Repeat-driven generation of antigenic diversity in a major human pathogen, *Trypanosoma cruzi*

Carlos Talavera-López[1], João Luís Reis-Cunha[3], Louisa A. Messenger[2], Michael D. Lewis[2], Matthew Yeo[2], Daniella C. Bartholomeu[3], José E. Calzada[4], Azael Saldaña[4], Juan David Ramírez[5], Felipe Guhl[6], Sofía Ocaña-Mayorga[7], Jaime A. Costales[7], Rodion Gorchakov[8], Kathryn Jones[8], Melissa Nolan Garcia[8], Edmundo C. Grisard[9], Santuza M. R. Teixeira[10], Hernán Carrasco[11], Maria Elena Bottazzi[8], Peter J. Hotez[8], Kristy O. Murray[8], Mario J. Grijalva[7, 12], Barbara Burleigh[13], Michael A. Miles[2], Björn Andersson[1*].

1 - Department of Cell and Molecular Biology; Karolinska Institutet; Stockholm, Sweden.

2 - London School of Hygiene and Tropical Medicine; London, United Kingdom.

3 – Departamento de Parasitologia, Universidade Federal de Minas Gerais; Belo Horizonte MG, Brazil.

4 - Departamento de Parasitología, Instituto Conmemorativo Gorgas de Estudios de la Salud; Ciudad de Panamá, Panamá.

5 - Facultad de Ciencias Naturales y Matemáticas, Universidad del Rosario; Bogotá, Colombia.

6 - Grupo de Investigaciones en Microbiología y Parasitología Tropical (CIMPAT), Universidad de Los Andes; Bogotá, Colombia.

7 - Centro de Investigación para la Salud en América Latina (CISeAL), Escuela de Ciencias Biológicas; Pontificia Universidad Católica del Ecuador; Quito, Ecuador.

8 - Sabin Vaccine Institute and Texas Children's Hospital Center for Vaccine Development, National School of Tropical Medicine; Baylor College of Medicine; Houston TX, United States.

9 – Departamento de Microbiologia, Imunologia e Parasitologia; Universidade Federal Santa Catarina; Florianópolis, SC, Brazil.

10 – Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais; Belo Horizonte MG, Brazil.

11- Universidad Central de Venezuela; Caracas, Venezuela.

28 12 - Infectious and Tropical Disease Institute, Department of Biomedical Sciences, Heritage College
29 of Osteopathic Medicine, Ohio University; Athens OH, United States.
30 13 - T.H. Chan School of Public Health; Harvard University; Boston MA, United States.
31
32 * - Corresponding author

ABSTRACT

Trypanosoma cruzi, a zoonotic kinetoplastid protozoan parasite with a complex genome, is the causative agent of American trypanosomiasis (Chagas disease). Having a very plastic and complex genome, the parasite uses a highly diverse repertoire of surface molecules, with pivotal roles in cell invasion, immune evasion and pathogenesis. Thus far, the complexity of the genomic regions containing these genes has impaired the assembling a genome at chromosomal level, turning impossible to study the structure and function of the several thousand repetitive genes encoding the surface molecules of the parasite. We here present the genome assembly of a *T. cruzi* clade I (TcI) strain at semi-chromosomal level using high coverage PacBio single molecule sequencing, together with whole genome Illumina sequencing of 34 *T. cruzi* TcI isolates and clones from different geographic locations, sample sources and clinical outcomes. Resolution of the surface molecule gene distribution reveals an unusual duality in the organisation of the parasite genome, a synteny of the core genomic region with related protozoa flanked by unique and highly plastic multigene families' clusters encoding surface antigens. The presence of abundant interspersed retrotransposons in these multigene families' clusters suggests that these elements are involved in a recombination mechanism for the generation of antigenic variation and evasion of the host immune response on these TcI strains . The comparative genomic analysis of the cohort of TcI strains revealed multiple cases of such recombination events involving surface molecule genes and has provided new insights into *T. cruzi* population structure.

53 **INTRODUCTION**

54 *Trypanosoma cruzi* is a kinetoplastid protozoan and the etiologic agent of Chagas disease,
55 considered one of the most important human parasitic disease in Latin America. The Global
56 Burden of Disease Study 2013 reported that almost 7 million people live with Chagas disease
57 in the Western Hemisphere [1], with the expectation that up to one third will progress to develop
58 chronic chagasic cardiomyopathy (CCC) or other life-threatening symptoms. In 2015,
59 5,742,167 people were estimated to be infected with *T. cruzi* in 21 Latin American countries
60 and around 13 % of the Latin American population is at risk of contracting *T. cruzi* infection
61 due to domicile infestation of triatomine bugs or due to non-vectorial transmission via blood
62 transfusion, organ transplant, oral, congenital or accidental infection [2]. Human Chagas
63 disease is not restricted to Latin America. Migration of infected humans to non-endemic areas
64 has made it a new public health threat in other geographic areas such as North America,
65 Europe and Asia [4]. Also, sylvatic *T. cruzi* transmission cycles, often associated with human
66 disease, have been described in areas formerly considered as free from this disease such as
67 in Texas (USA)[4].

68

69 The acute phase of the disease frequently lacks specific symptoms, is often undiagnosed
70 and usually resolves in a few weeks in immunocompetent individuals but may be fatal in
71 around 5% of diagnosed cases. Without successful treatment, a *T. cruzi* infection is normally
72 carried for life. The disease progresses to either a chronic indeterminate phase that is
73 asymptomatic, or to a chronic symptomatic phase with severe clinical syndromes such as
74 cardiomyopathy, megaesophagus and/or megacolon [5]; meningoencephalitis may occur,
75 especially in immunocompromised patients [4]. The current prolonged chemotherapy
76 (benznidazole or nifurtimox) is mostly effective only in the acute phase, particularly because
77 severe side effects may interrupt treatment of adults in the chronic phase. There is currently
78 no effective treatment for advanced chagasic cardiomyopathy [6], and there is an urgent need
79 to identify new potential drug and vaccine targets [7].

80

81 *T. cruzi* infection is a zoonosis, and the parasite has a complex life cycle; where transmission
82 to humans occurs most frequently by contamination with infected feces from triatomine
83 insect vectors (Subfamily Triatominae). The parasite evades the immune responses with the
84 aid of multiple surface molecules from three large diverse gene families (Trans-Sialidases,
85 Mucins and Mucin-Associated Surface Proteins - MASPs), which are also involved in cell
86 invasion and possibly pathogenicity [8].

87

88 Six distinct genetic clades of *T. cruzi* have been recognised, named TcI to TcVI (Discrete
89 Typing Units or DTU-I to VI). The first genome sequence for *T. cruzi* was produced using
90 Sanger sequencing technology from a hybrid, highly polymorphic, TcVI strain. The resultant
91 genome sequence, while extremely useful for the core regions of the genome, was highly
92 fragmented, especially in repetitive regions [9]. This sequence has been improved using
93 enhanced scaffolding algorithms, but the repetitive regions remained unresolved [10].
94 Subsequently, FLX 454 Titanium and Illumina sequencing were used to sequence a less
95 polymorphic TcI strain (Sylvio X10/1), which allowed the first comparative genomic studies of
96 *T. cruzi*, but correct assembly of repetitive regions was still impossible [11,12]. The thousands of
97 related genes that code for the surface proteins are generally located in large multigene
98 families clusters of the *T. cruzi* genome [13], in the form of extremely repetitive segments with
99 multiple gene copies and pseudogenes. These multigene families' clusters are distinct from
100 the core regions of the genome in synteny, gene content and diversity [22]. The repetitive nature
101 of the tandem arrays and the length of the repeat regions, have made correct assembly
102 impossible using short and medium-sized sequence reads. The available *T. cruzi* genome
103 sequences are therefore incomplete and erroneous in these important regions, making it
104 impossible to study the complex surface gene families.

105

106 The population structure of *T. cruzi* is complex, and there is a high degree of genetic and
107 phenotypic variation. The current TcI to TcVI clades are based on biochemical and molecular
108 markers [14], although there is substantial diversity even within these six groups [15]. The TcI
109 clade is widespread and can be found across the American continent, and has been
110 associated with CCC [16] and sudden death [17,18], among other clinical manifestations.
111
112 We have produced a complete and reliable reference sequence of the majority (~98.5%) of
113 the *T. cruzi* TcI Sylvio X10/1 genome, and we have generated Illumina whole genome
114 sequencing data for 34 *T. cruzi* TcI isolates and clones from different geographic locations
115 for comparative analyses. Thus, we have been able to decipher the majority of the
116 organisation of *T. cruzi* surface protein coding gene repertoire from TcI Sylvio X10/1 strain,
117 revealing large numbers of evenly spaced retrotransposons, which may play a role in
118 generating genomic structural diversity and antigenic variation. Furthermore, the comparative
119 data enabled the first exploration of whole-genome population genetics of *T. cruzi* in different
120 environments and geographic locations. We found patterns of active recombination
121 associated with generation of new surface molecule variants. Together, these results
122 contribute to answering longstanding questions on the biology of Chagas disease and
123 parasitism in general. The availability of the complete repertoire of genes encoding surface
124 molecules allows further research on virulence and pathogenesis, as well as the identification
125 of drug targets and vaccine candidates, focused on shared and conserved motifs present
126 within these variable families.
127
128
129
130
131
132

133 **RESULTS**:

134 **Genome sequence of *Trypanosoma cruzi* Sylvio X10/1**

135 The final Sylvio X10/1 (TcI) genome assembly reconstructed 98.5 % of the estimated strain
136 genome size and was contained in 47 scaffolds (**Figure 1a**) - which here will be referred to
137 as pseudomolecules - assembled from 210 X PacBio sequence data and a previous Illumina
138 data set (**Table 1**). Comparison with the available assembly of the TcVI strain CL Brener
139 revealed a conserved core of syntenic blocks composed of stretches of homologous
140 sequences separated by large gaps of sequence that were not reconstructed in the hybrid
141 TcVI strain. These gaps mostly corresponded to surface molecule gene arrays and simple
142 repeats in the Sylvio X10/1 genome (**Figure 1b**). The length of the PacBio reads and the high
143 coverage allowed the reconstruction of long stretches of repetitive sequences (**Figure 1b**)
144 that could not be resolved using shorter read data.

145

146 The coverage of genomic regions coding for surface molecules supported the correct
147 reconstruction of these areas (**Figure 1a**). To further investigate the quality of the new
148 assembly, Illumina short reads were mapped and analysed with FRC_bam, which revealed
149 assembly artefacts related to low coverage, wrong paired end read orientation, and higher
150 than expected sequencing coverage in regions with long stretches of simple repeats.

151

152 Repetitive elements comprised 18.43 % of the TcI Sylvio X10/1 genome, 2.18 % of which
153 cannot be classified using the repeat databases. LINE retroelements of the R1/Jockey group
154 (3.63 %) and VIPER LTRs (2.87 %) were found to be the most prevalent types of
155 retroelements, covering 6.89 % of the genome, which is much higher than the 2.57 %
156 estimated from the previously published Sylvio X10/1 draft assembly using short reads [12].

157

158 Although retrotransposons were found to be present throughout the genome, the frequency
159 of VIPER and L1Tc elements was markedly higher in multigene families clusters regions and

160 they were found within one kilobase of pseudogenes, hypothetical proteins and surface
161 molecule gene tandem arrays (One-sided Fisher exact test, *p-value* < 1.32 e-16). This
162 distribution indicates that these elements may play a role in the generation of new sequence
163 diversity in gene families clusters by providing a source of microhomology. It is compelling
164 to speculate that they could act as transcriptional regulators by the introduction of novel
165 transcription start sites, as has been proposed in other eukaryotes [19]; nevertheless, we do
166 not have experimental evidence for the activity of these retroelements in *T. cruzi*.
167
168 Simple and low complexity repeats were observed surrounding subtelomeric coding
169 sequences and were also more abundant in the subtelomeric regions (2.18 %), extending up
170 to 4 Kb, compared to core regions (0.98 %) where they were much shorter (10 - 120 bp). The
171 most prevalent type of simple repeat had the (C)n motif (11.7 %), (TG)n repeat motif (5.6 %)
172 and (CA)n repeat motif (5.1 %); each variable in length. This microhomology of the simple
173 subtelomeric repeats may facilitate recombination for the generation of new surface molecule
174 variants, as described in other parasitic protozoa, including *Trypanosoma brucei* and
175 *Plasmodium falciparum* [20,21]. However, it is noteworthy to mention that such subtelomeric
176 regions are far less complex and shorter in *T. brucei* (african, virulent) and *T. rangeli* [23]
177 (american, non-virulent).
178
179 Based on our annotation approach, a total of 19,096 genes were identified in the TcI Sylvio
180 X10/1 haploid genome sequence, compared to the higher estimate of 22,570 for TcVI CL
181 Brener, mostly due to the larger size of the multigene families' clusters in the TcVI hybrid
182 genome. The core regions of the genome were found to correspond well to results generated
183 previously using short read sequencing of the same strain [11], in both gene organisation and
184 content. Tandem repeated genes that were collapsed in previous *T. cruzi* genome assemblies
185 were now resolved. About 24.1% (n = 4,602) of the total annotated genes were truncated,

186 mostly due to the introduction of premature stop codons, and 67 % of these were located
187 within surface molecule gene arrays, sharing motifs of the complete genes.
188
189 The new assembly allowed an improved analysis of the *T. cruzi* surface molecule gene
190 repertoire. Genes of each of the three major surface molecules families were organised as
191 multiple tandem arrays. After genome annotation, the total number of such arrays were:
192 trans-sialidases, 312, with 2,048 complete gene copies and 201 pseudogenes; mucins 98,
193 with 2,466 complete copies and 111 pseudogenes; MASPs 264, with 1,888 complete copies
194 and 245 pseudogenes. These three surface molecule gene families comprised 16.02 Mbp
195 (39.04 %) of the TcI Sylvio X10/1 genome and presented a high level of sequence diversity
196 (**Figure 1a**). Sequence strand switches often delimited the surface molecule tandem arrays.
197 Commonly, these arrays had two to four complete copies immediately followed by two or
198 more truncated copies with motifs similar to the complete gene. The intergenic spaces
199 between arrays were rich in simple and low complexity repeats with no identifiable regulatory
200 elements. The VIPER and L1Tc retrotransposon elements, in clusters of two to four copies,
201 were found in the proximity of, or inside, tandem arrays containing trans-sialidases, mucins
202 and MASP genes. As the surface molecule genes are known to evolve rapidly and be highly
203 variable [22], the enrichment of VIPER and L1Tc elements in these regions supports the
204 hypothesis that they may be involved in generating new surface molecule gene variants via
205 recombination mediated by sequence homology.
206
207 Both, Ser/Thr kinases and DEAD-box RNA helicase genes were found at both extremes of
208 34 (10.81 %) trans-sialidase arrays located in pseudomolecules 1, 2 and 8. Searches against
209 the RFAM database identified 1,618 small RNAs in the TcI Sylvio X10/1 genome. These were
210 mostly ribosomal RNAs with the 5S rDNA subunit being the most common (31.9 %) followed
211 by ACA Box snoRNAs (30.9 %), SSU rDNA (12.2 %) and LSU rDNA (10.2 %) subunits. We
212 also found hits to telomerase RNA component (TERC), Catabolite Repression Control

213 sequester (CrcZ), Protozoa Signal Recognition Particle RNAs, spliceosomal RNA subunits
214 and miRNAs. The putative miRNAs identified in Sylvio X10/1 belong to the MIR2118 and
215 MIR1023 families, previously not found in protozoan parasites. The functional relevance of
216 these predicted small RNAs will need to be further validated *in vitro*. The miRNA segments
217 were located in both strands within 1 Kb of genes coding for DEAD-box RNA helicases
218 surrounding surface molecule gene tandem arrays.
219
220 **Genomic variation within the *Trypanosoma cruzi* TcI clade**:
221 Intra-TcI genomic diversity was examined among 34 samples from six countries: United
222 States, Mexico, Panama, Colombia, Venezuela and Ecuador, derived from a range of
223 triatomine vectors and human patients of different clinical stages (**Table 2** and
224 **Supplementary Table 1**). Our hybrid variant calling strategy allowed us to identify genomic
225 variants in the core and multigenic-families clusters in a reliable fashion (See methods).
226
227 A total of 1,031,785 SNPs and 279,772 INDELs shorter than 50 bp were called for all TcI
228 sequenced isolates relative to the Sylvio X10/1 genome. INDELs presented an average
229 density of 5.3 variants per Kb and SNPs 24.1 variants per Kb. An individual *T. cruzi* TcI isolate
230 was found to contain an average of 61,000 SNPs and 6,820 INDELs with a density of 31.8
231 variants per Kb. However, these measures fluctuated depending on the geographical and
232 biological source of the sample. Core regions had an average SNP density of 0.4 variants per
233 Kb, in contrast with multigene-families clusters where values of 10 variants per Kb were
234 found. It was not surprising that the bulk of the genomic variants were located in the
235 multigenic families clusters regions in all the isolates, with fewer differences in the core
236 regions. Although several studies using single gene markers have identified heterogeneity in
237 the TcI clade [15,24], the extent of this variation has not been previously assessed genome-wide.
238

239 The majority of INDELs (96 %) were found in intergenic or noncoding regions, and 81 % of
240 those were located in subtelomeric regions. INDELs within coding sequences were
241 exclusively found to cause frameshifts turning the affected coding sequence into a
242 pseudogene. This distribution of INDELs is a genomic signature that has been associated
243 with non-allelic homologous recombination due to unequal crossing over [25] or
244 microhomology-mediated end joining [26,27] (**Table 3**). Short insertions were more prevalent
245 than short deletions, a pattern common to all the analysed TcI genomes when compared to
246 Sylvio X10/1. In the subtelomeric regions, short insertions (1 - 3 bp) occurred within the
247 upstream and downstream portions of the coding sequences and usually involved the
248 addition of one or more cytosines or guanines. Deletions of 1 bp indicating the removal of an
249 adenine or thymine were also observed within these regions, but at a lower frequency. Longer
250 deletions (5 - 20 bp) and insertions (8 - 10 bp) were observed within trans-sialidases,
251 Retrotransposon Hot Spot (RHS), pseudogenes and, at a lower frequency, L1Tc
252 retroelements.
253
254 **Population genomics of the *Trypanosoma cruzi* TcI clade:**
255 We used the short genomic variants to analyse the population genomics of the *T. cruzi* TcI
256 clade, and where possible taking into account the different sample sources (insect vector or
257 human host), clinical outcome of the infected patients and geographic locations
258 (**Supplementary table 1**). This sampling strategy allowed comparison of parasite population
259 structure in different environments. Interestingly, a Bayesian PCA analysis using INDELs and
260 an IBD-based hierarchical clustering using only SNPs for all the samples showed a mostly
261 geography-specific population structure (**Figure 3a** and **Supplementary Figure 3**).
262
263 The analysis of the variation between two Colombian TcI isolates made it possible to compare
264 parasites from a HIV-positive patient with fatal cardiomyopathy (CG) and from an acute
265 chagasic patient infected by oral transmission (FcHc). For each strain, replicate clones from

266 the original sample were isolated and cultured under the same conditions, and five of the
267 replicates from each sample were sequenced in a single Illumina HiSeq 2500 run and 158,565
268 well-supported SNPs were called. Using this set of SNPs we calculated global and per-site
269 population genetic statistics. These samples displayed distinctive behaviour in a global
270 analysis of genomic diversity by separating into two well-defined clusters, as can be seen in
271 **Figure 3b**. Linkage Disequilibrium (LD) analyses were performed genome-wide for both
272 groups using the r2 statistic; revealing a fluctuating pattern of LD across the entire genome
273 with large blocks of low r2 values - implying a recombinatorial process - present at distinctive
274 chromosomal locations that were specific to each group of clones. Particularly, CG clones
275 had less genetic diversity than FcHc clones (**Figure 4a**) and displayed a trend towards LD,
276 whereas FcHc clones presented more dynamic LD pattern. Values of r2 near zero were more
277 common in LD sliding windows containing genes coding for surface molecules and r2 values
278 closer to one were present exclusively in core regions rich in housekeeping genes, indicating
279 that these regions are more stable. For the CG and FcHc clones we calculated a global
280 Fixation index (*Fst*) value of -0.9377958 and -0.1162212 respectively (**Figure 4b**). These
281 values are consistent with genetic differentiation in recombination hotspots in the
282 subtelomeric regions. The global Tajima's D value for the CG clones was 1.373 and 0.9906
283 for FcHc clones, suggesting the presence of multiple alleles at variable frequencies in both
284 populations (**Figure 4c**). This pattern was more evident in the subtelomeric regions, which is
285 consistent with balancing selection of surface molecules.
286
287 Analyses of genomic variation between samples isolated from humans and vectors from
288 Mexico, Panama and Ecuador revealed that the global genetic differentiation among samples
289 isolated from vectors was *Fst* = 0.1289547 whereas for samples isolated from humans the
290 observed was *Fst* = -0.05521983. The patterns of linkage disequilibrium between human and
291 vector derived isolates were similar to those observed in the Colombian clones. Estimates of
292 the Tajima's D statistic revealed a distinctive pattern of selection between the two groups.

293 Balancing selection was detected specifically in regions containing tandem gene arrays
294 coding for surface molecules in all the samples derived from vectors, regardless of their
295 geographical origin; whereas selective sweeps were present in the same regions in human-
296 derived samples. Large genomic areas (> 50 Kb) containing surface molecule genes
297 displayed negative Tajima's D values in human-derived isolates, in contrast with the pattern
298 observed in vector-derived isolates with long genomic stretches (> 70 Kb) of positive Tajima's
299 D values and short genomic blocks (< 5 Kb) with negative values.
300
301 **Genome Structural Variation**:
302 Genomic structural variants, such as deletions, tandem and interspersed duplications,
303 genomic inversions and chromosomal break-ends, were observed ubiquitously throughout
304 the genomes of the analysed TcI strains. The most common type of intrachromosomal
305 structural variant observed was tandem duplications followed by deletions larger than 50 Kb
306 (**Table 2**). Chromosomal break-ends, similar to the unbalanced chromosomal translocations
307 observed in higher eukaryotes, were the most abundant type of structural rearrangement and
308 they were only present in genomic regions that were statistically enriched with retroelements
309 and simple repeats. These areas presented a conserved pattern: they contained surface
310 molecule gene tandem arrays and their breakpoints were composed of simple repeats and
311 retrotransposons of the VIPER and L1Tc class.
312
313 These events were between 20 - 150 Kb in length and contained fragments or even complete
314 coding sequences for surface molecule genes, such as trans-sialidases, mucins and MASP
315 genes and surface glycoproteins (gp63/gp85). Housekeeping genes seemed to have not
316 been affected by these genomic rearrangements. The breakpoints were composed of simple
317 repeats, retrotransposons or both. Rearrangements affecting gene tandem arrays generated
318 longer coding sequences by superimposing fragments - or the entire coding sequence - on
319 genes of the same family located in a different genomic location. For instance, the Colombian

320 isolates generated longer trans-sialidase genes by moving coding sequences from
321 pseudomolecule 1 to pseudomolecule 8, while Texas isolates recombined trans-sialidases
322 between pseudomolecule 16 and pseudomolecule 21. In this way, surface molecule genes
323 were merged with another member of the same gene family - or a pseudogene - resulting in
324 a new mosaic gene sequence (**Figure 5a**).
325
326 Retroelements could be found within or near genomic regions containing surface molecule
327 gene tandem arrays (**Supplementary Tables 2, 3** and **4**) and L1Tc fragments or their entire
328 sequence were also included in the rearranged region in all the observed translocation spots,
329 where they were inserted into regions containing simple repeats composed by AT dimers
330 (**Figure 5b**).
331
332 Multiple examples of the generation of new surface molecule gene variants were identified in
333 TcI from diverse sources. It therefore appears that the parasite uses specific molecular
334 mechanisms of recombination that can rapidly generate surface molecule diversity, allowing
335 it to increase the genomic plasticity required to adapt to changing environments and evade
336 immune responses during short and long-term infections in various host species.
337
338 The sizes of the tandem duplications ranged from 6 - 75 Kb and mainly involved tandem
339 arrays coding for surface molecules, mostly trans-sialidases and mucins, but also Disperse
340 Gene Family 1 (DGF-1) and several hypothetical proteins. The breakpoints of these
341 duplications were surrounded by simple repeats and retroelements in subtelomeric regions.
342 A tandem duplication event could involve between four and 25 copies of a specific gene
343 when in the subtelomeric regions, whereas in core regions the number was between two and
344 eight. We observed that large deletions occurring in subtelomeric regions were surrounded
345 by simple repeats of the type (T)n and (AT)n and retrotransposons of the L1Tc class,

346 containing surface molecule gene tandem arrays. Deletions in these genomic regions tended
347 to be shorter (4 - 12 Kb) and sample specific.
348
349 **Distribution of Copy Number Variation (CNV) within the TcI clade**:
350 CNV varied extensively between *T. cruzi* TcI strains. Most notably, among the Colombian
351 strains, isolates derived from the same sample presented different gene copy numbers. There
352 have been previous attempts to assess CNV in the *T. cruzi* genome [28], but these studies were
353 performed using DNA tiling microarrays with probes designed using the TcVI CL Brener strain
354 assembly, in which subtelomeric regions are essentially absent.
355
356 The distribution of CNV in the genomes of the studied TcI samples was isolate-specific, and
357 involved segments of an average size of 5 Kb. In the samples analysed in our study we
358 observed blocks of segmental CNV within a chromosome with a pattern that was unique to
359 each sample. Notably, the Colombian clones presented individual profiles of CNV (**Figure 6a**
360 and **6b**) despite being derived from the same clinical isolates.
361
362 Sequence blocks affected by segmental CNVs contained retrotransposons of the VIPER and
363 L1Tc class, as well as surface molecule genes surrounded by simple repeats. The isolate-
364 specific nature of these CNV events demonstrates the high level of within-clade diversity of
365 the TcI samples. The distribution of CNV across the *T. cruzi* genome reinforces the dynamic
366 nature of the multigene families clusters and the surface molecule gene families. As
367 discussed below, it is important to note the association of structural and copy number
368 variation with the presence of retrotransposons and simple repeats and their putative
369 involvement in the generation of novel sequence variants via recombination.

370 **DISCUSSION**

371 Complete reconstruction of the *T. cruzi* genome to encompass the subtelomeric regions and
372 multigene families' clusters, has proved to be difficult to achieve using short reads, due to
373 sequencing library preparation biases and a genome architecture that is rich in long stretches
374 of simple repeats, large repetitive gene families and multiple retrotransposons. Here we have
375 used long PacBio sequencing reads to provide the most complete genome sequence of a *T.*
376 *cruzi* strain to date. This has allowed us to perform the first detailed analyses of the repertoire
377 of complex genes families that encode cell surface molecules, considered to be involved in
378 cell invasion and evasion of the host immune response. We have shown the duality in the
379 organisation of the parasite genome, comprised of a core genomic component with few
380 repetitive elements and a slow evolutionary rate, resembling that of other protozoa, and a
381 contrasting, highly plastic multigene families clusters encoding fast evolving surface
382 antigens, with abundant interspersed retrotransposons. The structural changes that generate
383 and maintain diversity in *T. cruzi* surface molecules have certain mechanistic parallels in other
384 protozoa such as those recently described in *Plasmodium falciparum* [28], but differing from
385 the shorter, less repetitive genome of the non virulent, human-infective *Trypanosoma* [22].

386

387 Early studies of the genetic diversity of *T. cruzi* using geographically disparate sampling and
388 restricted comparisons of genetic diversity suggested a clonal population structure [29,30];
389 however, population genetics with an expanded set of markers have now challenged this
390 view [15,31,32]. Nevertheless, there are still conflicting views as to which model best describes
391 the population structure of *T. cruzi* [33,34]. The newer Sylvio X10/1 genome sequence will now
392 enable extensive genome-wide comparative population genomics analyses, which may shed
393 light on this issue. Comparative analyses of 34 *T. cruzi* isolates and clones from the TcI clade
394 suggest many recombination events and population indices, normally associated with
395 genetic exchange between strains, which are more likely to be caused by the extensive
396 repeat-driven recombination in the subtelomeric regions. The extent of variation in the

397 multigene-families clusters rich in surface antigen coding genes and the geographical
398 clustering of strains based on geographic distribution indicates active, on-going adaptation
399 to host and vectors. This need for phenotypic - and thus genomic - versatility may impel the
400 active generation of sequence diversity in *T. cruzi*. Further analyses of the evolution of
401 multigene families' clusters will yield much more detailed understanding of diversity within
402 and between the six currently recognised genetic lineages of *T. cruzi* [23].
403
404 We have shown how the genome architecture and dynamic multigene families' clusters of *T.*
405 *cruzi* may provide a mechanism to rapidlygenerate sequence diversity, required to escape
406 the host immune response and adapt in response to new environments. It is the striking
407 richness in simple repeat, retrotransposons and motif conservation in the multigene families
408 clusters that renders these genomic areas susceptible to structural change, similar to yeast
409 and other pathogens (Weatherly 2016 - PMID: 27619017) [35,36,44]. Retrotransposons have been
410 associated with the generation of complexity in genomic regions in mammals and plants and
411 with control of gene expression [36,37]. In the case of *T. cruzi*, they appear to generate novel
412 variants via mechanisms that exploit sequence homology. The presence of the simple repeats
413 and retrotransposons near surface molecule coding genes provides the microhomology for
414 both mechanisms to operate in such regions. Besides retrotransposons, the modular
415 structure of the multigene families MASP and Trans-sialidase, where different genes share
416 conserved motifs, could also provide microhomology needed for this homologous
417 recombination (EL Sayed 2005 PMID: 16020725, Weatherly 2016 PMID: 27619017). Our
418 analysis of INDELs and chromosomal breakpoints in the subtelomeric regions confirmed that
419 a mechanism similar to NAHR or MMEJ operates as source of sequence diversity, for
420 example by transposition of trans-sialidase genes or pseudogenes to produce new sequence
421 mosaics. The required recombination machinery is conserved in *T. cruzi* [38]. Furthermore,
422 these mechanisms would explain the high level of pseudogenisation observed in *T. cruzi*.
423

Retrotransposons were first reported from *T. cruzi* in 1991 [39]. The presence of these elements may also partly account for the previously reported widespread observation of copy number variation in different *T. cruzi* strains [27]. Thus, we find that repeats near the surface molecule genes appear to drive recombination in *T. cruzi*. The apparent inability of *T. cruzi* to condense chromatin may facilitate transposition in a stochastic fashion, facilitating generation of sequence diversity in exposed regions of the genome. A similar process has been described in the neurons of higher eukaryotes [40] but not in any other unicellular organism. Retrotransposons may also have an important role as gene transcription regulators: they may either silence or promote gene expression, due to their susceptibility to DNA methylation or by providing potential binding sites respectively, as it has been observed in previous works [41]. This lack of a well-defined transcriptional regulation machinery in the *T. cruzi* genome may suggests a link to the requirement for retrotransposon closely associated with gene tandem arrays.

CONCLUSION

Here we have sequenced and assembled the complete genome of a *Trypanosoma cruzi* TcI strain. This has enabled the first resolution of the complex multiple gene families that encode *T. cruzi* surface molecules, and provided a basis for *T. cruzi* population genomics. We discover an extraordinary concentration of retrotransposons among the multigene families' clusters and indications of repeat-driven recombination and generation of antigenic diversity, providing the mechanisms for *T. cruzi* to evade the host immune response, and to facilitate the adaption to new host and vectors. This genome will provide an invaluable resource to facilitate the prospective discovery of novel drug targets and vaccine candidates for Chagas disease.

448 **METHODS**:

449 **Genome Sequencing and Assembly**: *Trypanosoma cruzi* Sylvio X10/1 strain was isolated
450 from an acute human case of Chagas disease in Brazil [11]. Total genomic DNA of this TcI strain
451 was obtained from culture epimastigotes as formerly described [11] and used to produce
452 PacBio CCS data according to standard protocols from the Genomic Facility of Science for
453 Life Laboratory (Sweden) and Pacific Biosciences (USA). Genomic DNA was sequenced to a
454 depth of 210X using the PacBio platform, supplying raw reads with an average length of 5.8
455 Kb. These reads were corrected by means of the PBcR v8.3 pipeline with the MHAP algorithm
456 [42] using the auto-correction parameters described to merge haplotypes and skipping the
457 assembly step, producing a total of 1,216 contigs (NG50 = 62 Kb). Later, the assembly was
458 scaffolded using the corrected PacBio reads with the SSPACE-Long scaffolder yielding 310
459 scaffolds (NG50 = 788 Kb); 118 gaps were filled using Illumina reads with GapFiller and
460 corrected PacBio reads with PBJelly2. Finally, the core regions of these scaffolds were
461 aligned against the core regions of the TcVI CL Brener reference genome using ABACAS
462 (http://abacas.sourceforge.net), producing 47 pseudomolecules. The quality of the new
463 assembly was assessed with FRC_bam with the Illumina paired end reads generated at the
464 Genomic Facility of Science for Life Laboratory (Sweden) using the same genomic DNA
465 extraction used for PacBio sequencing.

466

467 **Annotation of the *Trypanosoma cruzi* Sylvio X10/1 Genome**: The genome sequence was
468 annotated using a new kinetoplastid genome annotation pipeline combining homology-based
469 gene model transfer with *de novo* gene prediction. To allow for the sensitive identification of
470 partial genes, input sequences were split at stretches of undefined bases, effectively creating
471 a set of 'pseudocontigs', each of which does not contain any gaps. Gene finding was then
472 performed on both the original sequences and the pseudocontigs using AUGUSTUS, which
473 also calls partial genes at the boundaries of each pseudocontig. AUGUSTUS models were
474 trained on 800 genes randomly sampled from the 41 Esmeraldo-type (TcII) *T. cruzi* CL Brener

475 chromosomes in GeneDB. Protein-DNA alignments of reference proteins against the new *T.*
476 *cruzi* sequences, generated using Exonerate, were additionally used to improve the accuracy
477 of the gene prediction. In addition, the RATT software was used to transfer highly conserved
478 gene models from the *T. cruzi* CL Brener annotation to the target. A non-redundant set of
479 gene models was obtained by merging the results of both RATT and AUGUSTUS and, for
480 each maximal overlapping set of gene models, selecting the non-overlapping subset that
481 maximizes the total length of the interval covered by the models, weighted by varying levels
482 of *a priori* assigned confidence. Spurious low-confidence protein coding genes with a reading
483 direction in disagreement with the directions of the polycistronic transcriptional units were
484 removed automatically. The result of this integration process was then merged with ncRNA
485 annotations produced by specific tools such as ARAGORN and Infernal. Finally, protein-DNA
486 alignments with frame shifts produced by BLAST were used in a computational approach to
487 identify potential pseudogenes in the remaining sequence.
488
489 Downstream of the structural annotation phase, gene models were automatically assigned
490 IDs and further extended with product descriptions and GO terms, both transferred from CL
491 Brener orthologs and inferred from Pfam protein domain hits and represented as feature
492 attributes or Sequence Ontology-typed subfeatures tagged with appropriate evidence codes.
493 This annotation pipeline has been implemented in the Companion web server. The assembled
494 genome was scanned for small RNAs using INFERNAL against the curated RFAM database
495 using cmsearch with a minimum e-value of 1e-10, a GC-bias of 0 and a minimum alignment
496 length of 10 nt. This annotation process has been implemented into the web-based
497 annotation pipeline COMPANION [43] from the Wellcome Trust Sanger Institute.
498
499 Repetitive sequences were annotated using RepeatMasker with the NCBI+ search engine
500 and LTRHarvest. Using the genomic coordinates of the repetitive elements, the genome was
501 split in windows of 10 Kb to identify VIPER and L1Tc retroelements adjacent to surface

502 molecule genes (i.e: trans-sialidases, mucins and MASP). A one-sided Fisher's exact test was
503 used to evaluate if the retroelements were enriched in genomic segments containing surface
504 molecule genes.
505
506 **Identification of Single Nucleotide Polymorphism (SNP) and Insertion/Deletion events**
507 **(Indels):** An improved short-read mapping strategy was used to assign the reads to their
508 target sequences with high accuracy, especially in regions rich in simple and low complexity
509 repeats, by taking advantage of the statistical read placement implemented in the Stampy
510 read mapper to accurately call genomic variants from the mapped reads. Reads from all 34
511 *T. cruzi* TcI isolates (SRA BioProject accession number: PRJNA325924) were mapped against
512 the assembled *T. cruzi* Sylvio X10/1 genome using a two-step mapping process to improve
513 the mapping of Illumina data to highly repetitive regions: First, reads were mapped using
514 BWA MEM with default parameters; later, the BAM file produced by BWA was remapped with
515 Stampy (v1.23) using the *--bamkeepgoodreads* option. The final mapping file was sorted and
516 filtered for PCR duplicates using Picard Tools v1.137. Variants were called using FreeBayes
517 with a minimum per-base quality of 30, minimum mapping quality of 30 and minimum
518 coverage of 15 bases. Variants that were found in a potentially misassembled region were
519 excluded from the analysis. Additionally, genomic variants were called using FermiKit - which
520 is an assembly-based variant caller - to validate the genomic variants observed in
521 subtelomeric regions. A consensus of the two methods was used as a final set of variants for
522 downstream analyses. Haplotypes were phased using Beagle r1399. The phased markers
523 were used for downstream analyses with SNPrelate and VCFtools and the functional effect
524 of the identified variants was predicted using SnpEff.
525
526 **Identification of Genomic Structural Variants (SV)**: Genomic structural variants were
527 identified whiting TcI isolates, using a consensus of different methods: Delly2, Lumpy,
528 FermiKit and FindTranslocations (https://github.com/vezzi/FindTranslocations.git) using both

529 raw reads and realigned BAM files. For each method, a SV muts had a depth of coverage >
530 10 reads and a mapping quality of > 30. Later, a consensus was created with all the SV that
531 were supported by all the methods. SVs that were supported by FermiKit and at least one of
532 the mapping-based methods were also included but labelled as 'Low Confidence'. SVs
533 identified by only one method were not included. Breakpoint analysis was done with custom
534 Python scripts and their functional effect was predicted using SNPeff. Analyses of copy
535 number variation (CNV) were done using the BAM files for each sample with the *Control-*
536 *FREEC* package.
537
538 **Accessions**: Whole genome sequence Genbank accession: ADWP00000000 (CP015651-
539 CP015697). Reads for the 34 TcI isolates SRA accession: SRP076682.
540
541 **Authors contributions**: CT-L and BA conceived and designed the study. CT-L designed and
542 executed computational analyses. MY prepared Sylvio X10/1 genomic DNA for PacBio
543 sequencing and performed manual annotation of surface molecule genes. JEC, AS, JDR, FG,
544 SO-M, JAC, SMRT, HC, RG, KJ, MEB, PJH, KOM, MJG, BB provided genomic DNA for TcI
545 isolates. JLR-C and DCB created chromosome maps for surface molecules. MAM, LAM, ML,
546 JLR-C and ECG contributed to the interpretation of the results. CT-L, ECG, MAM, BA wrote
547 the manuscript. All authors read and approved the final version of the manuscript.
548
549
550
551
552
553
554
555

Acknowledgements: The authors wish to thank to John Kuijpers and Lawrence Hon from Pacific Biosciences (PacBio) for providing the data for the Sylvio X10/1 genome and preliminary assembly results. We would also like to acknowledge Eric Dumonteil, University of the Yucatan, Mexico, and Ed Wozniak, Texas Department of State Health Services, for past work on parasite collection. This research was funded by grants from the Knut and Alice Wallenberg Foundation, The Swedish Research Council and the European FP7 program. LAM was funded by a grant from the NIH; DCB, ECG and SMRT were funded by grants from CNPq (Brazilian Government Agency).

565 **TABLES**:

566

567 **Table 1: *Trypanosoma cruzi* Sylvio X10/1 strain (Tc-I) genome assembly.**

Metric	Value
Genome size	41.3 Mbp
Number of scaffolds	47
Percentage of reconstruction	98.5 %
Longest scaffold	3.1 Mbp
Shortest scaffold	404 Kb
NG50	1.0 Mbp
N50	1.1 Mbp

568

569 **Table 2: Genomic variants identified among the *Trypanosoma cruzi* TcI isolates.**

GROUP	SNPs	INDELs	DELETION	DUPLICATION	TRANSLOCATION
Colombia*	158565	59520	439	1231	4140
Colombiana**	105023	30697	23	86	273
Venezuela	77232	70086	43	183	614
Ecuador	122122	84201	40	164	354
Panama	620499	238833	225	605	2060
Texas	101771	78499	69	303	978

570 * - FcHc and CG clones from Colombia

571 ** - TcI Colombiana strain

572

573 **Table 3: Patterns of INDELs and their associated mechanisms of origin.**

INDEL type	Example	Mechanism	Frequency*
HR - deletion	GCATAAA*aa*AAAGC	NAHR	756 411
HR - insertion	CACA*AAAAAAAAAAA*GCTAC	NAHR	521 002
TR - mixed	ACACAC*acac*ACACAC*AC*AC	NAHR	118 432
Non-repetitive	TAGCAC*agt*GACTTCACA*GC*CTG	NHEJ	28 389
Long Insertion	C*GGCTAGACCAGGTACAGTC*A	MMEJ	32 666
Long Deletion	GC*acactgacacgacactgacacactgaa*A	MMEJ	31 712

574

575 HR = Homopolymer run

576 TR = Tandem Repeat

577 = Deletion

578 = Insertion

579 * For all the 34 TcI genomes compared against Sylvio X10/1

580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603

604 **SUPPLEMENTARY DATA:**

605

606 **Supplementary Table 1: *Trypanosoma cruzi* strains used in this study.**

Strain	Country	Source	Clinical form	Sequencing coverage*
H1a	Panama	Human	Chronic	31 X
H2	Panama	Human	Asymptomatic	28 X
H3	Panama	Human	Asymptomatic	32 X
H4	Panama	Human	No Disease	29 X
H5	Panama	Human	Chronic	31 X
H6	Panama	Human	Asymptomatic	31 X
H7	Panama	Human	Asymptomatic	28 X
H8	Panama	Human	No Disease	32 X
H12	Panama	Human	No Disease	31 X
H14	Panama	Human	Chronic	32 X
H15	Panama	Human	Chronic	31 X
V1	Panama	*P. geniculatus*	ND	28 X
V2	Panama	*R. pallescens*	ND	25 X
V3	Panama	*T. dimidiata*	ND	28 X
TBM3324	Ecuador	*R. ecuadoriensis*	ND	28 X
TBM3406B1	Ecuador	*R. ecuadoriensis*	ND	25 X
TBM3479B1	Ecuador	*R. ecuadoriensis*	ND	29 X
TBM3519W1	Ecuador	*R. ecuadoriensis*	ND	28 X
X10462-P1C9	Venezuela	Human	ND	31 X
X12422-P1C3	Venezuela	Human	ND	29 X
Colombiana	Colombia	Human	ND	30 X
CGI10	Colombia	Human	Acute Co-infection	30 X
CGI11	Colombia	Human	Acute Co-infection	30 X
CGI12	Colombia	Human	Acute Co-infection	30 X
CGI13	Colombia	Human	Acute Co-	30 X

			infection	
CGI14	Colombia	Human	Acute Co-infection	30 X
CGI15	Colombia	Human	Acute Co-infection	30 X
FcHcl1	Colombia	Human	Acute Oral	30 X
FcHcl2	Colombia	Human	Acute Oral	30 X
FcHcl3	Colombia	Human	Acute Oral	30 X
FcHcl4	Colombia	Human	Acute Oral	30 X
FcHcl5	Colombia	Human	Acute Oral	30 X
H1b	Mexico	Human	Acute	32 X
TD23	Texas	*T. dimidiata*	ND	30 X
TD25	Texas	*T. dimidiata*	ND	30 X

607

608 NA = Not determined.

609 * Illumina data

610

611 **FIGURE LEGENDS**:

612 **Figure 1**: **a**) Distribution of surface molecule gene tandem arrays in the 47 chromosomes of
613 the TcI Sylvio X10/1 genome. In this image, each line correspond to a pseudomolecule drawn
614 in proportion to its size, where genes corresponding to the largest *T. cruzi* multigene families
615 are represented by colored boxes, hypothetical genes are represented by grey boxes and
616 the other genes as black boxes. The position of the gene box above or below correspond
617 respectively to its presence in the + or – strand. **b)** Comparison of chromosome 15 from the
618 TcVI CL Brener and TcI Sylvio X10/1 assemblies. Grey lines between pseudomolecules
619 represent regions of synteny between orthologous genes, and yellow blocks represent gaps
620 in both genome assemblies. Coloured blocks represent paralogs of multicopy gene families.
621 **c**) Distribution of surface molecule gene tandem arrays in the 47 chromosomes of the TcI
622 Sylvio X10/1 genome.

623

624 **Figure 2**: **a**) Bayesian principal component analysis (PCA) of *T. cruzi* TcI strains using INDELs
625 and **b**) Identity by Descent (IBD) dendogram of *T. cruzi* TcI strains using SNPs. Both analyses,
626 using different markers, support the population structure of the analysed TcI samples.
627 Notably, the highly virulent TcI Colombiana and the Panamanian TcI H1 from a chronic patient
628 are presented as outliers (**b,** far left).

629

630 **Figure 3**: **a**) Linkage disequilibrium matrix (r2) of chromosome 2 for the Colombian CG and
631 FcHc clones. LD values range from 0 (recombination) to 1 (no recombination). **b**) Genome-
632 wide *Fst* distribution in 10 Kb bins displaying a state of panmixia for the CG clones and
633 moderate genetic differentiation in the FcHc clones, yellow dots represent outlier bins. **c**)
634 Distribution of subtelomeric *Tajima's D* selection test in both groups displaying overall
635 balancing selection ($D > 0$) in these regions for both clones.

636

637 **Figure 4**: **a**) Proposed mechanism of inter-chromosomal recombination between gene
638 tandem arrays for the generation of antigenic diversity. Here, VIPER retrotransposons (green)
639 are surrounding a tandem array, shown in the direction of transcripton, containing trans-
640 sialidases (blue) and pseudogenes (red) in pseudomolecules 1 and 8. The rearranged tandem
641 array is shown below displaying the merger of genes and pseudogenes to form a longer
642 coding sequence. **b**) Detail of simple repeats and retrotransposons within a trans-sialidase
643 tandem array in pseudomolecule 1.
644
645 **Figure 5**: Distribution of CNV changes in chromosome 3 of the Colombian **a**) CG clones and
646 **b**) FcHc clones. Black lines represent the reference genome sequence and cyan lines
647 represent the sample under study. Each sliding window for CNV evaluation is represented as
648 a dot. A drastic change in CNV can be noted in the FcHc clones whereas the CG clones seem
649 to be less affected.
650
651 **Supplementary Figure 1**: Distribution of coverage for **a**) PacBio and **b**) Illumina reads in a
652 subtelomeric end on chromosome 8 showing a well supported genome assembly of these
653 regions.
654
655 **Supplementary Figure 2**: Genome-wide Neighbour Joining tree for trans-sialidases genes
656 from the Sylvio X10/1 genome showing the high level of sequence diversity for these gene
657 family.
658
659 **Supplementary Figure 3**: Principal component variability for the genotype diversity of the 34
660 TcI isolates.
661

662 **REFERENCES**

663 1. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national
664 age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-
665 2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **385,**
666 117–171 (2015).
667 2. WHO | 6 February 2015, vol. 90, 6 (pp. 33–44). (2015).
668 3. Garcia, M. N. *et al.* Evidence of autochthonous Chagas disease in southeastern Texas.
669 *Am. J. Trop. Med. Hyg.* **92,** 325–330 (2015).
670 4. Bern, C. Chagas' Disease. *N. Engl. J. Med.* **373,** 456–466 (2015).
671 5. Rassi, A., Rassi, A. & Marin-Neto, J. A. Chagas disease. *Lancet* **375,** 1388–1402 (2010).
672 6. Morillo, C. A. *et al.* Randomized Trial of Benznidazole for Chronic Chagas'
673 Cardiomyopathy. *N. Engl. J. Med.* **373,** 1295–1306 (2015).
674 7. Pecoul, B. *et al.* The BENEFIT Trial: Where Do We Go from Here? *PLoS Negl. Trop. Dis.*
675 **10,** e0004343 (2016).
676 8. Osorio, L., Ríos, I., Gutiérrez, B. & González, J. Virulence factors of *Trypanosoma cruzi*:
677 who is who? *Microbes Infect.* **14,** 1390–1402 (2012).
678 9. El-Sayed, N. M. *et al.* The genome sequence of *Trypanosoma cruzi*, etiologic agent of
679 Chagas disease. *Science* **309,** 409–415 (2005).
680 10. Weatherly, D. B., Boehlke, C. & Tarleton, R. L. Chromosome level assembly of the
681 hybrid *Trypanosoma cruzi* genome. *BMC Genomics* **10,** 255 (2009).
682 11. Franzén, O. *et al.* Shotgun sequencing analysis of Trypanosoma cruzi I Sylvio X10/1
683 and comparison with T. cruzi VI CL Brener. *PLoS Negl. Trop. Dis.* **5,** e984 (2011).
684 12. Franzén, O. *et al.* Comparative genomic analysis of human infective Trypanosoma cruzi
685 lineages with the bat-restricted subspecies T. cruzi marinkellei. *BMC Genomics* **13,** 531
686 (2012).
687 13. Kim, D. *et al.* Telomere and subtelomere of Trypanosoma cruzi chromosomes are

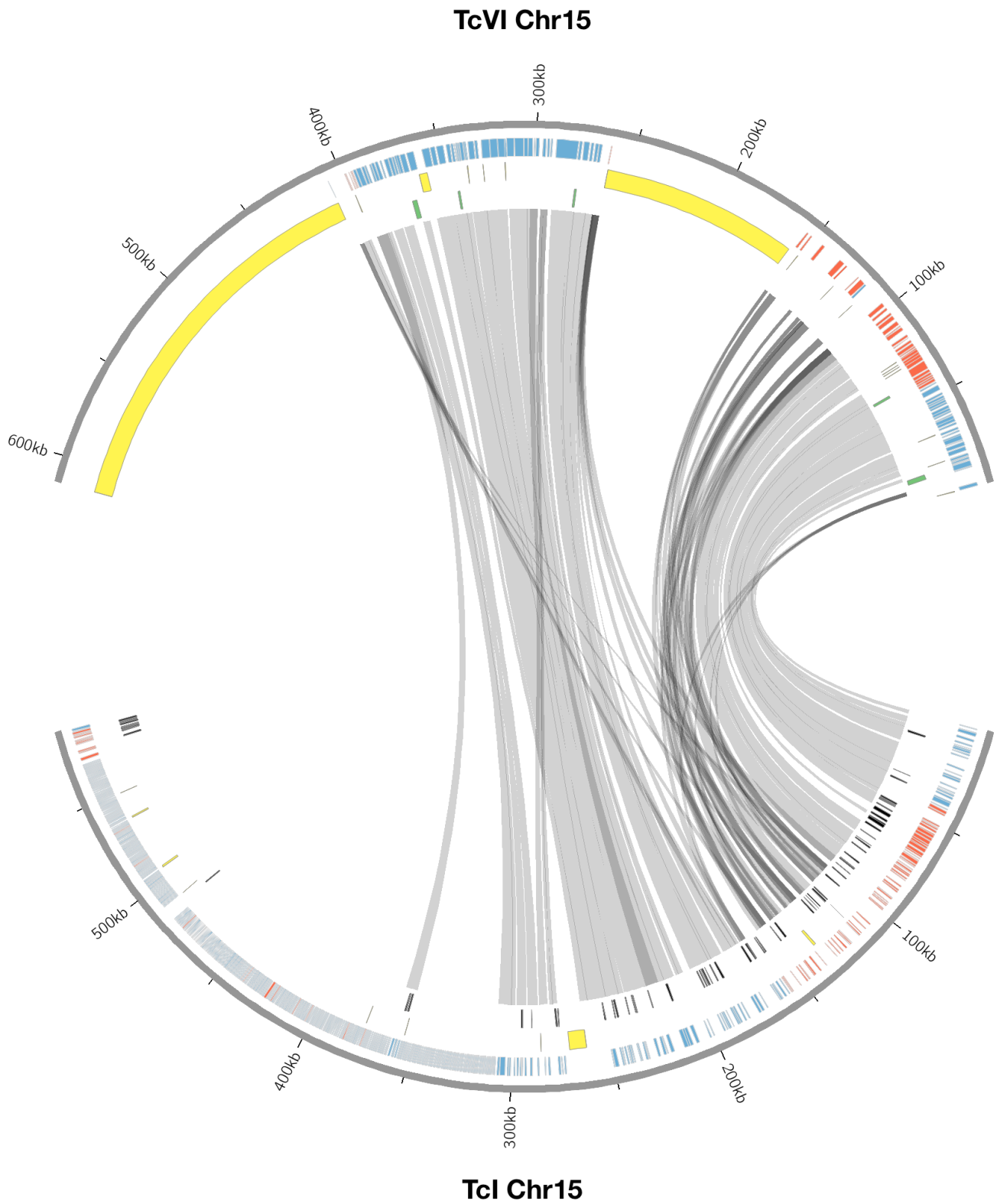
688 enriched in (pseudo)genes of retrotransposon hot spot and trans-sialidase-like gene
689 families: the origins of T. cruzi telomeres. *Gene* **346,** 153–161 (2005).
690 14. Zingales, B. *et al.* The revised Trypanosoma cruzi subspecific nomenclature: rationale,
691 epidemiological relevance and research applications. *Infect. Genet. Evol.* **12,** 240–253
692 (2012).
693 15. Llewellyn, M. S. *et al.* Genome-scale multilocus microsatellite typing of Trypanosoma
694 cruzi discrete typing unit I reveals phylogeographic structure and specific genotypes
695 linked to human infection. *PLoS Pathog.* **5,** e1000410 (2009).
696 16. Ramírez, J. D. *et al.* Chagas cardiomyopathy manifestations and Trypanosoma cruzi
697 genotypes circulating in chronic Chagasic patients. *PLoS Negl. Trop. Dis.* **4,** e899
698 (2010).
699 17. Montgomery, S. P., Starr, M. C., Cantey, P. T., Edwards, M. S. & Meymandi, S. K.
700 Neglected parasitic infections in the United States: Chagas disease. *Am. J. Trop. Med.*
701 *Hyg.* **90,** 814–818 (2014).
702 18. Bern, C., Kjos, S., Yabsley, M. J. & Montgomery, S. P. Trypanosoma cruzi and Chagas'
703 Disease in the United States. *Clin. Microbiol. Rev.* **24,** 655–681 (2011).
704 19. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells.
705 *Nat. Genet.* **41,** 563–571 (2009).
706 20. Claessens, A. *et al.* Generation of antigenic diversity in *Plasmodium falciparum* by
707 structured rearrangement of Var genes during mitosis. *PLoS Genet.* **10,** e1004812
708 (2014).
709 21. Hall, J. P. J., Wang, H. & Barry, J. D. Mosaic VSGs and the scale of *Trypanosoma*
710 *brucei* antigenic variation. *PLoS Pathog.* **9,** e1003502 (2013).
711 22. Andersson, B. The *Trypanosoma cruzi* genome; conserved core genes and extremely
712 variable surface molecule families. *Res. Microbiol.* **162,** 619–625 (2011).
713 23. Stoco, P.H., *et al*. Genome of the avirulent human-infective trypanosome--Trypanosoma

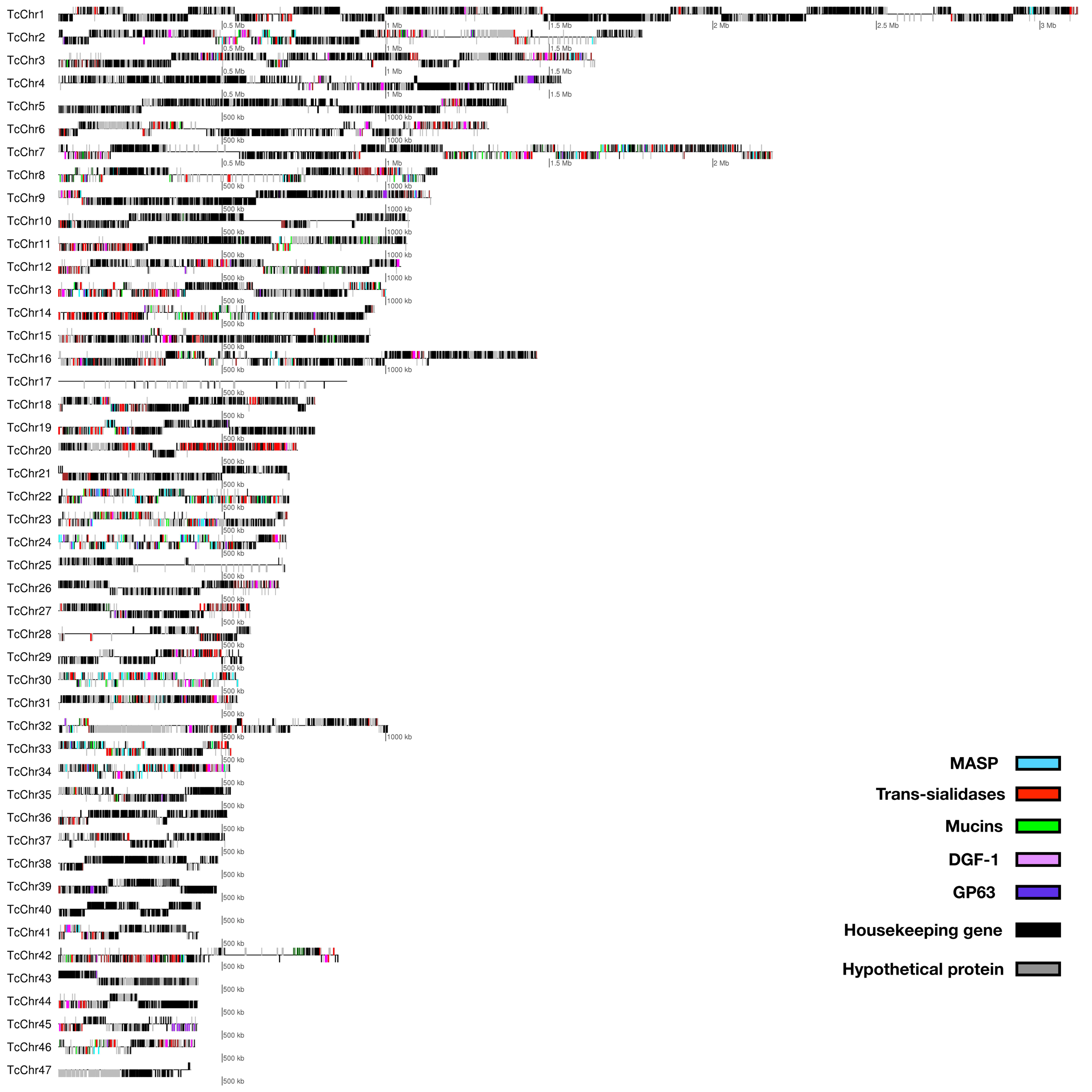
714 rangeli. *PLoS Neglected Tropical Diseases* **8(9)**, e3176 (2014).
715 23. Guhl, F. & Ramírez, J. D. Trypanosoma cruzi I diversity: towards the need of genetic
716 subdivision? *Acta Trop.* **119,** 1–4 (2011).
717 24. Parks, M. M., Lawrence, C. E. & Raphael, B. J. Detecting non-allelic homologous
718 recombination from high-throughput sequencing data. *Genome Biol.* **16,** 72 (2015).
719 25. Sfeir, A. & Symington, L. S. Microhomology-Mediated End Joining: A Back-up Survival
720 Mechanism or Dedicated Pathway? *Trends Biochem. Sci.* **40,** 701–714 (2015).
721 26. Weckselblatt, B., Hermetz, K. E. & Rudd, M. K. Unbalanced translocations arise from
722 diverse mutational mechanisms including chromothripsis. *Genome Res.* **25,** 937–947
723 (2015).
724 27. Minning, T. a., Weatherly, D. B., Flibotte, S. & Tarleton, R. L. Widespread, focal copy
725 number variations (CNV) and whole chromosome aneuploidies in *Trypanosoma cruzi*
726 strains revealed by array comparative genomic hybridization. *BMC Genomics* **12,** 139
727 (2011).
728 28. Miles, A. *et al.* Indels, structural variation, and recombination drive genomic diversity in
729 *Plasmodium falciparum*. *Genome Res.* **26,** 1288–1299 (2016).
730 29. Tibayrenc, M., Kjellberg, F. & Ayala, F. J. A clonal theory of parasitic protozoa: the
731 population structures of *Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium,*
732 *Trichomonas,* and *Trypanosoma* and their medical and taxonomical consequences.
733 *Proc. Natl. Acad. Sci. U. S. A.* **87,** 2414–2418 (1990).
734 30. Tibayrenc, M. & Ayala, F. J. Towards a population genetics of microorganisms: The
735 clonal theory of parasitic protozoa. *Parasitol. Today* **7,** 228–232 (1991).
736 31. Gaunt, M. W., Yeo, M., Frame, I. A. & Stothard, J. R. Mechanism of genetic exchange
737 in American trypanosomes. **421,** (2003).
738 32. Westenberger, S. J., Barnabé, C., Campbell, D. a. & Sturm, N. R. Two hybridization
739 events define the population structure of Trypanosoma cruzi. *Genetics* **171,** 527–543

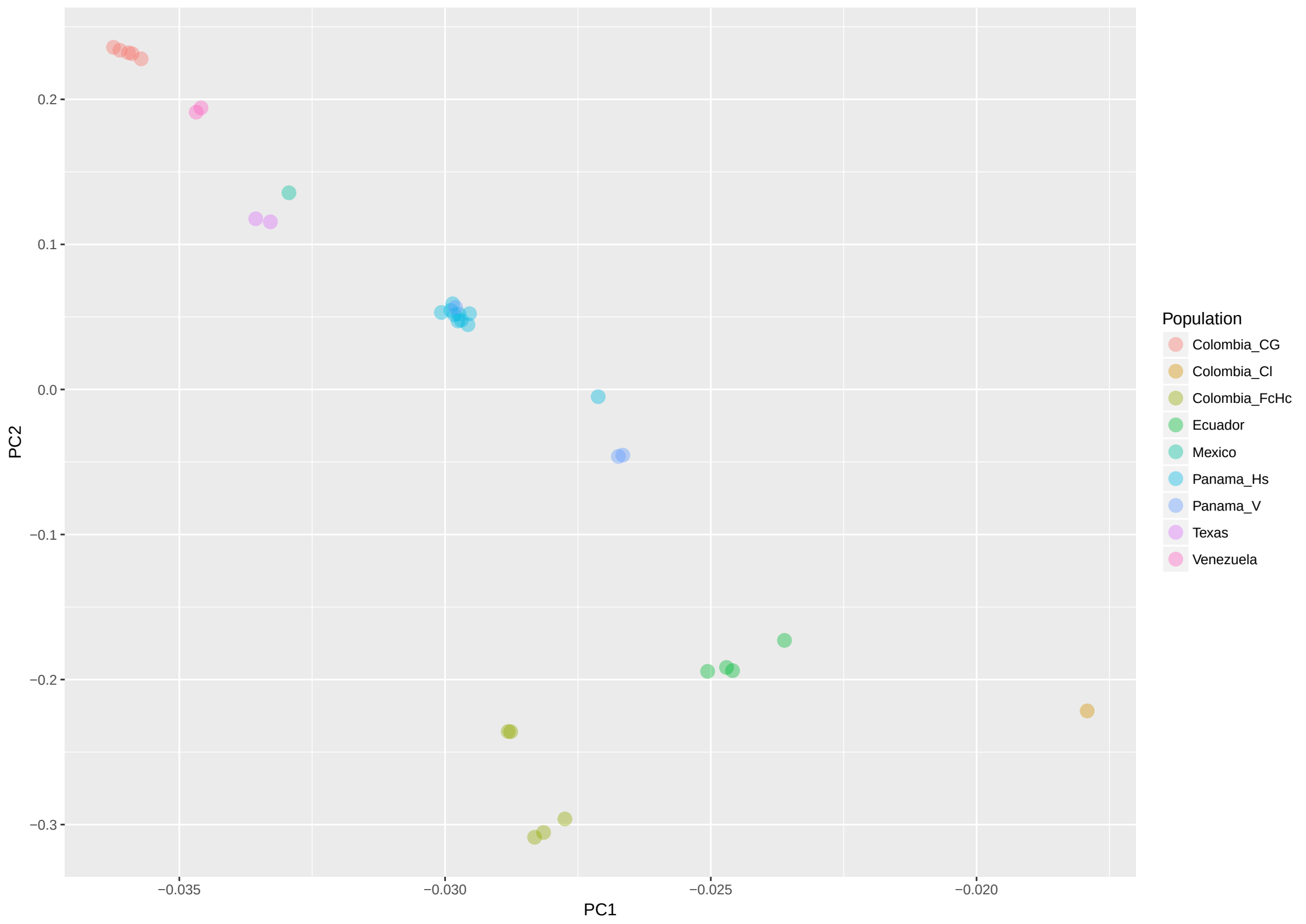
740 (2005).
741 33. Tibayrenc, M. & Ayala, F. J. The population genetics of Trypanosoma cruzi revisited in
742 the light of the predominant clonal evolution model. *Acta Trop.* **151,** 156–165 (2015).
743 34. Messenger, L. A. & Miles, M. A. Evidence and importance of genetic exchange among
744 field populations of Trypanosoma cruzi. *Acta Trop.* **151,** 150–155 (2015).
745 35. Aksenova, A. Y. *et al.* Genome rearrangements caused by interstitial telomeric
746 sequences in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 19866–19871 (2013).
747 36. de Jonge, R. *et al.* Extensive chromosomal reshuffling drives evolution of virulence in
748 an asexual pathogen. *Genome Res.* **23,** 1271–1282 (2013).
749 37. McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science* **342,**
750 632–637 (2013).
751 38. Ramesh, M. A., Malik, S.-B. & Logsdon, J. M., Jr. A Phylogenomic Inventory of Meiotic
752 Genes: Evidence for Sex in Giardia and an Early Eukaryotic Origin of Meiosis. *Curr.*
753 *Biol.* **15,** 185–191 (2005).
754 39. Villanueva, M. S., Williams, S. P., Beard, C. B., Richards, F. F. & Aksoy, S. A new
755 member of a family of site-specific retrotransposons is present in the spliced leader
756 RNA genes of Trypanosoma cruzi. *Mol. Cell. Biol.* **11,** 6139–6148 (1991).
757 40. Erwin, J. A., Marchetto, M. C. & Gage, F. H. Mobile DNA elements in the generation of
758 diversity and complexity in the brain. *Nat. Rev. Neurosci.* **15,** 497–506 (2014).
759 41. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene
760 expression. *Science* **351,** aac7247 (2016).
761 42. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and
762 locality-sensitive hashing. *Nat. Biotechnol.* 1–11 (2015).
763 43. Steinbiss, S. *et al.* Companion: a web server for annotation and analysis of parasite
764 genomes. *Nucleic Acids Res.* **44,** W29–34 (2016).

765 44. Faino, L. *et al*. Transposons actively and passively contributes to the evolution of the

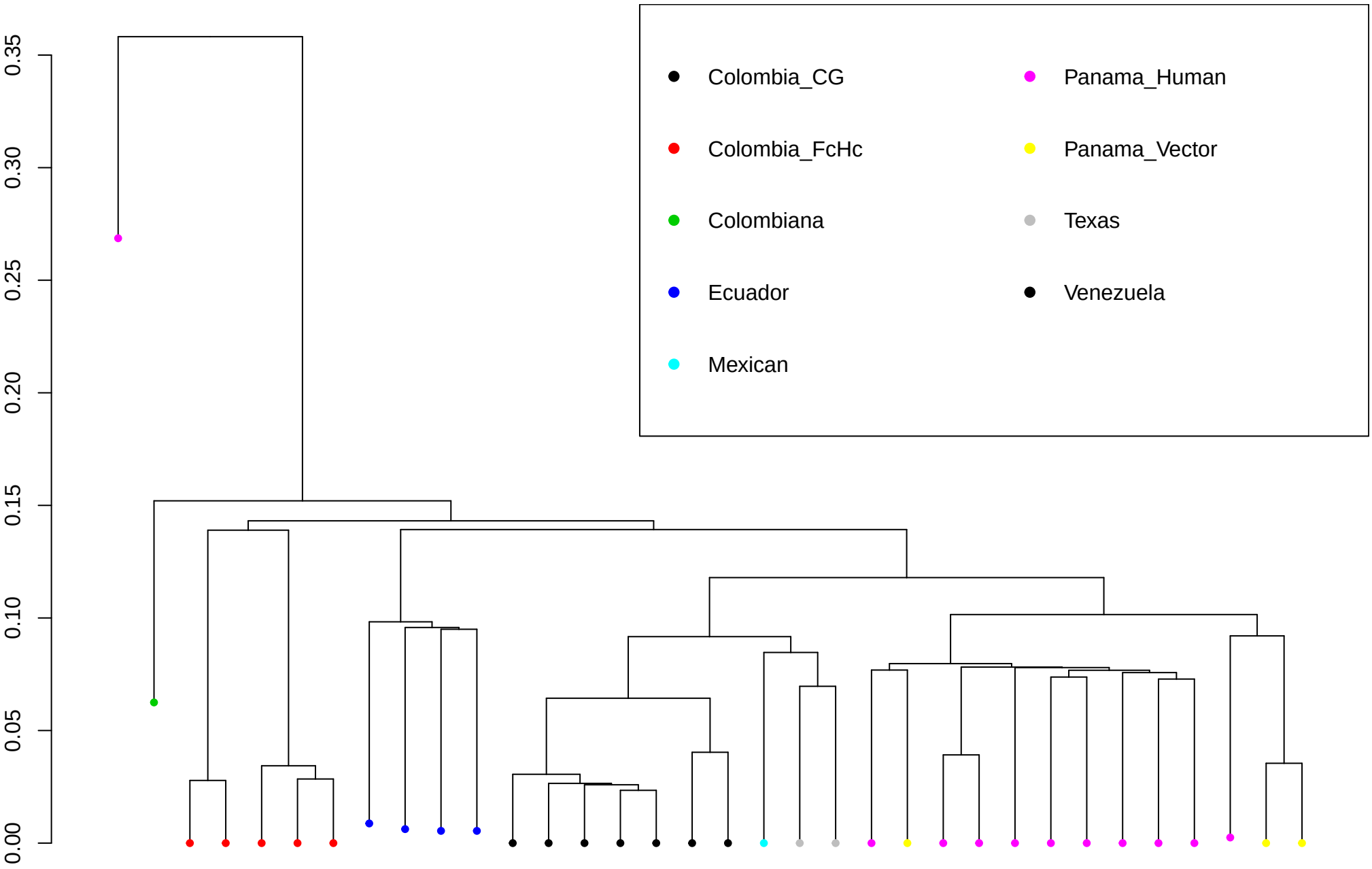
two-speed genome of a fungal pathogen. *Genome Res.* **26**: 1091 – 1100 (2016).

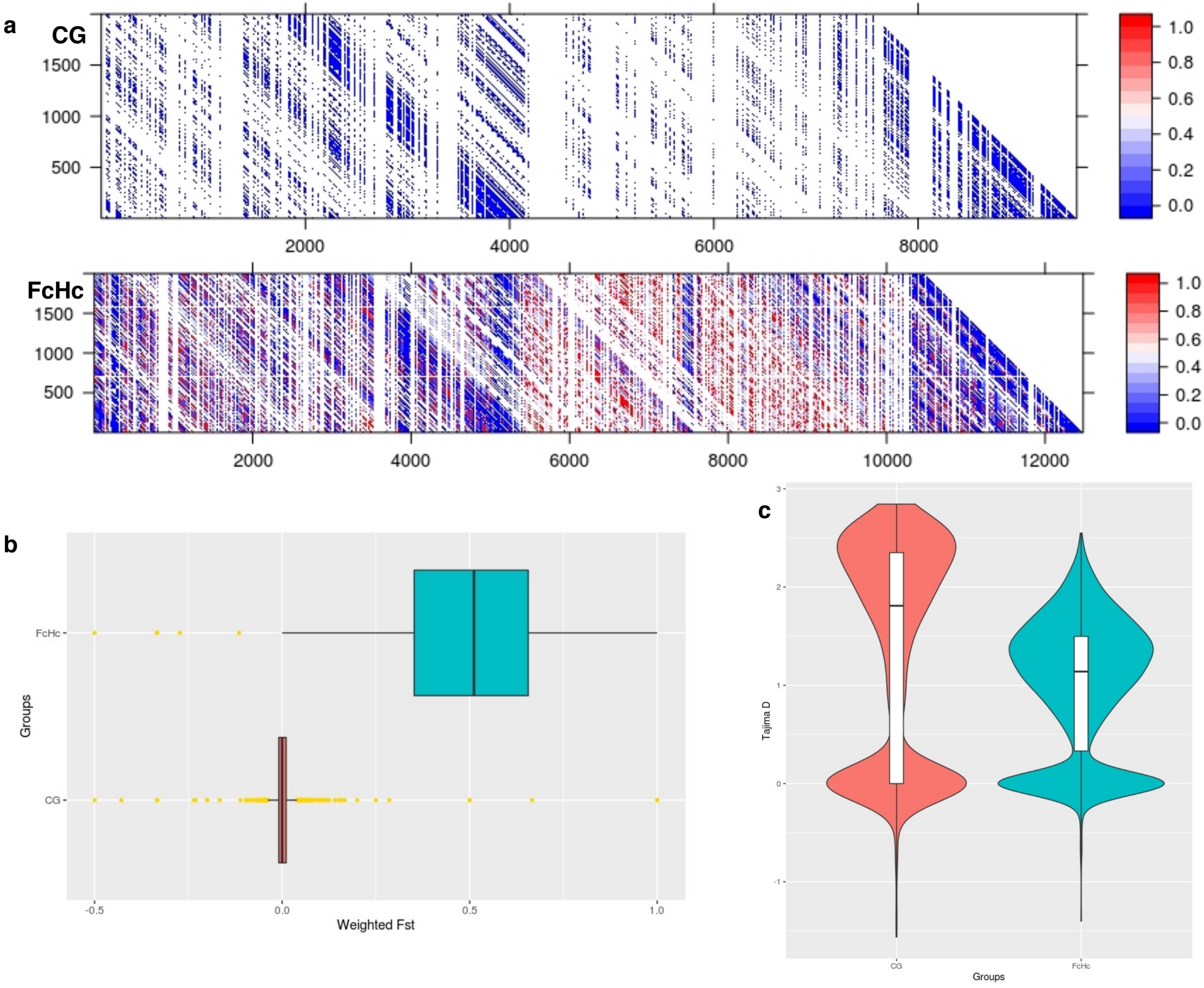

TcVI Chr15
300kb
200kb
100kb
400kb
500kb
600kb
500kb
400kb
300kb
200kb
100kb
TcI Chr15


TcChr1
TcChr2
TcChr3
TcChr4
TcChr5
TcChr6
TcChr7
TcChr8
TcChr9
TcChr10
TcChr11
TcChr12
TcChr13
TcChr14
TcChr15
TcChr16
TcChr17
TcChr18
TcChr19
TcChr20
TcChr21
TcChr22
TcChr23
TcChr24
TcChr25
TcChr26
TcChr27
TcChr28
TcChr29
TcChr30
TcChr31
TcChr32
TcChr33
TcChr34
TcChr35
TcChr36
TcChr37
TcChr38
TcChr39
TcChr40
TcChr41
TcChr42
TcChr43
TcChr44
TcChr45
TcChr46
TcChr47
0.5 Mb
1 Mb
1.5 Mb
2 Mb
2.5 Mb
3 Mb
500 kb
1000 kb
MASP
Trans-sialidases
Mucins
DGF-1
GP63
Housekeeping gene
Hypothetical protein

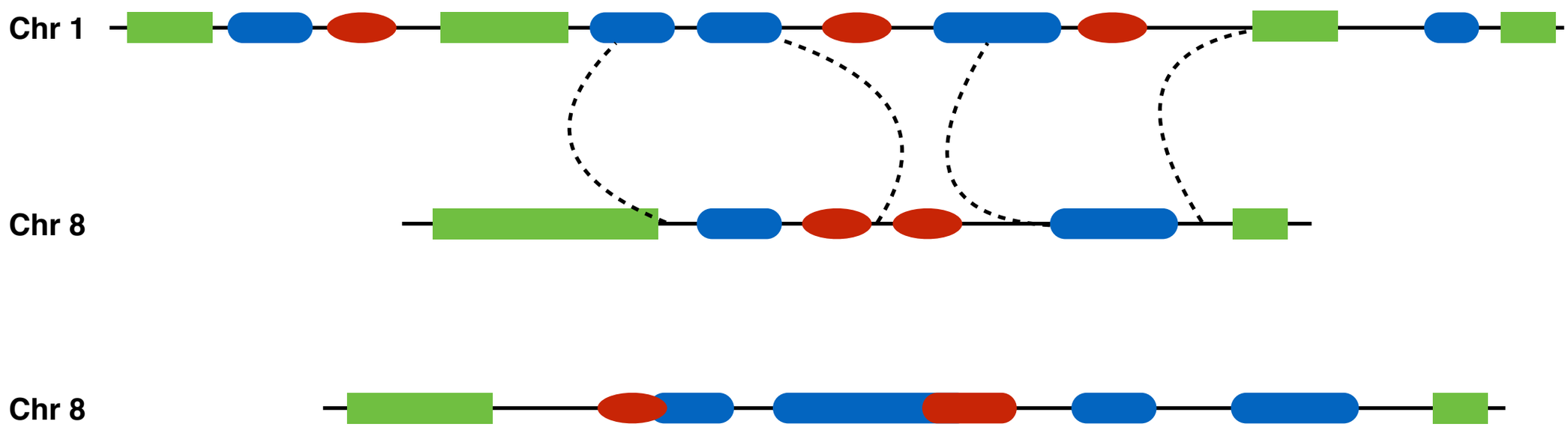

0.2
0.1
0.0
−0.1
−0.2
−0.3
PC2
−0.035
−0.030
−0.025
−0.020
PC1
Population
Colombia_CG
Colombia_CI
Colombia_FcHc
Ecuador
Mexico
Panama_Hs
Panama_V
Texas
Venezuela

Trypanosoma cruzi TcI IBD Clustering




a
CG
1500
1000
500
2000
4000
6000
8000
1.0
0.8
0.6
0.4
0.2
0.0
FcHc
1500
1000
500
2000
4000
6000
8000
10000
12000
1.0
0.8
0.6
0.4
0.2
0.0
b
FcHc
CG
Groups
-0.5
0.0
0.5
1.0
Weighted Fst
c
3
2
1
0
-1
Tajima D
CG
FcHc
Groups

a

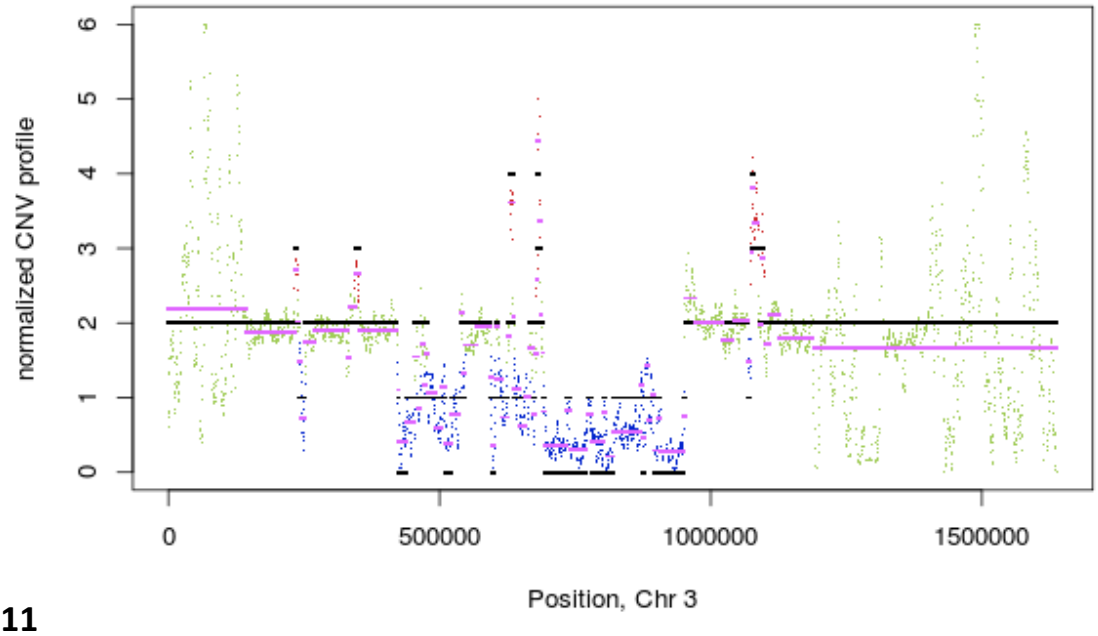

Chr 1
Chr 8
Chr 8

b

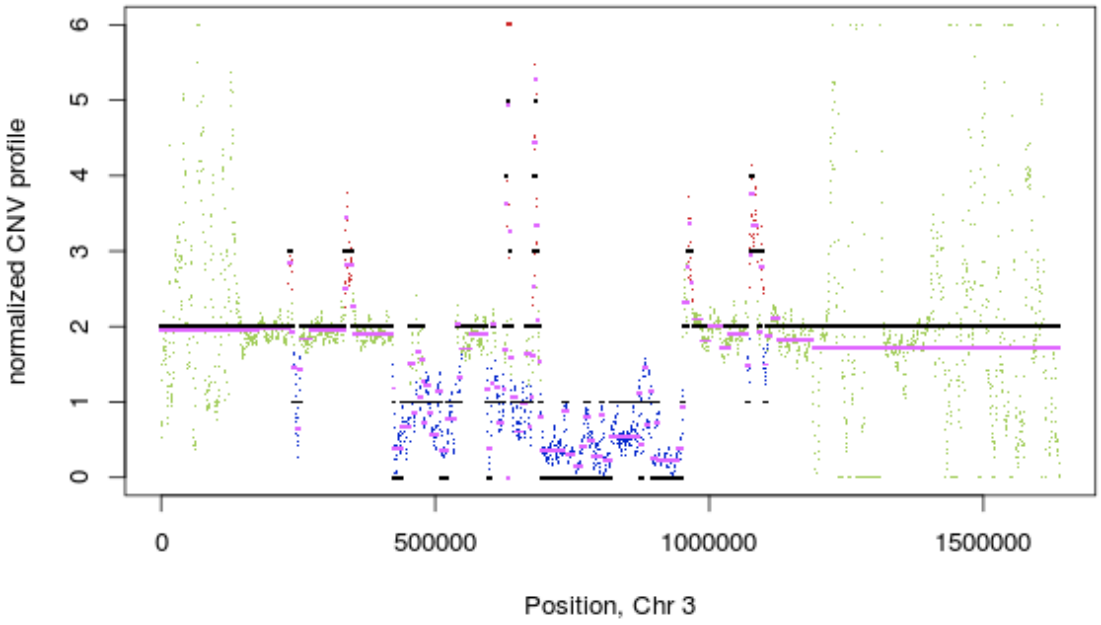
Trans-sialidase
AAAAATTTTTATATATATAT
LTR
VIPER
LTR
ACGACGACGACG

Distribution of CNV in chromosome 3 of the Colombian CG clones

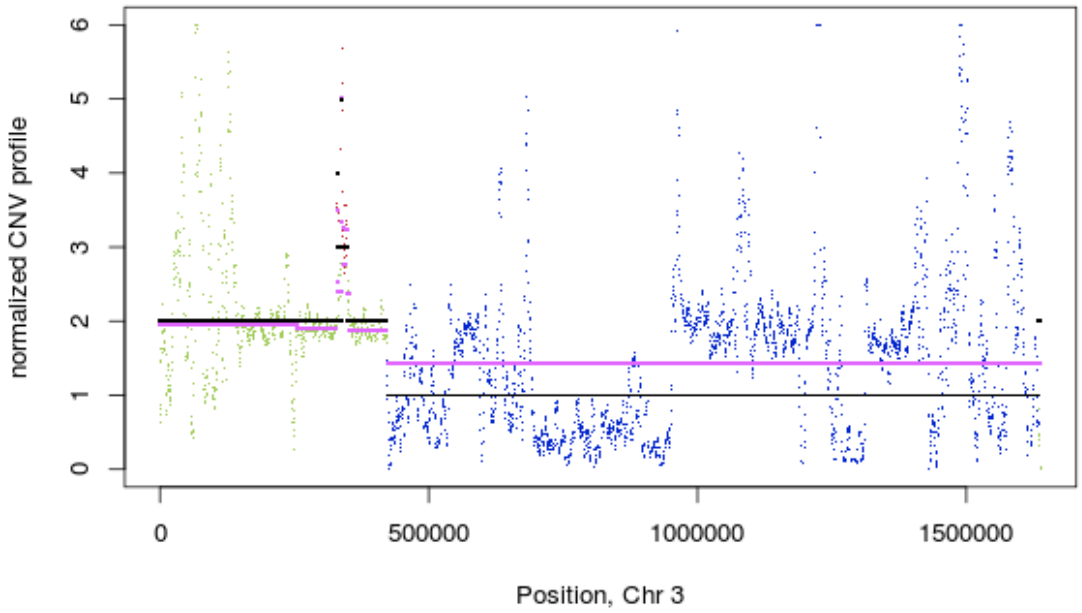
CGl10



CGl11

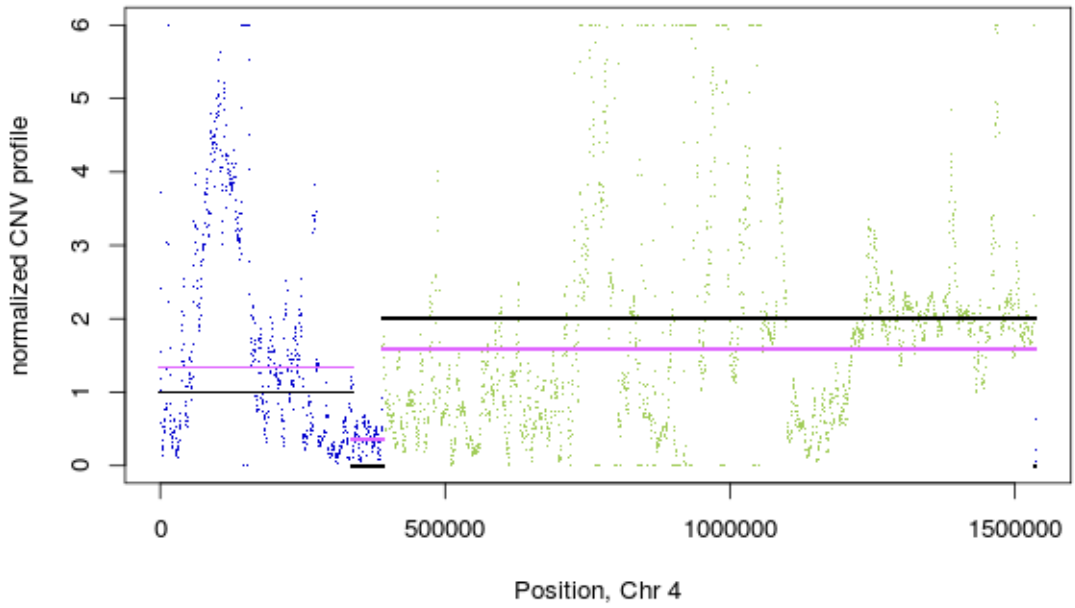


CGl13

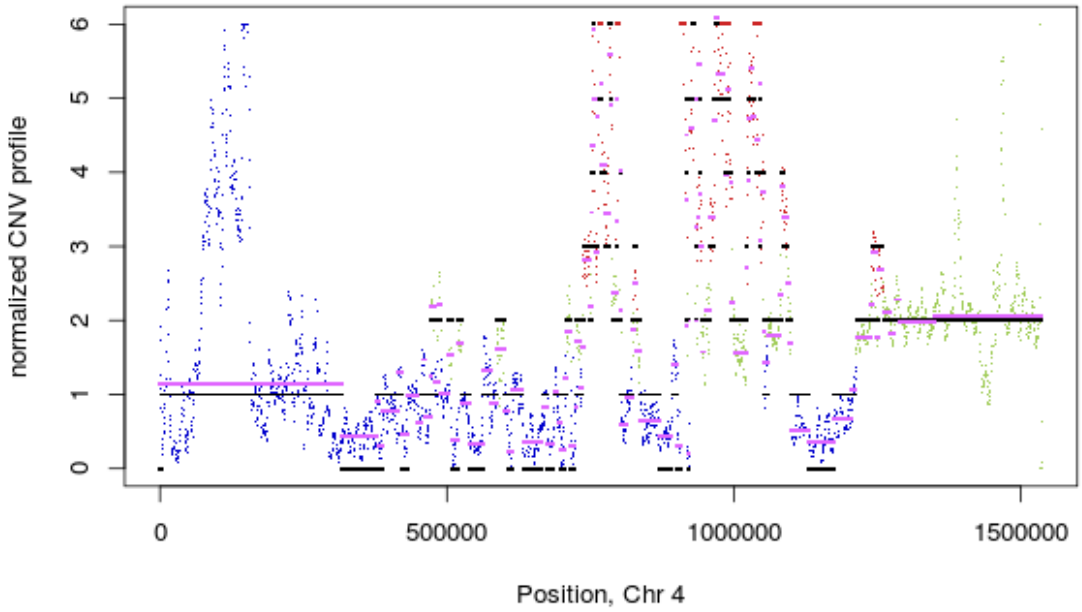
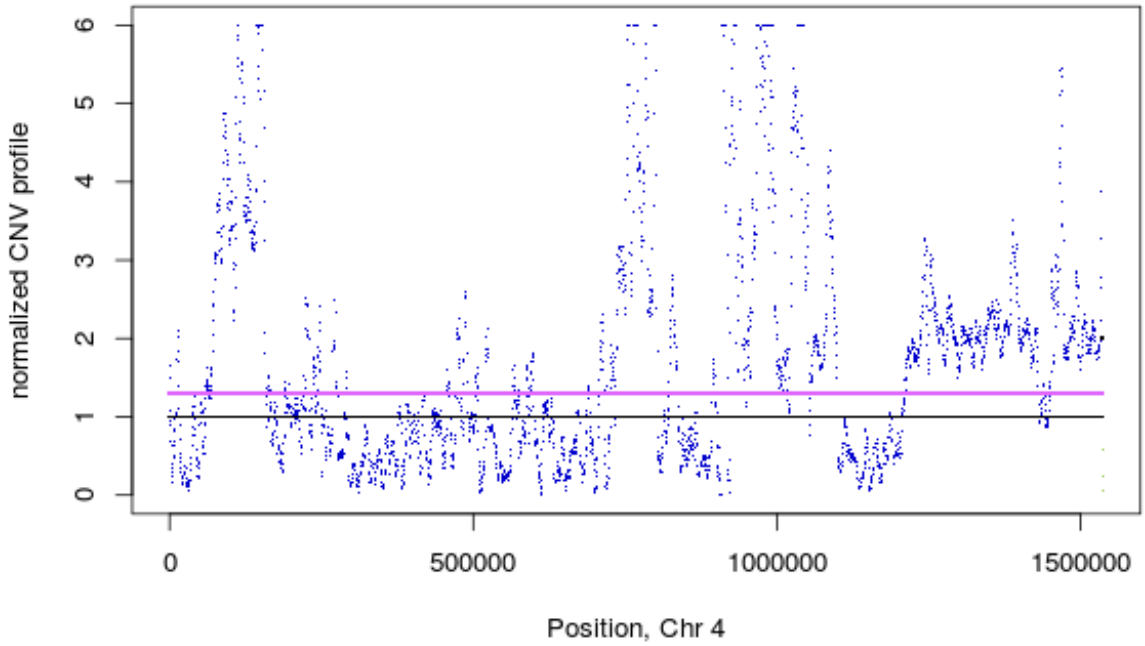


Distribution of CNV in chromosome 4 of the Colombian FcHc clones

FcHc1



FcHc2



FcHc4