 doi:10.20944/preprints202103.0070.v1

Illumination on the structure and characteristics of Entamoeba histolytica genome.

Musafer H. Al-Ardi

Al-Qadisiya General Directorate for education, Ministry of Education– Iraq, Mussafir78@yahoo.com.

Abstract: Entamoeba histolytica, like other Organismes, is characterized by diversity and heterogeneity in its genetic content, which is one of the most important reasons for survival, and the increase in susceptibility to infection.Non-condensation of chromosomes during the process of cell division and the ambiguity of the chromosomal ploidy makes predicting the exact chromosomal number difficult. Genes distributed across 14 chromosomes as well as many extra-chromosome elements. Most Genes composed of one axon only, with Introns in 25% of Genes. This genome is characterized by the presence of Polymorphic internal repeat regions, and several gene families, one of these large families encoding Transmembrane kinas, Cysteine protease (CP), SREHP protein, and others.

Keywords: Entamoeba histolytica ; Genome ; gene families ; Transposable elements.

Introduction

Quincks and Ross in 1893 discovered the cyst phase of Entamoeba, while Schaudinn in 1903 called it Entamoeba histolytica due to its ability to invasion tissues and distinguish it from E. coli1.

Clifford Dobell in 1919 mentioned the existence of several species of the amoeba genus, each of which had a cystic phase that had four nuclei2. Emile Brumpt in 1925 demonstrated the existence of two strains representing the same species3. Sargeaunt and Williams asserted that there could be two strains of this species, one capable of invading tissues and the other didn't2.In 1993 Diamond and Clark proved that the second strain is another species belonging to this genus, which was later called E. dispar 4.

In 2005, the complete genetic sequence of E.histolytica and the strain HM-1: IMSS was revealed, which opened up new horizons of research towards a broader and more comprehensive study of this parasite5.

Entamoeba spp. includes Several species that infect different species of hosts, E.muris that infect mice, E. bovis (cows), E. debliecki (goats and pigs), and other species6, addition to six different species found in the large intestine of humans (E.histolytica, E. dispar, E. hartmanni, E. polecki, E. moshkovskii, E. coli7.

In addition to the inability to induce disease, the non-wide spread, and the small size of the non-pathogenic species, there are many important differences between the two sections (invasive and non-invasive)8. The absence or difference of some genetic groups such as (SINE) in the non-pathogenic group which presence in E. histolytica, the presence and activity of some surface proteins such as cystine CP5, the activity of the amoebapore protein, all these variants are due to the genetic makeup differences among amoeba species genomes4.

Whole Genome

Entamoeba histolytica genome was the first precursor genome which sequences completed in 20059 using the Shotgun method which is the best method for sequencing. Its genome count estimated at 23.7 million base pairs compared with Plasmodium falciparum genome size (23 M bases), and the free-living amoeba Dictyostelium discoideum (34 M bases) 10.

The genome re-sequenced in 2010, with a genome size of 20.8 megabases (MB), distributed in 14-17 chromosomes. The difference occurs because of the removal of genes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0070.v1

with repeating regions and the deletion of genes less than 300 base pairs which there is no clear evidence of their action 11.

The complete genome distributed in an estimated number of chromosomes (about 14 chromosomes)12 as well as, many circular extra-chromosome elements that include a gene region of rDNA that can be replicated (Episome). E. histolytica genome does not have Microsatellites, so, measuring Genetic diversity and estimation of gene community composition depends on other genetic markers13 .

The genome of Entamoeba genus has 8300 genes, representing an average of 1260.9 base pairs (about 49.7%) of genome size. The shortest gene is 147 base pairs and the longest is 15,210 base pairs. An Introns occupy 24.4% of the expected genome size14, which mean that a quarter of E. histolytica genes contain introns, as well as 6% of the genes with double introns, third of these genes (about 31.8%), produce heterogenic proteins15 .

Most genes contain a single exon, however, 25% of them may have splicing. In general, the size of the genes is small due to the absence of introns, as these genes encode proteins up to 389 amino acids in length16 .

Among the genes, there are many polymorphic tandem repeat regions, these regions encoding the amoeba-rich serine (SREHP) and chitinase, as well as, many sites for short repeats of tRNA (tRNA-STR loci) that were used as genetic markers and to distinguish between the Genotypes patterns, which are associated with various clinical markers, also indicate high levels of diversity in the E. histolytica population12.

The size and number of genes in E. histolytica are proportional to the metabolic adaptations needed by the parasite, where the deletion or reduction of most cellular redox pathways in the mitochondria is observed with the presence of some enzymes similar to the oxidative enzymes in the eukaryotes16.

It is expected that a large portion of the genes was accidentally transferred from bacteria to the Amoeba genome, and there are shreds of evidence that indicate the functions of these genes in the amoeba metabolism17. Also, the genome encodes a large number of receptors such as kinases receptor and a variety of genes families that important in the parasite's virulence including Cysteine and Metallo-proteinases16.

Ploidy and chromosomes

Our knowledge about the structure and composition of the chromosomes in the amoeba is little because the chromosomes do not condense during the mitosis of cell division, therefore, a kind of ambiguity surrounds the chromosome group18. The huge variations between homologous chromosomes in different isolates makes predicting the accurate chromosome number is difficult15, as it cannot be confirmed that it is mono, diploid, or even quadruple with the presence of two or one groups of them in some studies19. This variation in the number and structure of chromosomes and the type of chromosomal ploidy even at the level of a single strain cell that growing in different conditions, whether in vivo or in vitro, maybe due to the ebb and flow of the subtelomeric repeats region as in the case of other Protista20. Interestingly, this region contains tRNA microarrays in E. histolytica15. The heterogeneity of chromosomal ploidy may also be a delusion that the parasites are multi-chromosomal21.

There is no reliable information about the size and nature of the centromeres, nor about the peripheral region, due to it does not exist or branched in a way that cannot be distinguished. despite the presence of genes encoding for histone proteins H1, H2A, H3, H4, chromatin surveyed with an electron microscope, it became evident the presence of nuclear particle-like structures consisting of a base protein bound with DNA, this protein differs from it in the other Eukaryotic cells13.

Extrachromosomes structures genes

Outside the chromatin, many rDNA molecules are circular in structures. These structures are important in the phenotype and this raises several questions, including whether the number of copies differs from it in the chromosome? Are these molecules

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2021 doi:10.20944/preprints202103.0070.v1

isolated in the same way as the isolation of the chromosome in the case of cell division?12.

Mukherjee et al. (2008) found that these episomes comprise about 10-20% of the total cellular DNA (genome), approximately 15% of the readable genomic sequences among E.histolytica genome belong to these molecules that include 200 copies occupying about 25 kilobases19. Lorenzi et al. (2010) suggested that the duplicated segments discovered along the genome represent some of these circular molecules22.

In addition to these molecules, there are many less widespread molecules of different sizes (5.12 and 50 kb), that exact their function4, nor their genetic sequences known10.

Various circular rDNA molecules in E. histolytica have been described, while all previous studies have focused on molecules carrying rRNA genes, of which there are approximately 200 copies encoded for small and large Ribosomal units and 5.8S rRNAs, but not encoded for 5S rRNA nor protein. The size of these Molecules are 24.5 kilobases4 .

When analyzing circular rDNA sequences in E. histolytica strains, two forms of arrangement are observed. In some strains, one rDNA unit is transcribed per cycle, while in other strains two units are arranged in opposite repeats. Also, these strains differ in the presence or absence of intergenic spacers (IGS). Where the upstream region is in two rDNA units strains are mismatched. These upstream sequences in the right direction of the single rDNA unit strains are present, while There are missing in the left direction23.

tRNA genes

The tRNA genes are organized as double and multiple arrays units that are separated from each other by rich repeats of thymine-adenine sequences. With evidence indicating their presence at the end of the chromosome and in the peripheral chromatin, giving them a function equivalent to the telomeres missing in E.histolytica. This suggests an effect of tRNA arrays in the structural organization of the nucleus24.

Most of the unique structural features that have been characterized in the E. histolytica genome are due to tRNA genes, as 10% of the readable sequences contain tRNA genes and these (with a few exceptions) are arranged in linear arrays10.

The number of copies of the tRNA gene is about 4,500, which is ten times more than the human genome. They are arranged in repeating arrays that make up 10% of the genome. 25 distinct arrays containing repeating units that encode 1-5 types of receptors for tRNA, Three arrays encode for 5S RNA, and one encodes for RNA which Later snRNA25.

Clark et al. (2007) confirmed that it is not possible to accurately guess the size of the arrays due to the convergence of the arrays that have been identified. In general, there are 25 distinct arrays with a unit size ranging from 500-1750 base pairs. The regular arrays of tRNAs observed in some cases represent more than one repetitive unit read in one direction and other cases both readings are observed in both directions10.

Intragenic regions among all the genetic arrays in E. histolytica contain complex structures of single tandem repeats (STRs). These repeats are variants in size (7-12 base pairs), although few reach more than 44 base pairs24. Some of these variances occur in the number of (STR) among the same array unit, but these variations are simple and do not appear when performing (PCR) for Inter-tRNA25. This variation is meaningful. When comparing different strains, this can be used as a method for genotyping in this organism24.

Gene families and their diversity

Efforts to know the complete genetic sequence of E.histolytica genome obtained amazing results, one of these is to distinguish some gene families, one of these large families encoding Transmembrane kinas (TMK) enzyme, which was previously found in high Eukaryotic cells, the presence of this protein in the plasma membrane reflects Participation in extra-cellular signals26, some members of this family affects phagocytosis Series of Oligonucleotide encodes this enzyme that gives the highest expression rate in invading tissues and exposure to inappropriate conditions, that is, it is an indicator of response to the environment14.

Another important gene family is a gene family encoding Cysteine protease (CP) with a total of 86 genes, as 50 genes encode to papain family, 22 genes encode to Metal-

 doi:10.20944/preprints202103.0070.v1

lo-protease (MP), 10 genes encode to Serine protease (SP), and 4genes encode to Aspartic protease (AP), although this group was not expressed in vivo27 .

SREHP protein, which is known as K2 membrane protein consists of the leading sequences at the amniotic end and an anchor basis of the hydrophobic sequences at the carboxylic end. The expression method of this group of genes vary according to the infection stage and the host type, so it is used to distinguish the isolated and growing strains28.

Gal / GalNAc lectin protein (260 KD) is a heterodimeric glycoprotein with a bi-sulfur bridge linking its two subunits. It is located on the Trophozoites plasma membrane. Two gene families are coded, one for the heavy unit (KD 170), and the other for the light (KD31-35). This gene group performs many functions, including attachment to the host cell surface and pathogenesis of the parasite29.

Lorenzi (2010) asserted that one of the large gene families encodes for the AIG1-like GTPases group. This gene family consists of 29 members distributed in three groups, 18 genes of which possess repetitive elements with no known function, but heterogeneity in the expression of these genes may be associated with virulence. Their products act as a heat shock protein22.

Other gene families include a family that encodes for the LRR protein9, and a family that encodes for Rab GTPases12.

Transposable elements (TES)

Transposable (jumping) elements are known as pieces of DNA that can be introduced into new locations of any chromosome, as well as they can produce duplicate copies of these elements. It is found in prokaryotic and eukaryotic cells. Discovered in eukaryotes by Barbara McClintock in the corn plant in 194030.

These elements play an important role in the biology of living organisms, as they can cause the mutation in the gene when inserted on it and can influence the genetic regulation if inserted in a place near the promoter in addition to being a basic material for the rearrangement of genes, which makes it an important indicator in the evolution of the genome31 .

E. histolytica, like other organisms, have two types of these elements (TES), long and short32, three independent gene families are long tandem repeat (LTR) called E. histolytica long interspersed nuclear elements (EhLINE), the others, three gene families are short tandem repeat (STR), non-independent, and called E. histolytica short interspersed nuclear elements (EhSINE). In E.dispar that have the same number of these large families just33.

LINEs have (4.8 kgb), while SINEs have (0.5 -0.7 kgb) in size, which make up 11.2% of the genome in E.histolytica33 . These elements are often found in intergenic regions, with thymine-rich sequences. 50 base pairs are found upstream of insertion site for a limited number of EhSINE1, those occupy different genome sites in E. histolytica and E. dispar, and vice versa, which may indicate the relationship of these elements to the virulence of the parasite and its ability to infect34 .

These transposable elements affect the gene expression of nearby or neighboring genes through various mechanisms, including alternative promotor processing, splicing, and others. SINEs are distinguished by being stable across generations, and attempts to insert them into the gene are rare, so genetic analyzes using SINEs are fairly accurate from RFLPs and Microsatellite sites35.

Conclusion:

Each organism has its own distinctive characteristics that reflect the image of its existence. Genes and the genome are the most expressive images of the activity and functions of the organism. E. histolytica shows a variety in the structure and characteristics of the genome, in terms of the number of chromosomes and the chromosome set, the number of genes that packed into these chromosomes, the structural details of these The genes, and the method by it these genes are used to keep the parasite survive and enable it to invade different hosts.Research manuscripts reporting large datasets that are deposited in a publicly available database should specify where the data have been depos-

ited and provide the relevant accession numbers. If the accession numbers have not yet been obtained at the time of submission, please state that they will be provided during review. They must be provided prior to publication.

Interventionary studies involving animals or humans, and other studies that require ethical approval, must list the authority that provided approval and the corresponding ethical approval code.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pinilla, E.; Consuelo, M. and Fernando, D. History of the Entamoeba histolytica protozoan. Rev. Méd. Chile .(2008).136(1): 118-124.
2. Marianne, L. Molecular diagnosis and characterization of two Intestinal Protozoa: Entamoeba histolytica & Giardia intestinalis . Karolinska Institutet , Stockholm. (2010).76 pp.
3. Dhaliwal, B. B. and Juyal, P. D . Parasitic Zoonoses . Springer, India. (2013). 157 pp.
4. Zaki, M. Characterisation of polymorphic DNA and its application to typing of Entamoeba histolytica and Entamoeba dispar . Ph.D. thesis , University of London . (2002). 278 pp.
5. Bhattacharya, S. and Bhattacharya, A. Amoebiasis and Entamoeba species: unexplored liaisons. Trop. Ga.; .(2013). 34(2):55–57.
6. Hooshyar, H. ; Rostamkhani, P. and Rezaeian, M. An annotated checklist of the Human and Animal Entamoeba) Amoebida: Endamoebidae) species- a review article. Iran J. parasitol. (2015)10(2):146-156.
7. Tanyukse, M. and Petri, W. A. Laboratory diagnosis of Amebiasis . Clin. Microbial. Rev .(2003). 16(4): 713–729.
8. Schuster, F.L. and Visvesvara, G.S. Amebae and ciliated protozoa as causal agents of waterborne Zoonotic disease. Vet. Parasitol. (2004). 126 (1-2):91–120.
9. Das, K. and Ganguly, S. Evolutionary genomics and population structure of Entamoeba histolytica .Compu. and Stru. Biotechn. J. 12 (2014): 26–33.
10. Clark, C.G. Alsmark, UC. ; Tazreiter, M. ; Saito-Nakano, Y. ; Ali, V. and Marion, S. Structure and Content of the Entamoeba histolytica Genome.In Advances in Parasitology. Vol. 65: R. Muller , R. and Tanabe, K. (2007). 451 pp.
11. López-Camarillo , C.; Zamorano, C.; E la Cruz, H.; Rosas, I. and Marchat, L. Genomics and proteomics approaches to understand virulence of Entamoeba histolytica. A Méndez-Vilas (Ed). (2011). 511 pp.
12. Weedall , G. D. and Hall, N. Evolutionary genomics of Entamoeba. Res. in Microbiol . (2011). 162(2-6): 637-645.
13. Black, S. J. and Seed, J. R. The pathogenic enteric protozoa:Giardia, Entamoeba, Cryptosporidium and Cyclospora . kluwer academic publishers. New York, (2004). 184 pp.
14. Clark, C.G. ; Johnson, p. j. and Adam, D. Anaerobic parasitic protozoa genomics and molecular biology .Caister academic press, UK, (2010). 562 pp.
15. Brendan , L. ; Anderson, I. ; Davies, R. ; Alsmark, U.C. ; Samuelson, J. ; Amedeo, P. ; Roncaglia, P. ; Berriman, M. ; Hirt, RP.; Mann, B.J. ; Nozaki, T. ; Suh, B. ; Pop, M. ; Duchene, M. and Ackers, J. The genome of the protist parasite Entamoeba histolytica. Natu. (2005). 43(3):865-868.
16. López-Camarillo , C.; Guillen, N .; Weber, C.; Orozco, E. and Marchat, L. A. Genomics approaches in the understanding of Entamoeba histolytica virulence and gene expression regulation . Afr. J. Bio. (2009). 8 (8):1363-1369.
17. Srivastava , S. Identification of strains of Entamoeba histolytica and .Entamoeba dispar from natural isolates . Ph.D. thesis ,school of life science, Jawaharlal Nehru University .India . (2005). 107 pp.
18. Chavez -Munguia , B. ; Tsutsumi , V. and Martenez-Palomo, A. Research brief Entamoeba histolytica: Ultrastructure of the chromosome and the mitotic spindle . Expe. Parasitol.; 114(2006): 235–239.
19. Mukherjee, C.; Clark, G. C. and Lohia, A. Entamoeba shows reversible variation in ploidy under different growth conditions and between life cycle phases . PLoS. Negl. Trop. Dis.; (2008). 2(8): 281.
20. Bagchi, A. Studies on structure and organisation of chromosomes in Entamoeba histolytica .Ph.D. thesis, School of Environmental Sciences , Jawaharlal Nehru University .India (2001).144 pp.
21. Ghosh, S. ; Frisardi, M. ; Ramirez-Avila, L. ; Descoteaux, S.; Sturm-Ramirez, K. ; Newton-Sanchez, A. ; Santos-Preciado, JI. ; Ganguly, C. ; Lohia, A.; Reed, S. and Samuelson, J. Molecular epidemiology of Entamoeba spp. evidence of a bottleneck (demographic sweep) and transcontinental spread of diploid parasites . J. Clin. Microbiol. (2000). 38(10): 3815–3821.
22. Lorenzi, H.A.; Puiu, D.; Miller, J.R.; Brinkac, L.M.; Amedeo, P.; Hall, N. and Caler, E.V. New assembly reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information. PLoS. Negl. Trop. Dis. (2010). 4(6): 716.
23. Bhattacharya, A. and Bhattacharya, P. J. Close sequence identity between Ribosomal DNA episomes of the nonpathogenic Entamoeba dispar and pathogenic Entamoeba histolytica. J. Bio.sci. (2002). 27 (3): 619–627.
24. Irmer , H.; Hennings, I.; Bruchhaus, I. and Tannich, E. tRNA Gene sequences are required for transcriptional silencing in Entamoeba histolytica. Euk. Cell. (2010). 9(2): 306–314.
25. Tawari, B.; Ali, M.; Scott, C.; Michael, A.; Quail, M. B.; Hall, N . and Clark, G. C. Patterns of Evolution in the unique tRNA Gene arrays of the genus Entamoeba. Mol. Biol. Evol. (2008). 25(1):187–198.

26. Ali, IK .; Solaymani-Mohammadi, S. ; Akhter, J. ; Roy, S. and Gorrini, C. Tissue Invasion by Entamoeba histolytica: Evidence of Genetic Selection and/or DNA Reorganization Events in Organ Tropism. PLoS. Neg. Trop. Dis. (2008). 2(4):219.
27. Tillack, M.; Biller, L.; Irmer, H.; Freitas, M.; Gomes, M. A.; Tannich, E. and Bruchhaus, I. The Entamoeba histolytica genome: primary structure and expression of proteolytic enzymes. BMC. Genom. (2007). 8(1):170 .
28. Zhiming, M . and Samuelson, J. A new gene family (ariel) encodes Asparagine-rich Entamoeba histolytica antigens, which resemble the Amebic vaccine candidate Serine-rich E. histolytica Protein. Inf. and Immu. (1998). 66(1): 353–355 .
29. Caron, M. and Seve, A. Lectins and pathology , Harwood academic publishers , British. 289 pp.
30. Pray, L. (2008). Transposons, or jumping genes: Not junk DNA? Nat. Edu. (2004). 1(1):32.
31. Griffiths, A. ; Miller, J. ; Suzuki, D. An Introduction to Genetic Analysis. 7th edition. New York. (2000). pp:157.
32. Dewannieux, M. and Heidmann, T. LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling. Cytog. Gen. Res. (2005). 110(1-4):35-48.
33. Mandal, P. K.; Bagchi, A. and Bhattacharya, S. An Entamoeba histolytica LINE/SINE pair inserts at common target sites cleaved by the restriction Enzyme-like LINE-encoded endonuclease. Ame. Soc. Microbiol. (2004). 3(1):170–179.
34. Kumari, V.; Sharma, R.; Yadav, V. P. ; Gupta, K.; Bhattacharya, Alok. and Bhattacharya, S. Differential distribution of a SINE element in the Entamoeba histolytica and Entamoeba dispar genomes: Role of the LINE-encoded endonuclease. BMC. Genom. (2011). 12(1):267.
35. Kumari, V.; Iyer, L. R.; Roy, R.; Bhargava,V.; Panda , S.; Paul, J. and Verweij, J. Genomic distribution of SINEs in Entamoeba histolytica strains: implication for genotyping . BMC. Genom. (2013). 14(1):432.