SCIENTIFIC REPORTS
natureresearch

Evaluation of whole genome amplification and bioinformatic methods for the characterization of *Leishmania* genomes at a single cell level

Hideo Imamura[1,4], Pieter Monsieurs[1,4], Marlene Jara[1], Mandy Sanders[2], Ilse Maes[1], Manu Vanaerschot[1], Matthew Berriman[2], James A. Cotton[2], Jean-Claude Dujardin[1,3]✉ & Malgorzata A. Domagalska[1]✉

Here, we report a pilot study paving the way for further single cell genomics studies in *Leishmania*. First, the performances of two commercially available kits for Whole Genome Amplification (WGA), PicoPLEX and RepliG were compared on small amounts of *Leishmania donovani* DNA, testing their ability to preserve specific genetic variations, including aneuploidy levels and SNPs. We show here that the choice of WGA method should be determined by the planned downstream genetic analysis, PicoPLEX and RepliG performing better for aneuploidy and SNP calling, respectively. This comparison allowed us to evaluate and optimize corresponding bio-informatic methods. As PicoPLEX was shown to be the preferred method for studying single cell aneuploidy, this method was applied in a second step, on single cells of *L. braziliensis*, which were sorted by fluorescence activated cell sorting (FACS). Even sequencing depth was achieved in 28 single cells, allowing accurate somy estimation. A dominant karyotype with three aneuploid chromosomes was observed in 25 cells, while two different minor karyotypes were observed in the other cells. Our method thus allowed the detection of aneuploidy mosaicism, and provides a solid basis which can be further refined to concur with higher-throughput single cell genomic methods.

Leishmania are unicellular protozoan parasites belonging to the family Trypanosomatidae[1] and causing a spectrum of diseases in tropical and sub-tropical regions, with an incidence estimated at 1.6 million cases per year[2]. The parasite has a dimorphic life cycle: extracellular flagellated promastigotes in the sand fly vector and intracellular amastigotes in macrophages of the vertebrate host. *Leishmania* have unique genetic and molecular properties that distinguish them from other unicellular and multicellular eukaryotes. Among others, there is no transcription regulation at initiation through promoters. Instead, genes are organized in long arrays of polycistronic units that are transcribed together, then trans-spliced and polyadenylated, and their expression is regulated post-transcriptionally[3]. In this context, gene dosage represents a straightforward strategy for the parasite to modify the expression of genes of interest[4]. This can occur through different mechanisms, like expansion/contraction of tandem arrays, episomal amplifications and aneuploidy[5].

We previously sequenced the genomes of 204 cultivated clinical isolates of *L. donovani* and found aneuploidy in all of them, often affecting half of the 36 chromosomes[6]. Experimental evolution studies suggest that aneuploidy changes constitute an adaptive mechanism to environmental changes like those occurring during the life cycle[7] or those associated with drug pressure[8]. In contrast to many organisms where it can be deleterious, aneuploidy thus appears to be crucial in *Leishmania*. Analyses of aneuploidy at individual promastigote cell level

[1]Institute of Tropical Medicine Antwerp, Molecular Parasitology Unit, Antwerp, Belgium. [2]Wellcome Sanger Institute, Hinxton, UK. [3]Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium. [4]These authors contributed equally: Hideo Imamura and Pieter Monsieurs. ✉email: jcdujardin@itg.be; mdomagalska@itg.be

by FISH revealed an additional dimension, i.e. the concept of mosaic aneuploidy: the tested chromosomes were present in two or more somy states (from monosomic to tetrasomic), varying between cells in a clonal line, so that the karyotype varies from cell to cell[9]. Mosaic aneuploidy could originate from segregation defects, but the currently favored hypothesis is a defect in regulation of chromosome replication[10].

So far, mosaic aneuploidy was only studied for a few *Leishmania* chromosomes, and these pioneering FISH-based studies should be complemented and refined by single cell genomic methods. A range of different methods exist, in which cell- or nucleus-sorting is coupled with whole genome amplification (WGA) and high-throughput sequencing[11]. Single cell genomics is well established in human genetics and cancer research among others[12], but it constitutes an emerging field in parasitology, with a few published studies in parasites like *Plasmodium*[13] or *Cryptosporidium*[14], never -to our knowledge- in *Leishmania* and other Trypanosomatids. We report here the first study on single cell genomics in *Leishmania*. In this pilot study, we first compared the performances of two WGA methods on their ability to preserve specific genetic variations, including aneuploidy levels and SNPs, using a dilution series of *Leishmania* cells. Alongside the assessment of both methods to detect aneuploidy and SNPs at the single cell level, this comparison allowed us to evaluate and optimize corresponding bioinformatic methods. The PicoPLEX approach was more adequate to assess chromosome number and we applied it to study aneuploidy of single cells sorted by fluorescence activated cell sorting (FACS). This allowed us to distinguish the dominating karyotype from minor ones at the single cell level.

Results

Genome coverage and base accuracy.

In a first part of the study, we used a series of samples with decreasing numbers of cells (from 1,000 to 10) of *L. donovani* BPK275 to compare the performance of RepliG and PicoPLEX WGA kits and optimize the downstream bioinformatic protocols. Illumina sequencing of the RepliG and PicoPLEX samples returned 222 million and 111 million 100 bp paired-end reads respectively. The average depth for RepliG was significantly higher than that of PicoPLEX. Specifically, for the five RepliG samples, the total number of reads was 222 million and 96.2% of the reads were mapped to the *L. donovani* BPK282 reference, resulting in average depth of 17.1 ± 14.7. For the five PicoPLEX samples, the total number of reads was 111 million and 41.1% of the reads were mapped, resulting in average depth 3.4 ± 2.4. (Supplementary data 1, Table S1). As somy calculation for RepliG_10 sample was not successful due to uneven sequencing depth resulting in chromosomes with almost no coverage, this sample was omitted in downstream processing.

In order to make an unbiased comparison on the genome coverage provided by both methods, all sequencing data sets were first subsampled to the number of reads from the sample with the lowest sequencing yield (i.e. the PicoPLEX sample starting from 10 cells, containing almost 16 million reads). When comparing the fraction of the genome covered by at least one read between RepliG and PicoPLEX, the former one was only slightly higher: the average ratio of those genome covering fraction for 1,000, 100, 50 and 25 cells was 1.07 ± 0.05 (standard error of the ratios, Supplementary Fig. S1). When considering the fraction of the genome covered by at least ten reads, this ratio RepliG /PicoPLEX increased to 1.53 ± 0.41 for these same samples. Overall a better genome coverage was thus achieved with RepliG than with PicoPLEX, as visualized in the Manhattan depth plots (Supplementary data 2). This higher genome coverage is due to the fact that on average a higher percentage of reads is mapping back to the reference genome for RepliG (96.2%) compared to PicoPLEX (41.1%). Main reason for this large difference in mapping percentage is the presence of a PicoPLEX related adapter sequence (between 16 and 41% of the reads), which could not be efficiently trimmed off. As those reads will not map to the reference, or will be filtered out due to a low mapping quality score, they will not have an impact on the further results in this work. However, when calculating the normalized standard deviation on the read depth – an indicator for evenness of coverage – PicoPLEX shows a lower variation (Supplementary data 1, Table S1): when comparing RepliG with PicoPLEX, the average ratio of those normalized standard deviations for 1,000, 100, 50 and 25 cells was 1.79 ± 0.16 (standard error). This trend can be confirmed by the lower Read Count Variation of PicoPLEX (see Methods). With the exception of the 1,000 cells sample where both methods show a similar value, the Read Count Variation is dramatically higher for RepliG compared to PicoPLEX, indicating a more even read coverage using the latter one.

The third sequence feature comparing both WGA methods was the base accuracy, which was measured by the allele frequency difference between sequence derived from a given number of cells and the control bulk DNA sample (the BPK275 control). Despite the fact that this strain is different from the BPK282 strain used to construct the reference genome sequence, the high genetic similarity between both strains (only 43 SNPs) ascertains that SNPs will not affect the base accuracy interpretation. Furthermore, each dataset was subsampled to prevent the influence of sequencing depth on the SNP prediction performance. Overall, the base accuracy was higher in RepliG samples than in PicoPLEX ones (Fig. 1). The base error rates were 16.1, 9.5, 5.5, 5.1 and 4.4 fold higher in PicoPLEX than RepliG at cell numbers of 1,000, 100, 50 and 25, respectively.

GC bias and read depth.

Both WGA methods here used were reported to be highly affected by GC content in terms of read depth distribution[15–17]. The GC content across the different chromosomes is expected to be uniform, but we found a strong negative correlation between the GC content and length of chromosomes in *L. donovani* BPK282 ($r^2 = 0.586$, *p* value: 5.47e−8, Supplementary Fig. S2 A-B*)* and *L. braziliensis* M2904 ($r^2 = 0.566$, *p* value 1.89e−7, Supplementary Fig. S2 C-D). This correlation was absent in the genomes of *Plasmodium*, *Cryptosporidium*, *Trypanosoma cruzi*, *Giardia* species deposited in EupathDB (https://eupathdb.org/eupathdb/).

Lowess (locally weighted scatterplot smoothing) curves were calculated for the different samples to assess the effect of GC bias on read depth. Normalized depth only moderately depends on the GC content in the BPK275 control, resulting in a straight line with minimal slope (ANOVA on regression slope *p* value < 1e−15, Tukey's post-hoc between BPK275 control and the closest regression line (PicoPLEX_1000) *)* value < 1e−06). However,

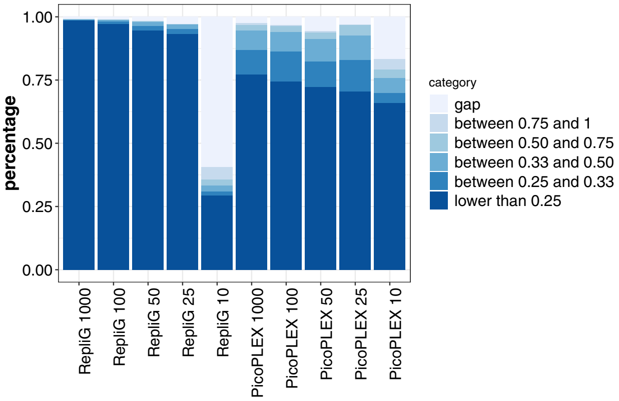


Figure 1. Base accuracy of the RepliG and PicoPLEX BPK275 samples. Genome equivalents are shown along the x-axis. The y-axis represents the percentage of the bases with the given classifications i.e. the alternative allele frequency of the detected SNPs.

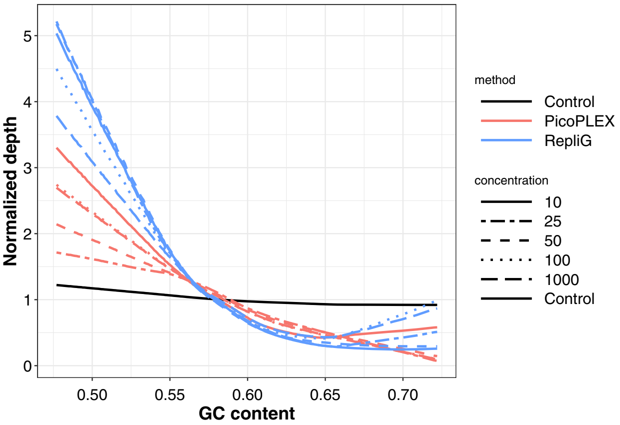


Figure 2. Depth dependency in relation to the GC content. Lowess curves showing the link between GC content and normalized depth for the RepliG samples (light blue), the PicoPLEX samples (light red) and the BPK275 control (black). The line type represents the different cell number equivalents.

the effect of GC content is more pronounced in the amplified samples, with a stronger bias for RepliG treated samples than for PicoPLEX treated ones (Fig. 2).

Somy determination. Boxplots were used to visualize somy values and their variation in the different samples, with and without GC correction, as described in the Methods section. The somy values of the BPK275 control had a negligible depth variability (Supplementary Fig. 3A) and the corresponding somy values corrected for GC bias were nearly identical to the values without GC correction for this control (Supplementary Fig. S4 A). For the RepliG samples, the boxplots suggest a high variability on the somy values (Supplementary Fig. 3 B). GC correction brings the somy values closer to those of the BPK275 control (Supplementary Fig. 4B). However, somy values were still imprecise, as characterized by the average somy deviation (ASD, ranging from 0.420 to 0.637, average = 0.57) and the somy difference count (SDC, ranging from 13 to 21, average = 17.80) as shown in Table 1. In contrast, the PicoPLEX samples showed smaller variation in somy values compared to RepliG (Supplementary Fig. 3C). After GC bias correction for these samples, somy values became closer to those of the control (Supplementary Fig. 4C), which is reflected in an ASD value ranging between 0.299 and 0.418 (average = 0.28), and a SDC value ranging from 2 to 11, average = 6.00 as shown in Table 1. The graphical summary of the ASD for each sample with and without GC bias correction is given in Fig. 3. It shows (i) the lower somy deviation in PicoPLEX samples (Mann–Whitney U = 0, *)* value = 4.05E–03) when compared to RepliG samples and (ii) the decrease in somy deviations after the GC correction is statistically significant for RepliG samples (Mann–Whitney U = 0, *p* value = 6.09E–03) but not for PicoPLEX samples (Mann–Whitney U = 13, *p* value = 2.36E–01). Since PicoPLEX treated samples gave the most accurate somy estimates, this approach was chosen for genome amplification for single cell sequencing.

Single cell genome sequencing. Based on more even genome coverage combined with a more accurate prediction of the real somy values in the above described experiments, we concluded that PicoPLEX is more ade-

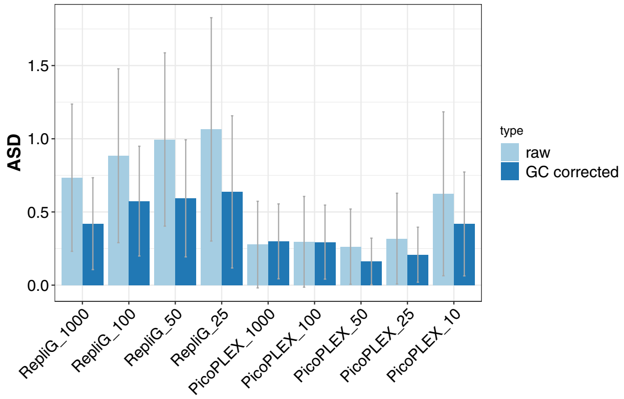


Figure 3. The average somy deviation (ASD) per chromosome between the BPK275 bulk control and the RepliG or the PicoPLEX samples, without (left bar) and with (right bar) GC bias correction. The error bars represent their standard deviation.

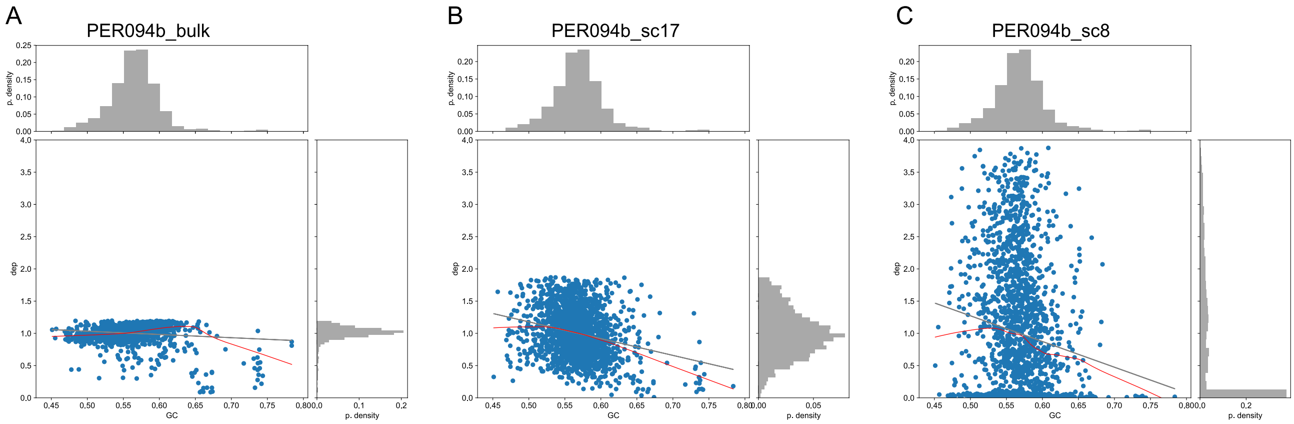


Figure 4. Impact of GC content on sequencing depth in the genome sequence from (**A**) PER094 control, (**B**) and (**C**) a PicoPLEX single cell sample with even depth and uneven depth, respectively. The red lines represent the Lowess curves, the grey lines represent linear regression curves, the top and right histograms represent the histograms for the GC content (GC) and the normalized depth (ND), respectively. PD stands probability density.

	Without GC correction			**With GC correction**		
Cell	**ASD**	**std**	**SDC**	**ASD**	**std**	**SDC**
RepliG_1000	0.73	0.50	20	0.42	0.31	13
RepliG_100	0.88	0.59	21	0.57	0.38	18
RepliG_50	0.99	0.59	28	0.59	0.40	19
RepliG_25	1.06	0.76	27	0.64	0.52	18
PicoPLEX_1000	0.28	0.30	4	0.30	0.26	8
PicoPLEX_100	0.30	0.31	7	0.29	0.25	7
PicoPLEX_50	0.26	0.26	5	0.16	0.16	2
PicoPLEX_25	0.32	0.31	6	0.21	0.19	3
PicoPLEX_10	0.62	0.56	20	0.42	0.36	11
ave_RepliG	0.92	0.61	24.00	0.56	0.40	17.00
ave_PicoPLEX	0.36	0.35	8.40	0.28	0.24	6.20

Table 1. Somy estimation in RepliG and PicoPLEX BPK275 samples, without and with GC bias correction. ASD (average somy deviation) between a sample derived from the respective cell number (ranging from 1,000 to 10) and BPK275 control and corresponding standard deviation. For somy difference count (SDC), we counted in each sample the number of chromosomes showing a somy value difference > 0.5 in comparison to BPK275 control. The average values were summarized at the bottom of the table.

quate than RepliG to accurately predict aneuploidy in single cells. To test this in real-life single cell applications, we applied this method to a line derived from another strain, *L. braziliensis* PER094, which is characterized by a much higher degree of heterozygosity with 57,402 heterozygous SNPs than the previously used *L. donovani* BPK275 with only 43 heterozygous SNPs. This increased number of heterozygous SNPs is essential for allele frequency analyses (see below). Sorting of individual cells of the line PER094 GFP was made by FACS using a GFP fluorescent *L. braziliensis* strain, as described in the Methods section. Two batches of cells were generated, respectively called PER094a (25 cells) and b (22 cells), each of them amplified and sequenced independently. For the batch of 25 PER094a cells, the total number of 151 bp paired-end reads was 183.2 million, of which 13.3% could be mapped to the *L. braziliensis* M2904 reference genome, resulting in average depth 2.46 ± 4.62. Similarly, for the set of 22 PER094b PicoPLEX samples, 188.9 million 151 bp paired-end reads were obtained of which 25.0% of the reads could be mapped to the reference genome, resulting in average depth of 9.4 ± 10.5. All the mapping statistics of total number of reads, mapped reads and the read depth statistics for both sets of PER094 samples were given in Supplementary data 1 (Table S2). General depth trends for PER094a and b samples could be seen in Manhattan depth plots (Supplementary data 3 and 4).

For some samples, mapping percentages are very low, which can be explained by a combination of three factors: 1) presence of a PicoPLEX specific adapter which hampers proper alignment to the reference genome (fraction of reads with adapter varying between 33.80% and 56.87%), 2) Low Phred quality score of the sequencing reads, where around 40% of the samples shows a significant drop in Phred quality score toward the 3′ end of the read. 3) contamination of the FACS instrument with human DNA, which causes between 2.42% and 48.63% (average of 17.59%) of the sequencing reads to be of human origin.

Similar to the BPK275 control, the Lowess approach was used to assess the effect of GC bias on the read depth. It only showed a very limited effect for undiluted and unamplified DNA of PER094 (further called PER094 control) (Fig. 4A). Remarkably, the PER094 control Lowess fit showed an opposite trend compared to the BPK275 control, the former showing a positive trend with increasing GC content. However, for *L. braziliensis* single cells the GC content had a clear impact on the read depth, represented by the negative slopes as shown in Fig. 4B,C for two single cells. Where the normalized depth histogram shows a normal distribution for the even depth case, the depth histogram of the uneven samples shows a peak at a normalized depth of zero, combined with a long tail distribution towards higher normalized depths, as can be seen in the dot plot and the right side depth histogram (Fig. 4). This visual inspection lead to an ad-hoc cut-off of normalized standard deviation of read depth of 0.31, to distinguish between even and uneven depth samples. The Lowess curves of all the single cell samples were shown in Supplementary Fig. 5A,B. Curves from samples with even depth are clearly distinguishable from the ones with uneven depth. Moreover, the trend of those even depth curves is comparable to the trend observed in the *L. donovani* BPK275 PicoPLEX samples.

In a next stage, boxplots were used to visualize the technical variability in the sequencing depth and somy estimates for all samples (Supplementary Figs. 6 and 7). Variability in somy values for the PER094 control, was very limited. For the single cells, variability was higher but this was more pronounced for samples with uneven depth while somy values could be accurately determined for samples with even depth: a single dominant karyotype (called kar1) was observed in 25 single cells and it was characterized by the same trisomy of chromosomes 11 and 25 and tetrasomy of chromosome 31, while all other chromosomes were disomic. This dominant karyotype matched perfectly with the 'average' karyotype observed in the PER094 control. In a few other single cells with even depth, divergent karyotypes were encountered: (i) kar2, with a disomic (instead of trisomic) chromosome 25 in PER094b-sc2 and -sc9 and (ii) kar3, with monosomic chromosome 1 (instead of disomic) in PER094a-sc2 (Supplementary Figs. 6, 7 and 8).

In a last step, we estimated the impact of GC bias correction on somy estimation. Among the even depth samples of PER094a and PER094b batches, GC correction could significantly lower the ASD value from 0.22 to 0.17 for PER094a cells (*p* value 6.08E−05) and from 0.27 to 0.16 for PER094b cells (*p* value 9.01E−05). The same trend was observed for the SDC value (2.14 and 4.4 to 1.0 and 1.6 for PER094a and PER094b cells respectively). (Table 2, Fig. 5A and B). In contrast, values observed in uneven depth samples were much higher, and the effect of GC correction was negligible. The overall impact of GC correction on these samples was illustrated in Supplementary data 5 and 6.

Analysis of somy variation with allele frequency distribution. Somy of a given chromosome can also be predicted by the allele frequency distribution, if heterozygosity is sufficiently high[18], which is the case for PER094. For chromosome 1, we compared the allele frequency patterns of PER094a_sc2 and PER094b_sc15, respectively shown to be monosomic and disomic by methods based on read depth (see above). We observed a typical normally distributed allele frequency of a disomic chromosome shown in PER094 control. In PER094a_sc2, however, we observed a skewed allele distribution for chromosome 1 with clear allele frequency shift toward 0 or 1 (Supplementary Fig. 9A): this was not observed for other chromosomes of that cell, hence this decrease of heterozygosity fits with monosomy of chromosome 1 in PER094a_sc2. In contrast, the chromosome 1 of PER094b_sc15 shows a disomic allele frequency pattern, leveled due to allele frequency dropout, common to single cell samples (Supplementary Fig. 9B). In the case of chromosome 25, the distribution of allele frequency is clearly bi-modal in PER094 control, which is a typical alternative allele frequency pattern of a trisomic chromosome in the bulk sequence (Supplementary Fig. 9C). Similarly, a pattern with two overlapping peaks was detected in chromosome 25 of PER094b_sc15, which is compatible with trisomy (Supplementary Fig. 9D). In contrast, the chromosome 25 of PER094b_sc9 (Supplementary Fig. 9C) showed a flat distribution similar to disomic chromosome 1 of PER094b_sc15 (Supplementary Fig. 9B).

The atypical allele frequency distribution shown above could be due to allele dropout. To verify this, we specifically measured the average pairwise base differences among a subset of PER094b PicoPLEX samples

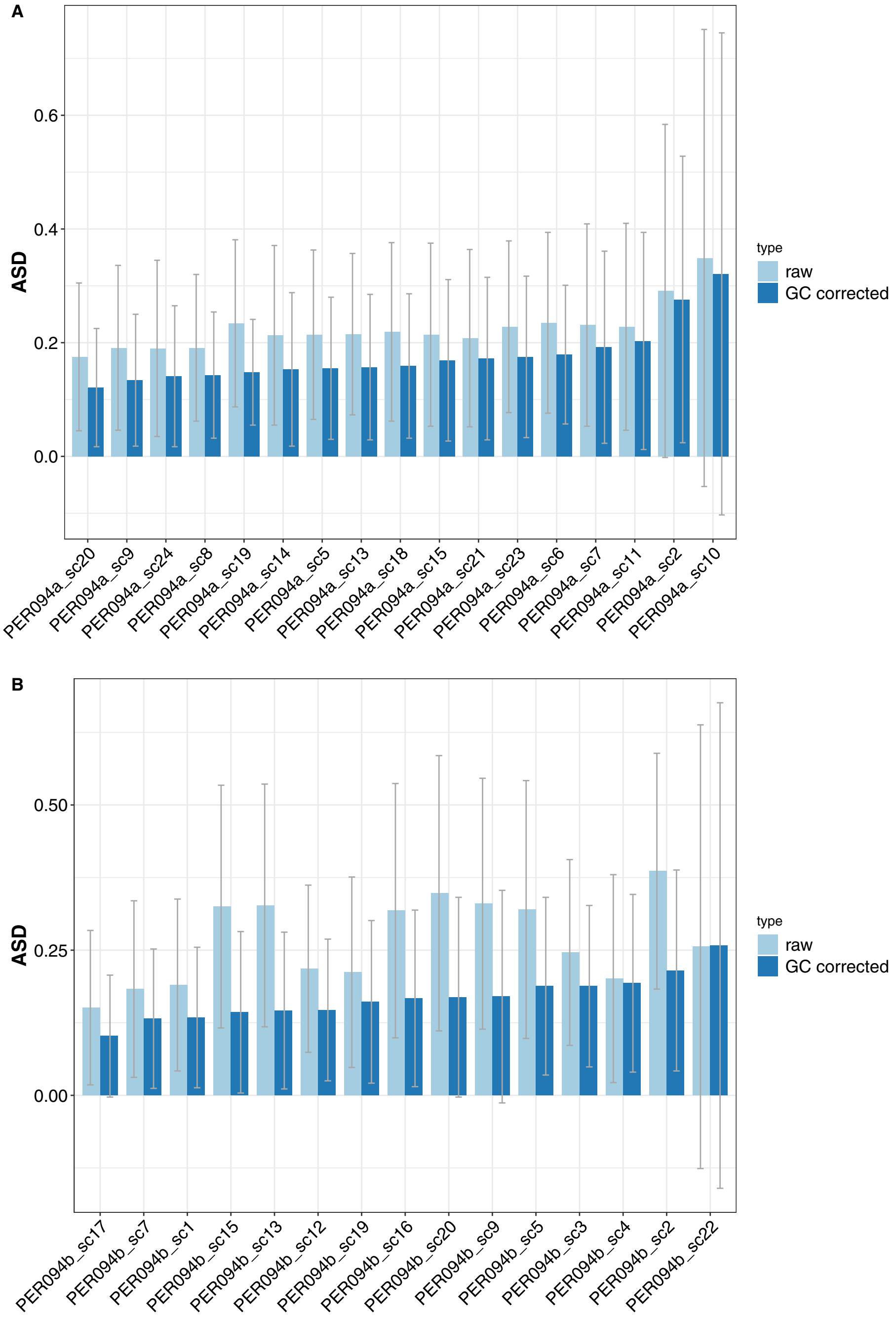
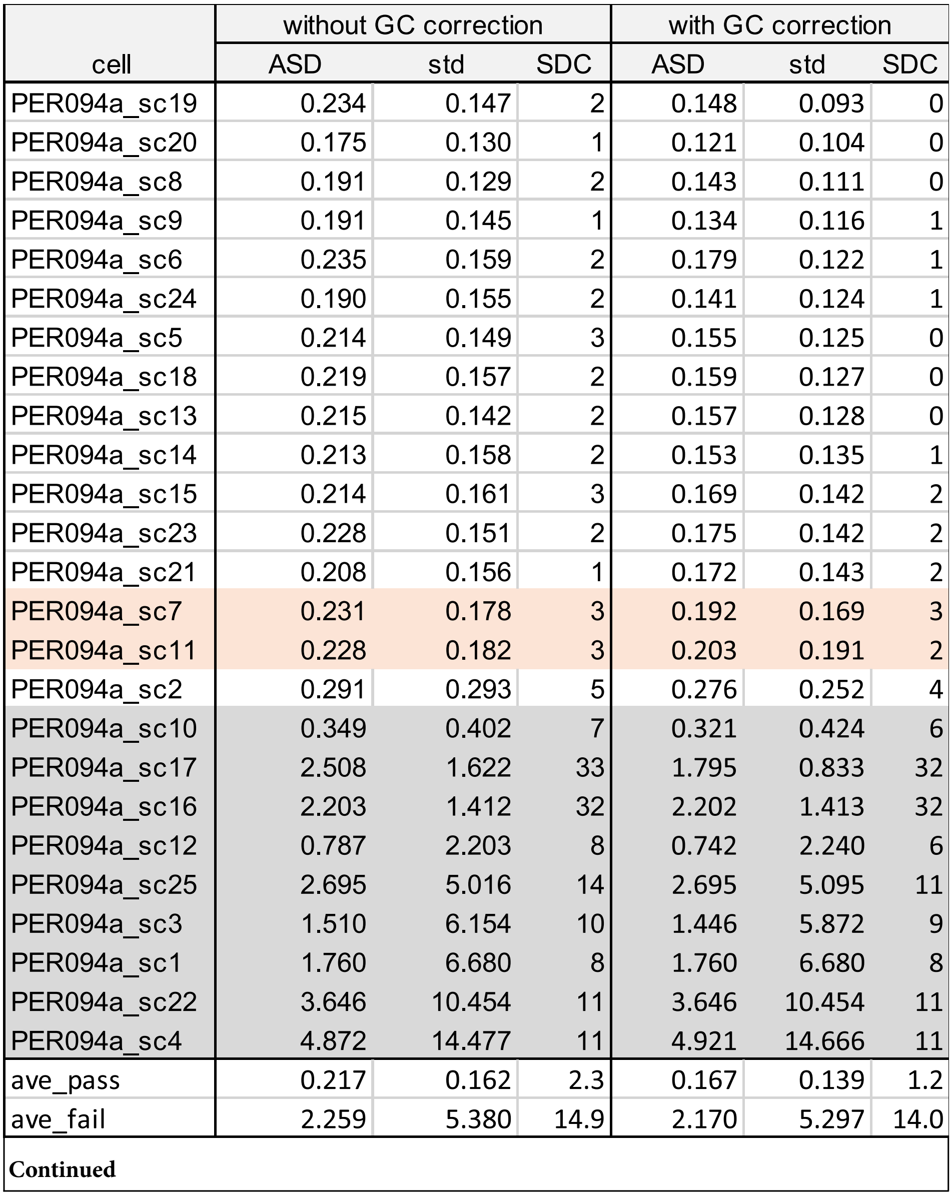


Figure 5. The average somy deviation (ASD) for PER094a and PER094b single cells with even depth coverage. (**A**) ASD for PER094a single cells without (left bar) and with (right bar) GC correction. (**B**) ASD for PER094b single cells without (left bar) and with (right bar) GC correction. The error lines represent the standard deviation. The uneven samples PER094a_sc10 and PER094b_sc22 were added as comparison at the most right side.

(PER094b_sc20, sc15, sc2, sc13, sc9, sc5, sc16). When sites with zero depth were excluded, the average pairwise base difference at 26,959 heterozygous SNP sites was 0.303 ± 0.020 and the corresponding pairwise difference between two bulk PER094 controls was 0.020. When missing depth was treated as a base mismatch, the average pairwise base difference at 57,402 heterozygous SNP sites was 0.349 ± 0.012 and the corresponding pairwise difference between two bulk PER094 controls was 0.019. This shows that allele dropout was quite high in PicoPLEX single cell samples.



	without GC correction			with GC correction		
cell	ASD	std	SDC	ASD	std	SDC
PER094a_sc19	0.234	0.147	2	0.148	0.093	0
PER094a_sc20	0.175	0.130	1	0.121	0.104	0
PER094a_sc8	0.191	0.129	2	0.143	0.111	0
PER094a_sc9	0.191	0.145	1	0.134	0.116	1
PER094a_sc6	0.235	0.159	2	0.179	0.122	1
PER094a_sc24	0.190	0.155	2	0.141	0.124	1
PER094a_sc5	0.214	0.149	3	0.155	0.125	0
PER094a_sc18	0.219	0.157	2	0.159	0.127	0
PER094a_sc13	0.215	0.142	2	0.157	0.128	0
PER094a_sc14	0.213	0.158	2	0.153	0.135	1
PER094a_sc15	0.214	0.161	3	0.169	0.142	2
PER094a_sc23	0.228	0.151	2	0.175	0.142	2
PER094a_sc21	0.208	0.156	1	0.172	0.143	2
PER094a_sc7	0.231	0.178	3	0.192	0.169	3
PER094a_sc11	0.228	0.182	3	0.203	0.191	2
PER094a_sc2	0.291	0.293	5	0.276	0.252	4
PER094a_sc10	0.349	0.402	7	0.321	0.424	6
PER094a_sc17	2.508	1.622	33	1.795	0.833	32
PER094a_sc16	2.203	1.412	32	2.202	1.413	32
PER094a_sc12	0.787	2.203	8	0.742	2.240	6
PER094a_sc25	2.695	5.016	14	2.695	5.095	11
PER094a_sc3	1.510	6.154	10	1.446	5.872	9
PER094a_sc1	1.760	6.680	8	1.760	6.680	8
PER094a_sc22	3.646	10.454	11	3.646	10.454	11
PER094a_sc4	4.872	14.477	11	4.921	14.666	11
ave_pass	0.217	0.162	2.3	0.167	0.139	1.2
ave_fail	2.259	5.380	14.9	2.170	5.297	14.0

Continued

	without GC correction			with GC correction		
cell	ASD	std	SDC	ASD	std	SDC
PER094b_sc17	0 151	0.133	1	0.102	0.105	1
PER094b_sc7	0.183	0.152	2	0.132	0.120	1
PER094b_sc1	0.190	0.148	2	0.134	0.121	1
PER094b_sc15	0.325	0.209	6	0.143	0.139	2
PER094b_sc13	0.327	0.209	7	0.146	0.135	1
PER094b_sc12	0.218	0.144	1	0.147	0.122	1
PER094b_sc19	0.212	0.164	1	0.161	0.140	1
PER094b_sc16	0.318	0.219	7	0.167	0.152	1
PER094b_sc20	0.348	0.237	10	0.169	0.172	2
PER094b_sc9	0.330	0.216	5	0.170	0.183	2
PER094b_sc5	0.320	0.222	6	0.188	0.153	2
PER094b_sc3	0.246	0.160	2	0.188	0.139	2
PER094b_sc4	0.201	0.179	2	0.193	0.153	3
PER094b_sc2	0.386	0.203	10	0.215	0.173	3
PER094b_sc22	0.256	0.382	5	0.258	0.418	6
PER094b_sc8	0.958	0.659	23	0.986	0.606	26
PER094b_sc10	1.481	5.473	8	1.511	5.626	9
PER094b_sc18	1.909	10.174	3	1.792	9.571	2
PER094b_sc6	7.345	16.187	10	7.296	16.173	9
PER094b_sc14	14.311	27.530	16	14.310	27.530	16
PER094b_sc11	18.400	41.837	15	19.219	43.741	14
PER094b_sc21	20.811	42.345	12	21.796	44.363	10
ave_pass	0.268	0.185	4.4	0.161	0.143	1.6
ave_fail	8.184	18.073	11.5	8.396	18.503	11.5

Table 2. Estimation of somy deviation in PER094a and b single cell samples, without and with GC bias correction. ASD and SDC between a single cell sample and BPK094 control and corresponding standard deviation (see Table 1 for explanation of ASD and SDC). The single cells with even depth are shown in white and the single cells with uneven depth are shown in gray. At the bottom of the table, average of high quality (even depth, ave_pass) and low quality (uneven depth, ave_fail) are indicated. Two intermediate depth quality samples PER094a_sc7 and PER094a_sc11 (indicated in orange) were not included in the summary statistics.

Discussion

In the present paper, we report the first study – to our knowledge – of the genomes of single *Leishmania* cells. Using *L. donovani* as model, we showed in a first step that the choice of WGA method should be determined by the planned downstream genetic analysis and that PicoPLEX and RepliG are more suitable for aneuploidy and SNP calling, respectively, as expected from previous work[11]. Given our interest in aneuploidy[6,7,19] and mosaicism[20], we evaluated and developed bioinformatic methods to assess the somy for each chromosome when only limited sequencing data is available, thereby taking into account and correcting for the GC-bias. In a second step, we used the PicoPLEX approach to sequence the genome of FACS-sorted single cells and determined their aneuploidy by the computational pipeline optimized in the first step. For this experiment, we used *L. braziliensis* as model, as its higher heterozygosity should permit to study allele frequency distribution, which is an complementary inferential method for somy estimation often applied in bulk sequencing[18]. We successfully called somy of all 35 chromosomes in 28 out of the 47 single cells, detected aneuploidy mosaicism by read-depth based methods and subsequently validated monosomy and trisomy of some chromosomes based on their signatory somy specific allele frequency distribution.

Bioinformatic pipelines were optimized to handle high technical read depth variation observed throughout the genome because of the required amplification process of WGA. We have developed and validated on low DNA amounts depth normalization methods. We tested a various range of window sizes and percentiles to quantify sequencing depth, greatly enhancing the sensitivity and specificity for somy detection. As indicated within the results, there is large variation on the sequencing coverage and evenness of sequencing depth between different samples or cells, but window of 5 kb seems to be the optimal value when considering all samples discussed in the manuscript. It also appeared critical to select the lower and upper depth bin cut offs for somy estimation; accordingly, these parameters must be optimized for each data set separately. In general we found that the combination of higher read depth and even read distribution always led to improved somy estimation based on depth and read allele frequency. Samples with even but low depth could still lead to accurate somy estimation. However, samples with higher intra-chromosomal read depth variation, regardless of its read depth size, always led to very poor somy accuracy. Therefore, based on those observations we advocate that quality filtering should primarily be focused on selecting those single cell sequencing data that have an even read depth, and only as a second and less important criterion look at the sequencing depth.

Previous *Leishmania* studies based on bulk sequencing have shown some correlation between read coverage depth and chromosome length[7,18,21]. In a previous study where we used SureSelect, a genome capture method, we observed the tendency that the shorter chromosomes had lower depth which resulted in more non-integer somy values compared to longer chromosomes in the sequencing results[19]. This bias was corrected using sequencing depth normalization using chromosomes with similar lengths. The observation of non-integer somy values in smaller chromosomes could be explained by a higher mosaicism among this size-class of chromosomes[18]. However, present study provides an alternative answer to these observations, as we found in reference genomes of both *L. donovani* and *L. braziliensis* a negative correlation between GC content and chromosome length. Interestingly, this appeared to be a *Leishmania*-specific feature as this GC-bias was not found in other tested genomes. The detection of this GC-bias has an impact for single-cell genomics of *Leishmania*, because WGA methods are reported to be particularly sensitive to GC content[15–17]. In our analysis, the slopes of the Lowess curves, used to represent the link between normalized depth and GC content were bigger in *L. donovani* RepliG samples than in PicoPLEX ones (Fig. 2). Accordingly, GC correction appears to be important for single cell genomics of *Leishmania*, but also for other species where the GC content is varying over the chromosomes like e.g. *Leptomonas pyrrhocoris* (data not shown).

In this study we sorted cells using FACS and we amplified and sequenced separately single genomes. Major disadvantages, previously reported to be associated with a FACS-based approach, are the high risk of contamination with foreign DNA during collection process and the high cost of WGA of each sample. First, the size of the microbial genome is much smaller than mammalian genomes and even small fragments of this DNA will efficiently compete with the microbial DNA during WGA. This was clearly illustrated for bacteria[22] and could represent an issue for *Leishmania* genome which is 100 fold smaller than human genome. To reduce the contamination risk, we cleaned and sterilized the FACS instrument, but treatment with a DNA removing agent was impossible. Second, the high cost of WGA in FACS-based approaches is related to the volume of sample collection (5 µl here) and reaction (75 µl here) and this may also have an effect on the efficiency of the amplification of minute amounts of DNA in the sample[22]. In our experiments, we tested the amplified *Leishmania* genomes for presence of human DNA by qPCR of only one human gene, RPL30 (see Supplementary methods). The presence of this gene was detected only in few samples, and these were excluded from further studies. However, different degree of contamination with human DNA was still detected by WGS, but this was not found to be a critical factor for the somy accuracy. Indeed, a large proportion of unmapped reads did not necessarily lead to the lower somy calling accuracy. Reciprocally, a large proportion of mapped reads did not necessarily lead to higher somy calling accuracy, the latter mainly because this type of samples often showed an uneven genome coverage. Using automated, microfluidics- or droplet-based single cell sorting and sequencing library preparation platforms would greatly reduce this risk and increase reproducibility of single cell whole genome analysis. Among the additional advantages of these methods, we mention: 1) the volume of the reaction (a few nanoliters), decreasing risk of contamination and cost and increasing reaction's efficacy; 2) the use of chips, which decreases the number of operations and the risk of contamination; 3) pre-staining or constructing parasites harboring a fluorescent marker is not needed, thus most types of unprocessed cells can be analyzed without lengthy preparation; 4) the high-throughput character of analyses and 5) the possibility -with some platforms- to check the number of cells present in each sample and sequence only the ones containing only one cell, hereby also reducing costs and ensuring that only individual cells are analyzed.

Despite the low number of *L. braziliensis* cells analysed here (28 with even depth), mosaicism could be detected, with 3 different karyotypes, all aneuploid: one dominant karyotype (kar1) in 25 cells and two others (kar2 and kar3) each one encountered in 2 and 1 cells respectively. Interestingly, kar2 and kar3 only differed slightly from kar1, by the somy of 1 chromosome (chr 1 and 25) respectively. Chromosome 31, shown to be tetrasomic in bulk sequences of all *Leishmania* species studied so far[6,18,21] was tetrasomic too in all 28 cells analyzed here. Probably because of allele drop-out, rather high in PicoPLEX samples, allele frequency distribution curves of individual chromosomes (expected to be monomodal, bi-modal and tri-modal for disomic, trisomic and tetrasomic chromosomes) were atypical in single cells. However, we could detect the allele frequency signatures of monosomic chromosome 1 and trisomic chromosome 25 and thus validate the read-depth based somy calling of these chromosomes. As such, our single cell sequencing data confirm the hypothesis of mosaic aneuploidy which was derived based on FISH data.

This study paves the way for further single cell genomics studies in *Leishmania*. The FACS-based approach described here is of interest for in-depth analysis of genomes in a small number of cells (for instance a plate of 96 cells), while different WGA methods should be used depending on the planned downstream genetic analysis (SNP, indel, CNV or aneuploidy). High-throughput analyses of single cells are needed to investigate the extent and dynamics of aneuploidy mosaicism in *Leishmania*, in both stable and variable experimental conditions. Therefore, microfluidics- and droplet-based platforms represent a promising alternative and several options exist (see recent review[23]).

Materials and methods

Preparation of samples for comparison of WGA methods.

The cloned line *L. donovani* MHOM/NP/03/BPK275/0-Cl18[21] was grown on HOMEM and harvested 20 passages after cloning. Parasites were washed and resuspended in PBS, to have 1,000, 100, 50, 25 and 10 promastigotes in 2 µl of PBS. For PicoPLEX (New England Biolabs), 2 µl of parasites were put in PCR tubes and flash-frozen in liquid nitrogen. For RepliG (Qiagen), 2 µl of parasites + 2 µl of PBS were put in PCR tubes and flash frozen in liquid nitrogen. PicoPLEX and RepliG samples were then further processed according to manufacturer's instructions. The average size of the DNA fragments after amplification was as expected, and differed between RepliG and PicoPLEX, being 2–100 kb and 0.1–1 kb, respectively. DNA of each sample was purified and concentrated with Genomic DNA Clean and Concentrator-25 (Zymo Research) according to the manufacturer's instructions. DNA from the same *L. donovani* line and the same batch was extracted with the QIAamp DNA Mini Kit (Qiagen) according to manufacturer's instructions and used as bulk control (further called BPK275 control). Samples were sent to Wellcome Trust Sanger Institute for whole genome sequencing: (i) in 40 µl, with a DNA concentration ranging between 63.3 and 91.3 ng/µl for RepliG samples, (ii) in 34 µl, with a DNA concentration ranging between 9.64 and 40.4 ng/µl for PicoPLEX samples and (iii) in 40 µl, with a concentration of 50.4 ng/µl for BPK275 control.

Preparation of samples for single cell analysis. A transgenic line of *L. braziliensis* strain MHOM/PE/02/PER094 with constitutive expression of the GFP reporter integrated within the ribosomal locus was generated as previously reported elsewhere[24]. After transfection and two weeks of selection a clonal/isogenic line was derived by the micro drop method. The resulting GFP fluorescent line PER094-GFP-Cl2 (further called PER094-GFP) was used to sort single cells within 96 well plates with the BD FACSAria II with a 85 µM nozzle and 45 PSI. Briefly, 2 ml of parasites were washed with 10 mL of PBS. Subsequently, the pellet was resuspended in another 10 mL of PBS and gently passed through a 5 µM filter (pipetting and gravity). The recovered cells were concentrated by centrifugation (1,500 g, 5 min) and brought to a new suspension of 20×10^6 parasites/mL in medium M199 + 100 units/mL of penicillin and 100 µg/mL of streptomycin. For sorting the single cells, gates were selected by using the side and forward scatter plots. A non-GFP wild type was included to establish the autofluorescence values and the gate corresponding to GFP positive cells. The single cells were sorted in a 96 well plate (containing 5 µL of lysis buffer) and immediately placed on dry ice until the next step. In parallel, DNA from the same PER094-GFP line and the same batch as the one used for single cells was extracted with the QIAamp DNA Mini Kit (Qiagen) according to manufacturer's instructions and used as bulk control (further called PER094 control).

Two sorting experiments were made and generated 2 batches of 39 and 40 single cells. A quality control was introduced before further selecting cells for WGA and WGS. DNA contained in each well was amplified by 4 qPCR assays targeting (i) kDNA, 18 s rDNA and G6PD to assess for the presence of good quality *Leishmania* DNA and (ii) RPL30 to assess for the presence of possible human DNA contamination (supplementary methods). After this process, the 2 batches resulted in 25 and 22 cells, further called PER094a and PER094b, respectively.

The performance of the cell sorter was also evaluated by sorting fluorescent beads (10 µM, Beckman Coulter) within 384 plates (optically clear flat bottom, Perkin Elmer). The presence of only one bead was corroborated by visualization of the whole well with a confocal microscope (Zeiss LSM 700). The percentage of wells with only one bead was 78.8% while the percentage of wells without or with more than one bead were 20.8 and 0.4% respectively.

Whole genome sequencing. For Illumina sequencing genomic DNA was sheared into 400–600 base pair (bp) fragments by focused ultrasonication (Covaris Adaptive Focused Acoustics technology, AFA Inc., Woburn, USA).

Amplification-free Illumina libraries were prepared[25], for PER094 samples 150 bp paired-end reads were generated on the Illumina HiSeq X, whilst for BPK275 samples 10% PhiX DNA was added to the library pool to increase complexity then 100 bp paired-end reads were generated on the Illumina HiSeq 2,500 following the manufacturer's standard sequencing protocols[26].

DNA read mapping, SNP calling. Paired-end reads from the RepliG and PicoPLEX samples of *L. donovani* and from BPK275 control were mapped to the improved reference *L. donovani* genome LdBPKv2[7] (available at ftp://ftp.sanger.ac.uk/pub/project/pathogens/Leishmania/donovani/LdBPKPAC2016beta/) using Smalt v7.4 (https://www.sanger.ac.uk/science/tools/smalt-0). Similarly for *L. braziliensis*, paired-end reads from 22 PER094b and 25 PER094a FACS-sorted single cells and the PER094 control were mapped to the *L. braziliensis* M2904 reference genome, which was recently improved based on PacBio SMRT sequencing. The parameters used in Smalt were described in the previous studies[7,19], which – apart from the default settings – implies activating exhaustive search for optimal alignments (*-x*), random mapping of multiple hit reads (*-r 3*), and requiring at least 80% of the nucleotides in the read being a perfect match *(-y 0.80*). Picard v1.85 (https://broadinstitute.github.io/picard/) was used for marking duplicated reads. For calculation of the genome coverage for the different samples, samtools was used to subsample all data to the number of reads obtained for the sample with the lowest yield. To assess the evenness of genome coverage, the variation on the genome coverage within 5 kb windows was calculated using the normalized standard deviation (also called coefficient of variation), i.e. the standard deviation of the sequencing depth within 5 kb windows divided by the average sequencing depth calculated over all 5 kb windows. The read count variation was calculated as specified using 5 kb windows[27].

For SNP calling for the evaluation of the accuracy of sequences derived from amplified and bulk DNA, we used the population SNP calling mode of UnifiedGenotyper in Genome Analysis Toolkit v3.4, with the SNP cut off (GATK QUAL score) (i) 1,500 for BPK275 samples and (ii) 2000 for PER094 samples (GATK: https://software.broadinstitute.org/gatk/)[28]. The SNP cut off for *L. braziliensis* was higher since more samples were used to call SNPs. The main goal of the SNP analysis in this study was to evaluate base accuracy and the base recovery rate, i.e. how many bases would be covered by at least one read. Therefore, no lower nor higher depth cut off was imposed in the population SNP calling. In *L. donovani* chr22, the base positions between position 736,224 and the end of that chromosome were excluded from the SNP calling as this region might be an erroneous region in the genome reference sequence, since many SNPs are also detected in other BPK275 related whole genome sequencing data (data not shown) .

GC content and GC bias correction.

With WGA methods like RepliG and PicoPLEX, normalized depth is reported to be affected by GC content bias[15–17]. Previous *Leishmania* studies based on bulk cell populations have suggested a correlation between read coverage depth and chromosome length[7,21]. However, the mechanism of depth bias was not fully understood. Neighboring chromosome normalization[7] was used to assess this effect. However, this will not be accurate for samples in which the variability of chromosome copy number is irregular and skewed. First the GC content was measured in 5 kb windows across each chromosome. To evaluate the impact of GC content bias on the current data sets, we selected the long disomic chromosomes 28, 29, 30, 32 and 34 of *L. donovani* and *L. braziliensis*. These disomic chromosomes were used to avoid the depth difference due to aneuploidy. For each sample, we fitted a Lowess (Locally Weighted Scatterplot Smoothing) curve in terms of the GC content and normalized depth, using a Python package statsmodels v0.9.0 (https://github.com/statsmodels/statsmodels). Depths greater than the 95 percentile were marked as outliers and removed.

In the next step this Lowess curve is used to correct the somy value for each chromosome separately. Based on the average GC content of a chromosome, its somy value is divided by the value of the Lowess curve for that specific GC percentage using a lookup hash table approach. In general, the Lowess curves properly represented the relationship between the GC content and depth within a GC content range between 55 and 65% for most of the high-quality samples. As the average GC content of each chromosome was between 56.2% and 61.5% for both *L. braziliensis* and *L. donovani* genomes, accurate GC correction could be guaranteed. Finally, somy values were renormalized using a median somy value after the initial GC correction to have the median somy value over all chromosomes close to 2. The method used here was analogous to the GC bias correction methods described in the previous studies[29,30]. Somy values with and without the GC correction were visualized with Gnuplot[31].

Somy estimation.

WGA methods are known to produce highly variable depth coverage[15,17]. To mitigate skewed, uneven read distribution[7], the chromosome median depth was calculated as a trimmed median of mean depths for 5000 bp bins, where bins with a depth less than the 10th percentile or greater than 90th percentile were removed. Somy values are visualized using boxplots as implemented using Matplotlib[32]. To assess depth across all the chromosomes, Manhattan plots across all the chromosomes were created based on median depth of 5000 bp windows. The upper limit of a Manhattan plot was set to be twice the value of a 98 percentile to focus on the informative range.

Somy range.

The range of monosomy, disomy, trisomy, tetrasomy, and pentasomy was defined to be the full cell-normalized chromosome 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 $4.5 \leq S < 5.5$, respectively[7]. However, single cell sequencing depth can be highly variable among cells unlike in bulk sequencing. To overcome this technical depth variability among individual cells and to characterize somy variability clearly, we defined the Average Somy Deviation (ASD): this is defined as the average difference between the calculated somy value and the true (integer) somy value, with the latter one calculated based on the bulk control sample. Similarly, the somy difference count (SDC) is defined as the number of chromosomes where the absolute difference between the predicted somy value and the true somy value is greater than 0.5 .

When defining a new karyotype (i.e. a cell having for at least one chromosome a somy value deviating from the somy value based on the bulk genome sequencing), a more stringent criterion is applied. First, the median somy value per chromosome is calculated over all cells with an even depth. Next, for each chromosome for each cell the absolute difference with this median value is calculated. If this absolute difference is higher than one, the cell is defined as having an aberrant karyotype.

Visualization of pairwise allele frequency difference.

Alternative base allele frequencies (allele frequencies) of variable sites were extracted from the GATK SNP vcf files. To quantify base recovery rate, a gap was considered to be homozygous mismatch of allele difference of one, instead of discarding missing sites or imputing the bases. For *L. braziliensis* single cells, we were particularly interested in identifying allele frequency shifts due amplification artifacts. Therefore we first identified high quality heterozygous sites in the PER094b and PER094a bulk samples, and only assessed allele frequency differences. We did not take this approach for the *L. donovani* samples since there are not enough heterozygous SNPs in BPK275. An allele frequency distribution of two samples was visualized with an allele frequency dot plot where the x- and y-coordinates represented their alternative allele frequencies, and their corresponding histogram was also given along each axis (Supplementary Fig. 9).

Data availability

Sequencing data are available in ENA (European Nucleotide Archive) study PRJEB8793.

Received: 21 February 2020; Accepted: 19 August 2020
Published online: 14 September 2020

References

1. Maslov, D. A. *et al.* Recent advances in trypanosomatid research: genome organization, expression, metabolism, taxonomy and evolution. *Parasitology* **146**, 1–27 (2019).
2. Alvar, J. *et al.* Leishmaniasis worldwide and global estimates of its incidence. *PLoS ONE* **7**, e35671 (2012).

3. Requena, J. M. Lights and shadows on gene organization and regulation of gene expression in Leishmania. *Front. Biosci. (Landmark Ed.)* **16**, 2069–2085 (2011).
4. Victoir, K. & Dujardin, J. C. How to succeed in parasitic life without sex? Asking Leishmania. *Trends Parasitol.* **18**, 81–85 (2002).
5. Berg, M., Mannaert, A., Vanaerschot, M., Van Der Auwera, G. & Dujardin, J. C. (Post-) Genomic approaches to tackle drug resistance in Leishmania. *Parasitology* **140**, 1492–1505 (2013).
6. Imamura, H. *et al.* Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent. *Elife* **5**, e12613 (2016).
7. Dumetz, F. *et al.* Modulation of aneuploidy in leishmania donovani during adaptation to different in vitro and in vivo environments and its impact on gene expression. *MBio* **8**, (2017).
8. Shaw, C. D. *et al.* In vitro selection of miltefosine resistance in promastigotes of Leishmania donovani from Nepal: Genomic and metabolomic characterization. *Mol. Microbiol.* **99**, 1134–1148 (2016).
9. Sterkers, Y. *et al.* Novel insights into genome plasticity in Eukaryotes: Mosaic aneuploidy in Leishmania. *Mol. Microbiol.* **86**, 15–23 (2012).
10. Sterkers, Y., Crobu, L., Lachaud, L., Pagès, M. & Bastien, P. Parasexuality and mosaic aneuploidy in Leishmania: alternative genetics. *Trends Parasitol.* **30**, 429–435 (2014).
11. Macaulay, I. C. & Voet, T. Single Cell Genomics: Advances and Future Perspectives. *PLoS Genet.* **10**, e1004126 (2014).
12. Chen, W. *et al.* Single cell omics: from assay design to biomedical application. *Biotechnol. J.* https://doi.org/10.1002/biot.201900262 (2019).
13. Nair, S. *et al.* Single-cell genomics for dissection of complex malaria infections. *Genome Res.* **24**, 1028–1038 (2014).
14. Troell, K. *et al.* Cryptosporidium as a testbed for single cell genome characterization of unicellular eukaryotes. *BMC Genomics* https://doi.org/10.1186/s12864-016-2815-y (2016).
15. Zhang, C. Z. *et al.* Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat. Commun.* **6**, 6822 (2015).
16. Chen, D. Y. *et al.* Comparison of single cell sequencing data between two whole genome amplification methods on two sequencing platforms. *Sci. Rep.* **8**, 4963 (2018).
17. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060 (2015).
18. 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).
19. Domagalska, M. A. *et al.* Genomes of Leishmania parasites directly sequenced from patients with visceral leishmaniasis in the Indian subcontinent. *PLoS Negl. Trop. Dis.* **13**, e0007900 (2019).
20. Barja, P. P. *et al.* Haplotype selection as an adaptive mechanism in the protozoan pathogen Leishmania donovani. *Nat. Ecol. Evol.* **1**, 1961–1969 (2017).
21. 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).
22. de Bourcy, C. F. A. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE* **9**, e105585 (2014).
23. Barea, J. S., Lee, J. & Kang, D. K. Recent advances in droplet-based microfluidic technologies for biochemistry and molecular biology. *Micromachines* **10**, 412 (2019).
24. Jara, M. *et al.* Tracking of quiescence in Leishmania by quantifying the expression of GFP in the ribosomal DNA locus. *bioRxiv* 641290 (2019) https://doi.org/10.1101/641290.
25. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6**, 291–295 (2009).
26. Bronner, I. F., Quail, M. A., Turner, D. J. & Swerdlow, H. Improved Protocols for Illumina Sequencing. *Curr. Protoc. Hum. Genet.* **80**, 18.2.1-42 (2014).
27. Deleye, L. *et al.* Performance of four modern whole genome amplification methods for copy number variant detection in single cells. *Sci. Rep.* **7**, 3422 (2017).
28. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
29. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gks001 (2012).
30. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res.* https://doi.org/10.1101/gr.175141.114 (2014).
31. Williams, T., Kelley, C. & many others. Gnuplot 5.2: an interactive plotting program. (2018).
32. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Acknowledgements

This study received financial support from the Flemish Ministry of Science and Innovation (SOFI Grants SINGLE and MADLEI), the Flemish Fund for Scientific Research (FWO, post-doctoral grant to MJ). JAC, MB and MS were supported by Wellcome via the core support of the Wellcome Sanger Institute (grant 206194).

Author contributions

All authors have approved the submitted version of this manuscript and have agreed both to be personally accountable for their own contributions and to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This work was conceived and designed by H.I., M.B., J.A.C., M.V., J.C.D. and M.A.D. Data were acquired and analyzed by H.I., P.M., M.J., M.S., I.M., J.A.C. and M.A.D. Data interpretation was made by H.I., P.M., M.J., J.A.C., J.C.D. and M.A.D. Paper was drafted by H.I., P.M., J.C.D. and M.A.D. and substantively revised by M.J., M.S., M.V., M.B. and J.A.C.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-71882-2.

Correspondence and requests for materials should be addressed to J.-C.D. or M.A.D.

Reprints and permissions information is available at www.nature.com/reprints.

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