

Accepted Manuscript

Integrated genomic and metabolomic profiling of ISC1, an emerging *Leishmania donovani* population in the Indian subcontinent

Bart Cuypers, Maya Berg, Hideo Imamura, Franck Dumetz, Géraldine De Muylder, Malgorzata A. Domagalska, Suman Rijal, Narayan Raj Bhattarai, Ilse Maes, Mandy Sanders, James A. Cotton, Pieter Meysman, Kris Laukens, Jean-Claude Dujardin



PII: S1567-1348(18)30200-4
DOI: doi:[10.1016/j.meegid.2018.04.021](https://doi.org/10.1016/j.meegid.2018.04.021)
Reference: MEEGID 3487
To appear in: *Infection, Genetics and Evolution*
Received date: 5 December 2017
Revised date: 14 April 2018
Accepted date: 17 April 2018

Please cite this article as: Bart Cuypers, Maya Berg, Hideo Imamura, Franck Dumetz, Géraldine De Muylder, Malgorzata A. Domagalska, Suman Rijal, Narayan Raj Bhattarai, Ilse Maes, Mandy Sanders, James A. Cotton, Pieter Meysman, Kris Laukens, Jean-Claude Dujardin , Integrated genomic and metabolomic profiling of ISC1, an emerging *Leishmania donovani* population in the Indian subcontinent. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. *Meegid*(2018), doi:[10.1016/j.meegid.2018.04.021](https://doi.org/10.1016/j.meegid.2018.04.021)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Integrated genomic and metabolomic profiling of ISC1, an emerging *Leishmania donovani* population in the Indian subcontinent

Bart Cuypers^{1,2}, Maya Berg^{1,#}, Hideo Imamura¹, Franck Dumetz¹, Géraldine De Muylder¹, Malgorzata A. Domagalska¹, Suman Rijal³, Narayan Raj Bhattarai³, Ilse Maes¹, Mandy Sanders⁴, James A. Cotton⁴, Pieter Meysman², Kris Laukens^{2*} & Jean-Claude Dujardin^{1,5,*}

¹Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

²Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium

³BP Koirala Institute of Health Sciences, Dharan, Nepal

⁴Wellcome Trust Sanger Institute, Hinxton, United Kingdom

⁵Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium

Current address: Department of Pharmaceutical Sciences, University of Antwerp, Antwerp, Belgium

*Shared senior authors

Corresponding Author: Prof. dr. Jean-Claude Dujardin: jcdujardin@itg.be

Abstract

Leishmania donovani is the responsible agent for visceral leishmaniasis (VL) in the Indian subcontinent (ISC). The disease is lethal without treatment and causes 0.2 to 0.4 million cases each year. Recently, reports of VL in Nepalese hilly districts have increased as well as VL cases caused by *L. donovani* from the ISC1 genetic group, a new and emerging genotype. In this study, we perform for the first time an integrated, untargeted genomics and metabolomics approach to characterize ISC1, in comparison with the Core Group (CG), main population that drove the last outbreak of VL in the ISC. We show that the ISC1 population is very different from the CG, both at genome and metabolome levels. These differences included SNPs, CNV and small indels in genes coding for known virulence factors, immunogens and surface proteins. Both genomic and metabolic approaches highlighted dissimilarities related to membrane lipids, the nucleotide salvage pathway and the urea cycle in ISC1 versus CG. Many of these pathways and molecules are important for the interaction with the host/extracellular environment. Altogether, our data predict major functional differences in ISC1 parasites, including virulence. Therefore, particular attention is required to monitor the fate of this emerging population in the ISC, especially in a post-VL elimination context.

Keywords: *Leishmania donovani*, Visceral leishmaniasis, Genome diversity, Metabolomics, Genomics, Multi-omics integration, Virulence

Highlights:

- ISC1 is an emerging *Leishmania donovani* population in the Nepalese highlands
- The genome and metabolome of ISC1 are distinct from strains from the last visceral leishmaniasis outbreak in the Indian sub-continent
- These differences include important host interaction factors
- The emergence of ISC1 should be closely monitored in the Indian subcontinent

ACCEPTED MANUSCRIPT

Introduction

Visceral Leishmaniasis (VL) or kala-azar is a neglected tropical disease globally responsible for 0.2 to 0.4 million cases each year (1). The disease is lethal without treatment and caused by protozoan parasites of the *Leishmania donovani* complex. Transmission occurs by sand flies of the *Phlebotomus* genus. In the Indian Subcontinent (ISC), VL is a major concern for public health with more than 200 million people at risk (2). The recent whole-genome sequencing and phylogenomic analysis of 204 *Leishmania* strains from the ISC (India, Nepal and Bangladesh), revealed that the last VL epidemics in the region were essentially driven by one large *Leishmania* population that emerged around 1850 and spread across the lowlands in the Ganges plain. Given its recent origin, this population is genetically very homogeneous (3) and was called the Core Group (CG).

Before 2000, sporadic VL cases were reported in the hilly regions of Nepal (1km above sea level) and were considered as the result of travel, rather than a local transmission cycle. However, more recently reports of VL have been increasing in these Nepalese highlands and transmission has been shown to occur locally (4). Interestingly, some isolates derived from VL patients in this region belong to a genetically distinct population of *L. donovani* (called ISC1), which diverged much earlier than 1850 (3). Rai et al. (5) investigated the geographic spread of different *L. donovani* genotypes over a period of 12 years (2002-2014) in Nepal and showed that the frequency of ISC1 was increasing in the country and that this emerging genotype was also present in the lowlands. Our earlier study showed that out of all sampled ISC1 isolates, 75% of them came from patients relapsing to miltefosine treatment, vs 50% for the CG strains (3). We have also demonstrated experimentally that ISC1 parasites need more time and molecular changes to develop antimonial (Sb) resistance *in vitro* than CG strains, which led to the hypothesis of Sb-resistance pre-adaptation of the CG, but not of ICS1 (Dumetz et al., accepted, mSphere 2018 3: p. e00548-17-e00548-17).

Since 2005, the kala-azar elimination programme (KAEP) has been running in India, Nepal and Bangladesh, based on vector control and detection/treatment of VL cases. In such a programme, it is

essential to track parasites and in particular to characterize emerging populations like ISC1 or the parasites recently reported in Sri Lanka (6), as these could react differently to control tools like diagnostics, drugs or vaccines. In this context, we report here the results of a deep molecular comparison between ISC1 and CG parasites. In the present study, we focused on parallel untargeted genomics and metabolomics, hereby undertaking the first *Leishmania* diversity study integrating these two 'omics layers.

ACCEPTED MANUSCRIPT

Methods

Parasites

Eight *L. donovani* strains cloned from seven Nepalese and two Indian clinical isolates were analyzed in this study (**Supplementary Tables S1.1**). Among these are five ISC1 strains (BPK026/0, BPK604/0, BPK612/0, BPK519/12, BPK406/6) and two Nepalese (BPK275/0, BPK282/0) as well as two Indian (BHU568/0, BHU573/0) strains from the CG. Promastigotes of all strains were grown in modified Eagle's medium (Invitrogen) supplemented with 20% (v/v) heat inactivated fetal calf serum (PAA Laboratories GmbH, Linz, Austria) pH 7.5 at 26°C. At each passage, all parasite lines were inoculated to a starting concentration of $5 \cdot 10^5$ parasites/mL and growth was followed for two passages enabling synchronized sampling of the different cultures and minimizing the effect of differences in growth rate between strains. For each strain, four growth replicates were cultured for metabolomics analysis and one for genome sequencing. Metabolites were extracted in stationary growth phase according to t'Kindt et al. (2010) (7) and DNA with the QIAamp DNA Blood Mini Kit (Qiagen).

Genomics

Library preparation and sequencing

Library preparation and sequencing for BPK026, BPK604, BPK612, BPK512, BPK406 was performed at the Wellcome Trust Sanger Institute (Hinxton, United Kingdom). Genomic DNA was sheared into 400–600-base pair fragments by focused ultrasonication (Covaris Adaptive Focused Acoustics technology (AFA Inc., Woburn, USA)) and standard Illumina libraries were prepared. 125 base pair paired end reads were generated on the HiSeq 2000 v4 according to the manufacturer's standard sequencing protocol (8). Libraries of BPK275, BPK282, BHU568 and BHU573 were prepared and sequenced at the Beijing Genomics Institute (BGI). Briefly, mechanical shearing with the Covaris (Illumina) was used to create 250-350bp fragments, after which the fragment ends were repaired and 3' A-tailed. Adapters were subsequently ligated and the target fragments were recovered from a

2% agarose gel after electrophoresis. Finally, the libraries were PCR amplified and 2 x 151 BP sequenced on an HiSeq 4000. All sequencing data of this study was submitted to the Sequence Read Archive (SRA) under accession number SRP125482. Individual accession numbers for each sample are available in **Supplementary Table S2**.

Data analysis

Genomic analysis was performed as described in Dumetz et al. (9). Briefly, reads were mapped to the *L. donovani* LdBPKv2 reference genome (9) using Smalt v7.4 (www.sanger.ac.uk/resources/software/smalt/) and options exhaustive searching (-x), read identity threshold of 80% (-y 0.8) and random mapping of multiple hits reads were used. Duplicate reads were marked with Picard v1.85 and SNPs and small indels were called using the GATK haplotypcaller v3.4 (10). Low quality SNPs were marked using GATK Variant Filtration with parameters: "QD < 2.0 || MQ < 40 || FS > 60.0 || ReadPosRankSum < -8.0" and removed from further analysis. The same tool was used for filtering indels, but with parameters "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0 || SOR > 200". Variants were annotated with SnpEff v4.1 (11). SNPs and small indels were considered significantly different between CG and ISC1, when the allele shift difference was at least 0.25 (12) and Mann–Whitney U test *p*-value < 0.05. Allele shifts larger than 0.80 were considered homozygous variants (12). Local copy number variation detection and chromosomal somy was calculated according to Downing et al. (13). For somy estimation, initially, the median read depth of each chromosome was calculated (d_i). All positions with a read depth more than 1 standard deviation away from this initial median were then removed and d_i was recalculated. This approach removed depth outliers by assembly errors, local CNVs or spurious high coverage regions impacting the final median. Subsequently, the median depth of the 36 chromosomes d_m was calculated and the somy (*s*-value) of each chromosome was obtained with the following formula $s = 2 * d_i / d_m$. Haploid gene copy number was defined to be a median depth of a gene divided by a median depth of the chromosome which it is on. We used two criteria to evaluate whether a gene or chromosome copy number

difference between the CG and ISC1 strains was biologically meaningful and statistically significant. The first requirement was that the absolute difference in gene/chromosome copy number between CG and ISC1 should be at least 0.5 (9). Secondly, the FDR adjusted p -value (Student t -test + Benjamini Hochberg correction) had to be lower than 0.05.

Metabolomics

Liquid Chromatography and Mass Spectrometry

Samples were analyzed with LC-MS using an Orbitrap Exactive mass spectrometer (Thermo Fisher) coupled to a 2.1 mm ZIC-HILIC column at Glasgow Polyomics (University of Glasgow, Scotland) (Sequant), exactly as previously described (14). Both negative and positive ion modes were run in parallel with rapid polarity switching. Several quality control samples were run at the start, middle and end of the LC-MS run to aid accurate metabolite identification and verify LC-MS stability (15) and these included: 1) Authentic standard mixes containing in total 217 metabolites (50–400 Da) representing a wide array of metabolic classes and pathways. 2) An amino acid standard mix (Sigma Product No. A9906). 3) Serial dilutions (Undiluted, 1/2, 1/4, 1/8 and 1/16) of a pooled sample of all extracts to filter out spurious signals that do not follow the dilution trend (16).

Data Preprocessing

Data analysis was performed using the XCMS 1.42.0 (17) and mzMatch 1.0.1 (18) packages in R as described previously (19). In summary, mzXML files were first subjected to retention time alignment using ObiWarp (20) and peak detection with the centWave algorithm (21). Corresponding peaks from the 4 biological replicates were subsequently combined and filtered on a maximum allowed reproducibility standard deviation (RSD) of 0.5. Peaksets from all samples were then combined and filtered further, requiring for each peak a minimal scaled CoDA-DW quality value of 0.8 (mzMatch noise filter), detection in at least 3 out of 4 replicates and a minimal peak intensity of 3000. The gapfiller tool was used to fetch peaks from the mzXML files that were missing in one or more

samples and derivative peaks were annotated (22). The peaks were then putatively identified allowing 2 ppm mass deviation using sequentially the LeishCyc (23), LipidMAPS (24), KEGG (25), the peptide database included in mzMatch and the Human Metabolome Database (26). Each time, only unmatched peaks were aligned against the next database, reducing the number of potential identifications to the most likely candidates (27). Peaks of which the intensity did not follow the dilution pattern in the pooled sample dilution series (Pearson correlation lower > 0.8 or p -value > 0.05) were removed from the analysis with the mzMatch Dilution Trend Filter tool (16). The remaining peaks were normalized for total ion count and visually verified before exporting them to a csv file containing all peak heights, retention times and relevant database annotations. The full analysis script and detailed parameter settings are available in **Supplementary Table S5**. The csv files from the positive and negative ionization mode were then merged. When metabolites were identified in both modes, only the results from the most intense peak were kept.

Data exploration and statistics

For the initial data exploration, principal component analysis (PCA) was performed in R directly on the metabolite intensity values of each sample. Subsequently, the relative expression between the CG and ISC1 strains was calculated for each individual metabolite (28). Briefly, first the average metabolite intensity was calculated per strain. Then, the obtained values were again averaged over the ISC1 strains and over the CG strains. Finally, these strain averages were converted into base-2 logarithm ratios between the two populations, henceforth called the 'Log2 Fold Change' (Log2FC). A two-sample t -test assuming unequal variance was applied to the strain averages to calculate whether or not this Log2FC was significant ($p < 0.05$).

To investigate the overall correlation between genome and metabolome, we calculated three types of pairwise distances between all strains: 1) The SNP distance (S-distance) was defined as the total number of SNP allele changes between a strain pair. In this calculation, a heterozygous SNP was counted as a distance of 1, while a homozygous SNP was counted as a distance of 2. 2) The gene

dosage distance (GD-distance) was determined by taking the Euclidean distance between two strains based on all absolute gene copy differences; as such, it combined the effect of CNV and aneuploidy.

3) The metabolic distance (M-distance) was defined as the Euclidean distance based on all metabolite intensities, however, since these intensities have a large, metabolite-specific dynamic range, they were first transformed into Z-scores. Analysis of the correlation between the respective distances was performed with non-parametric Mantel test (29) with 999 permutations, using the QIIME2 package in Python. Contrary to the classical correlation test, the Mantel test deals with the problem that distances in a distance matrix are not independent from one other. R^2 values were determined by linear regression analysis with the `lm` function in R.

Results

Genome sequence diversity

Promastigotes of five ISC1 (BPK026/0, BPK604/0, BPK612/0, BPK519/12, BPK406/6) and four CG strains (BPK275/0, BPK282/0, BHU568/0, BHU573/0) were sequenced with an average read depth of at least 38.5X and 97.1 % of the genome was covered at least 20X. Detailed mapping to BPK282v2 (9) and coverage statistics can be found in **Supplementary Table S2**. Using new SNP data of our strains and published ones of East African *L. donovani* LV9 and Sri Lanka *L. donovani* SRI-CLB, we performed a phylogenomic analysis (**Figure 1**). In correspondence with Imamura et al. (2016) (3), this showed that (i) ISC1 parasites cluster at a larger distance from CG than the Sri Lanka strain SRI_CLB did and (ii) in an intermediate position with respect to East African *L. donovani* LV9.

In total, we detected 53,101 SNPs between the CG and ISC1 clusters (allele shift > 0.25 & *pval* < 0.05) of which 51,668 were homozygous (allele shift > 0.8) and of which 19,776 were located in protein coding regions (**Supplementary Table S3**). 10,550 of these SNPs caused non-synonymous codon changes and 24 were flagged by SnpEff for having a probable high impact on the final protein: fifteen premature stop codons were gained, seven stop codons were lost, and two start codons were gained in ISC1 strains compared to CG ones (**Supplementary Table S1.2**). Fourteen of these high impact variants concerned hypothetical proteins of which the function and impact is unclear. However, of the remaining ten genes (with homozygous SNPs), three play key roles in the *Leishmania* virulence (autophagy protein Atg8, trypanothione synthetase and amastin-like surfaceprotein) and two are involved in phosphatidylinositol metabolism (CDP-diacylglycerol-inositol 3-phosphatidyl transferase and phosphatidylinositol 3- and 4-kinase). While a high impact is predicted by SnpEff, these genes could certainly still be functional, as for many of them (including trypanothione synthase, autophagy protein Atg8 and the amastin-like surface protein) the SNP was found towards the end of the gene.

Also at the indel level major differences were observed between ISC1 and CG with a total of 6,435 heterozygous and 5,780 homozygous indels (allele shift > 0.25 and > 0.80 respectively, $pval < 0.05$, **Supplementary Table S4**). 442 of these indels were located in coding regions and 83 concerned frameshift mutations, which are often highly disruptive. A large proportion (58%) of the frameshift indels were located in hypothetical proteins, while the 33 functionally annotated indels (**Supplementary Table S1.3**) concerned several important biological functions, including, but not limited to, transmembrane transport (3), translation (2), nucleic acid binding (3) and nucleotide salvage (4). 21 of these annotated indels were homozygous mutations. Remarkably, two genes encoding for (i) AAAfamily ATPase and (ii) dihydrouridine synthase domain protein respectively each had two indels. In addition, phosphatidylinositol metabolism was highlighted in both SNP and INDEL analysis: Phosphatidylinositol-specific phospholipase C,X/Y/C2 domain containing protein (indel), CDP-diacylglycerol-inositol3-phosphatidyl transferase (SNP) and Phosphatidylinositol3-and4-kinase (SNP).

Genome structure diversity

The genome structure was first investigated at chromosome level (**Figure 2**). Chromosomal copy estimates do not always correspond to integer values as these represent the average of a population of cells that do not necessarily show identical karyotypes (mosaicism). Many copy differences were found between chromosomes of different strains, but only two of these were consistent between ISC1 and CG populations ($p\text{-val} < 0.05$ & $|\text{copy difference}| > 0.5$). The ISC1 strains had 0.8 copies less of chromosome 8 and 0.7 copies less of chromosome 23.

Secondly, we analysed local gene copy number variations (CNVs) between ISC1 and CG isolates. These differences are the result of gene amplification or deletion events since the last common ancestor of ISC1 and CG. For every gene and sample, copy numbers per haploid genome are available in **Supplementary Table S4**. By expressing copy number differences per haploid genome, we can describe these changes independently of the aneuploidy phenomenon. As *Leishmania* is diploid for

most chromosomes, we used 0.5 copies/haploid genome as a cut-off, which corresponds with a single copy for diploid chromosomes. In total we found 62 loci containing genes with CNVs: for 29 of these loci, genes were more abundant in ISC1, while for the 33 remaining loci, genes were more abundant in CG. 29 loci concerned single gene duplications or deletions (**Supplementary Table S1.4**). The remaining 33 loci contained multiple genes showing CNVs: (i) 22 contained tandem arrays of the same gene family, expanding or contracting (**Supplementary Table S1.5**); (ii) 11 loci contained sets of genes of a different family, all amplified/deleted together (**Supplementary Table S1.6**).

Duplication/deletions of single copy genes concerned a wide array of biological functions (**Supplementary Table S1.4**). Noteworthy, 43% of the genes with known functions were coding for surface proteins, virulence factors or immunogens, including GP63, amastins, 60S acidic ribosomal protein P2 and surface antigen 2. With respect to tandem arrays of the same gene family (**Supplementary Table S1.5**), a large proportion (37%) of the genes with known functions were also coding for surface proteins, virulence factors or immunogens. More specifically, we found CNVs of genes coding for amastins, proteophosphoglycans, HSP70, HSP83, GP63, p1/s1 nucleases, loricin proteins, cysteine peptidase B and surface antigen 2 proteins. Interestingly, some of these CNVs were linked to the observations at sequence diversity level. Indeed, frameshift indels were found in both LdBPK_300020700 coding for a p1/s1 nuclease, and in LdBPK_120014400 coding for a surface antigen protein 2. For both genes the adjacent loci contained genes of the same gene family involved in a CNV. Other families that were targeted by both sequence and structure variants were amastins and autophagy protein Atg8. Finally, amplifications concerning groups of genes with different functions (**Supplementary Table S1.6**) are extensively reported to occur as circular episomes (30-32), but also as intra-chromosomal amplicons (3,33). For instance, the H- and M-loci are amplified in CG strains but have only a single copy per haploid genome in ISC1. Noteworthy, the H-locus contains among others the gene encoding for arginine-succinate synthase (ASS, see metabolomics section). Reciprocally, the R-locus (34) has a single copy per haploid genome in CG strains, but has 2.15 copies per haploid genome in ISC1 strains. Both the H and R locus have been associated with drug resistance

(30,35). Five of the other amplified gene groups are associated with virulence, including proteophosphoglycan, amastins, surface antigen protein 2 and acid phosphatases. While all three described types of CNVs concerned a large proportion of genes coding for surface proteins, virulence factors and immunogens there was no clear directionality to these CNVs since we observed both deletions (60%) and duplications (40%) in ISC1 versus CG.

Metabolomics

In total, 290 metabolites were identified with a mass accuracy < 2 ppm, covering a broad diversity of metabolic pathways and representing the following classes (in order of quantitative representation): glycerophospholipids (GPLs), amino acids and derivatives, fatty acyls and carbohydrates, nucleobases and nucleosides, sphingolipids, steroids and derivatives (**Supplementary Table S5 Supplementary Table S1.4**). As a first data exploration of the variation in our dataset, we performed Principal Component Analysis (PCA, **Figure 3**). The first principal component explained 34% of the total variation in our metabolomics experiment and clearly separated the CG (BPK275, BHU573, BHU568 and BPK282) from ISC1 strains (BPK026, BPK406, BPK519, BPK604, BPK612). The second principal component explained 23% of the variation, which separated CG strains in 3 groups: BPK275 (Nepal, SSG-resistant), BHU573-BHU568 (India, SSG-resistant) and BPK282 (Nepal, SSG-sensitive).

33 metabolites were more abundant in ISC1 strains ($\text{Log}_2\text{FC} > 1$, $p\text{-val} < 0.05$), while 12 were more abundant in CG strains ($\text{Log}_2\text{FC} > 1$, $p\text{-val} < 0.05$). Interestingly, 19 of these metabolites were glycerophospholipids (GPLs), providing evidence for major differences in lipid metabolism between CG and ISC1 (**Figure 4**). In addition, ISC1 parasites contained higher metabolite levels of phosphorylethanolamine, glycerylphosphorylethanolamine, and glycerophosphoinositol pointing further towards lipid changes between both groups. Secondly, changes were observed in the urea cycle in ISC1 vs CG. ISC1 had a 2.1 fold higher amounts of citrulline and 3.6 fold lower amounts of argininosuccinate. The reaction that involves these compounds, namely citrulline + aspartate + ATP -> argininosuccinate + AMP + PPi, is catalyzed by argininosuccinate synthase (ASS). These metabolite

differences therefore indicate a lower ASS activity in ISC1, which is in correspondence with the lower copy number of the respective gene in ISC1 (as described in the genome structure diversity section). Interestingly, the metabolite levels also indicate a difference within the nucleotide salvage pathway between both groups, as we found increased levels of guanine, uracil, thymine, AMP in ISC1 vs CG. Finally, concentrations of basic amino acids lysine (up) and histidine (down) were altered in ISC1 vs CG.

We have previously shown that significant metabolic variation is present between SSG-resistant strains (BPK275, BHU573 and BHU568: CG-R) and SSG sensitive strains (BPK282: CG-S) (19). Considering these metabolic differences, we also made individual comparisons of ISC1 vs CG-R, ISC1 vs CG-S and CG-S vs CG-R. As expected, this confirmed substantial differences in the metabolome between CG-S and CG-R, with 52 metabolites showing a significant ($\text{Log}_2\text{FC} > 1$, $p\text{-val} < 0.05$) difference in concentration in only one of both groups. Consequently, when comparing CG-R and CG-S to ISC1 individually, we found many metabolites only differential in one of both comparisons (62 and 51 for CG-R and CG-S respectively), that were not observed in the original ISC1 vs CG in comparison. However, when looking in detail to the functions of these newly identified differential metabolites, these related toward the same functional classes of metabolites as those found when comparing ISC1 to all CG strains (**Supplementary Table S1.7**). This shows that although significant metabolome variation within the CG is present, there are clear and consistent differences with ISC1 on the level of GPLs, salvage pathway and urea cycle.

Finally, we compared the metabolomic and genomic polymorphism to quantify the relationship between both data sets and assess if/how metabolomic variation could be explained by genomic variation. Therefore, we computed distances between each pair of isolates, based on M-distance, S-distance and GD-distance, as explained in the Methods section (). Unfortunately, for the S-distance the intra group variance was far smaller than the inter group variance, which biases the results. Therefore we did not analyse S-distance any further. However, this problem was absent for GD-distance and a Mantel test established a significant positive correlation between M-distance and GD-

distance ($p = 0.005$, Spearman $\rho=0.64$). This means that GD-distance explains a significant proportion of the metabolic distance, but the low spearman ρ indicates that much other variation is present as well.

ACCEPTED MANUSCRIPT

Discussion

In this work we have performed a genomic and metabolic profiling study of strains representing ISC1, a new *Leishmania donovani* population that is currently emerging in Nepal. We have demonstrated that these parasites are genomically and metabolically very distinct from the main population (CG) of *L. donovani*, which was the driver of the last VL epidemics in the Indian subcontinent (3).

From a genomic point of view, the differences between ISC1 and CG strongly contrast with the homogeneity of the latter group and explained by its young evolutionary history (3). First, we report a large number of SNPs and indels, many of them with a strong impact (stop codons gained/lost and frameshifts) on several important biological functions. Secondly, we also encountered differences in the karyotypes and more particularly two chromosomes (8 and 23) that showed a lower copy number in ISC1. Copy number variation is an adaptive strategy well known in *Leishmania*, for instance in response to experimental drug resistance selection (36), but its interpretation in clinical isolates needs to be done very carefully. Indeed, we showed that aneuploidy is strongly dependent on the environment and more abundant in promastigotes *in vitro* than in amastigotes within the vertebrate host (9). Strictly speaking, we can thus only conclude that cultivated promastigotes of ISC1 show significant karyotypic differences to those of CG. Further extrapolation of these karyotype differences to the situation *in vivo* should be tested by direct sequencing of clinical samples or in experimental animal infections. A third major class of genomic changes here observed concerned local copy number variations (CNVs). In contrast to aneuploidy, these were shown not to be affected during the life cycle of the parasite (9), and CNVs have been shown to be excellent markers for studies on *Leishmania* evolution (37), both in experimental (for instance, amplification of specific loci during drug resistance selection (34,38)) or in natural settings (for instance, decrease in copy number of gp63, mini-exon and rDNA genes in *Leishmania peruviana* (39-41)). In the absence of gene expression regulation at initiation (42), amplification/deletion of specific genes is a genomic solution to modulate the level of transcripts and corresponding products. As such, these CNVs may have an

important functional significance. In this context, it is striking to observe that many of the CNVs concerned genes that code for surface proteins, virulence factors or immunogens: most notably, lower copy numbers of amastins, surface antigens and GP63 genes were encountered in ISC1. Interestingly, many *Leishmania* virulence genes were the target of highly impactful non-synonymous SNPs (removed/introduced stop codons), such as autophagy protein Atg8, trypanothione synthetase and amastin-like surface proteins. Although it is impossible to predict now the exact consequences of these SNPs, they are likely to have a high impact at the protein level by extending/shortening the protein. Altogether, genomic differences between ISC1 and CG suggest a potential for differences in virulence, which should be explored by further work. A first experimental verification of this hypothesis was made in the context of antimony susceptibility showing CG isolates were pre-adapted to antimony while ISC1 ones were not. This particular feature could be linked to the presence of the H-locus amplification in the CG (Dumetz et al., accepted, mSphere 2018 3: p. e00548-17-e00548-17).

The metabolome being the molecular expression of the parasite's phenotype, its study is extremely relevant to detect important functional differences between organisms. Metabolic differences between ISC1 and CG were found in several functional groups and pathways. First, the most notable changes were found in the lipid metabolism, with 19 GPLs (predominantly GPCs), showing significantly different levels between both groups. Apart from their structural function, GPLs are involved in a wide array of cellular functions including protein and glycoconjugate anchoring, host cell infection, and apoptosis-like processes (43). A second notable metabolic difference concerned the urea cycle. In ISC1 versus CG we detected a higher concentration of citrulline and a lower concentration of argininosuccinate. In organisms with a complete urea cycle, argininosuccinate is then further converted to arginine and fumarate by argininosuccinate lyase (ASL), however, in *Leishmania* this enzyme is missing (45). This difference is of particular interest since both argininosuccinate and arginine deprivation have been linked with reduced thiol content (45,46) and ASS mutants have shown a lower virulence than WT parasites (47). Therefore, the importance of ASS should be further analyzed. Finally, a metabolic change that also came forward in our comparisons

between ISC1 and CG strains was the nucleotide salvage pathway. This pathway is essential, since *Leishmania* cannot synthesize the purine ring *de novo* and is therefore dependent on salvaging these from host purines (48). Our results suggest that ISC1 parasites might be better at salvaging nucleotides from their environment. Altogether, ISC1 and CG parasites appeared to show different levels of activity among different metabolic pathways, all of them susceptible to predict differences in virulence, which should be tested by further work. 'Notably, our observations were made in stationary phase axenic promastigotes, which contain a high degree of metacyclic, infectious promastigotes (49). This is of particular interest, since virulence is of essential importance to this life stage. However, it is certainly possible that more differences between ISC1 and CG might come forward when studying additional parasite life stages in follow-up studies. Ultimately, the rapidly evolving metabolomics technology will most likely allow metabolite quantifications of low amounts of cells *in vivo*, or even single-cells in the future (50).'

The parallel untargeted analysis of genome and metabolome of different *L. donovani* strains also allowed us to undertake a first integration of both datasets. A quantitative approach was attempted by directly correlating the pairwise GD-distance and metabolic distances between strains. Although this correlation was found to be significant, they also demonstrate that the genomic distances only explain a limited part of the variation observed at the metabolome level. This is not unexpected, since regulation of gene expression in Trypanosomatids is essentially post-transcriptional. Qualitative analysis of our results showed a few correspondences between genomic and metabolic markers (**Figure 5**). Correspondence could concern either an enzyme and the specific metabolites of the reaction it catalyses, or this enzyme and metabolites from the same pathway. The former is best illustrated by the argininosuccinate synthase (ASS) and argininosuccinate, the gene encoding ASS is located on the H-locus and had a 6.7 fold lower expression in ISC1 by the combined effect of some and local copy number and argininosuccinate is 3.6 times less abundant in the metabolome of ISC1. Two other examples of convergence between genomics and metabolomics concern pathways: genes and metabolites pointing towards a same altered pathway, even if they did not point towards the

same product. Firstly, the major changes revealed in the lipid metabolism by metabolomics were mirrored by CNVs. We found 1.55 copies per haploid genome more in ISC1 for diacyl glycerol acyltransferase, which is part of the GPC biosynthesis pathway. Interestingly, we found predominantly higher GPC levels in ISC1. A CNV was also found in a fatty acid elongase (1.65 copies per haploid genome more in ISC1), which is an enzyme responsible for the extension of fatty acids that are later used to build GPLs. Another CNV related to GPL building blocks was found in the glycerol uptake protein. Finally, a CNV was found in 3-ketoacyl-CoA thiolase, which is located in the glycosome and an enzyme of the β -oxidation pathway (breakdown of fatty acids) (49). We also detected high impact SNPs in the ISC1 group for two enzymes that are important for lipid signaling: a premature stop codon was in phosphatidylinositol 3&4-kinase and a stop codon was removed from CDP-diacylglycerol-inositol 3-phosphatidyl transferase. Secondly, the different activity of the salvage pathway revealed by metabolomics was also mirrored by genomics. We found indeed that ISC1 parasites have four copies more of the p1/s1 nuclease, which is part of the salvage pathway and responsible for digesting the ssRNA and ssDNA to nucleotides. This was accompanied by an increase of guanine, uracil, thymine and AMP in ISC1. *In summary, our data strongly suggest that gene copy number variation may be an important adaptation strategy of the parasite as shown by the gene dosage changes related to metabolic pathways where metabolites were altered (Figure 5). This implies corresponding changes in protein levels, which is somewhat surprising as transcript levels (driven by gene dosage) are generally assumed to match poorly with protein levels (50). However, these observations were made in context of differentiation, rather than inter-strains comparisons as we have performed here. Our data suggest that between strains, gene dosage levels might in fact drive protein levels, as was also found in yeast however further work is required to define the extent of agreement between variation of genome and proteome.*

The evolutionary history and origin of *L. donovani* ISC1 are still unclear. In a previous work, genomic data allowed us to estimate the origin of the CG around 1850, hereby fitting with the dates mentioned in the early reports of the first epidemics of 'Quinine-resistant malaria' in the ISC, which

later on appeared to be VL (3). However, when comparing the number of SNPs between ISC1 and CG and the number of SNPs within CG (respectively 45,743 and 2,418, as previously published with a larger sample and an earlier version of the reference genome (3)), we may conclude that divergence time between ISC1 and CG is much earlier than 1850, probably several thousands of years (or more). This suggests that ISC1 and CG represent two post-bottleneck groups of a historically much more polymorphic population endemic in the ISC, although we cannot exclude that one or both were recently introduced from another, currently unsampled VL region. The further fate of ISC1 in the Indian sub-continent remains to be determined by molecular tracking. In the context of the current elimination program, it is possible that new and strong evolutionary bottlenecks will be created. Among the possible scenarios, the size of the CG population could be strongly reduced, while emerging populations like ISC1 could be favored and spread over the region. Such a scenario is supported by (i) the presence of these variants in regions where the elimination program could be less active, like in the highlands and (ii) the functional differences predicted by the present work. In summary, we have shown that the ISC1 population is very different in genotype and phenotype from the CG population that drove last epidemics of VL in the ISC and that was affected by drug resistance. Genomic and metabolic approaches pointed out major changes in pathways and molecules that are important for the interaction with the host/extracellular environment. Therefore, further work should be performed to deeply characterize ISC1 parasites, especially at functional level. A particular attention is also required to monitor the fate of this emerging population in the ISC, especially in a post-elimination context.

References

1. Alvar, J., Vélez, I.D., Bern, C., Herrero, M., Desjeux, P., Cano, J., Jannin, J., Boer, M.d. and the, W.H.O.L.C.T. (2012) Leishmaniasis Worldwide and Global Estimates of Its Incidence. *PLOS ONE*, **7**, e35671.
2. Ostyn, B., Gidwani, K., Khanal, B., Picado, A., Chappuis, F., Singh, S.P., Rijal, S., Sundar, S. and Boelaert, M. (2011) Incidence of Symptomatic and Asymptomatic *Leishmania donovani* Infections in High-Endemic Foci in India and Nepal: A Prospective Study. *PLOS Neglected Tropical Diseases*, **5**, e1284.
3. Imamura, H., Downing, T., Van den Broeck, F., Sanders, M.J., Rijal, S., Sundar, S., Mannaert, A., Vanaerschot, M., Berg, M., De Muylder, G. *et al.* (2016) Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent. *eLife*, **5**.
4. Ostyn, B., Uranw, S., Bhattarai, N.R., Das, M.L., Rai, K., Tersago, K., Pokhrel, Y., Durnez, L., Marasini, B., Van der Auwera, G. *et al.* (2015) Transmission of *Leishmania donovani* in the Hills of Eastern Nepal, an Outbreak Investigation in Okhaldhunga and Bhojpur Districts. *PLOS Neglected Tropical Diseases*, **9**, e0003966.
5. Rai, K., Bhattarai, N.R., Vanaerschot, M., Imamura, H., Gebru, G., Khanal, B., Rijal, S., Boelaert, M., Pal, C., Karki, P. *et al.* (2017) Single locus genotyping to track *Leishmania donovani* in the Indian subcontinent: Application in Nepal. *PLOS Neglected Tropical Diseases*, **11**, e0005420.
6. Zhang, W.W., Ramasamy, G., McCall, L.-I., Haydock, A., Ranasinghe, S., Abeygunasekara, P., Sirimanna, G., Wickremasinghe, R., Myler, P. and Matlashewski, G. (2014) Genetic Analysis of *Leishmania donovani* Tropism Using a Naturally Attenuated Cutaneous Strain. *PLOS Pathogens*, **10**, e1004244.

7. t'Kindt, R., Scheltema, R.A., Jankevics, A., Brunker, K., Rijal, S., Dujardin, J.C., Breitling, R., Watson, D.G., Coombs, G.H. and Decuypere, S. (2010) Metabolomics to unveil and understand phenotypic diversity between pathogen populations. *PLoS Negl Trop Dis*, **4**, e904.
8. Bronner, I.F., Quail, M.A., Turner, D.J. and Swerdlow, H. (2014) Improved protocols for illumina sequencing. *Current Protocols in Human Genetics*, 18.12. 11-18.12. 42.
9. Dumetz, F., Imamura, H., Sanders, M., Seblova, V., Myskova, J., Pescher, P., Vanaerschot, M., Meehan, C.J., Cuypers, B., De Muylder, G. *et al.* (2017) Modulation of Aneuploidy in *Leishmania donovani* during Adaptation to Different In Vitro and In Vivo Environments and Its Impact on Gene Expression. *mBio*, **8**.
10. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297-1303.
11. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, **6**, 80-92.
12. Tihon, E., Imamura, H., Van den Broeck, F., Vermeiren, L., Dujardin, J.-C. and Van Den Abbeele, J. (2017) Genomic analysis of Isometamidium Chloride resistance in *Trypanosoma congolense*. *International Journal for Parasitology: Drugs and Drug Resistance*, **7**, 350-361.
13. Downing, T., Imamura, H., Decuypere, S., Clark, T.G., Coombs, G.H., Cotton, J.A., Hilley, J.D., de Doncker, S., Maes, I., Mottram, J.C. *et al.* (2011) Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome research*, **21**, 2143-2156.
14. Berg, M., García-Hernández, R., Cuypers, B., Vanaerschot, M., Manzano, J.I., Poveda, J.A., Ferragut, J.A., Castanys, S., Dujardin, J.-C. and Gamarro, F. (2015) Experimental resistance to drug

combinations in *Leishmania donovani*: metabolic and phenotypic adaptations. *Antimicrobial agents and chemotherapy*, **59**, 2242-2255.

15. Berg, M., Vanaerschot, M., Jankevics, A., Cuypers, B., Breitling, R. and Dujardin, J.C. (2013) LC-MS metabolomics from study design to data-analysis - using a versatile pathogen as a test case. *Computational and structural biotechnology journal*, **4**, e201301002.

16. Jankevics, A., Merlo, M.E., de Vries, M., Vonk, R.J., Takano, E. and Breitling, R. (2012) Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets. *Metabolomics : Official journal of the Metabolomic Society*, **8**, 29-36.

17. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*, **78**, 779-787.

18. Scheltema, R.A., Jankevics, A., Jansen, R.C., Swertz, M.A. and Breitling, R. (2011) PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Analytical chemistry*, **83**, 2786-2793.

19. Berg, M., Vanaerschot, M., Jankevics, A., Cuypers, B., Maes, I., Mukherjee, S., Khanal, B., Rijal, S., Roy, S., Opperdoes, F. *et al.* (2013) Metabolic adaptations of *Leishmania donovani* in relation to differentiation, drug resistance, and drug pressure. *Molecular microbiology*, **90**, 428-442.

20. Prince, J.T. and Marcotte, E.M. (2006) Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical chemistry*, **78**, 6140-6152.

21. Tautenhahn, R., Bottcher, C. and Neumann, S. (2008) Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*, **9**, 504.

22. Scheltema, R., Decuypere, S., Dujardin, J., Watson, D., Jansen, R. and Breitling, R. (2009) Simple data-reduction method for high-resolution LC-MS data in metabolomics. *Bioanalysis*, **1**, 1551-1557.

23. Doyle, M.A., MacRae, J.I., De Souza, D.P., Saunders, E.C., McConville, M.J. and Likic, V.A. (2009) LeishCyc: a biochemical pathways database for *Leishmania major*. *BMC systems biology*, **3**, 57.
24. Fahy, E., Sud, M., Cotter, D. and Subramaniam, S. (2007) LIPID MAPS online tools for lipid research. *Nucleic acids research*, **35**, W606-612.
25. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, **45**, D353-d361.
26. Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic acids research*, **35**, D521-526.
27. Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B.S., van Ommen, B., Pujos-Guillot, E., Verheij, E., Wishart, D. and Wopereis, S. (2009) Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics : Official journal of the Metabolomic Society*, **5**, 435-458.
28. Jara, M., Berg, M., Caljon, G., de Muylder, G., Cuypers, B., Castillo, D., Maes, I., Orozco, M.d.C., Vanaerschot, M., Dujardin, J.-C. *et al.* (2017) Macromolecular biosynthetic parameters and metabolic profile in different life stages of *Leishmania braziliensis*: Amastigotes as a functionally less active stage. *PLOS ONE*, **12**, e0180532.
29. Mantel, N. (1967) The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, **27**, 209-220.
30. Grondin, K., Papadopoulou, B. and Ouellette, M. (1993) Homologous recombination between direct repeat sequences yields P-glycoprotein containing amplicons in arsenite resistant *Leishmania*. *Nucleic acids research*, **21**, 1895-1901.
31. Mukherjee, A., Padmanabhan, P.K., Singh, S., Roy, G., Girard, I., Chatterjee, M., Ouellette, M. and Madhubala, R. (2007) Role of ABC transporter MRPA, γ -glutamylcysteine synthetase and

ornithine decarboxylase in natural antimony-resistant isolates of *Leishmania donovani*. *Journal of Antimicrobial Chemotherapy*, **59**, 204-211.

32. Chiquero, M.J., Olmo, A., Navarro, P., Ruiz-Perez, L.M., Castanys, S., Gonzalez-Pacanowska, D. and Gamarro, F. (1994) Amplification of the H locus in *Leishmania infantum*. *Biochimica et biophysica acta*, **1227**, 188-194.

33. Olmo, A., Arrebola, R., Bernier, V., Gonzalez-Pacanowska, D. and Ruiz-Perez, L.M. (1995) Co-existence of circular and multiple linear amplicons in methotrexate-resistant *Leishmania*. *Nucleic acids research*, **23**, 2856-2864.

34. Ubeda, J.M., Raymond, F., Mukherjee, A., Plourde, M., Gingras, H., Roy, G., Lapointe, A., Leprohon, P., Papadopoulou, B., Corbeil, J. *et al.* (2014) Genome-wide stochastic adaptive DNA amplification at direct and inverted DNA repeats in the parasite *Leishmania*. *PLoS biology*, **12**, e1001868.

35. Kundig, C., Leblanc, E., Papadopoulou, B. and Ouellette, M. (1999) Role of the locus and of the resistance gene on gene amplification frequency in methotrexate resistant *Leishmania tarentolae*. *Nucleic acids research*, **27**, 3653-3659.

36. Prieto Barja, P., Pescher, P., Bussotti, G., Dumetz, F., Imamura, H., Kedra, D., Domagalska, M., Chaumeau, V., Himmelbauer, H., Pages, M. *et al.* (2017) Haplotype selection as an adaptive mechanism in the protozoan pathogen *Leishmania donovani*. *Nature ecology & evolution*.

37. Dujardin, J.-C. (2009) Structure, dynamics and function of *Leishmania* genome: Resolving the puzzle of infection, genetics and evolution? *Infection, Genetics and Evolution*, **9**, 290-297.

38. Leprohon, P., Légaré, D., Raymond, F., Madore, É., Hardiman, G., Corbeil, J. and Ouellette, M. (2009) Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. *Nucleic acids research*, **37**, 1387-1399.

39. Victoir, K., Dujardin, J.C., De Doncker, S., Barker, D.C., Arevalo, J., Hamers, R. and Le Ray, D. (2009) Plasticity of gp63 gene organization in *Leishmania* (*Viannia*) *braziliensis* and *Leishmania* (*Viannia*) *peruviana*. *Parasitology*, **111**, 265-273.
40. Kebede, A., De Doncker, S., Arevalo, J., Le Ray, D. and Dujardin, J.C. (1999) Size-polymorphism of mini-exon gene-bearing chromosomes among natural populations of *Leishmania*, subgenus *Viannia*. *International Journal for Parasitology*, **29**, 549-557.
41. Inga, R., De Doncker, S., Gomez, J., Lopez, M., Garcia, R., Le Ray, D., Arevalo, J. and Dujardin, J.-C. (1998) Relation between variation in copy number of ribosomal RNA encoding genes and size of harbouring chromosomes in *Leishmania* of subgenus *Viannia*. *Molecular and Biochemical Parasitology*, **92**, 219-228.
42. Günzl, A. (2012) *RNA Metabolism in Trypanosomes* (ed. Albrecht Bindereif) Springer Berlin Heidelberg.
43. Pulido, S.A., Nguyen, V.H., Alzate, J.F., Cedeno, D.L., Makurath, M.A., Rios-Vasquez, A., Duque-Benitez, S.M., Jones, M.A., Robledo, S.M. and Friesen, J.A. (2017) Insights into the phosphatidylcholine and phosphatidylethanolamine biosynthetic pathways in *Leishmania* parasites and characterization of a choline kinase from *Leishmania infantum*. *Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology*, **213**, 45-54.
44. Zhang, K. and Beverley, S.M. (2010) Phospholipid and sphingolipid metabolism in *Leishmania*. *Molecular and biochemical parasitology*, **170**, 55.
45. Sardar, A.H., Jardim, A., Ghosh, A.K., Mandal, A., Das, S., Saini, S., Abhishek, K., Singh, R., Verma, S., Kumar, A. *et al.* (2016) Genetic Manipulation of *Leishmania donovani* to Explore the Involvement of Argininosuccinate Synthase in Oxidative Stress Management. *PLOS Neglected Tropical Diseases*, **10**, e0004308.
46. Mandal, A., Das, S., Roy, S., Ghosh, A.K., Sardar, A.H., Verma, S., Saini, S., Singh, R., Abhishek, K., Kumar, A. *et al.* (2016) Deprivation of L-Arginine Induces Oxidative Stress Mediated Apoptosis in

Leishmania donovani Promastigotes: Contribution of the Polyamine Pathway. *PLOS Neglected Tropical Diseases*, **10**, e0004373.

47. Lakhali-Naouar, I., Jardim, A., Strasser, R., Luo, S., Kozakai, Y., Nakhasi, H.L. and Duncan, R.C. (2012) *Leishmania donovani* Argininosuccinate Synthase Is an Active Enzyme Associated with Parasite Pathogenesis. *PLOS Neglected Tropical Diseases*, **6**, e1849.

48. Boitz, J.M., Ullman, B., Jardim, A. and Carter, N.S. (2012) Purine salvage in *Leishmania*: complex or simple by design? *Trends in parasitology*, **28**, 345-352.

49. Jamdhade, M.D., Pawar, H., Chavan, S., Sathe, G., Umasankar, P.K., Mahale, K.N., Dixit, T., Madugundu, A.K., Prasad, T.S.K., Gowda, H. *et al.* (2015) Comprehensive Proteomics Analysis of Glycosomes from *Leishmania donovani*. *OMICS : a Journal of Integrative Biology*, **19**, 157-170.

50. Madhubala, R., Myler, P., Papadopoulou, B., Fasel, N., McConville, M., Zilberstein, D., Opperdoes, F., Michels, P., Ouellette, M., Handman, E. *et al.* (2008) *Leishmania: After the genome*.

Acknowledgements

The authors would like to thank the Scottish Metabolomics Facility (ScotMet) for the purchase of the prepared authentic standard mix and for the analysis of the samples on their LC-MS platform. The authors would like to thank dr. Andris Jankevics for his support with the mzMatch metabolomics analysis pipeline.

ACCEPTED MANUSCRIPT

Funding

This work was supported by the Research Foundation Flanders [1101614N to B.C]; the Interuniversity Attraction Poles Program of Belgian Science Policy [P7/41 to J.C.D.], the InBev Baillet-Latour foundation and the Department of Economy, Science and Innovation in Flanders [ITM-SOFIB, SINGLE to J.C.D.]. JAC and MS are funded by the Wellcome Trust through their core funding of the Wellcome Trust Sanger Institute [Grant 206194].

Tables and Figures

Figures

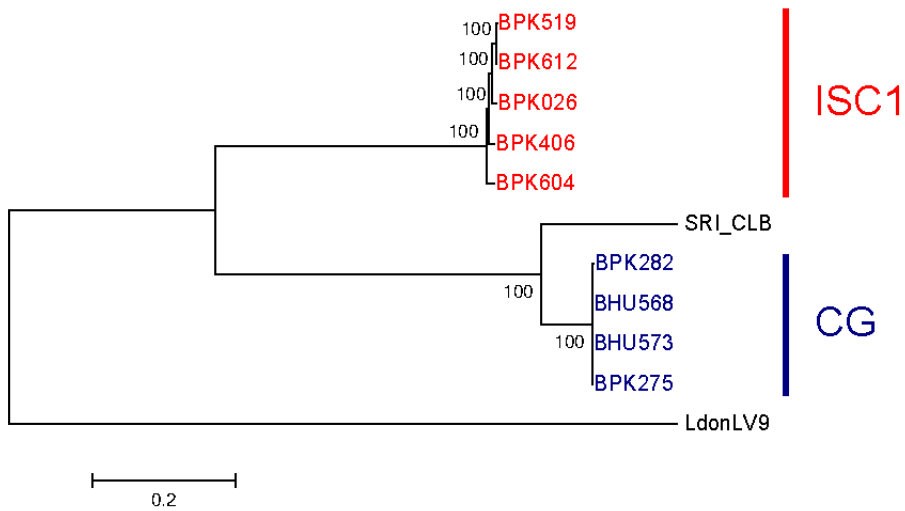


Figure 1: Neighbour joining tree based on whole genome SNP loci. Bootstrap support was calculated on 1000 bootstrap replicates. LdonLV9 = East African *L. donovani* LV9, SRI_CLB = Sri Lankan *L. donovani* SRI_CLB.

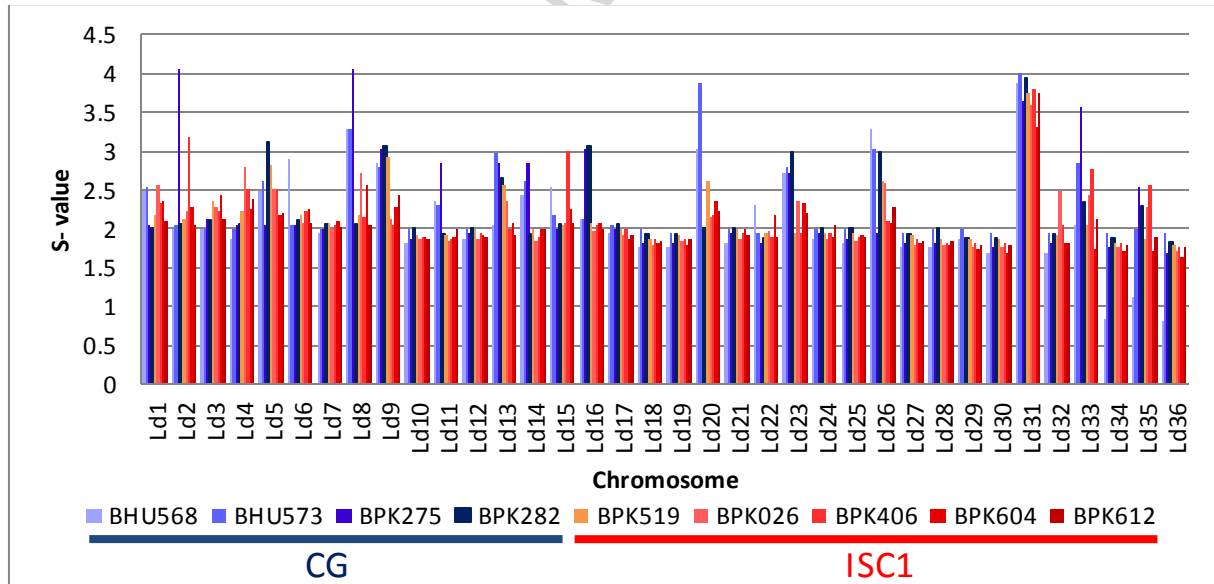


Figure 2: Somy levels in ISC1 and Core Group strains (CG). ISC1 strains are depicted in shades of red and yellow, while CG strains in shades of blue.

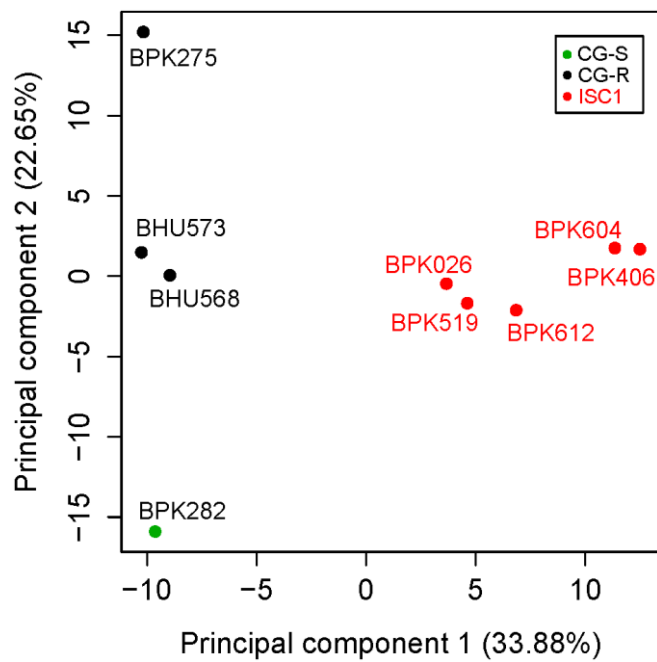


Figure 3: Principal component analysis of the metabolic profiles of the 5 *Leishmania donovani* ISC1 isolates and 4 CG strains in this study. This analysis was based on the quantitative measurements of all 290 putatively identified metabolites. In this PCA plot, average values are shown for each strain based on four biological replicates. For each principal component, the proportion of variation of the total dataset it explains is mentioned.

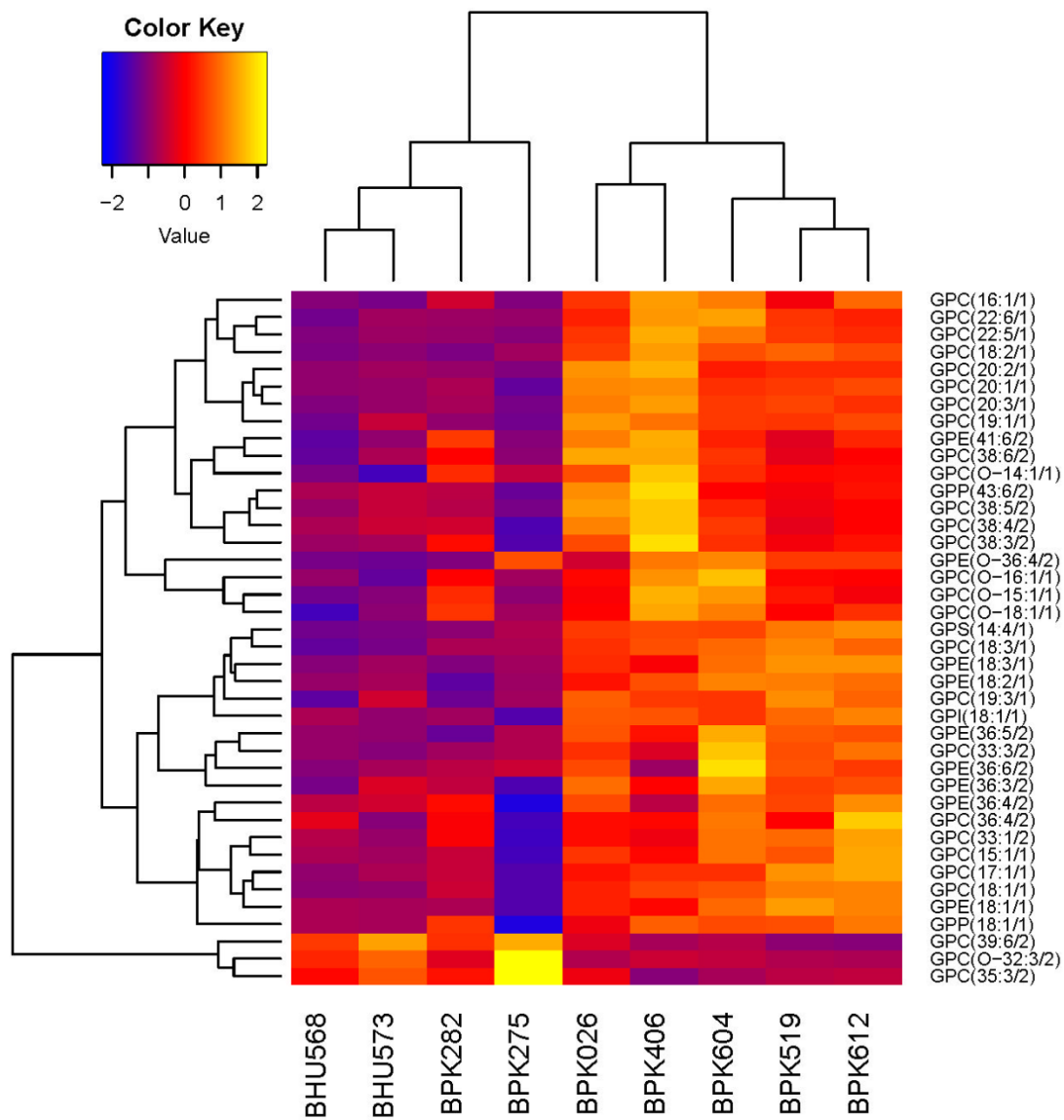


Figure 4: The 40 glycerophospholipids that showed differential expression ($|\text{Log}_2\text{FC}| > 1$ & $p < 0.05$) between ISC1 and CG parasites.

Key differences ISC1 vs CG shared between genome and metabolome

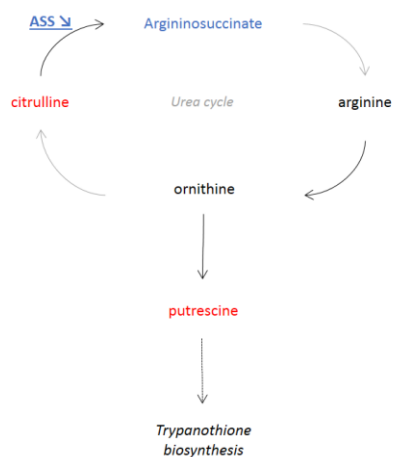
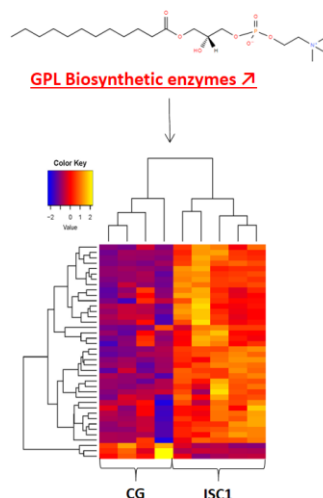
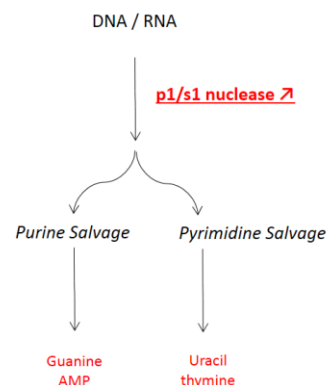
1. Argininosuccinate synthase \searrow 2. Glycerophospholipids \nearrow 3. Salvage Pathway \nearrow 

Figure 5: The 3 key differences between ISC1 and CG that are shared between genome and metabolome. 1. Argininosuccinate (ASS) has a lower genomic copy number in ISC1 vs CG. More citrulline and less argininosuccinate was detected in ISC1, suggesting a corresponding downregulation of the ASS protein. Although the urea cycle might be incomplete in *Leishmania*, links have been established with Trypanothione biosynthesis and virulence. 2. Several glycerophospholipid (or precursor) biosynthetic enzymes had higher gene copy numbers in ISC1 which could be linked to increased GPL amounts (See also Figure 4). 3. More copies of a p1/s1 nuclease gene were found in ISC1, associated with higher numbers of several salvage pathway metabolites. Genomic copy number information is underlined and bold. Pathway names are in italics. Arrows in grey represent potentially missing enzymes in *Leishmania* in the urea cycle. Blue = down in ISC1, Red = up in ISC1.

Supplementary Tables and Figures

Supplementary Tables:

Supplementary Tables S1

Summary of the strains used in this study as well as a summary of the main genomic sequence (SNPs, INDELS) and structure (CNV) differences between ISC1 and CG.

Supplementary Table S2

Coverage and mapping statistics of all genomes sequenced in this study. Includes also the Sequence Read Archive (SRA) accession numbers of the raw sequencing data.

Supplementary Table S3

Detailed SNPs and INDELS table for all strains in this study. The columns contain information about (in order): The chromosome on which the variant is located, the position within this chromosome, the reference and alternative allele, the variant quality, the variant frequency and variant sequencing depth in all studied strains, the average difference in allele frequency between the ISC1 and the CG, and associated Wilcoxon test p -value. The final columns contain information about the quality of the SNPs generated by GATK, and the functional annotations by SnpEff.

Supplementary Table S4

Table containing the somy estimations for each strain in this study (Tab 'Somy'), and table containing the haploid CNV depths for each gene and for each strain (Tab 'CNV'). For each CNV, the \log_2 fold change ($\log_2 \frac{\text{Average Expression Yeti}}{\text{Average Expression Core Group}}$) between ISC1 and ISC5 is also shown with its False Discovery Rate Adjusted (FDR) p -value.

Supplementary Table S5

All 290 putatively identified metabolites that were identified by our metabolomics pipeline. For each metabolite following parameters are given: ionization mode of the MS in which they were detected (Positive/Negative Mode), mass, retention time, putative identification, metabolic class and subclass, the peak intensity in each sample, ISC1 versus CG fold changes and p -values.

ACCEPTED MANUSCRIPT