Journal of Next Generation Sequencing & Applications

Alampalli et al., Next Generat Sequenc & Applic
2017, 4:3
DOI: 10.4172/2469-9853.1000152

Research Article Open Access

Genome Resequencing Revels Single Nucleotide Polymorphism and Repeat Regions in *Giardia lamblia* Indian Isolate

Shuba Varshini Alampalli[1], Akshay C Uttarkar[2], Suchithra Ventakesh[2], Sivarajan T Chettinar[1], Rishi Kumar Nageshan[2], Vidya Niranjan[2] and Utpal S Tatu[1*]

[1]*Department of Biochemistry, Indian Institute of Science, Bengaluru, Karnataka, India*

[2]*Department of Biotechnology, R V College of Engineering, Bengaluru, Karnataka, India*

[*]**Corresponding author:** Utpal S Tatu, Department of Biochemistry, Indian Institute of Science, Bengaluru, Karnataka, India, Tel: +080-22932823; E-mail: tatu@biochem.iisc.ernet.in

Rec date: September 11, 2017; **Acc date:** October 27, 2017; **Pub date:** October 30, 2017

Abstract

Giardia lamblia (syn. *Giardia duodenalis* and *Giardia intestinalis*) is a major human infecting intestinal parasite. It is a frequent cause of endemic and epidemic diarrhea. *G. lamblia* is divided into eight genotypes (A-H) which infect a wide range of mammals and humans, but human infections are caused by Genotypes A and B only. *Giardia lamblia* strains are speculated to undergo zoonotic transmission and also considerable genetic variation has been identified. To unambiguously determine the genotype of the Indian isolate, we sequenced the genome to high depth coverage and compared the assemblies with the nearly completed WB and DH genome and draft genome of Genotypes E (P15; pig isolate) and B (GS and GS-B; human isolate).

Our results identified the Indian isolate to be very closely related to isolate WB and ~97.71% of the sequencing reads aligned to the isolate WBC6 genome. This gave a list of SNPs, InDels and repeat motifs specific to the Indian isolate with the majority of the genome being conserved.

This work will layout a framework for the future studies. This study provides an opportunity to develop targeted and genome-wide genetic marker for better understanding of genotypes of *Giardia lamblia* prevalent in the developing world.

Keywords: *Giarida lamblia*; Genome sequencing; *De novo* assembly; Reference based assembly; Single nucleotide polymorphism

Abbreviations:

NGS: Next Generation Sequencing; SNP: Single Nucleotide Polymorphism; INDEL: Insertions/Deletions

Introduction

Giardia lamblia is causative agent of travelers' diarrhea which causes significant mortality and morbidity worldwide. The organism was initially thought to be primitive with less eukaryotic features like putative nucleolus [1], vestigial mitochondrial organelle also called as mitosome [2]. Giardia trophozoites have two nearly identical nuclei located anteriorly with respect to the long axis [3]. Both nuclei have equal number of DNA, approximately equal number of rDNA and are transcriptionally active [4]. Each nucleus has five chromosomes [5-8] inherited individually [3,9] and the genome size is estimated to be 11-12 Mb per haploid genome [7,10], and the chromosomes are monocentric [11], small, and identical to the size of chromomeres of higher eukaryotes [8]. *Giardia lamblia* only has four introns [7] and remnant of the essential organelles place this organism to be a simplistic in nature and entitled as "minimalistic organism" [7].

In 2000, only 9% of the *G. lamblia* genome was available and an urgent need to initiate the whole genome project for *G. lamblia* was required to understand the biology of Giardia species and define the phylogenetic relationship with other primitive organisms (PMID: 10731570). In 2007, the draft genome of *G. lamblia* assemblage A isolate WB was released that contained 90 scaffold which was later optically mapped to the physical map of all the five chromosomes (PMID: 24307482). In 2009, a draft sampling of *G. lamblia* assemblage B isolate GS was reported [12-14]. When scientist reclassified *G. lamblia* assemblages they observed many things that have yet remains unanswered in genomic prospective including genomic rearrangements that were genotype-specific, large differences in heterozygosity levels among isolates, population genetics, ascertain sexual stages, evidence of zoonotic transmission and species complexity among assemblages. Even with multiple groups sequencing the whole genome of *Giardia* yet the genome is incomplete [7]. This could be due to low complexity regions rich in AT or reoccurrence of repeat sequences that could be misguided while re-assembling existing genome. With many unanswered questions, there is a need to sequence many more *G. lamblia* clinical isolates to complete the genome. In this study, in order to enhance the opportunities for development of novel diagnostic and control strategies, whole genome re-sequencing of *G. lamblia* was undertaken which would support the development of targeted and genome-wide genetic markers such as single nucleotide polymorphisms (SNPs) and position enriched repeat motifs [15].

Methods

Cultivation of parasites

Indian isolate of *G. lamblia* parasites was cultured in TYI-S33 (Diamond, L. S., Harlow, D. R., and Cunnick, C. C. (1978) Trans. R.

Soc. Trop. Med. Hyg. 72, 431–432) supplemented with 12% fetal bovine serum and sub-cultured with 5 X 104 cells/tube from log phase parasites. The parasites were harvested by chilling on ice for 20 min followed by repeatedly inverting the tubes to dislodge the parasites and finally pelleted down at 700Xg for 5 min.

Sequencing and analysis

Whole genome sequencing was carried out with the paired end genomic sequencing methodology with 72-bp read lengths in the Illumina GA II X sequencer. More than 17 million reads were obtained, and 15.7 million high quality reads were used for alignment with the reference genome sequence (*G. lamblia* assemblage A, strain WB, version GiardiaDB-32). The genome coverage was calculated at 165X. The reads w ere used to *de novo* construct the genome using velvet 1.2.10 and mapped back to reference using MuMmer 3. The comparative studies were carried out with reference based assembly using PGA and features were analyzed using QUAST 4.5. MEGA6 was used to create phylogenetic trees. SNP and INDEL analysis was carried out in SnpEff v 4.1c. tabulating the results was carried out in R (Table 1).

Feature	*de novo* assembly	Reference based assembly
Tools	Velvet and bowtie 2.0	PGA
Number of contigs	5229	382
N50	31368	33697
Misassembled	5	3
Unaligned consigns	2+3 part	0+1 part
Indwells	14	4

Table 1: Comparing the results obtained by two dynamic methods to analyze NGS data.

Dot plot to predict repeat regions in the genome

300 nucleotides upstream of the genes and 300 nucleotides downstream of the genes of *Giardia lamblia* assemblage A isolate WB were extracted using R scripts. These reference sequences were used as the query in BLASTn to extract the upstream and downstream sequences of the *de novo* assembled genome of our isolate. The blast results were further used for repeat region analysis and motif prediction. YASS was used to plot the upstream sequences against the downstream sequences. MS Excel and R were used to analyze the sequences overlapping in either of the regions.

Gene ID	Protein Product	Scaffold	Type of Mutation	Mutation Description	Predicted Impact of the Mutation	Effect of the Mutation
GL50803_2902	Piwi proteins	CH991769.1	SNP	Nucleotide=166C>T, Protein=Gln56*	High	Stop codon gained in the transcript
GL50803_91451	Kinase, NEK	CH991763.1	Deletion	Nucleotide=801delA, Protein=Leu267fs	High	Frameshift variant leading to loss of function
GL50803_221689	ABC transporter	CH991767.1	SNP	Nucleotide=1958G>A, Protein=Cys653Tyr	Moderate	Missense variant leading to change in amino acid sequence
GL50803_41212	Hypothetical Protein	CH991767.1	SNP	Nucleotide=3826G>A, Protein=Ala1276Thr	Moderate	Missense variant leading to change in amino acid sequence
GL50803_137604	Variant-Specific Surface Protein 174	CH991767.1	SNP	Nucleotide=1696A>G, Protein=Thr566Ala	Moderate	Missense variant leading to change in amino acid sequence
GL50803_16910	Hypothetical Protein	CH991767.1	SNP	Nucleotide=779C>T, Protein=Ala260Val	Moderate	Missense variant leading to change in amino acid sequence
GL50803_96732	Hypothetical Protein	CH991779.1	SNP	Nucleotide=3115C>T, Protein=Arg1039Trp	Moderate	Missense variant leading to change in amino acid sequence
GL50803_113416	High cysteine membrane protein TMK-like	CH991763.1	SNP	Nucleotide=7160C>T, Protein=Pro2387Leu	Moderate	Missense variant leading to change in amino acid sequence
GL50803_114495	Kinase, NEK	CH991782.1	Insertion	Nucleotide=1084_1085insCTCTGA, Protein=Gln361_Ser362dup	Moderate	In-frame insertion leading to an extra amino acid, might affect the protein effectiveness

GL50803_114674	Hypothetical Protein	CH99176 7.1	SNP	Nucleotide=414C>T, Protein=Ser138Ser	Low	Synonymous variant
GL50803_21048	Hypothetical Protein	CH99176 7.1	SNP	Nucleotide=7848A>G, Protein=Ala2616Ala	Low	Synonymous variant
GL50803_103454	High cysteine membrane protein Group 1	CH99181 4.1	SNP	Nucleotide=207G>A, Protein=Leu69Leu	Low	Synonymous variant
GL50803_8983	Protein 21.1	CH99178 2.1	SNP	Nucleotide=2241T>C, Protein=Gly747Gly	Low	Synonymous variant
GL50803_101765	Variant-Specific Surface Protein 116 (VSP-116)	CH99176 7.1	SNP	Nucleotide=1854C>T, Protein=Ser618Ser	Low	Synonymous variant

Table 2: Overview of the SNP/InDels analysis.

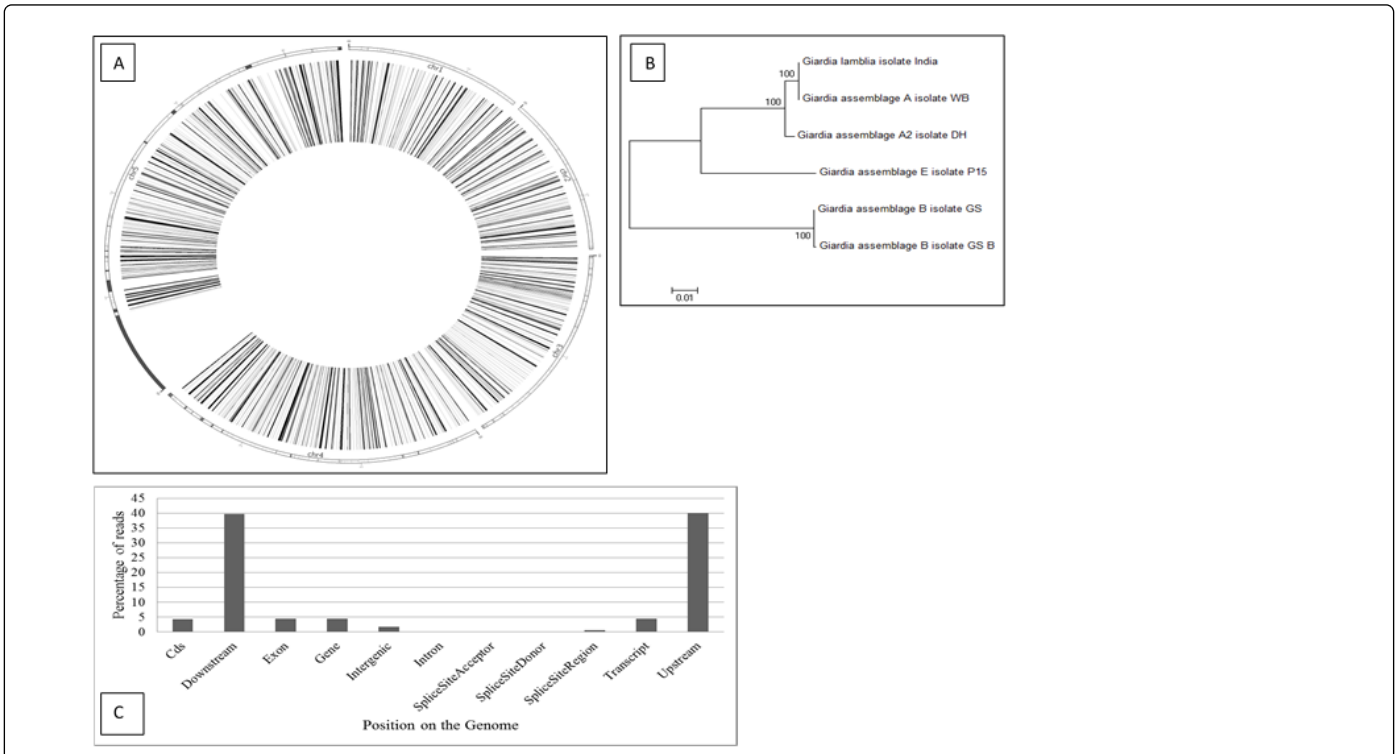

Figure 1: A) The circos plot is the *de novo* assembly of the lab strain mapped onto *G. lamblia* assemblage A strain WB genome. The outer circle is *G. lamblia* assemblage A strain WB chromosomes depicted as individual arcs of the circle and the bands inside the arcs are the genes loci in each chromosome. The black radiations are the regions mapped from the lab strain assembly. B) Phylogenetic analysis of the lab strain. From the *de novo* assembly of the lab strain, genes sequences of beta-giardin, NADP-dependent Glutamate dehydrogenase and Triosephosphate isomerase cytosolic were extracted and phylogenetic tree was constructed using MEGA6. With 100% bootstrap evidence, we can say the lab strain is very much closer to *G. lamblia* assemblage A strain WB. C) Distribution of the sequencing reads onto the reference genome. The lab strain whole genome sequencing reads were mapped onto *G. lamblia* assemblage A strain WB genome using Bowtie2. 97.71% of the reads aligned to the reference genome with most of the reads distributed in the upstream and downstream regions of the genome. This is possible as sequencing is a random process and also depends on the library preparations.

Motif search and enrichment

To remove the ambiguity of the gene sequence, motif analysis was carried out with 300 nucleotides upstream and downstream of the gene region (extracted as mentioned above). Tools from MEME suite were used to predict motifs. DREME was used to predict recurring, short and ungapped motifs that are relatively enriched in the upstream and

downstream regions respectively. The DREME output was used to compare the predicted motifs against JASPAR CORE 2014 database using TOMTOM. TOMTOM provides a list of similar motif/ transcription factors with Evalue>1. CentriMo identifies the predicted motifs that show significant preference for particular location in the upstream and downstream sequences respectively, thereby providing information regarding the local enrichment of each predicted motif (Table 2).

Results

After the QC of genomic sequencing reads, the high-quality reads were used to *de novo* assemble a genome (without reference genome) specific to *G. lamblia* Indian isolate. This step was important so as to extract specific gene loci (bg, tpi and gdh gene loci from the assembly and compare it with the existing *Giardia lamblia* assemblages (A, B, or E) and characterize the Indian isolate into a specific assemblage. *De novo* assembly of the reads resulted in 5229 contigs with N50 of 31,368 bases whereas the reference based assembly gave 382 contigs with N50 of 3369 (Figure 1A).

In order to classify the Indian isolate, phylogenetic analysis of bg, tpi and gdh gene loci were carried out using MEGA6. Figure 1B shows that the Indian isolate is closely related to *G. lamblia* assemblage A strain WB with 100% bootstrap support. This is held by ~98.71% of the reads of the Indian isolate that aligned to the reference genome, *G. lamblia* assemblage A strain WB (Figure 1C). The reference genome has large number of gaps in its genome.

The reference genome, *Giardia lamblia* assemblage A stain WB (GiardiaDB-32), is 12.827 Mb consisting 42.973% GC and 10% Ns. The genome is distributed into 5 chromosomes of 1.49 Mb, 1.52 Mb, 1.96 Mb, 2.76 Mb and 4.47 Mb. In addition to these five major chromosomes, variable minor (in copy number) or accessory chromosomes are present (doi:10.1016/j.pt.2010.07.002).

These accessory chromosomes appear to be duplications (or partial duplications) of major chromosomes, in some cases carrying long ribosomal RNA (reran) gene arrays. Chromosome 5 is annotated to transcribe maximum number of genes encoded by the organism. The *de novo* assembled scaffolds were compared onto the reference genome using Blast with E-value less than e-10. The BLAST results were plotted into a circus plot (Figure 1A) where in the reference genome with its genes as bands (outer circle) overlap with black ideograms (inner circle) that represent the regions covered by the *de novo* assembly.

The white ideograms in the circus plots might indicate repetitive regions as BLAST considers them as low intensity regions and thus reducing the E-value or they could be possible gaps introduced to make up the chromosome size. All the chromosomes seem to be partially covered by the *de novo* assembly due to large gaps filled with Ns. These gaps have been incorporated with evidence from PFGE analysis and optical mapping of the genome.

In order to retrieve the gapped sequences in chromosome 5 we carried out genome PCR walking for *G. lamblia* Indian isolate. The sequence data obtained from analyzing PCR products reviled sequence similarities between different regions of chromosome 5 and chromosome 2 (data not shown). We further analyzed to check whether there were many repeat regions globally in the *Giardia* genome.

Hence, we extracted 300 nucleotides upstream and downstream regions of the genes annotated in GiardiaDB-32 (as explained in Materials and Methods section) and aligned these regions to map similar regions with E-value greater than e-05. The dot plot (Figure 2) shows many regions away from the diagonal to be similar to both upstream and downstream regions of the genes.

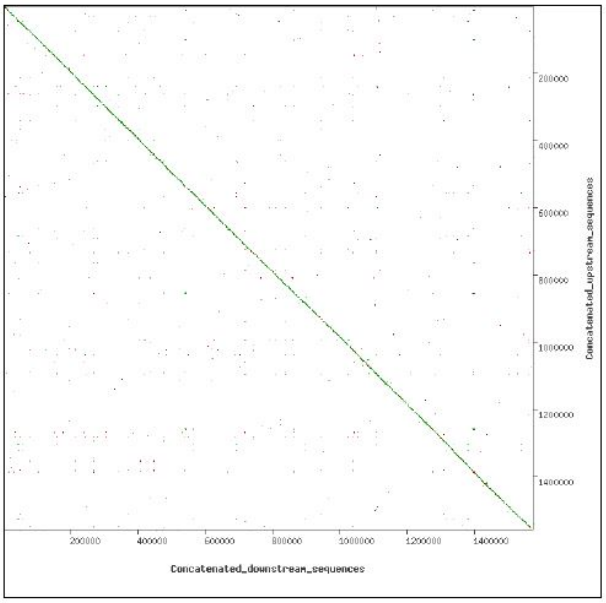

Figure 2: Dot plot of upstream and downstream regions.

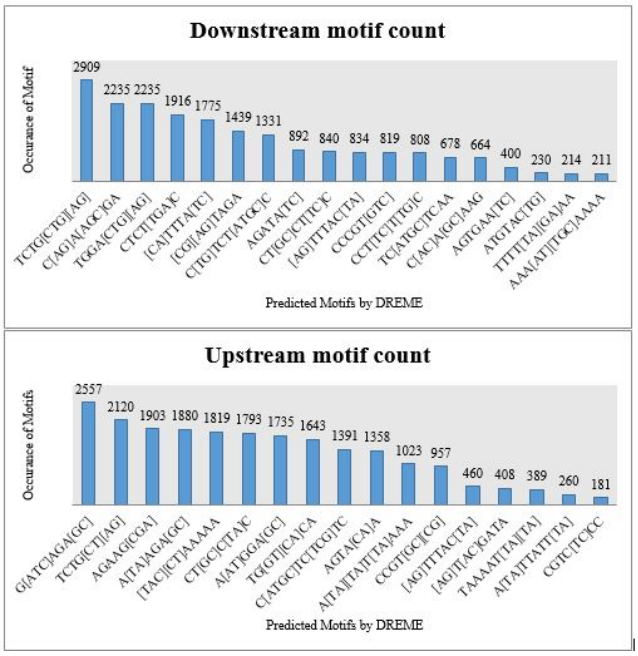

Figure 3: Motif distribution in upstream and downstream region of the Indian isolate genome.

This could be indicative of possible repeat sequences in the whole genome of the Indian isolate. Hence Motif search was carried out for

300 nucleotides upstream and downstream region of each gene. The gene sequences were not considered so as to remove ambiguity of the motif search. The search mostly reported the frequently occurring nucleotide repeats. Figure 3 shows the number of times each motif occurs in the specified regions.

Few of these motifs were similar to eukaryotic transcription factors and the others were specific to *Giardia lamblia*. In order to refine the large number of motifs, a local enrichment of the motifs was carried out using Centime. We found Indian isolate specific motifs that show significant preference for particular location in the upstream and downstream sequences respectively (Figure 4).

To avoid the haziness of the large gaps in the chromosomes, we extracted the 92-scaffold genome of *G. lamblia* ATCC 50803 (isolate WBC6) from NCBI. This 92-scaffold genome was used to analyze the SNPs and INDELs in WBC6 as phylogenetically the Indian isolate is very close to *G. lamblia* assemblage an isolate WB (Figure 1B). Overall 367 SNPs, 13 insertions and 12 deletions were predicted using SnpEff v 4.1c.

The transversions to transition ratio is about 4.4 (Tv/Ts ratio of observed SNPs) and missense to silent mutation ratio is 1.2. Across the entire genome the ratio of transitions to transversions is typically around 2. In protein coding regions, this ratio is typically higher, often a little above 3. The higher ratio occurs because, especially when they occur in the third base of a codon, transversions are much more likely to change the encoded amino acid. With *G. lamblia* being a small genome of 12.827 Mb and most of the SNPs reside in the non-coding region of the proteins, the Tv/Ts ratio (raw ratio) is expected to be high while considering only the protein coding regions.

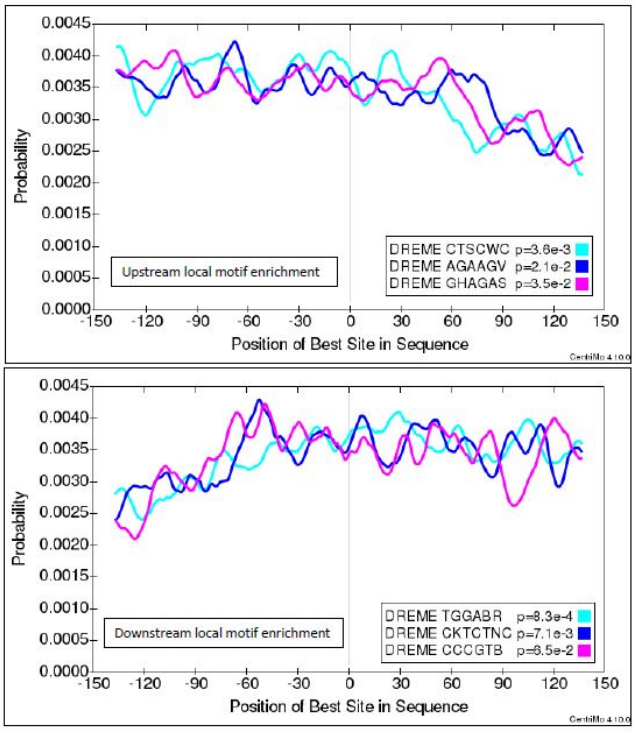

Figure 4: Motif enrichment by CentriMo.

Discussion

Two high impact factor variants were observed in genes coding for Piwi proteins on scaffold CH991769.1 (GL50803_2902) and NEF kinases on scaffold CH991763.1 (GL50803_91451). The SNP in Piwi protein genes leads to a stop codon; further evaluation of the genome indicates there might be a wrong annotation of the gene in the database (Figure 5). *Giardia lamblia* contains countable introns; hence polymorphisms in introns are hardly seen. The effect of polymorphisms is majorly seen if it occurs in the coding regions of the genome.

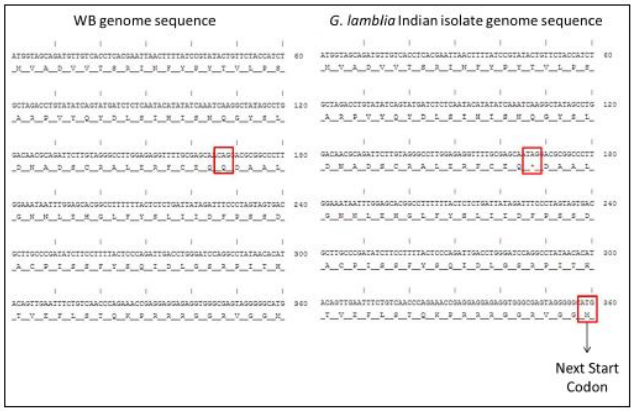

Figure 5: SNP in Piwi protein.

In our analysis we did find SNPs in exons (0.638%) but most of them did not affect the predicted protein function. Do these SNPs have a structural variation of the genome? It's hard to predict as the genome is incomplete. 99.178% of the SNP/InDels were predicted to be non-coding variants or variants affecting non-coding genes (3'UTR, 5'UTR, upstream or downstream of a gene) with no evidence of impact (Figure 6). This suggests that there may be a biological mechanism for maintaining genome fidelity and reducing heterozygosity between the four genome copies (tetraploid) [7].

A decade of improvement and technological advancements in genomics has led to faster improvements in parasitology with sequence assemblies becoming available for multiple genera including *Babesia, Cryptosporidium, Eimeria, Giardia, Leishmania, Neospora, Plasmodium, Theileria, Toxoplasma* and *Trypanosoma*.

The first partial genome sequence of *G. lamblia* was reported in 2000 [10], and in 2007, the full genome sequence was published and is available on *Giardia* DB [7] (PMID: 21209094). In this study we report whole genome re-sequencing of *Giardia lamblia* isolated in India. *Giardia lamblia* being a major cause of human giardiasis, it's important that the gaps in the genome be filled to understand the genome architecture of this simple minimalistic organism, which would answer topics like: (i) the search for functional aspects of gene arrangement such as breakpoints in chromosome rearrangements involved in drug resistance, (ii) clues to understanding the function of the two nuclei, (iii) genetic variation and (iv) chromosome specific sequencing.

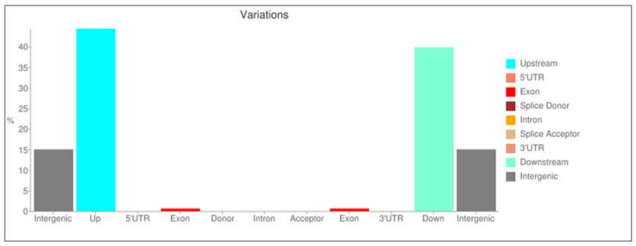

Figure 6: Percentage variation (SNP/INDEL) distributions in the *G. lamblia* ATCC 50806 isolate WBC6 genome. Overall 367 SNPs, 13 Insertions and 12 Deletions were predicted using SnpEff v 4.1c. The transversions to transition ratio is about 4.4 (Tv/Ts ratio of observed SNPs) and Missense to silent mutation ratio is 1.2. 99.178% of the effects were predicted to be non-coding variants or variants affecting non-coding genes (3'UTR, 5'UTR, upstream or downstream of a gene) with no predictable evidence of impact.

Conclusion

Trans splicing is a very unique kind of splicing activity that is being undergoing in this particular neglected human pathogen. Its challenging to understand and the so is the amount of knowledge about this trans spliced genes and their effects is still under investigation. NGS analysis of the Indian isolate and a comparative study and inference between two widely used approaches. Phylogeny identifies the position of Indian isolate with previously known isolates. Identification of the SNP's and INDEL's provide an insight for better understanding. This provides a lot of future scope for further studies to be carried out in this particular field of interest.

Acknowledgement

We are thankful for the grants from DBT-IISc partnership programme and acknowledge the support of Basic science grant from Department of Biotechnology, New Delhi, India.

Author's contribution

Conceiving of idea by UT. Methodology by VN. RKN performed the cultivation and harvesting of the parasites. SVA and STC performed *De novo* analysis, phylogeny and motif search. They also carried out SNP analysis and drafting of manuscript. ACU and SV performed comparative study by refining the *De Novo* assembly and perform reference based assembly, along with finalizing of manuscript. All authors read and approved the final manuscript.

References

1. Jiménez-García LF, Zavala G, Chávez-Munguía B, Ramos-Godínez Mdel P, López-Velázquez G, et al. (2008) Identification of nucleoli in the early branching protist *Giardia duodenalis*. Int J Parasitol 38: 1297-1304.
2. Tovar J, León-Avila G, Sánchez LB, Sutak R, Tachezy J, et al. (2003) Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation. Nature 426: 172-176.
3. Adam RD (2001) Biology of *Giardia lamblia*. Clin Microbiol Rev 14: 447-475.
4. Kabnick KS, Peattie DA (1990) *In situ* analyses reveal that the two nuclei of *Giardia lamblia* are equivalent. J Cell Sci 95: 353-360.
5. Adam RD (1991) The biology of *Giardia* spp. Microbiol rev 55: 706-732.
6. Le Blancq SM, Adam RD (1998) Structural basis of karyotype heterogeneity in *Giardia lamblia*. Mol Biochem Parasitol 97: 199-208.
7. Morrison HG, McArthur AG, Gillin FD (2007) Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science 317: 1921-1926.
8. Tůmová P, Uzlíková M, Wanner G, Nohýnková E (2015) Structural organization of very small chromosomes: Study on a single-celled evolutionary distant eukaryote *Giardia intestinalis*. Chromosoma 124: 81-94.
9. Bernander R, Palm JE, Svärd SG (2001) Genome ploidy in different stages of the *Giardia lamblia* life cycle. Cell Microbiol 3: 55-62.
10. Perry DA, Morrison HG, Adam RD (2011) Optical map of the genotype A1 WB C6 *Giardia lamblia* genome isolate. Mol Biochem Parasitol 180: 112-114.
11. Dawson SC, Sagolla MS, Cande WZ (2007) The cenH3 histone variant defines centromeres in Giardia intestinalis. Chromosoma 116: 175-184.
12. Adam RD (2000) The *Giardia Lamblia* genome. Int J Parasitol 30: 475-484.
13. Adam RD, Dahlstrom EW, Martens CA, Bruno DP, Barbian KD, et al. (2013) Genome sequencing of *Giardia lamblia* genotypes A2 and B isolates (DH and GS) and comparative analysis with the genomes of genotypes A1 and E (WB and Pig). Genome Biol Evol 5: 2498-2511.
14. Franzén O, Jerlström-Hultqvist J, Castro E, Sherwood E, Ankarklev J, et al (2009) Draft genome sequencing of *Giardia intestinalis* assemblage B isolate GS: Is human Giardiasis caused by two different species? Plos Pathog 5: e1000560.
15. Upcroft JA, Krauer KG, Upcroft P (2010) Chromosome sequence maps of the *Giardia lamblia* assemblage A isolate WB. Trends Parasitol 26: 484-491.