certain strain is usually considered as an avirulent commensal of *E. histolytica* [17] but a recent study suggested that related to cyst has been exploited to distinguish between species [15]. However, simplest morphological characteristic like the number of nuclei per bic liver abscess (ALA) in hamsters [16].

### 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 × 10⁷ 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 AsnTTT and LysTTT 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
5. Large gene families and their diversities

The genome of Entamoeba histolytica contains a number of large multi-gene families [14]. One such gene family encodes a group of AIG1 like proteins [23]. AIG1 protein family comprises of 29 members distributed in 3 clusters [23]. 18 of them are present near transposons, but whether their duplication and subsequent growth are encouraged by the proximity of transposons is required to be explored [23]. AIG1 proteins are associated with resistance to bacteria [40]. Another gene family encodes a group of leucine-rich-repeats (LRRs) containing proteins, homologous to bacterial fibronectin (BspA of Bacteroides forsythus) [41,23]. Lorenzi et al. identified 114 genes encoding for BspA-like proteins in the genome of E. histolytica strain HM1:IMSS. 41 of them are associated with transposable elements [23]. Proteins of the family contain conserved N-terminal domain. However, no classic membrane-targeting signal is present in the proteins [23]. Hence, it is tempting to speculate that conserved N-terminal domain of proteins might function as either an export signal or serve as a membrane-anchor domain or that export involves a non-classical transport mechanism, independent of the ER–Golgi pathway, similar to those that have been detected in yeast and mammalian cells [42]. At least one member of this family is expressed at the external surface of parasite [41]. Genome survey of E. invadens identified multiple copies of these leucine-rich-repeats (LRRs) containing genes and differential gene expression within gene families has also been reported [43]. However, it is quite unknown whether gene expression has been controlled in such a way that a single gene family is expressed at any one time, as observed in other parasites like Trypanosoma and Plasmodium [14].

Entamoeba also encodes a large number of Rab GTPase (like another protozoan parasite T. vaginalis), involved in vesicular trafficking in the cell [44,45]. A total of 102 Rab GTPase distributed in 16 subfamilies have been annotated in genome of E. histolytica [44,45]. Majority of them showed moderate similarity to Rab from other organisms, while only 22 amoebic Rab proteins including EhRab1, EhRab2, EhRab5, EhRab7, EhRab8, EhRab11, and EhRab21 showed significant similarity to Rab from other organisms [44]. E. invadens has over 100 Rab genes similar to E. histolytica [45]. A comparison of Rab GTPase from E. histolytica and E. invadens revealed that most Rab subfamilies are conserved among these two Entamoeba species [45]. This indicates that Rab GTPase-controlled vesicular trafficking machinery is well conserved among them and expansion of the gene family largely occurred before the divergence of these two species [45]. Rab GTPases have been involved in the regulation of cysteine protease secretion and transport [46,47]. E. histolytica differentially expressed their RabB protein (EhRabB) during phagocytosis of target cells, suggesting the potential role of EhRabB protein in phagocytosis process [48]. EhRabB protein has been mutated experimentally at 118 amino acid position and thus the resulted protein (RabBN118I) was unable to bind guanine nucleotide and became constitutively inactive [49]. Over-expression of such mutated RabB protein within E. histolytica trophozoites resulted in a significant reduction of parasite phagocytosis, cytopathic activity and ability to produce liver abscess in hamster [49]. Hence, Rab-regulated vesicular trafficking is important for parasite biology and pathogenesis. Gene families encoding heavy (hgl) and light (lgl) chain subunits of virulence determinant Gal/GalNAc lectin present in multiple Entamoeba species, but genes for intermediate chain subunit (igl) are only detected in E. histolytica and E. dispar [24]. Bioinformatics comparison among members of this gene family from E. histolytica and E. dispar identified the evidence of gene conversion within the lineages, which may play an important role in molecular evolution of these parasites [50]. Cysteine protease-5, the key virulence factor of E. histolytica is present as a pseudogene in E. dispar [14]. Over-expression of specific cysteine protease genes (ehcp-b8, ehcp-b9 and ehcp-c13) within parasite cells also confers pathogenicity to non-pathogenic E. histolytica clone A1 [51]. Southern blot analysis indicates that the surface proteins of E. histolytica are either not present or highly divergent in E. dispar [14].

6. Genetic diversity and population structure

Since E. histolytica genome does not appear to contain any microsatellite-like elements, measurement of genetic diversity and estimation of population structures greatly rely on other genetic markers like Serine Rich E. histolytica protein (SREHP) gene and chitinase [14]. SREHP is an immune dominant surface antigen, involved in phagocytosis of apoptotic host cells to prevent inflammatory responses by host [52] whereas chitinase is only expressed during encystations of amoeba [53]. Both genes contain tandem repeats which showed high degree of inter-isolate diversity based on their repeat types and arrangement patterns [2,3,54]. However, SREHP gene showed comparatively high degree of polymorphism than chitinase [3]. Since SREHP is highly immunogenic, such high genetic diversity within SREHP gene may suggest that it has a biological role like immune evasion [35]. However, PCR amplification of SREHP gene often produces multiple and mixed PCR bands from a single strain due to allelic variation [18]. Direct sequencing of such mixed PCR products (without cloning of PCR product into a vector prior to sequencing) gives rise to a chromatogram showing multiple peaks at a single nucleotide position. Multiple variations of a single sequence can be obtained from the analysis of such a sequence and this can be misinterpreted as genetic diversity. tRNA linked STR loci of E. histolytica has proved to be a useful population genetic marker and
has been used to identify the parasite genotypes associated with different disease outcomes [56]. Studies of genetic diversity based on 6 tRNA linked STR loci (i.e. D-A, STC-A-D, N-K2, R-R, A-L and S-Q) have identified few parasite genotypes associated with disease outcomes [8,13,57,12,58]. For example- SRR of R-R locus was associated with asymptomatic outcome, while 10RR was associated with symptomatic outcome [59]. J1DA and VEN2DA of D-A locus were associated with asymptomatic and symptomatic outcomes respectively [60]. Even though tRNA linked STR loci showed few associations with disease outcomes, they are actually surrogate marker and their variations are not at all directly linked to parasite virulence [59]. Moreover, these loci are frequently mutated to form new genotypes and hence any significant association of parasite genotype with disease outcome would be lost over time [18].

However, patterns of polymorphism within these repetitive DNA sometimes reflect the population structure of parasite [14]. For example, in Japan, diversity among parasite population infecting homosexual men was high, while diversity was much more limited among parasite infecting residents of institution [2]. Similarly, low diversity among parasite population infecting residents of institution was seen in the Philippines, where clear population structure was observed within and between locations [54]. In South Africa, genotypes clustered within households but showed extensive diversity among different households [61]. Recently, Zermeno et al. have proposed the worldwide genealogy and population structure of E. histolytica based on two tRNA linked STR loci (i.e. D-A and N-K2) [60]. Majority of these genotypes were found to be exclusive for a particular country. Only few were shared by isolates from different countries. For example- 18NK, 17NK, 10NK, and 11NK of N-K2 locus and 5DA and 6DA of D-A locus were the only genotypes distributed in many regions. Among them, 18NK and 6DA, corresponding to the genotype of E. histolytica strain HM1:IMSS were the most abundant and widely distributed in many countries like Mexico, Bangladesh, Japan, China and the USA. However, genealogies based on these two individual loci (i.e. D-A and N-K2) suggested that there were no parasite lineages related with a particular geographic region. Moreover, concatenated analysis of two tRNA linked STR loci (i.e. D-A and N-K2) revealed the possibility of genetic recombination among the population studied [60]. Genetic organization of E. histolytica population from stool and liver abscess samples of same patients were also studied [34]. The study revealed that E. histolytica population from stool and liver abscess samples were genetically distinct [34]. However, few opposite but interesting scenarios have also been reported. E. histolytica population isolated from amoebic liver abscess (ALA) patients was genetically identical with those isolated from asymptomatic patients [57]. This finding was further supported by recent STR loci based genotyping study of E. histolytica from India. E. histolytica isolates remaining asymptomatic are genetically closer to those causing liver abscess rather than the diarrheal isolates (Fig. 1) [12]. Repetitive DNA markers appear to be stable enough to link closely related parasites recently transmitted among members of a household, an institution or recent sexual partners [14]. However, extensive population diversity in limited geographic regions and frequent occurrences of novel genotypes limit the efficiency of repetitive loci to probe large scale, long term population structure of E. histolytica [14]. SNP (single nucleotide polymorphism) markers may be preferable in these situations.

SNPs within non-repetitive loci arising under neutral, positive and negative pressure are genetically stable and inherited by their descendants [11]. SNP analysis could be a successful strategy to identify the potential virulence marker of parasite linked to infection outcome [11,62]. Comparison between genome sequenc es of various E. histolytica strains deposited onto AmoebDB database version 4.1 [25, www.amoebadb.org] have identified a total of 2613 genes, which contain intra-species SNPs within them. Most of the proteins encoded by these genes are hypothetical in nature, while the functions of some genes are known. Few of such genes with known and hypothetical functions are listed in Table 1. A large number of SNPs have been identified in serine threonine isoleucine rich protein (EHI_073630), gene for Gal/Gal NAc lectin lgt2 (EHI_065330), heat shock protein70 (EHI_159140), pyrosine kinase (EHI_124500), gene for AIG1 family protein (EHI_144270), gene for Rab family GTPase (EHI_059670), etc. Gal/Gal NAc lectin is a surface antigen of Entamoeba and involved in parasite adhesion with intestinal epithelium [50]. AIG1 proteins are associated with resistance to bacteria [40]. Rab GTPases are involved in vesicular trafficking machinery of parasite [44,45]. However, further investigation is required to

Fig. 1. E. histolytica isolates remaining asymptomatic are genetically closer to those causing liver abscess: depicted by (A) phylogenetic tree, (B) graphical representation. Phylogeny was based on tRNA linked N-K2 (STR) locus. The sequences of all (a total of 22) representative STR patterns from N-K2 locus, obtained from the genetic analysis of 51 study isolates were aligned using ClustalW multiple alignment program of MEGA Version 4 software. Phylogenetic tree was constructed from the alignment through “Generalized Time Reversal (GTR) + gamma” substitution model of SeaView Graphical Interface Version 4 software using a maximum likelihood matrix algorithm. One distinct “D” group, one distinct “LA” group and one mixed “AS + LA” group can be assigned. ‘D’ group contains STR patterns found exclusively in diarrheal outcome, ‘LA’ group contains STR patterns found only in liver abscess outcome, ‘AS + LA’ group contains STR patterns exclusive for asymptomatic (AS) and liver abscess (LA) outcome.
determine the precise function of hypothetical proteins listed in Table 1. Homologs for some of these genes are also found in AmoebaDB database. Few genes of *E. histolytica* and their homologs are listed in Table 2. High degree of inter-species genetic variability is also observed among genes of *E. histolytica* and their homologs. A total of 326 SNPs have been identified within *E. dispar* hsp70 gene (EDI_012650) in comparison to that of *E. histolytica* (EHI_159140). Similarly, 520 inter-species SNPs have been detected in the homologous gene of lgl2 (EDI_244250) present in *E. dispar* SAW760 strain. Homologous gene for AIG1 family protein (EDI_001050) contains 144 inter-species SNPs. A total of 103 SNPs were identified in the homologous gene for actinin like protein (EDI_207850) present in *E. dispar* SAW760 strain. Homologous gene for elongation factor alpha 1 (EDI_134610) also contains a total of 90 SNPs. A total of 254 SNPs have also been identified in the homologous gene for inositol polyphosphate-5-phosphatase (EDI_159070). Another important virulence factor of *E. histolytica* is lysine and glutamic acid rich protein (KERP1). KERP1 is a surface-associated protein of

| Table 1 | Genes of *E. histolytica*, contain intra-species single nucleotide polymorphisms (SNPs). |
| --- | --- |
| AmoebaDB ID | Protein product for this gene | Total SNPs | Non-synonymous SNPs | Synonymous SNPs | Non-sense SNPs | Non-coding SNPs | Non-synonymous SNP/synonymous SNP ratio | SNPs per kb (CDS) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| EHI_073630 | Serine threonine isoleucine rich protein, putative | 70 | 46 | 24 | 0 | 0 | 1.92 | 4.6 |
| EHI_065330 | Gal/Gal NAc lectin lgl2 | 27 | 13 | 14 | 0 | 0 | 0.93 | 8.14 |
| EHI_159140 | Heat shock protein70, putative | 14 | 12 | 2 | 0 | 0 | 6 | 6.94 |
| EHI_006980 | Gal/Gal NAc lectin lgl1 | 13 | 9 | 4 | 0 | 0 | 2.25 | 3.89 |
| EHI_124500 | Tyrosine kinase, putative | 13 | 9 | 4 | 0 | 0 | 2.25 | 1.68 |
| EHI_144270 | Heat shock protein70, putative | 11 | 6 | 5 | 0 | 0 | 1.2 | 12.75 |
| EHI_164440 | Actinin like protein, putative | 9 | 0 | 9 | 0 | 0 | 0 | 5.74 |
| EHI_135220 | Phospholipid transporting p-type ATPase, putative | 9 | 3 | 0 | 0 | 0 | 0.5 | 3.05 |
| EHI_023050 | Protein kinase domain containing protein | 9 | 6 | 3 | 0 | 0 | 2 | 2.55 |
| EHI_05690 | Galactose inhibitable lectin 35 kDa subunit precursor | 8 | 4 | 4 | 0 | 0 | 1 | 8.57 |
| EHI_011210 | Elongation factor alpha 1 | 8 | 0 | 8 | 0 | 0 | 0 | 5.99 |
| EHI_139430 | Leucine rich repeat protein BspA family | 8 | 3 | 5 | 0 | 0 | 0.6 | 3.96 |
| EHI_023430 | Glycosyl hydrolase family 31 protein | 8 | 4 | 4 | 0 | 0 | 1 | 3.06 |
| EHI_042370 | Galactose specific adhesin 170 kDa subunit, putative | 8 | 4 | 2 | 0 | 0 | 3 | 2.05 |
| EHI_013980 | Phosphatidyl linositol 3 kinase, putative | 8 | 3 | 5 | 0 | 0 | 0.6 | 1.86 |
| EHI_119600 | Ubiquitin carboxyl terminal domain containing protein | 7 | 0 | 7 | 0 | 0 | 0 | 1.8 |
| EHI_059670 | Rab family GTPyase | 6 | 2 | 4 | 0 | 0 | 0.5 | 2.66 |
| EHI_160880 | Inositol polyphosphate 5 phosphatase, putative | 6 | 5 | 1 | 0 | 0 | 5 | 2.06 |
| EHI_012270 | Gal/Gal NAc lectin heavy subunit | 6 | 4 | 2 | 0 | 0 | 2 | 1.55 |
| EHI_045170 | US SnRNP specific 200 kDa protein, putative | 6 | 4 | 2 | 0 | 0 | 2 | 1.11 |
| EHI_164520 | Iron sulfur flavoprotein pseudogene | 5 | 4 | 1 | 0 | 0 | 4 | 11.36 |
| EHI_001160 | Plasma membrane calcium transporting ATPase, putative | 5 | 1 | 4 | 0 | 0 | 0.25 | 4.42 |
| EHI_000403 | Rap/Ran GTPyase activating protein, putative | 5 | 1 | 4 | 0 | 0 | 0.25 | 2.92 |
| EHI_188390 | Long chain fatty acid CoA ligase, putative | 5 | 1 | 4 | 0 | 0 | 0.25 | 2.57 |
| EHI_190880 | Thioredoxin domain containing protein 2, putative | 5 | 4 | 1 | 0 | 0 | 4 | 2.42 |
| EHI_061870 | Hypothetical protein | 22 | 13 | 9 | 0 | 0 | 1.44 | 3.4 |
| EHI_033550 | Hypothetical protein | 17 | 11 | 6 | 0 | 0 | 1.83 | 6.93 |
| EHI_023320 | Hypothetical protein | 17 | 12 | 5 | 0 | 0 | 2.4 | 5.51 |
| EHI_072500 | Hypothetical protein | 15 | 3 | 12 | 0 | 0 | 0.25 | 8.93 |
| EHI_013060 | Hypothetical protein | 15 | 10 | 5 | 0 | 0 | 2 | 3.11 |
| EHI_077250 | Hypothetical protein | 14 | 8 | 6 | 0 | 0 | 1.33 | 5.03 |
| EHI_018390 | Hypothetical protein | 13 | 8 | 5 | 0 | 0 | 1.6 | 4.73 |
| EHI_121060 | Hypothetical protein | 12 | 10 | 2 | 0 | 0 | 5 | 5.75 |
| EHI_059870 | Hypothetical protein | 12 | 8 | 4 | 0 | 0 | 2 | 4.54 |
| EHI_172000 | Hypothetical protein | 12 | 3 | 9 | 0 | 0 | 0.33 | 3.57 |
| EHI_159660 | Hypothetical protein | 12 | 5 | 7 | 0 | 0 | 0.71 | 2.28 |
| EHI_174540 | Hypothetical protein | 11 | 7 | 4 | 0 | 0 | 3 | 2.65 |
| EHI_196760 | Hypothetical protein | 10 | 9 | 1 | 0 | 0 | 9 | 8.71 |
| EHI_174560 | Hypothetical protein | 10 | 8 | 2 | 0 | 0 | 4 | 7.38 |
| EHI_111770 | Hypothetical protein | 10 | 9 | 1 | 0 | 0 | 9 | 5.45 |
| EHI_006990 | Hypothetical protein | 9 | 6 | 3 | 0 | 0 | 2 | 7.27 |
| EHI_025310 | Hypothetical protein | 9 | 4 | 5 | 0 | 0 | 0.8 | 4.34 |
| EHI_077750 | Hypothetical protein | 9 | 7 | 2 | 0 | 0 | 3.5 | 3.64 |
| EHI_103400 | Hypothetical protein | 9 | 0 | 3 | 0 | 6 | 0 | 3.62 |
| EHI_114110 | Hypothetical protein | 9 | 5 | 4 | 0 | 0 | 1.25 | 2.3 |
| EHI_016900 | Hypothetical protein | 8 | 5 | 3 | 0 | 0 | 1.67 | 6.79 |
| EHI_004180 | Hypothetical protein | 8 | 2 | 6 | 0 | 0 | 0.33 | 1.52 |
| EHI_119790 | Hypothetical protein | 7 | 3 | 4 | 0 | 0 | 0.75 | 21.28 |
| EHI_145460 | Hypothetical protein | 7 | 1 | 6 | 0 | 0 | 0.17 | 13.75 |
| EHI_106320 | Hypothetical protein | 7 | 4 | 3 | 0 | 0 | 1.33 | 11.69 |
| EHI_107040 | Hypothetical protein | 7 | 6 | 1 | 0 | 0 | 6 | 7.39 |
| EHI_144390 | Hypothetical protein | 7 | 6 | 1 | 0 | 0 | 6 | 7.39 |
| EHI_017780 | Hypothetical protein | 7 | 2 | 5 | 0 | 0 | 0.4 | 6.56 |
E. histolytica and has been shown to be involved in the parasite adherence to human enterocytes. It is also an important virulence factor of *E. histolytica* and has been shown to be involved in the parasite protein product for this gene. A total of 10 SNPs were identified within *kerp1* gene was present among *E. histolytica* species [as per AmoebaDB database (www.AmoebaDB.org)].

### Table 2

| E. histolytica HM1:IMSS | E. dispar SAW760 | E. invadens | E. moshkovskii Laredo |
|-------------------------|------------------|-------------|------------------------|
| **AmoebaDB ID**         | **Protein product for this gene** | **AmoebaDB ID** | **Protein product for this gene** |
| **AmoebaDB ID**         | **Protein product for this gene** | **AmoebaDB ID** | **Protein product for this gene** |
| EHI_073630              | Serine/threonine isoleucine rich protein, putative | EDI_083900 | Hypothetical protein |
| EHI_065330              | Gal/Nac lectin Ig2 | EDI_244250 | Furin repeat containing protein, putative |
| EHI_159140              | Heat shock protein 70, putative | EDI_012650 | Heat shock protein 70 kDa, putative |
| EHI_006980              | Gal/Nac lectin Ig1 | EDI_244250 | Furin repeat containing protein, putative |
| EHI_124500              | Tyrosine kinase, putative | EDI_004150 | Protein serine/threonine kinase HT1, putative |
| EHI_161490              | DNA polymerase, putative | EDI_056410 | Hypothetical protein, conserved |
| EHI_144270              | AIGI family protein | EDI_001050 | Hypothetical protein, conserved |
| EHI_164440              | Actin like protein, putative | EDI_207850 | Grainin, putative |
| EHI_135220              | Phospholipid transporting p-type ATPase, putative | EDI_018000 | Hypothetical protein, conserved |
| EHI_023050              | Protein kinase domain containing protein | EDI_123370 | Serine–threonine protein kinase, putative |
| EHI_035690              | Galactose-inhibitable lectin 35 kDa subunit precursor | EDI_023210 | Galactose-inhibitable lectin 35 kDa subunit precursor, putative |
| EHI_011210              | Elongation factor alpha 1 | EDI_134610 | Elongation factor 1-alpha |
| EHI_139430              | Leucine rich repeat protein BspA family | EDI_284090 | Hypothetical protein, conserved |
| EHI_023430              | Glycosyl hydrolyase family | EDI_137800 | Neutral alpha-glucosidase AB precursor, putative |
| EHI_042370              | Galactose specific adhesin 170 kDa subunit, putative | EDI_213670 | 170 kDa surface lectin precursor, putative |
| EHI_013980              | Phosphatidyl inositol 3 kinase, putative | EDI_147070 | Phosphatidylinositol 3-kinase catalytic subunit gamma, putative |
| EHI_119600              | Ubiquitin carboxyl terminal hydrolyase domain containing protein | EDI_023410 | Ubiquitin specific protease, putative |
| EHI_059670              | Rab family GTPase pseudogene | EDI_156940 | Trichohyalin, putative |
| EHI_160860              | Inositol polyphosphate 5 phosphatase, putative | EDI_159070 | Type II inositol-1,4,5-trisphosphate 5-phosphatase precursor, putative |
| EHI_012270              | Gal/Nac lectin heavy subunit | EDI_213670 | 170 kDa surface lectin precursor, putative |
| EHI_045170              | US SNBNC specific 200 kDa protein, putative | EDI_076220 | US small nuclear ribonucleoprotein 200 kDa helicase, putative |
| EHI_164520              | Iron sulfur flavoprotein pseudogene | EDI_064980 | Hypothetical protein, conserved |
| EHI_001160              | Plasma membrane calcium transporting ATPase, putative | EDI_013570 | Plasma membrane calcium-transporting ATPase, putative |
| EHI_000430              | Rap/Ran GTPase activating protein, putative | EDI_026850 | Rap GTPase-activating protein, putative |
| EHI_188590              | Long chain fatty acid CoA ligase, putative | EDI_093250 | Long-chain-fatty-acid—CoA ligase, putative |
| EHI_190880              | Thioredoxin domain containing protein 2, putative | EDI_197960 | Hypothetical protein |

* Homolog of the corresponding gene is not found in the particular Entamoeba species [as per AmoebaDB database (www.AmoebaDB.org)].
virulence determinant of *E. histolytica*. Studies of comparative genomic hybridization of *E. histolytica* and *E. dispar* strains suggested relatively low genomic diversity among *E. histolytica* [10]. A recent study by Weedall et al. has also identified a low level of single nucleotide diversity within *E. histolytica* populations [66]. Sequence analysis of defined regions also suggests similar observations [11,67]. Such low level of genetic diversity suggests a relatively recent common ancestor for *E. histolytica* [14]. However, this observation was quite incongruous with a recent report by Gilchrist et al. [62]. Multilocus sequence typing of *E. histolytica* clinical isolates identified extensive population diversity, suggesting that the genotypes of individual parasites do not contain consistent phylogenetic signals. They have blamed genetic recombination events for such a result, since it can break down the linkage between target loci and assist to form loci with different genealogies [62]. Hence, an important question regarding the population structure of *Entamoeba* is whether the parasite populations are predominantly clonal or sexual.

Sexual reproduction can help parasite to improve the fitness of their progeny [68]. Parasitic protists are continuously exposed to exogenous environmental factors and host immune pressure, which can alter the chemical structure and stability of their genome [68]. Parasites should repair structural alteration in their genome, since it can lead to mutations, deletions, insertions, translocation and loss of essential genetic information [68]. Parasites remove their DNA damage by recombinational DNA repair mechanism and this allows greater survival of offspring with undamaged DNA [68]. It is also an important mechanism to generate genetic diversity used by parasites to evade host immune response [68]. This particular feature of parasite is quite important, since sexual reproduction can exchange genes, responsible for drug resistance and parasite virulence. This could generate selectively advantageous genotypes that can spread very rapidly through host population [14]. Sexual reproduction can also help in the removal of deleterious genes. Current deleterious mutations brought together by sexual reproduction create unfit individuals that are eliminated from the population [68]. The genome of *E. histolytica* contains meiotic genes like SPO11, DMC1, and MND1 and many homologous recombination (HR) specific genes like MLH1, MSH2, RAD21 and RAD51 [22,69,68]. Moreover, ploidy changes and unscheduled gene amplification, which indicate the possibility of recombination have also been reported in *Entamoeba* [68]. *E. histolytica* contain a large number of retrotransposons in its genome, which also indicates their ability to reproduce by sexual means [68]. Organisms which reproduce solely by asexual means would eventually lose these retrotransposons from their genome [68]. However, Singh et al. recently provide the first direct demonstration of HR in *Entamoeba* using a construct with inverted repeats, which upon recombination results in sequence inversion. Increased rate of genetic recombination has been reported in *Entamoeba* under stress conditions and during encystation process [68]. Stage inter-conversion between cyst and trophozoite is crucial for disease transmission and pathogenesis in *E. histolytica* [68]. In addition to this, few indirect evidences of genetic recombination have also been identified in *Entamoeba* through population genetic studies. Complete genome sequencing of 10 axenic *E. histolytica* cell lines has identified pattern of polymorphism, indicates that recombination has occurred in the history of the population studied [66]. Concatenated genealogy based on repetitive loci (i.e. D-A and N-K2) also revealed the possibility of genetic recombination among *E. histolytica* population [60]. Bioinformatics comparison of Gal/GalNAc lectin among *E. histolytica* and its non-pathogenic sibling *E. dispar* also identified the evidence of gene conversion within the lineages [50].

Transposable elements constitute a significant portion of *E. histolytica* genome and they can affect the expression of adjacent genes [37]. Phenotypic characteristic of this parasite is also influenced by their genomic location [37]. Variability in genomic distribution of SINE1 and SINE2 among *E. histolytica* clinical isolates has been recently studied by Kumari et al. [70]. Several loci with extensive polymorphism of SINE occupancy among *E. histolytica* strains have been identified [70].

7. Conclusion
Queries related to evolution and population structure of *E. histolytica* still remains to be investigated. One of the concerning issue is whether *E. histolytica* population is sexual or clonal. Circumstantial evidence suggested that *Entamoeba* might engage in genetic recombination at some stage in their life-cycle. However, further detailed investigations with *Entamoeba* and other early branching protists are required to understand the origin of their sexual reproduction and to determine the variety of mechanisms by which these organisms exchange their DNA. Another major question that arises is whether *E. histolytica* population from ALA patients is genetically closer to that of asymptomatic individuals. If they are close (few studies suggested this), then individuals with persistent asymptomatic *E. histolytica* infection may be under high risk of developing ALA in the future. Prompt preventive measures should be undertaken for such individuals. Advanced whole genome sequencing of *E. histolytica* clinical isolates can be helpful to address this question.

Acknowledgments
This study was supported by grant from the Okayama University Program of Founding Research Centre for Emerging and Re-emerging Infectious Disease (OUP 2–5), Ministry of Education, Culture, Sports, Science and Technology of Japan. The authors would like to thank Mrs. Debarati Ganguly for her immense help regarding proof-reading of this manuscript.

References
[1] W. H. O., PAHO/UNESCO. A consultation with experts on amebiasis. Epidemiol Bull 1997;18:13–4.
[2] Haghjighi A, Kobayashi S, Takeuchi T, Masuda G, Nozaki T. Remarkable genetic polymorphism among *Entamoeba histolytica* isolates from a limited geographic area. J Clin Microbiol 2002;40:4081–90.
[3] Haghjighi A, Kobayashi S, Takeuchi T, Thammapalerd N, Nozaki T. Geographic diversity among genotypes of *Entamoeba histolytica* field isolates. J Clin Microbiol 2003;41:3748–56.
[4] Walsh JA. Problems in recognition and diagnosis of amebiasis: estimation of the global magnitude of morbidity and mortality. Rev Infect Dis 1986;8:228–38.
[5] Stanley Jr SL. Amebiasis. Lancet 2003;361:1023–4.
[6] Mehmet T, Petri J Jr. WA. Laboratory diagnosis of amebiasis. Clin Microbiol Rev 2003;16(4):713–29.
[7] D'Souza MM, Clark CG. A redescription of *Entamoeba histolytica*. Schaudinn, 1903 (Emended Walker, 1911) separating it from *Entamoeba dispar* Brunzt, 1935. J Eukaryot Microbiol 1993;40:346–4.
[8] Ali IBM, Mondal U, Roy S, Haque R, Petri JR Jr, et al. Evidence for a link between parasite genotype and outcome of infection with *Entamoeba histolytica*. J Clin Microbiol 2007;45:285–9.
[9] Duggal PR, Haque R, Roy S, Mondal D, Sack RB, et al. Influence of human leucocyte antigen class II alleles on susceptibility to *Entamoeba histolytica* infection in Bangladeshi children. J Infect Dis 2004;189:520–6.
[10] Shah PH, MacFarlane RC, Bhattacharya D, Matteo JC, Demeter J, et al. Comparative genomic hybridizations of *Entamoeba* strains reveal unique genetic fingerprints that correlate with virulence. Eukaryot Cell 2005;4:504–15.
[11] Bhattacharya D, Haque R, Singh U. Coding and noncoding genomic regions of *Entamoeba histolytica* have significantly different rates of sequence polymorphisms: implications for epidemiological studies. J Clin Microbiol 2005;43:4815–25.
[12] Das K, Mukherjee AK, Chowdhury P, Sehgal R, Bhattacharya MK, et al. Multilocus sequence typing system (MLST) reveals a significant association of *Entamoeba histolytica* genetic patterns with disease outcome. Parasitol Int 2014;63:308–14.
[13] Escure-dela Cadiz A, Kobayashi S, Takeuchi T, Tachibana H, Nozaki T. Identification of an avirulent *Entamoeba histolytica* strain with unique rDNA-linked short tandem repeat markers. Parasitol Int 2010;59:75–81.
[14] Weedall GD, Hall N. Evolutionary genomics of *Entamoeba*. Resmice, 162: 2011 637–45.
[15] Clark CG, Raffishian F, Tawari B, Windsor J, Twigg-Hesler A, et al. New insights into the phylogeny of *Entamoeba* species provided by analysis of four new small-subunit rRNA genes. J Evol Biol 2006;19:941–51.
[16] Dolabella SS, Serrano-Luna J, Navarro-Garcia F, Cerritos R, Jimenez C, et al. Amebic liver abscess production by *Entamoeba dispar*. Ann Hepatol 2012;11(1):107–17.
[17] Shirnukawa C, Kabir M, Taniuchi M, Mondal D, Kobayashi S, et al. *Entamoeba moshkovskii* is associated with diarrhea in infants and causes diarrhea and colitis in mice. J Infect Des 2012;206(5):744–51.
[18] Ali IBM, Clark CG, Petri JR Jr. Molecular epidemiology of amebiasis. Infect Genet Evol 2006;8(5):698–707.
Nakada-Tsukui K, Saito-Nakano Y, Nakada-Tsukui K, Saito-Nakano Y, Loftus BJ, Hall N, Amnedeo P, et al. New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information. PLoS Negl Trop Dis 2010;4:716.

Clark CG, Alsmark UC, Tazreiter M, Saito-Nakano Y, Ali IKM, Solaymani-Mohammadi S, Akhter J, Roy S, Gorrini C, et al. Tissue invasion by Entamoeba histolytica: the over expression of a mutated EhRabB protein produces a decrease of in vitro and in vivo virulence. Exp Parasitol 2013;133(3):339–45.

Weedall GD, Sherrington J, Paterson S, Hall N. Evidence of gene conversion in genes encoding the Gal/GalNac lectin complex of Entamoeba. PLoS Negl Trop Dis 2011;5(6):1209.

Matthiesen J, Bär AK, Bartels AK, Marien D, Ofsir S, et al. Overexpression of specific cysteine peptidases confers pathogenicity to a nonpathogenic Entamoeba histolytica clone. Mol Biol 2013;4(2):4.

Teixeira JE, Huston CD. Participation of the serine-rich Entamoeba histolytica protein in amebic phagocytosis of apoptotic host cells. Infect Immun 2008;76(3):959–66.

de la Vega H, Specht CA, Semino CE, Robbins PW, Eichinger D, et al. Cloning and expression of chitinases of Entamoeba. Mol Biochem Parasitol 1997;85:139–47.

Rivera WL, Santos SR, Kanbara H. Prevalence and genetic diversity of Entamoeba histolytica in an institution for the mentally retarded in the Philippines. Parasitol Res 2006;98:105–10.

Zhang T, Stanley JS. DNA vaccination with the serine-rich Entamoeba histolytica protein (SREHP) prevents amebic liver abscess in rodent models of disease. Vaccine 1999;18:868–74.

Ali IKM, Zaki M, Clark CG. Use of PCR amplification of tRNA gene-linked short tandem repeats for genotyping Entamoeba histolytica. J Clin Microbiol 2005;43:5842–7.

Feng M, Bai Y, Yang B, Fu Y, Xin X, et al. Unique short tandem repeat nucleotide sequences in Entamoeba histolytica isolates from China. Parasitol Res 2012;111:1137–42.

Jaiswal V, Ghosal U, Mittal B, Dhole TN, Ghoshal UC. Association between allelic variation due to short tandem repeats in tRNA gene of Entamoeba histolytica and clinical phenotypes of amoebiasis. Acta Trop 2014;133:1–7.

Ali IKM, Haque R, Alam F, Kabir M, Siddique A, et al. Evidence for a link between locus R-R sequence type and outcome of infection with Entamoeba histolytica. Clin Microbiol Infect 2012;18:235–7.

Zermeno V, Jimenez C, Mocan P, Valadez A, Valenzuela O, Racon E. Worldwide genealogy of Entamoeba histolytica: an overview to understand haplotypic distribution and infection outcome. Exp Parasitol 2013;137:243–52.

Zaki M, Reddy SG, Jackson TF, Ravdin JL, Clark CG. Genotyping of Entamoeba species in South Africa: diversity, stability, and transmission patterns within families. J Infect Dis 2003;187:1600–9.

Gilchrist CA, Ali IKM, Kabir M, Alam F, Scherbakova S, et al. A Multilocus Sequence Typing System (MLST) reveals a high level of diversity and a genetic component to Entamoeba histolytica virulence. BMC Microbiol 2012;12:151.

Seigneur M, Mounier J, Prevost MC, Guillen N. A lysine- and glutamic acid-rich protein (SREHP) prevents amebic liver abscess in rodent models of disease. Vaccine 1999;18:868–74.

Ali IKM, Zaki M, Clark CG. Use of PCR amplification of tRNA gene-linked short tandem repeats for genotyping Entamoeba histolytica. J Clin Microbiol 2005;43:5842–7.

Feng M, Bai Y, Yang B, Fu Y, Xin X, et al. Unique short tandem repeat nucleotide sequences in Entamoeba histolytica isolates from China. Parasitol Res 2012;111:1137–42.

Jaiswal V, Ghosal U, Mittal B, Dhole TN, Ghoshal UC. Association between allelic variation due to short tandem repeats in tRNA gene of Entamoeba histolytica and clinical phenotypes of amoebiasis. Acta Trop 2014;133:1–7.

Ali IKM, Haque R, Alam F, Kabir M, Siddique A, et al. Evidence for a link between locus R-R sequence type and outcome of infection with Entamoeba histolytica. Clin Microbiol Infect 2012;18:235–7.

Zermeno V, Jimenez C, Mocan P, Valadez A, Valenzuela O, Racon E. Worldwide genealogy of Entamoeba histolytica: an overview to understand haplotypic distribution and infection outcome. Exp Parasitol 2013;137:243–52.

Zaki M, Reddy SG, Jackson TF, Ravdin JL, Clark CG. Genotyping of Entamoeba species in South Africa: diversity, stability, and transmission patterns within families. J Infect Dis 2003;187:1600–9.

Gilchrist CA, Ali IKM, Kabir M, Alam F, Scherbakova S, et al. A Multilocus Sequence Typing System (MLST) reveals a high level of diversity and a genetic component to Entamoeba histolytica virulence. BMC Microbiol 2012;12:151.

Seigneur M, Mounier J, Prevost MC, Guillen N. A lysine- and glutamic acid-rich protein, KERP1, from Entamoeba histolytica binds to human enterocytes. Cell Microbiol 2005;7:569–79.

Santi-Rocca J, Weber C, Guigou G, Sismeiro O, Coppée JY, et al. The lysine- and glutamic acid-rich protein KERP1 plays a role in Entamoeba histolytica virulence. Cell Microbiol 2008;10:202.

Perdomo D, Baron B, Rojo-Domínguez A, Raynal B, England P, et al. The α-helical regions of KERP1 are important in Entamoeba histolytica adherence to human cells. Nat Sci Reports 2013;3:1171.

Weedall GD, Clark CG, Koldkjær P, Kay S, Bruchhaus I, et al. Genomic diversity of the human intestinal parasite Entamoeba histolytica. Genome Biol 2012;13:38.

Ghosal S, Pradhan M, Ramirez-Avila L, Descoteaux S, Sturm-Ramirez K, et al. Molecular epidemiology of Entamoeba spp.: evidence of a bottleneck (Demographic sweep) and transcontinental spread of diploid parasites. J Clin Microbiol 2000;38:3815–21.

Singh N, Bhattacharya A, Bhattacharya S. Homologous recombination occurs in Entamoeba and is enhanced during growth stress and stage conversion. PLoS One 2013;8(9):e74465.

Stanley Jr SL. The Entamoeba histolytica genome: something old, something new, something borrowed and sex too? Trends Parasitol 2005;21:451–3.

Kumar V, Iyer LR, Roy R, Bhargava V, Panda S, et al. Genomic distribution of SINEs in Entamoeba histolytica strains: implication for genotyping. BMC Genomics 2013;14:432.