Table 2. Overview of transferred annotation elements

Transferred annotation elements	
34 803	Elements were found on the reference
31 182	Elements were completely transferred
0	Elements were partially transferred
1266	Elements were split
3621*	Elements were not transferred
Coding sequences	
10 833	Gene models were transferred from the reference
10 592	Gene models were transferred correctly
0	Gene models were partially transferred
241*	Gene models were not transferred

*Multi-copy gene families, pseudogenes and some hypothetical proteins.

the regions shared between all *T. cruzi* reference DTUs, we identified protein sequences using BLASTX and searched for their orthologous groups in OrthoMCLDB.

A total of 6082 high-quality orthologous gene clusters were identified. After removing pseudogenes and multi-copy gene families that are phylogenetically unreliable due to increased variability, a total of 136 shared orthologous clusters were identified and subsequently used for evolutionary analysis (Fig. 4). From these 136 orthologous clusters, we identified 43 genes that were: (1) present in all genomes as single-exon genes with BLAST alignments covering >95 % of the *T. cruzi* reference sequences with an E-value <1e-30; and (2) also present in other trypanosomatids including *T. brucei* and *L. major*. These highly conserved gene sequences were used for subsequent phylogenetic analyses. All predicted orthologous groups, including paralogues and pseudo-genes, are available in the supplementary files.

Phylogenetic analysis and divergence time estimation

The predominantly clonal mode of *T. cruzi* propagation and the lack of evidence for intragenic recombination in the data, due its high degree of conservation among different isolates, permitted the use of nuclear gene sequences for reconstruction of the intraspecific phylogeny. The 43 nuclear loci analysed are randomly distributed in the genome. They are located in 18 of the 41 *T. cruzi* CL Brener-predicted *T. cruzi* chromosomes [58] (Fig. S2). We aligned the sequences for all 43 protein-encoding loci, and submitted them to an ML phylogenetic analysis. For each tree, we performed an approximate LRT to evaluate if they evolved under a molecular clock. Thirty loci had a chi-squared-based parametric branch value close to 1, supporting the hypothesis that observed outcome was likely to occur under a molecular clock model. Subsequently, the 30 loci that were evolving at a similar rate were concatenated and used to reconstruct a reference ML phylogenetic tree [50].

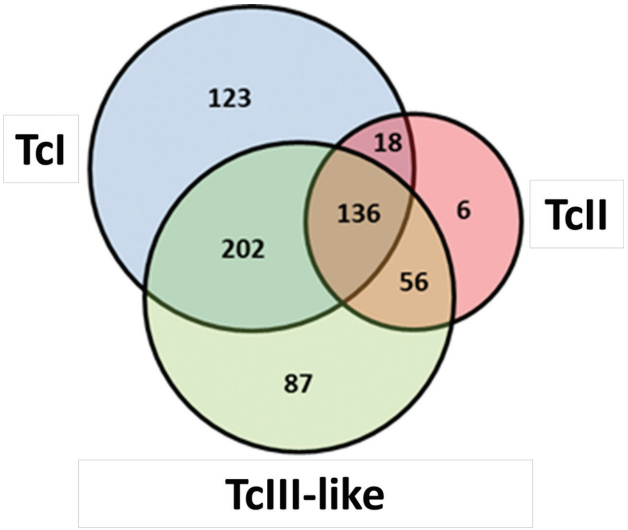

Fig. 4. Venn diagram of Tc231 sequences indicating the number of single-copy gene orthologous gene clusters that are specific or shared between TcI, TcII and Non-Esmeraldo (TcIII-like).

Analyses of the individual loci produced phylogenetic trees, the majority of which had the same topology as our reference tree (Fig. 5). This topology is consistent with a history of divergence in which *T. cruzi* II strains are in a separate clade to the other DTUs analysed [56].

Interestingly, 38 of the 43 loci are characterized by OrthoMCLDB as hypothetical proteins that are restricted to the phylum *Euglenozoa* (Table S2). Thus, these genes may be used to describe this phylum, like a barcode. As is shown in the ML nuclear tree, Fig. 5, TcIII is more closely related to TcI than to TcII. The same observation can be made looking at the Venn diagram (Figs 4 and 5).

To estimate times of divergence, the 30 loci that passed the LTR molecular clock test were used for Bayesian divergence

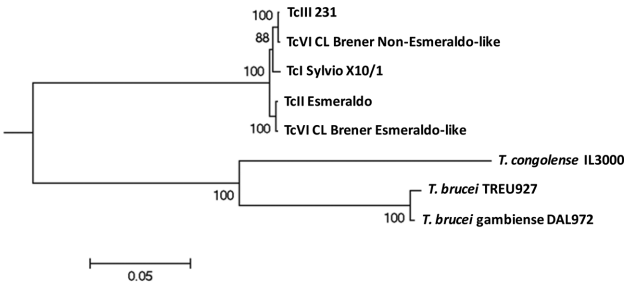

Fig. 5. ML nuclear tree obtained from 30 concatenated nuclear gene sequences. For the reconstruction, the best amino acid substitution model was JTT+G obtained by ProtTest, with 1000 bootstrap resampling used for statistical support. The scale bar shows length of branch that represents an amount genetic change of 0.05.