bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

Carlos Talavera-López[1Φ], Louisa A. Messenger[2], Michael D. Lewis[2], Matthew Yeo[2], João Luís Reis-Cunha[3], 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.

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

12 - Infectious and Tropical Disease Institute, Department of Biomedical Sciences, Heritage College of Osteopathic Medicine, Ohio University; Athens OH, United States.

13 - T.H. Chan School of Public Health; Harvard University; Boston MA, United States.

* - Corresponding author

Φ - Current address: The Francis Crick Institute; London, United Kingdom.

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ABSTRACT:

Trypanosoma cruzi, a zoonotic kinetoplastid protozoan with a complex genome, is the causative agent of American trypanosomiasis (Chagas disease). The parasite uses a highly diverse repertoire of surface molecules, with roles in cell invasion, immune evasion and pathogenesis. Thus far, the genomic regions containing these genes have been impossible to resolve and it has been impossible to study the structure and function of the several thousand repetitive genes encoding the surface molecules of the parasite. We here present an improved genome assembly of a *T. cruzi* clade I (TcI) strain using high coverage PacBio single molecule sequencing, together with 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 structure reveals an unusual duality in the organisation of the parasite genome, a core genomic region syntenous with related protozoa flanked by unique and highly plastic subtelomeric regions encoding surface antigens. The presence of abundant interspersed retrotransposons in the subtelomeres suggests that these elements are involved in a recombination mechanism for the generation of antigenic variation and evasion of the host immune response. 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.

INTRODUCTION:

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

109 *T. cruzi* infection is a zoonosis, and the parasite has a complex life cycle; where transmission to
110 humans occurs most frequently by contamination with infected feces from triatomine insect
111 vectors (Subfamily Triatominae). The parasite evades the immune responses with the aid of
112 multiple surface molecules from three large diverse gene families (Trans-Sialidases, Mucins and
113 Mucin-Associated Surface Proteins - MASPs), which are also involved in cell invasion and possibly
114 pathogenicity [8].
115
116 Six distinct genetic clades of *T. cruzi* have been recognised, named TcI to TcVI (discrete typing units,
117 DTU-I to VI). The first genome sequence for *T. cruzi* was produced using Sanger sequencing
118 technology from a hybrid, highly polymorphic, TcVI strain. The resultant genome sequence, while
119 extremely useful for the core regions of the genome, was highly fragmented, especially in repetitive
120 regions [9]. This sequence has been improved using enhanced scaffolding algorithms, but the
121 repetitive regions remained unresolved [10]. Subsequently, FLX 454 Titanium and Illumina sequencing
122 were used to sequence a less polymorphic TcI strain (Sylvio X10/1), which allowed the first
123 comparative genomic studies of *T. cruzi*, but correct assembly of repetitive regions was still
124 impossible [11,12]. The thousands of related genes that code for the surface proteins are generally
125 located in large, subtelomeric regions of the *T. cruzi* genome [13], in the form of extremely repetitive
126 segments with multiple gene copies and pseudogenes. These subtelomeric regions are distinct from
127 the core regions of the genome in synteny, gene content and diversity [22]. The repetitive nature of
128 the tandem arrays and the length of the subtelomeric repeat regions, have made correct assembly
129 impossible using short and medium-sized sequence reads. The available *T. cruzi* genome sequences
130 are therefore incomplete and erroneous in these important regions, making it impossible to study
131 the complex surface gene families.
132
133 The population structure of *T. cruzi* is complex, and there is a high degree of genetic and phenotypic
134 variation. The current TcI to TcVI clades are based on biochemical and molecular markers [14],

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

135 although there is substantial diversity even within these six groups [15]. The TcI clade is widespread
136 and can be found across the American continent, and has been associated with CCC [16] and sudden
137 death [17,18], among other clinical manifestations.
138
139 We have produced a complete and reliable reference sequence of the entire *T. cruzi* TcI Sylvio X10/1
140 genome, and we have generated Illumina sequencing data for 34 *T. cruzi* TcI isolates and clones
141 from different geographic locations for comparative analyses. Thus, we have been able to decipher
142 the complete organisation of the *T. cruzi* subtelomeric surface gene repertoire of the TcI Sylvio
143 X10/1 strain, revealing large numbers of evenly spaced retrotransposons, which may play a role in
144 generating genomic structural diversity and antigenic variation. Furthermore, the comparative data
145 enabled the first exploration of whole-genome population genetics of *T. cruzi* in different
146 environments and geographic locations. We found patterns of active recombination associated with
147 generation of new surface molecule variants. Together, these results contribute to answering
148 longstanding questions on the biology of Chagas disease and parasitism in general. The availability
149 of the complete repertoire of genes encoding surface molecules allows further research on virulence
150 and pathogenesis, as well as the identification of drug targets and vaccine candidates.
151
152
153
154
155
156
157 **RESULTS**:
158 **Genome sequence of *Trypanosoma cruzi* DTU-I Sylvio X10/1**:
159 The final Sylvio X10/1 genome assembly reconstructed 98.5 % of the estimated strain genome size
160 and was contained in 47 scaffolds - which here will be referred to as pseudomolecules - assembled

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

from 210 X PacBio sequence data and a previous Illumina data set (**Table 1**). Comparison with the available assembly of the TcVI strain CL Brener revealed a conserved core of syntenic blocks composed of stretches of homologous sequences separated by large gaps of sequence that were not reconstructed in TcVI. These gaps corresponded to most of the surface molecule gene arrays and simple repeats in the Sylvio X10/1 genome (**Figure 1a**). The length of the PacBio reads and the high coverage allowed the reconstruction of long stretches of repetitive sequences (**Figure 1b**) that could not be resolved using shorter read data.

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

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

Although retrotransposons were found to be present throughout the genome, the frequency of VIPER and L1Tc elements was markedly higher in subtelomeric regions and they were found within one kilobase of pseudogenes, hypothetical proteins and surface molecule gene tandem arrays (One-sided Fisher exact test, *p-value* < 1.32 e-16). This distribution indicates that these elements may play a role in the generation of new sequence diversity in the subtelomeric gene families by providing a source of microhomology. It is compelling to speculate that they could act as

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

187 transcriptional regulators by the introduction of novel transcription start sites, as has been
188 proposed in other eukaryotes [19]; nevertheless, we do not have experimental evidence for the
189 activity of these retroelements in *T. cruzi*.
190
191 Simple and low complexity repeats were observed surrounding subtelomeric coding sequences and
192 were also more abundant in the subtelomeric regions (2.18 %), extending up to 4 Kb, compared to
193 core regions (0.98 %) where they were much shorter (10 - 120 bp). The most prevalent type of
194 simple repeat had the (C)n motif (11.7 %), (TG)n repeat motif (5.6 %) and (CA)n repeat motif (5.1 %);
195 each variable in length. This microhomology of the simple subtelomeric repeats may facilitate
196 recombination for the generation of new surface molecule variants, as described in other parasitic
197 protozoa, including *Trypanosoma brucei* and *Plasmodium falciparum* [20,21].
198
199 A total of 19,096 evidence-supported genes were identified in the TcI Sylvio X10/1 haploid genome
200 sequence, compared to the higher estimate of 22,570 for TcVI CL Brener, mostly due to the larger
201 size of the subtelomeres in the TcVI hybrid genome. The core regions of the genome were found to
202 correspond well to results generated previously using short read sequencing of the same strain [11], in
203 both gene organisation and content. Tandemly repeated genes that were collapsed in previous *T.*
204 *cruzi* genome assemblies were now resolved. About 24.1% (n = 4,602) of the total annotated genes
205 were truncated, mostly due to the introduction of premature stop codons, and 67 % of these were in
206 subtelomeric regions located within surface molecule gene arrays, sharing motifs of the complete
207 genes.
208
209 The new assembly allowed the first analysis of the complete *T. cruzi* surface molecule gene
210 repertoire. Genes of each of the three major surface molecules families were organised as multiple
211 tandem arrays. After genome annotation, the total number of such arrays were: trans-sialidases,
212 312, with 2,048 complete gene copies and 201 pseudogenes; mucins 98, with 2,466 complete copies

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and 111 pseudogenes; MASPs 264, with 1,888 complete copies and 245 pseudogenes. These three surface molecule gene families comprised 16.02 Mbp (39.04 %) of the TcI Sylvio X10/1 genome and presented a high level of sequence diversity (**Supplementary figure 2**). Sequence strand switches often delimited the surface molecule tandem arrays. Commonly, these arrays had two to four complete copies immediately followed by two or more truncated copies with motifs similar to the complete gene. The intergenic spaces between arrays were rich in simple and low complexity repeats with no identifiable regulatory elements. The VIPER and L1Tc retrotransposon elements, in clusters of two to four copies, were found in the proximity of, or inside, tandem arrays containing trans-sialidases, mucins and MASP genes. As the surface molecule genes are known to evolve rapidly and be highly variable [22], the enrichment of VIPER and L1Tc elements in these regions supports the hypothesis that they may be involved in generating new surface molecule gene variants via recombination mediated by sequence homology.

Both, Ser/Thr kinases and DEAD-box RNA helicase genes were found at both extremes of 34 (10.81 %) trans-sialidase arrays located in pseudomolecules 1, 2 and 8. Searches against the RFAM database identified 1,618 small RNAs in the TcI Sylvio X10/1 genome. These were mostly ribosomal RNAs with the 5S rDNA subunit being the most common (31.9 %) followed by ACA Box snoRNAs (30.9 %), SSU rDNA (12.2 %) and LSU rDNA (10.2 %) subunits. We also found hits to telomerase RNA component (TERC), Catabolite Repression Control sequester (CrcZ), Protozoa Signal Recognition Particle RNAs, spliceosomal RNA subunits and miRNAs. The putative miRNAs identified in Sylvio X10/1 belong to the MIR2118 and MIR1023 families, previously not found in protozoan parasites. The functional relevance of these predicted small RNAs will need to be further validated *in vitro*. The miRNA segments were located in both strands within 1 Kb of genes coding for DEAD-box RNA helicases surrounding surface molecule gene tandem arrays.

Genomic variation within the *Trypanosoma cruzi* TcI clade:

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

239 Intra-TcI genomic diversity was examined among 34 samples from six countries: United States,
240 Mexico, Panama, Colombia, Venezuela and Ecuador, derived from a range of triatomine vectors and
241 human patients of different clinical stages (**Table 2** and **Supplementary table 1**). Our hybrid
242 variant calling strategy allowed us to identify genomic variants in the core and subtelomeric
243 regions in a reliable fashion (See methods).
244
245 A total of 1,031,785 SNPs and 279,772 INDELs shorter than 50 bp were called - relative to the Sylvio
246 X10/cl1 genome - for all the sequenced isolates. INDELs presented an average density of 5.3 variants
247 per Kb and SNPs 24.1 variants per Kb. An individual *T. cruzi* TcI isolate was found to contain an
248 average of 61,000 SNPs and 6,820 INDELs with a density of 31.8 variants per Kb. However, these
249 measures fluctuated depending on the geographical and biological source of the sample. Core
250 regions had an average SNP density of 0.4 variants per Kb, in contrast with subtelomeric regions
251 where values of 10 variants per Kb were found. It was not surprising that the bulk of the genomic
252 variants were located in the subtelomeric regions in all the isolates, with fewer differences in the
253 core regions. Although several studies using single gene markers have identified heterogeneity in
254 the TcI clade [15,23], the extent of this variation has not been assessed genome-wide.
255
256 The majority of INDELs (96 %) were found in intergenic or noncoding regions, and 81 % of those
257 were located in subtelomeric regions. INDELs within coding sequences were exclusively found to
258 cause frameshifts turning the affected coding sequence into a pseudogene. This distribution of
259 INDELs is a genomic signature that has been associated with non-allelic homologous recombination
260 due to unequal crossing over [24] or microhomology-mediated end joining [25,26] (**Table 3**). Short
261 insertions were more prevalent than short deletions, a pattern common to all the analysed TcI
262 genomes when compared to Sylvio X10/1. In the subtelomeric regions, short insertions (1 - 3 bp)
263 occurred within the upstream and downstream portions of the coding sequences and usually
264 involved the addition of one or more cytosines or guanines. Deletions of 1 bp indicating the removal

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of an adenine or thymine were also observed within these regions, but at a lower frequency. Longer deletions (5 - 20 bp) and insertions (8 - 10 bp) were observed within trans-sialidases, Retrotransposon Hot Spot (RHS), pseudogenes and, at a lower frequency, L1Tc retroelements.

Population genomics of the *Trypanosoma cruzi* TcI clade:

We used the short genomic variants to analyse the population genomics of the *T. cruzi* TcI clade, and where possible taking into account the different sample sources (insect vector or human host), clinical outcome of the infected patients and geographic locations (**Supplementary table 1**). This sampling strategy allowed comparison of parasite population structure in different environments. Interestingly, a Bayesian PCA analysis using INDELs and an IBD-based hierarchical clustering using only SNPs for all the samples showed a mostly geography-specific population structure (**Figure 2a** and **supplementary figure 3**).

The analysis of the variation between two Colombian TcI isolates made it possible to compare parasites from a HIV-positive patient with fatal cardiomyopathy (CG) and from an acute chagasic patient infected by oral transmission (FcHc). For each strain, replicate clones from the original sample were isolated and cultured under the same conditions, and five of the replicates from each sample were sequenced in a single Illumina HiSeq 2500 run and 158,565 well supported SNPs were called. Using this set of SNPs we calculated global and per-site population genetic statistics. These samples displayed distinctive behaviour in a global analysis of genomic diversity by separating into two well-defined clusters, as can be seen in **Figure 2b**. Linkage Disequilibrium (LD) analyses were performed genome-wide for both groups using the r2 statistic; revealing a fluctuating pattern of LD across the entire genome with large blocks of low r2 values - implying a recombinatorial process - present at distinctive chromosomal locations that were specific to each group of clones. Particularly, CG clones had less genetic diversity than FcHc clones (**Figure 3a**) and displayed a trend towards LD, whereas FcHc clones presented more dynamic LD pattern. Values of r2 near zero were

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

more common in LD sliding windows containing genes coding for surface molecules and r2 values closer to one were present exclusively in core regions rich in housekeeping genes, indicating that these regions are more stable. For the CG and FcHc clones we calculated a global Fixation index (*Fst*) value of -0.9377958 and -0.1162212 respectively (**Figure 3b**). These values are consistent with genetic differentiation in recombination hotspots in the subtelomeric regions. The global Tajima's D value for the CG clones was 1.373 and 0.9906 for FcHc clones, suggesting the presence of multiple alleles at variable frequencies in both populations (**Figure 3c**). This pattern was more evident in the subtelomeric regions, which is consistent with balancing selection of surface molecules.

Analyses of genomic variation between samples isolated from humans and vectors from Mexico, Panama and Ecuador revealed that the global genetic differentiation among samples isolated from vectors was *Fst* = 0.1289547 whereas for samples isolated from humans the observed was *Fst* = -0.05521983. The patterns of linkage disequilibrium between human and vector derived isolates were similar to those observed in the Colombian clones. Estimates of the Tajima's D statistic revealed a distinctive pattern of selection between the two groups. Balancing selection was detected specifically in regions containing tandem gene arrays coding for surface molecules in all the samples derived from vectors, regardless of their geographical origin; whereas selective sweeps were present in the same regions in human-derived samples. Large genomic areas (> 50 Kb) containing surface molecule genes displayed negative Tajima's D values in human-derived isolates, in contrast with the pattern observed in vector-derived isolates with long genomic stretches (> 70 Kb) of positive Tajima's D values and short genomic blocks (< 5 Kb) with negative values.

Genome structural variation:

Genomic structural variants, such as deletions, tandem and interspersed duplications, genomic inversions and chromosomal break-ends, were observed ubiquitously throughout the genomes of the analysed TcI strains. The most common type of intrachromosomal structural variant observed

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

317 was tandem duplications followed by deletions larger than 50 Kb (**Table 2**). Chromosomal break-
318 ends, similar to the unbalanced chromosomal translocations observed in higher eukaryotes, were
319 the most abundant type of structural rearrangement and they were only present in genomic regions
320 that were statistically enriched with retroelements and simple repeats. These areas presented a
321 conserved pattern: they contained surface molecule gene tandem arrays and their breakpoints were
322 composed of simple repeats and retrotransposons of the VIPER and L1Tc class.
323
324 These events were between 20 - 150 Kb in length and contained fragments or even complete coding
325 sequences for surface molecule genes, such as trans-sialidases, mucins and MASP genes and
326 surface glycoproteins (gp63/gp85). Housekeeping genes seemed to have not been affected by these
327 genomic rearrangements. The breakpoints were composed of simple repeats, retrotransposons or
328 both. Rearrangements affecting gene tandem arrays generated longer coding sequences by
329 superimposing fragments - or the entire coding sequence - on genes of the same family located in a
330 different genomic location. For instance, the Colombian isolates generated longer trans-sialidase
331 genes by moving coding sequences from pseudomolecule 1 to pseudomolecule 8, while Texas
332 isolates recombined trans-sialidases between pseudomolecule 16 and pseudomolecule 21. In this
333 way, surface molecule genes were merged with another member of the same gene family - or a
334 pseudogene - resulting in a new mosaic gene sequence (**Figure 4a**).
335
336 Retroelements could be found within or near genomic regions containing surface molecule gene
337 tandem arrays (**Supplementary tables 2, 3** and **4**) and L1Tc fragments or their entire sequence
338 were also included in the rearranged region in all the observed translocation spots, where they were
339 inserted into regions containing simple repeats composed by AT dimers (**Figure 4b**).
340
341 Multiple examples of the generation of new surface molecule gene variants were identified in TcI
342 from diverse sources. It therefore appears that the parasite uses specific molecular mechanisms of

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

343 recombination that can rapidly generate surface molecule diversity, allowing it to increase the
344 genomic plasticity required to adapt to changing environments and evade immune responses
345 during short and long-term infections in various host species.
346
347 The sizes of the tandem duplications ranged from 6 - 75 Kb and mainly involved tandem arrays
348 coding for surface molecules, mostly trans-sialidases and mucins, but also Disperse Gene Family 1
349 (DGF-1) and several hypothetical proteins. The breakpoints of these duplications were surrounded
350 by simple repeats and retroelements in subtelomeric regions. A tandem duplication event could
351 involve between four and 25 copies of a specific gene when in the subtelomeric regions, whereas in
352 core regions the number was between two and eight. We observed that large deletions occurring in
353 subtelomeric regions were surrounded by simple repeats of the type (T)n and (AT)n and
354 retrotransposons of the L1Tc class, containing surface molecule gene tandem arrays. Deletions in
355 these genomic regions tended to be shorter (4 - 12 Kb) and sample specific.
356
357 **CNV distribution in the TcI clade**:
358 CNV varied extensively between strains. Most notably, among the Colombian strains, isolates
359 derived from the same sample presented different gene copy numbers. There have been previous
360 attempts to assess CNV in the *T. cruzi* genome [27], but these studies were performed using DNA tiling
361 microarrays with probes designed using the TcVI CL Brener strain assembly, in which subtelomeric
362 regions are essentially absent.
363
364 The distribution of CNV in the genomes of the studied TcI samples was isolate-specific, and
365 involved segments of an average size of 5 Kb. In the samples analysed in our study we observed
366 blocks of segmental CNV within a chromosome with a pattern that was unique to each sample.
367 Notably, the Colombian clones presented individual profiles of CNV (**Figure 5a** and **5b**) despite
368 being derived from the same clinical isolates.

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Sequence blocks affected by segmental CNVs contained retrotransposons of the VIPER and L1Tc class, as well as surface molecule genes surrounded by simple repeats. The isolate-specific nature of these CNV events demonstrates the high level of within-clade diversity of the TcI samples. The distribution of CNV across the *T. cruzi* genome reinforces the dynamic nature of the subtelomeric regions and the surface molecule gene families. As discussed below, it is important to note the association of structural and copy number variation with the presence of retrotransposons and simple repeats and their putative involvement in the generation of novel sequence variants via recombination.

DISCUSSION:

Complete reconstruction of the *T. cruzi* genome to encompass the subtelomeric regions, has proved to be difficult to achieve using short reads, due to sequencing library preparation biases and a genome architecture that is rich in long stretches of simple repeats, large repetitive gene families and multiple retrotransposons. Here we have used long PacBio sequencing reads to provide the most complete genome sequence of a *T. cruzi* strain to date. This has allowed us to perform the first detailed analyses of the repertoire of complex genes families that encode cell surface molecules, considered to be involved in cell invasion and evasion of the host immune response. We have shown the duality in the organisation of the parasite genome, comprised of a core genomic component with few repetitive elements and a slow evolutionary rate, resembling that of other protozoa, and a contrasting, highly plastic subtelomeric region encoding fast evolving surface antigens, with abundant interspersed retrotransposons. The structural changes that generate and maintain diversity in *T. cruzi* surface molecules have certain mechanistic parallels in other protozoa such as those recently described in *Plasmodium falciparum* [28].

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Early studies of the genetic diversity of *T. cruzi* using geographically disparate sampling and restricted comparisons of genetic diversity suggested a clonal population structure [29,30]; however, population genetics with an expanded set of markers have now challenged this view [15,31,32]. Nevertheless, there are still conflicting views as to which model best describes the population structure of *T. cruzi* [33,34]. The newer Sylvio X10/cl1 genome sequence will now enable extensive genome-wide comparative population genomics analyses, which may shed light on this issue. Comparative analyses of 34 *T. cruzi* isolates and clones from the TcI clade suggest many recombination events and population indices, normally associated with genetic exchange between strains, are more likely to be caused by the extensive repeat-driven recombination in the subtelomeric regions. The extent of variation in the subtelomeric regions rich in surface antigen genes and the geographical clustering of strains within a region, indicates active, on-going adaptation to host and vectors. This need for phenotypic - and thus genomic - versatility may impel the active generation of sequence diversity in *T. cruzi*. Further analyses of the evolution of subtelomeric regions will yield much more detailed understanding of diversity within and between the six currently recognised genetic lineages of *T. cruzi* [22].

We have shown how the genome architecture and dynamic subtelomeric regions of *T. cruzi* may provide a mechanism to generate rapidly the sequence diversity required to escape the host immune response and adapt in response to new environments. It is the striking richness in simple repeats and retrotransposons in the subtelomeric regions that renders these genomic areas susceptible to structural change, similar to yeast and other pathogens [35,36,44]. Retrotransposons have been associated with the generation of complexity in genomic regions in mammals and plants and with control of gene expression [36,37]. In the case of *T. cruzi*, they appear to generate novel variants via mechanisms that exploit sequence homology. The presence of the simple repeats and retrotransposons near surface molecule genes provides the microhomology for both mechanisms to operate in such regions. Our analysis of INDELs and chromosomal breakpoints in the

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

421 subtelomeric regions confirmed that a mechanism similar to NAHR or MMEJ operates as source of
422 sequence diversity, for example by transposition of trans-sialidase genes or pseudogenes to
423 produce new sequence mosaics. The required recombination machinery is conserved in *T. cruzi* [38].
424 Furthermore, these mechanisms would explain the high level of pseudogenisation observed in *T.*
425 *cruzi*.
426
427 Retrotransposons were first reported from *T. cruzi* in 1991 [39]. The presence of these elements may
428 also partly account for the previously reported widespread observation of copy number variation in
429 different *T. cruzi* strains [27]. Thus, we find that repeats near the surface molecule genes appear to
430 drive recombination in *T. cruzi*. The apparent inability of *T. cruzi* to condense chromatin may
431 facilitate transposition in a stochastic fashion, facilitating generation of sequence diversity in
432 exposed regions of the genome. A similar process has been described in the neurons of higher
433 eukaryotes [40] but not in any other unicellular organism. Retrotransposons may also have an
434 important role as gene transcription regulators: they may either silence or promote gene
435 expression, due to their susceptibility to DNA methylation or by providing potential binding sites
436 respectively, as it has been observed in previous works [41]. This lack of a well-defined transcriptional
437 regulation machinery in the *T. cruzi* genome may suggests a link to the requirement for
438 retrotransposon closely associated with gene tandem arrays.
439
440 **CONCLUSION:**
441 Here we have sequenced and assembled the complete genome of a *Trypanosoma cruzi* TcI strain.
442 This has enabled the first resolution of the complex multiple gene families that encode *T. cruzi*
443 surface molecules, and provided a basis for *T. cruzi* population genomics. We discover an
444 extraordinary concentration of retrotransposons among the subtelomeric surface gene families and
445 indications of repeat-driven recombination and generation of antigenic diversity, providing the
446 mechanisms for *T. cruzi* to evade the host immune response, and to facilitate the adaption to new

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

METHODS:

Genome sequencing and assembly of *Trypanosoma cruzi* DTU-I Sylvio X10/1: Total genomic DNA from Sylvio X10/1 strain was used to produce PacBio CCS data according to the protocols from the Genomic Facility of Science for Life Laboratory, (Sweden) and Pacific Biosciences (USA). Genomic DNA was sequenced to a depth of 210X using the PacBio platform, supplying raw reads with an average length of 5.8 Kb. These reads were corrected by means of the PBcR v8.3 pipeline with the MHAP algorithm [42] using the auto-correction parameters described to merge haplotypes and skipping the assembly step, producing a total of 1,216 contigs (NG50 = 62 Kb). Later, the assembly was scaffolded using the corrected PacBio reads with the SSPACE-Long scaffolder yielding 310 scaffolds (NG50 = 788 Kb); 118 gaps were filled using Illumina reads with GapFiller and corrected PacBio reads with PBJelly2. Finally, the core regions of these scaffolds were aligned against the core regions of the TcVI CL Brener reference genome using ABACAS (http://abacas.sourceforge.net), producing 47 pseudomolecules. The quality of the new assembly was assessed with FRC_bam with the Illumina paired end reads.

Annotation of the *Trypanosoma cruzi* DTU-I Sylvio X10/1 Genome: The genome sequence was annotated using a new kinetoplastid genome annotation pipeline combining homology-based gene

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

model transfer with *de novo* gene prediction. To allow for the sensitive identification of partial genes, input sequences were split at stretches of undefined bases, effectively creating a set of 'pseudocontigs', each of which does not contain any gaps. Gene finding was then performed on both the original sequences and the pseudocontigs using AUGUSTUS, which also calls partial genes at the boundaries of each pseudocontig. AUGUSTUS models were trained on 800 genes randomly sampled from the 41 Esmeraldo-type *T. cruzi* CL Brener chromosomes in GeneDB. Protein-DNA alignments of reference proteins against the new *T. cruzi* sequences, generated using Exonerate, were used as additional hints to improve the accuracy of the gene prediction. In addition, the RATT software was used to transfer highly conserved gene models from the *T. cruzi* CL Brener annotation to the target. A non-redundant set of gene models was obtained by merging the results of both RATT and AUGUSTUS and, for each maximal overlapping set of gene models, selecting the non-overlapping subset that maximizes the total length of the interval covered by the models, weighted by varying levels of *a priori* assigned confidence. Spurious low-confidence protein coding genes with a reading direction in disagreement with the directions of the polycistronic transcriptional units were removed automatically. The result of this integration process was then merged with ncRNA annotations produced by specific tools such as ARAGORN and Infernal. Finally, protein-DNA alignments with frame shifts produced by LAST were used in a computational approach to identify potential pseudogenes in the remaining sequence.

Downstream of the structural annotation phase, gene models were automatically assigned IDs and further extended with product descriptions and GO terms, both transferred from CL Brener orthologs and inferred from Pfam protein domain hits and represented as feature attributes or Sequence Ontology-typed subfeatures tagged with appropriate evidence codes. This annotation pipeline has been implemented in the Companion web server. The assembled genome was scanned for small RNAs using INFERNAL against the curated RFAM database using cmsearch with a minimum e-value of 1e-10, a GC-bias of 0 and a minimum alignment length of 10 nt. This annotation

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

process has been implemented into the web-based annotation pipeline COMPANION [43] from the Wellcome Trust Sanger Institute.

Repetitive sequences were annotated using RepeatMasker with the NCBI+ search engine and LTRHarvest. Using the genomic coordinates of the repetitive elements, the genome was split in windows of 10 Kb to identify VIPER and L1Tc retroelements adjacent to surface molecule genes (i.e: trans-sialidases, mucins and MASP). A one-sided Fisher's exact test was used to evaluate if the retroelements were enriched in genomic segments containing surface molecule genes.

Identification of SNPs and Indels: An improved short-read mapping strategy was used to assign the reads to their target sequences with high accuracy, especially in regions rich in simple and low complexity repeats, by taking advantage of the statistical read placement implemented in the Stampy read mapper to accurately call genomic variants from the mapped reads. Reads from all *T. cruzi* TcI isolates were mapped against the assembled *T. cruzi* Sylvio X10/1 genome using a two-step mapping process to improve the mapping of Illumina data to highly repetitive regions: First, reads were mapped using BWA MEM with default parameters; later, the BAM file produced by BWA was remapped with Stampy (v1.23) using the *--bamkeepgoodreads* option. The final mapping file was sorted and filtered for PCR duplicates using Picard Tools v1.137. Variants were called using FreeBayes with a minimum per-base quality of 30, minimum mapping quality of 30 and minimum coverage of 15 bases. Variants that were found in a potentially misassembled region were excluded from the analysis. Additionally, genomic variants were called using FermiKit - which is an assembly-based variant caller - to validate the genomic variants observed in subtelomeric regions. A consensus of the two methods was used as a final set of variants for downstream analyses. Haplotypes were phased using Beagle r1399. The phased markers were used for downstream analyses with SNPrelate and VCFtools and the functional effect of the identified variants was predicted using SnpEff.

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Identification of genomic structural variants: Genomic structural variants (SV) were identified using a consensus of different methods: Delly2, Lumpy, FermiKit and FindTranslocations (https://github.com/vezzi/FindTranslocations.git) using both raw reads and realigned BAM files. For each method, a SV muts had a depth of coverage > 10 reads and a mapping quality of > 30. Later, a consensus was created with all the SV that were supported by all the methods. SVs that were supported by FermiKit and at least one of the mapping-based methods were also included but labelled as 'Low Confidence'. SVs identified by only one method were not included. Breakpoint analysis was done with custom Python scripts and their functional effect was predicted using SNPeff. Analysis of copy number variation (CNV) were done using the BAM files for each sample with the *Control-FREEC* package.

Accessions: Whole genome sequence Genbank accession: ADWP00000000 (CP015651-CP015697). Reads for the 34 TcI isolates SRA accession: SRP076682.

Authors contributions: CT-L and BA conceived and designed the study. CT-L designed and executed computational analyses. MY prepared Sylvio X10/cl1 genomic DNA for PacBio sequencing and performed manual annotation of surface molecule genes. JEC, AS, JDR, FG, SO-M, JAC, SMRT, HC, RG, KJ, MEB, PJH, KOM, MJG, BB provided genomic DNA for TcI isolates. JLR-C and DCB created chromosome maps for surface molecules. MAM, LAM, ML and ECG contributed to the interpretation of the results. CT-L, BA, MAM wrote the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements: The authors wish to thank to John Kuijpers and Lawrence Hon from Pacific Biosciences (PacBio) for providing the data for the TcI Sylvio X10/cl1 genome and preliminary assembly results. We would also like to acknowledge Eric Dumonteil, University

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

TABLES:

Table 1: ***Trypanosoma cruzi*** **DTU-I Sylvio X10/1 genome assembly**

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

Table 2: Genomic variants identified in the sequenced 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

* - FcHc and CG clones from Colombia

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

564 ** - TcI Colombiana strain

565

566

567

568

569

570 **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*GACTTCAC*AGC*CTG	NHEJ	28 389
Long Insertion	C*GGCTAGACCAGGTACAGTC*A	MMEJ	32 666
Long Deletion	GC*acactgacacgacactgacacactgaa*A	MMEJ	31 712

571

572 HR = Homopolymer run

573 TR = Tandem Repeat

574 = Deletion

575 = Insertion

576 * For all the 34 TcI genomes compared against Sylvio X10/cl1

577

578

579

580

581

582

583

584

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

585

586

587

588

589

590

591 **Supplementary table 1:** Samples used in this study

Sample	Country	Source	Chagas'	Seq. 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*	NA	28 X
V2	Panama	*R. pallescens*	NA	25 X
V3	Panama	*T. dimidiata*	NA	28 X
TBM3324	Ecuador	*R. ecuadoriensis*	NA	28 X
TBM3406B1	Ecuador	*R. ecuadoriensis*	NA	25 X
TBM3479B1	Ecuador	*R. ecuadoriensis*	NA	29 X
TBM3519W1	Ecuador	*R. ecuadoriensis*	NA	28 X
X10462-P1C9	Venezuela	Human	NA	31 X
X12422-P1C3	Venezuela	Human	NA	29 X

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Colombiana	Colombia	Human	NA	30 X
CG110	Colombia	Human	Acute Co-infection	30 X
CG111	Colombia	Human	Acute Co-infection	30 X
CG112	Colombia	Human	Acute Co-infection	30 X
CG113	Colombia	Human	Acute Co-infection	30 X
CG114	Colombia	Human	Acute Co-infection	30 X
CG115	Colombia	Human	Acute Co-infection	30 X
FcHc11	Colombia	Human	Acute Oral	30 X
FcHc12	Colombia	Human	Acute Oral	30 X
FcHc13	Colombia	Human	Acute Oral	30 X
FcHc14	Colombia	Human	Acute Oral	30 X
FcHc15	Colombia	Human	Acute Oral	30 X
H1b	Mexico	Human	Acute	32 X
TD23	Texas	*T. dimidiata*	NA	30 X
TD25	Texas	*T. dimidiata*	NA	30 X

NA = Not applicable

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

606

607

608

609

610

611

612

613 **FIGURE LEGENDS**:

614 **Figure 1**: **a**) Comparison of chromosome 15 from the TcVI CL Brener and TcI Sylvio X10/cl1
615 assemblies. Grey lines between pseudomolecules represent regions of synteny between
616 orthologous genes, and yellow blocks represent gaps in both genome assemblies. Coloured blocks
617 represent paralogs of multicopy gene families. **b**) Distribution of surface molecule gene tandem
618 arrays in the 47 chromosomes of the TcI Sylvio X10/cl1 genome.

619

620 **Figure 2**: **a**) Bayesian principal component analysis (PCA) of *T. cruzi* DTU-I strains using INDELs and
621 **b**) Identity by Descent (IBD) dendrogram of *T. cruzi* DTU-I strains using SNPs. Both analysis, using
622 different markers, support the population structure of the analysed TcI samples. Notably, the highly
623 virulent TcI Colombiana and the Panamanian TcI H1 from a chronic patient are presented as outliers
624 (**b,** far left).

625

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

632

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

633 **Figure 4**: **a**) Proposed mechanism of inter-chromosomal recombination between gene tandem
634 arrays for the generation of antigenic diversity. Here, VIPER retrotransposons (green) are
635 surrounding a tandem array, shown in the direction of transcripton, containing trans-sialidases
636 (blue) and pseudogenes (red) in pseudomolecules 1 and 8. The rearranged tandem array is shown
637 below displaying the merger of genes and pseudogenes to form a longer coding sequence. **b**) Detail
638 of simple repeats and retrotransposons within a trans-sialidase tandem array in pseudomolecule 1.
639
640 **Figure 5**: Distribution of CNV changes in chromosome 3 of the Colombian **a**) CG clones and **b**) FcHc
641 clones. Black lines represent the reference genome sequence and cyan lines represent the sample
642 under study. Each sliding window for CNV evaluation is represented as a dot. A drastic change in
643 CNV can be noted in the FcHc clones whereas the CG clones seem to be less affected.
644
645 **Supplementary figure 1**: Distribution of coverage for **a**) PacBio and **b**) Illumina reads in a
646 subtelomeric end on chromosome 8 showing a well supported genome assembly of these
647 regions.
648
649 **Supplementary figure 2**: Genome-wide Neighbour Joining tree for trans-sialidases genes
650 from the Sylvio X10/cl1 genome showing the high level of sequence diversity for these gene
651 family.
652
653 **Supplementary figure 3**: Principal component variability for the genotype diversity of the 34
654 TcI isolates.
655
656
657
658
659

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References:

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

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

12. Franzén, O. *et al.* Comparative genomic analysis of human infective Trypanosoma cruzi lineages with the bat-restricted subspecies T. cruzi marinkellei. *BMC Genomics* **13,** 531 (2012).

13. Kim, D. *et al.* Telomere and subtelomere of Trypanosoma cruzi chromosomes are enriched in (pseudo)genes of retrotransposon hot spot and trans-sialidase-like gene families: the origins of T. cruzi telomeres. *Gene* **346,** 153–161 (2005).

14. Zingales, B. *et al.* The revised Trypanosoma cruzi subspecific nomenclature: rationale, epidemiological relevance and research applications. *Infect. Genet. Evol.* **12,** 240–253 (2012).

15. Llewellyn, M. S. *et al.* Genome-scale multilocus microsatellite typing of Trypanosoma cruzi discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection. *PLoS Pathog.* **5,** e1000410 (2009).

16. Ramírez, J. D. *et al.* Chagas cardiomyopathy manifestations and Trypanosoma cruzi genotypes circulating in chronic Chagasic patients. *PLoS Negl. Trop. Dis.* **4,** e899 (2010).

17. Montgomery, S. P., Starr, M. C., Cantey, P. T., Edwards, M. S. & Meymandi, S. K. Neglected parasitic infections in the United States: Chagas disease. *Am. J. Trop. Med. Hyg.* **90,** 814–818 (2014).

18. Bern, C., Kjos, S., Yabsley, M. J. & Montgomery, S. P. Trypanosoma cruzi and Chagas' Disease in the United States. *Clin. Microbiol. Rev.* **24,** 655–681 (2011).

19. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41,** 563–571 (2009).

20. Claessens, A. *et al.* Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis. *PLoS Genet.* **10,** e1004812 (2014).

21. Hall, J. P. J., Wang, H. & Barry, J. D. Mosaic VSGs and the scale of Trypanosoma brucei antigenic variation. *PLoS Pathog.* **9,** e1003502 (2013).

22. Andersson, B. The Trypanosoma cruzi genome; conserved core genes and extremely variable surface molecule families. *Res. Microbiol.* **162,** 619–625 (2011).

23. Guhl, F. & Ramírez, J. D. Trypanosoma cruzi I diversity: towards the need of genetic

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

712 subdivision? *Acta Trop.* **119,** 1–4 (2011).

713 24. Parks, M. M., Lawrence, C. E. & Raphael, B. J. Detecting non-allelic homologous recombination
714 from high-throughput sequencing data. *Genome Biol.* **16,** 72 (2015).

715 25. Sfeir, A. & Symington, L. S. Microhomology-Mediated End Joining: A Back-up Survival
716 Mechanism or Dedicated Pathway? *Trends Biochem. Sci.* **40,** 701–714 (2015).

717 26. Weckselblatt, B., Hermetz, K. E. & Rudd, M. K. Unbalanced translocations arise from diverse
718 mutational mechanisms including chromothripsis. *Genome Res.* **25,** 937–947 (2015).

719 27. Minning, T. a., Weatherly, D. B., Flibotte, S. & Tarleton, R. L. Widespread, focal copy number
720 variations (CNV) and whole chromosome aneuploidies in Trypanosoma cruzi strains revealed
721 by array comparative genomic hybridization. *BMC Genomics* **12,** 139 (2011).

722 28. Miles, A. *et al.* Indels, structural variation, and recombination drive genomic diversity in
723 Plasmodium falciparum. *Genome Res.* **26,** 1288–1299 (2016).

724 29. Tibayrenc, M., Kjellberg, F. & Ayala, F. J. A clonal theory of parasitic protozoa: the population
725 structures of Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium, Trichomonas, and
726 Trypanosoma and their medical and taxonomical consequences. *Proc. Natl. Acad. Sci. U. S. A.*
727 **87,** 2414–2418 (1990).

728 30. Tibayrenc, M. & Ayala, F. J. Towards a population genetics of microorganisms: The clonal
729 theory of parasitic protozoa. *Parasitol. Today* **7,** 228–232 (1991).

730 31. Gaunt, M. W., Yeo, M., Frame, I. A. & Stothard, J. R. Mechanism of genetic exchange in
731 American trypanosomes. **421,** (2003).

732 32. Westenberger, S. J., Barnabé, C., Campbell, D. a. & Sturm, N. R. Two hybridization events
733 define the population structure of Trypanosoma cruzi. *Genetics* **171,** 527–543 (2005).

734 33. Tibayrenc, M. & Ayala, F. J. The population genetics of Trypanosoma cruzi revisited in the light
735 of the predominant clonal evolution model. *Acta Trop.* **151,** 156–165 (2015).

736 34. Messenger, L. A. & Miles, M. A. Evidence and importance of genetic exchange among field
737 populations of Trypanosoma cruzi. *Acta Trop.* **151,** 150–155 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

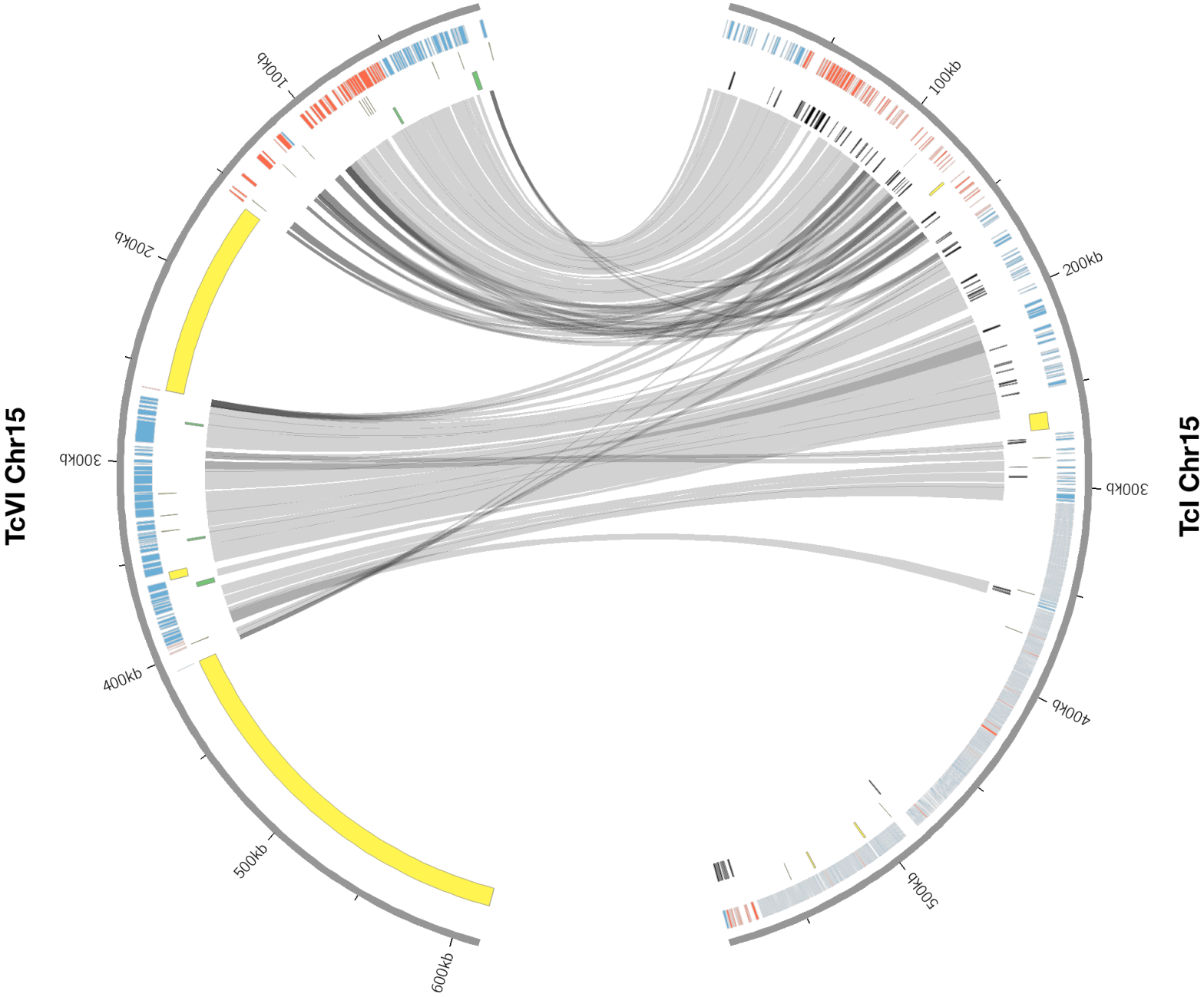
758 44. Faino, L. *et al.* Transposons actively and passively contributes to the evolution of the two-speed
759 genome of a fungal pathogen. *Genome Res.* **26**: 1091 – 1100 (2016).

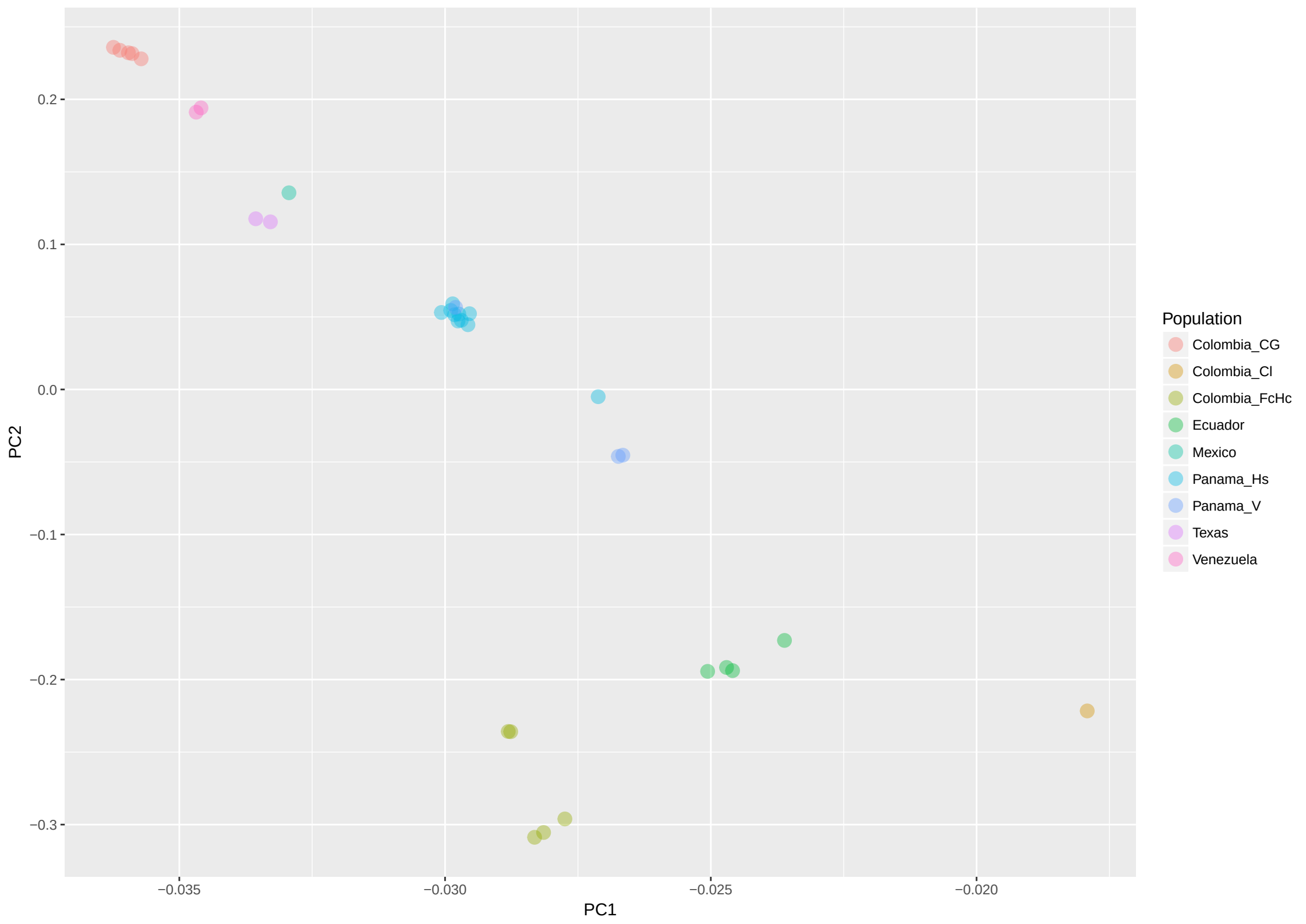
760
761
762
763
764

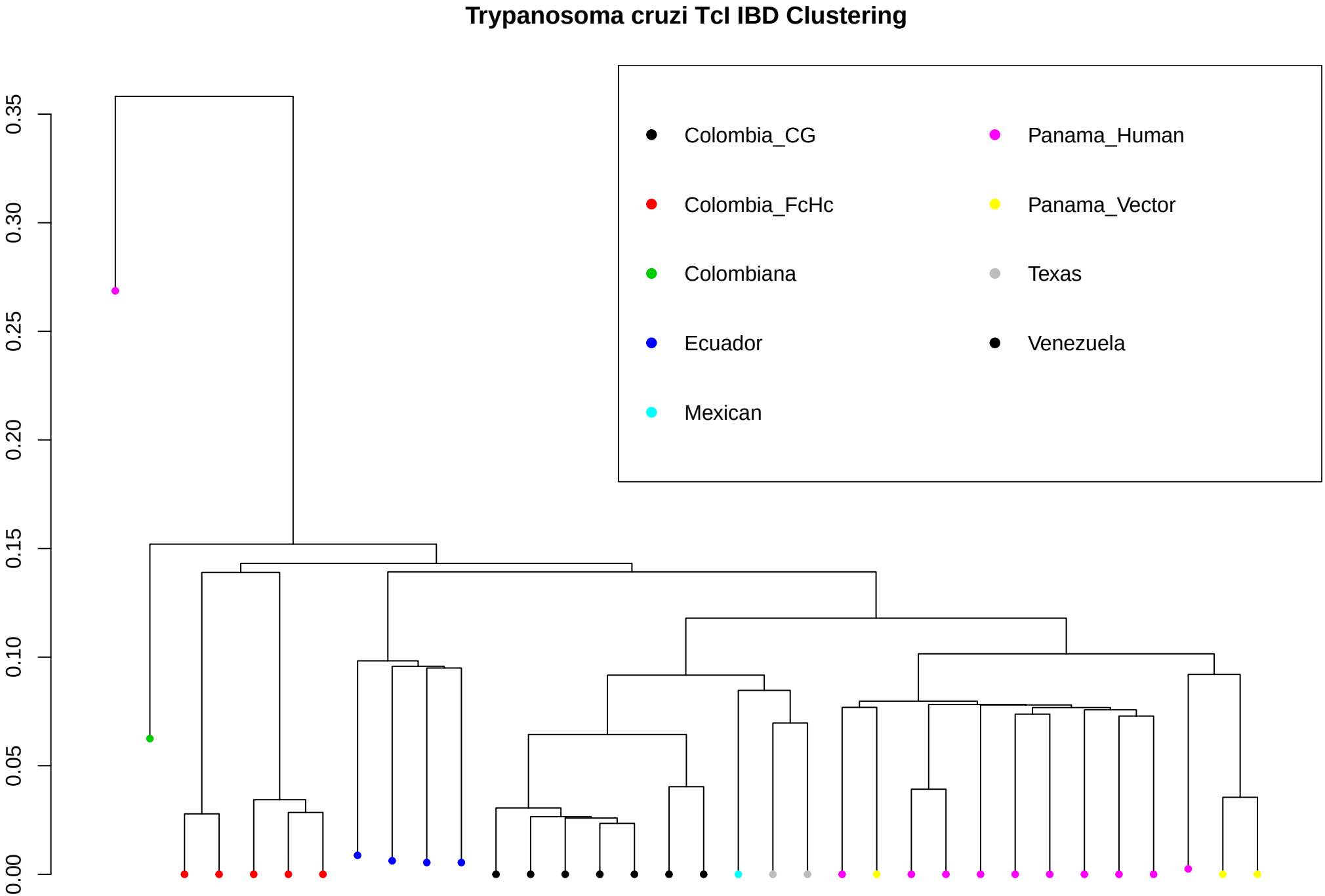
bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

765
766

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

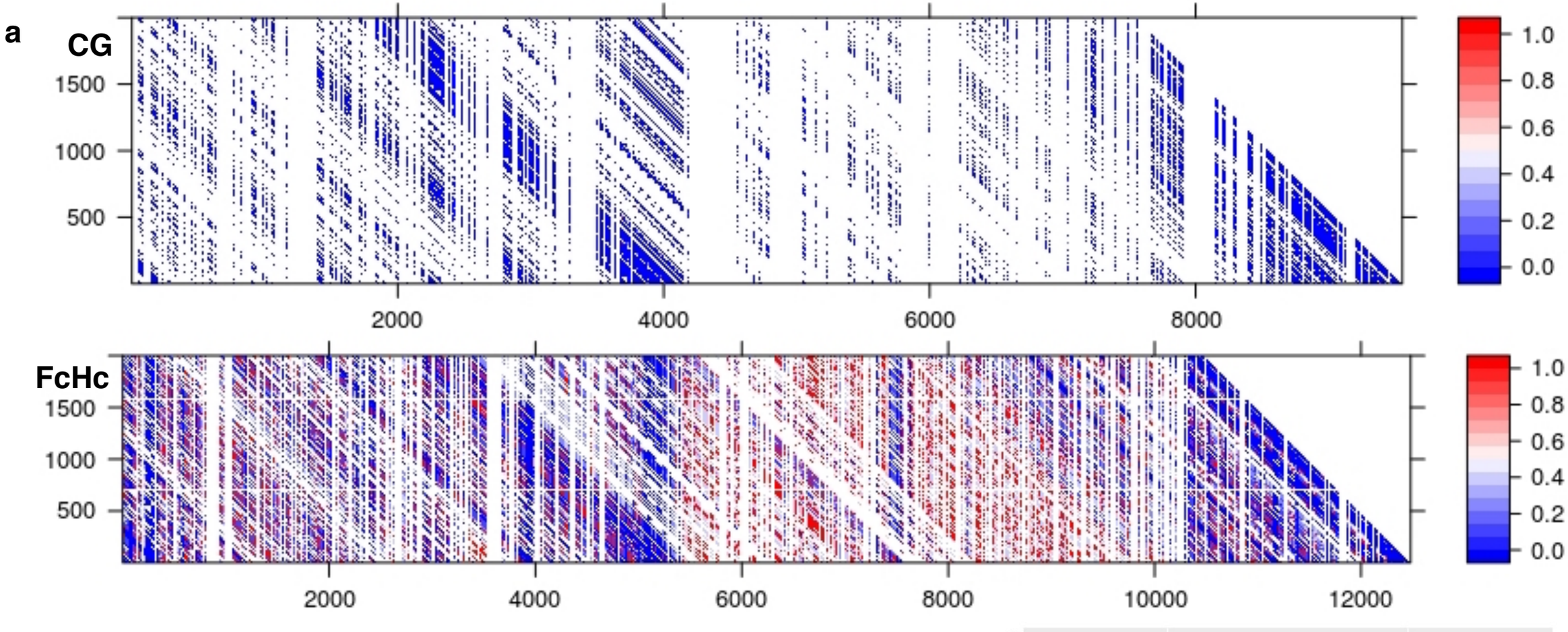



0.2
0.1
0.0
−0.1
−0.2
−0.3
PC2
−0.035
−0.030
−0.025
−0.020
PC1
Population
Colombia_CG
Colombia_CI
Colombia_FcHc
Ecuador
Mexico
Panama_Hs
Panama_V
Texas
Venezuela

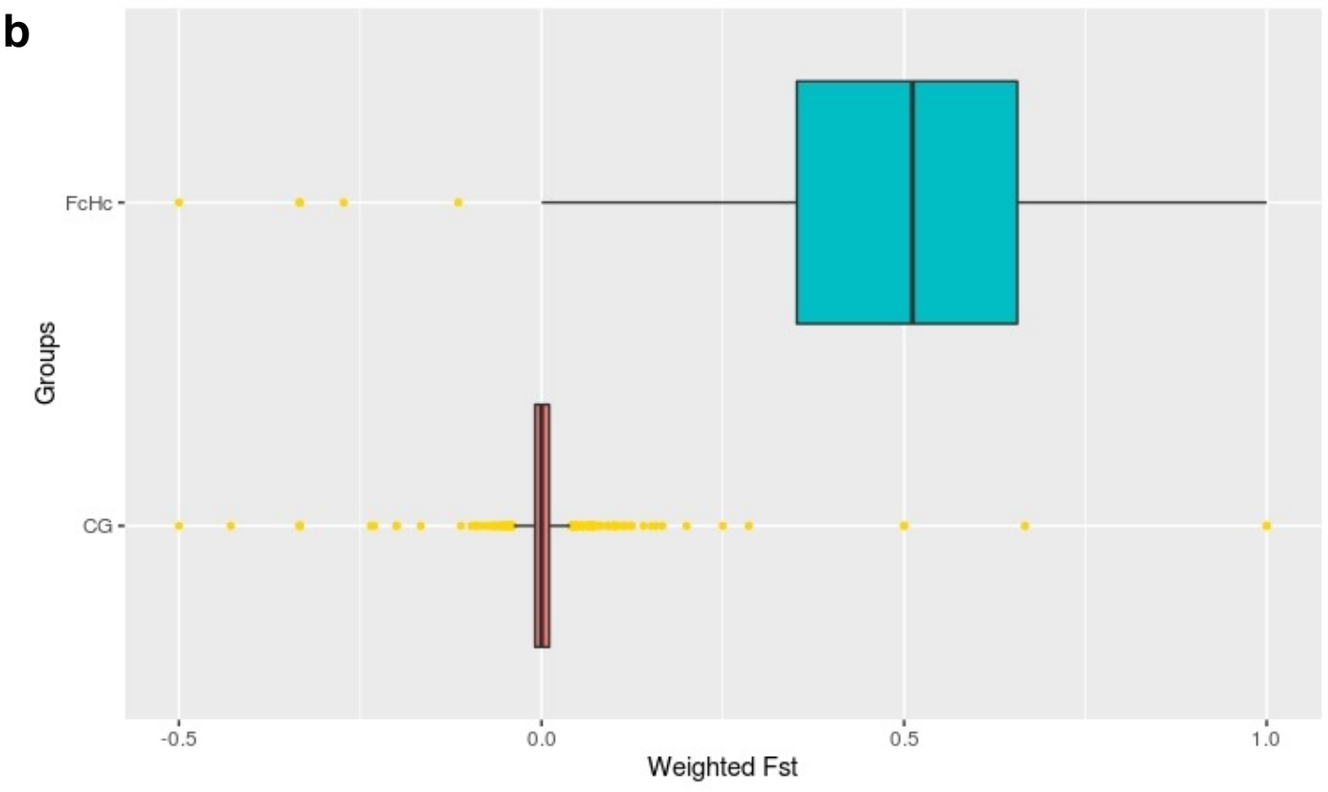

Trypanosoma cruzi TcI IBD Clustering
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Colombia_CG
Colombia_FcHc
Colombiana
Ecuador
Mexican
Panama_Human
Panama_Vector
Texas
Venezuela

aCC-BY-ND 4.0 International license.

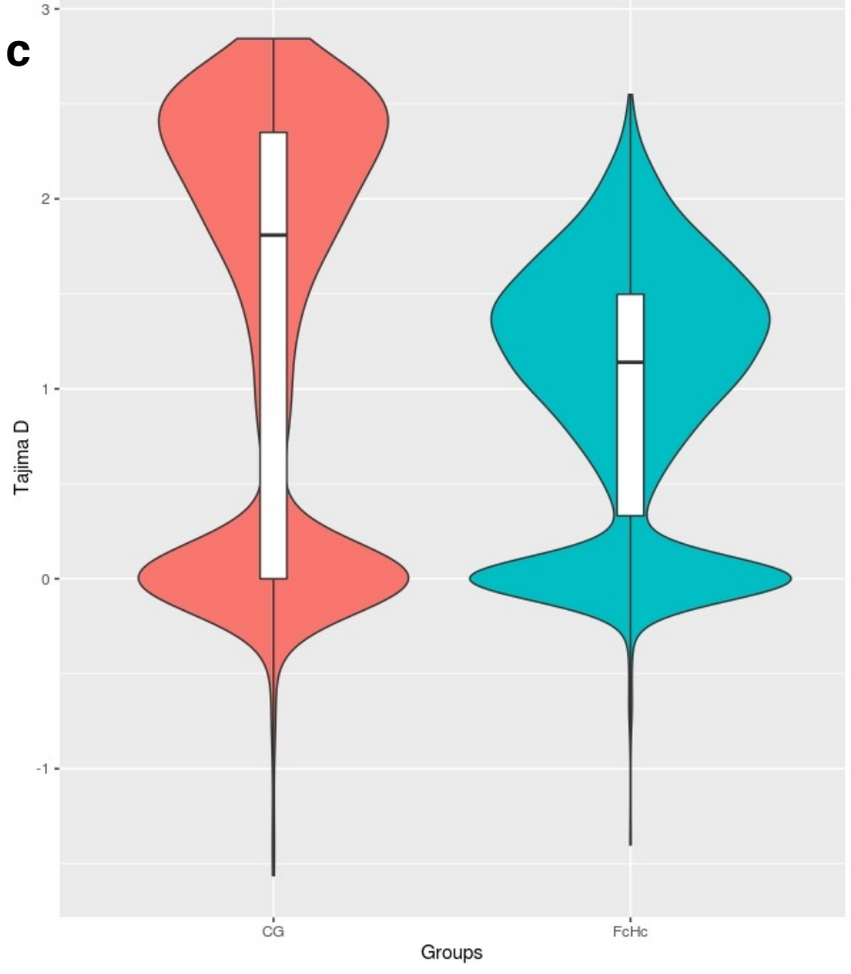
a



b

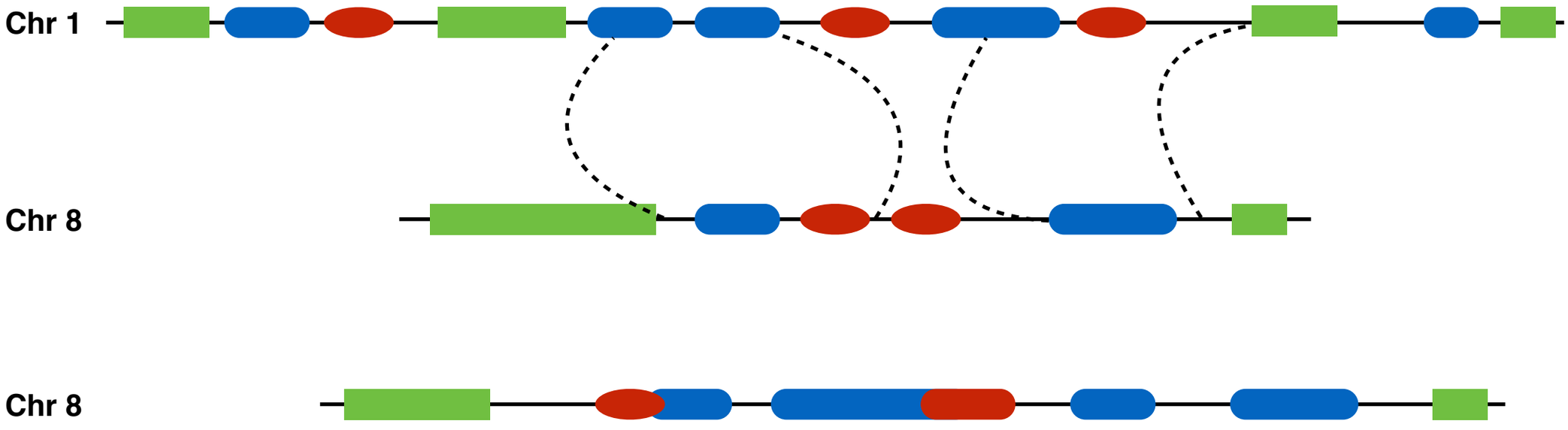


c



aCC-BY-ND 4.0 International license.

a

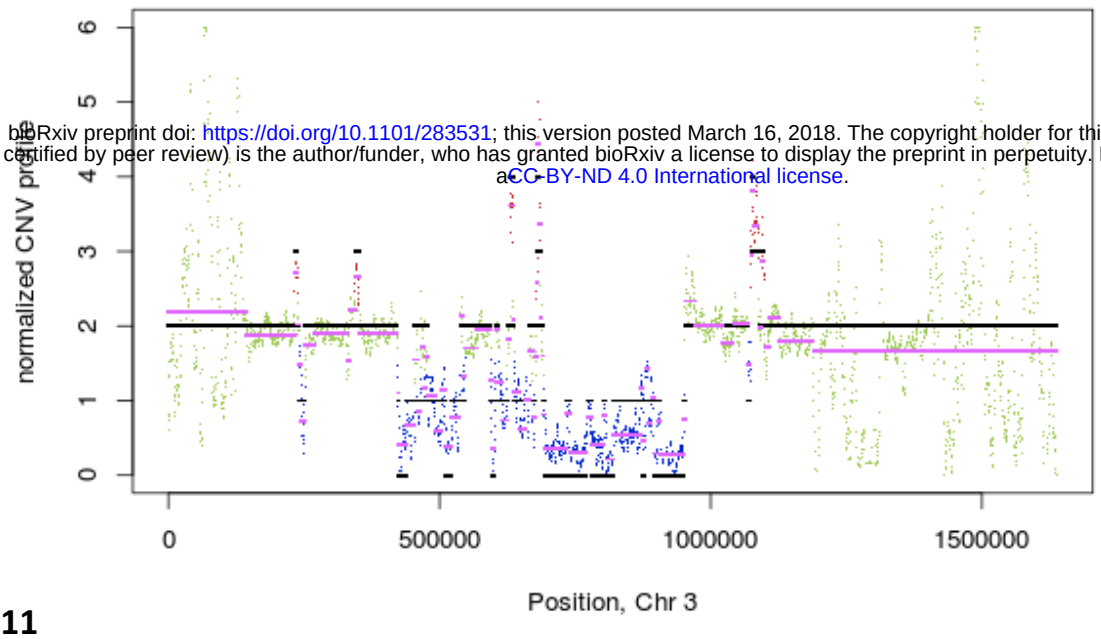


b

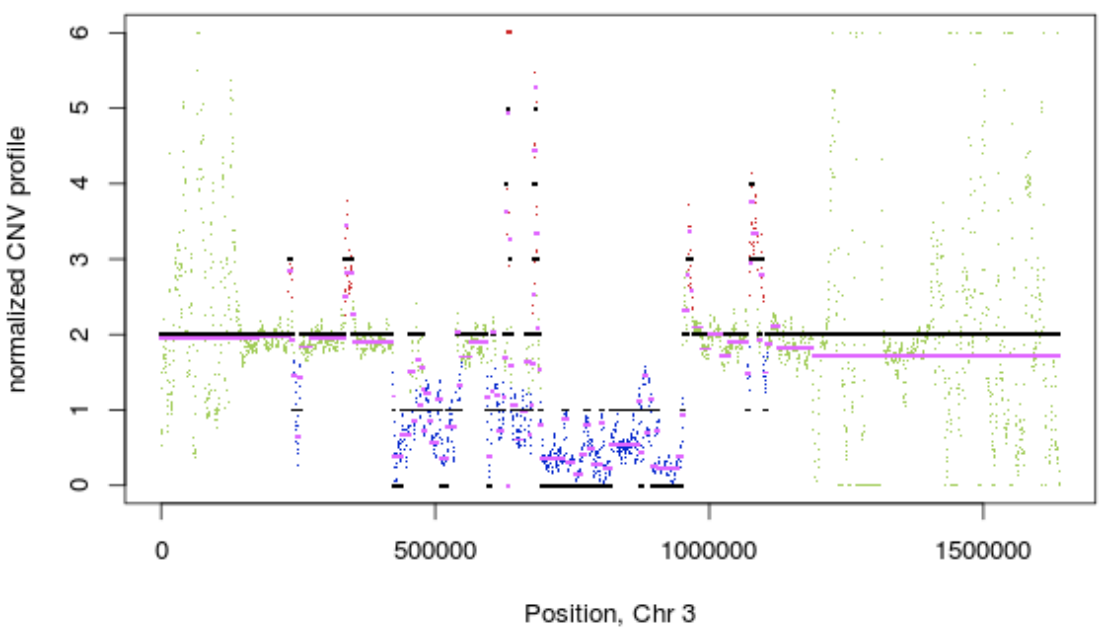
Distribution of CNV in chromosome 3 of the Colombian CG clones

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

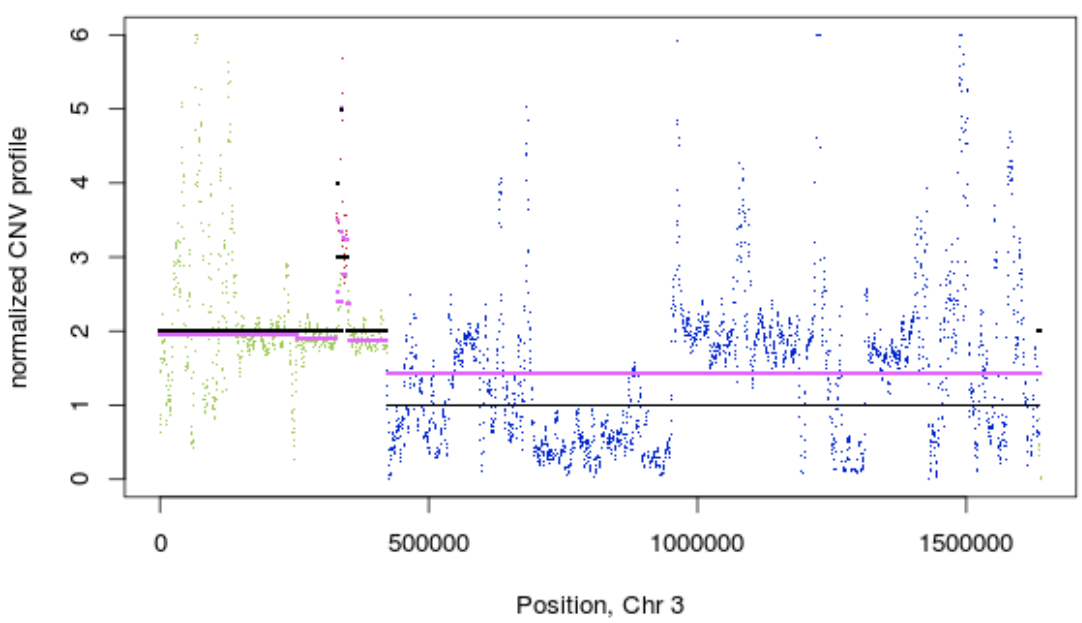
CGl10



CGl11



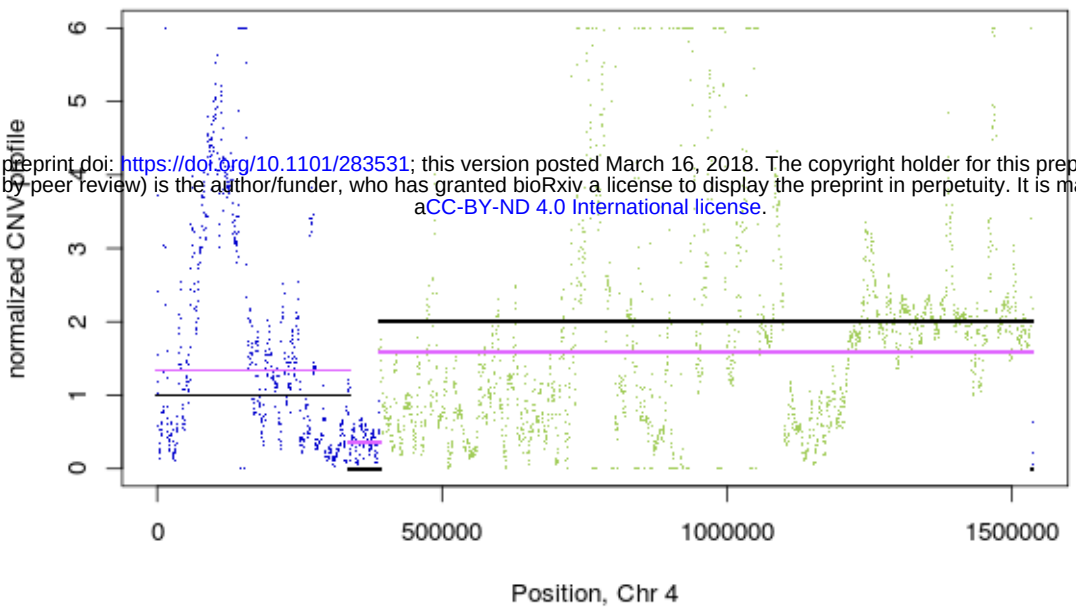
CGl13



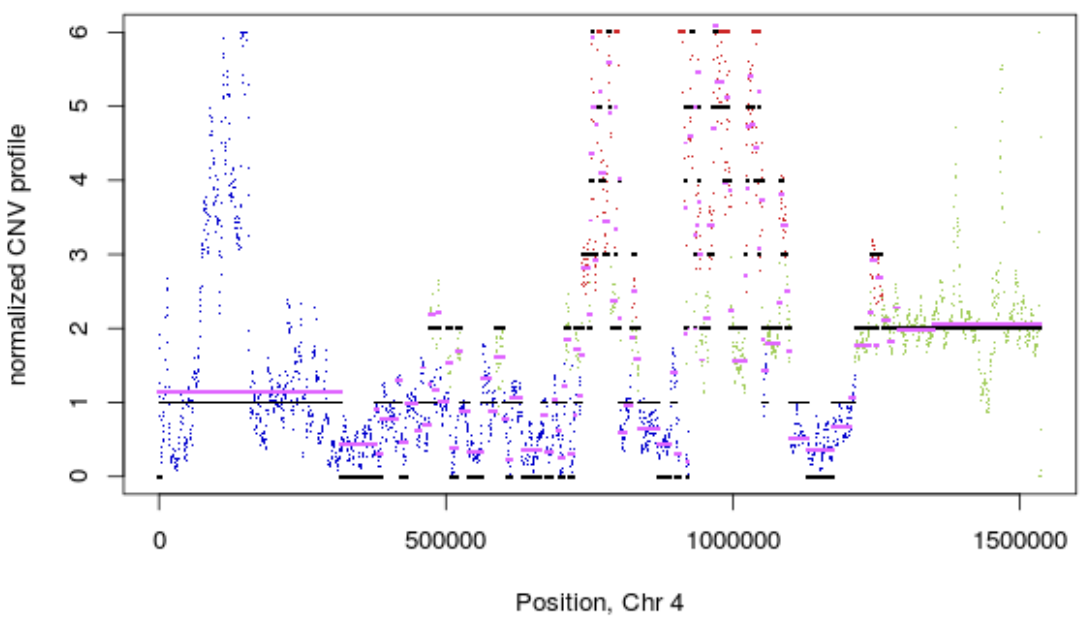
Distribution of CNV in chromosome 4 of the Colombian FcHc clones

bioRxiv preprint doi: https://doi.org/10.1101/283531; this version posted March 16, 2018. The copyright holder for this preprint (which was certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

FcHc1



FcHc2



FcHc4