like – DTU III-like), two *T. brucei* strains (strains: TREU927 and gambiense DAL972) and one *T. congolense* strain IL3000, using three different programs: MUSCLE v3.8 [45], MAFFT v7.0 [46] and CLUSTAL W v2.0 [47]. The three resulting alignments were combined into a consensus alignment using M-Coffee v9.03 [48], which was subsequently trimmed using trimAl v1.4 [49], with a consistency score cut-off of 0.1667 and a gap score cut-off of 0.1, to remove poorly aligned regions [50].

Maximum-likelihood (ML) phylogenetic reconstruction used two different input datasets. First, alignments were evaluated using a gene by gene approach and, second, amino acid sequences for all genes were concatenated and analysed as a single sequence. The ability to reconstruct the reference phylogeny was used to rank individual reconstructions by comparing them to the reference species phylogeny. All ML trees were reconstructed using PhyML v3.0 [51]. The best substitution model was set to JTT+G [52] determined by using ProtTest 3 [53] according to the agreement between the Akaike information criterion (AIC) and Bayesian information criterion (BIC).

We performed a Likelihood Ratio Test (LRT) on the final ML trees to evaluate the null hypothesis that each locus of the concatenated dataset evolved under a molecular clock [54]. All loci in which the molecular clock was not rejected, and that had a homologue in *T. brucei*, were concatenated for these analyses. Divergence dates were estimated using BEAST v.2 [55]. Both the strict and the relaxed lognormal clock models were used to estimate divergence times for the concatenated nuclear loci datasets. All analyses were conducted without any topological constraints using the best-fit substitution model selected by ProtTest3, with four gamma categories, as well as partitioning of codons into three positions. All priors were set to default values, except for the Yule speciation process as a tree prior and the divergence estimate between *T. cruzi* and *T. brucei*, which was set to 100 million years ago (mya) under a normal distribution with 10 mya as the standard deviation [56].

RESULTS AND DISCUSSION

Sequence generation and comparative analyses

The genome sequence of the Tc231 cloned strain was generated using Illumina HiSeq 2000 NGS technology, totalling 55 031 792 paired-end reads. In-house PERL scripts were used to estimate coverage and size of the Tc231 genome sequence. Briefly, this approach calculates the coverage of each nucleotide derived from 1594 *T. cruzi* single-copy genes to estimate the depth of genome coverage [57]. Using this approach, coverage was estimated as 41.7×. To estimate the expected size of the Tc231 genome we divided the total number of nucleotides used in the assembly (2 823 893 082) by the estimated genome depth coverage, resulting in a diploid genome size of 67.7 Mb, comparable to that of the non-hybrid *T. cruzi* strain Sylvio [20].

First, two different approaches to assemble the Tc231 genome were used: *de novo* and reference-based. In the *de novo* assembly, the trimmed paired-end reads were submitted to the VelvetOptimizer.pl script resulting in an assembly with a best K-mer size of 51. After filling gaps using IMAGE and performing corrections using ICORN, the final size of the haploid genome sequence was estimated at 28.4 Mb, represented by 13 482 scaffolds. The haploid genome size obtained was smaller than previously estimated by our calculations (67.7/2=33.9 Mb), suggesting that repetitive regions may not have been resolved by the *de novo* assembly. As a second approach, a reference-based assembly was performed. Tc231 reads were initially mapped to all available *T. cruzi* genome sequences (both CL Brener Esmeraldo-like and non-Esmeraldo-like haplotypes and Sylvio) to determine the best reference genome sequence. Since the CL Brener strain is a hybrid of TcII and TcIII, the CL Brener Non-Emeraldo-like haplotype resulted in the best coverage and highest similarity to the Tc231 genome sequence, and therefore was selected as the reference genome sequence for the analysis.

Using the programs BWA-MEM, SAMtools and BCFtools, a haploid genome of ~32.3 Mb was obtained. This genome size is close to the expected size of 33.9 Mb, and covers about 90 % of the reference sequence. The reference-based assembly, however, generated a highly fragmented genome sequence with approximately 21 464 contigs and scaffolds containing large regions of 'N's', mainly due to differences between the Tc231 genome sequence and the reference genome.

To overcome the intrinsic limitations of each strategy, an alternative assembly approach combining reference-based and *de novo* assembly was used. This combined approach resulted in 13 576 contigs with the shortest sequence length at 50 % of the genome (N50) of 5300 bp, 8471 scaffolds (N50=14 202) (Table S1), and an estimated haploid genome size of 35.36 Mb, close to but larger than the calculated genome size of 33.9 Mb. This approach combines the best elements of the *de novo* assembly approach (assembles sequences specific to the Tc231 genome) and of the reference-based assembly approach (repetitive content better resolved), bypassing the inherent limitations in each assembly strategy.

To evaluate the combined assembly strategy, we compared the metrics of the final combined assembly with those obtained from the *de novo* and reference-based assemblies, as well as with the assembly metrics of other *T. cruzi* genomes available in public databases (GenBank and TriTrypDB) [1, 2]. As shown in Table 1, the combined assembly strategy resulted in an improvement of all metrics compared to the separate *de novo* and reference-based genome assembly, and similar metrics when compared to other *T. cruzi* strain assemblies using longer reads (Roche 454). Some of these genome sequences, such as CL Brener, exhibit larger genome sizes when compared to the others, probably due to the hybrid nature of the two strains.