A novel high-throughput single-cell DNA sequencing method reveals hidden genomic heterogeneity in the unicellular eukaryote *Leishmania*

Gabriel H. Negreira[1], Pieter Monsieurs[1], Jean-Claude Dujardin[2], Malgorzata A. Domagalska[1,*]

[1] Experimental Parasitology Unit, Institute of Tropical Medicine Antwerp, Antwerp, Belgium.

[2] Molecular Parasitology Unit, Institute of Tropical Medicine Antwerp, Antwerp, Belgium.

* Corresponding Author. E-mail: mdomagalska@itg.be.

ABSTRACT

Genome instability is considered a major driver of adaptation in eukaryotic microorganisms, but its study is hampered by the limited availability of genomic technologies with single-cell resolution. Here we present a novel high throughput method to reconstruct both structural and nucleotide information at single-cell level in *Leishmania*, a protozoan parasite with a remarkable genome plasticity characterized by frequent gene copy number variations (CNVs) and high aneuploidy mosaicism. By combining the use of semi-permeable capsules with primary template-directed amplification, we could determine the karyotypes of hundreds of *Leishmania* parasites, detect distinct CNVs between different cell populations, and identify sub population of cells harboring distinct nucleotide variants, including in genes associated with drug resistance. This approach provides a powerful new framework to uncover hidden evolutionary potential in complex microbial populations, with application in studying adaptation, drug resistance, and genome evolution in *Leishmania* and other pathogens.

INTRODUCTION

Genome instability plays a paradoxical and yet pivotal role in the evolutionary success of many unicellular pathogens. While genomic integrity is typically vital for cellular function, moderate levels of instability – either through intra-chromosomal copy number variations (CNV) or through dosage changes of whole chromosomes (aneuploidy) – can facilitate survival under environmental stress by quickly diversifying phenotypes in a cell population[1–3]. For instance, aneuploidy promotes survival to drug pressure and host immunity in *Candida albicans*[4,5], and CNVs were linked to survival under nitrogen-limiting conditions in *Saccharomyces cerevisiae*[6]. Aneuploidy and/or CNVs are also commonly found in cancer and are linked to tumor progression and drug resistance [7,8]. Moreover, these types of structural variation are also found in protist parasites such as *Giardia intestinalis*, *Plasmodium spp.*, *Trypanosoma cruzi*, and *Leishmania*[9–13], representing potential adaptive mechanisms in clinically relevant organisms.

To fully understand the causes and impact of genome instability, as well as its adaptive role, methods with single-cell resolution are required, since bulk sequencing cannot resolve genomic heterogeneity, masking rare or transient variants that drive adaptation. While in recent years single-cell genomics has received increased attention, the main focus has been on transcription and chromatin accessibility, with single-cell DNA sequencing (scDNAseq) technologies lagging. Currently available options are often limited in throughput, as cells are usually manually processed in individual PCR plate wells, restricting the number of sequenced cells to the plate's capacity. In single-cell RNA workflows, this limitation is often overcome using microfluidic compartmentalization, which encapsulates cells into nanoliter emulsion droplets, allowing bulk processing while preserving single-cell resolution[14,15]. However, the only commercial emulsion droplet-based solution for scDNAseq – the Chromium™ Single Cell CNV solution from 10x Genomics™ – was discontinued in 2020[16], with no replacement currently available to our knowledge. This gap highlights the urgent need for novel, high-throughput scDNAseq approaches.

Although emulsion droplet approaches can increase throughput and decrease per-cell costs, they face a major limitation: their impermeable nature requires all reagents for genome barcoding and amplification – the 2 major steps in scDNAseq – to be co-encapsulated with the cells, complicating the development and optimization of custom workflows. In this sense, semi-permeable capsules (SPCs) offer a promising alternative: unlike standard emulsion droplets, these capsules have a porous shell that is selectively permeable: large biomolecules (such as genomic DNA or whole organelles) are retained inside the capsule, while small molecules (e.g. enzymes, buffers, metabolites) can diffuse in and out[17]. This property permits multi-step reactions to be performed in the pools of encapsulated cells by simply washing or exchanging the surrounding solution. Another major advantage of SPCs is their physical resistance, which allow the use of strong chemical agents for lysing bacterial and potentially fungal cell walls, as well as their application in techniques with extreme temperatures, such as sample preservation at -80 ˚C and enzymatic reactions requiring high temperatures (up to 100 ˚C). These features make SPCs a flexible solution to develop custom single-cell pipelines compared to emulsion droplets. It also allows the parallel isolation of up to 100.000 cells[18], increasing throughput while reducing per-cell cost. In addition, SPC

encapsulation can be achieved using relatively low-cost instruments. Importantly, SPC-based scDNAseq was originally developed for bacterial cells, with its applicability in eukaryotic systems still largely unexplored.

In this context, *Leishmania* – a protist parasite that cause leishmaniasis – represents an ideal system to test such a platform to study genome instability in eukaryotes, owing to its remarkable genomic plasticity and the central role that genome instability plays in its biology. First, while most eukaryotes maintain a euploid genome, i.e., with equal copies of each chromosome, *Leishmania* breaks this rule, as aneuploidy is not an exception but the genomic standard for these parasites. In its most basic form, aneuploidy is present as a characteristic polysomy of chromosome 31 contrasting with an otherwise fully disomic genome[19]. However, high levels of chromosome instability lead to frequent dosage changes involving virtually any of the 36 chromosomes, making any *Leishmania* population a 'mosaic' of distinct aneuploid karyotypes[20,21] . Aneuploidy patterns can shift in response to environmental stresses [22–25] and are directly reflected into RNA levels[22,26] and partially at protein levels[27]. In addition, *Leishmania* genome instability is also expressed in the form of frequent CNVs. Small segments of 20–70 kb can form extrachromosomal circular or linear amplicons carrying genes, including those encoding drug resistance factors[28]. Expansion of intra-chromosomal loci is also common. For instance, a common CNV found in chromosome 23 of some *L. donovani* strains was linked to resistance to antimonials in strains from Nepal[23,29]. In addition, the occurrence of mixed infections – where multiple *Leishmania* strains or even species coexist in the same host[30–32] – further highlight the need for genomic approaches that resolve parasite genotypes at single-cell resolution.

In a previous study our group established the first single-cell genomics study with *Leishmania*, using single parasites isolated in plates with fluorescence activated cell sorting (FACS)[33]. We could successfully sequence 47 parasites, of which 28 had their karyotypes determined. Moreover, we also demonstrated that the whole-genome amplification (WGA) method must be chosen based on the downstream analysis, with multiple displacement amplification (MDA) being more suited for nucleotide sequence information, and PicoPlex performing better for structural variations such as chromosome copy number assessment and CNVs[33]. Later, we described the application of the Chromium single-cell CNV solution (10X Genomics) as a high throughput, emulsion droplet-based approach that could reconstruct karyotypes in thousands of *Leishmania* cells, revealing potential paths towards karyotype diversification, but with no access to nucleotide variant detection due to shallow genomic coverage [21].

Here we developed a novel high throughput method tailored to detecting structural and nucleotide variants at single-cell level in eukaryotic parasites such as *Leishmania*. Our method is based on the SPC-based commercially available Single-Microbe DNA Barcoding solution (Atrandi Biosciences), which originally offers a workflow for SPC encapsulation, WGA and genome barcoding of bacterial cells. We compared the performance of the method with libraries previously made with the Chromium single-cell CNV, showing that the original SPC-based protocol successfully generates single-cell DNA libraries with *Leishmania*, but its uneven amplification prevents determination of structural variants. We solved this issue by using primary template-directed amplification (PTA)[34] as alternative WGA. This modification allowed us to fully

reconstruct the karyotype of hundreds of *L. donovani* parasites and to detect distinct CNV patterns between cells. Moreover, the higher coverage seen in PTA unlocked the ability to simultaneously retrieve nucleotide sequence information from single-cells, which was not possible before. We demonstrate the ability of the method to distinguish parasite subpopulations of the same species, as well as small populations in the same strain, revealing minor genotypic variants among cells in clinical isolates, including variants in genes related to drug-resistance in small subpopulations of parasites. Together, our results represent a significant advancement in single-cell genomics for *Leishmania* and other eukaryotes, providing a scalable, high-resolution tool to dissect both structural and sequence-level variation, and to deepen our understanding of parasite heterogeneity, evolution, and adaptation.

MATERIALS AND METHODS

Parasite Culture

In the present study, we refer to terms such as population, strain and clone, according to the nomenclature proposed for salivarian trypanosomes[35]. Accordingly: (i) a population is a group of *Leishmania* cells present at a given time in a given environment (e.g., culture or host); (ii) a strain is a population derived by serial passage *in vitro* from a primary isolate (in this case, from clinical samples) without any implication of homogeneity but with some degree of characterization (here, bulk genome sequencing); (iii) a clone is a population of cells derived from a single individual presumably by binary fission. Here, two *L. donovani* strains isolated from visceral leishmaniasis patients were used, the BPK081/0 cl8 (MHOM/NP/02/BPK081/0) strain, which was isolated in Nepal, and HU3 (MHOM/ET/67), isolated in Ethiopia. Both strains were retained at the cryobank of the Institute of Tropical Medicine Antwerp and were maintained in HOMEM medium supplemented with 20% fetal bovine serum at 26 ˚C with periodic passages done every 7 days in 1 in 50 dilutions for downstream experiments. The BPK081/0 cl8 strain was submitted to single-cell sequencing 7 passages after cloning, while HU3 had an unknown of passages (minimum of 15 passages).

Single-cell encapsulation into SPCs

Promastigotes from each strain were collected independently at late-stationary phase (day 7) by centrifugation at 1500 rcf for 5 minutes followed by 3 washes with PBS (magnesium and calcium-free) and passed through a 5 µm strainer to remove cell clumps. Then, cell concentration was adjusted to 10^7 parasites/ml in each strain, and both were combined at 50/50 ratio for downstream applications. The cell mixture was submitted to SPC encapsulation using the Flux device (Atrandi), following the Flux SPC Generator protocol (V6.3), with 6.3 µl of the cell suspension loaded, targeting a total of 15.000 encapsulated cells (corresponding to ~5.000 sequenced cells). After encapsulation, SPCs were washed 3 times with 1 ml of Capsule Wash Buffer (CWB - Atrandi) and were stored at -80˚C overnight. Afterwards, SPCs were thawed back to room temperature, washed 3 times with 1 ml of Wash Buffer (WB - Atrandi) and resuspended in 480 µl WB. The SPC suspension was subsequently divided into 6 samples with 80 µl each which were maintained on ice. Samples were named according to the WGA method applied: SPC-STD1-4 for the samples submitted to the standard WGA of the Single-Microbe Genome Barcoding Kit, and SPC-PTA1-2 for the samples submitted to PTA. A table summarizing the treatments that each sample was submitted to is provided in the results section (Table 1).

Cell lysis

Of the samples described above, 5 were submitted to different variations of the lysis protocol described in the Single-Microbe DNA Barcoding user guide (Atrandi) version 4.1 (P/N: CKP-BARK1): sample SPC-STD1 was first fixed by slowly adding 900 µl of ice-cold methanol to the SPC suspension on ice, following by an incubation at -20 ˚C for 30 minutes, 3 washes with 1 ml WB, and a final resuspension in 80 µl of WB. For samples SPC-STD2-4 and SPC-PTA1-2, methanol fixation was skipped. Then, the 5 samples received 80 µl of 2X Lysis Buffer (0.8 M KOH, 20 mM EDTA, 200 mM DTT - Atrandi), with

samples SPC-STD1, SPC-STD2, and SPC-PTA1 being incubated at 23 ˚C for 15 minutes, sample SPC-STD3 being incubated at 23 ˚C for 60 minutes, and sample SPC-STD4 being incubated at 99 ˚C for 15 minutes for lysis. After lysis, each sample was washed 3 times with 160 µl of Neutralization Buffer (NB - Atrandi) followed by 3 washes with WB, with final resuspension in 25 µl WB, with tubes being kept on ice afterwards.

For sample SPC-PTA2, the lysis protocol from the ResolveDNA kit (BioSkryb) was followed instead. In summary, SPCs were washed 3 times with 25 µl of Cell Buffer (CB – BioSkryb). Then, 25 µl of the MS mix (BioSkryb) were added, followed by an incubation at room temperature for 20 minutes at 1400 rpm in a thermomixer. Then, SPCs received 25 µl of SN1 reagent (BioSkryb) with another room temperature incubation for 1 minute at 1400 rpm followed the addition of 25 µl of the SDX reagent (BioSkryb) with a 10-minute incubation at room temperature without agitation. After lysis, SPCs were kept on ice until PTA. Since the components of the PTA lysis protocol could affect the downstream PTA, we also added the aforementioned lysis reagents from the PTA protocol to SPC-PTA1, skipping the incubation times.

Whole-genome amplification

Samples SPC-STD1-4 were submitted to the standard WGA method described in the Single-Microbe DNA Barcoding user guide (Atrandi). In summary, the 25 µl SPC suspensions were combined with 16,25 µl of nuclease-free water, 6.25 µl of 10X WGA Reaction Buffer (Atrandi), 6.25 µl of dNTP Mix (Atrandi), 3.125 µl Primer mix (Atrandi), 0.625 µl of 0.1 M DTT, 0.625 µl of 10% Pluronic F-68, 1.25 µl of WGA Enhancer (Atrandi), and 3.125 µl of WGA Polymerase (Atrandi) on ice. The WGA mixture was then submitted to an incubation at 45 ˚C for 15 minutes followed by 65 ˚C for 10 minutes in a thermocycler and were returned to ice.

For samples SPC-PTA1-2, the PTA protocol from the ResolveDNA kit version 1 (BioSkryb) was followed instead. In summary, a PTA master mix was prepared on ice by combining 85 µl of SB4 reagent, 17 µl of 1X SS2 reagent, 13,6 µl of SEZ1 reagent, and 20,4 µl of SEZ2 reagent (all from ResolveDNA kit). Then, 66,7 µl of the PTA master mix was added to the 100 µl of lysed SPCs of samples SPC-PTA1-2. These samples were then submitted to a 10-hour incubation at 30 ˚C followed by an inactivation at 65 ˚C for 3 minutes.

After WGA, all samples were washed 3 times with 200 µl of WB, with final resuspension done in 25 µl. DNA amplification was evaluated in all samples by staining 1 µl of SPCs with 5µM SYTO9 (ThermoFisher) and visualizing SPC positives in a fluorescence microscope. For samples SPC-STD1-4, amplified DNA was debranched by combining the 25 µl of SPC suspension with 17.5 µl of Nuclease-free water, 5 µl of 10X Debranching Buffer (Atrandi) and 2.5 µl of Debranching Enzyme (Atrandi) followed by an incubation at 37 ˚C for 1 hour at 1000 rpm agitation in a thermomixer. After debranching, SPCs were again washed 3 times with WB with final volume of 25 µl. The samples submitted to PTA SPC-PTA1-2 did not require debranching and were directly submitted to barcoding instead.

Barcoding of amplified genomes

Barcoding of amplified SPC-trapped DNA was done following the split-pool barcoding approach described in the Single-Microbe DNA Barcoding user guide (Atrandi) for all 6 samples. First, DNA end preparation was done to allow the subsequent ligation of barcodes to the amplified DNA molecules. This was done by adding 3.5 µl of End Prep Buffer (Atrandi), and 1.5 µl of End Prep Enzyme (Atrandi) to the 25 µl of SPC suspensions, followed by an incubation at 20 ˚C for 30 minutes and inactivation at 65 ˚C for 30 minutes. Each sample immediately received 25 µl of a ligation master mix (77 µl of nuclease-free water, 66 µl of 10X ligation buffer (Atrandi) and 22 µl of Ligation Enzyme (Atrandi) and was distributed along 4 wells (10 µl per well) of the first-round columns of the barcoding plate from the Single-Microbe DNA Barcoding kit (Atrandi). The well positions of each sample was recorded to be used as sample index later. Barcode ligation was carried out at 25 ˚C for 15 minutes and stopped with 40 µl of 1X Stop Buffer (SB - Atrandi) with a 5-minute incubation at room temperature. After that, SPCs from all wells were pooled in a single tube, washed 5 times with 1 ml WB (final volume of 150 µl), combined with 150 µl of a new ligation master mix, and redistributed along the 24 wells of the second-round columns in the barcoding plate (10 µl per well). Barcode ligation incubation and stop were done as described above. These steps were repeated 2 more times for a total of 4 barcoding rounds. Finally, the barcoded SPC pool was washed 5 times with SB and was stored at 4 ˚C with final volume of 200 µl.

Library preparation

The barcoded SPC pool was divided in 4 PCR tubes with ~1.250 expected barcoded genomes each (50 µl per tube). Each tube received 2 µl of release reagent (Atrandi) and was incubated at room temperature for 5 minutes to disrupt the SPCs and free the barcoded DNA. The pooled DNA was purified with 0.8X AmPure XP beads (Beckman Coulter) and was eluted in 27 µl of nuclease-free water per tube. DNA fragmentation, adapter ligation and library amplification were done using the NEB Next Ultra II FS DNA Library Prep kit for Illumina (New England Biolabs), with indexing primers UDI5-UDI8 from the Single Microbe DNA Barcoding Kit (Atrandi), and 12 PCR cycles. Final libraries were purified with 0.8X AMPure XP Beads. Of the 4 final libraries, 2 (~ 2.500 expected cells) were sent for sequencing. Sequencing was done at NovoGene using a NextSeq (Illumina) in high output mode with 150 pair-end reads.

Cell demultiplexing and read mapping of SPC-scDNA libraries

Demultiplexing of the raw sequencing data into individual single-cell FASTQ files was performed using custom Python scripts in a two-step procedure. Initially, the second read – containing four concatenated 8-nucleotide barcodes with structure [NNNNNNNN]AGAA[NNNNNNNN]ACTC[NNNNNNNN]AAGG[NNNNNNNN]T – was scanned for a perfect match against a list of expected barcode sequences of 8-nucleotide barcodes at the first, second, third, or fourth position. Barcodes were retained only if they perfectly matched the list. Barcodes observed at least 1,000 times were classified as dominant barcodes. The process of screening the second read was then repeated for the identified dominant barcodes, extending the dataset by permitting

up to one mismatch per 8nt barcode with the whitelist barcodes. This resulted in paired FASTQ files (R1 and R2) generated for each single cell.

These paired FASTQ files were then aligned to the *Leishmania donovani* BPK282 version 2 reference genome[22] using BWA[36]. Alignments were filtered for a minimum mapping quality of 30, duplicates were removed using Picard[37], and only properly paired reads were retained. The resulting BAM files served as input for the bamCoverage module of DeepTools[38], which calculated average coverage over 20 kb windows normalized to counts per million reads. This was used to generate a count matrix of cells (columns) by bins (rows) dimension. Counts were then randomly subsampled to a total of 100.000 reads per cell to allow cross-sample comparisons. The count matrix was used for cell quality control and karyotyping.

Single-cell sequencing with 10X-scDNA (10X Genomics)

ScDNAseq was already available for the BPK081 strain from a previous study using the 10X-scDNA solution[21]. For the HU3 strain, a new experiment was done using a kit bought before the discontinuation of the product. Cell encapsulation, WGA, barcoding and library prep were done as described in the Chromium Single-cell CNV user guide (rev C). Sequencing was carried out at the Genomics Core Leuven on a NovaSeq SP sequencer (Illumina) with 150 pair-end reads. Read demultiplexing, mapping, cell identification and read count matrix generation were done with Cellranger DNA (10X Genomics).

Compensation of GC bias and mappability

Compensation of GC bias was done by fitting a generalized additive model with bin GC content as explanatory variable and the mean normalized read counts as response variable using the mgcv package in R[39], after removal of outlier values for GC and count. The resulting model was used to predict expected counts for each bin based on their GC content, and a bin-specific correction factor was computed as the ratio between the median of the total predicted values to the predicted value for each bin. Raw counts were then adjusted by multiplying each bin's counts across all cells by its respective correction factor. Negative values arising from the correction were set to zero. The corrected count matrix was then used in downstream analysis.

To determine bin mappability, i.e., the fraction of reads that are uniquely aligned to a given genomic bin, a paired-end artificial library of 150 bp reads was simulated with ART Illumina[40] using a 400 bp insert size, 60 bp standard deviation, and 1x coverage. These reads were mapped back to the BPK282v2 reference genome with BWA-MEM, and the mappability of a bin was calculated as the fraction of reads mapping to it with MAPQ >30 over total mapped reads. Bins with mappability below 0.7 were considered non-mappable and excluded from downstream analyses for karyotyping and CNVs.

Data quality control

True cells were distinguished from background signal by ranking barcodes in descending order of total read count and plotting the log10-transformed read counts against log10-transformed barcode ranks. The inflection ("knee") point of this curve, representing the transition from true cells to background, was identified by detecting the point of

maximum perpendicular distance from a line connecting the curve's endpoints. Barcodes to the left of the knee were retained as true cells.

For estimation of coverage evenness, two metrics were used: (i) the gini index, which was calculated for each cell using the gc-corrected and cell-normalized read counts per 20 kb bin with the function `gini` from the ineq package[41], and (ii), the median absolute pairwise difference (MAPD) which corresponds to the median absolute difference between the gc-corrected normalized counts of two consecutive bins[42]. For karyotyping, two additional parameters were used to detect noisy cells in the dataset. First, an intra-chromosomal coefficient of variance (ICCV) was determined by calculating the mean normalized standard deviation of chromosomes in a cell. This summarizes how 'spread' the read coverage is around the chromosomal means. To quantify replication-associated coverage variation and distinguish replicating (S-phase) cells from non-replicative (G1/G2) cells, we developed a second score named the Intra-Chromosomal Fluctuation (ICF) score. For each cell, and for each chromosome, locally estimated scatterplot smoothing (LOESS) model is fitted to the coverage profile (read count per 20 kb bin as a function of genomic bin position). The fitted curve represents the smoothed intra-chromosomal coverage trend. Then, for each bin, we calculated the absolute distance between the LOESS fit and the mean coverage of the corresponding chromosome in that cell. This absolute deviation captures the extent of fluctuation around a flat (non-replicative) coverage profile. We computed the mean absolute deviation per chromosome, and for each cell, selected the five chromosomes with the highest deviations. The final ICF score for the cell is the mean of these five highest per-chromosome deviation values. High ICF values indicate structured, wave-like coverage profiles consistent with DNA replication activity, while low ICF values reflect flatter profiles characteristic of non-replicating cells.

Doublet detection

Doublet detection was done as previously reported[21]. In summary, previously generated bulk whole genome sequencing data of strains HU3 and BPK081 was first used to determine strain specific variants. This was done with the Genome Analysis Toolkit (GATK) version 4.1.4.1. This information was then used to attribute reads from the single-cell data to each strain. Cells exhibiting a mixture of diagnostic SNPs from both strains above 5% were classified as potential doublets.

Single-cell karyotype estimation

Karyotype estimation was done as previously described[21]. In summary, raw somies were determined by averaging the counts of each chromosome and multiplying them by the cell's scale factor which tries to infer the baseline ploidy of the cell by approximating the normalized values to integers. For this, we iteratively test, for each cell, scale factors within the range of 1.8 to 5 and compute, for each scale factor, the mean absolute distance of the scaled values to the nearest integer. The scale factor that minimized this distance is selected as the optimal estimate of the cell's ploidy. This method leverages the assumption that true somy values are integers, and that the optimal scale factor should result in values that are as close to integers as possible. Afterwards, raw somy values were converted to integers using Gaussian Mixture Models built at populational

level for each chromosome using the mixtools package in R[43], with one gaussian fitting each possible copy number state for that particular chromosome. The combined integer somy values of a cell defines its karyotype.

Copy number variations

Copy number variations were determined for two loci which differ between the two strains used in this study: a ~4 kb locus in chr19 which is present in HU3 but absent in BPK081, and the M-locus, a ~8 kb locus in chr36 which is present in BPK081 but absent in HU3. To estimate differences in the copy numbers, we generated a new count matrix with DeepTools as mentioned before, but this time with 100bp bins. Correction for GC bias and mappability was done as mentioned above. We then aggregated the corrected counts of the bins inside each locus and divided by the mean count of all other bins of the chromosome. This gave an estimated haploid copy number for the locus.

Nucleotide Variants

Additionally, all BAM files were used as input for variant calling with the Genome Analysis Toolkit (GATK) version 4.1.4.1 following the GATK best practices workflow: single-cell level variant calling with HaplotypeCaller generating GVCF files, combining GVCFs of single cells from the same sample using CombineGVCFs, joint genotyping with GenotypeGVCFs, and variant filtering with SelectVariants and VariantFiltration commands according to GATK recommendations. Known drug resistance marker regions were extracted from the VCF files using BCFtools and visualized with the ggheatmap from ggalign[44] package in R.

Phylogenetic reconstruction

For reconstruction of phylogenetic relationships between cells, the ScisTree2 package was used [45,46]. First the allele depth matrix generated with GATK was filtered in R with the following criteria: (i) Heterozyogous alleles (frequency < 0.9) were converted to homozygous reference allele, as the package expects only homozygous variants, (ii). Alternative alleles were only retained if mapped by two or more reads, otherwise they were also converted to reference, (iii), alternative alleles were only maintained if found in 2 or more cells, and (iv) loci with NA values in more than 40% of the cells were removed. The filtered allele depth matrix was then used to estimate genotype probabilities using the probability.from_reads function from ScisTree2 in Python, with default parameters (allelic dropout = 0.2, sequencing error = 0.01, and posterior = True). Trees were constructed with ScisTree2 function, with maximum number of iterations set to 1000. The generated trees were visualized in R using ggtree[47].

RESULTS

Two *L. donovani* strains, BPK081, and HU3, were chosen for this study based on the following criteria: (i) both strains have distinct genotypes, with HU3 displaying 122.495 small nucleotide variants compared to the reference BPK282v2, in contrast with BPK081 which has only 332 variants. This difference is useful to determine doublet rate as well as to assess the ability to distinguish different genotypes with our single-cell approach; (ii) BPK081 is a clonal strain, while HU3 is a not cloned and therefore potentially has a higher genomic diversity; (iii) both strains differ in (bulk) aneuploidy profiles, with BPK081 displaying aneuploidy only in chromosome 31, while HU3 presents several polysomic chromosomes, and lastly (iv), BPK081 has a known CNV in the form of an intra-chromosomal expansion in chromosome 36 in a region known as the M-locus which is absent in HU3[29], while HU3 display in bulk a CNV in chromosome 29 which is not detected in the bulk data of BPK081. These differences were used to estimate the method's ability to distinguish CNVs. Moreover, we have already generated high throughput single-cell DNA data for BPK081/0 cl8 in a previous study[21] using the Chromium single-cell CNV solution from 10X Genomics (referred here as 10X-scDNA), which is considered here as the gold standard technique. Although done on different cultures – which were originated from the same cryo-preserved sample (Figure 1A) – we could compare the single-cell karyotypes determined in the present study against those reported with the 10X-scDNA solution. In addition, here we also performed 10X-scDNA with HU3 as an additional comparison.

SPC-based scDNAseq

After validating the lysis and PTA protocols in bulk samples (see supplementary text, supplementary figures 1 and 2, and supplementary tables 1 and 2), we then tested their implementation in single-cell sequencing. For that, a 50/50 BPK081/HU3 mixture was submitted to SPC encapsulation, with trapped cells being visualized with SYTO5 staining. As expected, most SPCs appeared empty, while the few that contained parasites displayed only a single organism (supplementary figure 3A). No SPC containing multiplets were found by microscopy inspection. Noteworthy, this also confirmed that cells remained alive after encapsulation (supplementary video 1). The generated SPCs were then divided in 6 samples so different lysis conditions as well as the use of PTA as alternative WGA method could be tested. A summary of the different conditions and general statistics of each sample is provided in Table 1. After lysis and WGA, SPCs were stained again with SYTO5 to detect amplified DNA (supplementary figure 3B). Strong SYTO5 staining were seen in all samples except for sample SPC-STD4, which had no detectable signal, indicating WGA failure. Moreover, in sample SPC-STD1, although positive SPCs were entirely populated by amplified DNA, fixed parasites were still visible on some of them (supplementary Figure 3C), indicating partial lysis.

Sample	Methanol fixation	Lysis	WGA	Number of reads	Number of sequenced SPCs	Number of true cells	Number of doublets
SPC-STD1	Yes	Standard*	Standard**	33,099,195	184	147	6
SPC-STD2	No	Standard*	Standard**	55,474,658	257	205	4
SPC-STD3	No	1h incubation	Standard**	61,123,987	317	221	4
SPC-STD4	No	99 ˚C	Standard**	1,835,636	95	70	43
SPC-PTA1	No	Standard*	PTA	251,607,109	742	161	2
SPC-PTA2	No	PTA lysis	PTA	215,922,639	869	253	6

Table 1 - Summary of the different protocol modifications and results of each sample. * Standard lysis protocol of the Single-Microbe DNA Barcoding kit, i.e, 0.8 M KOH lysis for 15 minutes at 23˚C. ** Standard WGA of the Single-Microbe DNA Barcoding kit, which is based on MDA.

After confirmation of successful WGA, SPCs were pooled, barcoded and sequenced. Sequencing output was first assessed by read counts per sample (Figure 1B). Standard WGA samples showed a unimodal distribution around ~170,000 reads per cell, except sample SPC-STD4, which peaked at ~17,000 reads, consistent with its observed WGA failure. In contrast, PTA samples displayed bimodal/multimodal distributions, with peaks between 2,000–17,000 and a major peak at 600,000–900,000 reads, suggesting stronger DNA amplification.

To separate true cells from background signal, we applied a barcode rank approach. All samples exhibited a characteristic sharp drop-off in read counts, indicating a clear inflection point that distinguishes cell-associated barcodes from background (Figure 1C). The PTA samples contained the cells with highest sequencing depth, with thresholds at 656,209 reads (SPC-PTA1) and 448,715 reads (SPC-PTA2). Thresholds for the standard WGA samples were lower (109,851; 114,650; and 180,703 reads in samples SPC-STD1-3) and lowest in sample SPC-STD4 (11,413 reads), which also had the fewest SPCs flagged as true cells (70 SPCs). PTA samples had more total detectable SPCs, but a similar number of true cells (160 and 253 in SPC-PTA1 and SPC-PTA2) compared to standard WGA (135, 206, and 166 in SPC-PTA1-3 respectively) (Figure 1D and Table 1). Only true cells were retained for downstream analysis. We then assessed the presence of doublets by quantifying HU3-specific variants in each SPC, using a curated set of HU3-specific SNPs identified from bulk genomic sequencing. Classification was based on the proportion of detected SNPs matching this HU3-specific set. SPCs with >90% HU3-specific SNPs were classified as HU3, whereas SPCs with <5% were assigned to BPK081. Cells with in-between values were considered doublets . Successful WGA samples showed only 1–6 doublets, while most SPCs in sample SPC-STD4 contained a mixture of BPK081 and HU3 reads (Figure 1E and Table 1), suggesting that amplicons originated from mixed free-floating DNA. Because of the clear failure in WGA and the lack of single-cell signal, sample #SPC-STD4 was excluded from downstream analysis.

Nucleotide coverage was also higher in PTA samples (Figure 1F). On average, ~71% (SPC-PTA1) and ~67% (SPC-PTA2) of the genome was covered by ≥1 read (Figure 1G), with ~33% and ~18% being covered by ≥5 reads respectively (Figure 1H). Samples SPC-STD1-3 showed 35–39% coverage at ≥1 read and 2.4–3.5% at ≥5 reads. After equalizing each sample to 100,000 reads per cell (Figure 1I), PTA still performed best, with ~32% of the genome being mapped at least once, while in the standard WGA this value was ~27–29%. In addition, no conclusive difference was seen regarding the different lysis conditions. These results indicate that PTA achieves deeper and broader genome coverage than standard WGA.

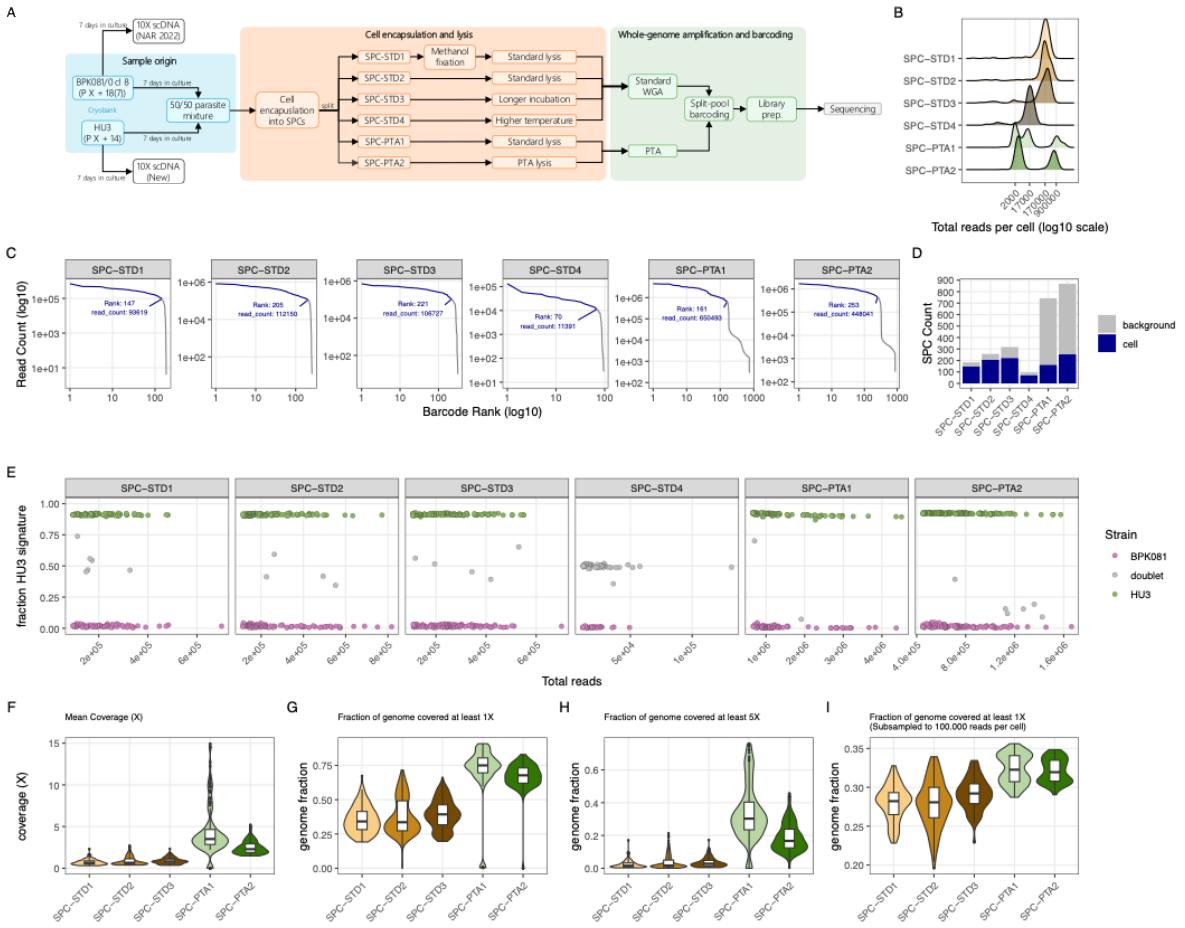

Figure 1 – Sequencing metrics and cell characterization in SPC-based scDNAseq. **A:** Schematic representing the experimental setup of the present study. Passage numbers of the two strains are indicated as P X (unknown number of passages) + passages since isolation from clinical sample (18 for BPK081, 14 for HU3), and in the case of BPK081, the number inside parenthesis indicate number of passages since cloning. Standard lysis was done with 0.8M of KOH for 15 minutes at 23 ˚C. Longer incubation was done for 1h instead of 15 minutes. Higher temperature was done at 99 ˚C instead of 23 ˚C. PTA lysis indicates the lysis protocol of the PTA kit (ResolveDNA™ – Bioskryb). **B:** Distribution of the total number of reads associated with each SPC barcode in all 6 samples. X-axis (log10 scale) show regions where some of the distributions peak. **C:** Distinction between 'true cells' and background signals. Read counts associated with each barcode were ranked in a log10 space and the 'elbow' point in the curve was set as the threshold for background signal. The labels show the rank and read count of the last SPC considered as 'true cell' **D:** Number of SPCs containing a 'true cell' (blue) or a background signal (grey) in each sample. **E:** Evaluation of HU3/BPK081 doublets (grey dots) in each sample. Y-axis show the percentage of variant reads with HU3 signature, which was predetermined based on bulk whole genome sequencing of HU3. **F-I**: analysis of nucleotide coverage depth in each sample. Coverage (F) is defined as the average number of reads mapping each nucleotide in the genome. Figures G and H show the percentage of the genome that was covered by at least 1 and 5 reads respectively, while figure I show the percentage of the genome covered by at least 1 read after cells were normalized to 100.000 reads per cell. Violin plots display the distribution of the data (kernel density), with embedded boxplots indicating the median (thick line), the interquartile range (IQR – shown by a box spanning the 25th to 75th percentile), and whiskers (lines) extending to 1.5×IQR. Points beyond the whiskers indicate outliers.

Evaluation of coverage evenness

Detecting structural variants such as CNVs and aneuploidy is a key application of scDNA-seq and requires uniform genome coverage to ensure that read counts reflect true copy number changes. To evaluate the suitability of our method, we compared coverage uniformity in our dataset with that of two libraries generated using the Chromium Single-Cell CNV solution (10X-scDNAseq), considered the gold standard for this application. After normalizing each sample to 100.000 reads per cell and correcting for GC bias, a Lorenz curve was made to evaluate global read distribution, indicating that the PTA samples had a similar profile compared to the 10X-scDNA libraries, while samples SPC-STD1-3 showed a skewer distribution (Figure 2A). This also reflected in lower gini coefficients compared to SPC-STD1-3 samples ($\tilde{x}$SPC-PTA1 = 0.245 and $\tilde{x}$SPC-PTA2 = 0.234, *versus* $\tilde{x}$SPC-STD1 = 0.321, $\tilde{x}$SPC-STD2 = 0.315, $\tilde{x}$SPC-STD3 = 0.319, $\tilde{x}$10X-BPK081 = 0.226, and $\tilde{x}$10X-HU3 = 0.217) as well as lower MAPD scores ($\tilde{x}$SPC-PTA1 = 0.508 and $\tilde{x}$SPC-PTA2 = 0.489, *versus* $\tilde{x}$SPC-STD1 = 0.634, $\tilde{x}$SPC-STD2 = 0.636, $\tilde{x}$SPC-STD3 = 0.629, $\tilde{x}$10X-BPK081 = 0.429, and $\tilde{x}$10X-HU3 = 0.401), achieving values comparable to the 10XscDNA libraries (Figure 2B-C). To estimate within-chromosome coverage dispersions, we introduced two complementary scores: (i) The intra-chromosomal coefficient of variance (ICCV) which reflects how spread the reads are in each chromosome, and (ii) the intra-chromosomal fluctuation (ICF) score, which is used to identify local intra-chromosomal biases which might skew the somy estimation, as, e.g., in cells under S-phase. The PTA samples displayed the lowest ICCV values ($\tilde{x}$SPC-PTA1 = 0.220 and $\tilde{x}$SPC-PTA2 = 0.203 *versus* $\tilde{x}$SPC-STD1 = 0.443, $\tilde{x}$SPC-STD2 = 0.428, $\tilde{x}$SPC-STD3 = 0.433, $\tilde{x}$10X-BPK081 = 0.317, and $\tilde{x}$10X-HU3 = 0.294), and similar ICF scores compared to the 10X-scDNA libraries, while samples SPC-STD1-3 had consistently the highest scores for both metrics ($\tilde{x}$SPC-PTA1 = 0.150 and $\tilde{x}$SPC-PTA2 = 0.145 *versus* $\tilde{x}$SPC-STD1 = 0.291, $\tilde{x}$SPC-STD2 = 0.276, $\tilde{x}$SPC-STD3 = 0.284, $\tilde{x}$10X-BPK081 = 0.163, and $\tilde{x}$10X-HU3 = 0.172) (Figure 2D-E). Importantly, no clear distinction was observed regarding lysis conditions. The improved uniformity in PTA was also noticeable when visualizing the normalized read coverage across all genomic bins (Figure 2F-H). In the standard WGA samples, coverage profiles were considerably noisier, with uneven coverage across cells that obscured clear aneuploidy patterns (Figure 2G). In contrast, the PTA samples exhibited much smoother profiles, with well-defined aneuploidy regions that closely resembled the high-quality profiles obtained with the 10X-scDNAseq reference (Figure 2H). Together, these analyses demonstrate that PTA improved coverage uniformity, achieving performance comparable to the 10X-scDNAseq gold standard.

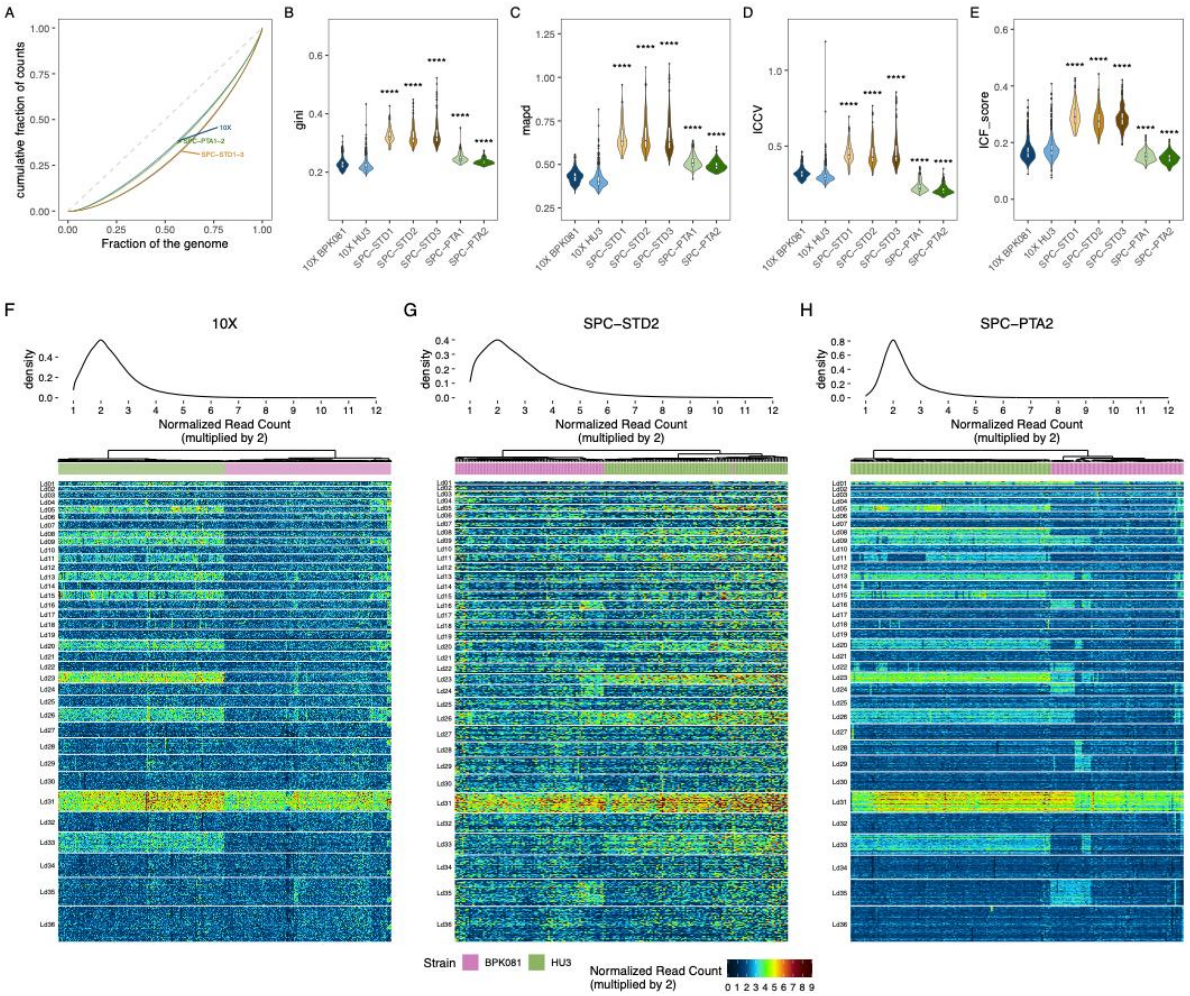

Figure 2 – Evaluation of the impact of the different WGA methods on coverage evenness. **A:** Lorenz curves showing the cumulative fraction of sequencing counts as a function of the cumulative fraction of the genome covered in the scDNAseq datasets. The dashed diagonal represents perfect uniform coverage. Curves deviating downward from the diagonal indicate increasing coverage unevenness across the genome. The blue line indicates the 10X-scDNA datasets, while the green lines show samples SPC-PTA1-2, and orange lines represent samples SPC-STD1-3. **B-E:** violin plots showing the gini (B), MAPD (C), ICCV (D) and ICF (E) scores for each dataset compared to the 10X-BPK081 and 10-HU3 libraries. Asteriscs indicate the p-values of a wilcox test of all samples against both 10X-scDNA libraries combined adjusted with the hochberg method. **** = pvalue < 0.0001. **F-H**: heatmap showing the read count of each 20k bin (y axis) in each cell (x – axis) in each chromosome (row groups). Colors indicate the read count of each bin normalized by the median read count in the cell. Annotation bars indicate the cell strain, and cells were arranged with the ward.D2 hyerarchical clustering algorithm based on the heatmap values. A density plot on top shows the distribution of values in the heatmap. Here sample SPC-STD2 is shown as a representative of standard WGA samples, while sample SPC-PTA2 represents the PTA samples.

Single-cell Karyotyping

To define the impact of the improved coverage uniformity on chromosome copy number estimation, we first checked the distribution of raw somy values obtained in each approach. In the standard WGA dataset, we observe a more continuous distribution, with a substantial number of values falling between integers (Figure 3A), likely due to the greater noise seen in this data. Conversely, PTA produces a markedly discrete distribution, with sharp peaks centered on integer values (e.g., 2, 3, and 4), and relatively few intermediate values (Figure 3B). Since chromosome copy numbers are inherently integers, PTA's output aligns more closely with biological expectations. This pattern indicates that PTA provides more accurate and reliable estimates of chromosomal copy numbers at the single-cell level. This also reflected into less cells being scaled to ploidies higher than 2, as noisy coverage seen in standard WGA distorted depth ratios and consequently led to incorrect scale factors. Cells scaled to ploidies higher than 2 in PTA could represent true polyploid cells which can spontaneously arise in culture[48], as well as cells at S/G2 phase.

Then, we compared the karyotype profiles of the SPC datasets with the 10XscDNA libraries of each respective strain. The karyotype profiles obtained from PTA samples were largely consistent with those generated by 10X-scDNA, whereas standard WGA samples showed poorer reproducibility. For the BPK081 strain, the dominant profile where only chromosome 31 was aneuploid was the most frequent in all samples, representing ~60% of cells in PTA samples, in close agreement with the ~80% reported in the 10X-scDNA dataset (Figure 3C). The highly aneuploid karyotypes 2 and 3 seen in 10X-scDNA were also detected in PTA samples 5 and 6, but not in samples SPC-STD1–3, likely due to underestimation of the chromosome 1 trisomy (Figure 3C-D). Importantly, the SPC-scDNA and 10X-scDNA datasets originated from different cultures derived from the same cryopreserved population, so some variation in karyotype content was expected, especially given the instability of the BPK081 dominant karyotype in culture[25]. Accordingly, an even more similar pattern was observed for the HU3 strain: PTA reproduced the same dominant karyotypes identified with 10X-scDNA with a remarkable similarity in proportions between them, while standard WGA produced divergent and inconsistent profiles (Figure 3E–F). Altogether, these results demonstrate that PTA-based amplification preserves karyotypic structure with high fidelity, enabling more accurate reconstruction of chromosome copy number profiles than the standard WGA.

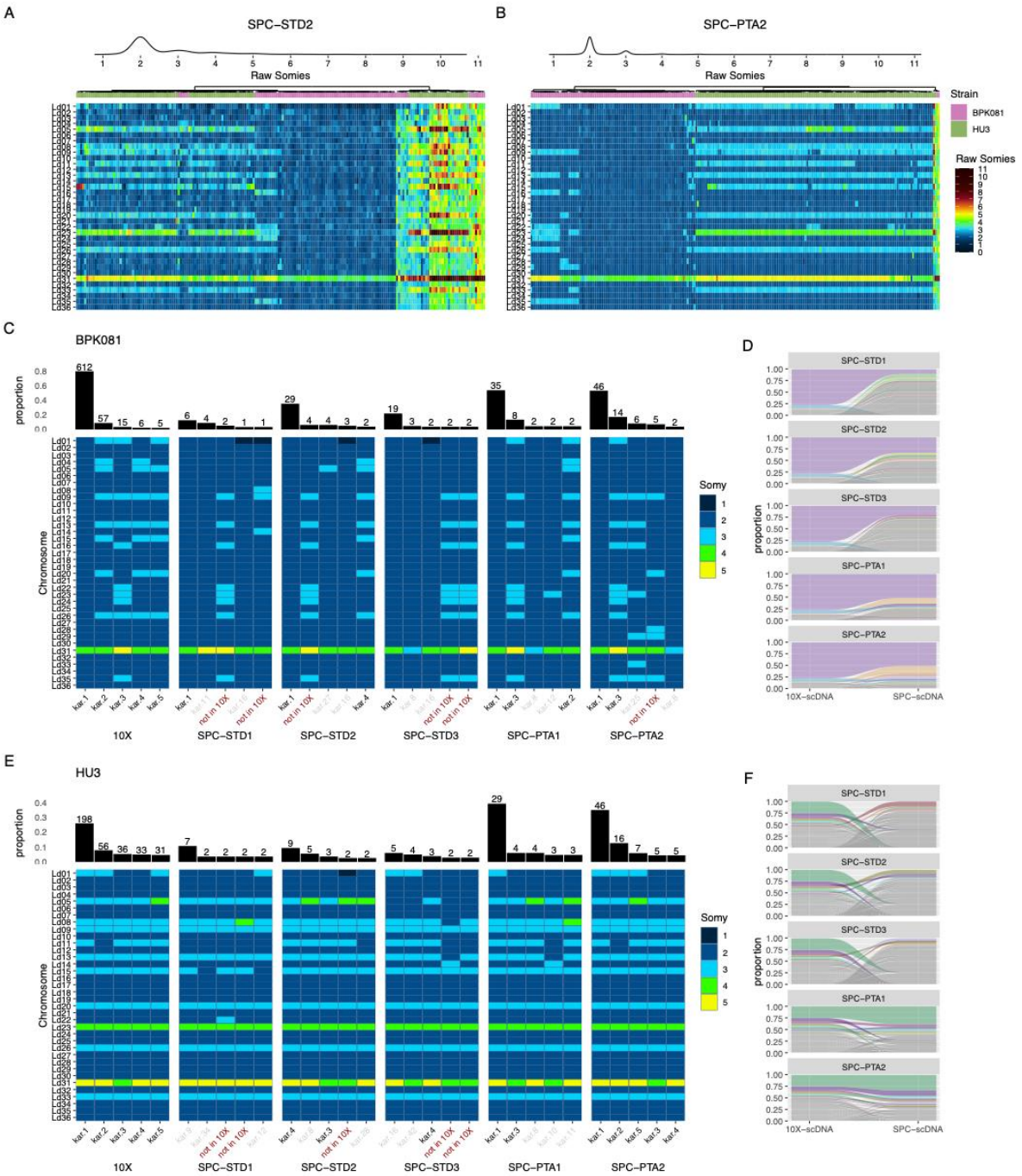

Figure 3 – Evaluation of the standard WGA PTA for determination of chromosome copy numbers in single-cells. **A-B:** Heatmaps showing the estimated raw somy values for each cell in sample SPC-STD2 (A) SPC-PTA2 (B). Raw somy values were calculated as the mean count of each chromosome (normalized by the cells mean) multiplied by the cell's scale factor. Annotation bars indicate the cell strain, and cells were arranged with the ward.D2 hyerarchical clustering algorithm based on the heatmap values. A density plot on top shows the distribution of values in the heatmap. **C:** heatmaps showing the 5 most frequent karyotypes identified the BPK081 cells of each dataset. Heatmap colors indicate the integer copy number of each chromosome (y-axis) in each of the 5 karyotypes (x-axis). Columns on top indicate the proportion of the total sequenced population displaying each karyotype. Numbers on top of the bars show the number of cells instead. **D:** Sankey diagram showing the proportion (y-axis) of each BPK081 karyotype in the SPC-scDNA samples compared to those found in the 10X-scDNA dataset. Each karyotype is represented by a line which width indicates its proportion and height indicating its position. Karyotypes which are not found in more than 1% of any dataset were not colored. **E-F**: The same as C-D but for the HU3 cells.

Detection of small structural variants

To test whether we could detect smaller structural variants, we took advantage of the existence two different CNVs which are unique to each strain – a ~8kb CNV (M-locus) in chromosome 36 unique to BPK081, and a ~4kb CNV in chromosome 29 characteristic of the HU3 strain. For this analysis, we increased the resolution to 100 bp bins, using the chromosome-normalized mean read count of the bins inside the CNV as an estimation of haploid copy numbers. In the PTA samples, all BPK081 cells displayed a clear and consistent increase in coverage depth at the M-locus compared to HU3, in contrast to the standard WGA samples where only a fraction of BPK081 cells had a detectable amplification of the M-locus (Figure 4A). Likewise, the CNV in chromosome 29 was consistently detected in all HU3 cells from PTA samples, with no BPK081 cell displaying them, while with standard WGA only a fraction of HU3 cells had a detectable amplification (Figure 4B). The distribution of estimated haploid copy numbers of both CNVs shows sharp, well-defined peaks at expected integer copy numbers for the SPC-PTA samples, with all BPK081 cells showing amplification at the M-locus and all HU3 cells showing amplification at chromosome 29 (Figure 4C). By contrast, the standard WGA samples (SPC-STD1–3) exhibited broad and noisy distributions, with many intermediate values and amplifications detected only in a subset of cells. Interestingly, in PTA samples, minor peaks at alternative integer values were observed for both loci, which could indicate the presence of subpopulations with distinct CNV states. Together, these results confirm that PTA retains better small structural information, enabling consistent detection and copy number estimation of specific small CNVs.

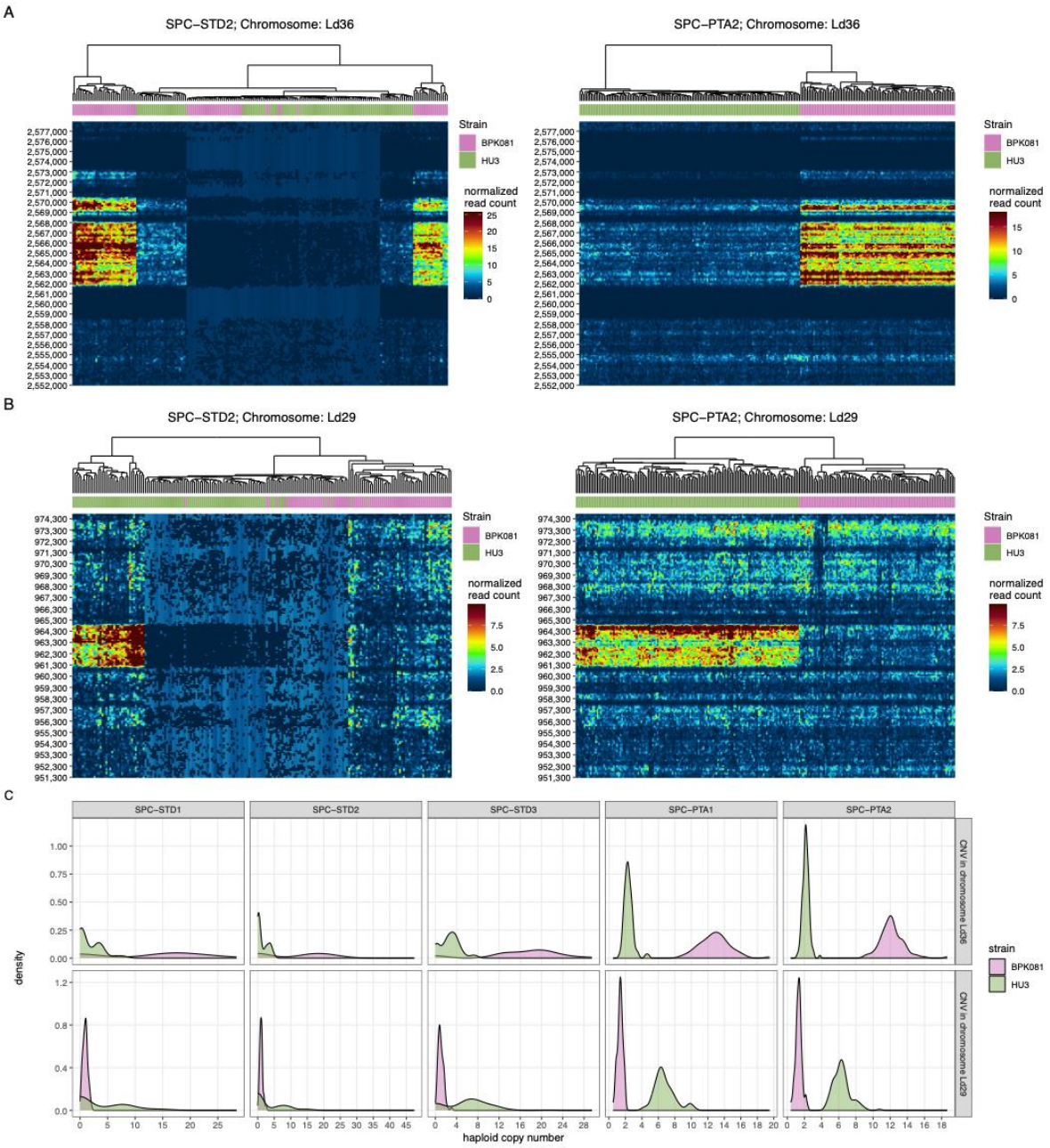

Figure 4 – Detection of strain-specific CNVs. **A:** heatmaps showing the read counts of samples SPC-STD2 (left) and SPC-PTA2 (right) across 100bp bins in the region surrounding the m-locus in chromosome 36. Heatmap colors indicate normalized read count for each bin. Annotation bars indicate the cell strain, and cells were arranged with the ward.D2 hierarchical clustering algorithm based on the heatmap values. **B:** same as A, but showing the normalized read counts for the CNV found in the chromosome 29. **C:** Distribution of the estimated haploid copy number of the m-locus (top row) and the CNV in chromosome 29 (bottom row) in each cell and in each sample. Haploid copy numbers were estimated by dividing the average count in the bins located in the m-locus by the average count of the other bins of the chromosome. Colors indicate the strain.

Nucleotide sequence variants

To investigate the applicability of our method in detecting heterogeneity at nucleotide sequence level, we first assessed its ability to distinguish the two *L. donovani* strains in a *de novo* way, i.e., without prior bulk sequencing data. This would mimic the scenario of samples with mixed infections. Principal component analysis (PCA) based on nucleotide variants showed a clear distinction between cells from HU3 and BPK081 strains in all samples (Figure 5A), indicating that both the standard WGA and PTA methods are good enough to separate parasite populations of the same species.

Next, we investigated if we could detect low-frequency nucleotide polymorphisms at particular loci of interest with single-cell resolution. We performed a global search for variants in all datasets and observed a greater proportion of confidently called sites in the PTA samples, where variants were resolved in ~80-90% of the cells on average, compared to standard WGA (~60 %) (Figure 5B).This is also noticeable when visualizing the estimated allele frequencies for each variant loci (Figure 5C). Because of the high prevalence of missing values in the standard WGA samples, we moved forward only with the PTA samples.

As proof-of-concept, we examined genes known to contribute to reduced susceptibility to leishmanicidal drugs to investigate potential presence of heteroresistance in the parasite population. We found a relative homogeneous genotype in BPK081, with few cells displaying heterozygous single nucleotide polymorphisms (SNPs) in some of the analyzed genes (Figure 5D), which is in accordance with the fact that this strain is a clone with relative few passages (7 since cloning). Conversely, HU3 cells displayed more diversity in these same loci, with small subset of cells displaying homozygous SNPs in genes related to resistance to antimony, miltefosine, and amphotericin B. This highlight the presence of spontaneous variants with potential clinical relevance which would likely have been missed by bulk sequencing.

Lastly, we used the nucleotide information to infer the short-term evolutionary relationship between cells along propagation in culture. The resulting tree revealed a compact structure with limited branch lengths and a maximum divergence of fewer than 70 nucleotide variants between cells in the BPK081 (Figure 5E). This restricted diversity indicates that the clonal strain retained a largely homogeneous genetic composition, with only minor sequence changes detectable across individual cells, in accordance with the fact that BPK081 is a clone with limited expansion time in culture. In contrast, the HU3 strain, which was never clonally derived and has been maintained for at least 15 passages (but likely many more passages), displayed higher heterogeneity (Figure 5F). The tree spanned over 30,000 nucleotide variants, with long branches and multiple subclades emerging within the population. A magnified view of the terminal region of the tree further highlights the presence of distinct subgroups differing by a few hundreds of variants, likely arising during ongoing diversification in culture.

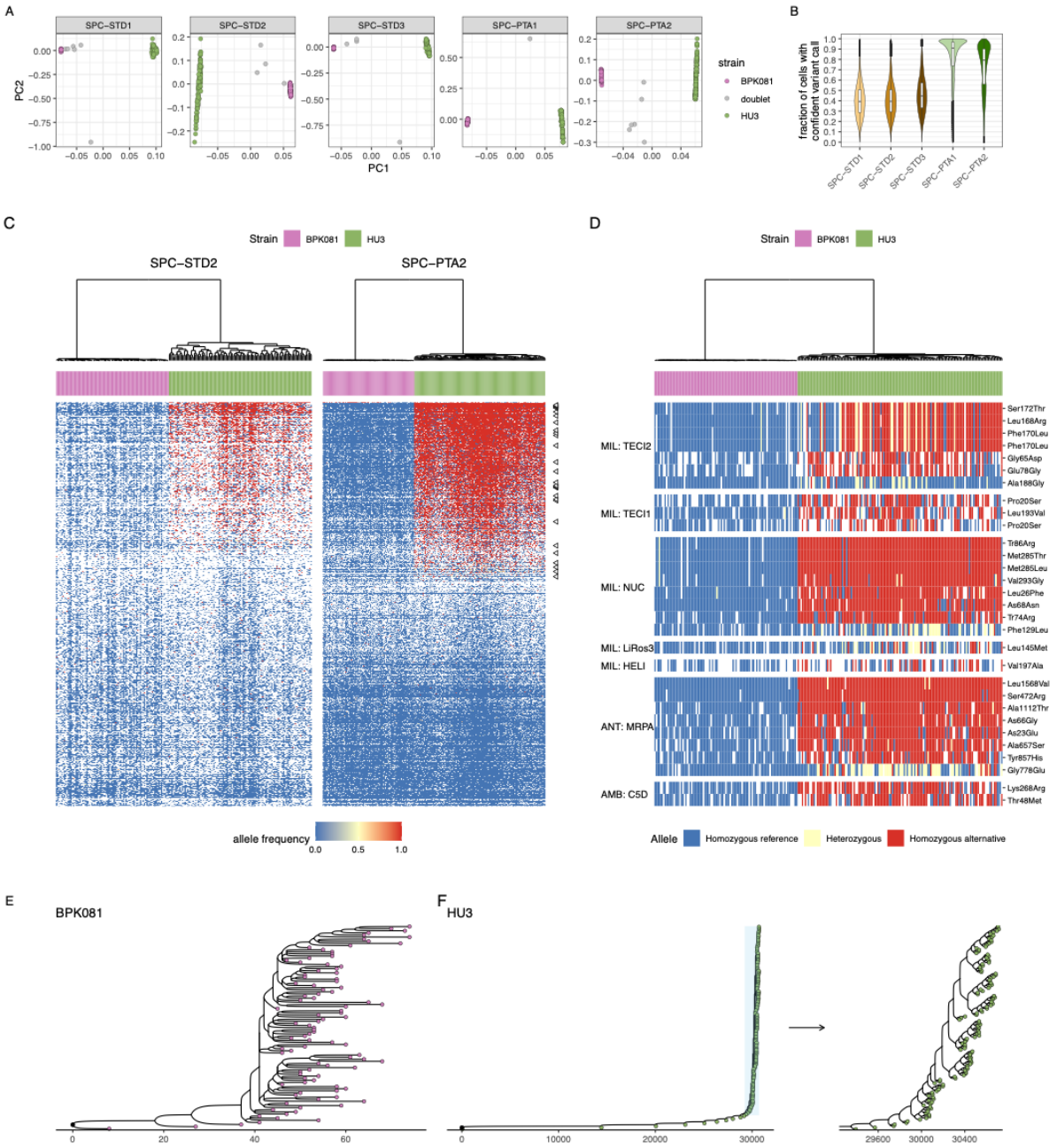

Figure 5 – Identification of nucleotide variants with the SPC-scDNA workflow. **A:** Principal component analysis (PCA) based on nucleotide variants identified in each single-cell dataset. PCA was done using PLINK2 after removal of linked variants. Colors of the dots indicate the strain. **B:** Violin plots indicating the fraction of cells in which each variant was confidently called. Included boxplots indicate the median (thick line), the interquartile range (IQR – shown by a box spanning the 25th to 75th percentile), and whiskers (lines) extending to 1.5*IQR of the data. **C:** heatmap showing the estimated allele frequency of 1000 randomly sampled variant loci in samples SPC-STD2 (left) and SPC-PTA2 (right). Color scale goes from blue (allele frequency of 0) to yellow (0.5) and red (1). White indicates missing values (NAs). Annotation bars indicate the cell strain, and cells were arranged with the ward.D2 hierarchical clustering algorithm based on the heatmap values. The black arrows to the right side of the heatmap of SPC-PTA2 show the drug resistance-associated loci. **D:** A zoom in the drug resistance-associated genes. Here only genes showing heterogeneity in nucleotide sequence are shown. The TEC1, TEC2, NUC, LiRos3, and HELI genes are associated with resistance against miltefosine (MIL), while MRPA is associated with Antimony (ANT) resistance, and C5D is related to amphotericin B (AMB). The right row labels indicate which type of mutation each variant is predicted to cause in the protein expressed by those genes. **E-F**: Phylogenetic trees built based on homozygous variants found in the dataset of SPC-PTA2. Trees were constructed with the ScisTree2 pipeline for the BPK081 strain (E) and HU3 (F). In F, an inset displays an expansion of the extreme right part of the tree. Numbers in the x-axis indicate the number of different variants between nodes.

DISCUSSION

In the present study, we developed and validated a high-throughput single-cell genomic approach using *Leishmania* as a model eukaryote. Using SPCs for high throughput cell encapsulation and PTA for WGA, we could simultaneously resolve karyotypes, detect intra-chromosomal CNVs, uncover low-frequency genotypes, and explore population structure. These findings represent a significant advance in dissecting genome instability and adaptation in eukaryotic pathogens at single-cell resolution.

Improved Whole-Genome Amplification with PTA vs Standard WGA

Leishmania display a highly segmented genome: its small size of 33 Mb is divided among 36 chromosomes[49]. This characteristic makes scDNAseq particularly challenging given that WGA methods are usually validated in mammal cells, with nuclear genomes about 100 times larger than *Leishmania*. Thus, to obtain sufficient amounts of material for sequencing, *Leishmania* DNA must be amplified orders of magnitude more compared to a mammalian counterpart, increasing the risk of artefactual biases. Our single-cell genome analysis of *Leishmania* applied primary template-directed amplification (PTA) to overcome the biases associated with other WGA methods such as MDA. PTA was developed as a modification of MDA by introducing the use of exonuclease-resistant terminator nucleotides which limit amplicon length and prevent reamplification of daughter molecules, changing the amplification kinetics from exponential to quasi-linear[34]. This dramatically reduces allelic dropout and amplification bias, thereby preserving the true allelic and structural representation of the single-cell genome[34]. Notably, this method has never been tested in organisms with small genomes such as *Leishmania*. Here, we demonstrated that *Leishmania* genomes amplified with PTA yielded consistent coverage across chromosomes and between replicates, outperforming the standard WGA – which is based on MDA – in all metrics. When benchmarking against an established 10X Genomics single-cell CNV dataset, the PTA-based workflow reached comparable coverage evenness and recapitulated expected karyotypes more in accordance with the expected distributions, whereas the standard WGA introduced major artifactual karyotypes not seen in the reference data. Altogether, our results demonstrate that PTA's uniform amplification preserves genomic structure far more faithfully, enabling more accurate reconstruction of chromosome copy number and sequence variants even from single cells with such a small genome.

SPCs as a viable tool for high throughput implementation of PTA

Although PTA offers good performance in WGA fidelity, its application has been limited to low throughput workflows, using cells isolated in plates which are manually processed for WGA in individual reactions [34,50,51]. Here we applied SPCs as an alternative solution for cell encapsulation in order to increase throughput. By applying this strategy to *Leishmania*, we could sequence a total of 399 cells (from samples SPC-PTA1-2) while using the equivalent of 17 PTA reactions (17 cells) with cells isolated in plates, reducing the cost per cell by more than 20 fold. This could be reduced even further by encapsulating more cells, as SCPs can encapsulate up to 100.000 cells (targeting a total of ~ 30.000 sequenced cells). Moreover, the split and pool barcoding strategy can also be exploited as a sample multiplexing system as was done here, by allocating cells of

different samples in specific pre-defined wells in the first barcoding round, reducing costs even further when dealing with multiple samples.

High-Resolution Karyotyping and CNV Detection

With the combination of SPCs and PTA, we achieved high-resolution single-cell karyotyping of *Leishmania* that surpasses prior methods. Each parasite's complete karyotype could be determined, and we could even resolve CNVs as subtle as a few kilobases. For example, in a mixed population of two *Leishmania* strains, our method could distinguish two known CNVs – spanning 4 kb to 8 kb in size – which were unique to each strain. Detecting such fine-scale CNVs at the single-cell level is a notable advance. Traditional single-cell CNV platforms like 10x Chromium yield copy number profiles at sparser resolution – on the order of ~2 Mb for mammalian cells, and at least ~200 kb for *Leishmania*[21] – and typically require pooling signals across multiple cells to detect smaller events[52,53]. In contrast, the high depth and uniformity of our PTA-amplified genomes enable detection of gene-level amplifications or deletions within individual cells. This could be particularly relevant in the context of genomic surveillance of drug resistance markers such as the amplification of the H-locus and M-locus, two CNVs found in some *L. donovani* strains which were linked to antimony resistance in the Indian Subcontinent[23]. Moreover, CNVs have been linked to drug resistance in other pathogenic protists, such as *Plasmodium falciparum*, where CNVs affecting the plasmepsin genes are linked to resistance to piperaquine resistance[54], highlighting the potential relevance of the method in other organisms.

Heterogeneity of nucleotide sequence

A further complication in *Leishmania* biology is the possible occurrence of mixed infections and within-host diversity. Field studies have shown that individual hosts (human or animal) can harbor multiple *Leishmania* genotypes or even different species simultaneously, with a potential impact on clinical outcome [30–32]. Mixed-genotype infections are likely under-detected by routine diagnostics[55], and confounds bulk genomic analyses. This was evident when cultured parasites showed genomic signatures distinct from their clinical counterparts, likely due to the expansion of initially undetectable genotypes in culture[56]. This complexity makes it challenging to link genotype to phenotype and to study evolutionary dynamics, especially for traits like drug resistance or virulence where minor variants can be clinically relevant.

Here we demonstrated the feasibility of distinguishing genotypically different cells by using artificial mixtures between two *L. donovani* strains. Cells were clearly separated without any prior knowledge between the genotypes composing the cell mixture. This highlights the applicability of this method to resolve genotypic diversity in parasite populations and might be applied to field samples in the future, e.g., to detect circulating low-frequency genotypes and to resolve mixed infection cases, with potential to transform molecular epidemiology. While sequencing single cells directly from host tissues may currently not be feasible, this method could readily be applied to fresh isolates (after only one or two passages) to detect circulating low-frequency genotypes and to monitor resistance-associated variants. As a proof-of-concept, we looked specifically for nucleotide polymorphisms in known drug resistance-associated genes in

our dataset. We revealed high heterogeneity in SNPs in a few subsets of genes in the HU3 strain, which constitutes so far undetected/hidden levels of genomic heterogeneity, potentially already present in the parasite infecting the human host. Such within-population variability in drug resistance markers might represent hallmarks of heteroresistance, where genetically distinct subpopulations coexist with differing levels of drug susceptibility[57,58]. Indeed, experimental selection of miltefosine resistance revealed that parasite adaptation to the drug was driven mainly by pre-existing mutations in a miltefosine transporter gene which were undetected in bulk genome sequencing[24].

Lastly, we also used the nucleotide information to infer the evolutionary relationship between cells of the two strains, showing that the clonal strain BPK081, maintained for only seven passages since cloning, exhibited minimal divergence among individual cells, while a higher diversity was observed in HU3. The ability to reconstruct phylogenetic trees at the single-cell level provides a powerful tool to study cellular evolution in both experimental and natural populations. For instance, using genetically introduced cellular barcoding, our group could track the clonal adaptation to two leishmanicidal drugs *in vitro*, revealing polyclonal origins for the genetic alterations associated with adaptation[24]. Similar information could be gathered with scDNAseq without the need for genetic engineering. More broadly, single-cell phylogenies have proven instrumental in understanding tumor evolution and drug resistance in cancer[59,60]. By resolving genetic lineages at cellular resolution, our method opens opportunities to monitor how eukaryotic cells diversify under selective pressures, whether in protozoan pathogens, model systems, or human disease.

In conclusion, our high-throughput single-cell genome sequencing method provides an unprecedented view of *Leishmania* genome variation at cellular resolution, confirming previous observations of mosaic aneuploidy as well as revealing novel insights on the extend of CNVs and nucleotide sequence diversity within parasite populations. By accurately capturing structural variations alongside sequence polymorphisms, this approach offers a promising tool to link genetic variation to phenotypic outcomes (such as drug resistance or virulence) in protozoan parasites. More broadly, the SPC-PTA workflow has the potential applicability to other pathogens and cell systems that exhibit genome instability (e.g. *Candida* spp. with aneuploidy, cancer cell subclones, etc.), suggesting a wide impact in fields ranging from microbiology to oncology. In summary, the SPC–PTA workflow fills a critical gap in our experimental arsenal to reveal the interplay between genome instability and the hidden diversity that underlies parasite adaptation, drug resistance, and survival in changing environments.

DATA AVAILABILITY

Raw sequencing data have been deposited in BioProject (NCBI) with accession number PRJNA1309081. Custom scripts used in this paper are available at https://github.com/gabrielnegreira/2025_scDNA_paper.

ACKNOWLEDGEMENTS

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

AUTHOR CONTRIBUTIONS

G.H.N.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, software, validation, visualization, writing – original draft, writing – review & editing. **P.M.**: Data curation, formal analysis, investigation, software, validation, visualization, writing – original draft, writing – review & editing. **J.C.D.**: conceptualization, funding acquisition, resources, writing – review & editing. **M.A.D.**: conceptualization, funding acquisition, methodology, project administration, resources, supervision, writing – review & editing.

FUNDING

This work was supported by the Medical Research Council [MR/Y001338/1]; the Departement Economie, Wetenschap en Innovatie (Department of Work, Economics, Science, Innovation & Social Economics of the Flemish Government – WEWIS); and the Fonds Wetenschappelijk Onderzoek (Flemish Fund for Scientific Research – FWO) [12AFI24N to G.H.N];

REFERENCES

1. Rustchenko, E. Chromosome instability in Candida albicans. *FEMS Yeast Res.* **7**, 2–11 (2007).

2. Gao, C. *et al.* Chromosome instability drives phenotypic switching to metastasis. *Proc. Natl. Acad. Sci.* **113**, 14793–14798 (2016).

3. Guérillot, R. *et al.* Unstable chromosome rearrangements in Staphylococcus aureus cause phenotype switching associated with persistent infections. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20135–20140 (2019).

4. Kakade, P., Sircaik, S., Maufrais, C., Ene, I. V. & Bennett, R. J. Aneuploidy and gene dosage regulate filamentation and host colonization by Candida albicans. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2218163120 (2023).

5. Yang, F. *et al.* Tunicamycin Potentiates Antifungal Drug Tolerance via Aneuploidy in Candida albicans. *mBio* **12**, e0227221 (2021).

6. Hong, J. & Gresham, D. Molecular Specificity, Convergence and Constraint Shape Adaptive Evolution in Nutrient-Poor Environments. *PLOS Genet.* **10**, e1004041 (2014).

7. Ben-David, U. & Amon, A. Context is everything: aneuploidy in cancer. *Nat. Rev. Genet.* **21**, 44–62 (2020).

8. Chen, X., Agustinus, A. S., Li, J., DiBona, M. & Bakhoum, S. F. Chromosomal instability as a driver of cancer progression. *Nat. Rev. Genet.* (2024) doi:10.1038/s41576-024-00761-7.

9. Korenková, V. *et al.* Comprehensive analysis of flavohemoprotein copy number variation in Giardia intestinalis: exploring links to metronidazole resistance. *Parasit. Vectors* **17**, 336 (2024).

10. Tůmová, P., Uzlíková, M., Jurczyk, T. & Nohýnková, E. Constitutive aneuploidy and genomic instability in the single-celled eukaryote Giardia intestinalis. *MicrobiologyOpen* **5**, 560–574 (2016).

11. Qidwai, T. Exploration of copy number variation in genes related to anti-malarial drug resistance in Plasmodium falciparum. *Gene* **736**, 144414 (2020).

12. Reis-Cunha, J. L. *et al.* Whole genome sequencing of Trypanosoma cruzi field isolates reveals extensive genomic variability and complex aneuploidy patterns within TcII DTU. *BMC Genomics* **19**, 816 (2018).

13. Mannaert, A., Downing, T., Imamura, H. & Dujardin, J.-C. Adaptive mechanisms in pathogens: universal aneuploidy in Leishmania. *Trends Parasitol.* **28**, 370–376 (2012).

14. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

15. Brouzes, E. *et al.* Droplet microfluidic technology for single-cell high-throughput screening. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 14195–14200 (2009).

16. 10X Genomics. Single Cell CNV. *10x Genomics* https://www.10xgenomics.com/products/single-cell-cnv.

17. Leonaviciene, G., Leonavicius, K., Meskys, R. & Mazutis, L. Multi-step processing of single cells using semi-permeable capsules. *Lab. Chip* **20**, 4052–4062 (2020).

18. Atrandi Biosciences. SPC Generation Kit (V6.3). (2024).

19. Reis-Cunha, J. L. *et al.* Ancestral aneuploidy and stable chromosomal duplication resulting in differential genome structure and gene expression control in trypanosomatid parasites. *Genome Res.* **34**, 441–453 (2024).

20. Sterkers, Y., Lachaud, L., Crobu, L., Bastien, P. & Pagès, M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major. *Cell. Microbiol.* **13**, 274–283 (2011).

21. Negreira, G. H. *et al.* High throughput single-cell genome sequencing gives insights into the generation and evolution of mosaic aneuploidy in Leishmania donovani. *Nucleic Acids Res.* **50**, 293–305 (2022).

22. 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**, 10.1128/mbio.00599-17 (2017).

23. Dumetz, F. *et al.* Molecular Preadaptation to Antimony Resistance in Leishmania donovani on the Indian Subcontinent. *mSphere* **3**, e00548-17 (2018).

24. Negreira, G. H. *et al.* The adaptive roles of aneuploidy and polyclonality in Leishmania in response to environmental stress. *EMBO Rep.* **24**, e57413 (2023).

25. Bussotti, G. *et al.* Leishmania Genome Dynamics during Environmental Adaptation Reveal Strain-Specific Differences in Gene Copy Number Variation, Karyotype Instability, and Telomeric Amplification. *mBio* **9**, e01399-18 (2018).

26. Iantorno, S. A. *et al.* Gene Expression in Leishmania Is Regulated Predominantly by Gene Dosage. *mBio* **8**, e01393-17 (2017).

27. Cuypers, B. *et al.* Four layer multi-omics reveals molecular responses to aneuploidy in Leishmania. *PLOS Pathog.* **18**, e1010848 (2022).

28. Laffitte, M.-C. N., Leprohon, P., Papadopoulou, B. & Ouellette, M. Plasticity of the Leishmania genome leading to gene copy number variations and drug resistance. *F1000Research* **5**, 2350 (2016).

29. Imamura, H. *et al.* Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent. *eLife* **5**, e12613 (2016).

30. Alves Souza, N. *et al.* Detection of mixed *Leishmania* infections in dogs from an endemic area in southeastern Brazil. *Acta Trop.* **193**, 12–17 (2019).

31. Shirian, S., Oryan, A., Hatam, G. R. & Daneshbod, Y. Mixed Mucosal Leishmaniasis Infection Caused by Leishmania tropica and Leishmania major. *J. Clin. Microbiol.* **50**, 3805–3808 (2012).

32. Antoniou, M., Doulgerakis, C., Pratlong, F., Dedet, J. P. & Tselentis, Y. Short Report: Treatment failure due to mixed infection by different strains of the parasite *Leishmania infantum*. *Am. J. Trop. Med. Hyg.* **71**, 71–72 (2004).

33. Imamura, H. *et al.* Evaluation of whole genome amplification and bioinformatic methods for the characterization of Leishmania genomes at a single cell level. *Sci. Rep.* **10**, 15043 (2020).

34. Gonzalez-Pena, V. *et al.* Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc. Natl. Acad. Sci.* **118**, e2024176118 (2021).

35. Organization, W. H. Proposals for the nomenclature of salivarian trypanosomes and for the maintenance of reference collections*. *Bull. World Health Organ.* **56**, 467–480 (1978).

36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

37. Picard toolkit. (2019).

38. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

39. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. B* **73**, 3–36 (2011).

40. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).

41. Zeileis, A. *Ineq: Measuring Inequality, Concentration, and Poverty*. (2014).

42. Theunis, K. *et al.* Single-cell genome and transcriptome sequencing without upfront whole-genome amplification reveals cell state plasticity of melanoma subclones. *Nucleic Acids Res.* **53**, gkaf173 (2025).

43. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. mixtools: An R Package for Analyzing Finite Mixture Models. *J. Stat. Softw.* **32**, 1–29 (2009).

44. Peng, Y. & Wang, S. *Ggalign: A 'ggplot2' Extension for Consistent Axis Alignment*. (2025).

45. Wu, Y. Accurate and efficient cell lineage tree inference from noisy single cell data: the maximum likelihood perfect phylogeny approach. *Bioinformatics* **36**, 742–750 (2020).

46. Zhang, H., Zhang, Y., Gao, T. & Wu, Y. Large-scale Inference of Cell Lineage Trees and Genotype Calling from Noisy Single-Cell Data Using Efficient Local Search. 2024.11.08.622704 Preprint at https://doi.org/10.1101/2024.11.08.622704 (2025).

47. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

48. Cruz, A. K., Titus, R. & Beverley, S. M. Plasticity in chromosome number and testing of essential genes in Leishmania by targeting. *Proc. Natl. Acad. Sci.* **90**, 1599–1603 (1993).

49. Späth, G. F., Piel, L. & Pescher, P. *Leishmania* genomic adaptation: more than just a 36-body problem. *Trends Parasitol.* (2025) doi:10.1016/j.pt.2025.04.002.

50. Luquette, L. J. *et al.* Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat. Genet.* **54**, 1564–1571 (2022).

51. Middelkamp, S. *et al.* Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox. *Cell Genomics* **3**, 100389 (2023).

52. Single cell copy number calling -Software -Single Cell CNV -Official 10x Genomics Support. https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/algorithms/cnv_calling.

53. Velazquez-Villarreal, E. I. *et al.* Single-cell sequencing of genomic DNA resolves sub-clonal heterogeneity in a melanoma cell line. *Commun. Biol.* **3**, 1–8 (2020).

54. Bopp, S. *et al.* Plasmepsin II–III copy number accounts for bimodal piperaquine resistance among Cambodian Plasmodium falciparum. *Nat. Commun.* **9**, 1769 (2018).

55. das Chagas, B. D. *et al.* Interspecies and Intrastrain Interplay among Leishmania spp. Parasites. *Microorganisms* **10**, 1883 (2022).

56. 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).

57. Pereira, C., Larsson, J., Hjort, K., Elf, J. & Andersson, D. I. The highly dynamic nature of bacterial heteroresistance impairs its clinical detection. *Commun. Biol.* **4**, 521 (2021).

58. Gow, N. A. R. *et al.* The importance of antimicrobial resistance in medical mycology. *Nat. Commun.* **13**, 5352 (2022).

59. Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single Cell Sequencing. *Cell* **173**, 879-893.e13 (2018).

60. Morita, K. *et al.* Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat. Commun.* **11**, 5327 (2020).

61. Engstler, M. & Beneke, T. Gene editing and scalable functional genomic screening in Leishmania species using the CRISPR/Cas9 cytosine base editor toolbox LeishBASEedit. *eLife* **12**, e85605 (2023).

62. Herrmann May, N. *et al.* Improved base editing and functional screening in Leishmania via co-expression of the AsCas12a ultra variant, a T7 RNA polymerase, and a cytosine base editor. *eLife* **13**, RP97437 (2025).