2016

Genomic comparisons and genome architecture of divergent Trypanosoma species

Katie Bradwell
Virginia Commonwealth University

Genomic comparisons and genome architecture of divergent *Trypanosoma* species

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Integrative Life Sciences, with a concentration in Bioinformatics and Genome Sciences, at Virginia Commonwealth University.

by

Katie Rebecca Bradwell
Professional Science Masters in Bioinformatics (Master of Bioinformatics),
Virginia Commonwealth University, 2015
Master of Science in Molecular and Cellular Biology and Genetics,
University of Valencia, 2010
Bachelor of Science in Microbiology (Hons),
University of Surrey, 2007

Director: Dr. Gregory Buck, Professor and Department Director, Center for the Study of Biological Complexity

Virginia Commonwealth University
Richmond, Virginia November 2016

Acknowledgment

The author wishes to thank several people and organizations. I would like to thank my family for their unending love and support. I would also like to thank Dr. Buck for his help and for his direction with this project. Last but not least, I would like to thank the national Science Foundation for funding and Virginia Commonwealth University for having a program and environment that allowed for such personalized development, independent thought, and innovation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DGF: dispersed gene family

DHFR: dihydrofolate reductase

DTU: discrete typing unit

GPI: glycophosphatidylinositol

KMP: kinetoplast membrane protein

KOG: clusters of orthologous groups for eukaryotes

LINE: non-long terminal repeat retrotransposon

MASP: mucin-associated surface protein

NGS: next-generation sequencing

PFGE: pulsed field gel electrophoresis

RHS: retrotransposon hot spot

RNAi: RNA interference

SCO: single copy ortholog

SNP: single nucleotide polymorphism

TS: trans-sialidase

VSG: variant surface glycoprotein

Abstract

Title of Dissertation: Genomic comparisons and genome architecture of divergent *Trypanosoma* species

By Katie R. Bradwell, Ph.D.

A dissertation submitted in partial fulfillment of the requirements for the Doctor of Philosophy in Integrative Life Sciences, with a concentration in Bioinformatics and Genome Sciences, at Virginia Commonwealth University.

Virginia Commonwealth University, 2016.

Major Director: Dr. Gregory Buck, Professor and Department Director, Center for the Study of Biological Complexity

Virulent *Trypanosoma cruzi*, and the non-pathogenic *Trypanosoma conorhini* and *Trypanosoma rangeli* are protozoan parasites with divergent lifestyles. *T. cruzi* and *T. rangeli* are endemic to Latin America, whereas *T. conorhini* is tropicopolitan. Reduviid bug vectors spread these parasites to mammalian hosts, within which *T. rangeli* and *T. conorhini* replicate extracellularly, while *T. cruzi* has intracellular stages. Firstly, this work compares the genomes of these parasites to understand their differing phenotypes. Secondly, genome architecture of *T. cruzi* is examined to address the effect of a complex hybridization history, polycistronic transcription, and genome plasticity on this organism, and study its highly repetitive nature and cryptic genome organization. Whole genome sequencing, assembly and comparison, as well as chromosome-scale genome mapping were employed.

This study presents the first comprehensive whole-genome maps of *Trypanosoma*, and the first *T. conorhini* strain ever sequenced. Original contributions

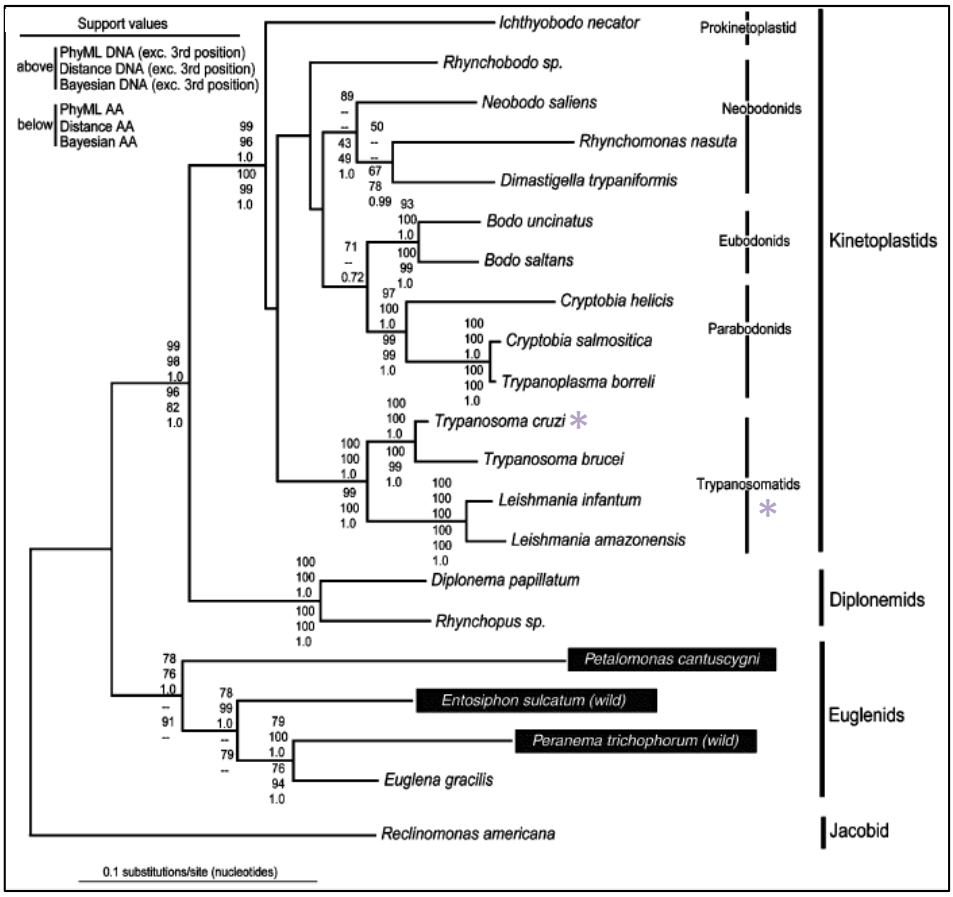
to knowledge include the ~21-25 Mbp assembled genomes of the less virulent *T. cruzi* G, *T. rangeli* AM80, and *T. conorhini* 025E, containing ~10,000 to 13,000 genes, and the ~36 Mbp genome assembly of highly virulent *T. cruzi* CL with ~24,000 genes. The *T. cruzi* strains exhibited ~74% identity to proteins of *T. rangeli* or *T. conorhini*. *T. rangeli* and *T. conorhini* displayed greater complex carbohydrate metabolic capabilities, and contained fewer retrotransposons and multigene family copies, e.g. mucins, DGF-1, and MASP, compared to *T. cruzi*. Although all four genomes appear highly syntenic, *T. rangeli* and *T. conorhini* exhibited greater karyotype conservation. *T. cruzi* genome architecture studies revealed 66 maps varying from 0.13 to 2.4 Mbp. At least 2.6% of the genome comprises highly repetitive repeat regions, and 7.4% exhibits repetitive regions barren of labels. The 66 putative chromosomes identified are likely diploid. However, 20 of these maps contained regions of up to 1.25 Mbp of homology to at least one other map, suggestive of widespread segmental duplication or an ancient hybridization event that resulted in a genome with significant redundancy.

Assembled genomes of these parasites closely reflect their phylogenetic relationships and give a greater context for understanding their divergent lifestyles. Genome mapping provides insight on the genomic evolution of these parasites.

Introduction

A Rationale for This Study

The genomic comparisons of *Trypanosoma* species and the study of *Trypanosoma cruzi* genome architecture described in this dissertation derive from the project Assembling the Tree of Life; Phylum Euglenozoa, referred to here as “ATOL”. The Euglenozoa are a group of diverse free-living and parasitic protozoa comprising the Kinetoplastids, Diplonemids, and Euglenids. ATOL was created to elucidate taxonomic and evolutionary relationships within this phylum, and resolve evolution of the lifestyle traits of its members. Of particular interest to this dissertation are the Trypanosomatids, including *T. cruzi* and closely related organisms. The position of *T. cruzi* within the phylum Euglenozoa is shown below.



Adapted from Breglia et al. *J. Eukaryot. Microbiol*., 54(1), 2007 pp. 86–92

T. cruzi is a major disease-causing parasite, which can have life-threatening consequences for the people it infects. To date, no effective and non-toxic vaccine exists against Chagas disease [1,2]. Additionally, it displays remarkable variation across strains, especially in terms of infectivity [3], karyotype [4,5], and multigene family expansion [6-8]. Some of these differences are thought to derive from a history of rare hybridization events across strains, which have produced the mixture of extant heterozygous hybrids, homozygous hybrids, and supposedly clonal strains. Sparse taxon sampling of organisms intermediate to *Trypanosoma brucei* (an African trypanosome, and causative agent of sleeping sickness) and *T. cruzi* also call for increased characterization of species such as *T. rangeli* and *T. conorhini*. The first goal of this study is thus to compare the non-pathogenic *T. conorhini* and *T. rangeli,*

and two strains of *T. cruzi* with high and low virulence. Given the highly variable karyotypes across strains, the unresolved hybridization history, and previous findings that suggest large shared segments across chromosomes [4,9], the second goal was to characterize the genome architecture of *T. cruzi*. Approaches originally envisaged for the ATOL project, such as whole genome shotgun sequencing, multigene phylogeny building, and comparative genomics, were supplemented with genome mapping technology, to achieve these two goals.

Objectives

i) Whole genome sequencing and genomic comparisons of *T. conorhini*, *T. rangeli* and *T. cruzi*

Assemble the genomes of these organisms to compare their genetic potential. This includes but is not limited to:

Karyotyping

- Define intra- and inter-species variations in chromosome size and distribution
- Explore levels of genome plasticity across species

Gene cluster analysis

- Perform OrthoMCL and KOG analysis to determine species-specific gene clustering and function enrichment and depletion

Phylogeny and percent identity

- Explore intra-strain phylogenetic relationships for *T. rangeli* and placement of *T. conorhini*, as well as overall phylogeny of the *T. cruzi* clade and outgroups

Heterozygosity

- Investigate the influence of the hybridization history of *T. cruzi* CL on heterozygosity compared to the other species

Expansion and contraction of gene families

- Define and discuss gene families and their role in shaping the life-cycles and infectivity of the species

rRNA copy number

- Reconcile the widely variable set of rRNA copy number estimate across *Trypanosoma* species
- Use bioinformatics estimates validated by qPCR (a more precise approach than previous studies)

Metabolic pathways analysis

- Explore the role of metabolic potential in defining why divergent lifestyles arose for these species

ii) Exploration of *T. cruzi* G genome architecture

- Perform genome mapping to determine overall chromosomal structure
- Elucidate relationships across chromosomes
- Define major novel structural features of the genome

Although most of the analysis in this study is carried out *in silico*, the power and utility of combining bioinformatics and wet lab approaches is demonstrated for genome size and SSU rRNA gene copy number estimates, two of the most basic and fundamental measures of each of the genomes described herein. The agreement between the *in silico* and wet lab approaches allowed for greater confidence in the methodologies used in downstream analysis, for example for multigene family copy number and genome repetitiveness estimates. Custom scripts used can be found at https://github.com/kbradwell/bioinformatics-scripts.

Literature review

The evolution of *Trypanosoma* species worldwide and their impact

Kinetoplastea is the largest class in the phylum Euglenozoa, and includes a broad range of free-living and parasitic species [10], all of which display unique features including trans-splicing, polycistronic transcription and RNA editing [11]. Trypanosomatids are the most medically important group with this class, and within Euglenozoa overall. The main genera within the trypanosomatids are *Trypanosoma* and *Leishmania*, which comprise organisms that cause African sleeping sickness (*T. brucei*), Chagas disases (*T. cruzi*) and Leishmaniasis (*Leishmania* spp.). The high number of small subunit rRNA divergences within the trypanosomatids indicate an ancient origin for the group, and it is likely that vertebrate parasitism arose multiple times in the trypanosomatids [12]. A common ancestor to *T. brucei* and *T. cruzi* likely diverged around 100 million years ago when Africa became isolated from other continents, allowing for the independent evolution and acquisition of parasitism of *T. brucei*. In terms of *T. cruzi* evolution there have been two major hypotheses: the southern supercontinent hypothesis, and the more recent bat-seeding hypothesis. The bat seeding hypothesis proposes that the most parsimonious explanation for the current geographical distribution and host range of *T. cruzi* is that its ancestor was originally a bat parasite that made switches to terrestrial mammals worldwide. One such switch gave rise to *T. cruzi* in Latin America, before the migration of humans to the New World [13].

To date, the genome sequences and annotation of *T. brucei* [14], *T. evansi* [15], *T. vivax* [16], *T. congolense* [16], *T. grayi* [17], *T. cruzi* [18] and *T.rangeli* [19] are available on publicly available databases such as GenBank and the Trypanosomatid-specific database TriTrypDB (www.tritrypdb.org). *T. brucei*, divided into the subspecies *T. b. gambiense*, *T. b. rhodesiense* and *T. b. brucei,* is an African trypanosome. *T. b. gambiense* and *T. b. rhodesiense* cause Human African Trypanosomiasis (HAT) and evade the host via switching of a variant surface glycoprotein (VSG) coat [20,21]. *T. evansi*, contended to be a subspecies of *T. brucei* [15], causes the disease surra in livestock and camels and has the widest geographical distribution of any disease-causing trypanosome. Unlike *T. brucei*, it has evolved to be independent of the tsetse fly as an obligatory vector [22] and only has a partial mitochondrial genome [23]. *T. vivax* and *T. congolense* are also closely related to *T. brucei*, although *T. vivax* displays a much more clonal population structure. Both *T. vivax* and *T. congolense* cause a huge medical and socioeconomic burden through their diseases of domestic animals [16]. The African crocodilian trypanosome, *T. grayi*, like *T. brucei*, *T. vivax* and *T. congolense* is transmitted by tsetse flies. However, it is more closely related to *T. cruzi* than it is to the African trypanosomes [17].

Finally, the species relevant to this study, *T. cruzi*, *T. rangeli* and *T. conorhini*, display very divergent lifestyles and pathogenic potentials. This study will present the first genome ever sequenced for *T. conorhini*, and the first genome sequences for selected strains of *T. cruzi* and *T. rangeli*. *T. conorhini* and *T. rangeli* are members of a phylogenetic group generally considered to be most closely related to the clade comprising *T. cruzi* and bat trypanosomes of the subgenus *Schizotrypanum*. Unlike the African trypanosomes, these organisms are spread by Reduviid bugs. *T. cruzi* is

spread by species of the *Triatoma* and *Rhodnius* genera [24], *T. rangeli* by species of the *Rhodnius* genus [25], and *T. conorhini* by *Triatoma rubrofasciata* [26]. Whereas *T. rangeli* and *T. cruzi* can be spread to a wide variety of mammals by their insect vectors, *T. conorhini* is restricted to rats, being transmitted in the feces of its vector after replication in the insect gut [27]. *T. cruzi*, like *T. conorhini*, is spread by the contaminative route, whereas *T. rangeli* is spread by inoculation of metacyclic trypomastigotes, the forms that infect vertebrate hosts, in the saliva during blood meals [28]. As discussed above, *T. cruzi* is a Latin American parasite, and its geographical distribution overlaps with *T. rangeli*, the other "American" trypanosome. Conversely, *T. conorhini* is a tropicopolitan parasite that is found worldwide, including in Latin America [26]. The three species differ in their replication modes after transmission by the vector to the mammalian host. *T. conorhini* and *T. rangeli* remain in the bloodstream during replication, whereas *T. cruzi* infects the tissues and organs, dividing inside the cells after differentiating into the amastigote form of the parasite [29]. This intracellular replication, and the adept immune evasion of the parasite, contributes to Chagas disease, which affects 6 to 7 million people according to World Health Organization estimates.

T. cruzi and *T. rangeli* as very diverse "species"

The American trypanosomes, *T. cruzi* and *T. rangeli*, have each been classified into discrete sub-populations. For *T. cruzi*, these populations are known as Discrete Typing Units (DTUs), with each DTU comprising multiple strains based on multilocus genotyping [30]. One of the major factors contributing to the diversity of the strains is thought to be the hybridization history of *T. cruzi*, which has produced some DTUs that bear homozygous hybrids and others that bear heterozygous hybrids. A theory

as to the mechanism of this natural hybridization proposes that two diploid genomes hybridize, yielding an initial polyploid heterozygous genome, which then undergoes loss of polyploidy over time due to inter-allelic recombination. The act of hybridization is thought to involve fusion of two parental nuceli, both in their diploid state, in a non-meiotic, “parasexual” encounter [31]. Homozygous hybrids, i.e. DTUs TcIV and TcIII, contain recombined alleles, bearing traits from the two parental lineages. Heterozygous hybrids, i.e. DTUs TcV and TcVI, display features of relatively recent hybridization, i.e. they are highly heterozygous, display minimal intra-lineage diversity and have accumulated very few *de novo* mutations [32]. I chose TcVI strain *T. cruzi* CL, and a supposedly non-hybrid TcI strain, *T. cruzi* G for the genomic comparisons described herein. *T. cruzi* CL is the parental strain of the clone *T. cruzi* CL Brener, which was chosen as the *T. cruzi* reference genome, and the first *T. cruzi* genome to be assembled [18]. Aside from presumed differences in heterozygosity, I expected differences in genome size and repetitive content, based on previous TcI and TcVI comparisons [7]. The two strains also exhibit an important difference in mammalian pathogenicity, since the G strain is highly susceptible to host interferon-γ [33]. *T. cruzi* G also displays a lower gene expression ratio of the virulence factor cruzipain to its negative modulator chagasin [34] compared to the CL strain. It is important to note that although *T. cruzi* G displays low virulence *in vivo*, it is able to infect cells *in vitro* [35-37], and many other strains from the TcI lineage are highly pathogenic [38].

Five lineages of *T. rangeli* have been identified; TrA, C, D and E are phylogenetically close, but TrB (which includes the AM80 strain reported herein) is a more divergent lineage positioned basal to the clade [39-41]. *T. rangeli* strains and local vectors have apparently co-evolved, with consequent lineage divergence in

concert with the insect complexes [42,43]. The TrD lineage, which includes the recently characterized *T. rangeli* SC-58 strain that was isolated from a rodent, has so far not been found in humans [44]. The *T. rangeli* AM80 strain was isolated from a human source in the Amazon, where the TrB lineage, the basal and most divergent of all known *T. rangeli* lineages, is highly prevalent [28]. Lineages TrA, prevalent from the northwestern region of South America (including Brazilian Amazonia) to Central America, and TrC spanning from the west of the Andes to Central America, were also found in humans. Lineages TrD and TrE have been rarely reported and so far only isolated from wild mammals and triatomines [43,45-48].

Emerging changes in trypanosome distribution

The ongoing rapid development of the Brazilian Amazon is likely to impact transmission of both *T. rangeli* and *T. cruzi* to humans, which are commonly co-infected with both parasites [40,41,47]. Additionally, non-endemic countries for Chagas disease, such as Spain and the United States, are experiencing an increase in cases imported from Latin America due to immigration [49,50]. *Triatoma rubrofasciata*, a vector for both *T. cruzi* and *T. conorhini*, has experienced recent increased spread throughout the tropical and subtropical world [51], raising questions about how this will affect the spread of these two parasites. These factors make the characterization of *T. conorhini*, *T. rangeli* and *T. cruzi* particularly timely.

Trypanosome comparative genomics

Because of their taxonomic positions and diverse lifestyles, these parasites present an opportunity to identify the genetic bases of their differing abilities to invade cells

and cause disease, as well as their diverse host ranges and life cycles in mammals and vectors. Comparisons of dixenous trypanosomatids to free-living bodonids have suggested that most differences lie within genes encoding metabolic and surface proteins [8]. Genome analysis of *Leishmania major* Friedlin, *T. brucei* TREU 927 and *T. cruzi* CL Brener [52], studies of lineage-specific features in *T. cruzi* Sylvio X10/1 (TcI) and *T. cruzi* CL Brener (TcVI) [7], and comparisons of *T. cruzi* and the bat-restricted *T. cruzi marinkellei* [6] suggest many differences are associated with differential multigene family expansion.

Genome architecture of trypanosomes

The extensive variation in chromosome sizes and chromosome distribution of protozoa, which is even higher than in higher eukaryotes, suggests that genome plasticity has played a major role in evolution. Trypanosomes are no exception, as even strains of the same species show extensive genomic heterogeneity [4,53-56]. However, despite this high level of structural variability, there is a high level of synteny even across genera of kinetoplastid protozoa [52,57], suggesting strong selective pressure to maintain gene order possibly associated with the polycistronic transcription process in these organisms.

T. cruzi is assumed to be mainly diploid [58,59], with some aneuploidy [53,60-62], and there is no evidence of a "euploid" state in any *T. cruzi* strain. The natural population undergoes mostly clonal evolution [63]. Sexual reproduction has not been demonstrated, although genetic exchange does occur as evidenced by the hybrid strains of distinct lineages described above. The hybrid nature of the CL Brener strain (the *T. cruzi* type strain) complicated its whole-genome shotgun

assembly [18,64]. Other *T. cruzi* strains have been assembled from whole genome shotgun sequence analysis [6,7,65]. While these other strains are not recent hybrids, they still contain a high degree of complexity and repetitiveness, which has largely prevented assembly into chromosome-sized scaffolds and thus impeded detailed study of genome architecture.

Plasticity in *T. cruzi* genomes has been observed as major chromosomal rearrangements affecting the sizes of homologous chromosomes among strains and the sizes of homologous chromosomes within a diploid pair of the same strain [5,56,59,60,66-69]. The latter appears to be a widespread phenomenon, with most strains apparently displaying a high number of size-polymorphic homologous chromosome pairs [5]. Moreover, *T. cruzi* is very permissive to copy number variation (CNV), including whole chromosome CNV. Gene-family rich chromosomal regions are prevalent in chromosomal segments displaying CNV and putative segmental duplication [70]. *T. cruzi* replicates via endodyogeny, whereby the nuclear membrane does not break down and chromosomes are poorly condensed during cell division, precluding cytogenetic analysis [71]. Karyotyping techniques on small and often similar-sized chromosomes is also challenging; a single pulsed field gel band can contain multiple heterologous chromosomes or a member of a homologous chromosome pair that differs in size. Therefore, estimations of chromosome numbers and relationships between homologous chromosomes are uncertain.

T. cruzi G, typical of other members of the TcI lineage, has a smaller genome and lower heterozygosity compared to most other DTUs [4,32]. Karyotyping on pulsed field gels with marker hybridization [4,9] has suggested that several large chromosomal segments are located in distinct pulsed field gel electrophoresis bands.

Thus, an approach that enables a greater overview of chromosomal structure has been lacking for trypanosome genomes. Given the previous karyotyping and marker hybridization of *T. cruzi* G, and its relative simplicity in terms of genome size and heterozygosity, this presents a good candidate for genome architecture studies.

Chapter I. Comparative genomics

In this work the genomes of *T. rangeli* AM80, *T. conorhini* 025E, *T. cruzi* G (TcI) and *T. cruzi* CL (TcVI) are sequenced and compared. As explained above, the lifestyles of the three species, summarized in Table 1, are highly divergent in terms of host range, geographic spread, and in the ability to infect mammalian host cells *in vivo* for intracellular replication.

Table 1: Lifestyle difference of *T. conorhini*, *T. rangeli* and *T. cruzi*

	Vectors	Hosts	Geographic location	Replication in host
T. conorhini	*Triatoma rubrofasciata*	Rat	Worldwide	Extracellular
T. rangeli	*Rhodnius* spp.	Mammals	Latin America (endemic)	Extracellular
T. cruzi	*Triatoma* spp., *Rhodnius* spp.	Mammals	Latin America (endemic)	Intracellular

Strain selection

All strains selected for genome-wide characterization and comparison originate from Brazil (Table 2). Their selection was based primarily on demographics, known host range, virulence, and relevance to the field. For example, *T. rangeli* AM80, which is from *T. rangeli* group TrB, is a human isolate from a region of Brazil under current development. It is thus likely to have an increasing impact on domestic transmission, and has the opportunity to participate in *T. cruzi* / *T. rangeli* mixed infections within the human population [28,72]. The *T. cruzi* G strain, which is less virulent than *T. cruzi* CL, was originially derived from an opossum in the Brazilian Amazon [35], and is associated with the sylvatic cycle of transmission. The CL strain was derived from

a triatomine bug in the residence of an infected person [73]. Parasites were obtained from the Trypanosomatid Culture Collection (TCC) at the University of Sao Paulo and Nobuko Yoshida (Universidade Federal de São Paulo).

Table 2: Strain information

Species	Strain	Source	ID[1]	Isolation date	Original host	Lineage	Geographical origin
T. conorhini	025E	TCC	TCC025E	1947	*Rattus rattus*	-	Brazil
T. rangeli	AM80	TCC	TCC086	1996	*Homo sapiens*	B	Brazil (Amazonas) Rio Negro
T. cruzi	G	Nobuko Yoshida[2]	TCC30	1983	*Didelphis marsupialis*	TCI	Amazonas/Brazil
T. cruzi	CL	Nobuko Yoshida[2]	TCC33	1963	*Triatoma infestans*	TCVI	Rio Grande do Sul, Brazil

[1]Trypanosomatid Culture Collection (TCC), Brazil
[2]Cultured in mice via cyclical passages

Genome assemblies

Parasites were cultured, and DNA was isolated and sequenced as shown in Figure 1. Briefly, epimastigote form parasites were cultured at 28°C in liver-infusion tryptose (LIT) medium, supplemented with 20% fetal bovine serum (FBS) with 20 ug/ml hemin for *T. rangeli* AM80, and 10% fetal bovine serum (FBS) with 10 ug/ml hemin for all other species, and harvested in log phase at ~1 X 107/ ml. Total DNA was isolated, and depleted of kinetoplast DNA (kDNA), by gel electrophoresis.

The purified DNA was used to prepare shotgun and 3 Kbp mate pair libraries (except for *T. cruzi* CL) for sequencing on the Roche 454 GS FLX+ platform as indicated by the manufacturer. Reads aligning with a minimum of 50% identity and over 50% length to kDNA from TriTrypDB were removed, and only those reads with at least 70% bases with a PHRED quality score greater than 25 and a minimum read length of 40 bp were kept using NGS QC toolkit [74] version 2.3. Assembly was performed using the Newbler version 2.9 assembler (Roche, Inc.), which limits the size of scaffolds to a minimum of 2 Kbp. Hence, all contigs larger than 500 bp that were not part of any scaffold were appended to the scaffolded assemblies for

completeness. The highly repetitive and heterozygous *T. cruzi* CL genome was left unscaffolded to minimize misassembly, and because leaving the assembly as individual contigs maximized the number of full-length gene hits to TriTrypDB (not shown). Contigs less than 500 bp in length were removed from the final assemblies.

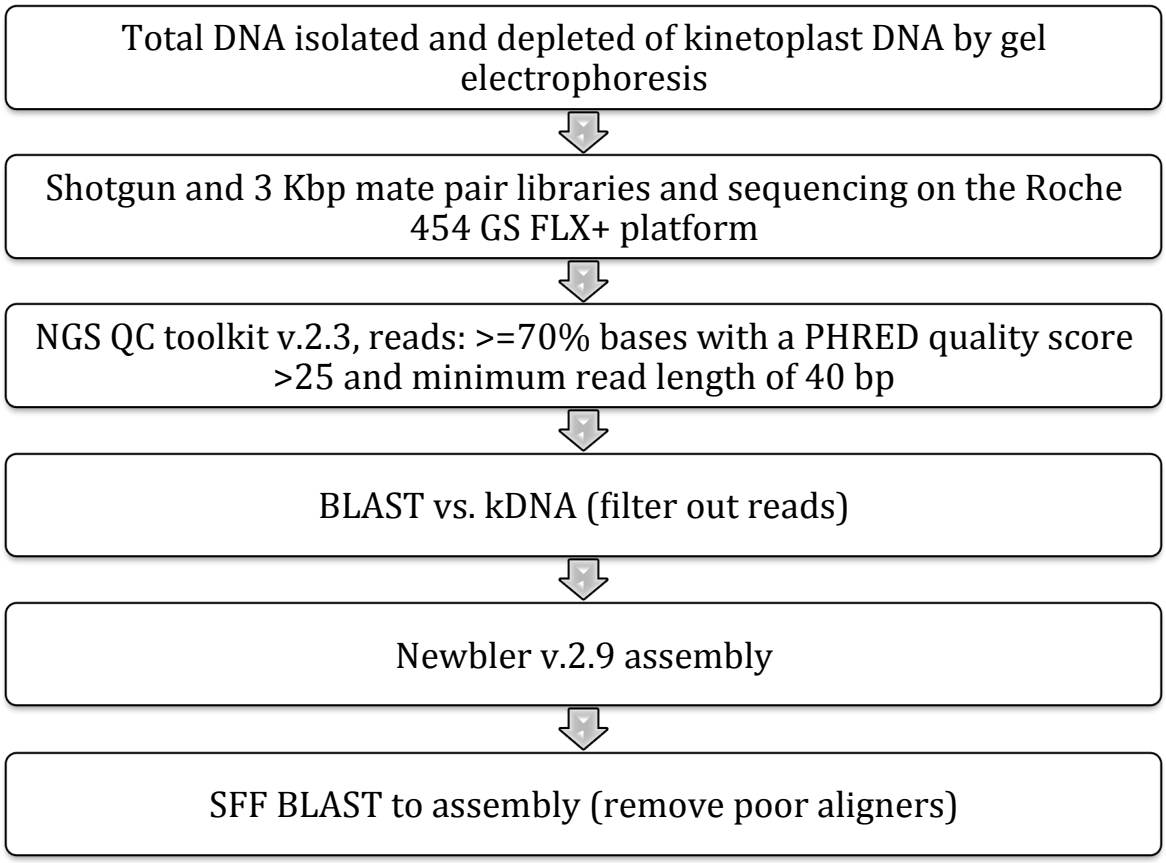


Figure 1: Genome sequencing steps. Library preparation and genome assembly for *T. conorhini*, *T. rangeli* and *T. cruzi*.

Reads were realigned to the final assemblies using BWA [75] and the average genome-wide coverage was calculated to range between 20-50X, depending on the strain (Table 3). The gigabytes of 454 Paired End (PE) data and non-PE data (SFF reads files), and number of runs to obtain the final assemblies are shown in Table 3. PE data allows for assembly scaffolding, as there is a known insert size between two sequenced gDNA fragment ends.

Table 3: Data generation and assembly coverage

	No. runs	No. PE runs	PE type	GB data (non-PE)	GB data PE	cov (X)
T. conorhini 025E	6	1	3 Kb	7.09	0.64	35.89
T. rangeli AM80	7	1	3 Kb	8.97	0.75	52.77
T. cruzi G	10	2	3 Kb	7.26	1.54	32.66
T. cruzi CL	10	6 (0 used)	3 Kb, 8 Kb	6.56	0	20.41

An in-house pipeline was used to estimate assembly completion and gene calling integrity (Figure 2). This pipeline performs two tasks: (i) randomly selects 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 99% of the reads and performs assemblies using Newbler with these read subsets; and (ii) uses tBLASTn to determine the presence of a curated set of 2,217 kinetoplastid orthologous single copy genes at 25, 50, 75, 90 and 99% alignment lengths (merging reference gene alignment lengths over multiple contigs or genes where necessary). BLASTn was used to further characterize whether these genes were complete or fragmented on contigs or gene calls, and tBLASTn was performed to predict frameshifts.

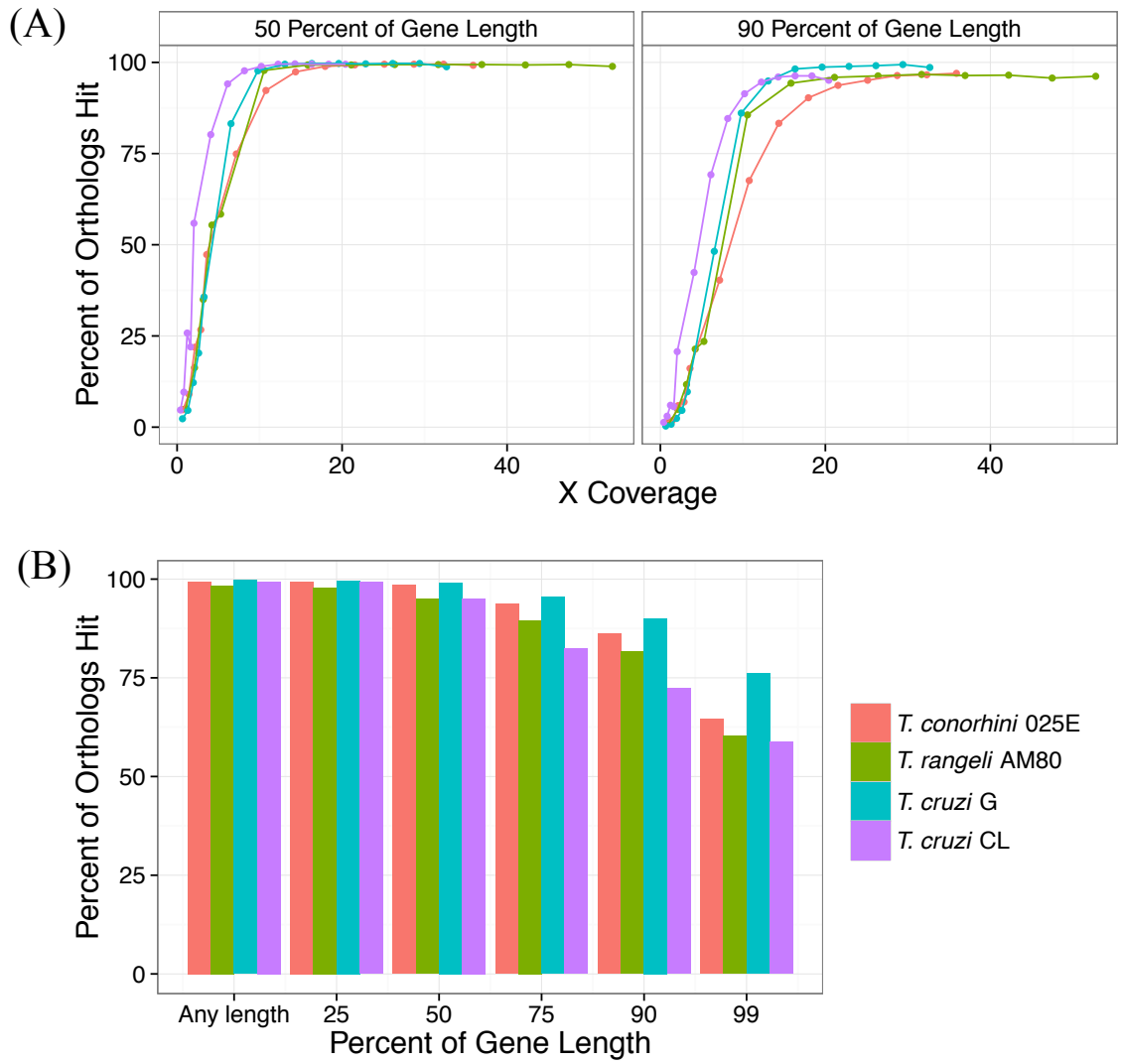

(A)
50 Percent of Gene Length
90 Percent of Gene Length
Percent of Orthologs Hit
100
75
50
25
0
0
20
40
X Coverage
(B)
Percent of Orthologs Hit
Any length
25
50
75
90
99
Percent of Gene Length
T. conorhini 025E
T. rangeli AM80
T. cruzi G
T. cruzi CL
(C)
<90% of reference length on contig
>90% of reference length on contig and
<90% of reference length in called gene
Count
Split gene call
Frameshift present
Split gene call
Percent of Orthologs Hit
15
10
5
0
(i)
(ii)
(iii)
(iv)

Figure 2: Genome completion assessment (A) Percentage of single copy ortholog genes found at 50% alignment length and 90% alignment length from assemblies performed using varying amounts of input reads data. (B) Percentages of single copy ortholog genes found and the percentage of their total length that is present in the assemblies (length is summed if the gene is found over multiple contigs). (C) i) The percentage of single copy ortholog genes that are found at <90% of their total length on a contig, ii) Percentage of single copy ortholog genes that contain split gene calls when they are are present on contigs at <90% of their length, iii) Percentage of single copy ortholog genes that contain frameshifts when they're present at >90% of their length on contigs yet <90% of their length in called genes, iv) Percentage of single copy orthologs that are present at >90% of their length on a contig yet have multiple gene calls, indicating a split gene.

Figure 2, part A demonstrates that all or nearly all of the genes from each of these organisms are represented full-length and intact in the assemblies, as above 20 X coverage most of the conserved ortholog genes are present across 90% of their length in all of the organisms. Figure 2 part B shows, using 100% of the data, how many conserved orthologs are found at various percentages of their lengths, and indicates that almost all the genes are found, albeit some may not be found in their entirety. The integrity of the *T. cruzi* CL genes was somewhat affected by gene calling errors due to frameshifts (Fig. 2, panel C (iii)). These frameshifts lead to split genes (Fig. 2, panel C (iv)) where an artificial stop codon is introduced by an indel followed by an ATG, which increases the number of genes predicted and lowers the average intergenic distance predicted.

Genome characteristics

The general characteristics of these genomes (Table 4) were determined using an in house Genome Annotation Pipeline (GAP). Briefly, genes were called using GeneMarkS v.4.7b [76]; tRNAscan-SE v.1.23 [77] was used to detect tRNAs; and 5S/18S/28S sequences were detected using RNAmmer v.1.2 [78]. SignalP v.4.1 [79]

(default settings) identified signal peptides and anchors in called genes. TMHMM v2.0 [80] (default settings) determined genes with at least one transmembrane domain. KOHGPI v.1.5 of GPI-SOM [81] was employed with the default training set and settings to predict genes with GPI anchors. BLAST [82] searches against Pfam [83], KOG [84], TriTrypDB and NCBI's nr databases were performed to determine validity and integrity of the gene calls, and ascertain probable gene functions and inferred annotations. Called genes were clustered using OrthoMCL v.2 [85] with a 1e-5 E-value cut-off for BLASTp, 50% match cut-off and 1.5 inflation value.

Table 4: Genome summaries for *T. conorhini*, *T. rangeli*, and *T. cruzi*

	T. conorhini 025E	*T. rangeli* AM80	*T. cruzi* G	*T. cruzi* CL
Genome Size (Mbp, # bases in all contigs)	21.34	21.16	25.18	34.54
Total number of contigs	1,660	1,080	1,452	20,109
Longest contig length	129,753	154,389	352,469	29,390
Shortest contig length	506	533	502	500
N50 length	24,561	43,151	74,655	2,154
Average Gene (bp)	1,403	1,346	1,277	926
Longest gene length (bp)	19,935	16,929	19,908	15,450
Shortest gene length (bp)	297	297	297	297
GC Content (%)	57.24	51.96	50.06	51.89
Coding Region (mbp)	14.25	13.61	16.23	22
Coding Region (%)	66.78	64.32	64.46	63.69
Number of Genes	10,154	10,109	12,712	23,763
Gene Clusters/Families*	9,419	9,310	11,665	19,177
Genes w/ Pfam hits	8,610	7,187	8,055	13,754
Genes w/ KOG hits	8,518	6,848	7,388	11,836
Genes w/ TriTrypDB hits	9,263	9,231	9,768	9,779
Genes w/ nr BLAST hits	9,227	9,191	9,725	9,724
Number of tRNAs	68	66	53	59
No. of rRNA features found	4	4	4	4
Genes w/ EC assignment	1,650	1,602	1,505	869
Genes w/ Signal Peptide	740	724	884	1,901
Genes w/Transmembrane Domain	1,817	1,873	2,233	3,849
Genes w/GPI anchor	1,309	1,369	1,817	4,197

*OrthoMCL prediction - orthologs, paralogs, singletons - from run using all 4 species.

The genome assembly sizes of these organisms, which range from ~21 Mbp for *T. conorhini* and *T. rangeli* to 25 – 36 Mbp for *T. cruzi* G and CL, respectively, are likely underestimates of total genome size due to the repetitive content of these genomes that collapsed into highly similar sequences in the assembly. The GC content of *T. conorhini* was slightly higher at ~57% than for the other three genomes, which ranged from ~50-52%. The number of genes in *T. conorhini* and *T. rangeli* (~10,000) is less than the number in *T. cruzi* G or CL, ~13,000 and ~24,000, respectively, and is largely consistent with that observed in other kinetoplastid protozoa [14,18,86]. *T. cruzi*, however, is recognized for having many large multigene families [7,18] likely explaining the expanded repertoire in the G and CL strains. Additionally, the CL strain is considered a hybrid [32], with a larger genome size than typically found in TcI strains [62,87], consistent with the observations of genome size and gene content described here. The genomes of each of these organisms apparently contain genes required for meiosis, suggesting likely capacity for sexual reproduction. Counts of glycosylphosphatidylinositol (GPI) anchored proteins and proteins with transmembrane domains are very similar between *T. conorhini* 025E and *T. rangeli* AM80, and highest in the *T. cruzi* strains.

OrthoMCL aims to find orthgroups, i.e. sets of genes that are descended from a single gene in the last common ancestor, which can include both paralogs (products of gene duplication) and orthologs. Clustering using OrthoMCL suggested that *T. cruzi* CL had the most gene clusters, which often corresponded to divergent members of multigene families, or sub-groups within large expanded multigene families. In every assembly the majority of orthogroup clusters only contain a single genes from that organism, i.e. clusters containing orthologs but only one gene per

organism, although the number of clusters containing two paralogs is higher in *T. cruzi* CL (Fig. 3).

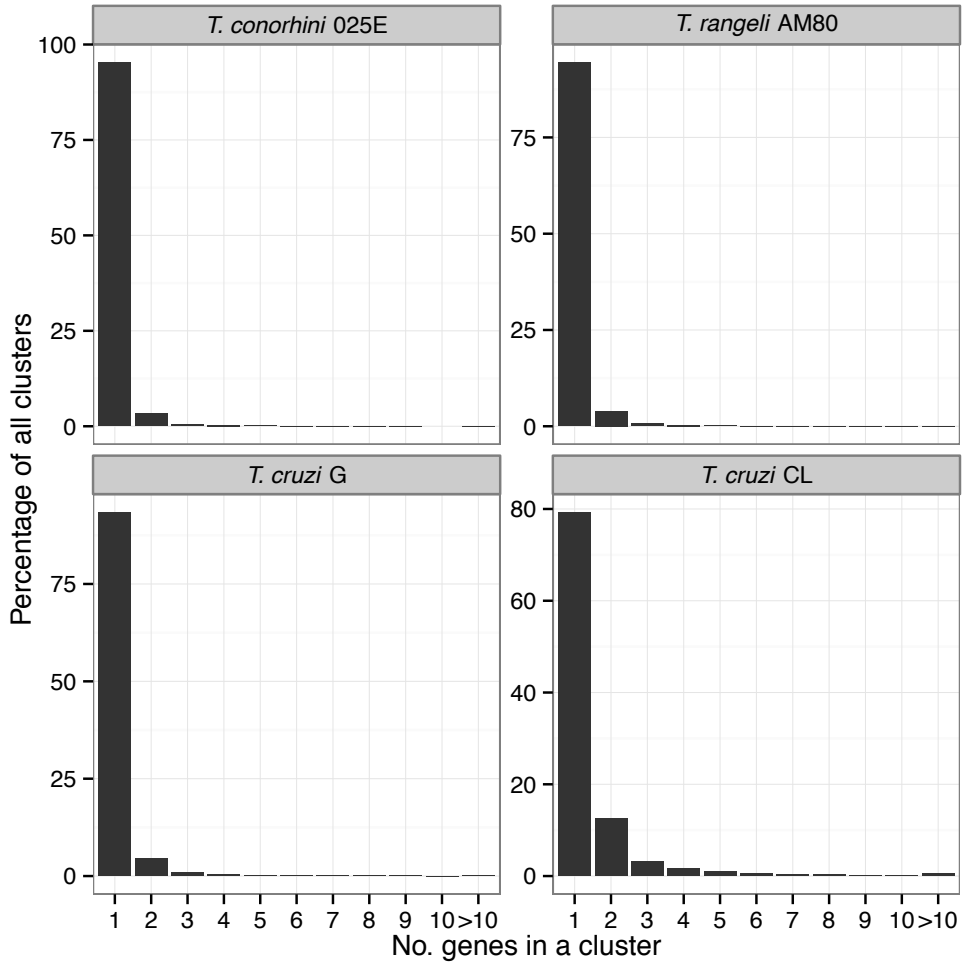


Figure 3: Number of genes per cluster based in OrthoMCL analysis. Percentage of clusters in terms of how many genes they contain. This analysis excludes counts of clusters of size 1 (genes without paralogs or orthologs).

Since *T. cruzi* CL is a likely hybrid between TcII/III lineages [32,61,88-90], this increase in gene clusters containing two genes may correspond to the two divergent haplotypes of single copy genes in the genome that have assembled separately.

Molecular karyotypes

The goals for karyotyping were to (i) define intra- and inter-species variations in chromosome size and distribution, (ii) explore levels of genome plasticity across species (iii) use densitometry to estimate chromosome numbers and genome size for validation against next generation sequencing estimates.

In addition to the four genomes described above, another strain of *T. conorhini* was obtained from the American Type Culture Collection (ATCC) for karyotype comparison (ATCC 30028). This strain was isolated from Hawaii in 1947 from *Triatoma rubrofasciata*. I also tried to use an ATCC strain deposited as a *T. conorhini* 1974 isolate from Malaysia by Weinman. However, this strain was determined by SSU rRNA analysis to group more closely to *T. cyclops*, ATCC was thus notified.

Genomic DNA isolation and pulsed field gel electrophoresis (PFGE) were performed as shown in Figure 4. 1% Megabase agarose gels (Bio-Rad) were loaded with agarose plugs bearing lysates of ~1 x 10^7 epimastigotes of each trypanosomatid strain for electrophoresis at 13.5°C using the CHEF DR III System (Bio-Rad). Run conditions used for karyotyping each species were based empirically on their individual distributions of chromosome sizes. For separation of smaller chromosome size ranges, I used the following program - Block 1: 5 V/cm, 20-200 seconds, 18 h, 120°. Block 2: 3 V/cm, 200-300 seconds, 32 h, 120°. Block 3: 1.5 V/cm 500-1,100 seconds, 12 h, 120°. The program used for separation of the largest chromosome size ranges was as follows - Block 1: 2 V/cm, 1,500 seconds, 12 h, 98°. Block 2: 2 V/cm, 1,800 seconds, 12 h, 106°. Block 3: 3 V/cm, 500 seconds, 38 h, 106°. Block 4: 5 V/cm, 20-200 seconds, 23 h, 120°. Block 5: 3 V/cm, 200-400 seconds, 34 h, 120°.

Band sizing based on standard curves of marker chromosome migration and densitometry for each gel was performed using GelAnalyzer v. 2010a [91], with rolling ball background subtraction. Pixel intensity and area under the curve of each band were plotted against the distance migrated in the gel and compared to standard curves of presumed single-copy diploid chromosomes to provide an estimate of the copy number of each chromosome band. Bands with estimated areas that were half of that expected for a chromosome pair were found in all species and were assumed to be due to size differences in chromosome pairs, as has previously been observed in *T. cruzi* [4,5,59,67,68].

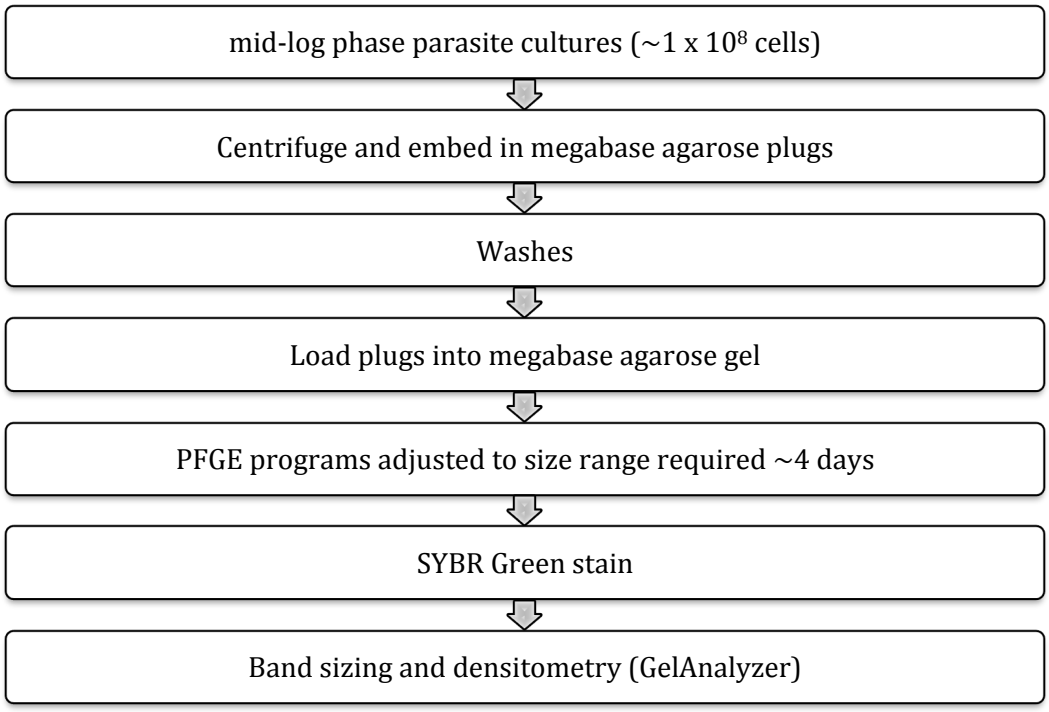


Figure 4: Pulsed field gel electrophoresis and densitometry steps. Sample preparation, PFGE and densitometry to obtain karyotype and genome size estimates.

Pulsed Field Gel Electrophoresis (PFGE) under multiple conditions provided an estimate of the sizes and numbers of chromosomes in the genomes of each of these

organisms (Fig. 5, Table 5). These analyses suggest that the *T. cruzi* CL strain bears an estimated ~37 chromosomes, similar to that previously estimated for *T. cruzi* CL Brener [54]. *T. cruzi* G, *T. rangeli* AM80, *T. conorhini* 30028, and *T. conorhini* 025E displayed 36.5, 40, 39.5 and 45 chromosomes, respectively. The chromosome numbers of the G and CL strains are quite conserved, but the sizes of the individual chromosomes are not, following a trend previously predicted for *T. cruzi* strains [54]. Genome sizes determined as described are estimates, but closely match estimates from Next Generation Sequencing (NGS) predictions. In two of the six PFGE runs of *T. rangeli* AM80 PFGE (not shown), I observed a faint band of ~5 Mbp that I excluded from the final the karyotype and chromosome or genome size estimates, although the presence of a chromosome of this size cannot be conclusively discounted. Previous studies have revealed significant variation in PFGE patterns specific to distinct lineages of *T. rangeli* [41,92].

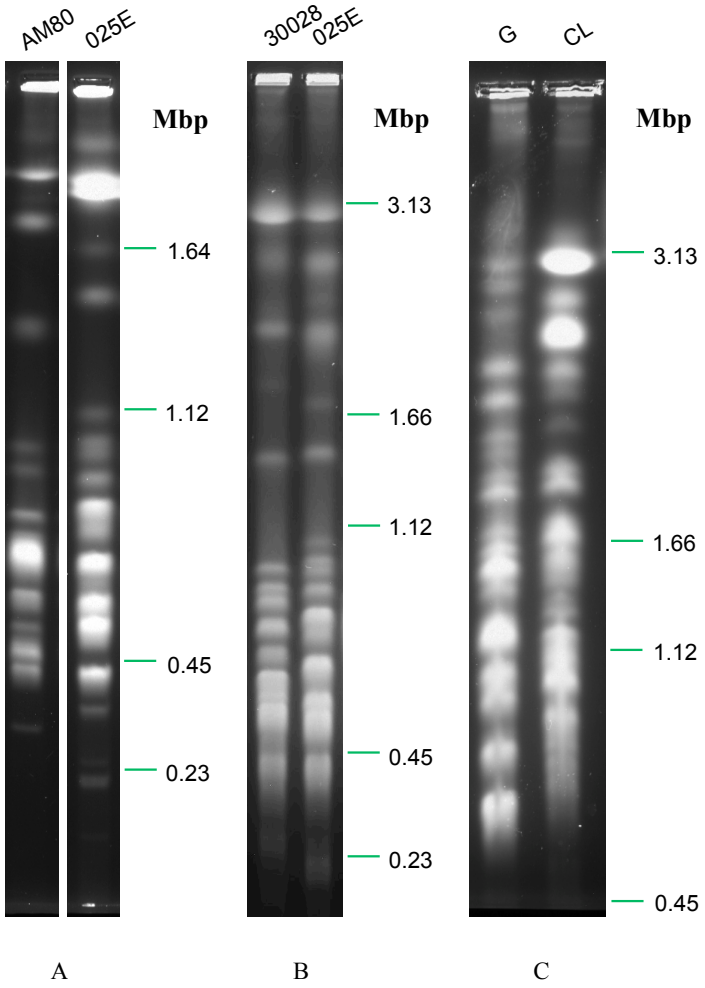


Figure 5: Karyotypes of *T. conorhini*, *T. rangeli* and *T. cruzi*. (A) *T. rangeli* AM80 vs. *T. conorhini* 025E using *Saccharomyces cerevisiae* chromosome size-markers (Bio-Rad). (B) *T. conorhini* 30028 vs. *T. conorhini* 025E. (C) *T. cruzi* G vs. *T. cruzi* CL. *Schizosaccharomyces pombe*, *Hansenula wingei* and *Saccharomyces cerevisiae* chromosomes (Bio-Rad) were used as markers for (B) and (C).

As expected, significant genome variability was observed among these karyotypes, although the two *T. conorhini* isolates show similar banding patterns. Clearly, major chromosomal rearrangements, expansions, or deletions seem to have occurred during the evolution of these parasites. Interestingly, the two *T. cruzi* strains appear to have at least double the number of megabase-sized chromosomes as *T. conorhini* or *T. rangeli*. Repeat expansions in *T. cruzi* have long been used to describe karyotype polymorphism across strains [4,5,62].

Table 5: Karyotyping and densitometry summary, and NGS genome size estimates

	T. conorhini 025E	*T. conorhini* 30028	*T. rangeli* AM80	*T. cruzi* G	*T. cruzi* CL
Total number of bands	21	20	19	18	21
Total number of predicted chromosomes	45	39.5	40	36.5	37
Number of megabase chromosomes (>1Mbp)	10	8	7	19.5	30
Number of chromosomes (300Kbp-1Mbp)	33.5	31.5	33	17	7
Number of chromosomes (<300Kbp)	1.5	0	0	0	0
Overall size range of chromosomes (Mbp)	3.22-0.21 (3.01)	3.24-0.33 (2.91)	3.24-0.31 (2.92)	3.09-0.66 (2.43)	3.48-0.74 (2.75)
Total predicted genome size (Mbp)[1]	39.703	38.423	34.847	44.0065	61.4795
NGS estimate for genome size (Mbp)[2]	41.04	n/a	30.33	48.75	64.33

[1] Sum of the total predicted number of Mbp at each band

[2] (Total number of bases) / (modal alignment depth of the assembly)

SSU rRNA copy number

Ribosomal gene copy number in Trypanosomatids presumably impacts ribosome synthesis, and therefore potentially modulates translation levels, although this has not been studied. A single repeat unit consists of 5' external transcribed spacer – 18S rRNA – internal transcribed spacer-1 (ITS-1) – 5.8S rRNA – ITS-2 and 28S rRNA. These repeat units have been found to be in 114 copies in *T. cruzi* via saturation hybridization studies, organized in tandem repeats across multiple chromosomes [93]. This contrasts to the predicted 56 rRNA copies in *T. brucei* [14], and 24 rRNA copies in *L. major* [94]. None of these estimates used more precise qPCR or bioinformatics reads mapping techniques, and copy numbers of SSU rRNA genes in each of these organisms genome assemblies is far lower than the estimates above: 6 and 9 full length copies for *L. major* Friedlin and *T. brucei* TREU927 respectively, and 12 partial copies for *T. cruzi* CL Brener in TriTrypDB v.29.

By determining SSU rRNA copy number via bioinformatics estimates, and validating these with qPCR measurements, this study aimed to (i) investigate whether the previous high rRNA copy number estimate for *T. cruzi* may be due to experimental technique or design (ii) compare the SSU rRNA copy numbers of *T. rangeli* and *T. conorhini* to *T. cruzi*.

Estimates of SSU rRNA copy number based on reads mapping were derived using the Equation 1, with gene coordinates predicted by RNAmmer v.1.2. Aligned bases were calculated at positions of q-score over 25 with a minimum two-fold coverage. Average per base coverage was calculated with a custom Python script and SAMtools v1.2, and divided by average coverage of a set of 6,479 single copy orthologs (SCOs). These were also checked using an alternative formula for copy number estimation based on mapped reads, which has been described previously [95]. Estimated copy numbers by the two methods were in agreement.

$$\frac{Average\ coverage\ of\ gene \times number\ of\ gene\ copies\ in\ assembly}{average\ coverage\ of\ SCOs} \qquad \textbf{Eqn 1}$$

qPCR and the relative threshold algorithm on ViiA 7 (Applied Biosystems, by Life Technologies) were employed using using the TaqMan® Universal PCR Master Mix Reagents Kit (P/N: 4304437). The probes and primers were designed using Primer Express® v.3.0 (Table 6). The probes were labelled in the 5' end with FAM (6-carboxyfluoresceine) and in the 3' end NFQ-MGB, and were specific for conserved regions of the 18S gene in each species where the highest number of reads mapped. Dihydrofolate reductase (DHFR) was used as a single copy reference gene for normalization. Estimations were taken at two different dilutions of gDNA sample, each in triplicate, with three biological replicates performed in separate runs. Wells

with no DNA served as no template controls, and standard curves indicated equivalency of primer/probe set efficiencies. The cycling conditions were: 50°C/2min; 95°C/10min; and 40 cycles of 95°C/15sec and 60°C/1min. The cycle threshold was determined to provide the optimal standard curve values (0.98 to 1.0).

Table 6: Primers and MGB probe sequences used for SSU rRNA copy number estimation

Organism	Gene	Forward primer	Probe	Reverse Primer
T. conorhini 025E	18S	TGCCTCAGAATCACTGCATTG	ATCTGCGCATGGCT	AGATTTTTGCGGCAGATTACG
	18S	CGCGCAACGAGGAATGTC	CGTAGGCGCAGCTC	GGACGTAATCGGCACAGTTTG
	DHFR	CGAGGCACACGTTTGAGTTG	CTGGCACGGCTCG	TGCCCACGTGTACAGTAATCACA
T. rangeli AM80	18S	GCGAAGGCATTCTTCAAGGA	ACCTTCCTCAATCAAG	TTCGATCCCCACACTTTGGT
	DHFR	CGTCCAAGAGGTTTTGTGTTGTT	ATTTGGACGACAGGACG	GCCTGGCCGCCTGAAC
T. cruzi G	18S	AAGCACTCGACTGTCCGATCA	TATTGCACCATCATCG	ACCGCCCGTCGTTGTTT
	18S	GAACTTTCGGTCAAGTGAAGCA	TCGACTGTCCGATCAC	TTCCGATGATGGTGCAATACA
	DHFR	CAATCTCCGAGGAGCTCACTTC	CAAACGGCAACGAGAC	CGTGGGATGAGTTTCTCAAAGTAG
T. cruzi CL	18S	AAGCACTCGACTGTCCGATCA	TATTGCACCATCATCG	ACCGCCCGTCGTTGTTT
	18S	GAACTTTCGGTCAAGTGAAGCA	TCGACTGTCCGATCAC	TTCCGATGATGGTGCAATACA
	DHFR	CAATCTCCGAGGAGCTCACGTC	CAAACGGCAACGAGAC	CGTGGGATGAGCTTCTCAAAGTAG

18S rRNA copy numbers show excellent agreement between bioinformatics estimates and measurements by quantitative PCR (Table 7).

Table 7: SSU rRNA copy number estimates

	T. conorhini 025E	*T. rangeli* AM80	*T. cruzi* G	*T. cruzi* CL
Estimated by read mapping	10	33	4	1
95% CI estimated by qPCR	13±0.5	28±5.9	7±1.5	3±0.8

I estimate that there are 4-7 copies in *T. cruzi* G and 1-3 copies in *T. cruzi* CL. The *T. cruzi* CL Brener genome assembly contains 12 fragments of 18S rRNA genes in

TriTrypDB [96] v.28, all less than half the size of the predicted 18S genes from the CL and G strains (~2,300 bp). My predictions are much lower than the previous estimate of 114 copies in *T. cruzi* [93], suggesting that experimental method may have played an important role in the observed differences. The SSU rRNA copy numbers are slightly higher in *T. rangeli* and *T. conorhini*, which have 28-33 and 10-13 copies respectively.

Synteny

As described above, karyotypes are very variable across strains of *T. cruzi*, and between *T. cruzi* and the two non-pathogenic parasites. There was however more conservation in chromosome sizes and distributions across *T. rangeli* and *T. conorhini*. Given these observations I aimed to determine whether gene order was conserved across all the species, despite karyotype differences. It has previously been reported that gene order is highly conserved across broadly divergent kinetoplastid protozoa [52,97].

A region of ~40 Kbp from the *T. cruzi* G genome was compared with homologous contigs from *T. conorhini* 025E, *T. rangeli* AM80 and *T. cruzi* CL. Twelve single copy orthologs from each organism were examined for presence, location and orientation on these genome segments. The genes were found in the same order and orientation in each of the organisms examined, although for *T. cruzi* CL two of the genes were found on separate contigs due to incomplete assembly of the syntenic region. Figure 6, generated using Circos [98], displays the locations of the twelve orthologous genes.

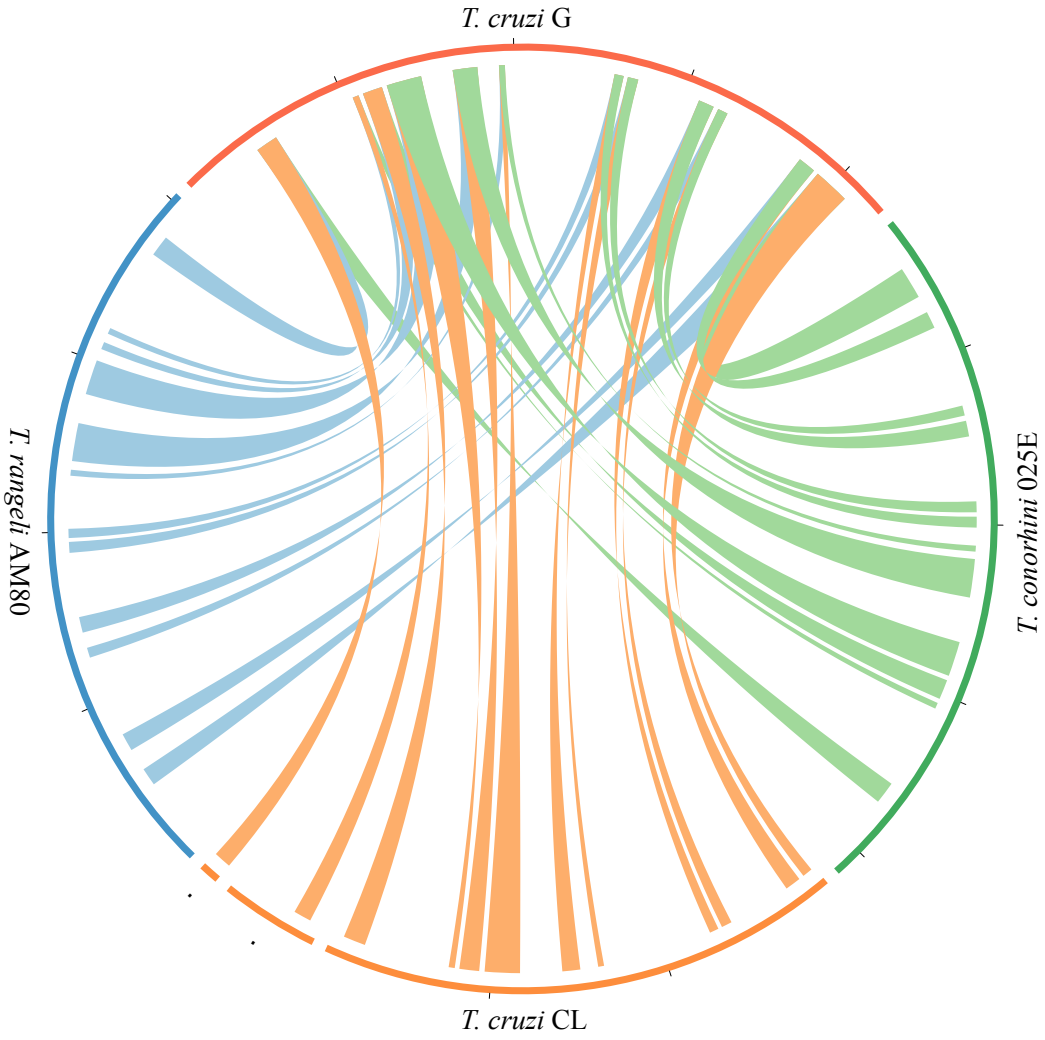


Figure 6: Synteny across twelve single copy orthologs. Gene order of orthologs on contigs of all four organisms.

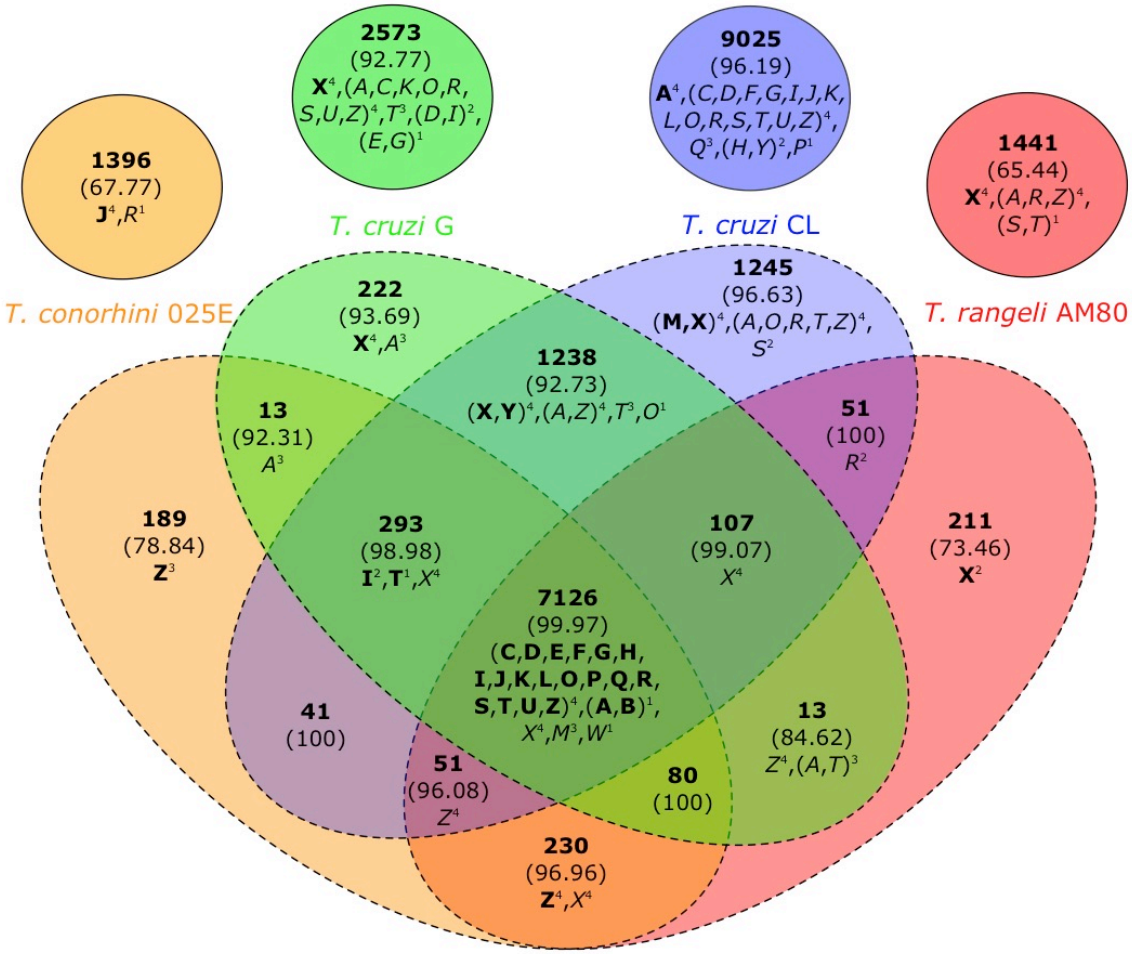
In this analysis, the genomes of *T. rangeli* and *T. conorhini* show a very high level of synteny to the genomes of *T. cruzi*. Figure 6 displays the gene order of twelve single copy orthologs in ~40 Kbp assembly contigs from each of the four genomes. The results show complete synteny between these genes in each of these organisms.

The genes on each of the contigs are on the same strand indicating a directional gene cluster and likely a single polycistronic transcription unit. Thus, although it is clear that intact chromosome architecture is not preserved across these species (c.f., pulsed field gels in Fig. 5), there is strong selection pressure, likely expressed at the level of polycistronic transcription units, for maintenance of gene order and orientation in these organisms.

Gene cluster diversity

To investigate gene cluster diversity the aims were (i) perform OrthoMCL and KOG analysis (ii) determine species-specific gene clustering (iii) explore functional enrichment and depletion with Fisher's Exact test.

Called genes from each species were clustered using OrthoMCL v.2 with a 1e-5 E-value cut-off for BLASTp, 50% match cut-off and 1.5 inflation value. This yielded a total of 11,110 clusters. Examining the species represented in each gene cluster (Fig. 7) revealed 7,126 gene clusters common to all four of the genomes. In contrast, there were 13,065 clusters unique to either the G or CL strains of *T. cruzi*, and 14,303 gene clusters unique to *T. cruzi*. *T. cruzi* CL alone exhibited 10,270 unique clusters, probably due to its hybrid genome. *T. rangeli* AM80 shares a total of 7,428 clusters with the two *T. cruzi* strains, and *T. conorhini* 025E shares 7,604. These results are consistent with previous observations that *T. cruzi* strains have undergone a significant level of gene amplification and divergence in contrast to other trypanosomatids [52].



A: RNA processing and modification
B: Chromatin structure and dynamics
C: Energy production and conversion
D: Cell cycle control, cell division, chromosome partitioning
E: Amino acid transport and metabolism
F: Nucleotide transport and metabolism
G: Carbohydrate transport and metabolism
H: Coenzyme transport and metabolism
I: Lipid transport and metabolism
J: Translation, ribosomal structure and biogenesis
K: Transcription
L: Replication, recombination and repair
M: Cell wall/membrane/envelope biogenesis
N: Cell motility
O: Posttranslational modification, protein turnover, chaperones
P: Inorganic ion transport and metabolism
Q: Secondary metabolites biosynthesis, transport and catabolism
R: General function prediction only
S: Function unknown
T: Signal transduction mechanisms
U: Intracellular trafficking, secretion, and vesicular transport
V: Defense mechanisms
W: Extracellular structures
Y: Nuclear structure
Z: Cytoskeleton

Custom function:
X: No significant KOG hit (E-value >1e-5)

Figure 7: Sequence diversity and functional enrichment across clustered genes. Genes clustered by OrthoMCL are shown here using draw.quad.venn from the VenDiagram package of Rstudio v.3.0.2, together with percent hits to TriTrypDB

v.24, KOG or PFAM databases in parentheses. Genes with the highest E-value KOG hit were chosen as representatives of each gene cluster for KOG enrichment and depletion analysis using the fisher.test of Rstudio with Bonferroni correction of P-values by p.adjust, with any clusters that lacked genes with E-value <1e-5 being described as having no significant KOG hit (category "X"). Enriched and depleted categories from Fisher's Exact test are indicated by bold and italic letters respectively. Circles with solid lines indicate singleton genes (no paralogs or orthologs, i.e. classified as a 'cluster of one' for this analysis), ovals with dashed lines represent clusters of genes that have orthologs and paralogs. p-value significance levels are shown as superscripts (1: 0.01-0.05, 2: 0.001-0.01, 3: 0.0001-0.001, 4: <0.0001).

Genes for each combination of organisms were examined for enrichment and depletion in functional KOG categories using Fisher's Exact test (Fig. 7). Fisher's Exact is a statistical test that is often used to determine whether genes/gene clusters in a specific experimental category are enriched for a specific functional category compared to counts of functional categories expected at random. In this case the experimental categories refer to gene clusters where specific combinations of species have orthologs or paralogs, e.g. *T. rangeli* specific gene clusters, or *T. cruzi* G/*T. conorhini* 025E specific gene clusters (clusters containing at least one gene from both *T. cruzi* G and *T. conorhini* 025E). Many KOG categories are enriched in the core set of 7,126 shared genes, and these shared genes, not surprisingly, include many housekeeping genes. However, notable absences included cell wall/membrane/envelope biogenesis and nuclear structure.

Interestingly, *T. cruzi* CL species-specific gene sets are the only sections enriched for cell wall/membrane/envelope biogenesis KOG functions. These KOG hits correspond mainly to the mucin family member TcMUCII and surface protease gp63. It can be hypothesized that this richer diversity in genes encoding cell surface molecules may be important for host immune evasion and therefore may play a role in the increased virulence and pathogenicity of this strain compared to *T. cruzi* G.

The 293 clusters shared among *T. conorhini* 025E, *T. cruzi* G and *T. cruzi* CL are enriched for lipid transport and metabolism and signal transduction mechanisms. The enriched lipid metabolism genes include acetyl-coA acetyltransferase and short-chain acyl-CoA dehydrogenase. The utilization of fatty acids by *T. cruzi* is known to be an important part of the parasite's metabolism [99,100]. *T. rangeli* AM80, which replicates in the bloodstream, hemolymph and salivary gland, may not use fatty acids as a significant source of energy, and may rely on carbohydrates to produce ATP. Such a preference for carbohydrates over fatty acids has also been demonstrated for *T. brucei* [101,102], which like *T. rangeli* undergoes only extracellular replication. Moreover, fatty acids are pro-inflammatory in intracellular amastigotes, e.g., in adipocytes [103], raising the possibility that this extra diversity in lipid metabolism genes is relevant to intracellular survival via mediation of the host pro-inflammatory response. Actin regulatory protein is the most prevalent member of the enriched signal transduction mechanisms category. Actin plays a central role in host cell invasion via endocytosis and phagocytosis/micropinocytosis [104]. Also present in this category are cysteine proteases, which are relevant to host cell invasion of *T. cruzi* and *T. cruzi*-like trypanosomes [105-107]. Enrichment of these categories in *T. conorhini* suggests the intriguing possibility that this parasite has an unknown intracellular stage.

Enrichment of category X in *T. rangeli* AM80, *T. cruzi* G and *T. cruzi* CL species or strain-specific genes but not *T. conorhini* 025E led to analysis of the functions of the clusters, using one gene from each cluster as a representative. The most common BLASTp hits for the *T. rangeli* AM80 paralog-containing clusters include hypothetical proteins, trans-sialidases, gp63 and (retrotransposon hotspot protein) RHS, with at least 5 clusters of each present. For *T. rangeli* AM80

singletons, the most frequent possible homologs in NCBI's non-redundant (nr) protein database are hypothetical proteins (274 clusters), trans-sialidase or *T. rangeli* sialidase (27 clusters), gp63 (13 clusters), protein kinase (8 clusters) and adenylate cyclase (7 clusters). The latter is important in the differentiation process of *T. cruzi*.

Over a quarter of the 1,245 *T. cruzi* CL strain-specific gene clusters containing paralogs are from trans-sialidase, RHS, dispersed gene family 1 (DGF-1), mucin-associated surface protein (MASP), mucin and gp63 multigene families. It is interesting to note in *T. rangeli* AM80 the complete absence of BLAST hits <1e-5 for MASP and DGF-1 for singletons and clusters containing paralogs, and much fewer mucin and RHS hits than for *T. cruzi* CL. This result implies that the *T. rangeli* AM80 MASP and DGF-1 genes have been lost or have diverged significantly. *T. cruzi* G, as for *T. cruzi* CL, contains many clusters with hits to trans-sialidase and RHS family members. The presence of many surface protein genes in this category for *T. rangeli* AM80 and *T. cruzi* potentially contributes to their wide host range and ability to sustain infection in the mammalian host.

Phylogenetic analysis

The main focuses of this analysis were to (i) resolve phylogeny for *T. brucei*, *T. cruzi*, *T. rangeli* and *T. conorhini* with large multigene alignment as opposed to single gene phylogeny (ii) define intra-strain phylogenetic relationships for *T. rangeli* (iii) determine phylogenetic distance of *T. conorhini* and *T. rangeli* from the *T. cruzi* clade (iv) estimate sequence identities.

Annotated proteins from *T. brucei* TREU 927, *T. rangeli* SC-58, *T. cruzi* Sylvio-X10, *Leishmania mexicana* MHOM/GT/2001/U1103, *Leishmania mexicana* Friedlin, *T. congolense* IL3000 and *T. vivax* Y486 were downloaded from TriTrypDB v.24. OrthoMCL v.2 [85] processing, with soft filtering for BLAST, 1e-5 E-value cut-off, 50% match cut-off and 1.5 inflation value, using the data from TriTrypDB and the gene calls from our sequenced genomes, identified 441 annotated single copy orthologs that are present in all species. Clustalo v.1.2 [108] alignments with Gblocks v0.91b [109] editing (parameters: b1=5, b2=6, b3=8, b4=5, b5=h) of these genes in 8 selected species were checked to ensure no alignment had >50% of positions filtered out or had a length of <100 amino acids. Two orthologs were removed in this analysis. EMBOSS infoalign [110] and a custom Python script were used to remove any edited alignments that contained a sequence >25% shorter than the median alignment length to avoid including partial or broken genes. Visual inspection of the remaining 242 edited alignments identified 71 that contained at least one poorly aligning sequence and were therefore removed, leaving a final set of 171 orthologous genes present in all 8 organisms. These gene alignments were concatenated using FASconCAT v1.0 [111], and the resulting supermatrices were used for phylogenetic reconstruction. ProtTest v.3.4 [112] Bayesian Information Criterion (BIC) determined that 88% of these proteins best fit the JTT substitution model, and 85% of proteins had gamma as the best model for rate heterogeneity. RAxML v.8.1.17 [113] PROTGAMMAJTT, which applies a gamma distribution with 4 discrete rate categories allowing for different rates of evolution at different sites, was used for building 200 maximum likelihood (ML) trees on distinct randomized stepwise addition parsimony starting trees to obtain the tree with the best likelihood. Support values for the tree were then obtained by rapid bootstrap analysis with 1000

replicates. Bootstrap values were then used to draw bipartitions on the best ML tree. TreeGraph 2.4.0 [114] and Inkscape 0.91 [115] were used for tree visualization and editing, with mid-point rooting on *T. brucei*. Figure 8 displays an overview of this multigene phylogeny process. The 171 amino acid alignments without Gblocks editing were used by PAL2NAL v.14 [116] to obtain corresponding codon alignments. These unedited amino acid and nucleotide alignments were each concatenated and used for average pairwise percent identity calculation in BioEdit v7.2.5 [117].

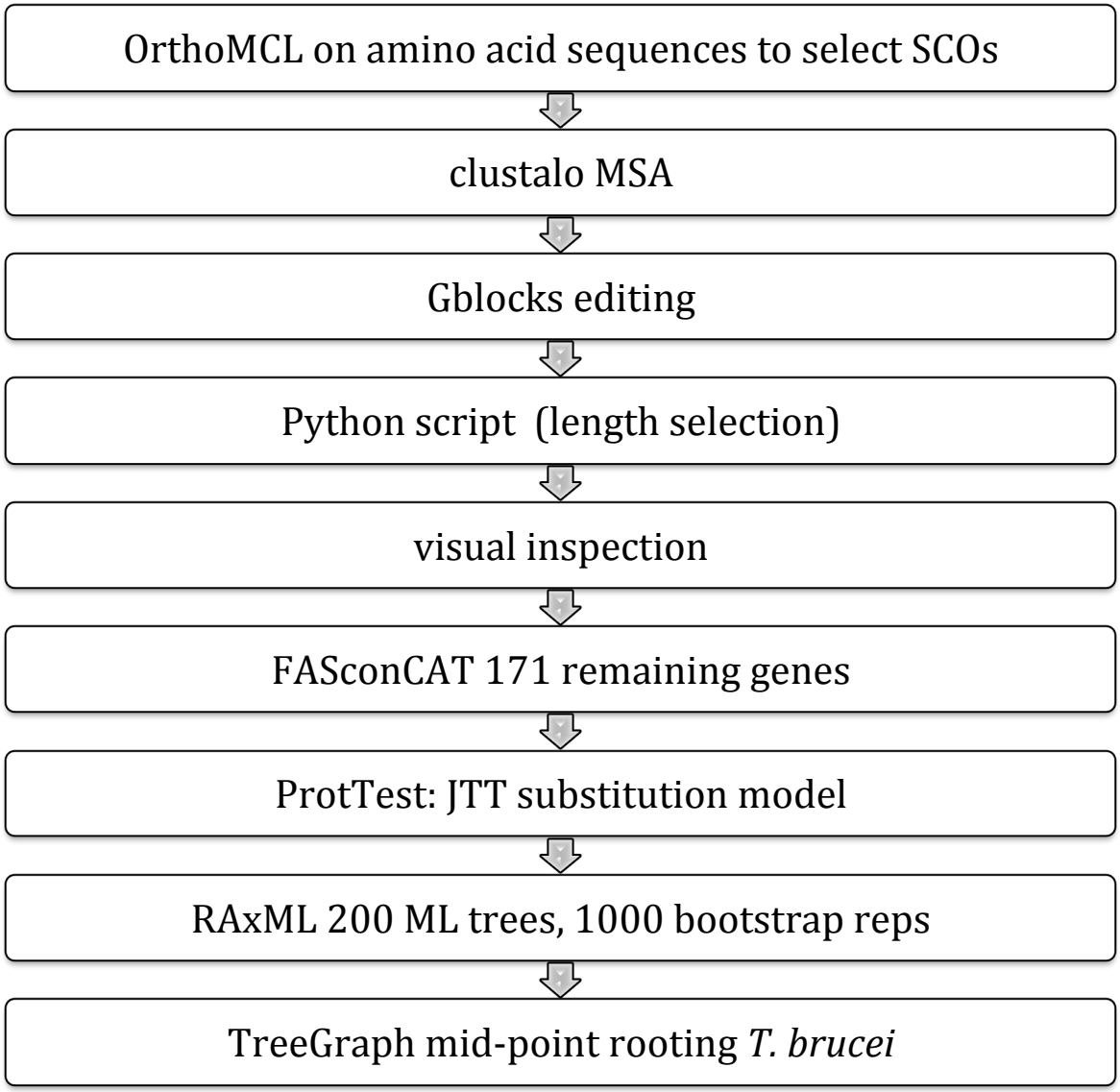


Figure 8: Multigene phylogeny steps. Multiple sequence alignment of orthologs, alignment editing, and Maximum Likelihood analysis to obtain a multigene phylogeny.

Percent identity analyses (Fig. 9) showed that, as expected, the highest identities observed, i.e., 94-97%, were between the *T. cruzi* isolates, with two DTU I isolates, G and Sylvio, being the most similar, aside from *T. cruzi* strains CL and CL Brener, which are clones from the same strain and exhibit near 100% identity (not shown). Percent nucleotide identity between the TraB (AM80) and TrD (SC58) isolates of *T. rangeli* was 91%. *T. conorhini* 025E and *T. rangeli* exhibited only ~81% identity to

each other, and ~74% identity to *T. cruzi* isolates. These observations support previous reports of *T. brucei* and *T. cruzi marinkellei* B7 percent identities to strains of *T. cruzi* [6,52]. Interestingly, the ~91% percent identity between *T. rangeli* AM80 and *T. rangeli* SC-58 was similar to the results comparing *T. c. marinkellei* and the *T. cruzi* strains.

T. brucei TREU 927	100							
T. rangeli AM80	56/54	100						
T. rangeli SC58	56/54	91/91	100					
T. conorhini 025E	56/54	82/81	80/79	100				
T. marinkellei B7	56/55	74/73	73/73	73/73	100			
T. cruzi CL	55/54	74/73	73/72	74/73	89/89	100		
T. cruzi G	56/55	75/74	74/73	74/74	91/91	95/95	100	
T. cruzi Sylvio-X10	56/54	75/74	74/73	74/74	91/91	94/94	97/97	100
	T. brucei TREU 927	*T. rangeli* AM80	*T. rangeli* SC58	*T. conorhini*	*T. marinkellei*	*T. cruzi* CL	*T. cruzi* G	*T. cruzi* Sylvio-X10

Figure 9: Nucleotide/amino acid percent identities of an ungapped alignment of 171 single copy genes. Genes used for multiple sequence alignments were concatenated and used to calculate percent sequence identity.

Phylogenetic analysis (Fig. 10) using amino acid sequences of these 171 orthologs confirmed that the *T. cruzi* strains are the most closely related, with the closest relationship between the two DTU I isolates, G and Sylvio.

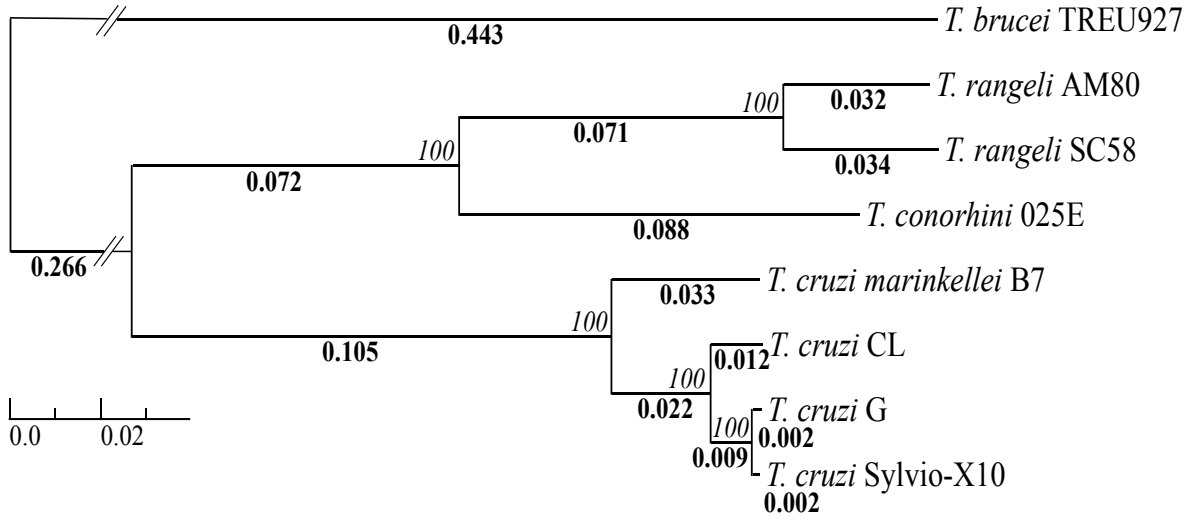


Figure 10: Multigene phylogeny Maximum Likelihood tree. Phylogenetic reconstruction and bootstrapping were performed using RAxML v.8.1.17. Bootstrap values are shown in italics. The scale bar indicates amino acid substitutions per site.

The subspecies *T. cruzi marinkellei* is clearly divergent from the *T. cruzi* strains. As expected, *T. conorhini* clusters with the *T. rangeli* strains, although the latter exhibit a greater distance from the *T. cruzi* strains, suggesting that there may be stronger evolutionary pressure for change in this taxon. The observed close relationship between *T. rangeli* and *T. conorhini*, and the greater evolutionary distance between them and *T. brucei* clades than the distance between the *T. cruzi* and *T. brucei* clades, confirm previous phylogenetic analyses based on a few genes [13,41,46,118,119]. Given the relatively higher number of genomes available within the *T. cruzi* clade, a greater taxon sampling of genomes closely related to *T. conorhini*, *T. rangeli* and species more closely related to the *T. brucei* clade, which are not yet available, would have allowed for more accurate and complete phylogenetic reconstruction. The above, however, represents the first phylogenetic

reconstruction of these species using a large multigene alignment. This arguably provides a less biased phylogenetic signal than basing phylogeny on a particular gene or small group of genes.

Heterozygosity

Heterozygosity measurements were used to (i) investigate the influence of the hybridization history of *T. cruzi* CL on heterozygosity compared to the other species (ii) determine whether there are any major heterozygosity differences across the other species, possibly providing insight into their hybridization history or reproductive mechanisms.

A custom Perl script, which calls SAMtools v1.2, was used to estimate heterozygosity. The script first aligns high quality sequence reads, i.e., reads with 70% of the bases with quality score ≥ 25, to each genome assembly. Polymorphic positions are then quantified in each of 6,479 single copy OrthoMCL v.2-generated orthologs present in each of the four genomes examined. A 25% cutoff of reads was used to confirm heterozygosity; i.e., at least 25% of reads must have a polymorphism at a site to be counted as a heterozygous position, and all positions considered for analysis had a read depth ≥ 10 and q-scores ≥ 25 to increase confidence. Synteny of the distribution of heterozygosity values at local areas of the genomes was assessed by Spearman's Rank followed by adjusting the p-values using the Bonferroni correction using Rstudio. Windows of distance (bp) and number of genes had to be similar for pairwise species comparisons.

The four organisms described herein are thought to be primarily diploid, although some *T. cruzi* strains, e.g., CL and CL Brener, are hybrid strains in which ploidy is less well defined [87,120,121]. I therefore examined levels of apparent heterozygosity in these strains using the set of 6,479 conserved single copy orthologs, covering ~9 million sites in each genome. As described above, a SNP was called if > 25% of the reads showed a variant call at that site. The number of heterozygous genes with at least one SNP varied from ~55% in *T. cruzi* G to ~92% in *T. cruzi* CL, and the average percent of heterozygous bases varied from ~0.2% in *T. cruzi* G to ~1.1% in *T. cruzi* CL, confirming that a single species can exhibit a wide range of sequence variability in gene alleles (Fig. 11). The high level of heterozygosity in *T. cruzi* CL is very likely mostly due to the fact that it is a hybrid in which a significant fraction of its genes are derived from two distantly related progenitors. Short read sequencing technologies do not permit the resolution of these alleles into separate contigs.

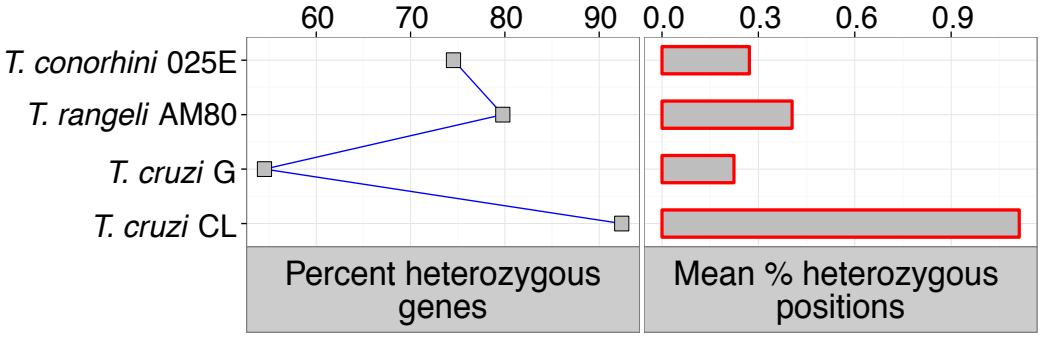


Figure 11: Heterozygosity of single copy orthologs. Summary values from 6,479 shared single copy ortholog genes. Percent heterozygous genes indicate percentage of genes with at least one heterozygous position, mean percent heterozygous positions were calculated by dividing the total number of positions by the number of heterozygous sites.

The percent heterozygous positions for *T. cruzi* G and *T. conorhini* 025E (~0.3%) are close to estimates of heterozygosity in *T. c. cruzi* Sylvio X10 (~0.22%) and *T. c. marinkellei* B7 (0.19%) where similar minimal coverage thresholds and methods were used [6], although those analyses were not restricted to single-copy orthologs, which may have positively biased their estimates. *T. rangeli* AM80 appears to have relatively high values for heterozygosity, with 80% of its genes having at least one heterozygous position and 0.4% heterozygous bases, possibly suggesting an elevated mutation rate, a recent hybridization or sexual recombination. The distributions of heterozygous genes falling into discrete mean levels of percent heterozygous positions were unimodal for all species (Fig. 12), suggesting that the genes examined were not of biased origin.

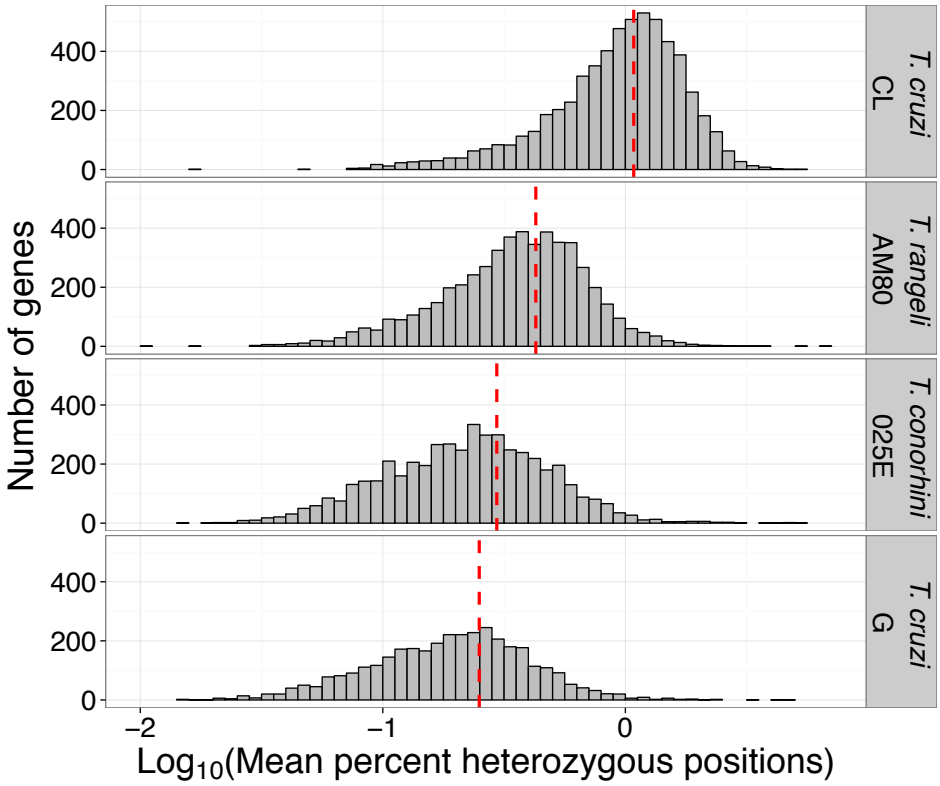


Figure 12: Heterozygosity distribution for single copy ortholog genes. Histogram showing the distribution of heterozygosity values among heterozygous genes. Red vertical dashed lines represent the mean values.

There was no significant enrichment of any of the 25 general KOG categories or enzyme E.C. numbers in genes with high or low heterozygosity values (data not shown). The overlap of highly heterozygous genes from the 6,479 orthologs in each species was limited, and I found no evidence of synteny in heterozygosity patterns across contigs in any pairwise species comparison (data not shown). Together, these observations suggest that generation of heterozygosity in these organisms is a stochastic process.

T. cruzi displays strong linkage disequilibrium and features of a mainly clonal species [122]. However, the presence of natural *T. cruzi* hybrids such as those of TcV and TcVI and conservation of meiosis-related orthologs in the genomes suggest the capacity for sexual reproduction. However, the genomes of *T. conorhini* 025E and *T. cruzi* G exhibit overall low levels of heterozygosity. These levels do not seem to fit with a strictly clonal model of evolution, where diversity is expected to accumulate independently between alleles in an individual over time (the “Meselson effect”). Moreover, long-term clonality without mechanisms to attenuate the impact of high mutational load (“Müller’s ratchet”) would seem to be detrimental to these species.

A comprehensive analysis of heterozygosity in sequences spanning the genomes compared to expected heterozygosity is beyond the scope of the present study. However, my analysis of single copy orthologs identified an apparent mosaic pattern of heterozygosity across these genomes, especially in *T. cruzi* G and *T. rangeli* AM80, where continuous regions of homozygosity often exceeding 50 Kbp interspersed with heterozygous clusters were identified. Mosaic heterozygosity has

been seen before in *Naegleria gruberi* [123], and clustering of heterozygosity has also been described in *T. c. marinkellei* [6]. Precise mechanisms that control heterozygosity in trypanosomes have yet to be elucidated. Many regions of low heterozygosity in all species have average coverage for single copy orthologs. Therefore, loss of heterozygosity via chromosome loss seems unlikely in these species, although this process cannot be ruled out. Mutational hotspots, mitotic recombination, mitotic gene conversion, and segmental duplication are also possible sources of differentially heterozygous regions of *T. cruzi* G and *T. rangeli* AM80 chromosomes.

Members of DTU TcVI, including hybrid strain *T. cruzi* CL, are reported to have a high degree of fixed heterozygosity but low intralineage diversity [30]. The higher heterozygosity originates at least in part from the distances between the Esmeraldo-like (TcII) and non-Esmeraldo-like (TcIII) alleles, provided by the 'parental strains' of DTU TcVI. As previously suggested [88], these hybrid genotypes were likely stabilized through long term asexual reproduction.

It is interesting to note that *T. cruzi* CL and *T. rangeli* AM80, two human-infecting parasites, exhibit higher heterozygosity. Increased heterozygosity has been linked to hybrid vigor, which has been reported in *Leishmania* [124], and is consistent with an enhanced host range and the ability to invade cells, replicate, and cause pathogenicity.

Multigene families

As described above, many differences across trypanosomatids are associated with differential multigene family expansion. For example, dispersed gene family 1 protein

(DGF-1), which has been implicated in the ability of parasites to bind to extracellular matrix proteins of host cells, was found in almost three-fold higher quantities in CL Brener strain (TcVI DTU) than the Sylvio X10/1 strain (TcI DTU) [7] . Additionally, unexpected expansions of multigene families have been observed given the lifestyle of certain species. For example, *T. rangeli* SC-58 was found to possess greater than two-fold the number of amastin copies than *T. cruzi* [19], despite this protein being implicated in the intracellular amastigote stage of *T. cruzi* mammalian infections and the extracellular replication of *T. rangeli*.

The aims for multigene families characterization in these organisms were to (i) define which gene families are present in each genome, and their role in shaping life-cycle and infectivity, (ii) analyze the motifs of the amastin gene family, and determine factors that may explain the presence of amastin in *T. rangeli*.

Thirteen multigene families were selected for analysis based on gene cluster diversity analyses of this study and literature searches. Copy numbers were calculated using an equation previously described [95]. Called genes by GeneMarkS v.4.7b [76] were grouped into multigene families based on annotation via BLASTp against NCBI's non-redundant protein database (E-value threshold 1e-5). Gene coordinates were then converted to GFF format and reads mapping was performed with BWA v.0.7.12 [75] (default parameters). BEDtools intersect and BEDtools coverage [125] were used to obtain the number of reads mapping to each multigene family, and the total number of reads was obtained using SAMtools v1.2 [126]. Total reads for *T. conorhini* 025E, *T. rangeli* AM80, *T. cruzi* G and *T. cruzi* CL were, respectively, 2,502,202, 3,028,172, 2,529,218 and 2,053,970. Haploid genome sizes used for the calculation were as estimated by averaging the NGS estimate (number

of bases / modal read depth of the assembly) and the densitometry estimate (Table 5). The complete gene length for each multigene family member was taken from UniProtKB full-length genes (Appendix 1). Fragmentation of genes is a common problem in copy number estimation of complex and incomplete genomes [127], and given that portions of genes in the genome may not assemble my estimates are likely conservative. As a validation for the read-based approach I obtained values of 1 for dihydrofolate reductase*,* poly(A) polymerase and DNA topoisomerase type IB, which are widely considered to be single copy genes in trypanosomatids, using the same methodology. A secondary validation using Equation 1 was performed, and similar copy number were obtained, thus also validating my estimate of haploid genome size for the organisms. Amastin motif prediction was performed on translated gene sequences using MEME [128] v.4.10.0 with the anr option and a maximum width of 20.

Like previous reports about *T. cruzi* [6,7,129], the four genomes described herein have variable representations of genes in multigene families (Fig. 13).

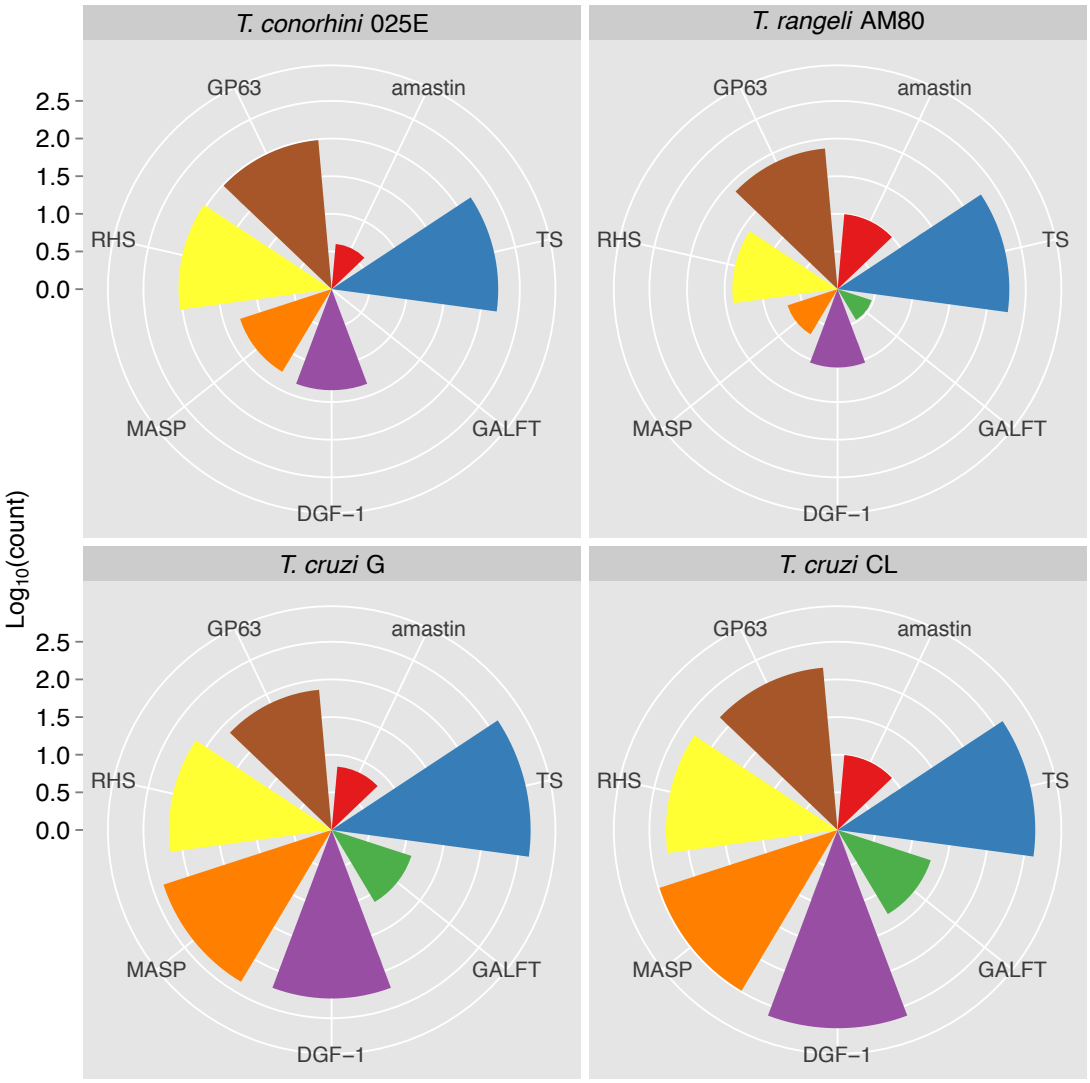


Figure 13: Multigene family copy numbers. Selected major multigene families shown are amastin, β-galactofuranosyl transferase (GALFT), surface protease GP63, retrotransposon hot spot (RHS) protein, mucin-associated surface protein (MASP), trans-sialidase (TS), and dispersed gene family protein 1 (DGF-1). Centers of plots represent 1 copy (0 in $\log_{10}$) and successive concentric circle values are shown by the $\log_{10}$ scale bar on the left.

Trans-sialidase (TS) and GP63 are highly expanded in all species. The TS family genes, which encode proteins that are linked to the cell membrane via GPI anchors, are very heterogeneous and form eight known groups [130-132]. The enzyme in *T. cruzi* transfers host sialic acids to parasite cell surface ligands, presenting a decoy to the host immune response and participating in the adhesion and internalization of

the parasites into host cells [133-136]. *T. rangeli* has a Group II sialidase that is a strict hydrolase lacking the ability to transfer sialic acid [19,130,137-139]. In both species, TS Group II enzymes likely participate in host cell adhesion and invasion, but for *T. rangeli* this activity is probably only relevant in the triatomine vector [137,139]. The sequences of *T. conorhini* TS were the most divergent compared to those from *T. cruzi, T. cruzi*-like species and *T. rangeli* [130]. The findings from this and previous studies uncovering TS genes in all *Trypanosoma* species suggest that, in addition to participation on host cell invasion and intracellular survival, TS may play some roles in parasite development in their insect vectors [129]. GP63 proteins are zinc-dependent metalloproteases that are highly expressed in *T. cruzi* amastigotes, where they contribute to cell infection [140,141]. This activity is consistent with my observation that pathogenic *T. cruzi* CL has the highest number of copies of this gene (~146), similar to the 174 copies predicted in *T. cruzi* CL Brener [18,99].

The most striking differences in copy number across the species are arguably in the mucin-associated surface protein (MASP) and dispersed gene family 1 (DGF-1) families, which are both highly repetitive in both *T. cruzi* strains, but less amplified in *T. rangeli* and *T. conorhini*. MASP genes are often found in clusters with mucin and other surface protein genes [18], and the protein is localized to the surface of infective forms of *T. cruzi* [142]. Polymorphism of MASP amino acid sequence is high, which likely contributes to immune system evasion [142] and the parasite's ability to infect multiple cell types [129,143]. Thus, the smaller size of the MASP gene family in *T. conorhini* 025E and *T. rangeli* AM80 (and *T. rangeli* SC-58 [19]), in contrast to *T. cruzi* strains, may be related to their lack of host cell infectivity, and their inability to induce acute infections with high levels of parasitemia or long chronic

infections. DGF-1 is less well represented in *T. rangeli* AM80 than in the other species, and shows lower diversity in gene cluster analysis in this study. Possession of only 11 copies of DGF-1 may contribute to the obligate extracellular nature of *T. rangeli* AM80 in the mammalian host, since this protein has been implicated in the ability of parasites to bind to extracellular matrix proteins of host cells [144]. *T. rangeli* SC-58 was estimated to have over 400 copies of this gene, despite less than 20 partial DGF-1 genes being annotated with genome coordinates [19]. If the latter copy number estimate is accurate, there is a striking inter-strain difference. *T. conorhini*, bearing only 22 copies, also shows significantly reduced numbers of DGF-1 compared to *T. cruzi*, which is consistent with its extracellular lifestyle.

Cruzipain, a key player in cell invasion, and glycosyltransfase enzyme, GALFT, which is involved in GPI anchor biosynthesis, are also highly differentially expanded. I find no evidence of cruzipain expansion in *T. rangeli* AM80 or *T. conorhini* 025E, although cruzipain homologs are present in these genomes. In *T. rangeli* the homolog is known as rangelipain and is known to be present in tandem repeats [41]. Amino acid identities for these genes are 76% between *T. rangeli* AM80 and *T. conorhini* 025E, 71% between *T. conorhini* 025E and the *T. cruzi* strains, and 69% between *T. rangeli* AM80 and the *T. cruzi* strains. Our group previously inferred network genealogies showing that cruzipain sequences of all DTUs of *T. cruzi* clustered tightly together and closer to *T. c. marinkellei* than to *T. dionisii* (*T. cruzi*-like species) while largely differed from homologs of *T. rangeli* and *T. brucei;* this study unveiled DTU- and species-specific polymorphisms [107]. Cruzipain precursors are activated upon removal of the N-terminal prodomain, resulting in proteins linked to the invasion process that are thought to play a larger role in *T. cruzi* CL, where expression levels are higher during infection, than *T. cruzi* G [34].

However, I do not see an expansion of cruzipain precursors in the CL strain, which has ~22 copies, compared to ~70 copies in *T. cruzi* G.

Kinetoplastid Membrane Protein-11 (KMP-11) is encoded in *Leishmania*, *T. brucei* and *T. cruzi.* The observation that the KMP-11 genes are expanded in *T. rangeli* SC-58 [19] represented an unexpected result in this non-pathogenic strain. This finding is more unusual given that the gene is found in low numbers across other trypanosomatids [145,146] and in this analysis I find just one copy in *T. conorhini* 025E, *T. rangeli* AM80 and *T. cruzi* CL, and none in *T. cruzi* G.

Mucin, a family thought to confer immune system protection [133,147], contains highly variable regions that make copy number estimation challenging. Although likely underestimated here, I observe a larger gene family in *T. cruzi* CL compared to the other species, presumably contributing to the poorer immune system clearance of this strain. Additionally, *T. rangeli* and *T. conorhini* appear to contain mostly mucin-like glycoproteins and little of the diversity of other mucin subgroups that is typical of *T. cruzi*, concurring with reports in other *T. rangeli* strains [148,149]. The low copy numbers of this gene family in these two species are also consistent with previous genomic [19] and transcriptomic [138] data from *T. rangeli,* and likely contribute to their inability to invade mammalian cells [129].

I find a lower copy number of amastin in *T. conorhini* 025E (~4 copies) compared to *T. rangeli* AM80 (~10 copies), *T. cruzi* G and CL (~7-10 copies), and *T. cruzi* CL Brener (14 copies) [150]. Although the exact function of amastins remains unclear, they are thought to be abundantly expressed on the surface of intracellular *T. cruzi* amastigotes and apparently support intracellular survival [129,151-155]. Since amastin is expressed in the intracellular mammalian amastigote stage of the parasite's life cycles in *T. cruzi* and *Leishmania*, finding expansion of this

immunogenic gene family in the extracellular *T. rangeli* AM80 (also previously reported in *T. rangeli* SC-58 [19]) was unexpected.

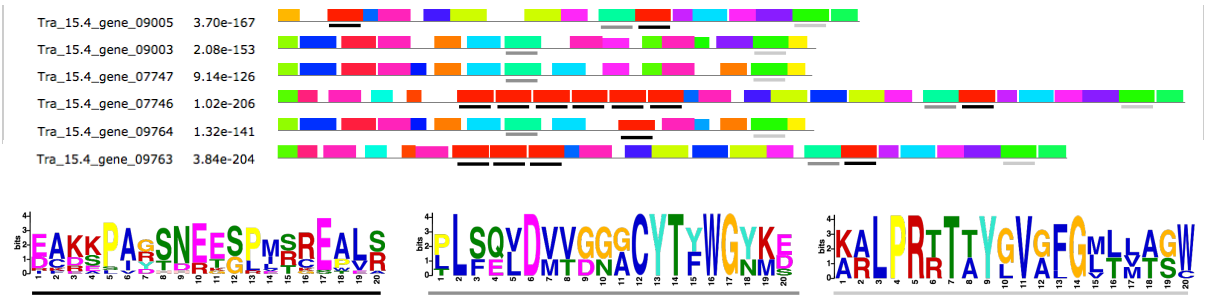


Figure 14: Repetitive motifs found in *T. rangeli* AM80 amastin multigene family. Motifs are linked between sequence logos and maps by underline in black, dark grey and light grey.

Motif analysis shows that the conserved amastin signature sequence of C-[IVLYF]-[TS]-[LFV]-[WF]-G-X-[KRQ]-X-[DENT]-C, which may be critical for amastin function [156], is present in all the species examined. Additionally, I found a motif, with consensus EAKKPAGSNEESPMSREALS, tandemly repeated 6 and 3 times respectively in two of the eight amastin genes analyzed from *T. rangeli* AM80 (Fig. 14). A combination of factors makes this result interesting: *T. rangeli* has high persistence in the mammalian host, and amastin is highly immunogenic. The function of this repeat is unknown, although I postulate that these degenerate repeats may aid recombination and antigenic reshuffling associated with evasion of the host immune system [157].

Pseudogenes

Aims for pseudogene analysis included (i) determine the percent pseudogenes out of total genes for each organism (ii) define the functions of predicted pseudogenes.

The steps for pseudogenes prediction are shown in Figure 15. Briefly, longest nr database hits for each genomic coordinate were predicted by gapped BLAST with the program lastal [158] v.744 (parameters -F15, -l5 –K20, -X 150 –P0), followed by selection of hits containing frameshifts or premature stop codons. Coordinates were converted to GFF format, removing any overlapping genes called by GeneMarkS v.4.7b [76] using BEDTools v.2.19.1 intersect. Since over 98% of 2,217 single copy orthologs shared between *T. brucei*, *T. vivax*, *T. congolense*, *T. dionisii*, *T. cruzi* and *Leishmania* species were present in the assemblies (Fig. 2), likelihood of finding false positives due to uncalled genes was low.

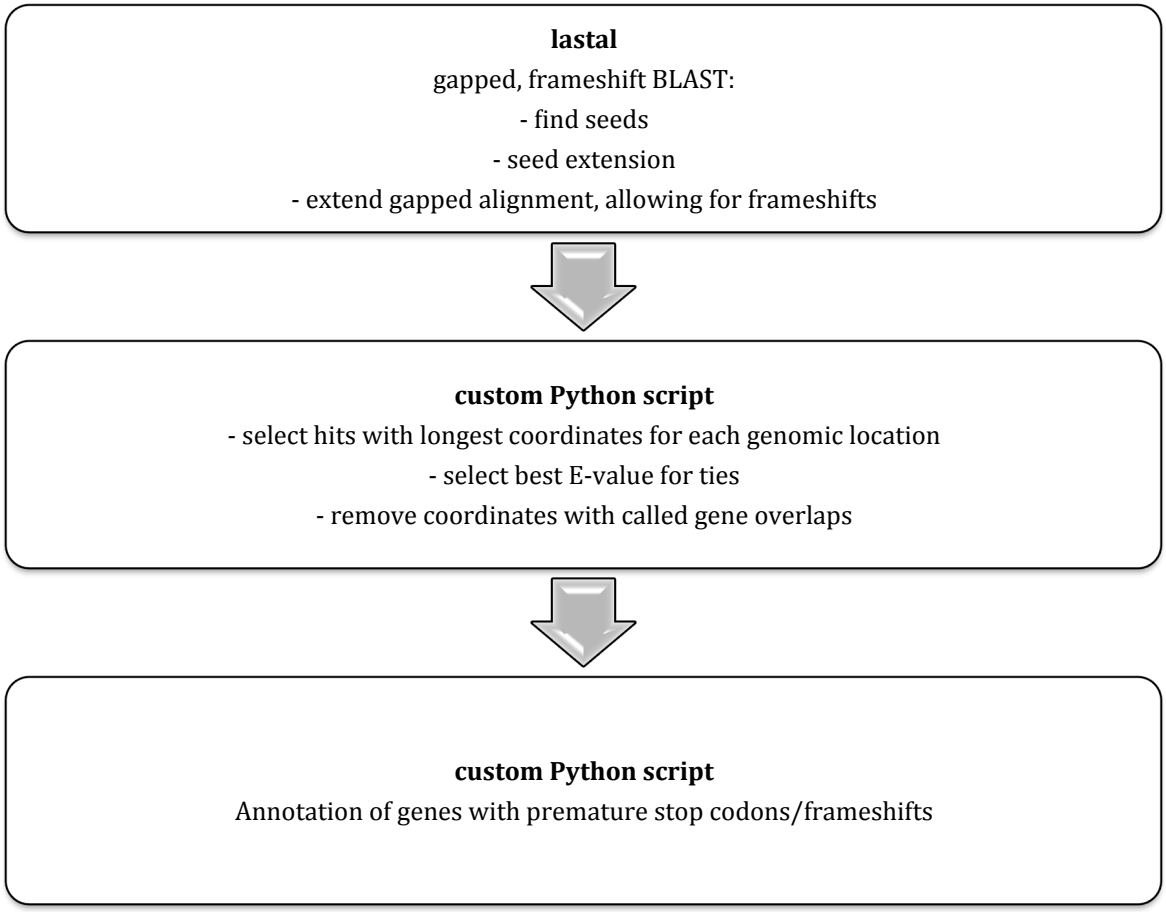


Figure 15: Steps in pseudogene prediction. A gapped frameshift BLAST was performed with the program lastal, followed by annotation using custom Python scripts.

Pseudogenes are defined herein as genes bearing in-frame stop codons or frameshifts, as well as the absence of features required for gene calling based on a non-supervised training model, such as upstream functional sites, start codons, nucleotide and amino acid composition, and length to the first in-frame stop codon.

The number of putative pseudogene coordinates predicted in the *T. conorhini, T. rangeli*, *T. cruzi* G, and *T. cruzi* CL genome assemblies were 113, 434, 942 and 4420, respectively (Table 8). The latter equates to 16% of total gene predictions (gene calls plus pseudogenes) in the CL strain. The *T. cruzi* CL Brener genome was

previously estimated to have 3,590 pseudogenes, or ~16% of all its genes. Over 2,000 of these were attributed to large multigene families [18].

Table 8: Pseudogene predictions summary. Total pseudogenes and percent pseudogenes out of total gene counts.

	T. conorhini 025E	*T. rangeli* AM80	*T. cruzi* G	*T. cruzi* CL
Number of pseudogenes	113	434	942	4420
% pseudogenes	1	4	7	16

NCBI nr annotated functions of the panels of predicted pseudogenes corresponded to 600 copies of putative pseudogenes from multigene families in *T. cruzi* CL. In both *T. cruzi* G and CL, the most frequent putative pseudogenes were of the trans-sialidase, RHS and MASP gene families, and hypothetical proteins.

The pseudogenes in these genomes may provide a repertoire of genetic information for producing variation, especially in multigene-families. *T. cruzi* was the first species in which a tandem array of pseudogenes, consisting of six mucin genes each with an in-frame stop codon, was discovered [159]. These were postulated to be selectively maintained in the genome, possibly to generate mucin gene diversity. A diversifying role has been suggested for the numerous pseudogenes of variable surface glycoproteins (VSGs) in *Trypanosoma equiperdum* and African trypanosomes that undergo rapid antigenic variation through gene recombination [160-163]. Additionally, TS gene and pseudogene organization, flanked by RHS genes at subtelomeric regions, in strain CL Brener is reminiscent of regions next to *T. brucei* VSG genes [164]. Pseudogenes could also play a role in post-transcriptional control of gene expression. Some pseudogenes transcribed in *T. brucei* have been proposed to participate in RNAi-based natural antisense suppression [165]. The genes responsible for RNAi machinery are absent in all

strains of *T. cruzi* examined to date, but present in both *T. rangeli* AM80 and *T. conorhini* 025E (A. Matveyev, personal communication). Analysis of the transcriptional activities and structural organization of these pseudogenes is beyond the scope of this study, but may clarify their roles in generation of protein diversity or post-transcriptional regulation.

Repetitive elements

Repetitive elements are thought to be very species-specific in parasitic protozoa [166]. The main repetitive elements in *T. cruzi* are:

L1Tc: a non-long terminal repeat retrotransposon from *Trypanosoma cruzi*

- 4.9-Kbp actively transcribed element
- It contains a single open reading frame coding for the machinery necessary for its autonomous retrotransposition.

SIRE: 428 bp short interspersed repetitive element of *T. cruzi*

VIPER: 2326 bp retroelement beginning with the first 182 bp of SIRE, and ending in the last 220 bp of SIRE. It contains machinery necessary for autonomous retrotransposition.

Thus, SIRE depends on VIPER for its mobility since it is a non-autonomous element.

Goals for this analysis included (i) determine the number of *T. cruzi* specific repetitive elements for each organism (ii) *de novo* repetitive element prediction.

Repeat counts in intergenic regions of the assemblies were identified by performing a Cross Match v. 0990329 search and categorization with Repeat Masker [167] v. 4.0.6. Repeat Masker library sequences of *Trypanosoma* species derived from Repbase (20150807 download) were used as a database for the search. *De novo* repeats were predicted using Repeat Masker with a library built using Repeat Modeler [168] v1.0.8. The latter identified and modeled *de novo* repeat families from the four genomes using RECON v.1.08, RepeatScout v.1.0.5 and Tandem Repeat Finder v.4.0.4 [169-171], with an RMBLASTn [168] v.1.2 search of Repbase.

Known trypanosome repeats in the genomes based on hits to the Repbase database [172], i.e., non-Long Terminal Repeat (non-LTR) elements, LTR elements, and satellites, are shown in Table 9.

Table 9: Predicted *Trypanosoma* repetitive elements

	T. conorhini 025E	*T. rangeli* AM80	*T. cruzi* G	*T. cruzi* CL
LTR retrotransposons:	--	--	--	--
SIRE (0.43 kb)	3	0	463	709
VIPER (4.5 kb)	0	2	146	287
Non-LTR retrotransposons:	--	--	--	--
CZAR (7.25 kb)	2	1	1	11
L1*Tc* (4.9 kb)	4	9	58	119
Other:	--	--	--	--
Satellite	8	27	138	620
Simple repeat	13715	11892	15523	17692

Repeat profiles are similar for *T. conorhini* 025E and *T. rangeli* AM80. The most common satellite sequence in all species is SZ23_TC. Retroelements are markedly increased in the *T. cruzi* strains. Analysis of *T. conorhini* 025E and *T. rangeli* AM80 LTR- and non-LTR elements identified only 22 and 35 elements, respectively. It is

tempting to hypothesize that these organisms may lack “copy and paste” type retrotransposons. To show that fewer repetitive element copies in the genomes of *T. rangeli and T. conorhini* was not just due to collapse of reads in highly similar repeats, I calculated the average coverage of the *de novo* repeat finder predictions in each repeat class. *T. cruzi* CL has coverage estimates close to the genomic average for every repeat class. *T. cruzi* G and *T. conorhini* 025E have around two-fold higher average coverage for the non-LTR and satellite sequences than the genomic average, and *T. rangeli* AM80 has two-fold higher coverage for just non-LTR sequences than the genomic average. Additionally, *de novo* predictions again confirmed that *T. rangeli* and *T. conorhini* possess much fewer LTR- and non-LTR repetitive elements. Retroelements may have wide-ranging implications on generation of genomic diversity, and their greater number in *T. cruzi* may have potentiated antigenic variation in this complex parasite [166].

Metabolic pathways

The main aim for metabolic pathway analysis was to explore the role of metabolic potential in defining why divergent lifestyles arose for these species, and why they may currently have different preferred environments.

The steps for metabolic pathway analysis are shown in Figure 16. Briefly, database reference genes from UniRef100 [173] and the Kyoto Encyclopedia of Genes and Genomes KEGG [174] were located on assembly contigs and mapped to metabolic pathways using ASGARD [175]. Enzymes found to be differentially present among the four species were subjected to an additional tBLASTn analysis of the sequencing

reads, requiring >60% of the reference gene sequence to be covered by at least four reads with an E-value <1e-5 to indicate presence.

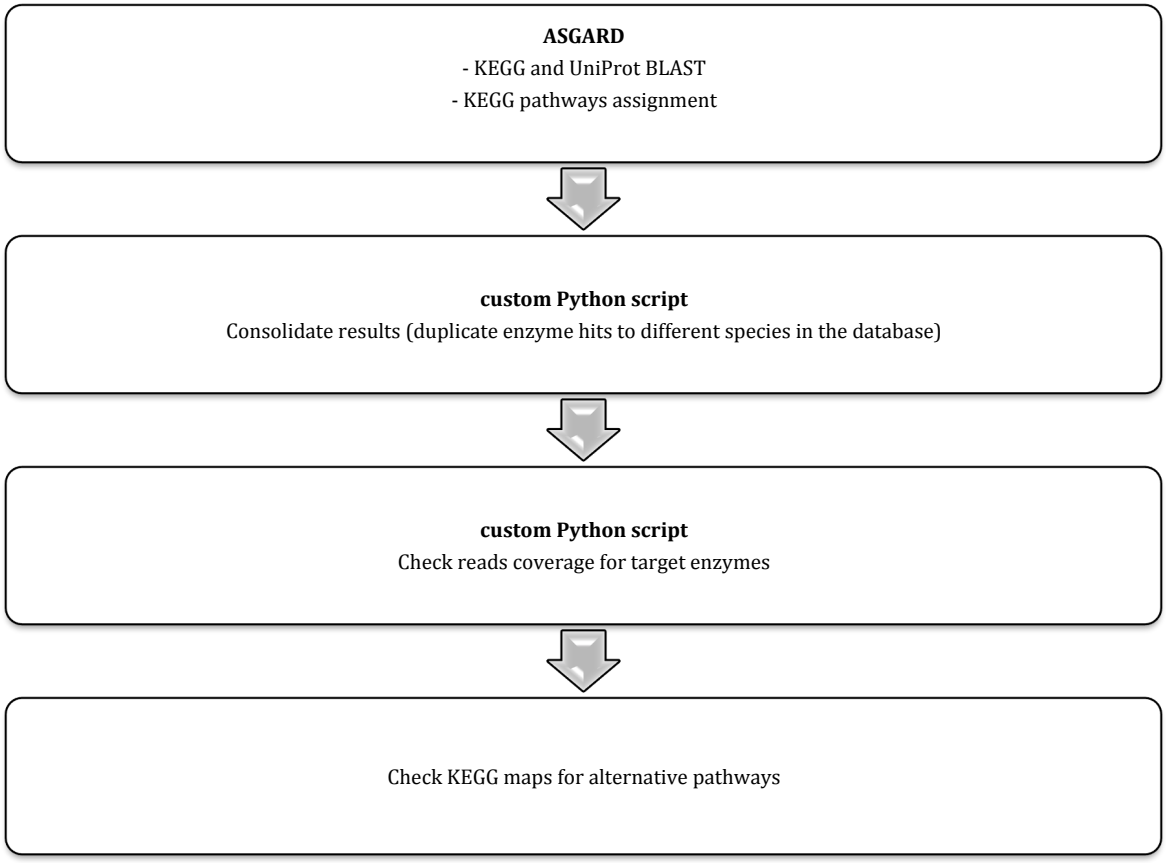


Figure 16: Metabolic pathways analysis steps. The program ASGARD was used to determine presence of enzymes in each organism and assign the enzymes to metabolic pathways.

Some of the more significant observations from these analyses are summarized in Table 10, and outlined below.

Table 10: Major metabolic pathways with differential gene presence

Enzyme	E.C.	*T. conorhini* 025E	*T. rangeli* AM80	*T. cruzi* G	*T. cruzi* CL	Reaction
Fatty acid metabolism						
Glycerol dehydrogenase	1.1.1.6	present	present	absent	absent	glycerol + NAD+ = glycerone + NADH + H+
Amino acid metabolism						
Proline racemase	5.1.1.4	present	absent	present	present	L-proline = D-proline
5-oxoprolinase	3.5.2.9	alternative	alternative	present	present	ATP + 5-oxo-L-proline + 2 H_2O = ADP + phosphate + L-glutamate
Primary-amine oxidase	1.4.3.21	present	absent	present	present	oxidize primary monoamines
Branched-chain amino acid aminotransferase	2.6.1.42	present	present	absent	absent	L-leucine/L-isoleucine/L-valine + 2-oxoglutarate = 4-methyl-2-oxopentanoate + L-glutamate
Glycine hydroxymethyltransferase	2.1.2.1	alternative	present	present	present	5,10-methylenetetrahydrofolate + glycine + H2O = tetrahydrofolate + L-serine
Carbohydrate metabolism						
NADP-alcohol dehydrogenase	1.1.1.2	alternative	alternative	present	present	an alcohol + NADP+ = an aldehyde + NADPH + H+
beta-glucosidase	3.2.1.21	present	present	absent	absent	Hydrolysis of terminal, non-reducing beta-D-glucosyl residues with release of beta-D-glucose
fructuronate reductase	1.1.1.57	present	present	absent	absent	D-mannonate + NAD+ = D-fructuronate + NADH + H+
xylulokinase	2.7.1.17	present	present	absent	absent	ATP + D-xylulose = ADP + D-xylulose 5-phosphate
mannitol 2-dehydrogenase	1.1.1.67	present	present	absent	absent	D-mannitol + NAD+ = D-fructose + NADH + H+
Terpenoid backbone synthesis						
CAAX prenyl protease 1	3.4.24.84	absent	absent	present	present	hydrolysis of peptide bond
Pyruvate metabolism						
D-lactate dehydrogenase	1.1.2.4	present	present	alternative	alternative	(R)-lactate + 2 ferricytochrome c = pyruvate + 2 ferrocytochrome c + 2 H+

absent
alternative pathways exist
present

Fatty acid metabolism

Trypanosomatids possess a unique set of elongase enzymes for *de novo* fatty acid synthesis [176]. *T. cruzi* amastigotes utilize lipid-dependent energy metabolism [99], but the functional importance of fatty acid oxidation in trypanosomatids is not fully understood. The organisms analyzed herein appear to be capable of synthesizing and oxidizing fatty acids. Glycerol dehydrogenase (EC 1.1.1.6), which is involved in converting glycerol to dihydroxyacetone, is absent in *T. cruzi* G and CL, but present in *T. rangeli* and *T. conorhini*. This enzyme was reportedly acquired by lateral gene transfer in *Leishmania*, *Crithidia* and *Leptomonas* spp. [177], enabling the parasites to use glycerol as a carbon source.

Amino acid metabolism

In *T. cruzi*, amino acids are relevant in energy metabolism [178-181], host-cell invasion [182], stress resistance [183,184], and differentiation [181,185]. Proline, in particular, plays a fundamental role in these processes, including energy support during the parasite's intracellular stages [181].

T. cruzi, unlike *T. brucei*, can metabolize D-proline using a putative proline racemase (PRAC). Although the *T. conorhini* genome bears a PRAC gene, as we have previously described [119] *T. rangeli* AM80, and strains of all other known *T. rangeli* lineages [119] contain only a pseudogene for this enzyme. Interestingly, genes for 5-oxoprolinase, which is involved in L-proline metabolism, are absent in *T. conorhini* 025E and *T. rangeli* AM80. Analysis of the available genomes of TriTrypDB [96] showed only intracellular-replicating species seem to possess this enzyme, which makes its apparent loss in these two species expected. L-proline metabolism via 5-oxoprolinase produces L-glutamate in the glutathione-mediated stress response pathway. *T. rangeli* AM80 also lacks enzymes that use oxygen as an acceptor (EC 1.4.3.-), which further limits the pathways available for glutamate synthesis. Absence of these enzymes may shed new light on the recent finding that *T. rangeli* SC-58 is particularly susceptible to oxidative stress [19]. However, there appear to be alternative pathways to produce glutamate and glutathione in both *T. rangeli* AM80 and *T. conorhini* 025E, e.g. glutamate dehydrogenase, which converts α-ketoglutarate to glutamate.

Consistent with reports of other kinetoplastid protozoa [186,187], all of these species lack ornithine and arginine decarboxylase genes, indicating that they are unable to generate putrescine or other polyamines and must salvage them from their hosts. Primary-amine oxidase, which is significant for amino acid metabolism and alkaloid biosynthesis, is absent in *T. rangeli* AM80. The *T. rangeli* AM80 and *T.*

conorhini 025E genomes encode branched-chain amino acid aminotransferase, which is required for synthesis and degradation of valine, leucine and isoleucine. That *T. cruzi* strains lack this gene [14] is interesting since leucine is reported to act as a negative regulator of proline-dependent metacylogenesis [188]. Additionally, the ability of this parasite to use the intact leucine skeleton, presumably obtained from the host [189], for isoprenoid and sterol formation would confer advantages in energy economy [190]. *T. cruzi* and *T. rangeli* can interconvert serine and glycine, a capacity not found in *T. brucei* [14]. *T. conorhini* 025E, like *T. brucei*, lacks the glycine hydroxymethyltransferase gene for conversion of glycine to L-serine and tetrahydrofolate or *vice versa*, although alternative routes exist in this organism for synthesis of these compounds.

Carbohydrate metabolism

Kinetoplastids compartmentalize carbohydrate metabolism in organelles known as glycosomes [191]. Glucose is the predominant carbohydrate utilized by *T. cruzi* [192] and *T. brucei* [193], although *Leishmania* spp. and *Phytomonas* spp. have developed adaptations to metabolize plant-derived carbon sources [14,194]. Genes for NADP-alcohol dehydrogenase (EC 1.1.1.2), which is required for degradation of the aromatic compound cyclohexanol, and participates along with other enzymes in acetaldehyde to ethanol interconversion in glycolysis, are present in both *T. cruzi* strains but absent in *T. conorhini* and *T. rangeli*. Several bacterial-type sugar kinases (glucokinase, galactokinase and L-ribulokinase), which contain targeting signals for import into glycosomes, are encoded in all four of the genomes of this study. However, genes for many other sugar metabolism enzymes, i.e. beta-glucosidase,

fructuronate reductase, xylulokinase, and mannitol 2-dehydrogenase are only present in *T. conorhini* and *T. rangeli*. Beta-glucosidase genes, for example, convert glucoside to α–D-glucose, and cellulose derivatives cellobiose and 1,4-β-D-Glucan to β –D-Glucose. These observations are consistent with the hypothesis that the latter two parasites exist in an environment higher in exogenous sugars and complex carbohydrates, an adaptation inconsistent with replication in the glucose-rich bloodstream. A nutritional role of plants in triatomines appears possible given demonstration of *Rhodnius* phytophagy [195]. Adaptation to vector diet may therefore have played an important role in the evolution of these species.

Overall metabolic potential

The metabolic potentials of *T. rangeli* and *T. conorhini* are more similar to each other than either is to *T. cruzi*. Each has around 20 differences in enzyme presence/absence compared to *T. cruzi*. All have complete pathways for glycolysis/gluconeogenesis, mannose metabolism, and pyruvate metabolism, although D-lactate dehydrogenase genes are absent in the *T. cruzi* strains. Glyoxylate and dicarboxylate metabolism appears deficient in all species, since isocitrate lyase and malate synthase, the two enzymes characteristic of the glyoxylate cycle, are absent. Interestingly, CAAX prenyl protease 1 (STE24 endopeptidase), presumably a membrane-associated protein [196] involved in terpenoid backbone synthesis, is present in the *T. cruzi* strains, but absent in *T. rangeli* AM80 (and also the *T. rangeli* SC-58 assembly of TriTrypDB v.24) and *T. conorhini* 025E. This gene is widely conserved in eukaryotes and highly diverged from CAAX prenyl protease 2, suggesting lack of redundancy. Terpenoids are

precursors of steroids and sterols, possibly suggesting a role in host-parasite interaction [197]. Several genes common to *T. rangeli*, *T. conorhini* and the *T. cruzi* strains i.e. genes encoding galactokinase, glutamate dehydrogenase (NADP), serine acetyltransferase and l-ribulokinase and 2-aminoethylphosphonate-pyruvate aminotransaminase (AEP transaminase) may be derived from horizontal gene transfer, and are absent in *T. brucei* [14]. Genes for aminoethylphosphonate (AEP) offer an alternative to ethanolamine phosphate for linkage of mucins to their GPI anchors. Enzymes for the synthesis of AEP from phosphoenol pyruvate are conserved in all four genomes.

Positive selection in multigene families

As previously observed with positive selection analysis using trypanosomatid comparisons [198], single copy ortholog genes (SCOs) across the species analyzed herein showed very little evidence of “gene-wide” selection (data not shown). I thus chose the approach of analyzing clusters of paralogous genes (CPGs), as duplicated genes are more likely to be under relaxed purifying selection and multigene families involved in host-parasite interactions are likely to be under positive selection.

The goals for this analysis were to (i) characterize positive selection in pairwise comparisons from within CPGs belonging to the major multigene families of these organisms (ii) determine site-specific positive selection in multigene families.

Gene-wide positive selection measures and site models were employed to achieve these goals.

(i) “Gene-wide” measures of positive selection: yn00 package of PAML [199].

This is based on taking a dN/dS measurement from pairwise gene alignments from each CPG, with dN and dS from alignments defined as follows:

$$dN = \frac{No.\ non-synonymous\ substitutions}{No.\ non-synonymous\ sites}$$

$$dS = \frac{No.\ synonymous\ substitutions}{No.\ synonymous\ sites}$$

with dN/dS>1 indicating positive selection.

(ii) Site models: codeml package of PAML.

This approach involves implementing site tests on each codon of a multiple sequence alignment, to find whether the genes in the alignment are likely to be under positive selection. Likelihood ratio test (LRT) statistics are produced for the alignment together with p-values. M1 (neutral) vs M2 (selection) and M7 vs M8 are two of the principal LRT tests. M7 specifies a neutral model with dN/dS ratios across a continuous beta distribution, i.e. dN/dS values fall between 0 and 1, and M8 specifies the beta distribution but also allows for sites that have dN/dS > 1, indicating positive selection. dN/dS ratios can also be obtained per site to estimate the specific sites under selection.

Based on estimated gene copy numbers, 11 of the largest multigene families (MGFs) across all species were chosen. OrthoMCL [85] v.2 was used to select clusters of paralogous genes (CPGs) for each species. Each gene in all clusters was assigned an annotation based on the best TriTrypDB hit. Each cluster was then kept if the most frequent annotation held by the genes was one of the 11 MGFs and there were at least 3 genes with that same annotation. 384 CPGs were found, whose amino acid sequences were aligned with clustalo [108] v.1.2. PAL2NAL [116] v.14 was then used to obtain the corresponding codon alignments, which were edited by Gblocks [109] v.0.91b with parameters –t=c, -b4=6, -b5=n, keeping only those alignments >300 nt (62 excluded) with no more than 50% of positions filtered out (0 excluded). These were assessed for positive selection via the yn00 program of the PAML package [199] v.4.8. All alignments with 4 or more sequences (163 CPGs) were additionally assessed for selection via the PAML program codeml. RAxML [113] v.8.1.17 was used with the GTRGAMMA model, building 50 maximum likelihood (ML) trees on distinct randomized stepwise addition parsimony starting trees to obtain the tree with the best likelihood. Based on these tree topologies and the alignments, LRT statistics for positive selection were inferred as described previously [198]. P-values were corrected for multiple testing in Rstudio v.0.98 using the “BH” method.

T. cruzi CL had many more CPGs overall, yet I find evidence of positive selection across many of the multigene families for all these species. I propose that positive selection is occurring widely in duplicated genes in these species, and thus the number of gene duplication events and CPGs, or percent of sites under selection pressure, may be the main differentiating factor in their adaptive abilities. In *T.*

rangeli AM80 enough sequences were present to carry out yn00 analysis of amastin genes, which indicated six pairwise gene comparisons with $d_N/d_S>1$ (Table 11). Given that the function of this gene in *T. rangeli* is unknown, the selection pressure present is intriguing.

Table 11: Pairwise positive selection across multigene families. Dashes indicate insufficient genes for alignment to perform the analysis.

	T. conorhini 025E			*T. rangeli* AM80			*T. cruzi* G			*T. cruzi* CL		
	dN/dS MEAN	dN/dS MEDIAN	# pairs	dN/dS MEAN	dN/dS MEDIAN	# pairs	dN/dS MEAN	dN/dS MEDIAN	# pairs	dN/dS MEAN	dN/dS MEDIAN	# pairs
amastin	-	-	-	1.55	1.51	6	-	-	-	-	-	-
beta	-	-	-	-	-	-	-	-	-	0.80	0.83	10
cruzipain	-	-	-	-	-	-	-	-	-	1.20	1.08	6
DGF-1	-	-	-	-	-	-	0.28	0.24	73	0.42	0.44	84
GP63	0.51	0.50	82	1.00	0.91	15	1.44	1.24	3	0.93	0.81	368
MASP	1.26	1.26	9	-	-	-	1.35	1.18	6	1.45	1.32	252
mucin	-	-	-	-	-	-	1.40	1.22	17	1.34	1.29	355
mucin-like	1.44	1.07	34	-	-	-	-	-	-	-	-	-
RHS	0.63	0.54	181	-	-	-	1.05	1.03	36	1.12	0.96	556
TASV	1.34	1.19	10	-	-	-	-			1.15	1.15	2
TS	0.80	0.73	47	0.96	0.97	323	0.98	0.94	107	0.95	0.83	620
UDP	-	-	-	-	-	-	-	-	-	0.42	0.42	94

Sufficient CPGs were present to allow percent of sites under selection from codeml site tests to be compared across all the species for trans-sialidase (Fig. 17). For this gene all species contain CPGs with >2% selected sites ($d_N/d_S>1$), although *T. rangeli* AM80 and *T. cruzi* CL each have a CPG containing ~7% selected sites, which is the maximum observed in sites with M2 and M8 agreement and posterior probability ≥0.95 in both models.

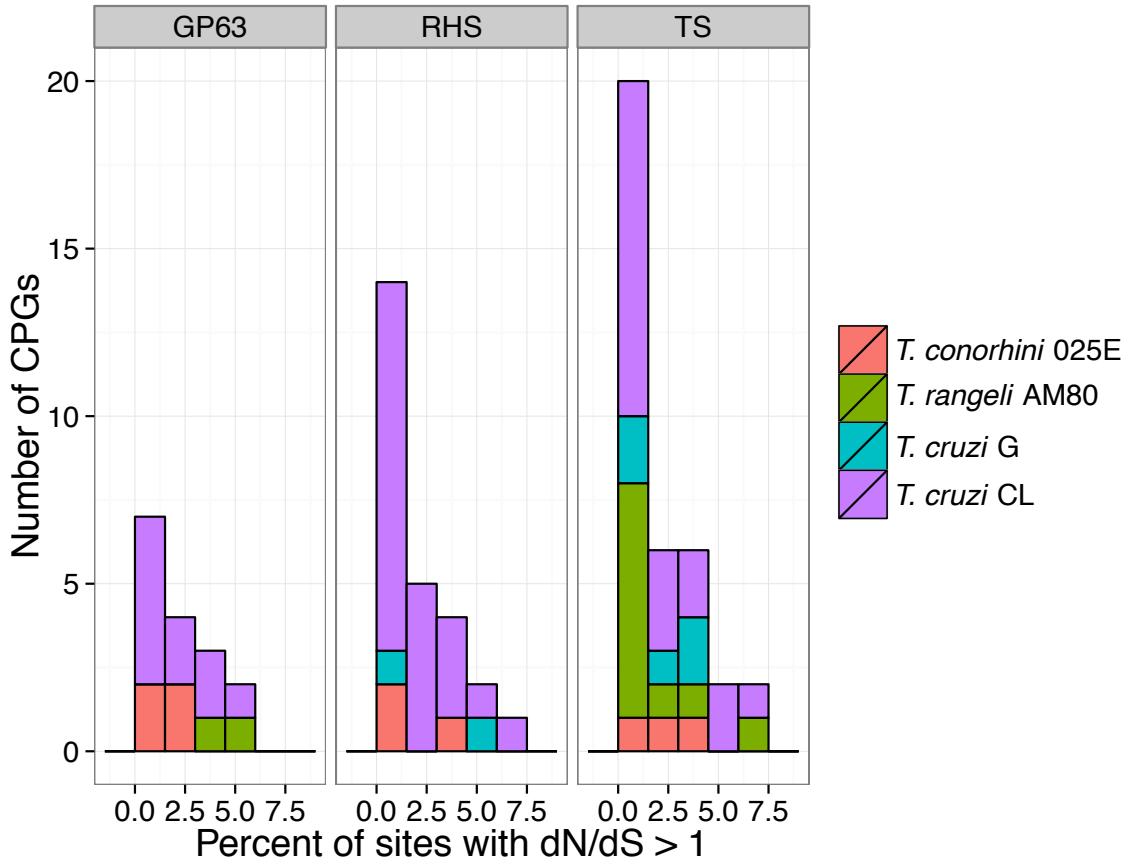


Figure 17: Sites under positive selection for GP63, RHS and trans-sialidase (TS) multigene families. CPGs where all three LRT statistics indicate positive selection at P<0.05 were chosen, and sites counted if both M2 and M8 models indicate a posterior probability ≥0.95.

The paralogous clusters of genes in multigene families shown here appear to contain as many as 7 % of sites under positive selection, with some even displaying gene-wide positive selection such as amastin in *T. rangeli*. Aside from this, MASP and mucin paralogous gene clusters of *T. cruzi* CL often displayed even higher percentages of sites under positive selection (not shown). However, one of the issues with this approach is that clustering the genes of a multigene family into paralogous groups is somewhat arbitrary, and in some cases insufficient alignments are produced for analysis due to vast differences across sub-groups or lack of sufficient numbers of genes in the assembly.

Conclusions

The genomes of *T. cruzi* strains G and CL, and the closely related species *T. rangeli* AM80 and *T. conorhini* 025E range from ~30-60 Mbp and contain between 10,000 and 24,000 genes. Characterization of the genomes in this work included multigene families, the heterozygosity, and pseudogene content, and used multi-gene strategies to explore their phylogenetic relationships. The results herein show that *T. cruzi* strains have more complex genomes with significantly more genes, a greater representation of multigene families, and more pseudogenes, than *T. rangeli* or *T. conorhini*. These observations generally are consistent with the more complex lifestyles of *T. cruzi* relative to the other parasites. Genes and gene families, including amastin, MASP, and DGF-1, and others, are represented in these parasites in ways that support their association with pathogenicity, intracellular life cycle and host range. The metabolic potentials of these organisms provide clues as to the basis of these biological capabilities, with *T. rangeli* and *T. conorhini* bearing a greater number of enzymes for utilizing complex carbohydrates and glycerol as carbon sources, and displaying highly divergent amino acid metabolism to *T. cruzi*. Based on these observations it seems likely that *T. rangeli* and *T. conorhini* are more adapted to life in the insect vector than their mammalian hosts. Heterozygosity levels suggest greater allelic diversity in *T. rangeli* AM80 and *T. cruzi* CL than in *T. conorhini* 025E and *T. cruzi* G. Phylogenetic distance in substitutions per site between the *T. rangeli* strains SC-58 and AM80 is about the same as *T. cruzi* strains to *T. c. marinkellei*, and the distance of *T. rangeli* AM80 to the *T. cruzi* strains is just over twice the distance between *T. rangeli* AM80 and *T. conorhini* 025E.

Chapter II. Genome architecture

Given the fragmented nature of *Trypanosoma* NGS assemblies, a complete assessment of genome architecture has not been possible using sequence data alone. As described above, *T. cruzi* G has a smaller, less heterozygous, and seemingly less repetitive genome than *T. cruzi* CL, and thus seems to present a simpler case for initial studies of genome-wide architecture in *Trypanosoma* species. To clarify the chromosome structure and genomic architecture of *T. cruzi* G, I employed BioNano, Inc. (San Diego, CA) genome mapping technology [200] to construct a whole-genome map of *T. cruzi* G. This microcapillary based optical mapping technology permits *de novo* probing of the genome architecture of an organism, in the absence of other cytogenetic or genome sequence data [201-206]. This analysis strategy generates high-resolution optical maps of molecules hundreds to thousands of Kbp.

As described above, some of the interesting aspects of *T. cruzi* that motivate studies of its genome architecture are:

- High intra-strain heterogeneity is observed in karyotypes
- The hybridization history of the *T. cruzi* DTUs in uncertain
- Mechanisms of genome rearrangements are unresolved
- Larger areas of sequence homology has been observed across bands during karyotype analysis and marker gene hybridization

The aims of this analysis were to (i) perform genome mapping to determine overall chromosomal structure (ii) elucidate relationships across chromosomes (iii) define major novel structural features of the genome.

Further *de novo* sequencing and assembly of *T. cruzi* G

Before genome mapping, the genome of *T. cruzi* G was re-sequenced and re-assembled using high coverage Illumina technology to obtain larger scaffolds. These were intended be large enough to align to genome maps, thus providing one form of validation and allowing functional assignments to regions of interest.

Parasites were obtained from Nobuko Yoshida (Universidade Federal de São Paulo), cultured and purified, and DNA was isolated and sequenced essentially as previously described [207]. Briefly, epimastigote form parasites were cultured at 28°C in liver-infusion tryptose (LIT) medium [208], supplemented with 20% fetal bovine serum (FBS) with 10 ug/ml hemin, and harvested in log phase at 1 X 10^7/ ml. Total DNA was isolated, and depleted of kinetoplast DNA (kDNA), by gel electrophoresis as previously described [207]. The purified DNA was then used to prepare 2 X 300 paired end and 2 X 300 mate paired reads on the Illumina MiSeq System, and 2 X 100 paired end reads on the HiSeq System for sequencing as indicated by the manufacturer. Adapters were removed from paired end reads using AdapterRemoval [209], and from mate paired reads using Nxtrim [210]. MiSeq reads were trimmed with meeptools [https://github.com/nisheth/meeptools] (meep=1) with minrl=90. HiSeq reads were trimmed with meeptools (meep=1) with minrl=75. Assembly was performed using the Newbler version 2.9 assembler (Roche, Inc.),

which limits the size of scaffolds to a minimum of 2 Kbp. Hence, all contigs larger than 500 bp that were not part of any scaffold were appended to the scaffolded assemblies for completeness. Reads were then realigned to the final assemblies using BWA [75] and an average genome-wide coverage of 469X was calculated. Genome size was estimated by dividing total number bases in reads by peak coverage of called genes, since the majority of assembled genes are thought to be single copy (Figure 3). To validate the assembly tBLASTn against *T. cruzi* CL Brener sequences was performed to determine the presence of a previously curated set of 2,217 kinetoplastid orthologous single copy genes. 98% of these genes were present at over 90 % alignment length.

General genome characteristics (Table 12) were determined using an in house Genome Annotation Pipeline (GAP). Briefly, genes were called using GeneMarkS v.4.6b [76]; tRNAscan-SE [77] v.1.23 was used to detect tRNAs; and 5S/18S/28S sequences were detected using RNAmmer [78] v.1.2. BLAST [82] against TriTrypDB v.24 and NCBI's nr databases was performed to determine validity and integrity of the gene calls, and ascertain probable gene functions and inferred annotations. Repeat counts in intergenic regions of the assemblies were identified by performing a Cross Match v. 0990329 search and categorization with Repeat Masker [167] v. 4.0.6. Repeat Masker library sequences of *Trypanosoma* species derived from Repbase (20150807 download) were used as a database for the search.

Table 12: NGS summary statistics for *T. cruzi* G assembly #2

Haploid Genome Size (mbp, # bases in all contigs)	23.5
Haploid Genome Size (mbp)[a]	56.98
Number of contigs	776
N50 length (mbp)	0.21
Longest contig length (mbp)	0.92
Shortest contig length (kbp)	0.5
Average coverage	469X
GC Content (%)	48.8
Coding Region (mbp)	14.1
Coding Region (%)	60
Number of Genes	10,105
Average intergenic distance (kbp)	0.93
Total repetitive content (%)	5.23
Intergenic repetitive element content (%)	4.23

[a]Total bases in reads / peak coverage of all genes

Differences in genome size estimates were observed between the 454 assembly described above (49 Mbp), and this Illumina assembly (57 Mbp), likely due to different biases in reads coverage across the two assemblies.

Data generation

In brief, nickase and polymerase enzymes incorporate fluorescent markers at specific restriction sites in genomic DNA. The gDNA is then strained and linearized, then passed through nanochannels on a chip and visualized. Each visualized molecule is then aligned and assembled into chromosome-sized maps (Fig. 18). This process is described in more detail below.

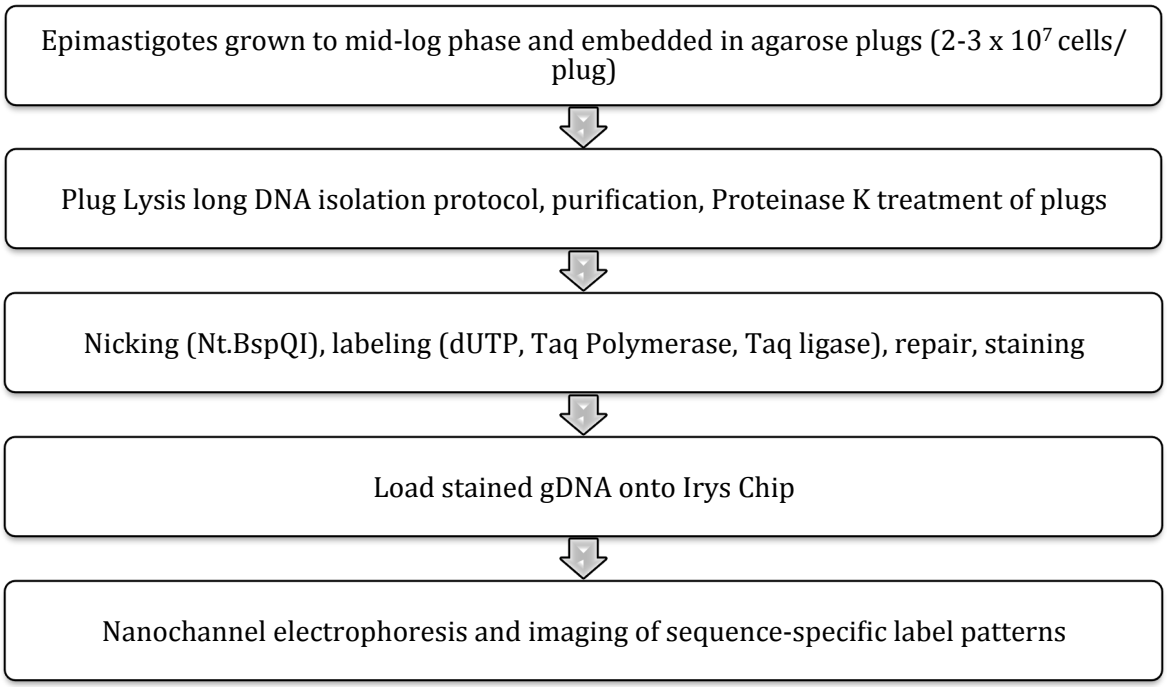


Figure 18: Sample preparation steps for BioNano genome mapping

Trypanosoma cruzi strain G epimastigotes were grown as previously described [208] up to a mid-log phase. Cells were harvested and embedded in agarose as plugs using the CHEF Genomic DNA Plug Kit (Bio-Rad), typically 2-3 x 10^7 cells per plug. Further processing of plugs containing *T. cruzi* G cells in order to obtain high molecular weight genomic DNA was performed according to the IrysPrepTM Plug Lysis Long DNA Isolation protocol (BioNano Genomics); treatment of plugs with Proteinase K resulted in higher purity of the final DNA without additional fragmentation when extended to 24 hours with 2-3 changes of the enzyme.

Among available modified restriction enzymes, nickase Nt.BspQ1 (recognition site GCTCTTCN) was selected based on *in-silico* analysis of available *T. cruzi* G DNA sequence in order to obtain the optimal fragment labeling density. Nicking of the DNA with Nt.BspQI enzyme (New England Biolabs), as well as subsequent steps including labeling, repair and staining were performed according to the IrysPrep®

Labeling-NLRS Experienced User Card (BioNano Genomics). Labeled long DNA molecules were loaded onto an IrysChip to be untwined and stretched on silicon wafer nanochannels through a gradient of micro- and nano- structures using electrophoretic forces. Thus linearized DNA molecules were imaged and high-resolution raw images were converted to digital representation of sequence-specific labeling patterns. Twenty-seven scans of a single flowcell on one chip generated 70,377 visualized molecules larger than 150 Kbp that were used as input for genome map assembly (Table 13). The total length of summed molecules was 17.5 Gbp. The average number of labels per 100 Kbp on molecules was 11.31, with a molecule average length of 249 Kbp. The largest molecule spanned 1.4 Mbp.

Table 13: Data generation overview

Total scan count	27
No. flow cells	1
Total no. molecules > 150 Kbp	70,377
Summed length of molecules > 150 Kbp (Gbp)	17.5
Molecule average length (Kbp)	249
Largest molecule (Mbp)	1.4
No. molecules assembled (> 150 Kbp in length)	57,752

Genome map assembly

Single molecules larger than 150 Kbp in length, with greater then 8 labels per 100 Kbp were assembled *de-novo* using “Irys” commercial software from BioNano Genomics. The mean label density across all maps was ~11.5 sites per 100 Kbp, which permitted accurate assembly and map-to-map comparison while limiting occurrence of ‘fragile sites’ caused by proximally located nicking sites on opposite

DNA strands. P-values cut-off thresholds were 1E-7 for initial assembly, and 1E-8 for extension and refinement. In order to have an independent assembly, the *de-novo* assembly was performed without using any reference. The software assembled 57,752 molecules into 66 consensus genome maps with a genome size of 44.075 Mbp, a map N50 of 0.884 Mbp, and an average depth of molecular coverage of 108.2 (Table 14).

Table 14: Genome map summary

Genome mapping	
Haploid Genome Size (mbp)	44
Number of maps	66
Map N50 (mbp)	0.88
Average coverage	108
Average label density / 100 kb	11.5
Total repetitive content (%)	2.6
Map size range (mbp)	2.4-0.13

The average coverage across all maps of ~108 fold, slightly exceeding the minimum 70-80 fold recommended for this technology [202,203], with an average mapped molecule size of 258 Kbp. A total of 66 maps were obtained, totaling ~44 Mbp. As described above, a new NGS assembly of *T. cruzi* G was performed using Illumina data, with a focus on obtaining longer contigs that I could align to genome maps using *in silico* restriction. Of the 66 assembled maps, 57 were readily aligned to next generation sequence contigs of 47 - 921 Kbp, with an average alignment confidence of 19.9. Conversely, over 80% of the sequences in largest NGS contigs (>300 Kbp) aligned to optical maps at an IrysView aligner confidence of 20 (Table 15). The false

positive and false negative labeling rates of molecules in comparison to the NGS reference were 2.8% and 7.6% respectively.

Table 15: Alignment of next-generation sequencing contigs to genome maps, ordered by contig size

NGS contig size	NGS contigs (count)	Summed NGS contig lengths (bp)	NGS contigs aligned to maps (count)	Summed alignment lengths in NGS contigs (bp)	Summed alignment lengths in NGS contigs / total (%)	Total labels in alignments (count)
> 100 kb	68	18113714	33	10727762	59	1962
> 150 kb	49	15672893	31	10473916	67	1919
> 200 kb	36	13360043	27	9727036	73	1769
> 250 kb	23	10582339	20	8211476	78	1508
> 300 kb	18	9163474	17	7641872	83	1385
> 350 kb	15	8205815	14	6640739	81	1214

The size estimates of the assembled optical maps ranged from 0.14 to 2.39 Mbp, the N50 was 0.88 Mbp, and the average map size was 0.67 Mbp. Figure 19 shows the sizes, average label density, and average coverage per map.

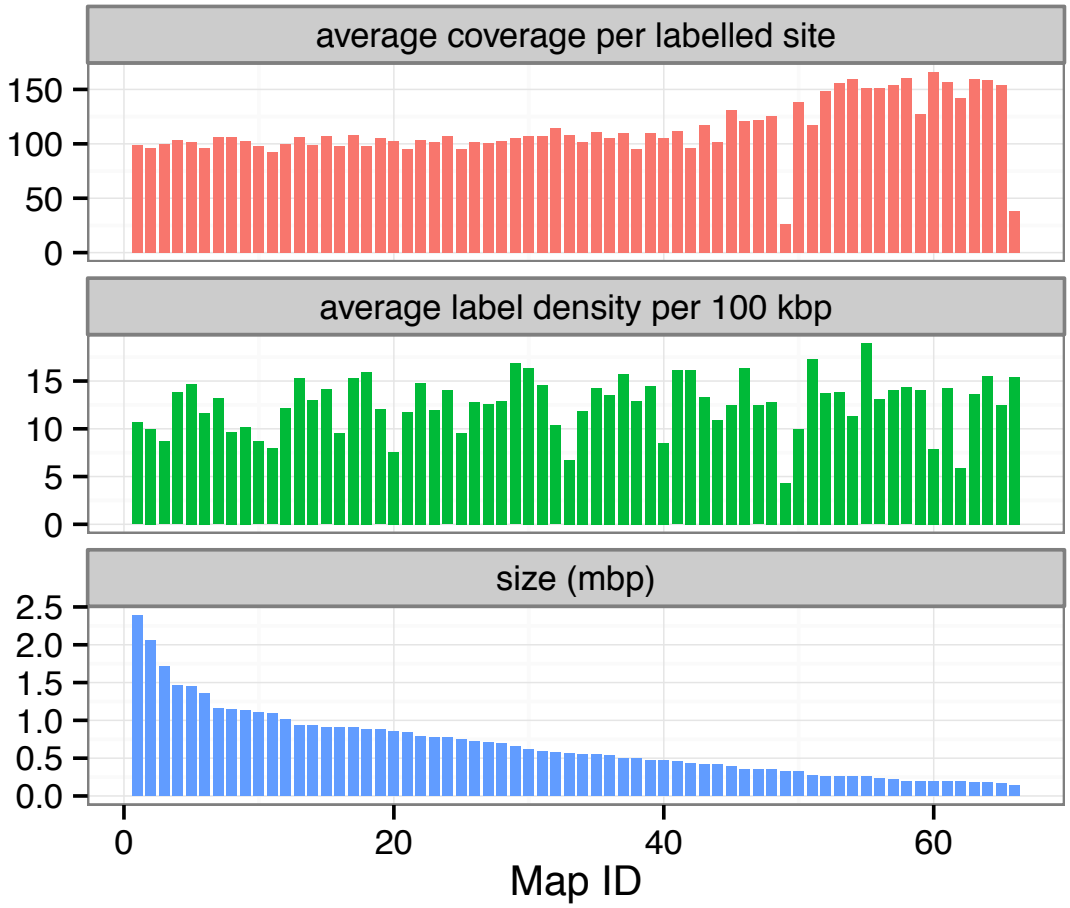


Figure 19: Genome map overview. The sizes, average label density, and average coverage in terms of mapped molecules for all 66 genome maps.

These results were consistent with the haploid genome size and number of chromosome bands observed in previously described pulsed field gel electrophoresis (PFGE) based molecular karyotype analyses from my data (Table 5) and previous studies [4,9]. I find ~18 chromosome bands and an estimated 37+ discrete chromosomes based on densitometry calculations from the pulsed field gels, including ~20 megabase-sized chromosomes (>1 Mbp) and ~10 intermediate chromosomes (300 Kbp – 1Mbp). A comparison of the optical and PFGE size estimates shows that the latter are slightly larger than the former. However, a general concordance between the PFGE and optical map size estimates suggests that many if not most of the maps represent full or nearly full-length chromosomes.

There are also 16 maps <300 Kbp present that are not evident in pulsed field gels. Mini-chromosomes of ~100 Kbp containing VSG genes are thought to be unique to *T. brucei* [211]. Only one map close to this size is observed in my analysis of the *T. cruzi* genome, which has a low level of coverage and is likely to be an assembly artifact. Overall, these maps apparently provide highly accurate representations of the *T. cruzi* chromosome structures.

Regions of map-to-map homology

The genome of *T. cruzi* G is primarily diploid [58,59] and its level of heterozygosity is low (~0.2% of bases) relative to levels in related *Trypanosoma* genomes (Fig. 11). Moreover, many *T. cruzi* isolates are hybrids that seem to have maintained a variable percentage of their genomes from their 'parental' strains [32], and molecular karyotype analyses of these strains have indicated significant chromosome size heterogeneity across strains [5]. Although *T. cruzi* G and other DTU1 strains display less heterozygosity, and are apparently more properly diploid than other *T. cruzi* strains, the molecular karyotypes of DTU1 strains vary significantly indicating a high degree of genomic flux [4,9,60]. Thus, it was not clear if the optical mapping strategy that was employed would resolve homologous chromosome pairs, or possibly chromosomes or fragments thereof that were derived from long ago hybridization events. Thus, I carefully examined the alignments in attempts to find evidence that the assembled optical maps represent homologous chromosome pairs, and to identify optical maps that show whole or partial homology.

I anticipated that homologous chromosome pairs would assemble as a single optical map. The haploid genome size estimates (44 Mbp) from summed optical map

lengths and PFGE densitometry (Table 5) [9], compared to the 57 Mbp genome size estimate provided by NGS (Table 12) are consistent with this interpretation. The average coverage of 64 of the 66 optical assemblies was very similar (Fig. 19), suggesting a similar level of ploidy for each of these chromosomes. However, probably because of the low level of heterozygosity observed in *T. cruzi* G, the molecules aligned to generate the optical maps did not segregate naturally into two groups that would be indicative of a true homologous chromosome pair. At an average label density of 11.5 sites per 100 Kbp, with 0.2% heterozygosity, and a 7 bp nickase site, only ~3 heterozygous sites per 2 Mbp are expected. This density was insufficient to determine ploidy.

Coverage of maps 49 and 66 was only 25-30% that of the average coverage (Fig. 19), suggesting that these DNA molecules are present at less than 1 copy per haploid genome and may be artifacts. Of the remaining assembled chromosome maps, only maps 10 and 16 align over nearly their full lengths (Fig. 20), indicating that these two chromosomes are similar and may be homologs, but represent haplotypes that are sufficiently divergent to assemble independently. However, the coverages of maps 10 and 16 are also essentially average, arguing that either these chromosomes have been duplicated and subsequently diverged or that these chromosomes are the remnants of a genomic hybridization event. I was able to identify significant alignments among many of the assembled chromosome maps (Fig. 20).

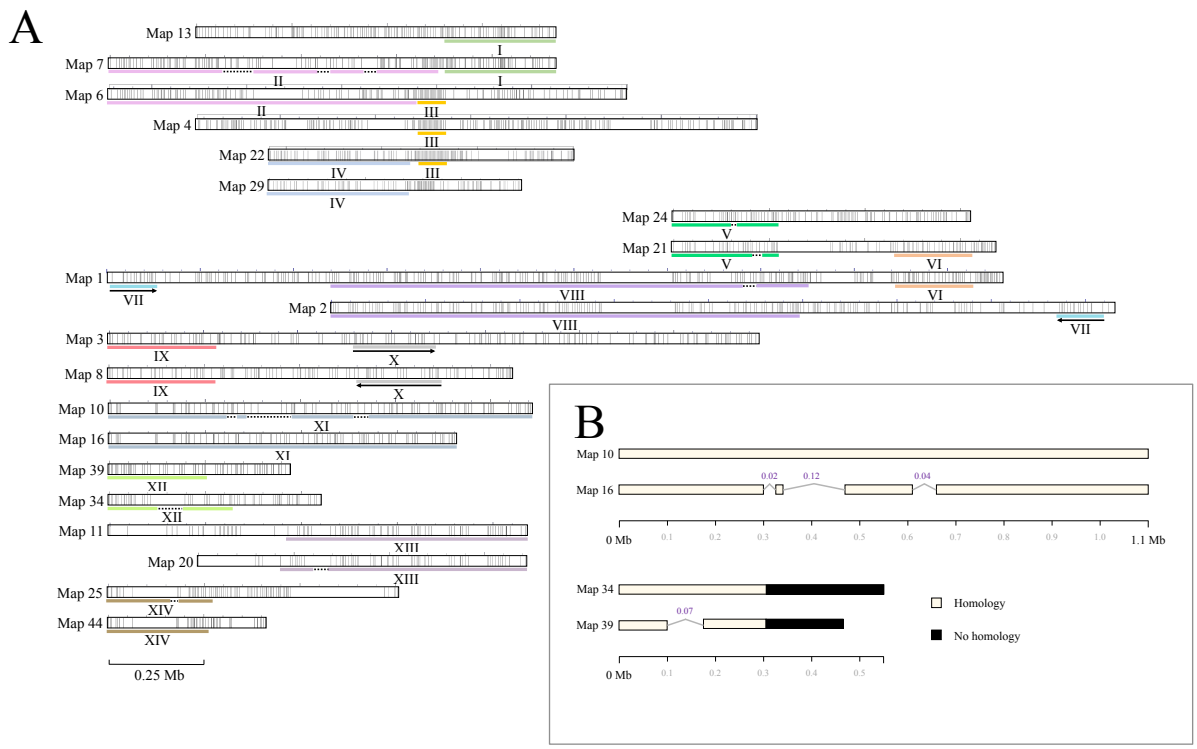


Figure 20: Homologous regions across genome maps of *T. cruzi* G. Homology is based on IrysView map-to-map alignment at confidence=20. (A) Regions of homology are indicated by Roman numerals and colored bars. Inversions and breaks between regions of homology are shown by arrows and dotted lines respectively. Maps without homology to at least one other map are not shown. Dotted lines (….) indicate absence of corresponding sequence to the sequence on the homologous map (referred to in the text as deletion/insertions). (B) Largest size differences in chromosomal regions of homology. *de novo* genome maps bearing regions of homology were aligned with IrysView (confidence for alignment set to 20). Maps containing size differences of >0.05 Mbp due to insertions, deletions or sequence expansions are shown.

Among 20 of the discrete genome maps there are 14 regions ranging from ~100 to ~1,250 Kbp with sufficient similarity to permit high confidence alignment to multiple genome maps (labelled I to XIV on Fig. 20). As described above, coverage of these maps (except maps 49 and 66, see above) is average with little variation over their lengths (not shown), consistent with each being present at a diploid level. These apparently scrambled chromosomes could be the result of segmental duplications

and translocations between different chromosomes and mitotic or meiotic nondisjunction or non-segregation within an ancestral *T. cruzi* G strain. Alternatively, the *T. cruzi* G strain could itself be a product of a hybridization between two ancestral *T. cruzi* strains which had previously undergone chromosome rearrangements followed by mitotic or meiotic non segregation. Interestingly, several of the apparently homologous regions in these maps; i.e. see regions II, III, IV,V and VIII, have endpoints proximal to SLRs, suggesting that the repetitive nature of the *T. cruzi* genome may be implicated in its apparent genetic instability.

Several of the *T. cruzi* G genome maps showing homology with other genome maps exhibit evidence of large insertions or deletions and inversions (Fig. 20). Homologous region XII, for example, present in maps 34 and 39 show clear evidence of an insertion/deletion of ~70 Kbp. It is not clear if this is a result of an insertion in map 34 or a deletion in map 39. Other homology regions; i.e., homology region II in maps 6 and 7, region V in maps 21 and 24, region VIII in maps 1 and 2, region XI in maps 10 and 16, region XIII in maps 11 and 20, and region XIV in maps 25 and 44, show similar insertion/deletion events. The largest of these deletion/insertions spanned ~120 Kbp.

Homology regions II in maps 6 and 7 and XI in maps 10 and 16 each showed three deletion insertion events (Fig. 20). Interestingly, the same maps (7 and 10) contained all three of the ‘deletions’ whereas the homologous map (6 and 16) contained all three of the ‘insertions’, possibly suggesting that the copies in maps 7 and 10 are experiencing negative selection.

Tandem repeats and “barren” regions

T. cruzi genomes, and to a lesser extent the genomes of other trypanosomatids, exhibit high levels of multicopy genes and repetitive sequences. Whole-genome sequencing of *T. cruzi* CL Brener has identified up to 50% repetitive DNA content [18], and previous studies using DNA reassociation kinetics suggest that the *T. cruzi* genome is highly repetitive [93,212]. The highly repetitive and tandemly organized 195 bp satellite DNA element alone may comprise about 9% of the *T. cruzi* genome [213-215]. Satellite DNA has been proposed to have a structural role in *T. cruzi* chromosomes, although the 195 bp element was found to be less frequent in TcI strains than other DTUs [216]. Intermediate repetitive sequences, such as tandem arrays of housekeeping genes, are also sometimes found in clusters on *T. cruzi* chromosomes [93,217]. Such repetitive sequences, and in particular highly repetitive tandem repeats, are often extremely difficult to resolve in short-read sequence assemblies.

Genome maps such as those described herein provide an alternative to identify, localize, and measure these sequence elements based on label distribution. Firstly, in order to check whether the overall label locations conformed to a random distribution label to label distances across all maps were compared to exponentially distributed random variates with the same mean, generated by the rtrunc function of the truncdist package of Rstudio v. 0.98 (Fig. 21). Both samples had a total of 5,000 data points. Left-truncation of both empirical and null distributions was applied at 1.5 Kbp, to avoid affects near the label resolution limit of ~1.5 Kbp previously established [200]. A two-sample, two-sided Kolmogorov-Smirnov (K-S) test was performed in Rstudio using the ks.test function.

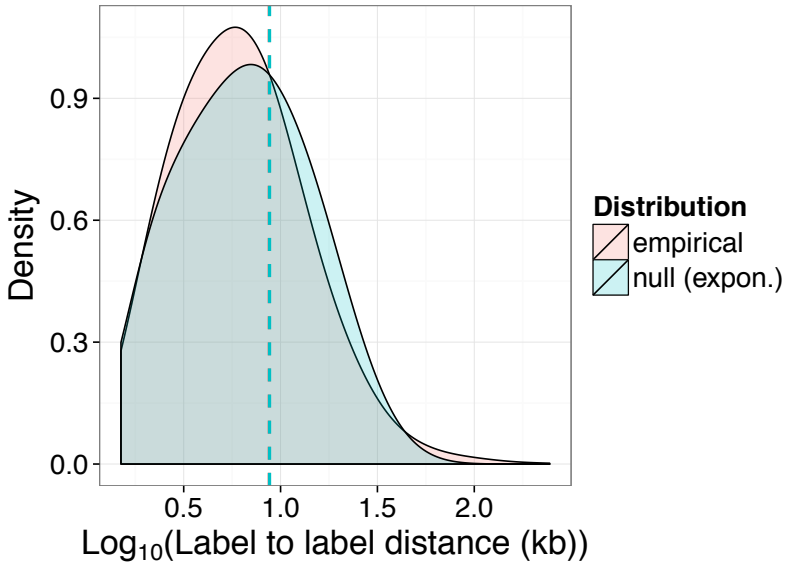


Figure 21: Label distribution. Label-to-label distances across genome maps of *T. cruzi* G compared to the null distribution, plotted on log10 scale. Random variates from the exponential distribution are plotted against observed label-to-label distances with the same mean (dashed vertical line) and lower bound for data points of 1.5 Kbp. A K-S test indicated that the distribution of label-to-label distances deviated significantly to that found at random (P = 8.663e-10).

The distribution of label to label distances in my analysis of the genome of *T. cruzi* G confirm the presence of a skewed distribution of sequence consistent with the influence of repetitive and/or other nonrandom sequences in the genome. Thus, Figure 21 shows plots of density curves of the label-to-label distance distribution across the *T. cruzi* G genome maps compared to the null exponential distribution. The results confirm the presence of a significantly biased representation of sequence in the genome maps; e.g., the maximum label-to-label distances in the null distribution and genome maps were 65 Kbp and 245 Kbp respectively. This is the result that would be expected for a genome bearing a high percentage of repetitive or otherwise less complex DNA sequence.

Single label repeats predicted by IrysView® GenomicAnalysis Viewer v.2.4.0.15879 were used for repeat statistics of molecules. Repeats coordinates on genome maps based on the corresponding aligned molecules were obtained with a custom Python script, with a molecule-to-map alignment threshold of confidence>40. Overlapping coordinates from molecule pile-up were merged using BEDTools v.2.19.1 merge. Barren region coordinates were located based on label-to-label distances greater than 65 Kbp. corr.test in Rstudio v. 0.98 was used to perform correlation analysis of repeat and barren regions to map length.

The genome maps reveal repetitive sequences, in particular single label repeats (SLRs, multiple consecutive equidistant labeled sites along a molecule or assembly), and long sequence segments barren of any labeled sites. Compound multi-label repeats may also be present in the maps but are not detectable due to their complexity and the technical capabilities of the IrysView technology. Thus, the representation of repetitive sequences in the *T. cruzi* genome herein may be underestimated. I found SLRs in 19,498 (20%) of the 98,211 molecules analyzed (Fig. 22). Approximately 2.7% of the sequence (~571 Mbp) in these molecules were present in SLR units ranging in size (distance between labels) from 1.25-33 Kbp. Interestingly, there seem to be two peaks in repeat unit size frequencies, one at ~3.25 Kbp and 6.25 Kbp, suggestive of a large family of similar repeat sequences (Fig. 22).

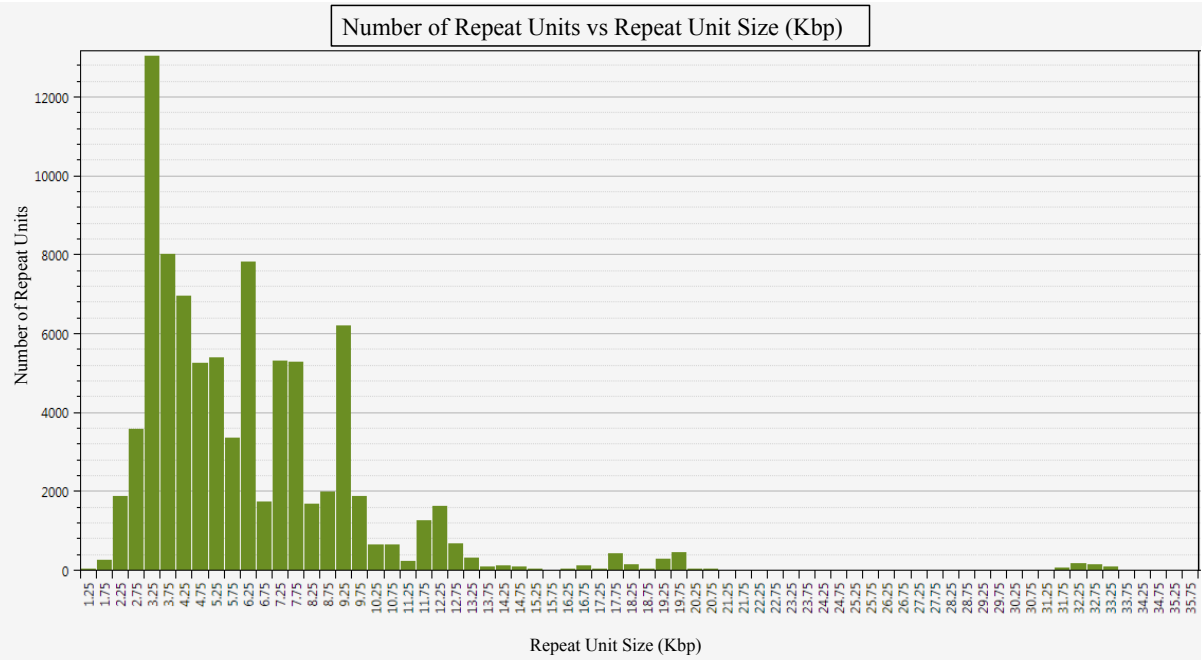


Figure 22: Frequencies of distinct repeat unit sizes on genome maps. Repeats are defined as at least three sequential labels with equidistant label-to-label distances. Repeat units are defined as one label-to-label distance within a repeat.

In the assembled genome maps, I identified ~28 SLR regions, representing ~2.6% (~1.13 Mbp) of the ~ 44 Mbp genome (Fig. 23). The SLR regions ranged from ~15 to ~89 Kbp. The average repeat region spanned ~43 Kbp, which explains the challenges they present to short-read sequencing technologies. Interestingly, SLR regions of ~70 Kbp on maps 4, 6 and 22 aligned using IrysView Software (Fig. 20), possibly suggesting common sequences and segmental duplication.

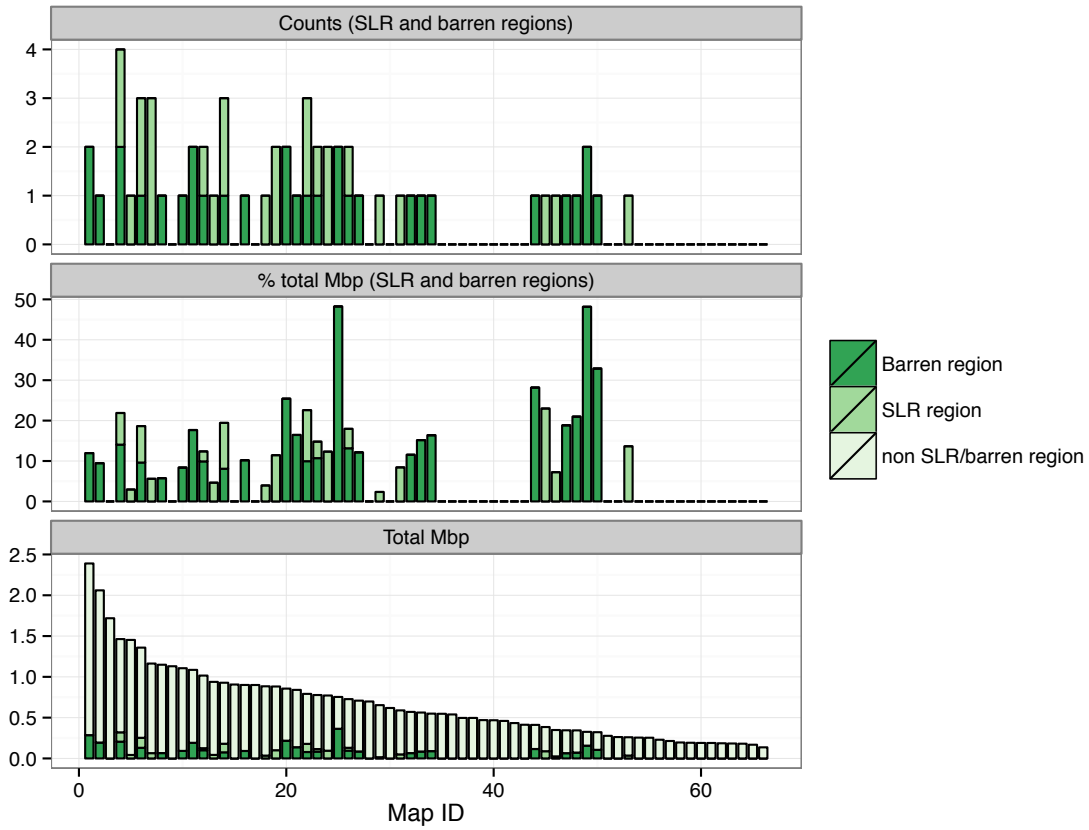


Figure 23: Single label repeat and barren region composition in genome maps. Barren regions are defined here as a label-to-label distance exceeding any length expected under the null distribution (65 Kbp).

Barren regions, defined as any region exceeding the maximum label-to-label distance expected in the null distribution (Fig. 21), and characterized by a lack of recognition sites for the restriction enzyme Nt.BspQ1, likely represent stretches of repetitive DNA in which the repeat unit lacks a recognition site or otherwise non-canonical sequences including homopolymeric runs or other non-random sequences. The distribution and frequency of these regions, together with single label repeat regions, was highly variable among the optical maps (Fig. 23). Two maps display barren regions over almost half of their length, and a third is ~25% SLRs. A higher content of repetitive DNA has been reported for larger chromosomes of *T. cruzi* [54], suggesting that map sizes would be positively correlated with content

of repetitive DNA. However, I found no evidence that increased map size is correlated with larger percent single label repeat content or barren regions. Despite this, I do observe differences in the presence or absence of these features based on map size. Approximately two-thirds of the ~32 optical maps smaller than ~500 Kbp lack both SLR repeat regions and barren regions, and the remaining third exhibit a higher percentage of sequence in SLR or barren regions. In contrast, all but 6 of the 34 maps greater than ~500 Kbp exhibit either SLR or barren regions, or both.

Table 16: Overview of barren, single label repeat, and map-to-map homology regions

	Mbp	% total map length (44 Mbp)	Count
Barren			
Total bp in barren regions	3.3	7.4	
Barren regions covered by NGS aln	0.5	1.1	
Total # barren regions w/ NGS aln	--		5
SLR			
Total bp in SLR regions	1.1	2.6	
SLR regions covered by NGS aln	0.3	0.7	
Total # SLR regions w/ NGS aln	--		12
Homology			
Total bp in homology regions	12.8	29.1	
Homology regions covered by NGS aln	4.5	10.3	
Total # homology regions w/ NGS aln	--		20

There are large areas of map-to-map homology (29% of the 44 Mbp genome), with possible implications on gene dosage (Table 16). Inference on hybridization history and genome rearrangements should also take these findings into account. Repetitive and barren regions are found on over two-thirds of the genome maps, covering 10% of the genome. Several megabases of the maps in these regions have an NGS contig alignment, indicating that further work to define the implications of the genes

present in these regions could yield interesting insights into the biology of these parasites.

T. cruzi genome architecture discussion

Assuming an almost complete genome reconstruction based on assembly of 44 Mbp into 66 genome maps, I sort to obtain an unbiased view of the chromosomal architecture of *T. cruzi* G via intra- and inter-map label distribution analysis. The distribution of labeling sites suggested combined content of SLR and barren regions among genome maps varied from complete absence to ~50 %. Although I cannot currently fully characterize the functions of these regions, I propose they may be (i) tandem arrays of housekeeping genes or repetitive gene families, or (ii) features with atypical sequence composition that have a functional role in chromosomal activity (e.g., mitosis, meiosis, etc.), with many of the smaller maps representing chromosomal fragments that have lost this function, possibly accounting for some of the instability or variability of chromosome content in these organisms. Better sequencing assemblies will be required to further characterize these sites and their potential functions.

The inter-map comparisons described here revealed large areas of homology among heterologous genome maps, supporting previous findings from gel hybridization [4,9] but providing much greater comprehensiveness. It is possible if not probable that genomic fragments that have apparently been duplicated in this genome contain genes for which a gene dosage effect confers a selective advantage to the organism. A more in depth analysis of these rearrangements awaits a more accurate alignment of these chromosome maps with accurate genome sequence

assemblies. However, it is clear from my observations that the *T. cruzi* genome has undergone significant genetic duplication, rearrangement and selective loss. Array-based comparative genomic hybridization revealed segmental aneuploidies including a 500 Kbp genome fragment bearing several genes in TcChr39 in a Brazilian TcI strain [70]. Putative segmental duplication accompanied by translocation was also implicated in *T. cruzi* CL Brener, SO3-cl5 and Y strains, as Chr7 markers were found on other chromosomal bands [4]. The model of interchromosomal exchange of large segments of DNA was also proposed for the G strain and its clone D11 [9], where TcChr37 markers covering a 1.1 Mbp region in the genome of clone D11 were found in two chromosomal bands in the G strain [4]. Similarly, on the heterologous genome maps of >2 Mbp; i.e., maps 1 and 2, I observe a region of homology spanning around 1.25 Mbp. I observe putative segmental duplications and translocations on several of the optical maps of the *T. cruzi* G genome (see in particular regions I, II, IV, V, VI, VII, IX, XII, XIII and XIV on Fig. 20). Assuming each of these individual maps represent two homologous chromosomes, each of these duplicative translocations seems to have involved translocations of telomeres. This observation, coupled with the observation that at least some of these translocations have end points proximal to repetitive sequences, suggests that a single homologous recombination event or gene conversion initiated by the repeat sequences may represent the mechanism by which at least some of these genome rearrangements have occurred.

I also considered the possibility that some of the maps may represent size-polymorphic homologous chromosomes. Differently sized homologous chromosomes within a single *T. cruzi* [5,54,58,67-69] or *T. brucei* [53,55,218] strain have previously been described. The majority of chromosome homologs differ in size

by <15% in *T. brucei*, and are due in large part to VSG expression site expansion/contraction [55]. In *T. cruzi*, this size polymorphism seems to widely affect the coding regions, rather than being confined to telomeric sequences, as for *T. brucei* [55,69]. TcI strains Sylvio X10 and DM28c have been reported to have very few homologous chromosomes that differ by > 75 Kbp (1 and 3 chromosomes respectively) suggesting few or mostly short size differences in homologous chromosomes in each of these strains [56]. Maps 10 and 16 exhibit broad homology across their entire 0.9-1.1 Mbp lengths, although there are three separate regions where insertions/deletions of 20, 120, and 40 Kbp are apparent (Fig. 20). Maps 1 and 2, and maps 6 and 7 also exhibit a high degree of homology across their lengths. It is possible that these six maps; i.e., 1 and 2, 6 and 7, and 10 and 16 do represent homologous pairs in a diploid organism. However, as discussed above, the empirically determined coverage of these maps is about the same (~108 fold) as it is for almost all of the other maps. Thus, it seems plausible that, as for the other maps, each of these maps represents a diploid pair but that these maps are the products of entire or nearly entire chromosomes that were derived from an ancestral hybridization event.

The widespread presence and size-polymorphism of segmentally duplicated or translocated regions, or apparently highly polymorphic members of homologous, and possibly duplicated, *T. cruzi* G diploid chromosomes, is consistent with the hypothesis that the *T. cruzi* genome exhibits extensive plasticity. The data are most consistent with the hypothesis that *T. cruzi* G chromosomes are diploid, but that some of the chromosomes; i.e., those that exhibit partial homology, are remnants of previous hybridizations or meiotic events during which the segregation and reduction phases were incomplete. In this scenario, each of the participants in the meiotic

event is assumed to have provided a complete set of chromosomes, but those chromosomes would have evolved and undergone rearrangements prior to the hybridization event. Although *T. cruzi* strains maintain genes required for meiosis [52], there is abundant evidence that hybrids exist in which diploid chromosomes are duplicated [120,121], suggesting that meiotic segregation events may be an imperfect event in these organisms. The result could be either a tetraploid organism, in which none of the chromosomes were properly segregated, or an aneuploid offspring in which a fraction of the chromosomes were tetraploid or reduced to triploid. Either way, presence of multiple copies of any chromosome, depending on the gene content of that chromosome, reduces the selective forces maintaining those genes, likely resulting in a higher rate of chromosome loss and/or a higher frequency of gene inactivation via mutation or rearrangement. Such a process could be responsible for generation of genomes like that of *T. cruzi* G, in which many of the diploid chromosomes bear partial homology with other diploid chromosomes. However, any hybridization or meiotic event that created the current *T. cruzi* G strain likely was a more ancient event in contrast to hybridizations that yielded other more clearly hybrid *T. cruzi* strains (e.g., Tulahuen or CL Brener [61,89,90,219]).

Of course, a canonical meiosis between two organisms with scrambled chromosomes would not be functional as most of the progeny would lack essential genes. In contrast, an imperfect meiosis, i.e., one in which segregation and reduction were not complete, between two related organisms with similar gene content but scrambled chromosomes would provide the opportunity for at least some viable progeny, in which some of the genetic material would be duplicated. Thus, chromosomes from both parental organisms might be present, ensuring that despite chromosomal rearrangement, all essential genes are present. It is possible therefore,

that *T. cruzi* and related organisms have evolved imperfect meiotic processes to permit sexual hybridization events between two distantly related organisms with rearranged chromosomes to generate viable progeny with new capabilities or altered, and possibly enhanced, capabilities due to gene dosage.

An obvious negative effect of large-scale loss of linkage or shuffling of the genome would presumably be the ensuing imbalance in gene dosage. In diploid organisms undergoing meiosis this would likely be selected against, and could be one reason why meiosis in *T. cruzi* lineages seems to be a rare event [63,71,122]. High clonality of *T. cruzi*, coupled with limited sexual exchange [63,122], has perhaps afforded the species greater freedom to reshuffle and gain/lose coding sequences across its genome. Maintenance of the co-direction of replication and transcription appears to be important however, and would seem to lead to selection against genome rearrangements such as those I have observed. Segmental duplication or CNV are perhaps permitted to a greater extent, as long as gene order and polycistronic transcription of nearby genes is not interrupted. Given the large areas of homology across heterologous chromosomes observed here, this seems plausible. Moreover, those non-viable recombinations are lost to history and we only see progeny of viable genetic exchanges.

Microsatellite analysis and serial cloning followed by PFGE and hybridization suggest that *T. cruzi* G actually has a monoclonal population [9], and *T. cruzi* strains appear to display karyotype stability during years of continuous cultivation [5,9,56]. Thus, a multiclonal population providing the observed diversity in otherwise homologous chromosomes of *T. cruzi* G is unlikely, although cannot be ruled out.

The analysis presented here provides a strong initial basis for analysis of the origins and mechanisms of chromosomal polymorphism among *T. cruzi* and related

organisms. Optical maps such as those reported here coupled with better whole genome sequence scaffolds will provide a functional context for the genomic regions bearing SLRs or barren regions, and a better understanding of the genes associated with the apparently duplicated chromosome segments identified herein. Analysis of the endpoints of these rearrangements may provide more insight into the mechanisms by which they were generated. Finally, identification of the genes associated with these events my provide insight into the selective pressures that led to their stabilization within the *T. cruzi* G genome and lead to a better understanding of the correlation between the genome composition of this organism, its phenotype, and its ability to invade, infect and cause disease.

T. cruzi genome architecture conclusions

This project has produced ~66 maps ranging from 130 Kbp to ~2.4 Mbp that represent chromosomes of *T. cruzi* strain G via assembly of optical maps from molecules 150 Kbp to 1.4 Mbp in length generated using the BioNano Irys® System. This approach has overcome many of the challenges associated with alternative short read technologies, which produce highly fragmented genomes for *Trypanosoma* species. The coverage of each map suggests that they each represent diploid chromosomes, but that many of these chromosomes bear homologous segments suggesting that the genome may be the result of a previous hybridization event, or large segmental duplications. This sheds new light on previous observations of putative large-scale regions of homology across chromosomal bands of *T. cruzi* G. This study shows for the first time the size ranges of these homologous regions across megabased-sized genomic regions, and demonstrates that these

mostly lie on otherwise heterologous maps. Some of the homologous regions observed seem to terminate in regions of repetitive sequence suggesting a mechanism of homologous exchange that may have yielded altered chromosomes, for example in the case of segmental duplication events. Plasticity and a high degree of rearrangements have occurred based on frequent insertions or deletions, which could also explain some of the chromosomes migrating as haploid bands in *T. cruzi* G PFGE analysis. There are also two putative inversions within regions of map-to-map homology. Higher resolution sequence analysis may reveal mechanisms for the events leading to these genomic structures and the selective forces that permitted them to be stabilized. The maps also reveal the repetitive nature of the genome, with some maps in particular displaying highly non-random label distributions.

Conclusions from comparative genomics and *T. cruzi* genome architecture analysis

Chapters I and II of this project contribute knowledge about the genomic features of *T. cruzi* and its closely related species *T. rangeli* and *T. conorhini*. Given the much more repetitive and obviously dynamic nature of *T. cruzi* chromosomes, based on analyses such as multigene family and retrotransposons expansion, and inter-strain karyotypes, we determined that genome mapping of *T. cruzi* would be valuable in explaining how this species has evolved these features. Large regions of homology were observed across the *T. cruzi* genome, suggesting that the genome is more complex than previously thought, and possibly the result of widespread segmental duplications or an ancient hybridization event. This sets the groundwork for comparative genome mapping analysis, and possible elucidation of the different mechanisms that have shaped the *T. cruzi* genome, and allowed it to become such a successful parasite and pathogen of mammals. Genomic comparisons revealed some aspects of *T. conorhini* and *T. rangeli* that possibly explain their lack of invasiveness and pathogenicity, such as reduced numbers of MASP, DGF-1 and GALFT genes compared to *T. cruzi*. On the other hand, we found some presence of *T. rangeli* amastin genes containing degenerate repeats, and overall diversity of gene family sequences of surface proteins (OrthoMCL and Fisher's Exact analysis), intriguing in terms of how they contribute to host cell immune response evasion and persistence in the mammalian host. The diversity between the recently published sequence of *T. rangeli* SC-58 and our *T. rangeli* AM80 strain highlights the fact that, like *T. cruzi*, *T. rangeli* strains can have many differences e.g. in terms of sequence identity and multigene family copy numbers. Geographic isolation of these strains

seems to be one factor in these observations, as well as evolution of divergent vector and mammalian host preferences, as explained in the literature review. The lack of pseudogenes in *T. conorhini* and *T. rangeli* compared to *T. cruzi* is also intriguing, suggesting that there may be an important role for pseudogenes in *T. cruzi*, such as to provide a repertoire for generating antigenic diversity in surface protein genes that are needed to evade the host immune response. The absence of retroelements in *T. rangeli* and *T. conorhini* is interesting, especially since *T. conorhini* still possesses RHS genes, whose orthologs in *T. cruzi* are known to enable retroelement integration. Whether retroelements are in the process of being lost or gained in this species is yet to be determined. Enrichment of cytoskeleton and lipid transport and metabolism genes in *T. conorhini*-specific gene clusters also raises questions as to how the lifecycle of *T. conorhini* may differ from *T. cruzi* or *T. rangeli*, since to date the full lifecycle of *T. conorhini* is unclear. For example, does higher diversity in genes involved in the cytoskeleton affect the morphology or possible migration of this parasite at specific lifecycle stages? This highlights the unique questions raised for each of these species, and provides a basis for developing studies to address the evolution of intracellular/extracellular lifestyles, genome rearrangements, and pathogenicity.

Literature Cited

Literature Cited

1. Quijano-Hernandez I, Dumonteil E. Advances and challenges towards a vaccine against Chagas disease. Hum Vaccin. 2011;7:1184–91.

2. Bustamante J, Tarleton R. Reaching for the Holy Grail: insights from infection/cure models on the prospects for vaccines for Trypanosoma cruzi infection. Mem. Inst. Oswaldo Cruz. 2015;110:445–51.

3. Messenger LA, Miles MA, Bern C. Between a bug and a hard place: Trypanosoma cruzi genetic diversity and the clinical outcomes of Chagas disease. Expert Rev Anti Infect Ther. 2015;13:995–1029.

4. Souza RT, Lima FM, Barros RM, Cortez DR, Santos MF, Cordero EM, et al. Genome Size, Karyotype Polymorphism and Chromosomal Evolution in *Trypanosoma cruzi*. Goldman GH, editor. PLoS ONE. 2011;6:e23042–14.

5. Henriksson J, Aslund L, Pettersson U. Karyotype Variability in *Trypanosoma cruzi*. Parasitology Today. 1996;12:1–7.

6. Franzén O, pez CT-L, Ochaya S, Butler CE, Messenger LA, Lewis MD, et al. Comparative genomic analysis of human infective *Trypanosoma cruzi* lineages with the bat-restricted subspecies *T. cruzi marinkellei*. BMC Genomics. BMC Genomics; 2012;13:1–1.

7. Franzén O, Ochaya S, Sherwood E, Lewis MD, Llewellyn MS, Miles MA, et al. Shotgun Sequencing Analysis of *Trypanosoma cruzi* I Sylvio X10/1 and Comparison with *T. cruzi* VI CL Brener. Bates PA, editor. PLoS Negl Trop Dis. 2011;5:e984–9.

8. JACKSON AP, Quail MA, Berriman M. Insights into the genome sequence of a free-living Kinetoplastid: *Bodo saltans* (Kinetoplastida: Euglenozoa). BMC Genomics. 2008;9:594.

9. Lima FM, Souza RT, Santori FR, Santos MF, Cortez DR, Barros RM, et al. Interclonal Variations in the Molecular Karyotype of *Trypanosoma cruzi*: Chromosome Rearrangements in a Single Cell-Derived Clone of the G Strain. Elias MC, editor. PLoS ONE. 2013;8:e63738–13.

10. Lukeš J, Skalický T, Týč J, Votýpka J, Yurchenko V. Evolution of parasitism in kinetoplastid flagellates. Molecular and Biochemical Parasitology. Elsevier B.V; 2014;195:115–22.

11. Clayton CE. Life without transcriptional control? From fly to man and back again. EMBO J. 2002;21:1881–8.

12. Fernandes AP, Nelson K, Beverley SM. Evolution of nuclear ribosomal RNAs in kinetoplastid protozoa: perspectives on the age and origins of parasitism. Proc. Natl. Acad. Sci. U.S.A. National Academy of Sciences; 1993;90:11608–12.

13. Hamilton PB, Teixeira MMG, Stevens JR. The evolution of *Trypanosoma cruzi*: the “bat seeding” hypothesis. Trends in Parasitology. Elsevier Ltd; 2012;28:136–41.

14. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome *Trypanosoma brucei*. Science. American Association for the Advancement of Science; 2005;309:416–22.

15. Carnes J, Anupama A, Balmer O, Jackson A, Lewis M, Brown R, et al. Genome and phylogenetic analyses of Trypanosoma evansi reveal extensive similarity to T. brucei and multiple independent origins for dyskinetoplasty. PLoS Negl Trop Dis. 2015;9:e3404.

16. JACKSON AP, Berry A, Aslett M, Allison HC, Burton P, Vavrova-Anderson J, et al. Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. Proc. Natl. Acad. Sci. U.S.A. 2012;109:3416–21.

17. Kelly S, Ivens A, Manna PT, Gibson W, Field MC. A draft genome for the African crocodilian trypanosome Trypanosoma grayi. Sci Data. 2014;1:140024.

18. El-Sayed NM. The Genome Sequence of *Trypanosoma cruzi*, Etiologic Agent of Chagas Disease. Science. 2005;309:409–15.

19. Stoco PH, Wagner G, Talavera-López C, Gerber A, Zaha A, Thompson CE, et al. Genome of the Avirulent Human-Infective Trypanosome—*Trypanosoma rangeli*. Kissinger JC, editor. PLoS Negl Trop Dis. 2014;8:e3176–17.

20. Borst P. Antigenic variation and allelic exclusion. Cell. 2002;109:5–8.

21. Pays E, Vanhamme L, Perez-Morga D. Antigenic variation in Trypanosoma brucei: facts, challenges and mysteries. Curr Opin Microbiol. 2004;7:369–74.

22. Desquesnes M, Dargantes A, Lai D-H, Lun Z-R, Holzmuller P, Jittapalapong S. Trypanosoma evansi and surra: a review and perspectives on transmission, epidemiology and control, impact, and zoonotic aspects. Biomed Res Int. 2013;2013:321237.

23. Schnaufer A, Domingo GJ, Stuart K. Natural and induced dyskinetoplastic trypanosomatids: how to live without mitochondrial DNA. International Journal for Parasitology. 2002;32:1071–84.

24. Garcia ES, Ratcliffe NA, Whitten MM, Gonzalez MS, Azambuja P. Exploring the role of insect host factors in the dynamics of *Trypanosoma cruzi-Rhodnius prolixus* interactions. J Insect Physiol. 2007;53:11–21.

25. Azambuja P, Garcia ES. *Trypanosoma rangeli* interactions within the vector Rhodnius prolixus: a mini review. Mem. Inst. Oswaldo Cruz. 2005;100:567–72.

26. DEANE MP, DEANE LM. Studies on the life cycle of *Trypanosoma conorrhini*. “In

vitro” development and multiplication of the bloodstream trypanosomes. Rev. Inst. Med. Trop. São Paulo. 1961;3:149–60.

27. Hoare CA. The Trypanosomes of Mammals. Blackwell Scientific Publications, Oxford and Edinburgh; 1972.

28. D 'Alessandro A, Saravia NG. *Trypanosoma rangeli*. In: Parasitic Protozoa. 2nd ed. Kreier J, Baker JR, editors. Academic Press, New York; 1992. pp. 1–54.

29. Fernandes MC, Andrews NW. Host cell invasion by *Trypanosoma cruzi*: a unique strategy that promotes persistence. FEMS Microbiol Rev. 2012;36:734–47.

30. Yeo M, Mauricio IL, Messenger LA, Lewis MD, Llewellyn MS, Acosta N, et al. Multilocus sequence typing (MLST) for lineage assignment and high resolution diversity studies in *Trypanosoma cruzi*. PLoS Negl Trop Dis. 2011;5:e1049.

31. Sturm NR, Campbell DA. Alternative lifestyles: the population structure of Trypanosoma cruzi. Acta Tropica. 2010;115:35–43.

32. Lewis MD, Llewellyn MS, Yeo M, Acosta N, Gaunt MW, Miles MA. Recent, independent and anthropogenic origins of *Trypanosoma cruzi* hybrids. PLoS Negl Trop Dis. 2011;5:e1363.

33. Rodrigues AA, Saosa JSS, da Silva GK, Martins FA, da Silva AA, Souza Neto CPDS, et al. IFN-γ Plays a Unique Role in Protection against Low Virulent *Trypanosoma cruzi* Strain. Milon G, editor. PLoS Negl Trop Dis. 2012;6:e1598–9.

34. Santos CC, Sant'anna C, Terres A, Cunha-e-Silva NL, Scharfstein J, de A Lima APC. Chagasin, the endogenous cysteine-protease inhibitor of *Trypanosoma cruzi*, modulates parasite differentiation and invasion of mammalian cells. Journal of Cell Science. 2005;118:901–15.

35. Yoshida N. Surface Antigens of Metacyclic Trypomastigotes of *Trypanosoma cruzi*. Infect. Immun. 1983;40:836–9.

36. da Silva CV, Kawashita SY, Probst CM, Dallagiovanna B, Cruz MC, da Silva EA, et al. Characterization of a 21kDa protein from *Trypanosoma cruzi* associated with mammalian cell invasion. Microbes and Infection. Elsevier Masson SAS; 2009;11:563–70.

37. da Silva CV, Luquetti AO, Rassi A, Mortara RA. Involvement of Ssp-4-related carbohydrate epitopes in mammalian cell invasion by Trypanosoma cruzi amastigotes. Microbes and Infection. 2006;8:2120–9.

38. Zago MP, Hosakote YM, Koo S-J, Dhiman M, Pineyro MD, Parodi-Talice A, et al. TcI Isolates of Trypanosoma cruzi Exploit the Antioxidant Network for Enhanced Intracellular Survival in Macrophages and Virulence in Mice. Infect. Immun. 2016;84:1842–56.

39. Grisard EC, Campbell DA, Romanha AJ. Mini-exon gene sequence polymorphism among *Trypanosoma rangeli* strains isolated from distinct geographical regions. Parasitology. 1999;118 (Pt 4):375–82.

40. Maia da Silva F, Rodrigues AC, Campaner M, Takata CSA, Brigido MC, JUNQUEIRA ACV, et al. Randomly amplified polymorphic DNA analysis of *Trypanosoma rangeli* and allied species from human, monkeys and other sylvatic mammals of the Brazilian Amazon disclosed a new group and a species-specific marker. Parasitology. 2004;128:283–94.

41. Ortiz PA, Maia da Silva F, Cortez AP, Lima L, Campaner M, Pral EMF, et al. Genes of cathepsin L-like proteases in *Trypanosoma rangeli* isolates: markers for diagnosis, genotyping and phylogenetic relationships. Acta Tropica. 2009;112:249–59.

42. Vallejo GA, Guhl F, Schaub GA. Triatominae-*Trypanosoma cruzi*/*T. rangeli*: Vector-parasite interactions. Acta Tropica. 2009;110:137–47.

43. Maia da Silva F, JUNQUEIRA ACV, Campaner M, Rodrigues AC, Crisante G, Ramirez LE, et al. Comparative phylogeography of *Trypanosoma rangeli* and *Rhodnius* (Hemiptera: Reduviidae) supports a long coexistence of parasite lineages and their sympatric vectors. Mol Ecol. 2007;16:3361–73.

44. Steindel M, Pinto JC, Toma HK, Mangia RH, Ribeiro-Rodrigues R, Romanha AJ. *Trypanosoma rangeli* (Tejera, 1920) isolated from a sylvatic rodent (*Echimys dasythrix*) in Santa Catarina Island, Santa Catarina State: first report of this trypanosome in southern Brazil. Mem. Inst. Oswaldo Cruz. 1991;86:73–9.

45. Grisard EC, Steindel M, Guarneri AA, Eger-Mangrich I, Campbell DA, Romanha AJ. Characterization of *Trypanosoma rangeli* strains isolated in Central and South America: an overview. Mem. Inst. Oswaldo Cruz. 1999;94:203–9.

46. Maia da Silva F, NOYES H, Campaner M, JUNQUEIRA ACV, COURA JR, Añez N, et al. Phylogeny, taxonomy and grouping of *Trypanosoma rangeli* isolates from man, triatomines and sylvatic mammals from widespread geographical origin based on SSU and ITS ribosomal sequences. Parasitology. 2004;129:549–61.

47. Maia da Silva F, Naiff RD, Marcili A, Gordo M, D'Affonseca Neto JA, Naiff MF, et al. Infection rates and genotypes of *Trypanosoma rangeli* and *T. cruzi* infecting free-ranging *Saguinus bicolor* (Callitrichidae), a critically endangered primate of the Amazon Rainforest. Acta Tropica. 2008;107:168–73.

48. Hamilton PB, Lewis MD, Cruickshank C, Gaunt MW, Yeo M, Llewellyn MS, et al. Identification and lineage genotyping of South American trypanosomes using fluorescent fragment length barcoding. Infection, Genetics and Evolution. Elsevier B.V; 2011;11:44–51.

49. Schmunis GA, Yadon ZE. Chagas disease: a Latin American health problem becoming a world health problem. Acta Tropica. 2010;115:14–21.

50. Martinez-Perez A, Poveda C, Ramirez JD, Norman F, Girones N, Guhl F, et al. Prevalence of Trypanosoma cruzi's Discrete Typing Units in a cohort of Latin American migrants in Spain. Acta Tropica. 2016;157:145–50.

51. Dujardin J-P, Lam TX, Khoa PT, Schofield CJ. The rising importance of Triatoma rubrofasciata. Mem. Inst. Oswaldo Cruz. 2015;110:319–23.

52. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, et al. Comparative genomics of trypanosomatid parasitic protozoa. Science. 2005;309:404–9.

53. Gibson WC, Miles MA. The karyotype and ploidy of *Trypanosoma cruzi*. EMBO J. 1986;5:1299–305.

54. Henriksson J, Porcel B, Rydaker M, Ruiz A, Sabaj V, Galanti N, et al. Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in Trypanosoma cruzi. Molecular and Biochemical Parasitology. 1995;73:63–74.

55. Melville SE, Leech V, Gerrard CS, Tait A, Blackwell JM. The molecular karyotype of the megabase chromosomes of Trypanosoma brucei and the assignment of chromosome markers. Molecular and Biochemical Parasitology. 1998;94:155–73.

56. Pedroso A, Cupolillo E, Zingales B. Evaluation of Trypanosoma cruzi hybrid stocks based on chromosomal size variation. Molecular and Biochemical Parasitology. 2003;129:79–90.

57. Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, et al. Gene synteny and evolution of genome architecture in trypanosomatids. Molecular and Biochemical Parasitology. 2004;134:183–91.

58. Frohme M, Hanke J, Aslund L, Pettersson U, Hoheisel JD. Selective generation of chromosomal cosmid libraries within the Trypanosoma cruzi genome project. Electrophoresis. 1998;19:478–81.

59. Henriksson J, Aslund L, Macina RA, Franke de Cazzulo BM, Cazzulo JJ, Frasch AC, et al. Chromosomal localization of seven cloned antigen genes provides evidence of diploidy and further demonstration of karyotype variability in Trypanosoma cruzi. Molecular and Biochemical Parasitology. 1990;42:213–23.

60. Vargas N, Pedroso A, Zingales B. Chromosomal polymorphism, gene synteny and genome size in T. cruzi I and T. cruzi II groups. Molecular and Biochemical Parasitology. 2004;138:131–41.

61. Westenberger SJ, Barnabé C, Campbell DA, Sturm NR. Two Hybridization Events Define the Population Structure of *Trypanosoma cruzi*. Genetics. 2005;171:527–43.

62. Reis-Cunha JL, Rodrigues-Luiz GF, Valdivia HO, Baptista RP, Mendes TAO, de Morais GL, et al. Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct Trypanosoma cruzi strains. BMC Genomics. 2015;16:499.

63. Tibayrenc M, Ward P, Moya A, Ayala FJ. Natural populations of Trypanosoma cruzi, the agent of Chagas disease, have a complex multiclonal structure. Proc. Natl. Acad. Sci. U.S.A. National Academy of Sciences; 1986;83:115–9.

64. Weatherly DB, Boehlke C, Tarleton RL. Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome. BMC Genomics. 2009;10:255–13.

65. Grisard EC, Teixeira SMR, de Almeida LGP, Stoco PH, Gerber AL, Talavera-López C, et al. Trypanosoma cruzi Clone Dm28c Draft Genome Sequence. Genome Announc. 2014;2.

66. Henriksson J, Pettersson U, Solari A. Trypanosoma cruzi: correlation between karyotype variability and isoenzyme classification. Experimental Parasitology. 1993;77:334–48.

67. Henriksson J, Dujardin JC, Barnabé C, Brisse S, Timperman G, Venegas J, et al. Chromosomal size variation in *Trypanosoma cruzi* is mainly progressive and is evolutionarily informative. Parasitology. 2002;124:277–86.

68. Cano MI, Gruber A, Vazquez M, Cortéz A, Levin MJ, González A, et al. Molecular karyotype of clone CL Brener chosen for the *Trypanosoma cruzi* Genome Project. Molecular and Biochemical Parasitology. 1995;71:273–8.

69. Santos MR, Cano MI, Schijman A, Lorenzi H, Vazquez M, Levin MJ, et al. The *Trypanosoma cruzi* genome project: nuclear karyotype and gene mapping of clone CL Brener. Mem. Inst. Oswaldo Cruz. 1997;92:821–8.

70. Minning TA, Weatherly DB, Flibotte S, Tarleton RL. Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in *Trypanosoma cruzi* strains revealed by array comparative genomic hybridization. BMC Genomics. BioMed Central Ltd; 2011;12:139.

71. Solari AJ. Mitosis and genome partition in trypanosomes. Biocell. 1995;19:65–84.

72. Cuba Cuba A. Review of the biologic and diagnostic aspects of *Trypanosoma* (*Herpetosoma*) *rangeli*. Rev Soc Bras Med Trop. 1998;31:207–20.

73. Brener Z, Chiari E. [MORPHOLOGICAL VARIATIONS OBSERVED IN DIFFERENT STRAINS OF TRYPANOSOMA CRUZI]. Rev. Inst. Med. Trop. São Paulo. 1963;5:220–4.

74. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS ONE. 2012;7:e30619.

75. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

76. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. Oxford University Press; 2001;29:2607–18.

77. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. Oxford University Press; 1997;25:955–64.

78. Lagesen K, Hallin P, Rodland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids

Res. 2007;35:3100–8.

79. Petersen TN, Brunak S, Heijne von G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6.

80. Krogh A, Larsson B, Heijne von G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305:567–80.

81. Fankhauser N, Maser P. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. Bioinformatics. 2005;21:1846–52.

82. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

83. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40:D290–301.

84. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4:41.

85. Li L, Stoeckert CJJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.

86. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, *Leishmania major*. Science. 2005;309:436–42.

87. Lewis MD, Llewellyn MS, Gaunt MW, Yeo M, Carrasco HJ, Miles MA. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in *Trypanosoma cruzi* populations and expose contrasts between natural and experimental hybrids. International Journal for Parasitology. 2009;39:1305–17.

88. Machado CA, Ayala FJ. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of Trypanosoma cruzi. Proc. Natl. Acad. Sci. U.S.A. National Acad Sciences; 2001;98:7396–401.

89. Brisse S, Henriksson J, Barnabe C, Douzery EJP, Berkvens D, Serrano M, et al. Evidence for genetic exchange and hybridization in *Trypanosoma cruzi* based on nucleotide sequences and molecular karyotype. Infection, Genetics and Evolution. 2003;2:173–83.

90. de Freitas JM, Augusto-Pinto L, Pimenta JR, Bastos-Rodrigues L, Gonçalves VF, Teixeira SMR, et al. Ancestral Genomes, Sex, and the Population Structure of Trypanosoma cruzi. J B, editor. PLoS Pathog. 2006;2.

91. Skosyrev VS, Vasil'eva GV, Lomaeva MG, Malakhova LV, Antipova VN, Bezlepkin VG. Specialized software product for comparative analysis of multicomponent DNA fingerprints. Genetika. 2013;49:531–7.

92. Henriksson J, Solari A, Rydaker M, Sousa OE, Pettersson U. Karyotype

variability in Trypanosoma rangeli. Parasitology. 1996;112 (Pt 4):385–91.

93. Castro C, Hernandez R, Castaneda M. *Trypanosoma cruzi* ribosomal RNA: internal break in the large-molecular-mass species and number of genes. Molecular and Biochemical Parasitology. 1981;2:219–33.

94. Martinez-Calvillo S, Sunkin SM, Yan S, Fox M, Stuart K, Myler PJ. Genomic organization and functional characterization of the Leishmania major Friedlin ribosomal RNA gene locus. Molecular and Biochemical Parasitology. 2001;116:147–57.

95. Agüero F, Verdún RE, Frasch ACC, Sánchez DO. A Random Sequencing Approach for the Analysis of the *Trypanosoma cruzi* Genome: General Structure, Large Gene and Repetitive DNA Families, and Gene Discovery. Genome Res. 2000;10:1996–2005.

96. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2010;38:D457–62.

97. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nat Genet. 2007;39:839–47.

98. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.

99. Atwood JA, Weatherly DB, Minning TA, Bundy B, Cavola C, Opperdoes FR, et al. The *Trypanosoma cruzi* proteome. Science. American Association for the Advancement of Science; 2005;309:473–6.

100. Caradonna KL, Engel JC, Jacobi D, Lee C-H, Burleigh BA. Host metabolism regulates intracellular growth of *Trypanosoma cruzi*. Cell Host Microbe. 2013;13:108–17.

101. Hannaert V, Bringaud F, Opperdoes FR, Michels PA. Evolution of energy metabolism and its compartmentation in Kinetoplastida. Kinetoplastid Biol Dis. 2003;2:11.

102. van Hellemond JJ, Tielens AGM. Adaptations in the lipid metabolism of the protozoan parasite *Trypanosoma brucei*. FEBS Letters. 2006;580:5552–8.

103. Miao Q, Ndao M. *Trypanosoma cruzi* Infection and Host Lipid Metabolism. Mediators of Inflammation. Hindawi Publishing Corporation; 2014;:1–10.

104. de Souza W, de Carvalho TMU, Barrias ES. Review on *Trypanosoma cruzi*: Host Cell Interaction. International Journal of Cell Biology. 2010;2010:1–18.

105. Maeda FY, Cortez C, Izidoro MA, Juliano L, Yoshida N. Fibronectin-Degrading Activity of *Trypanosoma cruzi* Cysteine Proteinase Plays a Role in Host Cell Invasion. Infect. Immun. 2014;82:5166–74.

106. McKerrow JH, Rosenthal PJ, Swenerton R, Doyle P. Development of protease inhibitors for protozoan infections. Current Opinion in Infectious Diseases. 2008;21:668–72.

107. Lima L, Ortiz PA, da Silva FM, Alves JMP, Serrano MG, Cortez AP, et al. Repertoire, Genealogy and Genomic Organization of Cruzipain and Homologous Genes in *Trypanosoma cruzi*, *T. cruzi*-Like and Other Trypanosome Species. Rodrigues MM, editor. PLoS ONE. 2012;7:e38385–15.

108. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol. 2014;1079:105–16.

109. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.

110. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7.

111. Kuck P, Meusemann K. FASconCAT: Convenient handling of data matrices. Molecular Phylogenetics and Evolution. 2010;56:1115–8.

112. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27:1164–5.

113. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90.

114. Stover BC, Muller KF. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics. 2010;11:7.

115. Inkscape. http://inkscape.org/, last accessed December 2015 [Internet]. Available from: http://inkscape.org/

116. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:W609–12.

117. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 1999;41:95–8.

118. Lima L, Espinosa-Alvarez O, Hamilton PB, Neves L, Takata CSA, Campaner M, et al. *Trypanosoma livingstonei*: a new species from African bats supports the bat seeding hypothesis for the *Trypanosoma cruzi* clade. Parasit Vectors. 2013;6:221.

119. Caballero ZC, Costa-Martins AG, Ferreira RC, P Alves JM, Serrano MG, Camargo EP, et al. Phylogenetic and syntenic data support a single horizontal transference to a *Trypanosoma* ancestor of a prokaryotic proline racemase implicated in parasite evasion from host defences. Parasit Vectors. 2015;8:115–8.

120. Branche C, Ochaya S, Aslund L, Andersson B. Comparative karyotyping as a tool for genome structure analysis of *Trypanosoma cruzi*. Molecular and Biochemical

Parasitology. 2006;147:30–8.

121. Obado SO, Taylor MC, Wilkinson SR, Bromley EV, Kelly JM. Functional mapping of a trypanosome centromere by chromosome fragmentation identifies a 16-kb GC-rich transcriptional “strand-switch” domain as a major feature. Genome Res. 2005;15:36–43.

122. Tibayrenc M, Ayala FJ. Reproductive clonality of pathogens: a perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. Proc. Natl. Acad. Sci. U.S.A. 2012;109:E3305–13.

123. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, et al. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell. 2010;140:631–42.

124. Volf P, Benkova I, Myskova J, Sadlova J, Campino L, Ravel C. Increased transmission potential of Leishmania major/Leishmania infantum hybrids. International Journal for Parasitology. 2007;37:589–93.

125. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

126. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

127. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. Guigo R, editor. PLoS Comput Biol. 2014;10:e1003998–9.

128. Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Mach Learn. 1995;21:51–80.

129. Bartholomeu DC, de Paiva RMC, Mendes TAO, DaRocha WD, Teixeira SMR. Unveiling the intracellular survival gene kit of trypanosomatid parasites. PLoS Pathog. 2014;10:e1004399.

130. Chiurillo MA, Cortez DR, Lima FM, Cortez C, Ramírez JL, Martins AG, et al. The diversity and expansion of the trans-sialidase gene family is a common feature in Trypanosoma cruzi clade members. Infection, Genetics and Evolution. 2016;37:266–74.

131. Burgos JM, Risso MG, Breniere SF, Barnabe C, Campetella O, Leguizamon MS. Differential distribution of genes encoding the virulence factor trans-sialidase along Trypanosoma cruzi Discrete typing units. PLoS ONE. 2013;8:e58967.

132. Freitas LM, Santos dos SL, Rodrigues-Luiz GF, Mendes TAO, Rodrigues TS, Gazzinelli RT, et al. Genomic analyses, gene expression and antigenic profile of the trans-sialidase superfamily of Trypanosoma cruzi reveal an undetected level of complexity. PLoS ONE. 2011;6:e25914.

133. Buscaglia CA, Campo VA, Frasch ACC, Di Noia JM. Trypanosoma cruzi surface mucins: host-dependent coat diversity. Nat Rev Microbiol. 2006;4:229–36.

134. Scudder P, Doom JP, Chuenkova M, Manger ID, Pereira ME. Enzymatic characterization of beta-D-galactoside alpha 2,3-trans-sialidase from *Trypanosoma cruzi*. J. Biol. Chem. 1993;268:9886–91.

135. Ferrero-Garcia MA, Trombetta SE, Sanchez DO, Reglero A, Frasch AC, Parodi AJ. The action of *Trypanosoma cruzi* trans-sialidase on glycolipids and glycoproteins. Eur J Biochem. 1993;213:765–71.

136. Buschiazzo A, Amaya MF, Cremona ML, Frasch AC, Alzari PM. The crystal structure and mode of action of trans-sialidase, a key enzyme in *Trypanosoma cruzi* pathogenesis. Molecular Cell. 2002;10:757–68.

137. Añez-Rojas N, Peralta A, Crisante G, Rojas A, Añez N, Ramírez JL, et al. *Trypanosoma rangeli* expresses a gene of the group II trans-sialidase superfamily. Molecular and Biochemical Parasitology. 2005;142:133–6.

138. Grisard EC, Stoco PH, Wagner G, Sincero TCM, Rotava G, Rodrigues JB, et al. Transcriptomic analyses of the avirulent protozoan parasite *Trypanosoma rangeli*. Molecular and Biochemical Parasitology. 2010;174:18–25.

139. Peña CP, Lander N, Rodriguez E, Crisante G, Añez N, Ramirez JL, et al. Molecular analysis of surface glycoprotein multigene family TrGP expressed on the plasma membrane of *Trypanosoma rangeli* epimastigotes forms. Acta Tropica. 2009;111:255–62.

140. Cuevas IC, Cazzulo JJ, Sánchez DO. gp63 homologues in Trypanosoma cruzi: surface antigens with metalloprotease activity and a possible role in host cell infection. Infect. Immun. American Society for Microbiology; 2003;71:5739–49.

141. Kulkarni MM, Olson CL, Engman DM, McGwire BS. *Trypanosoma cruzi* GP63 Proteins Undergo Stage-Specific Differential Posttranslational Modification and Are Important for Host Cell Infection. Infect. Immun. 2009;77:2193–200.

142. Bartholomeu DC, Cerqueira GC, Leão ACA, DaRocha WD, Pais FS, Macedo C, et al. Genomic organization and expression profile of the mucin-associated surface protein (*masp*) family of the human pathogen *Trypanosoma cruzi*. Nucleic Acids Res. Oxford University Press; 2009;37:3407–17.

143. Burleigh BA, Woolsey AM. Cell signalling and *Trypanosoma cruzi* invasion. Cell Microbiol. 2002;4:701–11.

144. Kawashita SY, da Silva CV, Mortara RA, Burleigh BA, Briones MRS. Homology, paralogy and function of DGF-1, a highly dispersed *Trypanosoma cruzi* specific gene family and its implications for information entropy of its encoded proteins. Molecular and Biochemical Parasitology. 2009;165:19–31.

145. Diez H, Thomas MC, Uruena CP, Santander SP, Cuervo CL, Lopez MC, et al. Molecular characterization of the kinetoplastid membrane protein-11 genes from the parasite Trypanosoma rangeli. Parasitology. 2005;130:643–51.

146. Ramirez JR, Berberich C, Jaramillo A, Alonso C, Velez IV. Molecular and antigenic characterization of the Leishmania (Viannia) panamensis kinetoplastid

membrane protein-11. Mem. Inst. Oswaldo Cruz. 1998;93:247–54.

147. Schenkman S, Ferguson MA, Heise N, de Almeida ML, Mortara RA, Yoshida N. Mucin-like glycoproteins linked to the membrane by glycosylphosphatidylinositol anchor are the major acceptors of sialic acid in a reaction catalyzed by trans-sialidase in metacyclic forms of *Trypanosoma cruzi*. Molecular and Biochemical Parasitology. 1993;59:293–303.

148. Abate T, Rincon M, Diaz-Bello Z, Spencer L, Rodriguez-Acosta A. A mucin like gene different from the previously reported members of the mucin like gene families is transcribed in Trypanosoma cruzi but not in Trypanosoma rangeli. Mem. Inst. Oswaldo Cruz. 2005;100:391–5.

149. Wagner G, Yamanaka LE, Moura H, Lückemeyer DD, Schlindwein AD, Stoco PH, et al. The *Trypanosoma rangeli* trypomastigote surfaceome reveals novel proteins and targets for specific diagnosis. Journal of Proteomics. Elsevier B.V; 2013;82:52–63.

150. Kangussu-Marcolino MM, Cardoso de Paiva RM, Araújo PR, de Mendonça-Neto RP, Lemos L, Bartholomeu DC, et al. Distinct genomic organization, mRNA expression and cellular localization of members of two amastin sub-families present in *Trypanosoma cruzi*. BMC Microbiology. 2013;13:1–11.

151. Cruz MC, Souza-Melo N, da Silva CV, DaRocha WD, Bahia D, Araújo PR, et al. *Trypanosoma cruzi*: Role of δ-Amastin on Extracellular Amastigote Cell Invasion and Differentiation. Blader I, editor. PLoS ONE. 2012;7:e51804–11.

152. JACKSON AP. The evolution of amastin surface glycoproteins in trypanosomatid parasites. Mol Biol Evol. 2010;27:33–45.

153. Rafati S, Hassani N, Taslimi Y, Movassagh H, Rochette A, Papadopoulou B. Amastin peptide-binding antibodies as biomarkers of active human visceral leishmaniasis. Clin Vaccine Immunol. 2006;13:1104–10.

154. Salotra P, Duncan RC, Singh R, Subba Raju BV, Sreenivas G, Nakhasi HL. Upregulation of surface proteins in *Leishmania donovani* isolated from patients of post kala-azar dermal leishmaniasis. Microbes and Infection. 2006;8:637–44.

155. Stober CB, Lange UG, Roberts MTM, Gilmartin B, Francis R, Almeida R, et al. From genome to vaccines for leishmaniasis: screening 100 novel vaccine candidates against murine *Leishmania major* infection. Vaccine. 2006;24:2602–16.

156. Rochette A, McNicoll F, Girard J, Breton M, Leblanc E, Bergeron MG, et al. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in Leishmania spp. Molecular and Biochemical Parasitology. 2005;140:205–20.

157. Mendes TAO, Lobo FP, Rodrigues TS, Rodrigues-Luiz GF, daRocha WD, Fujiwara RT, et al. Repeat-enriched proteins are related to host cell invasion and immune evasion in parasitic protozoa. Mol Biol Evol. 2013;30:951–63.

158. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame

genomic sequence comparison. Genome Res. 2011;21:487–93.

159. Allen CL, Kelly JM. *Trypanosoma cruzi*: Mucin Pseudogenes Organized in a Tandem Array. Experimental Parasitology. 2001;97:173–7.

160. Thon G, Baltz T, Eisen H. Antigenic diversity by the recombination of pseudogenes. Genes Dev. 1989;3:1247–54.

161. Barnes RL, McCulloch R. *Trypanosoma brucei* homologous recombination is dependent on substrate length and homology, though displays a differential dependence on mismatch repair as substrate length decreases. Nucleic Acids Res. Oxford University Press; 2007;35:3478–93.

162. Borst P, Bitter W, Blundell PA, Chaves I, Cross M, Gerrits H, et al. Control of VSG gene expression sites in *Trypanosoma brucei*. Molecular and Biochemical Parasitology. 1998;91:67–76.

163. Boothroyd CE, Dreesen O, Leonova T, Ly KI, Figueiredo LM, Cross GAM, et al. A yeast-endonuclease-generated DNA break induces antigenic switching in *Trypanosoma brucei*. Nature. 2009;459:278–81.

164. Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, Bason N, et al. Telomeric expression sites are highly conserved in *Trypanosoma brucei*. PLoS ONE. 2008;3:e3527.

165. Wen Y-Z, Zheng L-L, Qu L-H, Ayala FJ, Lun Z-R. Pseudogenes are not pseudo any more. RNA Biology. 2012;9:27–32.

166. Wickstead B, Ersfeld K, Gull K. Repetitive elements in genomes of parasitic protozoa. Microbiol Mol Biol Rev. 2003;67:360–75–tableofcontents.

167. Smit AFA, Hubley R, Green P. *RepeatMasker Open-4.0*. 2013-2015 [Internet]. Available from: http://www.repeatmasker.org

168. Smit AFA, Hubley R. *RepeatModeler Open-1.0*. 2008-2015 [Internet]. Available from: http://www.repeatmasker.org

169. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.

170. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76.

171. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21 Suppl 1:i351–8.

172. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.

173. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics.

2007;23:1282–8.

174. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999;27:29–34.

175. Alves JMP, Buck GA. Automated system for gene annotation and metabolic pathway reconstruction using general sequence databases. Chem Biodivers. 2007;4:2593–602.

176. Lee SH, Stephens JL, Englund PT. A fatty-acid synthesis mechanism specialized for parasitism. Nat Rev Microbiol. 2007;5:287–97.

177. Kraeva N, Butenko A, Hlavacova J, Kostygov A, Myskova J, Grybchuk D, et al. Leptomonas seymouri: Adaptations to the Dixenous Life Cycle Analyzed by Genome Sequencing, Transcriptome Profiling and Co-infection with Leishmania donovani. PLoS Pathog. 2015;11:e1005127.

178. Cazzulo JJ. Intermediate metabolism in Trypanosoma cruzi. J Bioenerg Biomembr. 1994;26:157–65.

179. Lisvane Silva P, Mantilla BS, Barison MJ, Wrenger C, Silber AM. The uniqueness of the Trypanosoma cruzi mitochondrion: opportunities to identify new drug target for the treatment of Chagas disease. Curr Pharm Des. 2011;17:2074–99.

180. Silber AM, Colli W, Ulrich H, Alves MJM, Pereira CA. Amino acid metabolic routes in Trypanosoma cruzi: possible therapeutic targets against Chagas' disease. Curr Drug Targets Infect Disord. 2005;5:53–64.

181. Tonelli RR, Silber AM, Almeida-de-Faria M, Hirata IY, Colli W, Alves MJM. L-proline is essential for the intracellular differentiation of Trypanosoma cruzi. Cell Microbiol. 2004;6:733–41.

182. Martins RM, Covarrubias C, Rojas RG, Silber AM, Yoshida N. Use of L-proline and ATP production by Trypanosoma cruzi metacyclic forms as requirements for host cell invasion. Infect. Immun. 2009;77:3023–32.

183. Pereira CA, Alonso GD, Torres HN, Flawia MM. Arginine kinase: a common feature for management of energy reserves in African and American flagellated trypanosomatids. J Eukaryot Microbiol. 2002;49:82–5.

184. Pereira CA, Alonso GD, Ivaldi S, Silber AM, Alves MJM, Torres HN, et al. Arginine kinase overexpression improves Trypanosoma cruzi survival capability. FEBS Letters. 2003;554:201–5.

185. Contreras VT, Salles JM, Thomas N, Morel CM, Goldenberg S. In vitro differentiation of Trypanosoma cruzi under chemically defined conditions. Molecular and Biochemical Parasitology. 1985;16:315–27.

186. Carrillo C, Cejas S, Huber A, Gonzalez NS, Algranati ID. Lack of arginine decarboxylase in Trypanosoma cruzi epimastigotes. J Eukaryot Microbiol. 2003;50:312–6.

187. Ariyanayagam MR, Oza SL, Mehlert A, Fairlamb AH. Bis(glutathionyl)spermine and other novel trypanothione analogues in Trypanosoma cruzi. J. Biol. Chem. 2003;278:27612–9.

188. Homsy JJ, Granger B, Krassner SM. Some factors inducing formation of metacyclic stages of Trypanosoma cruzi. J Protozool. 1989;36:150–3.

189. Manchola NC, Rapado LN, Barison MJ, Silber AM. Biochemical Characterization of Branched Chain Amino Acids Uptake in Trypanosoma cruzi. J Eukaryot Microbiol. 2016;63:299–308.

190. Ginger ML, Prescott MC, Reynolds DG, Chance ML, Goad LJ. Utilization of leucine and acetate as carbon sources for sterol and fatty acid biosynthesis by Old and New World Leishmania species, Endotrypanum monterogeii and Trypanosoma cruzi. Eur J Biochem. 2000;267:2555–66.

191. Opperdoes FR, Borst P. Localization of nine glycolytic enzymes in a microbody-like organelle in Trypanosoma brucei: the glycosome. FEBS Letters. 1977;80:360–4.

192. Cannata JJ, Cazzulo JJ. The aerobic fermentation of glucose by Trypanosoma cruzi. Comp Biochem Physiol B. 1984;79:297–308.

193. Fairlamb AH, Opperdoes FR. Carbohydrate Metabolism in African Trypanosomes, with Special Reference to the Glycosome. In: Morgan MJ, editor. Carbohydrate Metabolism in Cultured Cells. Boston, MA: Springer US; 1986. pp. 183–224.

194. Chaumont F, Schanck AN, Blum JJ, Opperdoes FR. Aerobic and anaerobic glucose metabolism of Phytomonas sp. isolated from Euphorbia characias. Molecular and Biochemical Parasitology. 1994;67:321–31.

195. Diaz-Albiter HM, Ferreira TN, Costa SG, Rivas GB, Gumiel M, Cavalcante DR, et al. Everybody loves sugar: first report of plant feeding in triatomines. Parasit Vectors. 2016;9:114.

196. Pryor EEJ, Horanyi PS, Clark KM, Fedoriw N, Connelly SM, Koszelak-Rosenblum M, et al. Structure of the integral membrane protein CAAX protease Ste24p. Science. 2013;339:1600–4.

197. Romano MC, Jiménez P, Miranda C, Valdez RA. Parasites and steroid hormones: corticosteroid and sex steroid synthesis, their role in the parasite physiology and development. Front Neurosci. Frontiers; 2015;9:1–12.

198. Emes RD, Yang Z. Duplicated paralogous genes subject to positive selection in the genome of Trypanosoma brucei. PLoS ONE. 2008;3:e2295.

199. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.

200. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol. 2012;30:771–6.

201. Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, et al. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 2014;24:688–96.

202. Cao H, Hastie AR, Cao D, Lam ET, Sun Y, Huang H, et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience. 2014;3:34.

203. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzén O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015;12:780–6.

204. Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, et al. Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. PLoS ONE. 2013;8:e55864.

205. Stankova H, Hastie AR, Chan S, Vrana J, Tulpova Z, Kubalakova M, et al. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol J. 2016;14:1523–31.

206. Xiao S, Li J, Ma F, Fang L, Xu S, Chen W, et al. Rapid construction of genome map for large yellow croaker (Larimichthys crocea) by the whole-genome mapping in BioNano Genomics Irys system. BMC Genomics. 2015;16:670.

207. Alves JMP, Serrano MG, Maia da Silva F, Voegtly LJ, Matveyev AV, Teixeira MMG, et al. Genome evolution and phylogenomic analysis of *Candidatus Kinetoplastibacterium*, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*. Genome Biol Evol. Oxford University Press; 2013;5:338–50.

208. Camargo EP. GROWTH AND DIFFERENTIATION IN TRYPANOSOMA CRUZI. I. ORIGIN OF METACYCLIC TRYPANOSOMES IN LIQUID MEDIA. Rev. Inst. Med. Trop. São Paulo. 1964;6:93–100.

209. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88.

210. O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, Cox AJ. NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics. 2015;31:2035–7.

211. Van der Ploeg LH, Cornelissen AW. The contribution of chromosomal translocations to antigenic variation in Trypanosoma brucei. Philos Trans R Soc Lond B Biol Sci. 1984;307:13–26.

212. Lanar DE, Levy LS, Manning JE. Complexity and content of the DNA and RNA in Trypanosoma cruzi. Molecular and Biochemical Parasitology. 1981;3:327–41.

213. Sloof P, Bos JL, Konings AF, Menke HH, Borst P, Gutteridge WE, et al. Characterization of satellite DNA in Trypanosoma brucei and Trypanosoma cruzi. J Mol Biol. 1983;167:1–21.

214. González A, Prediger E, Huecas ME, Nogueira N, Lizardi PM. Minichromosomal repetitive DNA in Trypanosoma cruzi: its use in a high-sensitivity parasite detection assay. Proc. Natl. Acad. Sci. U.S.A. National Academy of Sciences; 1984;81:3356–60.

215. Requena JM, Jimenez-Ruiz A, Soto M, Lopez MC, Alonso C. Characterization of a highly repeated interspersed DNA sequence of Trypanosoma cruzi: its potential use in diagnosis and strain classification. Molecular and Biochemical Parasitology. 1992;51:271–80.

216. Elias MCQB, Vargas NS, Zingales B, Schenkman S. Organization of satellite DNA in the genome of Trypanosoma cruzi. Molecular and Biochemical Parasitology. 2003;129:1–9.

217. Campetella O, Henriksson J, Aslund L, Frasch AC, Pettersson U, Cazzulo JJ. The major cysteine proteinase (cruzipain) from Trypanosoma cruzi is encoded by multiple polymorphic tandemly organized genes located on different chromosomes. Molecular and Biochemical Parasitology. 1992;50:225–34.

218. Gottesdiener K, Garciá-Anoveros J, Lee MG, Van der Ploeg LH. Chromosome Organization of the Protozoan *Trypanosoma brucei*. Molecular and Cellular Biology. 1990;10:6079–83.

219. Machado PE, Eger-Mangrich I, Rosa G, Koerich LB, Grisard EC, Steindel M. Differential susceptibility of triatomines of the genus *Rhodnius* to *Trypanosoma rangeli* strains from different geographical origins. International Journal for Parasitology. 2001;31:632–4.

Appendices

Appendix 1

Gene lengths used for multigene copy number calculation

	Reference genes	
	gene length (nt)**	UniProtKB ID
amastin	1161	K2PC41
beta galactofuranosyl glycosyltransferase	1659	Q4D109
cruzipain-precursor	1401	P25779
dispersed gene family protein 1 (DGF-1)	10560	Q4D375
mucin_TcMUCI*	1356	Q4CL35
mucin_TcMUCII*	5826	K4DWN9
mucin_TcSMUGL*	423	Q4CN03
mucin_TcSMUGS*	654	Q4E365
mucin-like_glycoprotein	1512	Q4DH37
mucin-associated surface protein (MASP)	2352	V5ALA8
retrotransposon hot spot (RHS) protein	4509	Q4CRR5
surface protease GP63	2742	Q4D292
syntaxin binding protein	1977	K2NVB0
target of rapamycin (TOR) kinase 1	7776	K2NXT0
trans-sialidase group I*	3612	Q4CVS5
trans-sialidase group II*	2814	Q4DKM3
trans-sialidase group III*	3126	Q4CZ80
trans-sialidase group IV*	5244	Q4DGT2
trans-sialidase group V*	2970	Q4E5U7
trans-sialidase group VI*	2901	Q4DP77
trans-sialidase group VII*	3441	Q4D6R7
trans-sialidase group VIII*	3348	Q4DE10
UDP-Gal or UDP-GlcNAc-dependent glycosyltransferase	1881	K2MXL9

* putative, as subgroups within the same multigene family may have high sequence similarity

** UniProtKB longest full length gene record for all *Trypanosoma cruzi*, *Trypanosoma conorhini* and *T. rangeli* strains (www.uniprot.org/uniprot/; accessed 21 Dec 2015). Trans-sialidase subgroups defined by table S3 Freitas LM, dos Santos SL, Rodrigues-Luiz GF, Mendes TAO, Rodrigues TS, Gazzinelli RT, et al. (2011) Genomic Analyses, Gene Expression and Antigenic Profile of the Trans-Sialidase Superfamily of Trypanosoma cruzi Reveal an Undetected Level of Complexity. PLoS ONE 6(10): e25914. doi:10.1371/journal.pone.0025914

Vita

Katie Rebecca Bradwell was born on February 2, 1984, in London, England, and is a British citizen. She graduated from Farnham College, Surrey in 2002, and received her Bachelor of Science in Microbiology from the University of Surrey, Guildford in 2007, which included a year in industry at Oxoid Ltd., Basingstoke. She subsequently taught in a public school in Madrid, Spain and worked in a Madrid-based patent firm, each for a year. She received a Master of Science in Molecular and Cellular Biology and Genetics from the University of Valencia, Spain in 2010, and then obtained a short-term position as a visiting researcher at Imperial College London in trypanosome biology. She received a Professional Science Masters in Bioinformatics from Virginia Commonwealth University in 2015.