A complete *Leishmania donovani* reference genome identifies novel genetic variations associated with virulence

Patrick Lypaczewski1, Johanna Hoshizaki1, Wen-Wei Zhang1, Laura-Isobel McCall1,2, John Torcivia-Rodriguez3, Vahan Simonyan3, Amanpreet Kaur4,5, Ken Dewar4,5 & Greg Matlashewski1

*Leishmania donovani* is responsible for visceral leishmaniasis, a neglected and lethal parasitic disease with limited treatment options and no vaccine. The study of *L. donovani* has been hindered by the lack of a high-quality reference genome and this can impact experimental outcomes including the identification of virulence genes, drug targets and vaccine development. We therefore generated a complete genome assembly by deep sequencing using a combination of second generation (Illumina) and third generation (PacBio) sequencing technologies. Compared to the current *L. donovani* assembly, the genome assembly reported within resulted in the closure over 2,000 gaps, the extension of several chromosomes up to telomeric repeats and the re-annotation of close to 15% of protein coding genes and the annotation of hundreds of non-coding RNA genes. It was possible to correctly assemble the highly repetitive A2 and Amastin virulence gene clusters. A comparative sequence analysis using the improved reference genome confirmed 70 published and identified 15 novel genomic differences between closely related visceral and atypical cutaneous disease-causing *L. donovani* strains providing a more complete map of genes associated with virulence and visceral organ tropism. Bioinformatic tools including protein variation effect analyzer and basic local alignment search tool were used to prioritize a list of potential virulence genes based on mutation severity, gene conservation and function. This complete genome assembly and novel information on virulence factors will support the identification of new drug targets and the development of a vaccine for *L. donovani*.

Visceral Leishmaniasis (VL) is the second most lethal parasitic disease following malaria and is prevalent throughout underdeveloped and tropical regions of the world. There are some 300,000 new cases each year1 and *Leishmania donovani*, transmitted by the infected sand fly, is the major causative agent of VL in the Indian and African continents. Although *L. donovani* is extensively studied, its genome remains poorly annotated because it is heavily fragmented and a complete assembly is crucial to understanding this parasite’s biology, metabolic pathways, tissue tropism and disease pathology.

The pathology of leishmaniasis is predominantly parasite species-specific, such as for example *L. major* that causes cutaneous leishmaniasis (CL) whereas *L. donovani* typically causes lethal visceral leishmaniasis (VL). Previous studies have compared genomes of *L. major* and *L. donovani* parasites to study virulence and disease tropism and have identified a number of species specific genes including A2 present in *L. donovani* that is a pseudogene in *L. major*2,3. More recently, as cases of atypical CL caused by *L. donovani* have emerged, studies have compared cutaneous and visceral disease-causing strains of *L. donovani*, as these strains provide a unique opportunity to study the genetic determinants of disease pathogenesis using more recently diverged strains4.

1Department of Microbiology and Immunology, McGill University, 3775 University Street, H3A 2B4, Montreal, Quebec, Canada. 2Present address: Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK, USA. 3Center for Biologics Evaluation and Research, Food and Drug Administration, Maryland, USA. 4Department of Human Genetics, McGill University, Montreal, Quebec, Canada. 5McGill University and Genome Quebec Innovation Center, McGill University, Montreal, Quebec, Canada. Correspondence and requests for materials should be addressed to G.M. (email: greg.matlashewski@mcgill.ca)
Second-generation sequencing technologies including Illumina, have made the sequencing of large genomes feasible through the mapping of short sequence reads of 50 to 250 nucleotides (nt) to a reference genome. While human and many other well studied higher vertebrates have better assembled reference genomes, the kinetoplastids suffer in this regard because most *Leishmania* species either lack sequencing information altogether or have incomplete reference genomes with sometimes thousands of sequence gaps. The current *L. donovani* reference genome (ASM22713v2 from strain BPK282A) was generated using second generation technologies and contains over 2,000 gaps and therefore there are many incomplete or inaccurate protein coding sequences. The first complete *Leishmania* genome generated is that of *L. major* by a consortium of laboratories employing large insert clone tiling paths to sequence each chromosome individually. This genome was later improved by the reassembly of complex collapsed loci that were incorrect in the original reference genome.

Since then however, advances in sequencing technologies have drastically reduced the cost of sequencing and eased genome assembly tasks by increasing the length of the generated sequences. Long read sequencing or “third-generation” sequencing refers to more recent technologies including Oxford NanoPore and Pacific Biosciences (PacBio) that can result in reads ranging up to 50 kb or 100 kb that are capable of generating more complete genomic assemblies, provided the read lengths traverse across repetitive elements. One such highly repetitive cluster is the A2 gene family from *L. donovani* considered to be an important virulence factor and is necessary for survival in visceral organs and protection against host response stress. Due to its repetitive nature, the A2 gene cluster is misassembled in all *Leishmania* genomes generated using second generation sequencing, and only resolved in a recent resequencing effort targeted to *L. infantum* exploiting the long-read capabilities of PacBio sequencing which resulted in a complete genome assembly. The current *L. donovani* genome however was obtained from second generation sequencing and consequently, no precise DNA or complete protein sequences are available for any A2 protein in *L. donovani*, hindering the comparison of A2 genes in visceral disease-causing strains or using mass spectrometry to identify A2 proteins which relies on accurate genome sequences for protein identification.

In this study, we have combined second and third generation sequencing to generate a complete assembly of the *L. donovani* genome from the strain responsible for cutaneous leishmaniasis (CL) in Sri Lanka. This new assembly enabled the generation of an improved genome annotation and an unbiased analysis of chromosome synteny comparing *L. donovani* and *L. major* genes and strand switch transcription units. We have used this complete assembly to re-interrogate the genetic makeup of the visceral and cutaneous disease-causing *L. donovani* strains resulting in the identification of novel SNPs and indels generating a more complete and accurate chromosome map of the genetic differences between these phenotypically distinct *L. donovani* strains. This study further enabled re-annotation of much of the genome highlighting the importance of a complete reference assembly to support future functional genomic and proteomic studies involving the *L. donovani* pathogen.

Results

**A complete *L. donovani* genome assembly.** The currently available assembly for *L. donovani* (ASM22713v2 from strain BPK282A) contains over 2,000 gaps due to the presence of low complexity regions and the highly repetitive nature of the *Leishmania* genome. This incomplete assembly makes it difficult to compare *L. donovani* genomes from strains with different phenotypic properties. DNA was therefore isolated from the attenuated cutaneous disease-causing strain of *L. donovani* from Sri Lanka and was subjected to deep sequencing using second and third generation sequencing. We reasoned that a complete assembly of the genome from this attenuated *L. donovani* strain will identify a more complete complement of genetic changes associated with loss of virulence of this strain. A total of 9 PacBio sequencing runs were performed generating 712,443 reads representing an estimated 107-fold coverage of the estimated 35 Mb genome. Importantly, there were 51,484 reads longer than 12 kb, representing a 20-fold coverage in very long reads. The long-read sequencing data was assembled using various assemblers as described in methods and merged using the longest chromosomes produced by each assembler followed by refinement using the high-quality short-read Illumina-generated data and iterative edge extension to close the remaining gaps.

The previous *L. donovani* reference assembly (ASM22713v2 from strain BPK282A) had over 2000 gaps spread across the 36 chromosomes. Figure 1 depicts the location on each chromosome of the gaps that have been closed in the new assembly reported here. The new assembly now contains contiguous DNA sequences in all 36 chromosomes and a corresponding 22-fold increase in N50 indicating that a larger proportion of the data has been assembled into large contigs as 50% of the genome is contained in contigs greater than 12 kb, representing a 20-fold coverage in very long reads. The long-read sequencing data was assembled using an estimated 107-fold coverage of the estimated 35 Mb genome. Importantly, there were 51,484 reads longer than 12 kb, representing a 20-fold coverage in very long reads. The long-read sequencing data was assembled using various assemblers as described in methods and merged using the longest chromosomes produced by each assembler followed by refinement using the high-quality short-read Illumina-generated data and iterative edge extension to close the remaining gaps.

Assembly of the A2 virulence gene cluster and synteny comparison between *L. major* and *L. donovani.* A2 is a major virulence factor required for *L. donovani* survival in visceral organs. The A2 gene family cluster on chromosome 22 has recently been assembled for *L. infantum*, however has not been for *L. donovani*. We therefore investigated whether the structure of this region could be determined with this revised assembly. It was advantageous that the attenuated cutaneous *L. donovani* strain used in this assembly has fewer copies of the A2 genes than other virulent strains of *L. donovani*. As shown in Fig. 2a, the new assembly could read-through the entire cluster of highly repetitive A2 and flanking sequences and could position the A2 genes and interspersed flanking 3′ A2-rel and 5′ A2-rel genes. The A2 genes are contained in two opposite facing clusters.
on either side of a strand-switch locus consisting of one cluster of 3 copies and one cluster with a single A2 gene. The long sequence reads generated by the PacBio sequencing were crucial in generating the assembly of the A2 genes where reads of 11 kb and longer are shown spanning the repetitive cluster (Fig. 2b).

To generate supporting evidence for this A2 gene assembly, Western blot analysis of the A2 proteins from this strain was performed to compare the number and sizes of the A2 proteins with the predicted molecular weights from this assembly (ORFs; Supplementary Fig. S2). As shown in Fig. 2c, the apparent molecular weights from Western blotting correspond to the sizes predicted from the sequenced ORFs. The 3 bands on the Western blot are consistent with the molecular weights of the 4 gene products as the A2 gene copies 2 and 3 encode proteins of a similar size (Supplementary Fig. S2). This represents the first complete structure and sequence for A2 genes in *L. donovani*, a prototype virulence factor. The difficulty in assembling this complex region is demonstrated in Fig. 2d, where a deviation from the average read coverage can be seen around the 300,000 bp position, in and around the A2 cluster, due to difficulties in the aligner assigning a unique position to similar reads across a repetitive region.

Directly comparing synteny at the chromosomal level was not possible with the previous *L. donovani* assembly due to the heavy fragmentation of the genome. With the new *L. donovani* assembly, it was possible to accurately compare chromosome synteny between *L. donovani* and *L. major*. As shown in Fig. 3, the genome of *L. donovani*, exhibited a very high level of synteny with the *L. major*. Chromosome 22 was highlighted here because this is the location of the A2 genes that have become pseudogenes in *L. major* and have therefore diverged between these old-world species15. The level of synteny demonstrated here for chromosome 22 was maintained on all other chromosomes (Supplementary Fig. S3). These results indicate that evolution between cutaneous and visceral pathologies by different *Leishmania* species resulted largely from SNPs, pseudogenes and copy number variation and not from large changes such as chromosome rearrangements or complete gene deletions/insertions.

Identification of new genes and improvements in annotations. As this assembly was larger in terms of total number of bases covered and more contiguous due to the removal of sequence gaps, the impact this had on gene annotations was investigated. The genome from the new assembly was annotated using the Companion pipeline23 and the new and previous annotations (GenBank: GCF_000227135.1) were then aligned together and overlapping annotations were removed. Remarkably, close to 15% of the *L. donovani* protein coding genes had new or edited annotations as shown in Fig. 4a. Part of this increase in number of annotations resulted from the expansions of multi-copy gene families beyond the copy numbers in the previous annotation. An example is shown in Fig. 4b where there are 10 amastin genes identified in this new assembly compared to the previously identified 2. These results support the use of this assembly as the reference for bioinformatic analysis as it provides a more complete and accurate annotation of the *L. donovani* genome ORFs.

---

**Table 1.** Quality assessment metrics of the previous and current assemblies. Old assembly refers to ASM22713v2 from strain BPK282A, new assembly refers to the assembly presented in this work. Contigs denotes the number of genomic fragments uninterrupted by stretches of unknown bases (Ns) or chromosome ends. N50 is used as a measure of contiguity; 50% of the genome is contained in contigs of size N50 and above. Annotated genes were broken down into protein coding, transfer-RNA (tRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA), spliced leader RNA (SLRNA) and small nucleolar RNA (snoRNA) genes. The number of genes mapped indicates the number of annotated genes along the genome.

|          | Contigs | N50 (bp) | Protein coding | tRNA | rRNA | snRNA | SLRNA | snoRNA | Genes mapped |
|----------|---------|----------|----------------|------|------|-------|-------|--------|--------------|
| Old Assembly | 2,154   | 45,436   | 7,969          | 64   | 11   | 4     | —     | 31     | 8,081        |
| New Assembly | 36      | 1,067,468| 8,633          | 90   | 51   | 6     | 68    | 910    | 9,758        |

---

**Figure 1.** Location of the gaps along 36 chromosomes that have been closed in this new assembly. Chromosomal locations of gaps are indicated in red. No gaps remain in the current assembly.
Comparison of virulent and attenuated *L. donovani* parasites. As indicated above, there are 2 distinct strains of *L. donovani* in Sri Lanka where one is responsible for visceral leishmaniasis (VL) and the other for cutaneous leishmaniasis (CL)⁴. Subsequently, the CL strain was experimentally passaged through the visceral...
organs of BALB/c mice to select for a gain-of-function strain with increased virulence (IV strain) for survival in visceral organs where it was revealed through proteomic analysis that the resulting IV strain had an increase in stress response and antioxidant proteins\(^2\). Illumina whole genome sequencing and comparative genomic analysis of the VL, CL and IV strains was performed to identify SNPs associated with a change in virulence for survival in the visceral organs. As shown in Fig. 5, all 70 of the previously identified homozygous SNP differences between the VL and CL strains\(^4\) were confirmed in this new assembly and an additional 15 novel SNPs within protein coding genes were found using this complete assembly. In addition, there were 12 mutations associated with the IV strain with gain-of-function for increased survival in visceral organs not labeled in Fig. 5; four were heterozygous but with frequencies changing towards the VL genotype (IV $\rightarrow$ VL), four were heterozygous but present only in the IV strain and four were homozygous deletions in the IV strain. The newly identified differences between the VL and CL strains and the ones contributed from the IV strain are summarized in Table 2.

Combining data from the previous and current analysis, all the genes with genetic differences were organized into priority clusters based on the likelihood to affect protein function and phenotype (Fig. 6). A detailed list of the genes and cluster assignments is shown in Table 3. From the 66 genes containing 70 SNPs previously identified, 7 were previously experimentally assessed using gene replacement with a wildtype copy for virulence in visceral organs and one gene was identified as a misannotation and was therefore removed from the list. In decreasing order of priority, 13 genes in the highest impact cluster (red) were characterized as potentially having a major effect on protein function due to large amino acid changes or co-occurrence of mutations in both the VL and IV strains. SNPs in common between the IV and VL strains (4 IV $\rightarrow$ VL in the red cluster) indicate a selection associated with survival in visceral organs during the experimental passaging of the CL strain in mouse visceral organ\(^2\). Nine genes either with multiple co-occurring SNPs or non-conservative amino acid changes in conserved domains with a high score as assessed by PROVEAN software were placed in the second highest priority cluster. As detailed in methods, PROVEAN is a bioinformatic tool that classifies the significance of specific genetic mutations with respect to protein function\(^5\). Eighteen genes with non-conservative amino acid changes occurring in conserved domains but scored as unlikely to have a large effect on protein function by PROVEAN were placed in Cluster 3. Twenty-one genes with conservative amino acid changes in conserved domains were divided between Cluster 4 and Cluster 5 based on PROVEAN scores and 14 genes with mutations in domains with higher variability were placed in Cluster 6. Four mutations seen solely in the IV strain but not in the VL or CL strains are likely to be the result of random mutation or adaptation specific to the murine host were placed in Cluster 7.

A 25 kb region on chromosome 36 containing 4 genes was found to be missing in the IV strain but present in the VL and CL strains. This deletion did not occur in a location previously identified on this chromosome where a fission can occur as seen in \emph{L. alderi}\(^6\). Upon experimental verification, this deleted region was present in a

Figure 4. The new \emph{L. donovani} genome assembly results in a significant change in gene annotations. (a) New or improved gene annotations are highlighted in Blue along the 36 chromosomes. Compared to the previous \emph{L. donovani} reference assembly (ASM22713v2 from strain BPK282A1), there were 1,087 protein coding genes unannotated or differently annotated in the current assembly. Unannotated or differently annotated genes were obtained by removing all annotations generated from our assembly that shared 95% or greater similarity to those previously available\(^8\). (b) Expansion of the amastin gene cluster on chromosome 8. Top track contains the previously two known coding sequences aligned to the previous \emph{L. donovani} reference assembly (ASM22713v2 from strain BPK282A1). Gaps in the previous assembly depicted as dotted lines. Bottom track contains 10 amastin genes identified in the updated assembly. One previously identified Amastin gene has been aligned, 1 has been expanded and 8 have been annotated de novo.
major chromosomal rearrangements or deletions are apparent between phenotypically different Leishmania species as previously reported, and including this study, suggesting that virulence and tropism can be acquired or used in future sequencing efforts aimed at comparing genes and synteny of genomes of other Leishmania species. This assembly can be assumed to be syntenic and only gene synteny was possible. In contrast, the contiguous assembly presented within used an entirely reference-free and by extension, bias-free generation process. This assembly will significantly improve genomic, functional genomic and proteomic research outcomes and support the identification of drug targets and the development of vaccines. This assembly further provides a larger repertoire of target DNA sequences to identify diagnostic and prognostic disease progression markers. Given the recent interest in generating genetically modified live attenuated parasites as vaccine candidates, a complete genome assembly will permit the verification that genetic modifications target intended genes with no off target mutations.

Discussion

It has been possible to generate a complete genome assembly for L. donovani through combining second and third generation sequencing technologies, similarly to a recent resequencing of the L. infantum genome resulting in a complete assembly, highlighting the usefulness of PacBio sequencing in regards to Leishmania genomes. This resulted in a more accurate annotation of the genome increasing the number of potential protein coding genes and identifying novel mutations/polymorphisms associated with virulence. It was remarkable that the present assembly resulted in annotation changes in close to 15% of the genome representing 1087 protein coding genes. Although 13 degenerate pseudogenes are identified in Supplementary Table S1 more do exist since our annotations derived from functional genes in L. major and therefore genes functional in other species were not identified. Through this updated genome annotation, additional SNPs have been identified including in genes potentially involved in visceral disease and several non-coding genes have been annotated allowing future L. donovani research beyond protein coding genes. It has also been possible to assemble known virulence factor gene families in L. donovani including the A2 and Amastin gene families. This version of the L. donovani genome assembly will significantly improve genomic, functional genomic and proteomic research outcomes and support the identification of drug targets and the development of vaccines. This assembly further provides a larger repertoire of target DNA sequences to identify diagnostic and prognostic disease progression markers. Given the recent interest in generating genetically modified live attenuated parasites as vaccine candidates, a complete genome assembly will permit the verification that genetic modifications target intended genes with no off target mutations.

Supported by a de novo assembly, this study provides the first direct evidence for synteny between chromosomes in L. donovani and L. major, two old world parasite species with different pathologies and reservoirs. Previously, due to the large number of gaps in the L. donovani genome, the segments were aligned to a reference assembly assumed to be syntenic and only gene synteny was possible. In contrast, the contiguous assembly presented within used an entirely reference-free and by extension, bias-free generation process. This assembly can be used in future sequencing efforts aimed at comparing genes and synteny of genomes of other Leishmania species with L. donovani. The strong gene level synteny further highlights the major phenotypic effects of SNPs and indel mutations when comparing genomes from L. donovani strains causing visceral and cutaneous pathologies. As no major chromosomal rearrangements or deletions are apparent between phenotypically different Leishmania species as previously reported, and including this study, suggesting that virulence and tropism can be acquired or lost through relatively small coding changes at the amino acid level such as SNPs, indels and frameshifts without the need for chromosomal scale events.

Figure 5. Verification of previously identified SNPs and location of new SNPs that differ between the virulent VL and attenuated CL strains of L. donovani. Chromosomal location of previously identified homozygous non-synonymous SNPs between the cutaneous and visceral disease derived L. donovani strains (Red) compared to the novel SNPs identified only in this study (Blue) (synonymous and heterozygous codon changes identified are not labeled). Note that all the previously identified SNPs were also identified, or confirmed, in this study. 70 SNPs were previously identified across 66 genes. The same 70 SNPs were identified in this study, with an additional 15 novel SNPs not previously seen specific to the cutaneous strain. Genomic locations of SNPs identified in the previous study were translated to new genomic coordinates based on the new assembly for consistency. Arrows in yellow highlight the position of the previously identified RagC SNP on chromosome 36 and the A2 copy number difference on chromosome 22.
with unknown function are of high priority for future studies. Nevertheless, although the gain of function IV fifteen attenuated cutaneous strain specific mutations identified in this study. Mutations marked with a appear previous gene ID in italic, genes annotated only in this assembly list only one gene ID. The top segment lists type amino acids and CL as the mutated amino acids. Genes with annotations in the previous assembly list the Table 2. Summary of novel mutations identified in this study. All mutations are annotated using VL as the wild type amino acids and CL as the mutated amino acids. Genes with annotations in the previous assembly list the previous gene ID in italic, genes annotated only in this assembly list only one gene ID. The top segment lists fifteen attenuated cutaneous strain specific mutations identified in this study. Mutations marked with a appear at 50% but also co-occur with gene duplication event and are therefore possibly homozygous on one copy. 'INS' denotes amino acid insertions, 'F/S' denotes frameshifts, 'DEL' denotes amino acid deletions. The middle segment lists four mutations where the gain-of-function IV strain's genotype changed towards that of the visceral genotype. The bottom segment lists eight mutations present only in the gain-of-function IV strain and likely represents adaptations specific to the murine host. Calculated PROVEAN scores are shown in the fourth column, scores below the −2.5 threshold for deleterious mutations are highlighted in bold.

This study reports the complete A2 gene continuous sequence and an assembly of an entire A2 cluster including the A2-rel flanking genes in *L. donovani*. While the organization of this A2 gene cluster was previously theorized based on available sequences and Southern blot analysis, no sequencing technology could accurately read-through an entire cluster prior to third-generation sequencers. Interestingly, a similar organization of A2 and A2-rel flanking genes was obtained during the resequencing and assembly of the *L. infantum* genome, further supporting the genomic arrangement of this important virulence cluster. The present assembly contains entire A2 ORFs that are consistent with the corresponding protein sizes determined by Western blot analysis and further supporting the genomic arrangement of this important virulence cluster. The present assembly contains

| Chr | Gene   | Mutation                  | PROVEAN | Protein Name                          |
|-----|--------|---------------------------|---------|---------------------------------------|
| 7   | LdBPK_070700 LdCL_070011900 | Ala282Val                | −0.743  | vascular-type Ca2± ATPase, putative   |
| 12  | LdBPK_120275 LdCL_120008300 | Gln115Asp                 | −0.258  | Myotubularin-related protein, putative|
| 14  | LdBCL_140017600             | Ser291ts                  | N/A     | kinesin k39                           |
| 14  | LdBPK_141190 LdCL_140017700 | Gln1034Asp                | −1.06   | kinesin K39                           |
| 22  | LdBPK_220840 LdCL_220015800 | Pro219FS                  | N/A     | hypothetical protein                  |
| 23  | LdBCL_230017500             | INS:446Glu*               | −12.453 | sucrose hydrolase-like protein        |
| 25  | LdBPK_250620 LdCL_250011400 | Ala969Glu                 | 0.736   | Raptor N-terminal CASPase like domain |
| 25  | LdBPK_250790 LdCL_250013200 | INS:110 Ala, Asn, Ser, Ala, Ala, Ala, Ala | N/A | hypothetical protein                  |
| 27  | LdBPK_270830 LdCL_270014900 | Ala1493Thr                | −0.25   | ATP-binding cassette protein subfamily A|
| 29  | LdBCL_290028400             | Thr208Ala                 | 0.4     | VIT family putative                   |
| 30  | LdBPK_301640 LdCL_300021700 | Gln334STOP                | N/A     | hypothetical protein                  |
| 31  | LdBPK_311390 LdCL_310020800 | STOP:1486 Leu, Ser, His   | 0       | hypothetical protein                  |
| 31  | LdBPK_311470 LdCL_310021600 | Thr498Ala                 | −0.15   | hypothetical protein                  |
| 31  | LdBPK_311470 LdCL_310021600 | His497Arg*                | 0.942   | hypothetical protein                  |
| 31  | LdBPK_311470 LdCL_310021600 | Gly380Asp*                | −0.383  | hypothetical protein                  |

**IV — VL mutations**

| Chr | Gene   | Mutation                  | PROVEAN | Protein Name                          |
|-----|--------|---------------------------|---------|---------------------------------------|
| 33  | LdBPK_230014900 | Asp712Glu                | −1.625  | hypothetical protein, unknown function|
| 31  | LdBPK_3123870 LdCL_310037100 | Met189Thr                | −4      | hypothetical protein, unknown function|
| 31  | LdBPK_313290 LdCL_310041200 | Val187Phe                | −0.634  | Hypothetical protein                  |
| 34  | LdBPK_342210 LdCL_340029800 | Thr116DEL                | −1.098  | hypothetical protein                  |

**IV-Only mutations**

| Chr | Gene   | Mutation                  | PROVEAN | Protein Name                          |
|-----|--------|---------------------------|---------|---------------------------------------|
| 14  | LdBPK_140470 LdCL_140010000 | Gln89Lys                | −0.044  | cystathionine beta-lyase-like protein |
| 31  | LdBPK_312810 LdCL_310036400 | Cys173Phe                | −9.5    | regulator of chromosome condensation (RCC1) repeat, putative |
| 32  | LdBPK_312270 LdCL_310035800 | Gly675Ser                | −1.292  | hypothetical protein                  |
| 32  | LdBPK_324000 LdCL_320046000 | Val250Ile                | 0       | hypothetical protein, unknown function|
| 36  | LdBPK_361580 LdCL_360021300 | Gene deletion | N/A     | Serine/Threonine Kinase, putative      |
| 36  | LdBPK_361590 LdCL_360021400 | Gene deletion | N/A     | Serine/Threonine Kinase, putative      |
| 36  | LdBPK_361600 LdCL_360021500 | Gene deletion | N/A     | Engulfment and cell motility domain 2, putative |
| 36  | LdBPK_361610 LdCL_360021600 | Gene deletion | N/A     | Predicted tripartite motif protein     |
strain had significantly increased survival in visceral organs\(^2\)\(^4\), it was surprising that this strain did not have more genetic differences associated with the increased virulence. It is possible that the selection process for survival in the visceral organs of mice is different from that in humans.

The Illumina sequence analysis of the cutaneous (CL) and visceral (VL) disease associated strains using the complete assembly identified 15 novel homozygous SNPs beyond the previously identified 70 SNPs\(^4\)(Fig. 5). One of these new SNPs was in the Raptor gene that is part of the highly conserved Target Of Rapamycin (TOR) signaling pathway\(^29\). There are three TOR gene homologs in the \textit{Leishmania} genome\(^30\) revealing this pathway is conserved in kinetoplastids. Interestingly, the RagC GTPase which is a binding partner of Raptor in the TOR pathway is also mutated in the attenuated cutaneous \textit{L. donovani} strain and restoration of the wildtype RagC GTPase increased virulence in visceral organs\(^4\). Considering that there are two mutated genes (RagC and Raptor) in the attenuated cutaneous \textit{L. donovani} strain strongly highlights this pathway as playing a role in determining disease tropism and virulence.

As both HIVE and VarScan were used to identify SNPs and indels, we are confident that the expanded list of 83 variable genes shown in Fig. 6 contains most if not all the genes associated with visceral disease, with the exception of UTR mutations that may influence protein expression levels. Since this number of genes is relatively small, we are currently investigating all genes in clusters 1–4 with respect to their involving in visceral organ virulence using CRISPR-Cas9 gene editing recently developed for use in \textit{Leishmania}\(^31,32\). It is noteworthy that the correct selection of gRNA sequences for CRISPR-Cas9 gene editing requires a complete genome and accurate annotations for precise gene editing with no off-target mutations that is now possible with the complete assembly reported here.

**Methods**

**Whole genome sequencing.** DNA extraction. \textit{Leishmania} DNA for both Illumina and PacBio sequencing was derived from the attenuated cutaneous strain of \textit{L. donovani} from Sri Lanka\(^4\) that was passaged through mice to increase survival in visceral organs\(^2\)\(^4\). DNA was extracted following the previously described phenol-chloroform methods for isolation of \textit{Trypanosomatid} genomic material\(^33\).

**Illumina.** Sequencing library preparation (Kapa HTP) and 250 nt paired-end sequencing (Illumina MiSeq) was performed using manufacturers’ protocols.

**PacBio sequencing.** A total of 9 sequencing cells were prepared. 7 cells were prepared using the DNA Template kit v2.0 (3–10 kb) with DNA/Polymerase Binding Kit P4 and 2 using the DNA Template Prep Kit 3.0 with DNA/Polymerase Binding Kit P5. The cells were sequenced on a PacBio II RS instrument with BaseCaller v1 protocol.

**Genome assembly.** HGAP assembler. Raw reads from the 9 sequencing cells were loaded into the SMRT Analysis portal (Pacific Biosciences) in HD5 format. The Hierarchical Genome Assembly Process (HGAP) version 2 with Quiver polishing was chosen as version 3 is stated to improve speed at the detriment of assembly quality. Expected genome size was set to 36Mbp, minimum read length for pre-assembly was set to 500 bp and minimum read length for full assembly was set to 100 bp. Minimum Polymerase Read quality was set to 0.80, and the remainder of options remained at default settings.
| Cluster Number | Cluster Mutation Type                                                                 | New annotation                  | Equivalents (when available) |
|---------------|---------------------------------------------------------------------------------------|--------------------------------|------------------------------|
|               | Nonsense, Frameshift, Insertions, Deletions, IV to VL                                  | LdCL_300021700 LdBPK_301640    |                              |
|               |                                                                                       | LdCL_310020800 LdBPK_311390    |                              |
|               |                                                                                       | LdCL_250013200 LdBPK_250790    |                              |
|               |                                                                                       | LdCL_310020100 LdBPK_311320    |                              |
|               |                                                                                       | LdCL_310022200 LdBPK_311510    |                              |
|               |                                                                                       | LdCL_080011700 LdBPK_080670    |                              |
|               |                                                                                       | LdCL_340029800 LdBPK_342210    |                              |
|               |                                                                                       | LdCL_230014900 LdBPK_230830    |                              |
|               |                                                                                       | LdCL_310037100 LdBPK_312870    |                              |
|               |                                                                                       | LdCL_220015800 —                |                              |
|               |                                                                                       | LdCL_140017600 —                |                              |
|               |                                                                                       | LdCL_230017500 —                |                              |
|               |                                                                                       | LdCL_310041200 LdBPK_313290    |                              |
|               | Multiple SNPs in the same gene, Non-conservative amino acid change in conserved region with good PROVEAN score | LdCL_270015000 LdBPK_270840    |                              |
|               |                                                                                       | LdCL_290026900 LdBPK_292100    |                              |
|               |                                                                                       | LdCL_310021600 LdBPK_311470    |                              |
|               |                                                                                       | LdCL_310028800 LdBPK_312080    |                              |
|               |                                                                                       | LdCL_340046300 LdBPK_343690    |                              |
|               |                                                                                       | LdCL_290022800 LdBPK_291720    |                              |
|               |                                                                                       | LdCL_310024300 LdBPK_311710    |                              |
|               |                                                                                       | LdCL_360060000 LdBPK_360120    |                              |
|               |                                                                                       | LdCL_360062000 LdBPK_365480    |                              |
|               | Non-conservative amino acid change in conserved region with poor PROVEAN score         | LdCL_070018300 LdBPK_071330    |                              |
|               |                                                                                       | LdCL_320013800 LdBPK_320820    |                              |
|               |                                                                                       | LdCL_250016900 LdBPK_251150    |                              |
|               |                                                                                       | LdCL_220022000 LdBPK_221470    |                              |
|               |                                                                                       | LdCL_250015300 LdBPK_251000    |                              |
|               |                                                                                       | LdCL_040011100 LdBPK_040560    |                              |
|               |                                                                                       | LdCL_360016300 LdBPK_361120    |                              |
|               |                                                                                       | LdCL_200014300 LdBPK_200960    |                              |
|               |                                                                                       | LdCL_090011700 LdBPK_090660    |                              |
|               |                                                                                       | LdCL_130016200 LdBPK_131090    |                              |
|               |                                                                                       | LdCL_340009000 LdBPK_340390    |                              |
|               |                                                                                       | LdCL_130017800 LdBPK_131230    |                              |
|               |                                                                                       | LdCL_230009900 LdBPK_230440    |                              |
|               |                                                                                       | LdCL_220018100 LdBPK_221070    |                              |
|               |                                                                                       | LdCL_340044900 LdBPK_343550    |                              |
|               |                                                                                       | LdCL_290028400 —                |                              |
|               |                                                                                       | LdCL_250011400 LdBPK_250620    |                              |
|               |                                                                                       | LdCL_270014900 LdBPK_270830    |                              |
|               | Conservative amino acid change in conserved region with good PROVEAN score            | LdCL_350013100 LdBPK_350830    |                              |
|               |                                                                                       | LdCL_360052700 LdBPK_364550    |                              |
|               |                                                                                       | LdCL_230026600 LdBPK_231940    |                              |
|               |                                                                                       | LdCL_230009400 LdBPK_230400    |                              |
|               |                                                                                       | LdCL_320031100 LdBPK_322560    |                              |
|               |                                                                                       | LdCL_020008200 LdBPK_020280    |                              |
|               |                                                                                       | LdCL_310027700 LdBPK_311990    |                              |
|               |                                                                                       | LdCL_320031200 LdBPK_322570    |                              |
|               | Conservative amino acid change in conserved region with poor PROVEAN score           | LdCL_330011900 LdBPK_330640    |                              |
|               |                                                                                       | LdCL_170010200 LdBPK_170470    |                              |
|               |                                                                                       | LdCL_070019900 LdBPK_070700    |                              |
|               |                                                                                       | LdCL_210025000 LdBPK_211930    |                              |
|               |                                                                                       | LdCL_290029900 LdBPK_291730    |                              |
|               |                                                                                       | LdCL_060083000 LdBPK_060140    |                              |
|               |                                                                                       | LdCL_290029000 LdBPK_292290    |                              |
|               |                                                                                       | LdCL_03007500 LdBPK_030250     |                              |
|               |                                                                                       | LdCL_250006200 LdBPK_250110    |                              |
|               |                                                                                       | LdCL_360015800 LdBPK_361070    |                              |
|               |                                                                                       | LdCL_310023400 LdBPK_311630    |                              |
|               |                                                                                       | LdCL_360062700 LdBPK_365540    |                              |
|               |                                                                                       | LdCL_140017700 LdBPK_141190    |                              |

Continued
Celera assembler. The PacBio corrected Reads (PBcR) module of the celera-assembler version 8.3 was used to assemble the long reads data. The subreads were first extracted from the PacBio H5 files to FASTQ using bash5tools.py. The Bogart unitigger was used by specifying the "unitigger = bogart" option in the spec file. The consensus caller module was PBDAGCON. Due to the sequences originating from a non-clonal sample and the use of the DNA/Polymerase Binding kit P4 in some PacBio sequencing cells which produces lower quality data than P5 kits, error rate limits were relaxed for various variables, listed in the full spec file available in supplementary information (Supplementary Methods S1).

Canu assembler. The Canu v1.0 assembler is a modified version of the Celera Assembler designed to handle high noise data such as NanoPore and PacBio sequencing data. Canu has both the ability to assemble raw PacBio data by performing error correction using consensus sequence or assemble data in a hybrid mode where PacBio reads are pre-error corrected using short read Illumina data. In Raw mode, the trimmed PacBio reads were given to the assembler using default settings except for the expected genome size option which was set to 35Mbp using the option "genomeSize = 35 m". In hybrid mode, the Illumina reads were first error corrected by internal consensus using Pollux\textsuperscript{35}, the paired end reads were then merged together to form longer sequences with a high confidence core region using FLASh\textsuperscript{36}, and used to correct the PacBio reads using Proovread\textsuperscript{37}. The error-corrected PacBio data was then used by Canu to generate a draft assembly.

Pilon. The Pilon error correcting software was used to fix small errors present in the PacBio based assemblies using high depth and high accuracy Illumina data\textsuperscript{38}. The entire MiSeq dataset in FASTQ format from the corresponding sample was aligned to the draft assembly using the Burrows-Wheeler Aligner (bwa) to generate SAM alignment files. Samtools was then used to convert and sort the files to a binary usable format as described in the Samtools section. This alignment was then passed to the Java Pilon executable for correction of small indels, SNPs, gap filling and assembly of unmapped reads using the command "java -Jar pilon.jar -genome [new-assembly.fasta] -frags [alignment.bam] -fix all,novel".

GMcloser. GMcloser was used to merge the assemblies generated using different assemblers\textsuperscript{39}. Short read Illumina data was aligned to the contigs resulting from the different contigs from different assemblers with identical reads mapped to them were assumed to be part of the same chromosome. When a contig from one assembly encompassed a gap present in another assembly, the gap was filled with the missing information to generate a merged assembly with the least number of gaps. All the alignment and merging steps are handled internally to GMcloser using the command "gmcloser -t [assembly1.fasta] -q [assembly2.fasta] -r mismq_R1.fastq mismq_R2.fastq -et".

### Table 3. Summary of all genes containing mutations in the cutaneous isolates and classification into clusters.

Entries were not repeated in multiple lists. Identified mutations were further classified into priority clusters for effect on protein function and future analysis for genes associated with survival in visceral organs. Mutations were prioritized by likelihood of contributing to visceral tissue tropism by severity of the coding change, accumulation of secondary mutations and conservation. Gene loci listed from the current assembly as well as previous ID numbers when available.

| Cluster Number | Cluster Mutation Type | New annotation | Equivalents (when available) |
|----------------|-----------------------|----------------|-----------------------------|
| Cluster 6 (14) | Non-conservative amino acid change in less conserved region, Conservative amino acid change in less conserved region | LdCL_340022100 LdBPK_341580 | |
| | | LdCL_060016600 LdBPK_060650 | |
| | | LdCL_210015400 LdBPK_211040 | |
| | | LdCL_050010900 LdBPK_050580 | |
| | | LdCL_230011600 LdBPK_230610 | |
| | | LdCL_250024100 LdBPK_251840 | |
| | | LdCL_070015100 LdBPK_071060 | |
| | | LdCL_250005300 LdBPK_250840 | |
| | | LdCL_310022100 LdBPK_311500 | |
| | | LdCL_200006800 LdBPK_200200 | |
| | | LdCL_250014400 LdBPK_250910 | |
| | | LdCL_230010400 LdBPK_230500 | |
| | | LdCL_120008300 LdBPK_120275 | |
| | | LdCL_290028100 LdBPK_292210 | |
| Cluster 7 (4) | IV-only mutations | LdCL_140010000 LdBPK_140470 | |
| | | LdCL_310035800 LdBPK_312770 | |
| | | LdCL_310036400 LdBPK_312810 | |
| | | LdCL_320046000 LdBPK_324000 | |
| | | LdCL_360021300 LdBPK_361580 | |
| | | LdCL_360021400 LdBPK_361590 | |
| | | LdCL_360021500 LdBPK_361600 | |
| | | LdCL_360021600 LdBPK_361610 | |
Necessary from the sequences available for nosomes. The snRNAs U1 through U6, ribosomal RNAs and the spliced leader RNA were manually annotated as jar mpileup2snp –output-vcf 1 [mpileup.file] (VCF) using the mpileup file generated by samtools as described above using the command “java -jar VarScan.

Specific position sorted BAM file the base distribution at every position using the command “samtools mpileup -B -f [reference assembly] [strain sorted by alignment location for compatibility with downstream analysis software using "samtools sort -@ 30 -o view -b” was used to convert the BW A generated SAM file to the binary alignment BAM format. The file was then

Synteny. The online SynMap2 software was used to generate the synteny dotplot across the entire genome using annotations from L. major and the annotations generated by Companion in this study. The chromosome to chromosome circular charts were generated by Companion as part of the annotation process.

Comparison of visceral (VL), cutaneous (CL) and increased virulence (IV) L. donovani strains. BWA. The Burrows-Wheeler Aligner (BWA) was used to process the FASTQ Illumina sequencing files obtained from Genome Quebec. The maximal exact match algorithm was used in paired-end mode using the command “bwa mem” and providing the matched pair read files and reference sequence as arguments in order to generate a SAM format alignment file of the reads on the reference.

Genomic annotations from the Companion Pipeline were downloaded in General Feature Format (GFF) and gene annotations were extracted using the Galaxy tool “Extract features” set to look for the “CDS” keyword in column #3 of the GFF file. Known coding regions from the reference L. donovani strain BPK282A1, assembly ASM22713v2 were downloaded from GenBank and aligned to our improved assembly in BED format. Bedtools intersect intervals through Galaxy was used to identify annotations that were unique to our annotations or were not at least 95% covered previously using settings “-wa -f 0.95 -v -r”.

Galaxy. The Galaxy webtool (https://usegalaxy.org/) was used to perform file conversions and data extraction such as moving a chromosome's FASTA sequence from one assembly to another.

Identification of new genes. Genomic annotations from the Companion Pipeline were downloaded in General Feature Format (GFF) and gene annotations were extracted using the Galaxy tool “Extract features” set to look for the “CDS” keyword in column #3 of the GFF file. Known coding regions from the reference L. donovani strain BPK282A1, assembly ASM22713v2 were downloaded from GenBank and aligned to our improved assembly in BED format. Bedtools intersect intervals through Galaxy was used to identify annotations that were unique to our annotations or were not at least 95% covered previously using settings “-wa -f 0.95 -v -r”.

Synteny. The online SynMap2 software was used to generate the synteny dotplot across the entire genome using annotations from L. major and the annotations generated by Companion in this study. The chromosome to chromosome circular charts were generated by Companion as part of the annotation process.

Comparison of visceral (VL), cutaneous (CL) and increased virulence (IV) L. donovani strains. BWA. The Burrows-Wheeler Aligner (BWA) was used to process the FASTQ Illumina sequencing files obtained from Genome Quebec. The maximal exact match algorithm was used in paired-end mode using the command “bwa mem” and providing the matched pair read files and reference sequence as arguments in order to generate a SAM format alignment file of the reads on the reference.

Samtools. The samtools package was used for file manipulations and conversions. The commands “samtools view -b” was used to convert the BWA generated SAM file to the binary alignment BAM format. The file was then sorted by alignment location for compatibility with downstream analysis software using “samtools sort -@ 30 -o [output.file]”. The alignment files where then prepared for analysis using the mpileup modules which tabulates the base distribution at every position using the command “samtools mpileup -B -f [reference assembly] [strain specific position sorted BAM file] > [output.file]”.

VarScan. The VarScan v2 mutation caller was used to generate a list of mutations in Variant Call Format (VCF) using the mpileup file generated by samtools as described above using the command “java -jar VarScan.jar mpileup2snp –output-vcf 1 [mpileup.file] > [output.VCF]”. We also used VarScan to generate indel locations based on the same mpileup file using the command “java -jar VarScan.jar mpileup2indel –output-vcf 1 [mpileup. file] > [output.vcf]”.

SnpEff. To filter the VCF files generated by VarScan to a list of non-synonymous SNP, we used the SnpEff software. The oriented and annotated assembly was downloaded from the Companion tool as described above along with the gene annotation file in GFF format containing the names, locations and amino acid sequences of identified genes. This GFF file was used to build a SnpEff database using the SnpEff.jar command “build” with argument “-gff” after installing the genome and GFF file in the appropriate locations according to the software instructions.

The SnpEff software was then used to annotate the 10th column of the VCF file with mutation effect codes. All the mutations were then examined manually for accuracy using the Integrative Genomics Viewer (IGV) with all raw Illumina data loaded.

Classification. Non-synonymous mutations were clustered according to the mutation effect in order to prioritize further gene function studies. Each cluster was further broken down based on the mutation’s PROVEAN score. The PROVEAN software was designed to predict the magnitude of a mutation’s impact on protein function. To generate PROVEAN scores, we retrieved homologous sequences from other Leishmania species and kinetoplasts and generated a multiple sequence alignment (MSA). The MSA was then passed to the PROVEAN software which scored each SNP based on the alignment. We used PROVEAN scores below a threshold of –2.5 as an indication a SNP is likely to affect protein function. Cluster assignments were as follows:
1. Mutations likely to have the largest impact on protein function were included; non-sense, frameshift and amino acid insertion/deletions as well as all SNPs in the gain-of-function IV strain that were the same in the virulent visceral strain (VL) allele, indicating a selection pressure on those genes for visceral organ survival. 

2. Genes in with multiple SNPs and genes where non-conservative mutations occurred in highly conserved Leishmania/Kinetoplastida regions. 

3. Due to the high number of genes in cluster 2, split off poor PROVEAN scoring genes. 

4. Cluster 4 comprised genes with conservative amino acid changes but occurring in Leishmania/Kinetoplastida conserved regions. 

5. Due to the high number of genes in cluster 4, split off poor PROVEAN scoring genes. 

6. Conservative and non-conservative amino acid changes in less conserved regions. 

7. Changes present only in the gain of function IV strain. This cluster was considered low probability as it likely contains either random mutations or adaptations specific to survival in the murine host. 

**HIVE.** HIVE\(^52\) was used to perform differential profiling of genomes from visceral (VL), gain of function increased virulence (IV), and (cutaneous) CL strains. 

1. Reads from all the samples were aligned to the assembly of the genome using HIVE-hexagon\(^53\) parameterized for parasitic eukaryotic species and specifically adjusted to work with Leishmania analysis as demonstrated in previous studies\(^27\). 

2. Coverage and variant calling analysis was performed using HIVE-heptagon\(^54\) to produce variant call frequencies and coverages for every genomic position. 

3. HIVE differential profiler\(^52\) was used to analyze relative differences in SNP calls and variant coverages for multiple samples. 

**A2 Immunoblotting.** A2 Immunoblotting was performed as previously described\(^45\). Briefly, 1 × 10⁷ cutaneous CL strain promastigotes were collected at mid log-phase and resuspended in 1 mL fresh medium. The cells were then heat-shocked for 4 h at 40 °C to induce A2 protein expression, washed, lysed in SDS-PAGE loading buffer and loaded on a 10% (w/v) acrylamide gel. The proteins were transferred to nitrocellulose at 25 V overnight at 4 °C. The membrane was blocked for 1 h in 10% (w/v) skim milk powder dissolved in PBS with 0.1% (v/v) Tween-20. The membrane was then incubated for 1 h at RT with a 1:10,000 dilution of C9 Ascites fluid (anti-A2 Mab) in blocking solution followed by 6 × 5 min washes in PBS-T. Secondary HRP labeled anti-mouse IgG antibody (Thermo Fisher Scientific) was incubated at 1:10,000 in blocking buffer for 1 h at RT followed by 6 × 5 min washes in PBS-T. The membrane was incubated in ECL reagent (Zm Tech) for 1 min at RT before being exposed to x-ray film (Denville Scientific). Film images were captured using a Gel-Doc XR documentation system with Quantity One software (BioRad Laboratories). 

**Data Availability** 

All data used in this study have been deposited online with GenBank. Raw PacBio reads for the IV strain, Illumina MiSeq reads for the CL, VL and IV strains, the new genome assembly and the annotations generated in this study can be found under the PRJNA450813 BioProject accession number. 

**Accession codes.** GenBank BioProject PRJNA450813. 

**References** 

1. Alvar, J. et al. Leishmaniasis worldwide and global estimates of its incidence. PLoS One 7, e35671 (2012). 
2. Peacock, C. S. et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nat. Genet. 39, 839–847 (2007). 
3. Zhang, W. W. & Matlashewski, G. Screening Leishmania donovani-specific genes required for visceral infection. Mol. Microbiol. 77, 505–517 (2010). 
4. Zhang, W. W. et al. Genetic Analysis of Leishmania donovani Tropism Using a Naturally Attenuated Cutaneous Strain. PLoS Pathog. 10, e1004244 (2014). 
5. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351 (2016). 
6. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864 (2017). 
7. Grisard, E. C. et al. Trypanosoma cruzi Clone Dm28c Draft Genome Sequence. Genome Announc. 2, 2–3 (2014). 
8. Downing, T. et al. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res. 21, 2143–2156 (2011). 
9. Ivens, A. C. et al. The genome of the kinetoplastid parasite, Leishmania major. Science (80-.). 309, 436–442 (2005). 
10. Rogers, M. B. et al. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. Genome Res. 21, 2129–2142 (2011). 
11. Alonso, G., Rastrojo, A., López-Pérez, S., Requena, J. M. & Aguado, B. Resequencing and assembly of seven complex loci to improve the Leishmania major (Friedlin strain) reference genome. Parasites and Vectors 9 (2016). 
12. Mikhayev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. Mol. Ecol. Resour. 14, 1097–1102 (2014). 
13. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. Genomics, Proteomics Bioinforma. 13, 278–289 (2015). 
14. Zhang, W. W. & Matlashewski, G. Loss of virulence in Leishmania donovani deficient in an amastigote-specific protein, A2. Proc. Natl. Acad. Sci. 94, 8897–8811 (1997). 
15. Zhang, W. W. et al. Comparison of the A2 gene locus in Leishmania donovani and Leishmania major and its control over cutaneous infection. J. Biol. Chem. 278, 35508–35515 (2003).
16. McCall, L. I. & Matlashewski, G. Localization and induction of the A2 virulence factor in Leishmania: Evidence that A2 is a stress response protein. Mol. Microbiol. 77, 518–530 (2010).
17. McCall, L. I. & Matlashewski, G. Involvement of the Leishmania donovani virulence factor A2 in protection against heat and oxidative stress. Exp. Parasitol. 132, 109–115 (2012).
18. González-De La Fuente, S. et al. Resequencing of the Leishmania infantum (strain JPCM5) genome and de novo assembly into 36 contigs. Sci. Rep. 7, (2017).
19. Karunaweerla, N. D., Pratlong, F., Siriwandane, H. Y. V. D., Ihalamulla, R. L. & Dedet, J. P. Sri Lankan cutaneous leishmaniasis is caused by Leishmania donovani zymodeme MON-37. Trans. R. Soc. Trop. Med. Hyg. 97, 380–381 (2003).
20. Ranasinghe, S. et al. Leishmania donovani zymodeme MON-37 isolated from an autochthonous visceral leishmaniasis patient in Sri Lanka. Pathog. Glob. Health 106, 421–424 (2012).
21. Singh, N., Chikara, S. & Sundar, S. SOLID™ Sequencing of Genomes of Clinical Isolates of Leishmania donovani from India Confirm Leptomonas Co-Infection and Raise Some Key Questions. PLoS ONE 8, e55738 (2013).
22. McCall, L. I., Zhang, W. W. & Matlashewski, G. Determinants for the Development of Visceral Leishmaniasis Disease. PLoS Pathogens 9, (2013).
23. Steinbiss, S. et al. Companion: a web server for annotation and analysis of parasite genomes. Nucleic Acids Res. 44, W29–W34 (2016).
24. McCall, L. I. et al. Adaptation of Leishmania donovani to cutaneous and visceral environments: In vivo selection and proteomic analysis. J. Proteome Res. 14, 1033–1059 (2015).
25. Choi, Y. & Chan, A. P. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31, 2745–2747 (2015).
26. Coughlan, S. et al. The genome of Leishmania adleri from a mammalian host highlights chromosome fission in Sauroleishmania. Sci. Rep. 7 (2017).
27. Gannavaram, S. et al. Whole genome sequencing of live attenuated Leishmania donovani parasites reveals novel biomarkers of attenuation and enables product characterization. Sci. Rep. 7 (2017).
28. Zhang, W. W. & Matlashewski, G. Characterization of the A2-A2rel gene cluster in Leishmania donovani: Involvement of A2 in viralization during infection. Mol. Microbiol. 39, 935–948 (2001).
29. Sancak, Y. et al. The rag GTPases bind raptor and mediate amino acid signaling to mTORC1. Science (80-.). 320, 1496–1501 (2008).
30. Madeira da Silva, L. & Beverley, S. M. Expansion of the target of rapamycin (TOR) kinase family and function in Leishmania shows that TOR3 is required for acidocalcisome biogenesis and animal infectivity. Proc. Natl. Acad. Sci. 107, 11965–11970 (2010).
31. Zhang, W. W. & Matlashewski, G. CRISPR-Cas9-mediated genome editing in Leishmania donovani. MolBio 6, e00861–15 (2015).
32. Zhang, W.-W., Lypaczewski, P. & Matlashewski, G. Optimized CRISPR-Cas9 Genome Editing for Leishmania and Its Use To Target a Multigene Family, Induce Chromosomal Translocation, and Study DNA Replication Mechanisms. mSphere 2, 1–15 (2017).
33. Medina-Acosta, E. & Cross, G. A. M. Rapid isolation of DNA from trypanosomatid protozoa using a simple ‘mini-prep’ procedure. Mol. Biochem. Parasitol. 59, 327–329 (1993).
34. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33, 623–630 (2015).
35. Marinier, E., Brown, D. G. & McComkey, R. J. Pollux: Platform independent error correction of single and mixed genomes. BMC Bioinformatics 16, (2015).
36. Magoč, T. & Salzberg, S. L. FLASH: Fast adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–2963 (2011).
37. Hackl, T., Hedrich, R., Schultz, J. & Förster, F. Proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 30, 3004–3011 (2014).
38. Walker, B. J. et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, (2014).
39. Kosugi, S., Hirakawa, H. & Tabata, S. GMcloser: Closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. Bioinformatics 31, 3733–3741 (2015).
40. Robinson, T. et al. Integrative genomics viewer. Nature Biotechnology 29, 24–26 (2011).
41. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192 (2013).
42. Aslett, M., Aurrecoechea, C., Berriman, M. & Al, E. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 38, D457–62 (2010).
43. Elia, D. et al. Genome-wide analysis of small nucleolar RNAs of leishmania major reveals a rich repertoire of RNAs involved in modification and processing of rRNA. RNA Biol. 12, 1222–1255 (2015).
44. Afgan, E. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 44, W3–W10 (2016).
45. Quintan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
46. Gruening, B. A. Galaxy wrapper (2014).
47. Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E. & Lyons, E. SynMap2 and SynMap3D: Web-based whole-genome synteny browsers. Bioinformatics 33, 2197–2198 (2017).
48. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3384v1 (2013).
49. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
50. Koboldt, D. C. et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576 (2012).
51. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnPEff. Fly (Austin) 6, 80–92 (2012).
52. Simonyan, V. & Mazumder, R. High-performance integrated virtual environment (hive) tools and applications for big data analysis. Genes (Basel). 5, 957–981 (2014).
53. Santana-Quintero, L., Dingerdissen, H., Thierry-Mieg, J., Mazumder, R. & Simonyan, V. HIVE-heptagon: High-performance, parallelized sequence alignment for next-generation sequencing data analysis. PLoS ONE 9 (2014).
54. Simonyan, V. et al. HIVE-heptagon: A sensible variant-calling algorithm with post-alignment quality controls. Genomics 109, 131–140 (2017).

Acknowledgements
GM acknowledges the support of the Canadian Institutes of Health Research and the Global Health Innovative Technologies Fund. PL acknowledges receiving a doctoral training award from the Fonds de Recherche du Quebec Santé. LIM was funded by a graduate fellowship from the Canadian Institutes of Health Research (#MOP235928). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Author Contributions
P.L. wrote the main manuscript and performed the analysis. J.H., J.T.R. and A.M. performed analysis. W.Z., L.I.M., V.S. and K.D. provided materials and insight. G.M. planned the experiment and wrote the manuscript. All authors reviewed the manuscript.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-34812-x.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018