Genomes of intracellular *Leishmania* parasites directly sequenced from patients

Running title: Genomes of *Leishmania* sequenced in clinical samples

Malgorzata A. Domagalska[a#], Hideo Imamura[a], Mandy Sanders[b], Frederik Van den Broeck[a], Narayan Raj Bhattarai[c], Manu Vanaerschot[a*], Ilse Maes[a], Erika D'Haenens[a**], Keshav Rai[c], Suman Rijal[c], Matthew Berriman[b], James A. Cotton[b#], Jean-Claude Dujardin[a,d#].

Author Affiliations:

[a] Institute of Tropical Medicine Antwerp, Molecular Parasitology Unit, Antwerp, Belgium;

[b] Wellcome Sanger Institute, Hinxton, United Kingdom

[c] BP Koirala Institute of Health Sciences, Dharan, Nepal

[d] University of Antwerp, Department of Biomedical Sciences, Antwerp, Belgium.

[#] Corresponding authors: email address: Malgorzata A. Domagalska (mdomagalska@itg.be), Jean-Claude Dujardin (jcdujardin@itg.be), James A. Cotton (jc17@sanger.ac.uk)

* current affiliation: Columbia University, Department of Microbiology and Immunology, New York, USA

** current affiliation: University Hospital of Ghent, Center for Medical Genetics, Ghent, Belgium

Keywords: Leishmania, whole genome sequencing, genome capture, aneuploidy

Text: 5778 words; abstract: 214 words

Abstract

Whole genome sequencing (WGS) is increasingly used for molecular diagnosis and epidemiology of infectious diseases. Current *Leishmania* genomic studies rely on DNA extracted from cultured parasites, which might introduce sampling and biological biases into the subsequent analyses. Up to now, direct analysis of *Leishmania* genome in clinical samples is hampered by high levels of human DNA and large variation in parasite load in patient samples. Here, we present a method, based on target enrichment of *Leishmania donovani* DNA with Agilent SureSelect technology, that allows the analysis of *Leishmania* genomes directly in clinical samples. We validated our protocol with a set of artificially mixed samples, followed by the analysis of 63 clinical samples (bone marrow or spleen aspirates) from visceral leishmaniasis patients in Nepal. We were able to identify genotypes using a set of diagnostic SNPs in almost all of these samples (97%) and access comprehensive genome-wide information in most (83%). This allowed us to perform phylogenomic analysis, assess chromosome copy number and identify large copy number variants (CNVs). Pairwise comparisons between the parasite genomes in clinical samples and derived *in vitro* cultured promastigotes showed a lower aneuploidy in amastigotes as well as genomic differences, suggesting polyclonal infections in patients. Altogether our results underline the need for sequencing parasite genomes directly in the host samples.

48 **Importance**

49

50 Visceral leishmaniasis (VL) is caused by parasitic protozoa of the *Leishmania donovani*
51 complex and is lethal in the absence of treatment. Whole Genome Sequencing (WGS) of *L.*
52 *donovani* clinical isolates revealed hitherto cryptic population structure in the Indian Sub-
53 Continent and provided insights into the epidemiology and potential mechanisms of drug
54 resistance. However, several biases are likely introduced during the culture step. We report
55 here the development of a method that allows determination of parasite genomes directly
56 in clinical samples, and validate it on bone marrow and splenic aspirates of VL patients in
57 Nepal. We find genomic differences between the parasites analyzed directly in clinical
58 samples and the derived isolates, further supporting the need for direct analyses in host
59 samples. We believe this method is most suitable for clinical studies and for molecular
60 tracking in the context of elimination programs.

61

62 **Introduction**

63 *Leishmania* (Kinetoplastida, Trypanosomatidae) is a genus of parasitic protozoa
64 transmitted by sand flies to a variety of mammal hosts including humans. Within mammals,
65 the amastigote-stage parasites are strictly intracellular and infect a range of professional
66 phagocytes. *Leishmania* cause a broad spectrum of clinical presentations, the most severe
67 being visceral leishmaniasis (VL) which annually affects up to 90,000 new individuals (1) and
68 has a 10% mortality rate (2). Ongoing control programs, like the one running in the Indian
69 subcontinent (ISC) (3) could be challenged by several factors, including drug resistance (4,5)
70 and emergence of new parasite variants (6,7). More than ever, close surveillance is required
71 to avoid new epidemics, including molecular tracking of the etiological agent.

72 Whole genome sequencing (WGS) is increasingly used for highly discriminatory
73 genotyping of parasites. Using that method, we previously resolved the hitherto cryptic
74 population structure of *Leishmania donovani* in the ISC (8). We identified a core group (CG)
75 – associated with the successive VL epidemics in the lowlands – itself structured into (i) six
76 congruent monophyletic groups (ISC2-7) together with (ii) three other groups (ISC8-10) and
77 several ungrouped isolates, with a less certain evolutionary history. We found in the
78 Nepalese highlands an emerging population (ISC1), quite different from the CG (6,9). We
79 described an unprecedented level of aneuploidy and found genetic variants likely to
80 underpin reduced efficacy of antimonial drugs in that region.

81 However, such studies suffer a major limitation: they currently rely on sequencing
82 parasites isolated from patients and grown axenically *in vitro*, with two possible biases. First,
83 there could be a major representation bias, because of the variable isolation success and
84 the high frequency (up to 90%) of asymptomatic cases (10), currently unsampled. Second,
85 we have shown experimentally major differences in the genome of *L. donovani* during the

86 life cycle: in particular, aneuploidy was much lower and affected other chromosomes in the
87 intra-cellular life stages (amastigotes) of infected hamsters, in comparison to cultivated
88 promastigotes (11). This genome plasticity may affect gene dosage and therefore have a
89 biological impact via transcriptome and proteome changes (12). If the phenomenon also
90 occurs between amastigotes from human subjects and *in vitro* cultivated promastigotes, it
91 could have a major impact on the search for molecular bio-markers for clinically important
92 traits such as drug resistance or virulence, as relevant characters present in the host could
93 be diluted or disappear in culture.

94 Applying parasite WGS directly in host samples thus appears critical for next
95 generation genetic analyses of natural populations of *Leishmania*. Such applications are
96 challenged by the low amounts of *Leishmania* DNA in biological samples and the high
97 abundance of DNA in the nucleated host cells. Accordingly, methods for enrichment of
98 parasite DNA are required before undertaking WGS. A series of approaches have been used
99 for other pathogens, including (i) removal of methylated host DNA (13), (ii) selective whole
100 genome amplification (14) or (iii) targeted genome enrichment (15). Here, we explored the
101 last solution. Using Agilent SureSelect technology, we designed an array for application in
102 the *L. donovani* species complex. In this method an Illumina-based NGS sequencing library is
103 prepared from the mixed gDNA sample, which is later hybridized with biotinylated RNA
104 baits complementary to the *Leishmania* genome and the fraction of the input gDNA that
105 anneals to the baits is then sequenced (further called SuSL-seq). We validated and
106 optimized this genome capture procedure with artificial mixtures of *Leishmania* and host
107 DNA to simulate the proportions expected in clinical samples. Finally, we applied the
108 method on a set of 63 bone marrow and splenic aspirates from Nepalese VL patients. In 12

109 of them, we compared the karyotype and genome sequence with that of isolates derived
110 from the same tissue samples.
111
112 **Results**
113
114 **SuSL-seq analysis of clinical samples: sequencing statistics**
115 Details on the SureSelect bait design are described in Text S1 and in Fig. S1A. SuSL-
116 seq was optimized on artificial mixtures (*Leishmania* DNA diluted in mammalian DNA in
117 different ratios: see details and metrics in Text S1 and Figs S1B-C and S2): a percentage of
118 *Leishmania* DNA of 0.006% was found to be the lowest limit for suitable analysis of genome
119 diversity. SuSL-seq was then applied to 63 clinical samples with *Leishmania* DNA percentage
120 of at least 0.006%, 59 bone marrow and 4 splenic aspirate samples collected from VL
121 patients in Nepal (Table A1) in the same region where we previously studied genome
122 diversity of *L. donovani* isolates (Fig. S3).
123 Among all SuSL-seq samples, the median depth of reads mapped against the *L.*
124 *donovani* reference genome was 22.5, the 5th and 95th percentiles of median depth were 2
125 and 74.16, respectively, while the maximum depth was 126.5 (in BPK89BM) and the
126 minimum average depth was 0.074 (in BPK161BM). Across all clinical SureSelect libraries,
127 the mean percentage of reference genome bases covered by at least 5 reads was 61.1%
128 and, the mean percentage of bases covered by at least 1 read was 78.4%. This coverage
129 average was skewed by a small number of lower-quality samples. For instance, within a
130 sequencing batch of 24 samples, over 90% of the genome was covered by at least one read
131 in 14 samples; in all cases better genome representation was obtained following SureSelect
132 enrichment (Fig. S4A). Overall, as for the artificial mixtures, we obtained the highest

133 enrichment for samples with lowest parasite loads. An approximately linear relationship was
134 found between Leishmania DNA percentage in clinical samples and the proportion of reads
135 mapping to *Leishmania* after SuSL-seq for samples with up to 1 % of *Leishmania* DNA (Fig.
136 1A). For samples with higher *Leishmania* content, no further increase was observed beyond
137 around 60-70% of reads aligning to BPK282v2 reference genome (Fig. 1A). Eleven samples
138 had low coverage, with less than 10% of the genome covered at a depth of 5x (Fig. 1B).
139 These low-coverage samples also had the lowest parasite load, and were excluded from the
140 analyses described below, so 52 samples were studied in greater detail.
141
142 **SuSL-seq analysis of clinical samples: genetic variants**
143 Using an improved version of the BPK282 reference genome BPK282v2 (11), we
144 were able to call SNPs in these 52 samples, of which 51 were identified as belonging to CG
145 and one belonging to the genetically distinct ISC1 sub-population (8). We identified 197
146 variable sites that had sequence coverage and could therefore be reliably genotyped in all
147 51 CG clinical samples; these sites were also characterized in 191 previously sequenced
148 parasite isolates from the CG (8) and 9 newly sequenced isolates (see below). The 197 sites
149 were sufficient to reconstruct a phylogeny consistent with our previously published analysis
150 (8), placing all clinical SureSelect samples within the previously identified ISC diversity for *L.*
151 *donovani* (dots in Fig. 2A). Out of the 51 CG samples, 40 clustered in one of the previously
152 defined groups (ISC3-6, and ISC9). Five samples clustered at the base of ISC4. Six clinical
153 samples, clustered with isolates previously designated as 'ungrouped' (Fig.2A and Figs S5A
154 and B), forming a novel monophyletic group that we have named ISC11. We verified the
155 accuracy of our experimental and bio-informatic pipeline by including in the phylogenomic
156 analyses SuSL-seq data from artificial mixtures (DNA of BPK282 diluted in human DNA): as

157 expected, they clustered in the ISC6 group with the original BPK282 sequence data (not
158 shown).

159 Notably, among the 11 samples that were excluded from the phylogenomic analysis
160 because of poor genome-wide coverage, nine could be attributed to a specific ISC group, on
161 the basis of diagnostic SNPs specific to these groups (labeled with * in Table A1). The
162 remaining two samples, BPK211BM and BPK161BM, had sequence data with too low
163 coverage for any genotyping (Dataset S1A). Analysis of genotype distribution per village
164 revealed sympatric genotypic diversity: for instance, in Itahari, four different ISC types were
165 sampled over 11 days in June 2002 and three were sampled over four months in 2003
166 (Dataset S1B). Interestingly, in this village, two cohabiting members of the same family were
167 sampled during each of these two periods: in both cases, two different ISC types were
168 identified. We also searched for small indels that were unique to clinical samples. We
169 identified six unique indels: five were located in non-coding regions, and one introduced a
170 1-bp frameshift in the LdAQP1 gene (Dataset S1C).

171 Next, we examined chromosome copy numbers. Consistent with our observations in
172 artificially mixed samples (Text S1), the apparent accuracy of our somy estimates in clinical
173 samples depended on the *Leishmania* DNA percentages, as those with low percentage had
174 lower normalized read depth and genome coverage (Dataset S1A). For example,
175 chromosome 31 is known to be tetrasomic in all *Leishmania* species and isolates sequenced
176 so far, except under experimental drug resistance selection where somy can be even higher
177 (9). As expected, S-values (normalized coverage depth, see Materials and Methods) of
178 chromosome 31 were close to 4 in samples with high genome coverage. These values,
179 however, were lower in samples with lower genome coverage, even though the somy of
180 chromosome 31 was distinctively higher than 2 except for the very lowest coverage

181 samples. To distinguish real somy variation and technical artefacts, we classified samples
182 into group of similar genome coverage into those with high (36 samples, genome coverage
183 above 89.1%), medium (16 samples, coverage 56.6-86.9%) and low quality (9 samples,
184 coverage 9.8%-45.8%) somy calls (Dataset S1A and Fig. S6A). Two samples (coverage 1.1%
185 and 0.8%), were excluded from this analysis because their depth was too low to quantify
186 somy. The high-quality samples provided a clear picture of the average ploidy of amastigote
187 populations in human hosts: most chromosomes showed an S-value around 2 (thus overall
188 disomic), with only chromosome 31 showing consistently high somy around 4 (Fig. 2B).
189 Other exceptions were chromosomes 16 and 18 which showed S-value of 1 (thus overall
190 monosomic) in BPK89BM, and in BPK146BM, respectively (Fig. 2B and Fig. S6A1).
191 Chromosome 17 showed a S-value of 1 or lower than 2 (suggestive of mosaicism between
192 monosomic and disomic variants) in seven samples: BPK276BM, BPK104BM, BPK146BM,
193 BPK161BM, BPK198BM, BPK191BM, BPK4BM (Fig. 2B and Fig. S6A1-3). We detected S-
194 values between 2 and 3 (potentially indicating the presence of mosaicism between disomic
195 and trisomic variants) of chromosome 20, 23, 32 and 33 in BPK821BM. (Fig. S6A1).

196 We detected two main local copy number variants (CNVs) specific of CG (termed M-
197 and H-locus, 8) in all 51 SuSL-seq CG samples. When we compared average normalized
198 depth for these two CNVs between clinical and previously published promastigote samples
199 for each genetic group, we found good agreement for M-locus with correlation R^2=0.941
200 and p-value=2.96 x 10^{-4}, and slightly lower depth for H-locus in clinical samples with
201 correlation R^2=0.571, p-value=4.84 x 10^{-2} (Fig. 2C). The ISC1 group contained many other
202 CNVs that were used to test the accuracy of local depth measured in SureSelect platform.
203 We found correlation (R^2=0.9921, p-value=5.3449e-14) between the depth in the ISC1-

204 specific CNVs in the previously analyzed promastigote samples and the depth in the
205 BPK72BM clinical sample (Fig. S6B).
206
207 **Comparison of paired clinical and cultured isolate samples**
208 For 12 of the SuSL-seq clinical samples, paired cultured parasite isolates derived
209 from the same clinical material were available. We compared the genotypes within each of
210 these 12 pairs, at three levels: SNPs, indels and chromosome copy number. SNP analysis
211 revealed that nine SuSL-seq samples were assigned to different ISC genotype groups to the
212 matched cultured isolates (Table 1), with samples assigned to different core group
213 genotypes differing by 9-26 homozygous sites and 3-13 heterozygous positions. In two
214 cases, the clinical samples were genotyped as being in the CG and the isolate was assigned
215 to ISC1; these pairs of samples were far more distinct, differing in 29942 and 34146
216 homozygous SNPs and 4370 and 167 heterozygous sites respectively. For the 3 paired
217 samples with matching ISC genotypes, we found only minor nucleotide differences (0-9
218 homozygous and 3-7 heterozygous SNPs). We found differences in Indel variants in 10 out of
219 12 samples: in one of them (BPK081) a single base insertion was observed in LdAQP1
220 (described in detail above) in the bone marrow, which was not present in the paired culture
221 (Table 1). Finally, chromosome copy number differed in 10 paired samples, while in BPK087
222 and BPK081 we found no significant differences in somy values, and parasites remained
223 overall disomic in culture (Fig. 3A). Accordingly, none of the 12 paired samples revealed
224 identical genomic features between the clinical samples and the derived isolates.
225 We then focused on the nine paired samples for which different ISC types were
226 detected in the bone marrow and the derived isolates. Considering the sympatric ISC
227 genotype diversity highlighted above, it is possible that some patients harbor polyclonal

228 infections, with a given genotype dominant in a human host and another one dominant in
229 culture, because of fitness differences in the respective environments. This possibility was
230 explored by further genetic analyses and by competition experiments.

231 First, we found a second and minor genotype in the SuSL-seq mapped reads of two
232 clinical samples (BPK471BM and BPK296BM), at a proportion estimated as 9.2%, and 6%,
233 respectively (Fig.3B). Secondly, we tested for the presence of more than one genotype in
234 un-enriched clinical samples by applying ISC single locus genotyping (ISC-SLG) combined
235 with NGS-based deep amplicon sequencing using four previously described diagnostic
236 SNPs/insertion allowing discrimination between the ISC3, ISC4, ISC5 and ISC6 genotypes
237 (16). Short regions flanking the SNP/indel were amplified, sequenced and the frequency of
238 both alleles in all the tested samples was estimated. In order to increase the chance of
239 detecting the presence of minor genotypes, we selected clinical samples with the highest
240 parasite loads and/or highest total DNA contents. In addition, we included polyclonal clinical
241 samples BPK471BM and BPK296BM (see above), two clinical sample-isolate pairs (BPK077
242 and BPK276) and cloned lines, which served as positive or negative controls of the presence
243 of polymorphism at the diagnostic SNP/indel sites. The estimated precentage of *Leishmania*
244 DNA and number of *Leishmania* genome equivalents in the clinical samples varied between
245 0.05-2.85%, and 12-60901, respectively (Dataset S1D). We could confirm the principal ISC
246 genotypes defined by SuSL-seq in all clinical samples (Dataset S1D). Alternative alleles
247 containing SNP variants could be observed in most of the clinical samples, but their
248 frequency was not significantly different from the frequency found in the control cloned
249 lines (Fig. S7 and Dataset S1D). For the AQP1 insertion, the situation was similar with the
250 exception of the sample BPK120BM (defined as belonging to ISC9), which indicated the
251 presence of a small population of ISC5 genotype (z-score:3.2; p-value: 0.0014). The

252 reciprocal was found in BPK276BM, an ISC5 sample in which a small proportion of WT alleles
253 was encountered and we found no alleles containing the ISC5-specific AQP1 insertion in the
254 paired clinical isolate. Although these results are not significant (p-value: 0.139), they are
255 consistent with the observed ISC5 genotype in BPK276BM, and ISC6 genotype in the derived
256 isolate (Dataset S1D).

257 Finally, to test whether different growth rates could explain the different genotypes
258 observed between direct clinical samples and those grown in vitro, we set up a growth-
259 competition experiment. We inoculated flasks with cloned strains BPK091 and BPK077 (ISC4
260 and ISC6, respectively), and BPK067 and BPK091 (ISC3, and ISC4, respectively) together in
261 the following ratios (strain1/strain2): 90/10, 50/50, and 10/90. Using Sanger sequencing to
262 monitor the presence of diagnostic SNPs that could distinguish the two genetic groups at
263 different time points during *in vitro* cultivation, we were able to detect which of the two co-
264 cultured strains was dominant. We reproducibly observed for both parasite combinations
265 that one strain consistently dominated the culture, even if it was only present as 10% of the
266 initial inoculum (Fig. S8). Interestingly, BPK091, which was present in both mixes, became
267 dominant after 21 passages when co-cultured with BPK077 but after only five passages it
268 was out-competed by BPK067.

269

270

271 **Discussion**

272 We demonstrated here proof-of-principle for the sequencing of *Leishmania* genomes
273 directly from clinical samples of patients with VL. Tested on 63 clinical samples from Nepal,
274 SuSL-seq showed a high analytical sensitivity: we were able to (i) perform phylogenomic
275 analyses on 82.5% of samples, (ii) assign 97% of samples to previously defined ISC
276 genotypes using a set of diagnostic SNPs, (iii) estimate chromosome copy number in 82.5%
277 of samples and (iv) identify large local CNVs in 83%. With the current design of the baits, our
278 method should be applicable to all the parasites causing visceral leishmaniasis i.e. *L.*
279 *infantum* and *L. donovani*, in East Africa, the Mediterranean basin and Latin America.
280 Further work is required to test SuSL-seq on other host samples (animal reservoir, insect
281 vectors, or different host tissues). It would be desirable to improve its performance in
282 *Leishmania* samples with low parasite load and/or low amount of total DNA: here, the
283 combination of SureSelect with other enrichment methods might prove useful.

284 Our study sheds new light on the biology of *Leishmania* in the human host. A first
285 aspect concerned the karyotype of amastigotes. Massive aneuploidy has been described in
286 cultivated *Leishmania* promastigotes of all species (17). In particular, sequencing of 204 *L.*
287 *donovani* genomes (8) revealed strong differences in karyotype between parasite strains,
288 with a baseline overall disomy and up to 22 polysomic chromosomes (essentially 3N and 4N)
289 in single strains. In sharp contrast, we did not find high levels of aneuploidy in the
290 amastigotes present in clinical samples of the same region. Besides chromosome 31 which is
291 constitutively tetrasomic in all *Leishmania* species, most chromosomes in amastigotes of
292 clinical samples were inferred to be disomic, and only a few monosomic or with
293 intermediate coverage depth, the latter likely reflecting mosaic aneuploidy in the analyzed
294 parasite populations, with mixtures of cells of different karyotypes (18). Overall, the

295 differences in aneuploidy observed between clinical samples and derived isolates are
296 consistent with our previous findings where we reported close to diploid karyotype in
297 parasites adapted to Syrian golden hamster, and highly dynamic aneuploidy during
298 propagation of *L. donovani* in vitro (11). This genomic plasticity was mirrored at
299 transcriptome level and might provide a rapid adaptation in response to environmental
300 cues.

301 A second potential finding is related to the observed genotypic differences between
302 12 clinical samples and derived cultured isolates. Strictly speaking and considering different
303 types of genetic variants (SNPs, indels an somy), none of the 12 isolates showed an identical
304 genotype to the parasites present in the paired bone marrow samples. We cannot exclude
305 that the 1-nt indel in the LdAQP1 of BPK81BM 'mutated' during culture maintenance and
306 restored the reading frame in the corresponding isolate's genome. However, knowing that
307 frame shifts in this gene have been linked to *Leishmania* antimony resistance (9), this result
308 shows the risk of losing drug resistance markers when genotyping isolates. The differences
309 in karyotype as observed in 11/12 paired samples could also be acquired during culture, but
310 could also have resulted from selection from a mosaic background, as shown experimentally
311 (18). More generally, however, the SNP differences that led to isolates and SuSL-seq
312 samples representing different ISC genotypes did not appear to be explained by mutations
313 appearing *in vitro*; each ISC group is characterized by a set of diagnostic SNPs and the
314 observed switch of ISC genotypes concerned all diagnostic SNPs in these sample pairs. In
315 two cases, the isolate and clinical samples differed at tens of thousands of sites, and clearly
316 represented the two very different populations of parasites circulating in VL patients in
317 Nepal. We excluded SNP-calling artefacts, and the clinical work followed specific procedures
318 for labeling, processing and tracking all samples. Thus, excluding experimental mistakes as

319 the reason for these results, the most likely explanation is polyclonality of the infection in
320 the host. In such a scenario, different clones would be most abundant in the host and the
321 promastigote culture medium, due to their differences in respective fitness in these two
322 environments.

323 We have attempted to verify the polyclonality hypothesis through a set of direct and
324 indirect experiments. On one hand, in two clinical samples we detected minor SuSL-seq
325 reads (above 6%) with the signature of a second ISC genotype. We also tried to detect the
326 second genotype directly in selected bone marrow samples with amplicon sequencing of 4
327 loci, which contain SNPs/AQP1 indel specific for different genetic groups from ISC. However,
328 we were unable to find any statistically robust evidence for presence of a second genotype
329 with this method, except in one case, BPK120BM. This is most likely due to a sensitivity issue
330 of the amplicon sequencing method which was applied: (i) without enrichment (in most
331 cases, in presence of more than 99% of human DNA, Dataset S1D), (ii) to find minor variants
332 and (iii) targeting one single variant within (iv) a single copy gene. The two clinical samples
333 BPK296BM and BPK471BM illustrate this problem well: they were clearly shown to be
334 polyclonal from SuSL-seq data, while the second clone could not be detected from amplicon
335 sequencing, most likely due to the low number of *Leishmania* genome equivalents
336 (respectively, 43 and 12). On the other hand, the plausibility of this hypothesis was
337 indirectly validated by a competition experiment in which we co-cultivated strains of
338 different ISC genotypes in the same tube: even if a strain constituted only 10% of the initial
339 inoculum, it could rapidly and in a reproducible way dominate over time in culture.

340 Further work is required to assess the extent of polyclonality among *L. donovani*
341 infections. The phenomenon is common in *Plasmodium*, especially in high transmission
342 areas (19). Among trypanosomatids, it has been reported in *Trypanosoma cruzi*, where

343 among others multiple genotypes were found to be transmitted congenitally (20) and
344 temporal variation of genotypes was observed in asymptomatic carriers (21). In *Leishmania*,
345 mixed infections of different species were reported in wild rodents (22), horses (23) and
346 humans (24,25). A recent study documented the possibility of polyclonality in *L. donovani*
347 from East Africa, as different genotypes were isolated from different organs of the same
348 patient (26). A similar observation was made on a Spanish dog in which two different
349 zymodemes of *L. infantum* were collected from the skin and a lymph node (27).
350 Furthermore, natural *Leishmania* hybrids are regularly reported all over the world (see for
351 instance 28,29) and in our previous study in the ISC (8), we found 8 isolates of hybrid origin
352 (out of 192 isolates of the CG, i.e. 4.2%). This could only occur in sand flies as a result of
353 polyclonal infections.

354 The present work highlights the complexity of *L. donovani* infections in the human
355 host. It appears that a patient can be infected by multiple clones showing different
356 genotypes, which possibly vary at the somy level. This genetic diversity could offer an
357 adaptive advantage to the whole population of cells, by providing 'individual' solutions to
358 different environments, leading to different dominant genomes in the patient and in
359 culture. We showed that this did not have a major impact so far for phylogenomic studies.
360 However, if any link is to be made between parasite genome variation and clinical
361 phenotypes – for example treatment outcome or pathogenicity – analyzing *Leishmania*
362 directly from clinical samples is necessary, as both the genotype and aneuploidy in the
363 mammalian host and in vitro culture can differ.

364

365

366 **Material and Methods**

367

368 **DNA extraction and next generation sequencing.** DNA from parasite cultures and
369 clinical samples (bone marrow and spleen aspirates) was extracted using QiaAmp DNA
370 blood mini kit (Qiagen, Venlo, the Netherlands) following the manufacturer's instructions,
371 and DNA concentrations were verified using the Qubit broad-range DNA quantification kit
372 (Thermo Fisher Scientific, Waltham, USA). Sequencing was done following the
373 manufacturer's standard cluster generation and sequencing protocols

374 Standard Whole Genome Sequencing was applied to isolates. Genomic DNA was
375 sheared into 400bp fragments by focused ultrasonication (Covaris Inc., Woburn, USA).
376 Amplification-free Illumina libraries were prepared (30) and 125bp PE reads were generated
377 on the Illumina HiSeq v4.

378 Deep sequencing was applied to quantify parasite load in batch 1 of Bone Marrow
379 (BM) samples. Genomic DNA was sheared into 150bp fragments by focused ultrasonication
380 (Covaris Inc., Woburn, USA) and libraries were prepared using the SureSelect Automated
381 Library Prep Kit (Agilent, Santa Clara, USA). Index tagged samples were amplified using KAPA
382 HiFi DNA polymerase, (KAPA Biosystems, Wilmington, USA) and quantified using an
383 Accuclear DNA Kit (Biotium, Fremont, USA), 75bp PE reads were generated on the Illumina
384 HiSeq 2500.

385 SureSelect was used for *Leishmania* genome capture. Genomic DNA was sheared
386 into 150bp fragments by focused ultrasonication (Covaris Inc., Woburn, USA) and libraries
387 were prepared, either as described above, or using the NEBNext Ultra II DNA Library prep Kit
388 (New England Biolabs, Ipswich, USA). Samples were tagged with a unique index, amplified
389 and quantified by qPCR (Text S1). Samples were then pooled into batches for SureSelect

390 enrichment, based on qPCR estimates of parasite load, so that each batch had as little
391 variation in parasite load as possible: we aimed for only a 2-fold variation within a batch,
392 while optimising the use of the sequencing data. Based on our experience of enrichment
393 levels obtained in the experimental samples, batch sizes varied from 6 or 7 samples per
394 batch where they had less than 0.05% *Leishmania* gDNA content up to 8 samples for higher
395 parasite load. Samples were pooled within batches in an equimolar fashion. Pooled material
396 was taken forward for hybridization, capture and enrichment using the standard SureSelect
397 Target Enrichment System XT2 protocol recommended by the manufacturer (Agilent
398 Technologies https://www.agilent.com/cs/library/usermanuals/Public/G9630-90000.pdf).
399 Custom designed oligonucleotide baits were used at a 1:10 stock dilution. 75bp paired-end
400 reads were generated on the Illumina HiSeq 2500.

401 Amplicon sequencing was used to explore the presence of minor genetic variants in
402 clinical samples. In the first step we used the primers listed in Text S1 to amplify fragments
403 that contained diagnostic SNPs/insertion for ISC3, ISC4, ISC5 and ISC5 groups using
404 Advantage 2 Polymerase Mix (TAKARA, Kusatsu, Japan). 2µl of template with varying DNA
405 concentration were used in a total of 50µl PCR reaction (Dataset S1D). Touch down PCR with
406 the following conditions was used: 95˚C for 5 mins followed by 16 cycles of 95˚C for 30s,
407 68˚C to 6O˚C (decrease of 0.5˚C with each cycle) for 30s, 68˚C for 30sec, followed by 19
408 cycles of 95˚C for 30s, 60˚C for 30s, 68˚C for 30s and a final step at 68˚C for 10min. Then, a
409 second round amplification was performed, using 5µl of template, to add indexed Illumina
410 adapters (31). PCR conditions of the second round were 95˚C for 3 mins followed by 10
411 rounds of 98˚C for 20s, 65˚C for 30s, 72˚C for 30s finishing with 5 mins at 72˚C. The final
412 products were purified using AMPure XP SPRI beads and quantified on a plate reader using
413 an AccuClear DNA quantification kit (Biotium, Fremont, USA) followed by equimolar pooling.

414 Due to low base diversity of the amplicons, 20% PhiX was spiked into the resulting pool and
415 300bp PE reads were generated on the Illumina MiSeq.
416
417 **DNA read mapping, SNP and indel calling.** SuSL-seq reads from clinical samples and
418 data from 204 (54 strains, 150 isolates) previously sequenced *L. donovani* genomes (8) were
419 mapped to the improved reference *L. donovani* genome LdBPKv2 (11), using Smalt v7.4
420 (http://www.sanger.ac.uk/science/tools/smalt-0). The methods were reported elsewhere
421 (11). Briefly, Smalt options for exhaustive searching for optimal alignments (-x) and random
422 mapping of multiple hit reads were used. Reads were mapped when mapping base identity
423 was greater than or equal to 80% to the reference genome (-y 0.8), to eliminate spuriously
424 mapped reads mainly from human DNA. For the sequence analyses, all reads mapped to the
425 reference, not limited to bait locations, were analyzed. For phylogenetic analyses and for
426 evaluating the accuracy of SuSL-seq data, we used the population SNP and indel calling
427 mode of UnifiedGenotyper in Genome Analysis Toolkit v3.4 with the SNP cut off 4000 used
428 (GATK: https://software.broadinstitute.org/gatk/). No lower or higher depth cut off was
429 used in the population SNP calling since the depth of some SuSL-seq samples was quite low.
430
431 **Reference genome masking and SNP screening.** To mask repetitive regions and
432 regions prone to generate false positive SNPs due to the presence of non-*Leishmania* DNA in
433 clinical samples, we masked 7,453 regions that spanned a total of 43,200 bases. These
434 regions were identified by blast similarity hits (8) and by using SNPs that were generated by
435 pure human DNA SureSelect analysis and samples with lower amount of *Leishmania* DNA.
436 The masked positions were provided in MaskSuSL4.bed file. We also excluded a SNP cluster
437 containing more than 7 SNPs within 60 neighboring bases to exclude false SNPs from

438 simulated *Leishmania* DNA samples mixed with mouse DNA Mpd1 and Mpd2. This was
439 effective to eliminate non-*Leishmania* reads with a dozen of SNPs but with highest mapping
440 score 60.
441
442 **Somy estimation.** For somy estimation, 2 different methods were used. For samples
443 with consistently high-quality mapping against *Leishmania*, such as genomic libraries from
444 cultured promastigotes we used the median depth across each chromosome (11,32). For
445 samples whose read coverage depth was more variable, such as amplified and unamplified
446 clinical samples, depth was calculated in 1000bp windows to allow a more uniform analysis
447 of coverage variation (11). Previous studies have shown some correlation between read
448 coverage depth and chromosome length for some sequencing runs (11). To correct this bias,
449 we applied local median correction (11). This method calculates somy based on the median
450 coverage values of chromosomes with similar lengths. In the calculation of somy, the first
451 7,000 bases and last 2,000 bases of each chromosome were excluded because these regions
452 tend to be repetitive telomeric regions and their depth was not reliable. Similarly, certain
453 over-amplified chromosome regions were excluded from depth calculation for SuSL-seq
454 samples: these regions were listed in Dataset S1E. The range of monosomy, disomy,
455 trisomy, tetrasomy, and pentasomy was defined to be the full cell-normalized chromosome
456 depth or S-value: $S < 1.5$, $1.5 \leq S < 2.5$, $2.5 \leq S < 3.5$, $3.5 \leq S < 4.5$, and $4.5 \leq S < 5.5$,
457 respectively (11).
458
459 **Gene copy number variation.** Local copy number variants were detected as
460 described elsewhere (8). In brief, we calculated an average normalised haploid depth of H-
461 locus (Ld23: 90426 - 104470) and M-locus (Ld36: 2558460 - 2576450) for each strain and

462 then obtained an average normalised haploid depth and standard deviation for each group.
463 For ISC1-specific CNVs, we calculated an average normalised haploid depth for each gene
464 for each ISC1 strain and then obtained an average normalised haploid depth and standard
465 deviation for each CG group, using the depth of 191 sequenced strains (promastigotes). ISC1
466 specific gene CNVs were then defined to be genes whose normalized haploid depth differed
467 more than $\pm$ 0.3 from the baseline average depth of other ISC groups. The value 0.3 was
468 chosen here because we have observed many long CNVs in trisomic chromosomes, which
469 were amplified or deleted in 1 of the 3 homologous chromosomes.
470

471 **Phylogenetic analysis.** For phylogenetic analysis, 197 variables sites represented by
472 394 bases were used to perform neighbor-joining analysis using the p-distance model with
473 1000 bootstrap replicates as performed in MEGA 7.0.21 (33) and and Splitstree v4 (34).
474 These analyses used 191 promastigote strains plus 51 clinical samples (one clinical sample
475 from ISC1 was excluded because phylogenetically too divergent). Artificial mixtures of
476 *Leishmania* and human DNA (Leishmania DNA percentage of 0.06%) were used to measure
477 the accuracy to multiple infections of the phylogenetic analyses.

478

479 **Identifying minor ISC genotypes and polyclonal infections with SuSL-seq reads.** Two
480 methods were used to identify minor ISC genotypes, respectively based on (i) SNP motifs
481 specific of ISC groups and (ii) alternative allele frequency from sequenced reads.
482 We first extracted genotype specific 23-mer motifs that contained a SNP or
483 reference base at the center. For ISC3, ISC4, ISC5, ISC6 and ISC7 groups, we created 26, 53,
484 38, 26 and 14 motifs that matched SNPs common to these groups, respectively. These group

485 specific SNPs were identified using 191 samples analyzed in the previous study (8) and
486 mapped to the pacbio reference (11). To identify minor genotype reads which may have
487 been missed by the standard SNP calling, we counted the reads in the alignment files that
488 contained genotype specific motifs. These motifs were listed in Dataset S1D and used to
489 estimate the proportion of ISC3, ISC4, ISC5, ISC6 and ISC7 in each sample.

490 In the second approach, a GATK population VCF file for bone marrow and
491 promastigotes were converted to a table that contained alternative allele frequency. In
492 addition, total read depth, reference read depth, alternative allele read depth and genotype
493 information created for VCF file were kept along alternative allele frequency. We checked
494 alternative allele frequencies at group specific SNP sites to identify polyclonal samples. A
495 mean of alternative allele frequency was calculated by averaging alternative allele
496 frequencies at all ISC specific SNP sites per sampling, excluding missing depth sites.

497

498 **Identifying minor ISC groups using Amplicon sequencing.** We applied a read count
499 method and an exact motif matching method to identify minor ISC groups using amplicon
500 sequencing. The read count method was based on samtools pileup and included reads
501 marked as PCR duplicates, counting all four bases to access back ground base error rates.
502 The exact motif matching method was based on a 17 mer for each diagnostic SNP and a 22
503 mer for the diagnostic insertion and they were listed in Dataset S1F and G.

504

505 **Parasite cultures and competition experiment.** Twelve clinical isolates paired to
506 bone marrow samples were cultivated on HOMEM. Three *L. donovani* cloned strains were
507 used in this experiment BPK067/0/cl2, BPK077/0/cl5, BPK091/0/cl9, respectively belonging
508 to the ISC3, ISC6, and ISC4 genetic groups (8). The parasites were grown on Tobie's blood

509 agar (35) with a saline overlay at 26°C to mimic the growth conditions during parasite
510 isolation from clinical samples (36). The parasites were inoculated to the concentration of
511 $5x10^5 ml^{-1}$, and passaged to fresh medium every 3-4 days. For each strain mixture, 3
512 replicates were prepared, and the pure strains were grown in duplicates. The proportion
513 between two parasites strains was roughly verified at the start of the experiment, and was
514 further assessed approximately every 5 passages.

515 To follow the proportions between two different parasite strains in culture, a
516 diagnostic SNPs, which allow distinction among different ISC genetic groups were used (16).
517 Specifically, the Ld36_p5263_F1* and Ld36_p5263_R1, and Ld33_p3644_F1 and
518 Ld33_p3644_R1* primer pairs were used to distinguish between ISC3 and reference, and
519 ISC6 and reference genotypes, respectively. The presence of these alleles was verified by
520 standard Sanger sequencing, where primers mentioned above with * were used to
521 sequence the amplified products. PCR amplicons were sequenced at Baseclear (Leiden, The
522 Netherlands) with ABI 3730 (XL) DNA Analyzer, and the chromatograms were analyzed using
523 the 4peaks software (Nucleobytes). The ability to detect both alleles was first verified by
524 sequencing the PCR products amplified from DNA mixed in the same proportions as the
525 parasites in the competition experiment. With this method, it was not possible to detect the
526 exact proportions of the two alleles, but the relations between them (i.e. which allele was
527 present in higher proportion) was assessed.

528

529

530 **Data Release.** Raw data was deposited in the European Nucleotide Archive (ENA)
531 with the accession number ERP110990

532

533

534

535

536 **Acknowledgements**

537 This study was supported by Belgian Science Policy Office (TRIT, P7/41), Flemish Fund for
538 Scientific Research (G.0.B81.12), and Department of Economy, Science and Innovation in
539 Flanders ITM-SOFIB (SINGLE project, to J.C.D., M.V.A. and M.A.D.). MB, MJS and JAC are
540 supported by Wellcome via their core support for the Wellcome Sanger Institute (grant
541 206194). We acknowledge the support of the core pipeline staff at WSI, in particular Peter
542 Ellis and Sara Widaa, and the help of Owen Hardy from Agilent technologies for assistance in
543 designing the SureSelect array. The funders had no role in study design, data collection and
544 analysis, or preparation of the manuscript.

545

546 **References**

547 1. Burza S, Croft SL, Boelaert M. 2018. Leishmaniasis. Lancet 392:951-970.
548 https://doi.org/10.1016/S0140-6736(18)31204-2
549 2. Alvar J, Vélez ID, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, den Boer M; WHO
550 Leishmaniasis Control Team. 2012. Leishmaniasis worldwide and global estimates of
551 its incidence. PLoS One 7:e35671. https://doi.org/10.1371/journal.pone.0035671
552 3. Rijal S, Sundar S, Mondal D, Das P, Alvar J, Boelaert M. 2019. Eliminating visceral
553 leishmaniasis in South Asia: the road ahead. BMJ 364:k5224.
554 https://www.bmj.com/content/364/bmj.k5224
555 4. Dujardin JC, Decuypere S. 2013. Epidemiology of leishmaniasis in the time of drug
556 resistance. In: Drug Resistance in *Leishmania* Parasites. Consequences, Molecular

557 Mechanism and Possible Treatments. Eds Ponte-Sucre A, Padron-Nieves M, Diaz E.
558 Spinger-Verlag, Wienen. pp 65-83. https://doi.org/10.1007/978-3-7091-1125-3_4
559 5. Dujardin JC. 2018. Epidemiology of leishmaniasis in the time of drug resistance (the
560 Miltefosine era). In: Drug resistance in *Leishmania* parasites: consequences,
561 molecular mechanisms and possible treatments. Eds Ponte-Sucre A & Padron-Nieves
562 M. Springer International Publishing, pp 85-107. https://doi.org/10.1007/978-3-319-
563 74186-4_4
564 6. Cuypers B, Berg M, Imamura H, Dumetz F, De Muylder G, Domagalska MA, Rijal S,
565 Bhattarai NR, Maes I, Sanders M, Cotton JA, Meysman P, Laukens K, Dujardin JC.
566 2018. Integrated genomic and metabolomic profiling of ISC1, an emerging
567 *Leishmania donovani* population in the Indian subcontinent. Infect Genet Evol
568 62:170-178. https://doi.org/10.1016/j.meegid.2018.04.021
569 7. Karunaweera ND, Ferreira MU. 2018. Leishmaniasis: current challenges and
570 prospects for elimination with special focus on the South Asian region. Parasitology.
571 145:425-429. https://doi.org/10.1017/S0031182018000471
572 8. Imamura H, Downing T, Van den Broeck F, Sanders MJ, Rijal S, Sundar S, Mannaert A,
573 Vanaerschot M, Berg M, De Muylder G, Dumetz F, Cuypers B, Maes I, Domagalska M,
574 Decuypere S, Rai K, Uranw S, Bhattarai NR, Khanal B, Prajapati VK, Sharma S, Stark O,
575 Schönian G, De Koning HP, Settimo L, Vanhollebeke B, Roy S, Ostyn B, Boelaert M,
576 Maes L, Berriman M, Dujardin JC, Cotton JA. 2016. Evolutionary genomics of
577 epidemic visceral leishmaniasis in the Indian subcontinent. Elife 5:e12613.
578 https://doi.org/10.7554/eLife.12613.001
579 9. Dumetz F, Cuypers B, Imamura H, Zander D, D'Haenens E, Maes I, Domagalska MA,
580 Clos J, Dujardin JC, De Muylder G. 2018. Molecular Preadaptation to Antimony

581 Resistance in *Leishmania donovani* on the Indian Subcontinent. mSphere 3:e00548-
582 17. https://doi.org/10.1128/mSphere.00548-17
583 10. Ostyn B, Gidwani K, Khanal B, Picado A, Chappuis F, Singh SP, Rijal S, Sundar S,
584 Boelaert M. 2011. Incidence of symptomatic and asymptomatic *Leishmania donovani*
585 infections in high-endemic foci in India and Nepal: a prospective study. PLoS Negl
586 Trop Dis 5:e1284. https://doi.org/10.1371/journal.pntd.0001284
587 11. Dumetz F, Imamura H, Sanders M, Seblova V, Myskova J, Pescher P, Vanaerschot M,
588 Meehan CJ, Cuypers B, De Muylder G, Späth GF, Bussotti G, Vermeesch JR, Berriman
589 M, Cotton JA, Volf P, Dujardin JC, Domagalska MA. 2017. Modulation of Aneuploidy
590 in *Leishmania donovani* during Adaptation to Different In Vitro and In Vivo
591 Environments and Its Impact on Gene Expression. MBio 8:e00599-17.
592 https://doi.org/10.1128/mBio.00599-17
593 12. Cuypers B. 2018. A systems biology approach for a comprehensive understanding of
594 molecular adaptation in *Leishmania donovani*. PhD thesis, University of Antwerp.
595 https://anet.be/record/opacuantwerpen/c:lvd:14740884/E
596 13. Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, Dimalanta ET,
597 Amaral-Zettler LA, Davis T, Quail MA, Pradhan S. 2013. A method for selectively
598 enriching microbial DNA from contaminating vertebrate host DNA. PLoS One
599 8:e76096. https://doi.org/10.1371/journal.pone.0076096
600 14. Oyola SO, Ariani CV, Hamilton WL, Kekre M, Amenga-Etego LN, Ghansah A, Rutledge
601 GG, Redmond S, Manske M, Jyothi D, Jacob CG, Otto TD, Rockett K, Newbold CI,
602 Berriman M, Kwiatkowski DP. 2016. Whole genome sequencing of *Plasmodium*
603 *falciparum* from dried blood spots using selective whole genome amplification.
604 Malar J 15:597. https://doi.org/10.1186/s12936-016-1641-7

605 15. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, Bryant JM, Chan J,
606 Creer D, Holdstock J, Kunst H, Lozewicz S, Platt G, Romero EY, Speight G, Tiberi S,
607 Abubakar I, Lipman M, McHugh TD, Breuer J. 2018. Direct Whole-Genome
608 Sequencing of Sputum Accurately Identifies Drug-Resistant *Mycobacterium*
609 *tuberculosis* Faster than MGIT Culture Sequencing. J Clin Microbiol. 56:e00666-18.
610 https://doi.org/10.1128/JCM.00666-18
611 16. Rai K, Bhattarai NR, Vanaerschot M, Imamura H, Gebru G, Khanal B, Rijal S, Boelaert
612 M, Pal C, Karki P, Dujardin JC, Van der Auwera G. 2017. Single locus genotyping to
613 track *Leishmania donovani* in the Indian subcontinent: Application in Nepal. PLoS
614 Negl Trop Dis 11:e0005420. https://doi.org/10.1371/journal.pntd.0005420
615 17. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, Harris D, Her Y,
616 Herzyk P, Imamura H, Otto TD, Sanders M, Seeger K, Dujardin JC, Berriman M, Smith
617 DF, Hertz-Fowler C, Mottram JC. 2011. Chromosome and gene copy number
618 variation allow major structural change between species and strains of *Leishmania*.
619 Genome Res 2011 21:2129-42. https://doi.org/10.1101/gr.122945.111
620 18. Prieto Barja P, Pescher P, Bussotti G, Dumetz F, Imamura H, Kedra D, Domagalska M,
621 Chaumeau V, Himmelbauer H, Pages M, Sterkers Y, Dujardin JC, Notredame C, Späth
622 GF. 2017. Haplotype selection as an adaptive mechanism in the protozoan pathogen
623 *Leishmania donovani*. Nat Ecol Evol. 1:1961-1969. https://doi.org/10.1038/s41559-
624 017-0361-x
625 19. Pacheco MA, Lopez-Perez M, Vallejo AF, Herrera S, Arévalo-Herrera M, Escalante AA.
626 2016. Multiplicity of Infection and Disease Severity in *Plasmodium vivax*. PLoS Negl
627 Trop Dis 10:e0004355. https://doi.org/10.1371/journal.pntd.0004355

628 20. Llewellyn MS, Messenger LA, Luquetti AO, Garcia L, Torrico F, Tavares SB, Cheaib B,
629 Derome N, Delepine M, Baulard C, Deleuze JF, Sauer S, Miles MA. 2015. Deep
630 sequencing of the *Trypanosoma cruzi* GP63 surface proteases reveals diversity and
631 diversifying selection among chronic and congenital Chagas disease patients. PLoS
632 Negl Trop Dis 9:e0003458. https://doi.org/10.1371/journal.pntd.0003458
633 21. Sánchez LV, Bautista DC, Corredor AF, Herrera VM, Martinez LX, Villar JC, Ramírez JD.
634 2013. Temporal variation of *Trypanosoma cruzi* discrete typing units in
635 asymptomatic Chagas disease patients. Microbes Infect 15:745-8.
636 https://doi.org/10.1016/j.micinf.2013.06.008
637 22. Ferreira Ede C, Cruz I, Cañavate C, de Melo LA, Pereira AA, Madeira FA, Valério SA,
638 Cunha HM, Paglia AP, Gontijo CM. 2015. Mixed infection of *Leishmania infantum* and
639 *Leishmania braziliensis* in rodents from endemic urban area of the New World. BMC
640 Vet Res 11:71. https://doi.org/10.1186/s12917-015-0392-y
641 23. Soares IR, Silva SO, Moreira FM, Prado LG, Fantini P, Maranhão Rde P, da Silva Filho
642 JM, Melo MN, Palhares MS. 2013. First evidence of autochthonous cases of
643 *Leishmania (Leishmania) infantum* in horse (Equus caballus) in the Americas and
644 mixed infection of *Leishmania infantum* and *Leishmania (Viannia) braziliensis*. Vet
645 Parasitol 197:665-9. https://doi.org/10.1016/j.vetpar.2013.06.014
646 24. Veland N, Valencia BM, Alba M, Adaui V, Llanos-Cuentas A, Arevalo J, Boggild AK.
647 2013. Simultaneous infection with *Leishmania (Viannia) braziliensis* and *L. (V.)*
648 *lainsoni* in a Peruvian patient with cutaneous leishmaniasis. Am J Trop Med Hyg
649 88:774-7 https://doi.org/10.4269/ajtmh.12-0594

650 25. Shirian S, Oryan A, Hatam GR, Daneshbod Y. 2012. Mixed mucosal leishmaniasis
651 infection caused by *Leishmania tropica* and *Leishmania major*. J Clin Microbiol
652 50:3805-8. https://doi.org/10.1128/JCM.01469-12
653 26. Zackay A, Cotton JA, Sanders M, Hailu A, Nasereddin A, Warburg A, Jaffe CL. 2018.
654 Genome wide comparison of Ethiopian *Leishmania donovani* strains reveals
655 differences potentially related to parasite survival. PLoS Genet 14:e1007133.
656 https://doi.org/10.1371/journal.pgen.1007133
657 27. Pratlong F, Portus M, Rispail P, Moreno G, Bastien P, Rioux JA. 1989. Simultaneous
658 presence in dogs of 2 zymodemes of the *Leishmania infantum* complex. Ann
659 Parasitol Hum Comp 64:312-4. https://doi.org/10.1051/parasite/1989644312
660 28. Rogers MB, Downing T, Smith BA, Imamura H, Sanders M, Svobodova M, Volf P,
661 Berriman M, Cotton JA, Smith DF. 2014. Genomic confirmation of hybridisation and
662 recent inbreeding in a vector-isolated *Leishmania* population. PLoS Genet
663 10:e1004092. https://doi.org/10.1371/journal.pgen.1004092
664 29. Odiwuor S, De Doncker S, Maes I, Dujardin JC, Van der Auwera G. 2011. Natural
665 *Leishmania donovani/Leishmania aethiopica* hybrids identified from Ethiopia. Infect
666 Genet Evol 11:2113-8. https://doi.org/10.1016/j.meegid.2011.04.026
667 30. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009.
668 Amplification-free Illumina sequencing-library preparation facilitates improved
669 mapping and assembly of (G+C)-biased genomes. Nat Methods 6:291-5.
670 https://doi.org/10.1038/nmeth.1311
671 31. Bronner IF, Quail MA, Turner DJ, Swerdlow H. 2014. Improved Protocols for Illumina
672 Sequencing. Curr Protoc Hum Genet 80:18.2.1-42.
673 https://doi.org/10.1002/0471142905.hg1802s80

674 32. Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, Hilley JD, de
675 Doncker S, Maes I, Mottram JC, Quail MA, Rijal S, Sanders M, Schönian G, Stark O,
676 Sundar S, Vanaerschot M, Hertz-Fowler C, Dujardin JC, Berriman M. 2011. Whole
677 genome sequencing of multiple *Leishmania donovani* clinical isolates provides
678 insights into population structure and mechanisms of drug resistance. Genome Res
679 21:2143-56. https://doi.org/10.1101/gr.123430.111
680 33. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5:
681 molecular evolutionary genetics analysis using maximum likelihood, evolutionary
682 distance, and maximum parsimony methods. Mol Biol Evol 28:2731-9.
683 https://doi.org/10.1093/molbev/msr121
684 34. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary
685 studies. Mol Biol Evol. 23:254-67. https://doi.org/10.1093/cid/cit102
686 35. Tobie, E. J., Von Brand, T. & Mehlman, B. 1950. Cultural and physiological
687 observations on *Trypanosoma rhodesiense* and *Trypanosoma gambiense*. Journal of
688 Parasitology 36:48-54.
689 36. Yardley V, Croft SL, De Doncker S, Dujardin JC, Koirala S, Rijal S, Miranda C, Llanos-
690 Cuentas A, Chappuis F. 2005. The sensitivity of clinical isolates of *Leishmania* from
691 Peru and Nepal to miltefosine. Am J Trop Med Hyg 73:272-5.
692
693
694
695
696

697 **Legends**

698

699 **Fig. 1. Performance of SuSL-sequencing on clinical samples**. **A**. Enrichment of *Leishmania*
700 DNA from clinical samples by SureSelect; Y axis shows percentage of reads mapping to the *L.*
701 *donovani* reference genome following SureSelect enrichment. X axis is the pre-enrichment
702 percentage of *Leishmania* DNA, estimated by either whole-genome sequencing or qPCR;
703 dotted lines indicate no enrichment and 10-, 100- and 1000-fold enrichment. **B**. Read
704 mapping statistics in the 63 samples after SureSelect enrichment. Upper panel shows
705 evenness of genome coverage as the proportion of bases covered with a minimum of
706 number of sequencing reads from 1 to 100 as indicated by the seven different lines. Lower
707 panel shows the percentage of reads mapping to the *Leishmania* reference genome for each
708 sample, shown on the x-axis. The x-axis is common to both panels. BM, Bone Marrow; SP,
709 Spleen aspirate.

710

711 **Fig. 2. Genomic diversity among clinical samples. A.** SNPs. Phylogenetic tree based on
712 SureSelect enriched clinical samples (black circle) and previously sequenced isolates (8).
713 Diagram is a neighbor-joining phylogeny based on 197 variable sites. ISC2-ISC10 were sub-
714 populations previously defined (8); ISC11 is a novel group that includes clinical samples and
715 isolates previously designated as 'ungrouped'. Full sample identifiers are shown in Fig. S5A.
716 **B.** Karyotypes (High quality samples). Y axis shows normalized somy estimate for each
717 chromosome (x axis). Points show central estimate and bars show one standard deviation
718 around these estimates, calculated by the binned depth method. **C.** Local CNVs. Average
719 copy number per cell of H- and M-loci per ISC group in SureSelect enriched clinical samples

720 (Bone Marrow or Spleen aspirate; BM/SP) and in cultured isolates (promastigotes, Prom).
721 Error bars show one standard deviation around the mean estimate. Only a single sample
722 was available from bone marrow aspirates in ISC1, hence no standard deviation is shown.

723

724 **Fig. 3. Somy comparison between matched clinical samples and culture parasites.** A**.**
725 Inferred somy for 3 samples for which SureSelect-enriched bone marrow (BM) samples and
726 cultured isolates (MO) from the same patients had matching ISC genotypes. Y axis shows
727 normalized somy estimate for each chromosome (x axis). Points show central estimate and
728 bars show one standard deviation around these estimates. Somy estimates and standard
729 deviations for the bone marrow samples were based on the binned depth method while
730 corresponding values for isolates were based on depth of each position. **B.** Evidence for
731 polyclonal infections. Bars represent the proportion of sequencing reads showing the ISC-
732 specific genotype or the reference (REF) genotype at loci with ISC-specific alleles. 1)
733 Genotypes at ISC9 and ISC4 loci for SureSelect-enriched bone marrow sample (BPK471BM).
734 2) Genotypes at ISC4 and ISC3 loci for SureSelect-enriched bone marrow sample BPK296BM.

735

736 **Table 1. Genotype differences observed in 12 paired samples: clinical samples analyzed by**
737 **SureSelect and derived isolates sequenced classically.** Total base difference* =
738 Homozygous difference x 2 + Heterozygous difference. Almost diploid**: S-value was 2 in all
739 chromosomes except the tetrasomic chromosome 31. Notes: (1) the S-values of
740 chromosome 11 in all bone marrow samples were slightly higher due to amplification
741 artefact and therefore, their values were assumed to be disomic. (2) The S-values of

742 chromosome 31 of BPK87BM and BPK80BM were lower than the expected tetrasomic
743 values due to lower genome coverage.

744 **Supplementary material (10 items)**

745 Fig. S1 to Fig. S8

746 Dataset S1

747 Text S1

748

749 **Legends**

750 **Fig. S1. A. Distribution of probes across *L. donovani* BPK282 reference genome. Most of**
751 **the reference genome is covered by only a single probe.** Some regions are not covered to
752 avoid repetitive regions and regions with homology to mammalian hosts. Changes in the
753 genome assembly subsequent to the bait design and the presence of additional probes to
754 capture *L. infantum* have led to some regions being covered by multiple probes sequences.
755 **B.** Much of the variation in depth in coverage in clinical samples is due to the presence of
756 multiple probe sequences. Plots show the total coverage density across all clinical samples
757 for regions of the reference genome covered by 0, 1, 2 or more probe sequences. **C.** Depth
758 of read coverage in SureSelect enriched samples across two regions of the *Leishmania*
759 genome. Lines show normalized coverage depth (actual coverage per base pair divided by
760 genome-wide mean coverage) for one clinical sample (BPK543, blue) and the sum for all
761 clinical samples (orange). Grey shading shows regions where SureSelect probes map to the
762 genome. Mean coverage for BPK452 was 9.2 reads, for the summed clinical samples 2447.8.

763

764 **Fig. S2. Performance of SuSL-sequencing on artificial mixtures of** *Leishmania* and mammal
765 DNA at three different *Leishmania* DNA percentages: 0.06, 0.006 and 0.0006 %. **A**.
766 Summary statistics for sequencing data from these experiments, showing the total number
767 of reads in each library, the proportion of reads mapping to the *L. donovani* reference
768 genome, and the enrichment factor, calculated as the ratio of the proportion of reads
769 mapping to *L. donovani* in the SureSelect libraries to the proportion of *Leishmania*
770 promastigote DNA included in the pre-pulldown DNA mixtures. **B.** Evenness of genome
771 coverage. The y-axis shows proportion of bases covered with a minimum number of
772 sequencing reads-- read depth along the x-axis -- by reads samples from SureSelect
773 sequencing data from artificial mixtures. 104M, 78M, 52M, 26M and 13M indicate the total
774 number of reads (in millions) sampled from each library; 0.06, 0.006 and 0.0006 are the
775 input proportions of *Leishmania* DNA in percent. For lower *Leishmania* DNA percentages,
776 the higher read depth did not result in higher genome base coverage. **C**. Inferred somy from
777 SureSelect on artificial mixtures. BPK282, pure DNA from promastigote; LD06-0006,
778 simulated clinical samples at three *Leishmania* DNA percentages (0.06%, 0.006% and
779 0.0006% respectively). Y axis shows normalized somy estimate for each chromosome (x
780 axis). Points show central estimate and bars show one standard deviation around these
781 estimates from four replicates of each *Leishmania* DNA percentages 0.06%, 0.006%,
782 0.0006%, respectively.

783

784 **Fig. S3. Map of Nepal showing geographical origin of clinical samples and cultured**
785 **promastigote isolates.** Colours represent different genotype groups defined elsewhere (8).

786 Circles represent Nepalese promastigote isolates from ref 8, crosses bone marrow and
787 spleen samples.

788

789 **Fig. S4. A. Genome sequencing of clinical samples: percentage of the genome covered by**
790 **more than 1 read, with and without SuSL enrichment. B**. Relationship between qPCR
791 estimate of *Leishmania* DNA concentration (y axis) and the proportion of sequencing reads
792 mapping to the *L. donovani* reference genome before SureSelect enrichment in clinical
793 samples for which both qPCR and pre-enrichment sequence data are available, confirming
794 accuracy of the qPCR estimates. Batches were formed of samples with similar proportions of
795 *Leishmania* DNA pre-enrichment.

796

797 **Fig. S5. A. Phylogenetic tree (Neighbor-joining) based on bone marrow (BM), spleen (SP),**
798 **isolates paired to clinical samples (MO) and previously sequenced lines (labelled A1 or B2,**
799 **ref 8).** The tree is identical to the main one (Fig 2A), except that the labels of the samples
800 are indicated. Bone marrow (BM) and spleen (SP) samples are labelled in blue; isolates that
801 are paired to clinical samples are labelled in red, previously sequenced lines are labelled in
802 black. ISC2-ISC10 were sub-populations previously defined (ref 8). See Fig.2A for comments
803 on ISC11. **B.** Results from STRUCTURE analyses from five replicate runs ($2x10^6$ MCMC chains
804 following 10^6 burn-in steps) under the Admixture model assuming 1-15 *K* clusters. The plot
805 on the left shows the estimated loglikelihood of all runs for each *K* cluster. Barplots on the
806 right summarize the assignment probabilities of every *L. donovani* sequence to each
807 inferred cluster assuming 5, 6, 7, 8 or 9 clusters. Arrows denote the CG samples sequenced
808 in this study.

809

810 **Fig. S6. A. Somy in bone marrow and spleen samples.** BM=Bone marrow and SP=spleen.
811 The y-axis and x-axis represent somy and chromosome, respectively. S-values and standard
812 deviation error bars are given in black filled circle and vertical line, respectively. The S- and
813 standard deviation values for the bone marrow samples were based on the binned depth
814 method. The samples were sorted in the order of higher genome coverage, the 3 categories
815 reflect resolution of somy estimation 1) High: whose genome coverage ranging from 99.45%
816 to 89.10%. 2) Medium: whose genome coverage ranging from 86.93% to 56.60%. 3) Low:
817 whose genome coverage ranging from 45.80% to 9.80%. In the lower Leishmania DNA
818 concentration samples, it is still possible to identify higher copy number in chromosome 31.
819 S-values of chromosome 2 and 11 were higher for almost all samples but these higher values
820 were also observed in artificial mixtures of BPK282 where chromosome 2 and 11 were
821 definitively disomic. Close inspection of read depth did not reveal any particular evidences
822 to support their S-values to be higher than 2. **B.** Correlation between gene depth at 303
823 ISC1-specific CNV genes identified in promastigote samples. Correlation was moderately
824 high r^2 = 0.68, p-value = 10^{-76} and slope = 0.882 indicating that SuSL-seq can be used to
825 confirm existing group-specific CNVs. On the plot, colored dots represent counts in log10
826 scale and probability density distributions were given on the corresponding axes.

827

828 **Fig. S7. Allele frequency for the amplicon diagnostic of ISC5 genotype estimated with NGS-**
829 **based deep amplicon sequencing.** Blue line represents the frequency for the non-ISC5
830 allele, while the orange line represents the ISC5-specific allele, which contains a 2bp-
831 insertion in the AQP1 locus. Arrows indicate the two clinical samples in which a second
832 allele was detected above the background, being statistically significant for

833 BPK120_ISC5_BM. BM, Bone Marrow; SP, Spleen aspirate; MO, cultured isolate; NC,
834 Negative Control, a cloned strain where the diagnostic ISC5-specific allele is absent; PC,
835 Positive Control, a cloned strain characterized by the presence of the diagnostic ISC5-
836 specific allele.
837
838 **Fig. S8. Growth competition experiment.** Flasks were inoculated with individual cloned
839 strains or mixtures of two cloned strains in the following ratios: 90/10, 50/50 and 10/90.
840 Cultures were analyzed at the beginning of the experiment (Start), after 5, 15, 20 and 24
841 passages (P5, P15, P20 and P24 respectively). Diagnostic PCR that allows to distinguish
842 between the two compared strains was used to monitor the presence of each strain in
843 culture. A selected fragment of chromatogram containing the variable nucleotide (labelled
844 with red circle) is shown. Arrows point at the samples were the changes in dominance
845 occurred. **A**. Competition between the BPK091, and BPK077 strains (ISC4, and ISC6,
846 respectively). **B**. Competition between the BPK091, and BPK067 strains (ISC4, and ISC3,
847 respectively).
848
849 **Dataset S1. A.** The sequence depth and coverage statistics: The average values of median
850 depth of 36 chromosomes, the depth standard deviation, the genome coverage of
851 SureSelect enriched and raw samples were given for each sample. Low sequence depth
852 samples were marked as LOWDEP, indicating that these samples were not included in the
853 main phylogenetic tree because of their missing depth and coverage. Somy detection
854 accuracy classification was assigned to each sample. High corresponds to 36 high genome
855 coverage samples, Medium corresponds to 16 medium range genome coverage samples
856 and Low corresponds to 9 low genome coverage samples. NA corresponds to 2 samples with

857 almost no informative genome coverage. Further descriptions were found in the text. **B.**
858 Sympatric genotypic diversity at VDC (village development committee, the lowest
859 administrative unit in Nepal) level. VDCs where 2 different ISC genotypes (or more) were
860 identified. See more details on the corresponding samples in Supplementary Table 1. *,
861 because of lower sequence quality, ISC genotype was defined by a set of diagnostic SNPs. [1]
862 Patients 80, 207 and 208 belong to the same family: they presented 2 different genotypes
863 sampled at 10 months of interval. [2] Patients 89 and 91 belong to the same family: they
864 presented 2 different genotypes sampled at the same time. **C.** New insertions identified in
865 SureSelect samples: These indels were in non-repetitive regions and supported with depth
866 greater than 60. They were only identified in SureSelect samples. This illustrated that
867 previously unknown indels can be detected in SureSelect samples when there was sufficient
868 depth. These indel calls were verified on the IGV alignment viewer to ensure the accuracy of
869 indel alternative allele frequencies. **D**. *Leishmania* genome equivalents in samples used for
870 NGS-based amplicon sequencing. The table shows the percentage of Leishmania DNA in the
871 sample, the total amount of DNA, the amount of *Leishmania* DNA and the corresponding
872 *Leishmania* genome equivalents used in each reaction to amplify short DNA fragments
873 containing the diagnostic SNPs/AQP1 insertion for specific ISC genotypes. SuSL-seq
874 genotype denotes the genotype assessed based on the phylogenomic clustering, while
875 amplicon-seq genotype is based on the 4 diagnostic SNPs described by Rai et al. (ref 16). $,
876 there are no ISC group-specific SNPs for ISC9 and ISC11, hence we could only report the
877 negative results for ISC3-6 diagnostic SNPs. **E.** Excluded regions for somy estimation. For all
878 chromosomes, the first 7000 bases and the last 2000 bases were excluded for depth
879 evaluation. In addition, positions (x) in the following ranges were excluded because of the
880 depth errors associated with SureSelect enrichment and because of the known major

881 repetitive or amplified regions (see remarks). **F.** Genotyping using group specific SNP
882 motifs: The table shows the ISC group, chromosome, position, reference allele, SNP allele,
883 SNP score based on GATK, reference base motif and SNP base motif. All the diagnostic SNPs
884 were homozygous SNPs except those of ISC7. Here ISC7 genotype was not detected: in a
885 previous study (8), we found that genotype exclusively in India. A diagnostic SNP and
886 reference base are located at the 12th base. **G.** ISC5 AQP1 insertion motifs: The reference
887 base motif and AQP1 insertion base motif, along with neighboring motives that includes 2
888 base insertions of any base combinations at a given position are shown. The frequency
889 counts of these motives were given for each sample and these neighboring motives were
890 used to estimate the error rates for detecting any two-base insertion nearby. These data
891 indicated that the insertion detection error rates were extremely low. ISC5 strains showed
892 higher insertion error rate but these were likely associated with base errors in the detection
893 of the AQP1 insertion.

894

895 **Appendix (1 item)**

896 **Table A1.** List of Nepalese clinical samples studied here: date of sampling and geographical
897 origin, type of sample, treatment received (AmB, Amphotericin B; SAG, Sodium Antimony
898 Gluconate; MIL, Miltefosine) and outcome, ISC group (according to the results of current
899 study), *because of lower sequence quality, ISC genotype was defined based on a set of
900 diagnostic SNPs, remarks about (i) treatment provided before admission or (ii) family links
901 between patients, isolation outcome and code of successful paired isolates.

902

903

904

905

906

907

908

909

910 **Table 1. Genotype differences observed in 12 paired samples: clinical samples analyzed by**

911 **SureSelect and derived isolates sequenced classically.** Total base difference* =

912 Homozygous difference x 2 + Heterozygous difference. Almost diploid**: S-value was 2 in all

913 chromosomes except the tetrasomic chromosome 31. Notes: (1) the S-values of

914 chromosome 11 in all bone marrow samples were slightly higher due to amplification

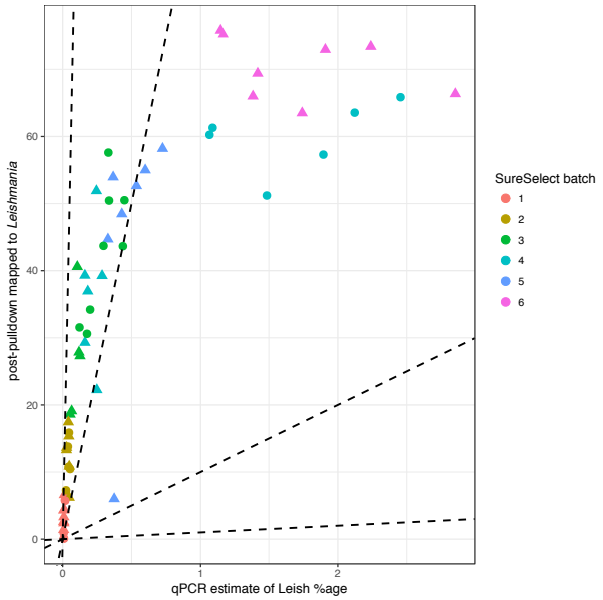
915 artefact and therefore, their values were assumed to be disomic. (2) The S-values of

916 chromosome 31 of BPK87BM and BPK80BM were lower than the expected tetrasomic

917 values due to lower genome coverage.

918

Homozygous difference	Heterozygous difference	Indel difference	Karyotype (bone marrow)	Karyotype (derived isolate)
20	7	4	almost diploid**	trisomy on ch5 and 11
17	8	3	almost diploid	trisomic on ch2, 4, 7, 8, 9, 11, 12, 20, 26 ,33 and 35. tetrasomic on ch23
21	13	4	almost diploid	regular
14	8	8	almost diploid	trisomic on ch20, 26 and 33
34146	167	6275	almost diploid	monosomic on ch17
29942	4370	8281	almost diploid	trisomic on ch23, 33 and 35
26	3	3	almost diploid	trisomic on ch2, 5, 8, 9, 11, 12, 13, 14, 20, 23 ,27 and 32. tetrasomic on ch33
18	4	1	monosomic on ch17	trisomic on ch4, 11, 20, 23. tetrasomic on ch5
9	3	4	almost diploid	trisomic on ch1, 7, 8, 10, 13. 15, 20, 22, 23. tetrasomic on ch9
3	7	LdAQP1: frameshift G/GA insertion	almost diploid	regular
0	5	0	almost diploid	trisomic on ch8, 9, 13 and 23
1	9	0	almost diploid	trisomic on ch1, 7, 8, 10, 13, 15, 20, 22 and 23. tetrasomic on ch9



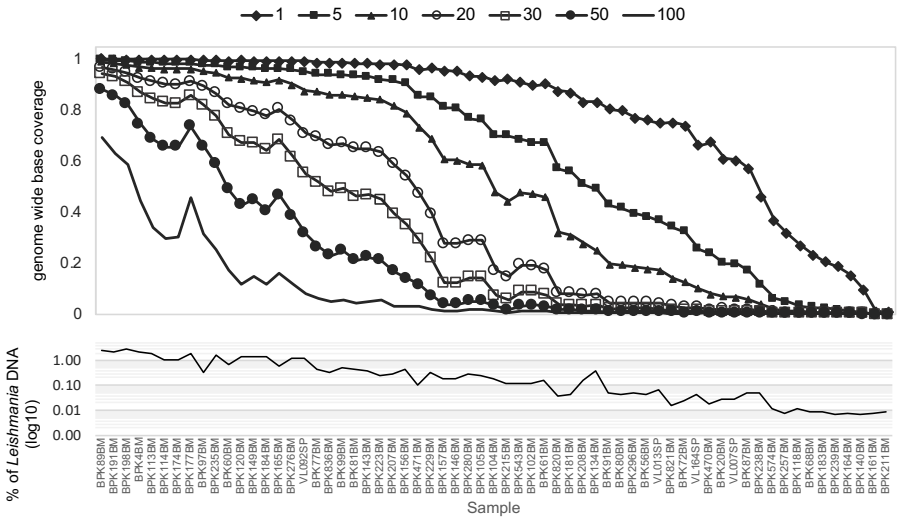


Fig. 1. Performance of SuSL-sequencing on clinical samples. A. Enrichment of *Leishmania* DNA from clinical samples by SureSelect; The y-axis shows percentage of reads mapping to the *L. donovani* reference genome following SureSelect enrichment. The x-axis is the pre-enrichment percentage of *Leishmania* DNA, estimated by either whole-genome sequencing or qPCR; dotted lines indicate no enrichment and 10-, 100- and 1000-fold enrichment. **B.** Read mapping statistics in the 63 samples after SureSelect enrichment. The upper panel shows evenness of genome coverage as the proportion of bases covered with a minimum of number of sequencing reads from 1 to 100 as indicated by the seven different lines. The lower panel shows the percentage of reads mapping to the *Leishmania* reference genome for each sample, shown on the x-axis. The x-axis is common to both panels. BM, Bone Marrow; SP, Spleen aspirate.

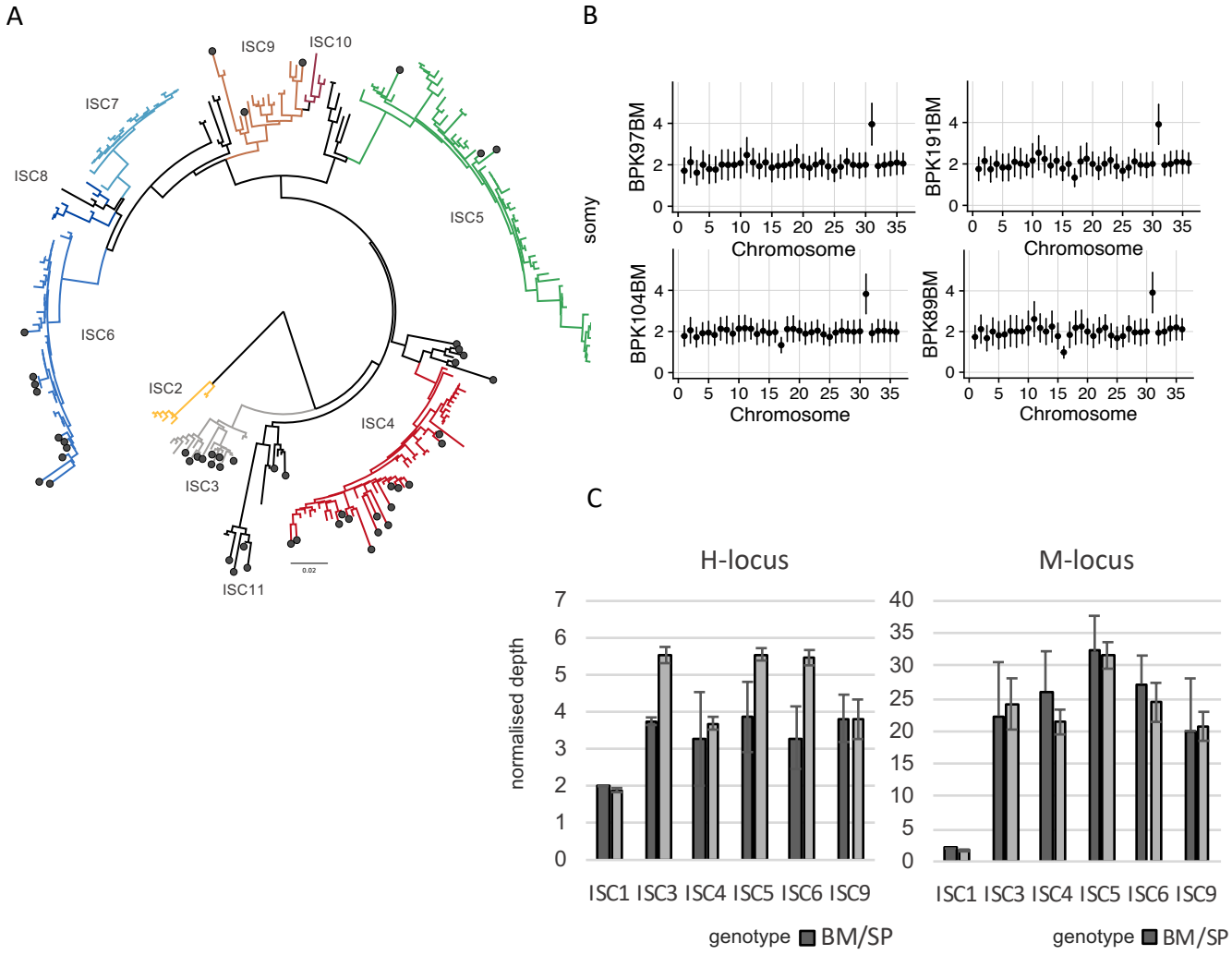


Fig. 2. Genomic diversity among clinical samples. A. SNPs. Phylogenetic tree based on SureSelect enriched clinical samples (black circle) and previously sequenced isolates (8). Diagram is a neighbor-joining phylogeny based on 197 variable sites. ISC2-ISC10 were sub-populations previously defined (8); ISC11 is a novel group that includes clinical samples and isolates previously designated as 'ungrouped'. Full sample identifiers are shown in Fig. S5A. **B.** Karyotypes (High quality samples). Y axis shows normalized somy estimate for each chromosome (x axis). Points show central estimate and bars show one standard deviation around these estimates, calculated by the binned depth method. **C.** Local CNVs. Average copy number per cell of H- and M-loci per ISC group in SureSelect enriched clinical samples (Bone Marrow or Spleen aspirate; BM/SP) and in cultured isolates (promastigotes, Prom). Error bars show one standard deviation around the mean estimate. Only a single sample was available from bone marrow aspirates in ISC1, hence no standard deviation is shown.

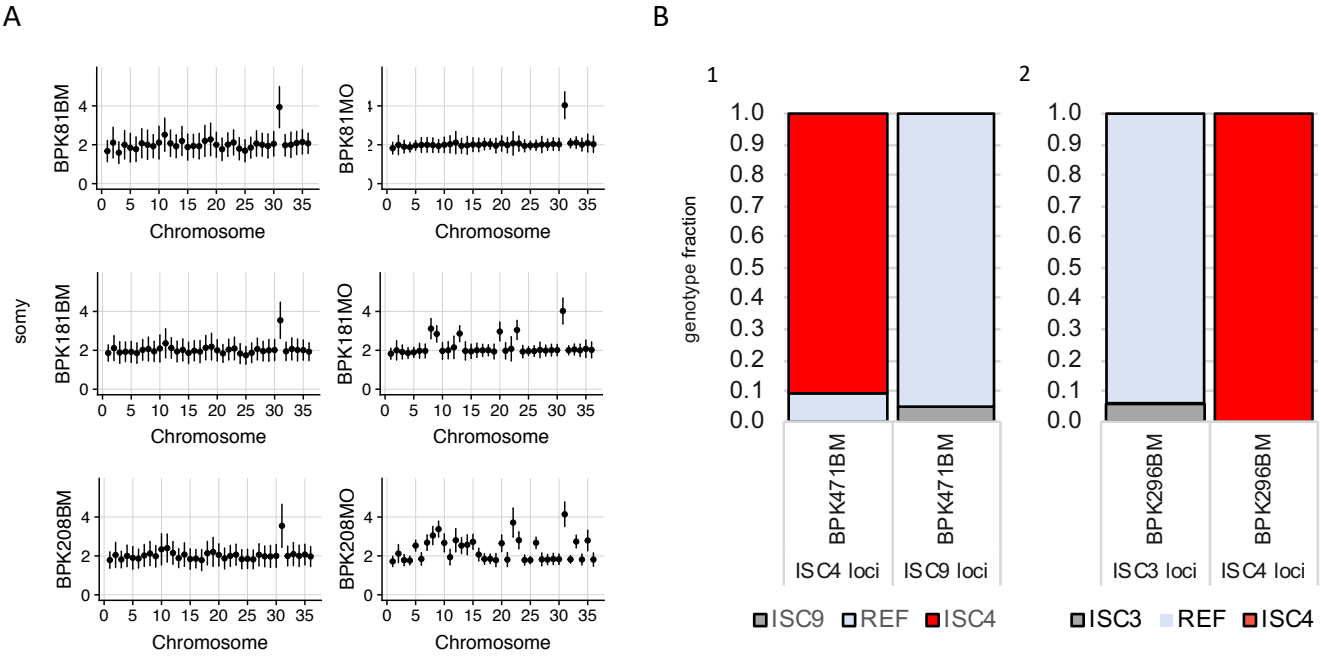


Fig. 3. Somy comparison between matched clinical samples and culture parasites. A. Inferred somy for 3 samples for which SureSelect-enriched bone marrow (BM) samples and cultured isolates (MO) from the same patients had matching ISC genotypes. The y-axis shows normalized somy estimate for each chromosome (the x-axis). Points show central estimate and bars show one standard deviation around these estimates. Somy estimates and standard deviations for the bone marrow samples were based on the binned depth method while corresponding values for isolates were based on depth of each position. **B.** Evidence for polyclonal infections. The bars represent the proportion of sequencing reads showing the ISC-specific genotype or the reference (REF) genotype at loci with ISC-specific alleles. 1) Genotypes at ISC9 and ISC4 loci for SureSelect-enriched bone marrow sample BPK471BM. 2) Genotypes at ISC4 and ISC3 loci for SureSelect-enriched bone marrow sample BPK296BM.