International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Genome Analysis of WGS of Leishmania Major

EL Bakri Fatima Zahrae[1], Azeddine Ibrahimi[2], Driss belghyti[3], Meryam Lemrani[4]

[1,2]Med Biotech Laboratory, Medical and pharmacy school of Rabat, University Mohamed V
[1,3]Department of Durable Development and natural resources, faculty of science, University of Ibn Tofail
[4]Department of Transmissible Disease, Pasteur Institute of Casablanca, Morocco August 17, 2024

Abstract
Leishmaniasis represent a major public health problem, because of the number of cases recorded each year and the wide distribution of the disease. it is a parasitic disease of flagellated protozoa trans- mitted by the bite of certain species of sandfly, causing a spectrum of clinical pathology in humans ranging from disfiguring skin lesions to fatal vis- ceral leishmaniasis. Cutaneous leishmaniasis due to *Leishmania major* is a polymorphic disease, in fact the infection can be asymptomatic, localized, or disseminated. The objective of this work is to determine the genomic diversity that contributes to clinical variability, by trying to identify the varia- tion in chromosome number and to extract SNPs and SNPs and InDels, it is based on four sequences (WGS) of *Leishmania major* available on NCBI in Fastq form, from three countries: Tunisia, Algeria and Israel, the analysis is set up from a pipeline to facilitate the discovery of genetic diversity, in particular SNP and chromosomal somy.

Keywords: Leishmania Major, Cutaneous Leish- Maniasis, NGS, Bio-Informatique, Somy, Variant-Calling

1. Introduction

Leishmaniasis is a vector-borne parasitic disease, transmitted by the hematophagous females of small dipters (phlebotomus). They are caused by a number of species of the Leishmania genus. Clinical manifestations are broad-spectrum, ranging from self-curative lesions to gross disfigurement and po- tentially fatal visceral disease (1). Broadly speak- ing, there are three clinical syndromes: visceral leishmaniasis (VL), cutaneous leishmaniasis (CL) and mucocutaneous leishmaniasis (MCL). (2) Cur- rently, according to the World Health Organiza- tion (WHO), there are around 12 million cases of leishmaniasis worldwide, with between 50,000 and 90,000 new cases of VL and between 600,000 and 1 million new cases of CL occurring each year (WHO). The most common clinical form is LC, which is often considered as a group of diseases due to the varied spectrum of clinical manifestations, ranging from small cutaneous nodules to macro- scopic destruction of mucosal tissues. The wide spectrum of clinical manifestations of LC may be explained by the fact that it is caused by several species of Leishmania, which are transmitted to hu- man and mammalian hosts by species of sandfly vectors. (3) The virulence of Leishmania species is one of the determining factors in the long-term outcome of infection. Human interaction with der- motropic Leishmania ranges from asymptomatic to severe cutaneous leishmaniasis, depending on the genetic diversity of Leishmania species (4). The virulence of strains from different

endemic regions poses a problem for monitoring cases diagnosed worldwide. This is because the presence of a two- loop life cycle requires adaptation mechanisms to different environments. (5,6). High-throughput sequencing, also known as Next Generation Sequenc- ing (NGS), enables the whole genome to be se- quenced (WGS: Whole Genome Sequencing). NGS enables the identification of genetic variants re- sponsible for clinical characteristics relating to the biology of leishmaniasis (7,8)

2. Materials and methods

NGS technology has become an indispensable tool for Leishmania researchers. Recent genomic analyses of Leishmania have facilitated discovery, ge- nomic diversity, including SNPs, CNVs, Somy variations and structural variations in detail and pro- vided valuable insights into genome complexity and gene regulation..

2.1 data base, leishmania sample

The National Center for Biotechnology Information (NCBI), is a Genome Browsers database that provides access to genome data from different species, in addition to sequence data.(9) The sequences analyzed in this study are in fastq format, available from NCBI , FASTQ is the text file format for storing biological sequences and associated qual- ity scores. Strain selection is based on information from the fastq data, which are specific to the study of Leishmania , they classify into country of iso- lation, year of collection, sequencer and sequencer data, and even parasite host. t

Strain	GC	Size	Country/Host
SRR6369642	57,9	338Mbp/152.1M	Algeria-Homo
ERR439247	55,5	2,9Gbp/1,3G	Israeil/Homo
SRR6369657	56,9	2,6Gbp/2,6	Tunisie/Homo
SRR6369659	57	4,3Gbp/2,8G	Tunisie/Homo

Table 1: details of strains studied

2.2 Read mapping to reference genome

The analysis of NGS data involves grouping a set of bioinformatics tools according to well-established efficiency criteria so that they form a ant discovery and genotyping. Local realignment around InDels allows us to correct mapping errors made by genome aligners and make read align- ments more consistent across regions. (10) There are two steps in the realignment process: The first is to identify those regions where alignments can potentially be improved and create a list of target intervals using GATK's RealignerTargetCreator tool. Inputs to this tool are the genome reference

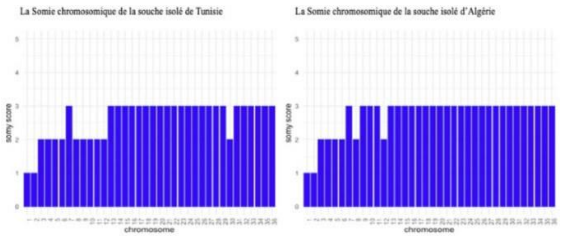

Figure 1: Chromosome number variation of strain isolated from Alg´erie et Tunisie

file and BAM file sorted by coordinate order. As for the second step, it realigns the reads in these

regions using a consensus model that takes all the reads in the alignment context together by GATK's IndelRealigner tool.(11,12,13)

Chromosome number variation of strain isolated from Alg´erie et Tunisie since this tool takes the BAM and the list of target intervals (output from the Realigner Target Creator tool) as input, it generates a realigned ver- sion of the input BAM file. The seventh step is to detect variants (SNPs, InDels and CNVs). Based on the evaluation of BAM files after their valida- tion, we deviated into two points: identifying vari- ants in terms of structure (SNPs, InDels), and in terms of chromosome number (CNVs). Identifica- tion of SNPs and InDels: to identify these two types of variants, we used GATK's HaplotypeCaller tool, which takes the BAM file and the reference genome file as input to produce the VCF (VariantCallFor- mat) file, a generic format for storing DNA poly- morphism data such as SNPs, insertions, deletions and structural variants. CNV identification: depth values were calculated using the Samtools tool. The input to this tool is the BAM file sorted in coor- dinate order. The output file has three columns: the first is the name of the contig or chromosome, the second is the position and the third is the num- ber of Reads aligned to that position. The eighth step aims to filter genetic variants using different criteria such as inheritance pattern, consequences of amino acid change, minor allele frequencies in human populations, splice site strength, conserva- tion, etc. This filtration will be carried out via the Variant Filtration tool, which takes two parame- ters as input: the reference genome file and the VCF file. This filtration will be carried out using GATK's Variant Filtration tool, which takes two parameters as input: the reference genome file and the VCF file, and the result of this tool are well-filtered VCF format files.

2.3 Snp filtring and Variant calling

The ninth step is dedicated to annotating the VCFs. SnpEff is an efficient annotation tool. It annotates and predicts the effects of genetic vari- ants (such as amino acid changes, non-synonymous changes, synonymous changes and inter-gene muta- tions. SNPs and InDels were compiled in a VCF file of population genetic variation). SnpEff is used in the following steps: - Edit the SnpEff configuration file. - Create a directory under "snpEff/data /". - Store the FASTA file and the GFF file in "snpEff / data / ". - Run the command in the terminal to create the database. - After building the database, the SnpEff tool takes as input: the VCF format file, the database and the configuration file. Out- put: an EFF.VCF format file used to store variant annotations. The R script The comparison of these variants is based on the development of a compara- tive analysis method using the R programming lan- guage, integrating different types of data (type of variant detected, position, annotation, etc.).

2.4 Chromosome number variation analysis

We studied aneuploidy by calculating Somy val- ues. Frequent aneuploidy is one of the main differ- ences between genome analyses of Leishmania and Trypanosomatidae. To calculate the Somy in each chromosome. We worked on the files containing the depth of each position, which we obtained via the Shell script (files containing the depth). For each chromosome, removing outliers by eliminating all higher positions (mean + 2 * standard deviation) and all lower positions (mean - 2 * standard devi- ation), we calculated the median Read depth (di), we calculated the median depth of the 36 chromo- somes (dm) and we obtained the Somy (value s) with the following formula: $s = 2 \times di / dm$.

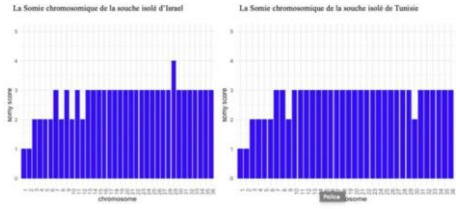

Figure 2: Chromosome number variation of strain isolated from Israeil et Tunisie.

3. Results

3.1 Chromosome copy number vari- ationl

The CNV study shows chromosomal copy number variation for the four Leishmania major strains analyzed. Analysis of the median read depth for the four L. major genomes shows estimated Somy val- ues for all 36 chromosomes**(**figure1 and 2)

Chromosome number variation of strain isolated from Israeil et Tunisie.

Chromosome 29 is the only chromosome with a consistently high copy variation of the disomic and trisomic forms present for the other chromo- somes, with tetrasomy manifested for the strain collected in Israel. A trisomic form of chromo- some 29 is present in strains collected from Alge- ria and Tunisia. The strain isolated in Algeria is predominantly trisomic, with six of the 36 chromo- somes disomic and the first two monosomic. The two strains isolated from Tunisia show copy vari- ation, the variance being evident for chromosomes 8,10,11,12, which are trisomic in one strain and di- somic in the other. Chromosome 30 is disomic in both Tunisian strains, while in Algeria and Israel it is trisomic. Both strains also show a trisomic form for chromosome 9 and 11. The four somy profiles show a diversity of chromosome copy number ex- pression, a mosaic of disomic and trisomic forms.

3.2 Nucleotide diversity

SNPs and InDels are annotated according to their position in terms of genomic region type and genes. The prediction of genomic variants (SNPs, InDels) in L.major strains (figure 9), shows information on the effect by type and region, as well as an assess- ment of the impact of the variant. According to the HTML report of the annotated files, Figure 10 shows the number of variants per type. The num- ber of SNPs is variable for the four strains of Leish- mania major, the strain collected in Israel has the highest number of all four strains **(**figure 3), for the number of Insertions, the strains from Tunisia has more than 18,561. The Algerian strain has a minimum number of InDels of 9,370. Variations in the number of SNPs, InDels and insertions are also detected in strains from the same country.

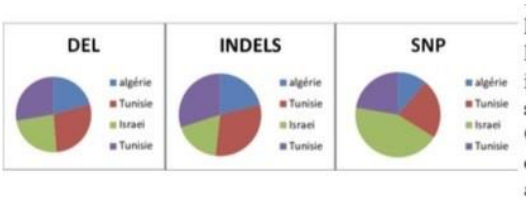

Figure 3: percentage of SNPs, Inserts and Indels of the four strains analyzed

4. Disccussion

The parasite undergoes genomic plasticity, selected by variation in the number of genetic copies (figure1,2). This structural genomic variation has been described by Imamura as a primary strategy for adapting to environmental change(9,10). The leish- mania genome displays ploidy dynamics, and the strains studied from three countries show a diver- sity of chromosomal somy expression. CNVs can

apply either to entire chromosomes, leading to ane- uploidy, or to specific genomic regions. For the latter, amplification of chromosomal regions oc- curs at direct or inverted homologous repeat se- quences, leading to extra-chromosomal amplified DNA. This ability of Leishmania to respond to drug pressure and stress situations (14,15). Bu- tenco's study shows the contraction of a family of oxygen-sensitive "adenylate cyclase" genes that manage cAMP O2-dependent signaling via protein A, essential for the cell survival and proliferation of Leishmania promastigotes under low oxygen con- centration, suggesting that this parasite relies on different mechanisms to cope with hypoxia where subjected to different environmental signals during development.(16) Dumetz et al also demonstrated the ability of leishmania to pre-adapt to different ssstress conditions. Mosaic aneuploidy is thought to provide a strong adaptive advantage for the whole population rather than the single cell (17). Somy variation, affecting almost all chromosomes, has been described as leading to heterozygosity (18,19). At the end of the annotation results of this study, an analysis of SNPs is recommended in order to ob- tain expressive and more meaningful results on the genetic and genomic diversity that contributes to phenotypic expression variability. Amal Ghouila et al in 2016 (20,21) are based on heterozygosity and homozygosity analysis of SNPs and InDels by al- lele frequency counting. Anzhelika Butenko et al in 2019 studied a comparative analysis of parasite species, the tool used for SNPs analysis is Gene Ontology(GO), which is a bioinformatics approach consists of exploiting variant data and managing annotations in order to intend to analyze genes and gene products, genetic variability "SNPs" revealed by analyzing gene families gaining, lost, expanded and contracted by identifying unique orthologous group to facilitate comparison. Based on the re- sults of this study, an increase in the number of genomes analyzed is recommended in order to ob- tain expressive and more meaningful results on the clinical variability of cutaneous leishmaniasis. The genomic variance of strain genomes studied by the various high-throughput sequencing (NGS) tech- niques needs to be monitored. We also need to develop new therapeutic targets for future drugs, as parasite eradication by antiparasitic treatment is often threatened by antibiotic resistance.

5. Conclusion

Leishmania major infection is a major public health problem. Cutaneous leishmaniasis presents a phenotypic diversity of clinical forms due to ge- nomic variability. The data presented suggest that the parasite exhibits intra-species somy variation, which is a structural variation leading to intra-chromosomal amplification is considered a mech- anism for modifying allele expression levels. Leish- mania virulence is probably influenced by the ex- pression of genes present in duplicated or tripli- cated form reported in evidence on adaptation and preadaptation to environmental change as well as response to treatment. These results should be backed up by a larger number of samples and functional expression studies.

References

6. Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, et al. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res. 2011;21: 2143–2156.doi:10.1101/gr.123430.111
7. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al. Comparative ge- nomic analysis of three Leishmania species that cause diverse human disease. Nat Genet. 2007;39: 839–847. doi:10.1038/ng2053

8. Samarasinghe SR, Samaranayake N, Kariyawasam UL, Siriwardana YD, Imamura H, Karunaweera ND. Genomic insights into virulence mechanisms of Leishmania dono- vani: evidence from an atypical strain. BMC Genomics. 2018;19: 843. doi:10.1186/s12864- 018-5271-z
9. Mahnaz T, Al-Jawabreh A, Kuhls K, Schönian G. Multilocus microsatellite typ- ing shows three different genetic clusters of Leishmania major in Iran. Microbes and in- fection / Institut Pasteur. 2011;13: 937–42. doi:10.1016/j.micinf.2011.05.005
10. Bañuls AL, Bastien P, Pomares C, Arevalo J, Fisa R, Hide M. Clinical pleiomorphism in human leishmaniases, with special mention of asymptomatic infection. Clin Microbiol In- fect. 2011;17: 1451–1461. doi:10.1111/j.1469- 0691.2011.03640.x
11. Stuart K, Brun R, Croft S, Fairlamb A, Gürtler RE, McKerrow J, et al. Kinetoplas- tids: related protozoan pathogens, different diseases. J Clin Invest. 2008;118: 1301–1310. doi:10.1172/JCI33945
12. Cantacessi C, Dantas-Torres F, Nolan MJ, Otranto D. The past, present, and fu- ture of Leishmania genomics and transcrip- tomics. Trends Parasitol. 2015;31: 100–108. doi:10.1016/j.pt.2014.12.012
13. Ivens AC, Peacock CS, Worthey EA, Mur- phy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, Leish- mania major. Science. 2005;309: 436–442. doi:10.1126/science.1112680
14. Leishmania major,
15. NCBI,https://www.ncbi.nlm.nih. gov/genome/?term=leishmania
16. Jacob.R et al, 2018 , Performance bench- marking of GATK3.8 and GATK4, https://doi.org/10.1101/348565
17. Imamura H, Dujardin J-C. A Guide to Next Generation Sequence Analysis of Leishma- nia Genomes. Methods Mol Biol. 2019;1971: 69–94. doi:10.1007/978-1-4939-9210-2-3
18. Anzhelika Butenko , Alexei Y. Kostygov
19. , Jovana Sadlova , Yuliya Kleschenko , Tomas Becvar , Lucie Podesvova , Diego H. Macedo , David Zihala , Julius Lukes , Paul

A. Bates , Petr Volf , Fred R.Opperdoes and Vyacheslav Yurchenk, Comparative genomics of Leishmania, (2019) 20:726 https://doi.org/10.1186/s12864-019-6126-y.

20. Kazemi B. Genomic Organization of Leishmania Species. Iran J Par- asitol. 2011;6: 1–18. Available: https://www.ncbi.nlm.nih.gov/pmc/articles
21. /PMC3279888/.
22. Barker, D.C., DNA Diagnosis of Hu- man Leishmaniasis. 1987, Parasitol. Today, https://doi.org/10.1016/0169- 4758(87)90174-8
23. Imamura H, Downing T, Van den Broeck F, Sanders MJ, Rijal S, Sundar S, et al. Evolu- tionary genomics of epidemic visceral leish- maniasis in the Indian subcontinent. Elife. 2016;5. doi:10.7554/eLife.12613
24. Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology. 2008;26: 1135–1145. doi:10.1038/nbt1486
25. Lighthall GK, Giannini SH. The chro- mosomes of Leishmania. Parasitology To- day. 1992;8: 192–199. doi:10.1016/0169-
26. 4758(92)90263-2
27. Bard, E., Molecular biology of Leishma- nia. Bioehem. Cell. Biol., 67, 516-524, 1989 Sep;67(9):516-24.doi: 10.1139/o89-083.

28. Dumetz F, Imamura H, Sanders M, Se- blova V, Myskova J, Pescher P, et al. Mod- ulation of Aneuploidy in Leishmania dono- vani during Adaptation to Different In Vitro and In Vivo Environments and Its Im- pact on Gene Expression. mBio. 2017;8. doi:10.1128/mBio.00599-17
29. Laffitte M-CN, Leprohon P, Papadopoulou B, Ouellette M. Plasticity of the Leishmania genome leading to gene copy number varia- tions and drug resistance. F1000Res. 2016;5. doi:10.12688/f1000research.9218.1
30. Ghouila A, Guerfali FZ, Atri C, Bali A, Attia H, Sghaier RM, et al. Comparative genomics of Tunisian Leishmania major isolates caus- ing human cutaneous leishmaniasis with con- trasting clinical severity. Infect Genet Evol. 2017;50: 110–120.