Genetic characterization of the *Entamoeba moshkovskii* population based on different potential genetic markers

Sanjib K. Sardar[1], Ajanta Ghosal[1], Tapas Haldar[1], Akash Prasad[1], Sweety Mal[1], Yumiko Saito-Nakano[2], Seiki Kobayashi[2], Shanta Dutta[3], Tomoyoshi Nozaki[4], Sandipan Ganguly[1]*

1. Division of Parasitology, ICMR-National Institute of Cholera and Enteric Diseases (ICMR-NICED), Kolkata, India

2. Department of Parasitology, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8640, Japan

3. Division of Bacteriology, ICMR-National Institute of Cholera and Enteric Diseases (ICMR-NICED), Kolkata, India.

4. Department of Biomedical Chemistry, School of International Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan.

Corresponding author: Dr. Sandipan Ganguly. Email: sandipanganguly@hotmail.com

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Abstract

Entamoeba moshkovskii, according to recent studies, appears to exert a more significant impact on diarrheal infections than previously believed. The efficient identification and genetic characterization of *E. moshkovskii* isolates from endemic areas worldwide are crucial for understanding the impact of parasite genomes on amoebic infections. In this study, we employed an MLST system to characterize *E. moshkovskii* isolates, with the aim of assessing the role of genetic variation in the pathogenic potential of *E. moshkovskii*. We incorporated three potential genetic markers – KERP1; amoebapore C (apc); and Chitinase. Sequencing was attempted for all target loci in 68 positive *E. moshkovskii* samples, and successfully sequenced a total of 33 samples for all three loci. The analysis revealed 17 distinct genotypes, labelled M1-M17, across the tested samples when combining all loci. Notably, genotype M1 demonstrated a statistically significant association with diarrheal incidence within *E. moshkovskii* infection (p=0.0394). This suggests that M1 may represent a pathogenic strain with the highest potential for causing diarrheal symptoms. Additionally, we have identified a few SNPs in the studied loci that can be utilized as genetic markers for recognizing the most potentially pathogenic *E. moshkovskii* isolates. In our genetic diversity study, the apc locus demonstrated the highest Hd value and π value, indicating its pivotal role in reflecting the evolutionary history and adaptation of the *E. moshkovskii* population. Furthermore, analyses of linkage disequilibrium(LD) and recombination within the *E. moshkovskii* population suggested that the apc locus could play a crucial role in determining the virulence of *E. moshkovskii*.

Keywords: *Entamoeba moshkovskii,* Multilocus Sequence Typing, KERP1, amoebapore C, Chitinase, linkage disequilibrium.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Introduction

Amoebic infection is a complex issue as several species are morphologically indistinguishable from each other, including *Entamoeba histolytica*, *E. dispar, E. bangladeshi*, and *E. moshkovskii* (Fotedar *et al.,* 2007). This makes it challenging to accurately estimate the prevalence of each species and its potential to cause disease in humans. Adding to the confusion is the fact that cysts of a non-pathogenic amoeba, *E. hartmanni*, can also be mistaken with the pathogenic *E. histolytica* under a microscope (Burrows *et al.,* 1959). While *E. histolytica* is known to cause pathogenicity in amoebic infections, the actual prevalence of this species is likely overestimated due to these morphological overlaps. Recent research has indicated that *E. moshkovskii* might have a more significant impact on human infections than previously believed. This species has been detected in multiple countries, including the United States, Italy, Iran, Turkey, Indonesia, Colombia, Bangladesh, India, Kenya, Australia, Malaysia, Tanzania, Tunisia, and Brazil (Ali *et al.,* 2003; Fotedar *et al.,* 2007; Khairnar and Parija 2007; Ayed *et al.,* 2008; Beck *et al.,* 2008; Delialioglu *et al.,* 2008; Ngui *et al.,* 2012; Shimokawa *et al.,* 2012; Anuar *et al.*; 2012, Kyany'a *et al.*, 2019; Fonseca *et al.*, 2016; Al-Areeqi *et al.*, 2017). In addition, *E. moshkovskii* has been identified in farm animals like pigs, showcasing its potential for zoonotic transmission (Sardar *et al.,* 2022). Moreover, it has been documented in non-human primates (NHP) as well (Levecke *et al.,* 2010). A study conducted in eastern India by Sardar *et al.*, 2023 also revealed that *E. moshkovskii* is one of the causative agents of diarrheal incidents in humans. The research found that many patients suffering from diarrhoea, infected with *E. moshkovskii*, tested negative for other common enteric pathogens such as bacteria and viruses (Sardar *et al.,* 2023). These findings, combined with various studies conducted in different regions, suggest the potential pathogenicity of *E. moshkovskii* in humans. Therefore, diagnosing diarrheal patients should include consideration of *E. moshkovskii* as a potential pathogen to ensure accurate identification of the causative agent.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Neglecting this can result in undetermined cases of diarrheal illness, leading to improper drug treatments for patients.

Although, *E. moshkovskii* has demonstrated pathogenicity, not all genotypes within the species are linked to diarrheal incidence, similar to *E. histolytica*, where only 10% of infections exhibit pathogenicity (Sardar *et al.,* 2023). Hence, it is essential to conduct genotyping of *E. moshkovskii* to accurately identify its pathogenic genotypes. In our previous study, we have identified some significant single nucleotide polymorphisms (SNPs) that were linked to clinical outcomes (Sardar *et al.,* 2023). However, it is important to note that the correlation between SNPs and clinical features does not necessarily mean that they directly impact pathogenicity (Sardar *et al.,* 2023). While genotyping using 18S rRNA is essential in phylogenetic analysis, it does not directly affect pathogenicity. Therefore, we need to explore alternative approaches to understand how *E. moshkovskii* genotypes control pathogenicity.

Efficient identification and genetic characterization of clinical isolates from endemic areas worldwide play a crucial role in understanding the impact of parasite genomes on amoebic infections. Multilocus Sequence Typing (MLST) is a valuable tool widely used in various studies, offering a convenient and reproducible system for typing (Klint *et al.,* 2007; Bom *et al.,* 2011; Xia and Xiong, 2014; de Vries *et al.,* 2015; Urwin and Maiden, 2003). There have been successful developments of several MLST systems for characterizing strains of *E. histolytica* (Sardar *et al.,* 2023b). The selection of appropriate genetic markers is vital for genotype analysis. In our MLST study, we have integrated three potential genetic markers associated with the incidence of diarrheal *E. moshkovskii* infection. These markers include KERP1, a protein rich in lysine and glutamic acid, amoebapore C (apc) and Chitinases. By incorporating these markers into our genotyping system, we aim to enhance the characterization of *E. moshkovskii* strains and gain further insights into their role in diarrhoeal infections.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

KERP1 is a protein abundant in lysine and glutamic acid. This protein is found on the exterior of the *E. histolytica* parasite in the form of a trimeric protein complex. KERP1 is a significant factor associated with the virulence of *E. histolytica* and possesses unique characteristics that differentiate it from other known proteins. This unique protein comprises 25% lysine and 19% glutamic acid residues. Its initial discovery was prompted by its interaction with the brush border of human enterocytes (Perdomo *et al.,* 2013). KERP1 localizes at the trophozoite plasma membrane and exhibits close association to intracellular vesicles (Seigneur *et al.,* 2005). Analyses of gene expression indicate elevated KERP1 transcript levels in virulent strains, while nonvirulent *E. histolytica* strains display lower protein levels (Santi-Rocca *et al.,* 2008). In vivo investigations employing the hamster model of amoebic liver infection further supported the significance of KERP1 as a virulence factor (Santi-Rocca *et al.,* 2008, Perdomo *et al.,* 2013). Using antisense methods to decrease KERP1 expression stopped liver abscess formation, highlighting the significance of the protein in amoebic pathogenicity (Baxt and Singh 2008). While the precise function of KERP1 during infection remains unclear, it is undoubtedly engaged in trophozoite interactions, promoting host cell death, phagocytosis, and initiating inflammation in ALA development (Nozaki and Bhattacharya 2015). Therefore, KERP1 is considered a crucial virulence factor for *Entamoeba*. The KERP1 gene from *E. moshkovskii* displays homology with the corresponding gene in *E. histolytica*. The kerp1 genes of *E. histolytica* (EHI_098210), *E. nuttalli* (ENU1_189420), and *E. moshkovskii* (EMO_099600) exhibit noteworthy similarities. Specifically, there is a 100% self-match in *E. histolytica*, a 97% amino acid identity across the entire protein in *E. nuttalli*, and a 45% amino acid identity over a portion of the protein in *E. moshkovskii* (Weedall 2020).

Amoebapore C (apc) is another protein implicated in the virulence of *E. histolytica*. Earlier research has unveiled the presence of important SNPs in the upstream region of the Amoebapore C protein gene within *E. histolytica* (Bhattacharya *et al.,* 2005). These SNPs

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

exhibit a notable connection with the disease outcomes of Amoebiasis (Bhattacharya *et al.,* 2005). Although the fact that the precise role of Amoebapore C remains somewhat subtle, its influence on the severity of the disease is becoming clearer through these genetic associations. This gene exhibits homologous counterparts in various other *Entamoeba* species, including *E. dispar* strain SAW760 (EDI_206610), *E. invadens* strain IP1 (EIN_133650), and *E. moshkovskii* Laredo (EMO_119370) (Das *et al.,* 2021). Given its association with disease severity, Amoebapore C presents itself as a promising candidate for inclusion in our genotyping investigation.

The third gene analyzed in the MLST study is the Chitinase of *E. moshkovskii*. Within *Entamoeba* species, there are multiple Chitinases that share a conserved type 18 glycohydrolase domain (Vega *et al.,* 1997). The process of amoebic encystation involves the expression of Chitinase (Vega *et al.,* 1997). These Chitinase genes contain repetitive DNA sequences that display notable variations among isolates. Specifically, the repeat types and arrangement patterns within *E. histolytica* show considerable inter-isolate diversity. While the involvement of Chitinase (EC 3.2.1.14) in cyst wall formation is plausible, its role remains unverified. Chitinase functions by breaking down Chitinase, a polymer made up of N-acetyl D-glucosamine units joined by β-1,4 linkages. Although there is a suggestion that *Entamoeba* Chitinase contributes to cyst wall modification during encystation, supporting evidence is limited (Mi-Ichi *et al.,* 2021). In our MLST study, we have included a potential genetic marker: a Chitinase gene from *E. moshkovskii* (EMO_056190), which shares sequence similarity with *E. histolytica* (KM1_098160).

The objective of the MLST analysis was to identify genotypes that exhibit a statistical correlation with the co-infection status of *E. moshkovskii*, drawing insights from our epidemiological dataset. We conducted comparative genetic assessments of distinct *E. moshkovskii* populations within diverse co-infection subgroups. We also intend to uncover

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

genetic markers, such as SNPs, that display significant connections with the occurrence of diarrhoeal incidence attributed to *E. moshkovskii* infections. It is essential to investigate the correlation between *E. moshkovskii* genotypes and infection status to acquire a better understanding of the molecular mechanisms that play a role in *E. moshkovskii* pathogenesis. Furthermore, it is important to gather genome information of infecting strains from endemic areas worldwide to expand our understanding of this relationship. This study aimed to identify specific genetic variations associated with *E. moshkovskii* infection. Furthermore, the research aimed to explore the impact of genetic variations within *E. moshkovskii* subgroups on their infection outcomes, specifically concerning diarrheal incidence. This includes cases of sole *E. moshkovskii* infection and those with co-infection involving other enteric pathogens, using a multi-locus sequence typing (MLST) approach.

Materials and methods

Ethical statement

This study received ethical approval from the Institutional Human Ethics Committee of the ICMR-National Institute of Cholera and Enteric Diseases (IRB Number: A-1/2015-IEC). All participants provided informed consent, and for children, voluntary consent was obtained from their caregivers (parent/guardian).

Samples

The study utilizes the 68 samples that tested positive for *E. moshkovskii*. These specific samples had been identified using microscopy and PCR techniques targeting the 18S rRNA locus, as described in a prior publication (Sardar *et al.,* 2023). This study involved the characterization of five subgroups within *E. moshkovskii* populations: Sole D, IOEP, IEH, ISTH, and IB/V. Sole D represents patients experiencing diarrhoea solely due to *E. moshkovskii* infection. IEH

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

signifies *E. moshkovskii* positive samples co-infected with *E. histolytica.* IOEP denotes *E. moshkovskii* positive samples co-infected with other Enteric Parasites, including *G. lamblia* and *Cryptosporidium* spp. ISTH corresponds to *E. moshkovskii* positive samples co-infected with soil-transmitted helminths. Lastly, IB/V includes *E. moshkovskii* positive samples co-infected with other diarrheal agents, such as *E. coli*, *Shigella* spp, *V. cholera*, or the virus Rotavirus. For a comprehensive understanding of the study area, stool sample collection, DNA extraction process, detection of *E. moshkovskii* in the specimens, and data collection on co-infection status, we recommend referring to the methodology section of our previous study by Sardar *et al*., 2023. (Sardar *et al.,* 2023).

PCR amplification:

Positive samples for *E. moshkovskii* were chosen to amplify three target genes: KERP1, amoebapore C (apc), and Chitinase. The amplification process was conducted using a reaction mixture with a volume of 50 µl. This mixture included 5 units of TaKaRa Ex-Taq polymerase, PCR buffer at a 1X concentration, 0.2 µM of both forward and reverse primers, and 3 µl of stool DNA samples with a concentration of 50 ng/µl. The amplification reactions were performed using a thermal cycler PCR system from Applied Biosystems. The PCR cycling procedure commenced with an initial denaturation phase at 94°C, lasting for 5 minutes. This was followed by 35 amplification cycles, each comprised of distinct steps. These cycles consisted of a denaturation step at 94°C for a duration of 30 seconds, an annealing phase at 56°C (for KERP1) for 25 seconds, a polymerization step at 72°C lasting 45 seconds, and a concluding extension stage at 72°C, maintained for 7 minutes. The amplification process for apc and Chitinase followed a similar pattern, except the annealing temperature was set at 57°C, and the polymerization time was reduced to 35 seconds.

Primer sequences, expected PCR product sizes, and annealing temperatures employed

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

are provided in **Table 1.1**. Following amplification, the PCR products were subjected to electrophoresis on agarose gel (Seakem® LE Agarose, Lonza) and subsequently visualized under a UV-transilluminator following staining with 0.5µm/ml ethidium bromide.

DNA sequencing

PCR products of the expected sizes were extracted using a Roche Gel Extraction Kit following the manufacturer's protocols. Their yield was subsequently verified through gel electrophoresis. The purified PCR products were then subjected to direct sequencing using the corresponding amplification primers in both forward and reverse directions. This sequencing process employed the BigDye Terminator v3.1 Cycle Sequencing Kit from Applied Biosystems, USA. The obtained sequences were analyzed using an ABI3730 sequencer.

Sequence analysis

The DNA sequences obtained were aligned using the clustalW multiple sequence alignment program from GenomeNet Bioinformatics resources and manually edited. Subsequently, the DNA sequences from the three loci were combined to determine a genotype. Alphanumerical codes were then assigned to the obtained genotypes. We also identify the SNPs present in the obtained local isolates. All the sequences were aligned with reference sequences obtained from AmoebaDB using the MultAlin online tool and thereafter the SNPs were identified. The representative nucleotide sequences of each haplotype reported in this study have been deposited to NCBI GenBank. We also inferred relationships between haplotypes by constructing a minimal-spanning haplotype network using Pop-ART v1.7. To elucidate the connections within the sequence data, we employed a color-coded representation of isolates based on their co-infection status.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Statistical analysis

Categorical data analysis was conducted using GraphPad Prism 9, CA, USA. The association between genotypes/repeat patterns and clinical phenotypes was assessed using Fisher's exact test. Statistical significance was defined as a p-value less than 0.05 in all instances.

Results

Successful amplification of the target loci

Out of the total 68 samples, successful amplification was achieved in 33 samples across three designated target loci (**Support file 1**). Nevertheless, the remaining samples did not yield successful amplification for all three of these loci due to the presence of low DNA concentration in the stool samples. Some samples exhibited faint amplification, which was inadequate for sequencing, while others displayed no distinct bands upon agarose gel electrophoresis. The lack of amplification observed in certain samples could be attributed to either a low cyst concentration of *E. moshkovskii* in faecal samples or issues with the quality of DNA. Furthermore, the existence of genetic polymorphisms at the specified target sites could also contribute to the observed lack of amplification. Although implementing a nested PCR technique has the potential to enhance the amplification rate, we did not prefer to adopt this approach to prevent the risk of cross-contamination. The representative haplotype sequences obtained in this study have been submitted to NCBI GenBank under accession numbers OR621050-OR621064.

Single nucleotide polymorphisms (SNPs)

A total of twenty-two SNPs and three deletions were detected within three genes. We successfully obtained complete gene sequences for the kerp1 and Chitinase genes. In contrast, our analysis of the apc gene involved a partial sequence, encompassing both a partial

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

coding region and a 322 bp-long intron region, as predicted in AmoebaDB.

In Kerp1, we have identified a total of 9 SNPs along with three specific deletions denoted as 33A, 42A, and 52A. These deletions resulted in the removal of a sequence of three base pairs, consequently altering the corresponding amino acid sequence. Specifically, the sequence **-Glu-Val-Val-Gln-His-Arg-Ala-Ser-** was substituted with **-Glu-Trp-Phe-Thr-Gln-Ser-Ser- (Fig. 1)**. As a consequence of the deletion, one amino acid was lost in the replaced sequence. Interestingly, despite this alteration in the amino acid sequence, our analysis did not reveal any statistically significant associations between this deleted amino acid stretch and various coinfected groups. Furthermore, we have identified five SNPs labeled as 23A/C, 25 A/C, 26C/T, 27T/G, and 28G/A. These variants have shown a significant correlation with diarrheal incidences in cases of sole infection with *E. moshkovskii*, with corresponding p-values of 0.0261 for each SNP. Conversely, the remaining three SNPs did not exhibit any statistically significant connections with the coinfected groups.

In the apc gene, a total of 10 SNPs were identified. Among these, 9 SNPs were located within the intronic region of the gene. Four intronic SNPs exhibited a statistically significant correlation with the occurrence of sole diarrhoea, which we refer to as the Sole D group (as shown in Table 6.2). Within these four SNPs, three - namely 420T/A (p=0.0224), 564T/A (p=0.2916), and 523A/T (p=0.0538) - demonstrated a positive statistical correlation with the incidence of sole diarrhoea. Additionally, one SNP, 299A/T, displayed a negative association (p=0.0315) with sole diarrhoeal incidence. These intronic SNPs may have the capacity to decrease protein levels by potentially influencing splicing processes (Wang and Sadee 2016). Moreover, the identified SNPs can also serve as genetic markers. It is important to note that the intron region reported in this study was based on the predicted genomic sequence of the apc gene (EMO_119370) sourced from AmoebaDB.

The Chitinase locus analysis revealed the presence of three SNPs, all of which lacked

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

statistical significance when correlated with specific subgroups of *E. moshkovskii.* Details of these identified SNPs within the target loci of the studied isolates can be found in **Table 2**.

In the obtained DNA sequences of the three loci, a majority of the SNPs were characterized as non-synonymous mutations.

Association between genotype and co-infection status:

After conducting sequencing on all the samples, we identified a total of six distinct haplotypes in the Kerp1 group. Notably, one of these haplotypes, Emk3, displayed a complete 100% match with the reference sequence EMO_099600. In contrast, the apc group also exhibited six distinct haplotypes, all of which differed from the reference sequence EMO_119370. Within the Chitinase gene sequences, we observed three distinct haplotypes. Notably, one of these haplotypes, Emch1, exhibited similarity to the reference sequence EMO_056190. These individual haplotypes were subsequently pooled together to construct the respective genotypes.

After combining three distinct loci, this study has successfully identified 17 distinct genotypes labeled as M1 through M17. Among these genotypes, seven, specifically M1, M3, M5, M8, M9, and M17, were identified in multiple isolates (**Table 3**). Notably, genotype M1 exhibited a statistically significant association with the sole diarrhoeal group within *E. moshkovskii* infection, as indicated by a p-value of p=0.0394. While M9 and M17 were found in Sole D and IOEP subgroups with multiple occurrences, their presence did not show any statistically significant associations with their respective groups. The M12 genotype was detected in several groups, except for the Sole D group. The remaining genotypes were found in both the Sole D group and other co-infected groups. This research demonstrated that the M1 genotype holds the highest potential for being a pathogenic strain of *E. moshkovskii.*

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Genetic diversity:

The haplotype diversity (Hd) across individual polymorphic loci ranges from 0.572 to 0.833. Within the 33 samples of *E. moshkovskii*, the number of haplotypes ranges from 2 to 6. Among the three examined loci, the apc locus showed with the highest Hd value of 0.833 and the highest number of haplotypes (6). Additionally, the apc locus exhibits the highest observed nucleotide diversity. Moderate levels of haplotype diversity were revealed within the kerp1 and Chitinase loci, with Hd values of 0.572 and 0.589, respectively (**Table 4**). Additionally, these loci exhibited moderate levels of nucleotide diversity (π), with kerp1 exhibiting 0.0048 and *Chitinase* exhibiting 0.0035. Among the loci examined, the apc locus showed the greatest number of polymorphic sites (10), while the Chitinase locus showed the fewest (3). Tajima's D statistics revealed positive values for all of the loci. These positive values could imply the presence of either a population bottleneck or balancing selection. The presence of a substantial variety of genotypes within this group, coupled with the positive Tajima's D value, may support the hypothesis of balancing selection (**Table 4**). However, given that these values did not show statistical significance, confirming these results would require a larger sample size.

The combined nucleotide sequences of the three target loci were either 1259 bp (Due to 3 deletions in Kerp1) or 1256 bp long and included 22 variable sites. The Hd value was 0.93. Tajima's D statistic for the concatenated sequences was 1.42, supporting the idea of balancing selection (**Table 5**). To validate these findings further, a larger sample size would be needed.

Linkage disequilibrium (LD) and recombination analyses of target loci

We assessed intragenic LD and the count of potential recombination events for each target locus. At the apc locus, an incomplete intragenic LD value was observed (|D′| Y = 0.9779 - 0.1659X), with Y representing the LD value and X indicating the nucleotide distance in kilobases. The incomplete intragenic LD value at the apc locus suggests a non-random

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

distribution of its alleles within the *E. moshkovskii* population. Conversely, complete LD values (|D′| Y = 1.0000 – 0.0000X) were discovered at the Kerp1 and Chitinase loci, indicating a random distribution of alleles for these two genes. The analysis of intragenic recombination revealed a single potential event (Rm) exclusively at the apc locus. Conversely, no recombination events were observed at the Kerp1 and Chitinase loci, signifying that the alleles of these two genes are distributed randomly in the studied population (**Table 4)**.

Our study isolates were analyzed for an overall interlocus LD and the number of potential recombination events. The concatenated multilocus sequence data was used for this purpose. The analysis revealed an incomplete LD value (|D′| Y=0.8169 + 0.1059X) with a single potential recombination event in the population. This was observed when the concatenated sequences of all three loci were analyzed. An incomplete interlocus LD value (|D′| Y = Y=0.7682 + 0.4469X) was discovered between Chitinase and Kerp1, indicating a single recombination event. However, upon further analysis of concatenated sequences among apc + Chitinase and apc + Kerp1, two potential recombination events with incomplete interlocus LD values were observed (**Table 6**). This interesting finding suggests a possible non-random association of the apc locus with both Chitinase and Kerp1.

Interlocus LD and recombination analyses of E. moshkovskii *population from different co-infection/Sole infection groups*

We conducted Interlocus LD and recombination analyses on the *E. moshkovskii* population from four groups (Sole D, IBV, IEH, and IOEP) using concatenated multilocus sequence data. Our analysis revealed that only the apc locus showed a single recombination event in intralocus LD analysis. To further analyze interlocus LD values, we looked at concatenated multilocus sequences both including and excluding the apc locus. Interestingly, we found that the inclusion of the apc locus increased recombination events. However, LD analysis of

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

concatenated multilocus sequences from the IEH and IOEP groups produced a complete interlocus LD value (| D′| Y = 1.0000 + 0.0000×) with no recombination events in either case. Based on the complete interlocus LD value (|D′| Y = 1.0000 + 0.0000×) and absence of recombination events in the IEH and IOEP groups, it appears that these groups might be isolated compared to the others. However, since most concatenated sequences show at least one recombination event, inter-population genetic recombination may occur among the different subpopulations (**Table 7 and Table 8**). One notable finding from this study is that the IEH and IOEP populations of *E. moshkovskii* are undergoing a speciation process due to their isolation.

Haplotype network construction

Haplotype grouping was conducted using the KERP1, amoebapore C (apc), and Chitinase markers. The inclusion of various co-infected subgroups in the haplotype network did not result in any apparent impact on isolate grouping (**Figs. 2a, 2b** and **2c**). In Kerp1, Em1kerp1/Hap1 has emerged as the predominant haplotype within all subgroups and is likely the ancestral haplotype. However, in our analysis, the Em1kerp1/Hap3 genotype is reported as the original DNA code (prototype) for the Kerp1 protein. This determination is based on its perfect alignment with the reference sequence EMO_099600 from AmoebaDB, as well as its significantly high prevalence across all co-infected subgroups, supporting its prototype status. A frameshift was observed due to the deletion of three bases, leading to the origin of other mutant types from Em1kerp1/Hap3. The remarkable adaptability of Em1kerp1/Hap3, as indicated by its prevalence across all subgroups, further underscores its prototype status. However, it was observed that Em1kerp1/Hap6 originated from either Em1kerp1/Hap1 or Em1kerp1/Hap3 and was specifically found in the IOEP subgroups. The Em1kerp1/Hap2 was at one mutational step from Em1kerp1/Hap1 and was not observed in the IOEP subgroup. The

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Em1kerp1/Hap6 haplotype forms a branch with four descendant haplotypes, namely Em1kerp1/Hap3, Em1kerp1/Hap4, and Em1kerp1/Hap5.

In Amoebapore C, six haplotypes (Emapc/hap1 to Emapc/hap6) have been identified. The most frequent haplotype, Emapc/hap1, is likely the ancestral haplotype and is predominant in all four subgroups. Emapc/hap1 is six mutational steps away from Emapc/hap6. However, Emapc/hap2 and Emapc/hap4 are one and two mutational steps, respectively, from Emapc/hap6. Emapc/hap5 has emerged as a descendant of both Emapc/hap2 and Emapc/hap4 haplotypes. Interestingly, Emapc/hap5 is exclusively observed in parasites from the Sole D group. In of Chitinase, we detected three distinct haplotypes: Emch/hap1, Emch/hap2, and Emch/hap3. Among these, Emch/hap1 was the most prevalent and presumed to be the ancestral haplotype. Emch/hap2 and Emch/hap3 were found to be at two and one mutational step away from Emch/hap1, respectively. Notably, Emch/hap2 was not observed within the IEH subgroups.

Discussion

Our goal was to analyze the genetic makeup of various isolates of *E. moshkovskii* and its association with virulence factors found in *E. histolytica*. We focused specifically on the loci of lysine and glutamic acid-rich protein 1 (KERP1), amoebapore C (pore-forming peptides), and Chitinase. KERP1 is found on the trophozoite plasma membrane and internal vesicles, where it plays a crucial role in establishing amoeba-cell contacts and the development of liver abscesses (Perdomo et al., 2013, Santi-Rocca *et al.,* 2008). Amoebapore forms ion channels or pores in lipid membranes, depolarizing target cells (Leippe *et al.,* 1991). The expression of amoebapores is necessary for the complete manifestation of virulence in *E. histolytica*, particularly in the context of amebic liver abscesses (Zhang *et al.,* 2004). Chitinase, on the other hand, breaks down chitinase, a β-1,4-linked polymer of N-acetyl-d-glucosamine, and is

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

believed to be involved in remodelling the cyst wall during encystation in *Entamoeba* (Mi-Ichi *et al.,* 2021, Chatterjee *et al.,* 2009). In our study, we employed PCR amplification using specific primers designed for targeting these genetic loci. This approach has the potential to provide novel insights into the co-infection dynamics of *E. moshkovskii.*

Accurate identification and genetic characterization of clinical isolates from endemic regions worldwide provides a valuable tool for understanding the impact of parasite genome on the outcomes of amoebic infections. Previous research has established that tRNA-linked STR loci serve as surrogate markers for determining disease outcomes (Ali *et al.,* 2012). In our current research, we have genetically characterized *E. moshkovskii* populations with varying co-infected groups using the above mentioned coding genes. The Kerp1 gene exhibited the highest number of SNPs, with five of them being associated with diarrhoea incidence and potentially serving as genetic markers. These SNPs may also play a role in modulating the pathogenicity of *E. moshkovskii.* The Apc gene displayed a number of significant SNPs, the exact impact of which remains to be determined. However, the presence of these SNPs can serve as a genetic marker for diarrhoeal diseases caused by isolates similar to those observed in Kerp1. Most of the SNPs observed in our study were non-synonymous, which aligns with previous findings for Mycobacterium tuberculosis, where only 36 out of 101 identified SNPs were synonymous (Baker *et al.,* 2004). This trend was also noted by Das *et al.,* 2021 in their study of Multilocus sequence typing (MLST) of *Entamoeba histolytica* (Das *et al.,* 2021). Interestingly, the Chitinase gene showed only synonymous SNPs and was found to be the most conserved gene among the studied loci, with only three SNPs observed. In contrast, the Chitinase of *E. histolytica* is highly polymorphic and contains STR units, whereas the Chitinase of *E. moshkovskii* is not a highly polymorphic gene as observed in this study (Das *et al.,* 2014). The reported SNPs could potentially play a role in drug sensitivity in *E. moshkovskii*, similar to how certain SNPs have been linked to multidrug resistance in *Plasmodium falciparum*, as

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

reported by Coulibaly *et al.,* in 2022 (Coulibaly *et al.,* 2022).

Recombination events were only identified in the apc locus. The presence of co-infection specific SNPs, potential recombination events within the apc locus, and various non-synonymous base changes all suggest that this region of the genome is under selection pressure. As such, these observations may indicate that apc could play a crucial role in determining the virulence of *E. moshkovskii*. The precise function of apc in *E. moshkovskii* pathogenesis remains unknown. While investigating another gene, Kerp1, we identified a significant number of SNPs, despite the absence of detected potential recombination events. Therefore, further studies with a larger sample size are required to gain a better understanding of role of Kerp1 in population structure of *E. moshkovskii*. It is important to note that this is the first molecular epidemiology based study to date conducted on the role of different genes in the pathogenicity of *E. moshkovskii*.

The findings of the study suggest a correlation between the parasite genotypes and *E. moshkovskii* infection status. This study is the first to explore the direct link between parasite factors and the infection dynamics of *E. moshkovskii*. However, further biomarkers are necessary to comprehensively understand the role of the parasite genome.

Conclusion

The latest research has unveiled the genetic composition of *E. moshkovskii* isolates under investigation, establishing their link to infection dynamics. Through the analysis, numerous significant SNPs within specific genetic regions have been detected. These SNPs exhibit the potential to influence, either directly or indirectly, the pathogenicity and drug sensitivity of *E. moshkovskii*. The investigation has also pinpointed distinct clusters of isolates that display genetic segregation. Moreover, the study supports the hypothesis that a connection exists between parasite genotypes and infection dynamics.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Acknowledgements. The authors would like to express their appreciation to all patients who participated in the study. We are also grateful for the support extended by the hospital staff in facilitating the collection of samples.

Ethical Approval. The ethical clearance of this study was reviewed and approved by the Institutional Human Ethics Committee of the ICMR-NICED. Informed consent statements were obtained from the participants.

Competing interests. The authors affirm that they do not have any competing interests related to the publication of the article.

Authors' contributions. The authors confirm their contribution to the article as follows: **Sanjib K. Sardar:** Conceptualization, Visualization, Methodology, Data curation, Writing original draft, Formal analysis. **Ajanta Ghosal**: Methodology. **Tapas Haldar**: Methodology. **Akash Prasad**: Methodology. **Sweety Mal:** Methodology. **Yumiko Saito Nakano:** Validation and investigation **Seiki Kobayashi:** Validation. **Shanta Dutta:** Project administration. **Tomoyoshi Nozaki:** Validation. **Sandipan Ganguly:** Conceptualization, Visualization, Validation, Funding acquisition, Review and editing, Supervision.

Funding. This study was funded by the Indian Council of Medical Research (ICMR), the Government of India and Japan Agency for Medical Research and Development (AMED, grant number: JP23fk0108683), Japan. The first author received financial support through the ICMR Fellowship.

Availability of data and materials. Representative sequences obtained in this study were

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

deposited in GenBank under the accession numbers OR621050-OR621064.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

References:

Al-Areeqi, MA, Sady, H, Al-Mekhlafi, HM, Anuar, TS, Al-Adhroey, AH, Atroosh, WM, Dawaki, S, Elyana, FN, Nasr, NA, Ithoi, I, Lau, YL and Surin J (2017) First molecular epidemiology of *Entamoeba histolytica, E. dispar* and *E. moshkovskii* infections in Yemen: different species-specific associated risk factors. *Tropical medicine & international health: TM & IH* **22(4)**, 493-504. doi:10.1111/tmi.12848.

Ali, IK, Haque, R, Alam, F, Kabir, M, Siddique, A and Petri WA Jr (2012) Evidence for a link between locus R-R sequence type and outcome of infection with *Entamoeba histolytica. Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases* **18(7)**, E235-7. doi:10.1111/j.1469-0691.2012.03826.x.

Ali, IK, Hossain, MB, Roy, S, Ayeh-Kumi, PF, Petri, WA, Jr, Haque, R and Clark CG (2003) *Entamoeba moshkovskii* infections in children, Bangladesh. *Emerging infectious diseases* **9(5)**, 580-4. doi:10.3201/eid0905.020548.

Anuar, TS, Al-Mekhlafi, HM, Ghani, MK, Azreen, SN, Salleh, FM, Ghazali, N, Bernadus, M, and Moktar N (2012) First molecular identification of *Entamoeba moshkovskii* in Malaysia. *Parasitology* **139(12)**, 1521-5. doi:10.1017/S0031182012001485.

Ayed, SB, Aoun, K, Maamouri, N, Abdallah, RB and Bouratbine A (2008) First molecular identification of *Entamoeba moshkovskii* in human stool samples in Tunisia. *The American journal of tropical medicine and hygiene* **79(5)**, 706-7. PMID: 18981508.

Baker, L, Brown, T, Maiden, MC and Drobniewski F (2004) Silent nucleotide polymorphisms and a phylogeny for Mycobacterium tuberculosis. *Emerging infectious diseases* **10(9)**, 1568-77. doi:10.3201/eid1009.040046.

Baxt, LA and Singh U (2008) New insights into *Entamoeba histolytica* pathogenesis. *Current opinion in infectious diseases* **21(5)**, 489-94. doi:10.1097/QCO.0b013e32830ce75f.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Beck, DL, Doğan, N, Maro, V, Sam, NE, Shao, J and Houpt ER (2008) High prevalence of *Entamoeba moshkovskii* in a Tanzanian HIV population. *Acta tropica* **107(1)**, 48-9. doi: 10.1016/j.actatropica.2008.03.013. Epub 2008 Apr 7. PMID: 18471796; PMCID: PMC2459240.

Bhattacharya, D, Haque, R, Singh U (2005) Coding and noncoding genomic regions of *Entamoeba histolytica* have significantly different rates of sequence polymorphisms: implications for epidemiological studies. *Journal of clinical microbiology*. **43(9),** 4815-9. doi: 10.1128/JCM.43.9.4815-4819.2005.

Bom, RJ, Christerson, L, Schim van der Loeff, MF, Coutinho, RA, Herrmann, B and Bruisten SM (2011) Evaluation of high-resolution typing methods for Chlamydia trachomatis in samples from heterosexual couples. *Journal of clinical microbiology* **49(8)**, 2844-53. doi:10.1128/JCM.00128-11.

Burrows, RB (1959) Morphological differentiation of *Entamoeba hartmanni* and *E. polecki* from *E. histolytica*. *The American journal of tropical medicine and hygiene* **8**, 583-9. doi:10.4269/ajtmh.1959.8.583.

Chatterjee, A, Ghosh, SK, Jang, K, Bullitt, E, Moore, L, Robbins, PW and Samuelson J (2009) Evidence for a "wattle and daub" model of the cyst wall of *Entamoeba*. *PLoS pathogens* **5(7)**, e1000498. doi:10.1371/journal.ppat.1000498.

Coulibaly, A, Diop, MF, Kone, A, Dara, A, Ouattara, A, Mulder, N, Miotto, O, Diakite, M, Djimde, A and Amambua-Ngwa A (2022) Genome-wide SNP analysis of *Plasmodium falciparum* shows differentiation at drug-resistance-associated loci among malaria transmission settings in southern Mali. *Frontiers in genetics* **13**, 943445. doi:10.3389/fgene.2022.943445.

Das, K, Mukherjee, AK, Chowdhury, P, Sehgal, R, Bhattacharya, MK, Hashimoto, T, Nozaki, T and Ganguly S (2014) Multilocus sequence typing system (MLST) reveals a

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

significant association of *Entamoeba histolytica* genetic patterns with disease outcome. *Parasitology international* **63(2)**, 308-14. doi:10.1016/j.parint.2013.11.014

Das, K, Sardar, SK, Ghosal, A, Saito-Nakano, Y, Dutta, S, Nozaki T and Ganguly S (2021) Multilocus sequence typing (MLST) of *Entamoeba histolytica* identifies kerp2 as a genetic marker associated with disease outcomes. *Parasitology international* **83**, 102370. doi:10.1016/j.parint.2021.102370.

de la Vega, H, Specht, CA, Semino, CE, Robbins, PW, Eichinger, D, Caplivski, D, Ghosh, S, and Samuelson J (1997) Cloning and expression of chitinases of *Entamoebae*. *Molecular and biochemical parasitology* **85(2)**, 139-47. doi:10.1016/s0166-6851(96)02817-4.

de Vries, HJ, Schim van der Loeff, MF and Bruisten SM (2015) High-resolution typing of Chlamydia trachomatis: epidemiological and clinical uses. *Current opinion in infectious diseases* **28(1)**, 61-71. doi:10.1097/QCO.0000000000000129.

Delialioglu, N, Aslan, G, Ozturk, C, Ozturhan, H, Sen, S and Emekdas G (2008) Detection of *Entamoeba histolytica* antigen in stool samples in Mersin, Turkey. *The Journal of parasitology* **94(2)**, 530-2. doi:10.1645/GE-1355.1.

Fonseca, JA, Heredia, RD, Ortiz, C, Mazo, M, Clavijo-Ramírez, CA and Lopez MC (2016) Identification of *Entamoeba moshkovskii* in Treated Waste Water Used for Agriculture. *Ecohealth* **13(1)**, 156-60. doi:10.1007/s10393-015-1084-6.

Fotedar, R, Stark, D, Beebe, N, Marriott, D, Ellis, J and Harkness, J (2007) PCR detection of *Entamoeba histolytica, Entamoeba dispar,* and *Entamoeba moshkovskii* in stool samples from Sydney, Australia. *Journal of clinical microbiology* **45(3)**, 1035-7. doi: 10.1128/JCM.02144-06. Epub 2007 Jan 17.

Khairnar, K and Parija SC (2007) A novel nested multiplex polymerase chain reaction (PCR) assay for differential detection of *Entamoeba histolytica, E. moshkovskii* and *E.*

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

dispar DNA in stool samples. *BMC microbiology* **7, 47**. doi:10.1186/1471-2180-7-47.

Klint, M, Fuxelius, HH, Goldkuhl, RR, Skarin, H, Rutemark, C, Andersson, SG, Persson, K, and Herrmann B (2007) High-resolution genotyping of Chlamydia trachomatis strains by multilocus sequence analysis. *Journal of clinical microbiology* **45(5)**, 1410-4. doi:10.1128/JCM.02301-06.

Kyany'a, C, Eyase, F, Odundo, E, Kipkirui, E, Kipkemoi, N, Kirera, R, Philip, C, Ndonye, J, Kirui, M, Ombogo, A, Koech, M, Bulimo, W and Hulseberg CE (2019) First report of *Entamoeba moshkovskii* in human stool samples from symptomatic and asymptomatic participants in Kenya. *Tropical diseases, travel medicine and vaccines* **5, 23**. doi:10.1186/s40794-019-0098-4.

Leippe, M, Ebel, S, Schoenberger, OL, Horstmann, RD and Müller-Eberhard HJ (1991) Pore-forming peptide of pathogenic *Entamoeba histolytica*. *Proceedings of the National Academy of Sciences of the United States of America* **88(17)**, 7659-63. doi:10.1073/pnas.88.17.7659.

Levecke, B, Dreesen, L, Dorny, P, Verweij, JJ, Vercammen, F, Casaert, S, Vercruysse, J and Geldhof P (2010) Molecular identification of *Entamoeba* spp. in captive nonhuman primates. *Journal of clinical microbiology* **48(8)**, 2988-90. doi:10.1128/JCM.00013-10.

Mi-Ichi, F, Sakaguchi, M, Hamano, S and Yoshida H (2021) *Entamoeba* Chitinase is Required for Mature Round Cyst Formation. *Microbiology spectrum* **9(1)**, e0051121. doi: 10.1128/Spectrum.00511-21. Epub 2021 Aug 4.

Mi-Ichi, F, Tsugawa, H, Arita, M and Yoshida H (2022) Pleiotropic Roles of Cholesteryl Sulfate during *Entamoeba* Encystation: Involvement in Cell Rounding and Development of Membrane Impermeability. *mSphere* **7(4)**, e0029922. doi:10.1128/msphere.00299-22.

Ngui, R, Angal, L, Fakhrurrazi, SA, Lian, YL, Ling, LY, Ibrahim, J and Mahmud R (2012) Differentiating *Entamoeba histolytica, Entamoeba dispar* and *Entamoeba*

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

moshkovskii using nested polymerase chain reaction (PCR) in rural communities in Malaysia. *Parasites & vectors* **5, 187**. doi:10.1186/1756-3305-5-187.

Perdomo, D, Baron, B, Rojo-Domínguez, A, Raynal, B, England, P and Guillén, N (2013) The α-helical regions of KERP1 are important in *Entamoeba histolytica* adherence to human cells. *Scientific reports* **3, 1171**. doi: 10.1038/srep01171.

Santi-Rocca, J, Weber, C, Guigon, G, Sismeiro, O, Coppée, JY and Guillén N (2008) The lysine- and glutamic acid-rich protein KERP1 plays a role in *Entamoeba histolytica* liver abscess pathogenesis. *Cellular microbiology* **10(1)**, 202-17. doi:10.1111/j.1462-5822.2007.01030.x.

Sardar, SK, Das, K, Maruf, M, Haldar, T, Saito-Nakano, Y, Kobayashi, S, Dutta, S and Ganguly S (2022) Molecular evidence suggests the occurrence of *Entamoeba moshkovskii* in pigs with zoonotic potential from eastern India. *Folia parasitologica* **69**, 2022.012. doi:10.14411/fp.2022.012.

Sardar, SK, Ghosal, A, Haldar, T, Maruf, M, Das, K, Saito-Nakano, Y, Kobayashi, S, Dutta, S, Nozaki, T and Ganguly S (2023) Prevalence and molecular characterization of *Entamoeba moshkovskii* in diarrheal patients from Eastern India. *PLoS neglected tropical diseases* **17(5)**, e0011287. doi:10.1371/journal.pntd.0011287.

Sardar, SK, Ghosal, A, Haldar, T, Das, K, Saito-Nakano, Y, Kobayashi, S, Dutta, S, Nozaki, T and Ganguly S (2023) Investigating genetic polymorphism in *E. histolytica* isolates with distinct clinical phenotypes. *Parasitology research* **122(11)**, 2525-2537. doi:10.1007/s00436-023-07952-x.

Seigneur, M, Mounier, J, Prevost, MC and Guillén N (2005) A lysine- and glutamic acid-rich protein, KERP1, from *Entamoeba histolytica* binds to human enterocytes. *Cellular microbiology* **7(4)**, 569-79. doi:10.1111/j.1462-5822.2005.00487.x.

Shimokawa, C, Kabir, M, Taniuchi, M, Mondal, D, Kobayashi, S, Ali, IK, Sobuz, SU,

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Senba, M, Houpt, E, Haque, R, Petri, WA Jr and Hamano S (2012) *Entamoeba moshkovskii* is associated with diarrhea in infants and causes diarrhea and colitis in mice. *The Journal of infectious diseases* **206(5)**, 744-51. doi:10.1093/infdis/jis414.

Urwin, R and Maiden MC (2003) Multi-locus sequence typing: a tool for global epidemiology. *Trends in microbiology* **11(10)**, 479-87. doi:10.1016/j.tim.2003.08.006.

Wang D, Sadee W (2016) CYP3A4 intronic SNP rs35599367 (CYP3A4*22) alters RNA splicing. Pharmacogenet Genomics **26(1)**, 40-3. doi: 10.1097/FPC.0000000000000183.

Weedall, GD (2020) The *Entamoeba* lysine and glutamic acid rich protein (KERP1) virulence factor gene is present in the genomes of *Entamoeba nuttalli, Entamoeba dispar* and *Entamoeba moshkovskii. Molecular and biochemical parasitology* **238**, 111293. doi:10.1016/j.molbiopara.2020.111293.

Xia, Y and Xiong L (2014) Progress in genotyping of *Chlamydia trachomatis*. *Chinese medical journal* **127(22)**, 3980-6. PMID: 25421201.

Zhang, X, Zhang, Z, Alexander, D, Bracha, R, Mirelman, D and Stanley SL Jr (2004) Expression of amoebapores is required for full expression of *Entamoeba histolytica* virulence in amebic liver abscess but is not necessary for the induction of inflammation or tissue damage in amebic colitis. *Infection and immunity* **72(2)**, 678-83. doi:10.1128/IAI.72.2.678-683.2004.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Table 1. Primer sequences, expected PCR product sizes, and annealing temperatures of the targeted loci.

Name of the primer	**Primer sequence (5'-3')**	**Annealing temperature**	**Product size**
Emkerp1_F	TATGAGCGTTGGGGAGATTC	56°C	594 bp
Emkerp2_R	CTTCCCGCCATCAAAAATAA		
Emapc_F	TCTTGAAAGTCTTTGCGCCA	57°C	449 bp
Emapc_R	TCCTCCTCTCGTAGTCCAAA		
EmChitinase_F	TGTGGTGTTTCAAAAGTTTCCA	57°C	357 bp
EmChitinase_R	CAACACAAAAATAAATAGTCATTCACG		

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Table 2. Identified SNPs within the target loci of the studied isolates.

Target loci	**Base pair (bp) analysed**	***SNP position including deletion**	**Amino acid substitution**	**Coinfection status**			
				Sole D	**IEH**	**IOEP**	**IB/V**
Kerp1	546	Deletion 33A, 42A, 52A	12-17 VVQHRA/ WFT|QS	X^C	X^C	X^C	X^C
		23A/C	8Q/P	p^b=0.0261 X^2=4.90	X^C	X^C	X^C
		25 A/C	9T/L	p^b=0.0261 X^2=4.90	X^C	X^C	X^C
		26C/T	9 T/L	p^b=0.0261 X^2=4.90	X^C	X^C	X^C
		27T/G	11 T/L	p^b=0.0261 X^2=4.90	X^C	X^C	X^C
		28G/A	10V/I	p^b=0.0261 X^2=4.90	X^C	X^C	X^C
		374 A/G	125 E/G	X^C	X^C	X^C	X^C
		405 G/T	135N/K	X^C	X^C	X^C	X^C
		432C/G	144D/E	X^C	X^C	X^C	X^C
		441T/A	147D/E	X^C	X^C	X^C	X^C
Amoebapore C (apc)	407	299A/T	Intron region	p^b=0.0315 X^2= 4.626 (negative association)	X^C	X^C	X^C
		325A/T	Intron region	X^C	X^C	X^C	X^C
		327C/T	Intron region	X^C	X^C	X^C	X^C

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

		360T/A	Intron region	X^C	X^C	X^C	X^C
		370T/A	Intron region	X^C	X^C	X^C	X^C
		398A/C	Intron region	X^C	X^C	X^C	X^C
		420T/A	Intron region	$p^b=0.0224$ $X^2= 5.215$	X^C	X^C	X^C
		564T/A	Intron region	$p^b= 0.2916$ $X^2= 1.112$	X^C	X^C	X^C
		523A/T	Intron region	$p^b= 0.0538$ $X^2= 3.718$	X^C	X^C	X^C
		607A/T	55K/N	X^C	X^C	X^C	X^C
Chitinase	306	75T/C	25Y/Y(Synonymous)	X^C	X^C	X^C	X^C
		223A/T	75I/L	X^C	X^C	X^C	X^C
		286T/C	96S/P	X^C	$p^b= 0.0001$ $X^2= 17.22$ (negative association)	X^C	X^C

[b] Probability value of the particular association

[c] Does not have any association with disease outcomes,

[d] Synonymous SNPs, do not affect the protein sequences,

[e] intron region.

X^2=Chi-square value

***The positions of all SNPs were reported with reference to first base of the start codon, which is designated as position 1.**

Table 3. List of identified genotypes (M1 to M17) resulting from the combination of three independent studied loci

Co-infection status	**Sample ID**	**Sequence pattern**			**Genotype**
		Kerp1	**Amoebapore C (apc)**	**Chitinase**	
Sole D	EM_IND/1	Emk2	Emapc3	Emch1	[b]M1
	EM_IND/3	Emk1	Emapc5	Emch1	M2
	EM_IND/4	Emk2	Emapc3	Emch1	[b]M1
	EM_IND/5	Emk2	Emapc3	Emch1	[b]M1
	EM_ID/6	Emk1	Emapc6	Emch2	M3
	EM_IND/10	Emk1	Emapc4	Emch1	M4
	EM_IND/11	Emk3	Emapc4	Emch1	[a]M5
	EM_IND/12	Emk2	Emapc3	Emch3	M6
	EM_IND/15	Emk1	Emapc5	Emch2	M7
	EM_IND/16	Emk2	Emapc3	Emch1	[b]M1
	EM_IND/17	Emk1	Emapc1	Emch3	M8
	EM_IND/19	Emk3	Emapc4	Emch2	[a]M5
	EM_IND/21	Emk2	Emapc5	Emch1	[a]M9
	EM_IND/22	Emk2	Emapc3	Emch1	[b]M1
	EM_IND/26	Emk2	Emapc5	Emch1	[a]M9
IB/V	EM_IND/34	Emk1	Emapc6	Emch2	[a]M3
	EM_IND/36	Emk2	Emapc3	Emch1	[b]M1
	EM_IND/37	Emk3	Emapc4	Emch2	[a]M5
	EM_IND/39	Emk1	Emapc3	Emch1	M10
	EM_IND/40	Emk4	Emapc6	Emch1	M11
	EM_IND/47	Emk1	Emapc1	Emch1	[b]M12
	EM_IND/48	Emk3	Emapc1	Emch1	M13
	EM_IND/49	Emk1	Emapc1	Emch1	[b]M12

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

	EM_IND/50	Emk5	Emapc6	Emch1	M14
	EM_IND/51	Emk2	Emapc6	Emch2	M15
IEH	EM_IND/57	Emk1	Emapc1	Emch3	[a]M8
	EM_IND/60	Emk2	Emapc1	Emch3	M16
	EM_IND/61	Emk3	Emapc4	Emch1	[a]M5
IOEP	EM_IND/63	Emk3	Emapc2	Emch2	[a]M17
	EM_IND/64	Emk1	Emapc1	Emch3	[b]M8
	EM_IND/65	Emk1	Emapc1	Emch1	[b]M12
	EM_IND/67	Emk6	Emapc2	Emch2	[a]M17
	EM_IND/68	Emk1	Emapc6	Emch2	[a]M3

[a] multiple occurrences but not statistically significant.

[b] Statistically associated with the Sole D group, p=0.0394 (X^2=4.244).

Table 4. Different genetic diversity indices of *E. moshkovskii* population based on three target loci.

	Haplotype details		**Nucleotide details**		**Number of segregating sites (S)**	**Tajima's D**	***LD* (|*D′*|)**	**Rm**
	Number of haplotypes	*Hd*	*K*	*π*				
Kerp1	5	0.572	2.59	0.0048	9	0.51	Y=1.0000 + 0.000X	0
apc	6	0.833	3.97	0.0097	10	1.90	Y=0.9779 - 0.1659X	1
Chitinase	3	0.589	1.08	0.0035	3	1.06	Y=1.0000 + 0.000X	0

Assessment of intragenic recombination revealed that apc exhibited one recombination event (Rms), and LD was incomplete in this loci, whereas the remaining two markers (with complete LD) did not have Rms.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Table 5. Different genetic diversity indices of *E. moshkovskii* population based on using concatenated multilocus sequences.

	Haplotype details		**Nucleotide details**		**Number of segregating sites (S)**	**Tajima's D**	***LD* (|*D′*|)**	**Rm**
	Number of haplotypes	*Hd*	*K*	*π*				
Kerp1+apc+ Chitinase	17	0.93	7.65	0.0060	22	1.42	Y=0.8169 + 0.1059X	1

Table 6. Different genetic diversity indices, interlocus linkage disequilibrium (LD) and recombination analyses of *E. moshkovskii* population using concatenated multilocus sequences.

		Haplotype details		**Nucleotide details**		**Number of segregating sites (S)**	**Tajima's D**	***LD (|D′|)***	**Rm**
		Number of haplotypes	*Hd*	*K*	*π*				
Including apc	**apc+ Chitinase**	11	0.90	3.97	0.0071	13	1.87	Y=0.9708 - 0.4183X	2
	apc+ Kerp1	13	0.89	6.56	0.0069	19	1.37	Y=1.067 - 0.5529X	2
Excluding apc	**Chitinase+ Kerp1**	9	0.83	3.67	0.0043	12	0.776	Y=0.7682 + 0.4469X	1

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press

Table 7. Interlocus LD and recombination investigations within the *E. moshkovskii* population across various co-infection/sole infection groups, utilizing combined multilocus sequences excluding apc loci.

	Populations	No. of samples	No. of polymorphic sites analyzed	No. of pairwise comparisons	No. of significant pairwise comparisons (Fisher's exact test)	*LD* (|*D′*|)	Rm
Excluding apc	**All**	33	12	66	12	Y=0.7682 - 0.4469X	1
	Sole D + IBV	25	10	45	11	Y=0.7121 - 0.5057X	1
	Sole D + IEH	18	8	28	11	Y=0.7423 - 0.8854X	1
	Sole D + IOEP	20	10	45	21	Y=0.8196 - 0.4000X	1
	IBV + IEH	13	10	45	11	Y=0.6397 - 0.4699X	1
	IBV + IOEP	15	12	66	11	Y=0.8170 - 0.2243X	1
	IEH + IOEP	8	10	45	1	Y=1.0000 - 0.0000X	0

Table 8. Interlocus LD and recombination investigations within *E. moshkovskii* population across various co-infection/sole infection groups, utilizing combined multilocus sequences after inclusion of apc loci.

	Populations	No. of samples	No. of polymorphic sites analyzed	No. of pairwise comparisons	No. of significant pairwise comparisons (Fisher's exact test)	*LD* ($|D'|$)	Rm
Including apc	**All**	33	22	231	51	Y=0.8169 + 0.0398X	3
	Sole D + IBV	25	20	190	40	Y=0.8280 + 0.0717X	3
	Sole D + IEH	18	18	153	46	Y=0.9681 - 0.6813X	3
	Sole D + IOEP	20	20	190	60	Y=0.8645 + 0.0426X	3
	IBV + IEH	13	20	190	28	Y=0.8253 - 0.1884X	2
	IBV + IOEP	15	22	231	38	Y=0.8806 + 0.1126X	2
	IEH + IOEP	8	19	171	16	Y=1.0000 - 0.0000X	0

A clear observation from the analyses of both intragenic and interlocus LD was that genetic recombination predominantly took place at the apc locus.

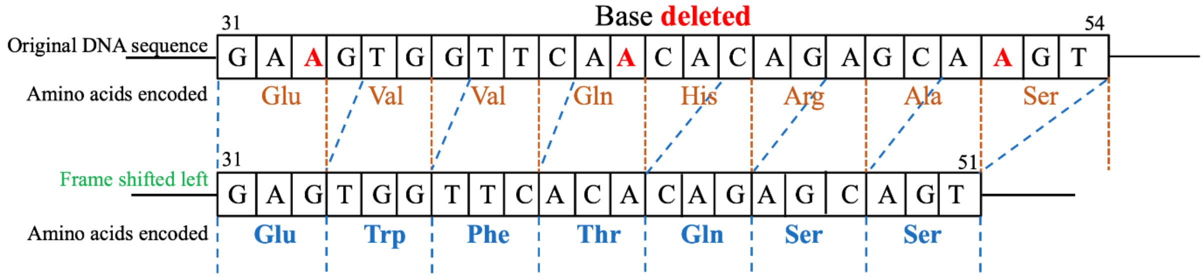
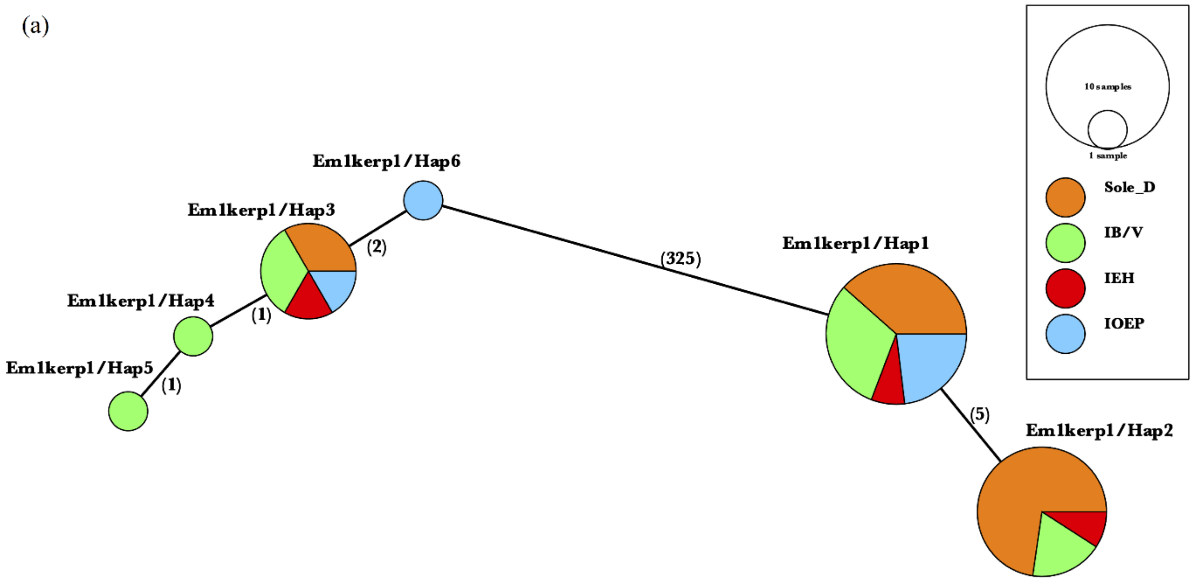

Base deleted
31
54
Original DNA sequence
G A A G T G G T T C A A C A C A G A G C A A G T
Amino acids encoded
Glu Val Val Gln His Arg Ala Ser
31
51
Frame shifted left
G A G T G G T T C A C A C A G A G C A G T
Amino acids encoded
Glu Trp Phe Thr Gln Ser Ser

Figure 1. Illustration of the frameshift mutation identified in the kerp1 locus of *E. moshkovskii* in Eastern India. Deletion of three adenine (A) bases occurred at positions 33, 42, and 52 in the original DNA sequence. As a result, the amino acid sequence -Glu-Val-Val-Gln-His-Arg-Ala-Ser- was replaced with -Glu-Trp-Phe-Thr-Gln-Ser-Ser-, leading to the loss of one amino acid in the altered sequence.



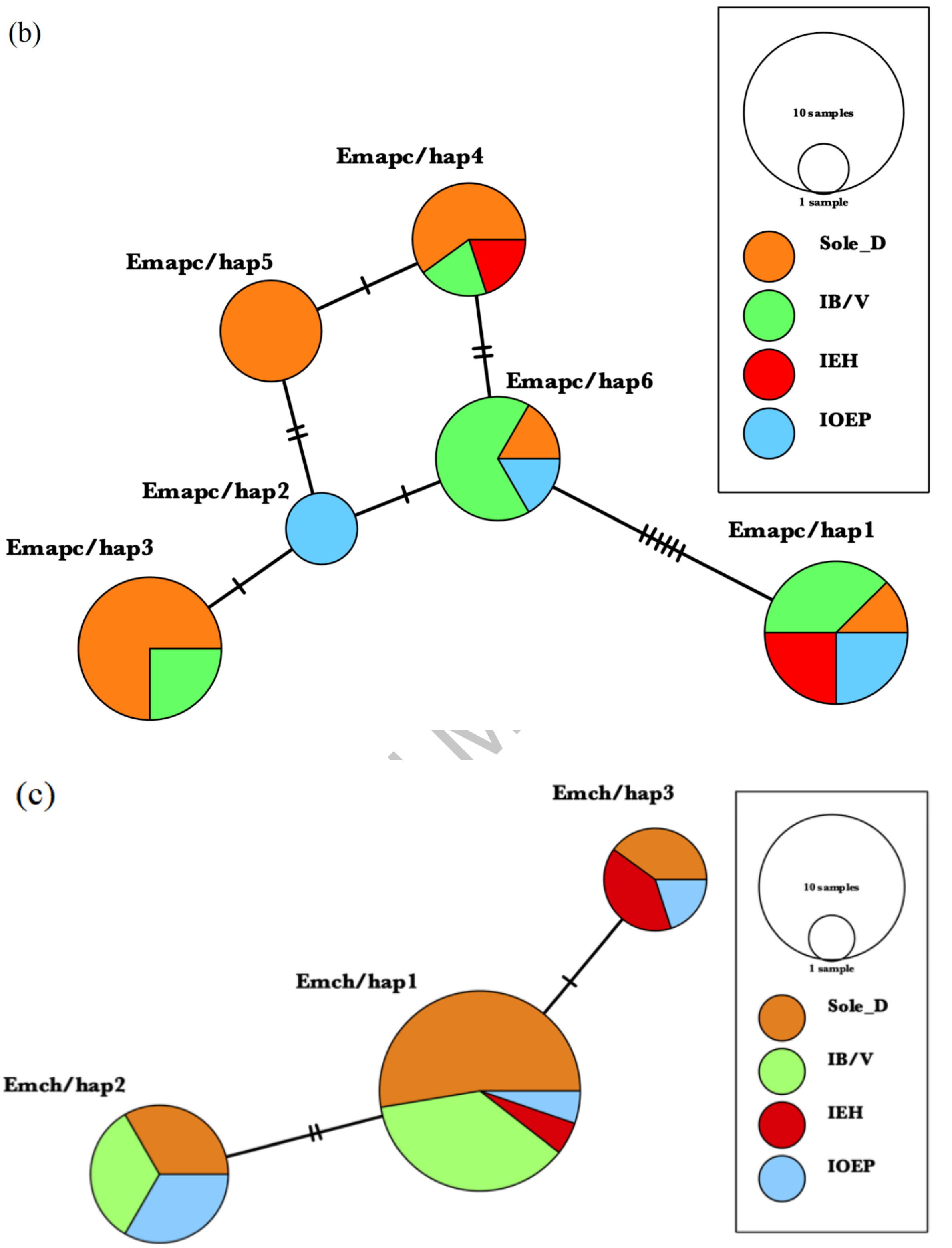


Figure 2. Minimal spanning haplotype network of *E. moshkovskii* haplotypes obtained from individuals with different co-infected sub-groups. The network illustrates the genetic relationships between the haplotypes, with each circle representing a unique haplotype and the

size of the circle indicating its frequency. The colours of the circles correspond to the coinfection status. (a) Lysine and glutamic acid-rich protein 1 (KERP1); (b) Amoebapore C (pore-forming peptides); (c) Chitinase.

https://doi.org/10.1017/S003118202400026X Published online by Cambridge University Press