

1 Genome-wide locus sequence typing (GLST) of eukaryotic pathogens

2
3 Philipp Schwabl^a, Jalil Maiguashca Sánchez^b, Jaime A. Costales^b, Sofía Ocaña^b, Maikell Segovia^c, Hernán J. Carrasco^c, Carolina Hernández^d, Juan David Ramírez^d, Michael D. Lewis^e, Mario J. Grijalva^{b,f} and Martin S. Llewellyn^a

6
7 ^aInstitute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK

9 ^bCentro de Investigación para la Salud en América Latina, Pontificia Universidad Católica del Ecuador, Quito, Ecuador

11 ^cLaboratorio de Biología Molecular de Protozoarios, Instituto de Medicina Tropical, Universidad Central de Venezuela, Caracas, Venezuela

13 ^dGrupo de Investigaciones Microbiológicas, Programa de Biología, Universidad del Rosario, Bogotá, Colombia

15 ^eLondon School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

16 ^fInfectious and Tropical Disease Institute, Biomedical Sciences Department, Heritage College of Osteopathic Medicine, Ohio University, 45701 Athens, OH, USA

20 Abstract

21 Analysis of genetic polymorphism is a powerful tool for epidemiological surveillance and research. Powerful
22 inference from pathogen genetic variation, however, is often restrained by limited access to representative
23 target DNA, especially in the study of obligate parasitic species for which *ex vivo* culture is resource-intensive
24 or bias-prone. Modern sequence capture methods enable pathogen genetic variation to be analyzed directly
25 from vector/host material but are often too complex and expensive for resource-poor settings where infectious
26 diseases prevail. This study proposes a simple, cost-effective ‘genome-wide locus sequence typing’ (GLST)
27 tool based on massive parallel amplification of information hotspots throughout the target pathogen genome.
28 The multiplexed polymerase chain reaction amplifies hundreds of different, user-defined genetic targets in a
29 single reaction tube, and subsequent agarose gel-based clean-up and barcoding completes library preparation
30 at under 4 USD per sample. Approximately 100 libraries can be sequenced together in one Illumina MiSeq
31 run. Our study generates a flexible GLST primer panel design workflow for *Trypanosoma cruzi*, the parasitic
32 agent of Chagas disease. We successfully apply our 203-target GLST panel to direct, culture-free
33 metagenomic extracts from triatomine vectors containing a minimum of 3.69 pg/μl *T. cruzi* DNA and further
34 elaborate on method performance by sequencing GLST libraries from *T. cruzi* reference clones representing
35 discrete typing units (DTUs) TcI, TcIII, TcIV, and TcVI. The 780 SNP sites we identify in the sample set
36 repeatably distinguish parasites infecting sympatric vectors and detect correlations between genetic and
37 geographic distances at regional (< 150 km) as well as continental scales. The markers also clearly separate
38 DTUs. We discuss the advantages, limitations and prospects of our method across a spectrum of
39 epidemiological research.

Introduction

Genome-wide single nucleotide polymorphism (SNP) analysis is a powerful and increasingly common approach in the study and surveillance of infectious disease. Understanding patterns of SNP diversity within pathogen genomes and across pathogen populations can resolve fundamental biological questions (e.g., reproductive mechanisms in *T. cruzi*¹, reconstruct past² and present transmission networks (e.g., *Staphylococcus* infections within hospitals)³ or identify the genetic bases of virulence^{4,5} and resistance to drugs (see examples from *Plasmodium* spp.^{6,7}). A number of obstacles, however, complicate access to representative, genome-wide SNP information using modern sequencing tools. Micro-pathogens are often sampled in low quantities and together with large amounts of host/vector tissue, microbiota, or environmental DNA. Sequencing is rarely viable directly from the infection source and studies have often found it necessary to isolate and culture the target organism to higher densities before extracting DNA. These additional steps, however, are resource-intensive and bias-prone. Pathogen isolation is less often attempted on asymptomatic infections and is less likely to succeed when levels of parasitaemia in a sample are low. Genomic sequencing data on the protozoan parasite *Leishmania infantum*, for example, has for such reasons come to exhibit major selection bias towards aggressive strains isolated by invasive sampling from canine hosts. A short look into the limited number of whole-genome sequencing (WGS) datasets available for *L. infantum* at the European Nucleotide Archive (ENA) quickly confirms this statement. Vector-isolated genomes have yet to be reported from the Americas and only a single study claims to have sequenced *L. infantum* from asymptomatic hosts⁸. Selection bias also often occurs due to competition among isolated strains. Studies on the kinetoplastid *Trypanosoma cruzi*, for example, are time and again confounded by growth and survival rate differences among genotypes in culture^{9–11}, and gradual reductions to genetic diversity are often observed over time¹². Karyotypic changes are also known to arise during *T. cruzi* micromanipulation and axenic growth^{13,14}.

A variety of approaches therefore aim to obtain genome-wide SNP information without first performing pathogen isolation and culturing steps. Some studies separate target sequences from total DNA or RNA by exploiting base modifications or transcriptional properties specific to the pathogen¹⁵, vector¹⁶ or host^{17,18}. Others describe the use of biotinylated hybridization probes^{19–22} or selective whole-genome amplification, e.g., based on the strand displacement function of phi29 DNA polymerase²³. Such techniques are costly and often excessive when a study's primary objective is to evaluate genetic distances and diversity among samples rather than to reconstruct complete haplotypes or investigate structural genetic traits. Epidemiological tracking and source attribution studies, for example, often benefit little from measuring invariant sequence areas or defining the complete architecture of sample genomes. Also pathogen typing or population assignment objectives primarily require information on polymorphic sites. It is nevertheless quite common to see such studies to undertake expensive WGS procedures only for final analyses to take place 'post-VCF'²⁴, i.e., using a list of diagnostic markers compiled from a small fraction of polymorphic reads.

Highly multiplexed polymerase chain reaction (PCR) amplicon sequencing offers a much more efficient option when obtaining genome-wide SNP information is the primary goal. First marketed under the name Ion

AmpliSeq by Thermo Fisher Scientific²⁵, the method consists in the simultaneous amplification of dozens to hundreds of DNA targets known or hypothesized to contain sequence polymorphism in the sample set. Each sample's resultant amplicon pool is then prepared for sequencing by index/adaptor ligation or in a subsequent 'barcoding' PCR. Panel construction is highly flexible, requiring only that the primers exhibit similar melting/annealing temperatures and a low propensity to cross-react. As such, target selection can be tailored to specific research goals, for example, to profile resistance markers²⁶ or to genotype neutral SNP variation for landscape genetic techniques²⁷. The potential to isolate and genotype pathogen DNA at high-resolution directly from uncultured sample types by multiplexed amplicon sequencing has however received little attention thus far. Simultaneous PCR-based detection of multiple pathogen species or genotypes is certainly common²⁸, but multiplexable primer panels are rarely designed for subsequent sequencing and polymorphism analysis. The Ion AmpliSeq brand currently offers pre-designed panels for studies on ebola²⁹ and tuberculosis³⁰ but the use of custom panels for other pathogen species (e.g., *Bifidobacterium*³¹ or human papilloma virus³²) remains surprisingly rare in the literature.

In this study we describe the design and implementation of a large multiplexable primer panel for *T. cruzi*, parasitic agent of Chagas disease. In contrast to past multi-locus sequence typing (MLST) methods involving at most 32 (individually amplified) gene fragments, our 'genome-wide locus typing' (GLST) tool simultaneously amplifies 203 sequence targets across 33 (of 47) *T. cruzi* chromosomes. We apply GLST to metagenomic DNA extracts from triatomine vectors collected in Colombia, Venezuela and Ecuador and further describe method sensitivity/specificity by sequencing GLST libraries from *T. cruzi* clones representing discrete typing units (DTUs) TcI, TcIII, TcIV, and TcVI. The 780 SNP sites identified from GLST amplicon sequencing repeatably distinguish parasites infecting sympatric vectors and detect correlations between genetic and geographic distances at regional (< 150 km) and continental scales. The markers also clearly separate DTUs. We discuss the advantages and limitations of our method for epidemiological studies in resource-poor settings where Chagas and other 'neglected tropical diseases' prevail.

Methods

Triatomine samples and *T. cruzi* reference clones

T. cruzi-infected intestinal tract and/or faeces samples of *Rhodnius ecuadoriensis* and *Panstrongylus chinai* were collected by the Centro de Investigación para la Salud en América Latina (CISeAL) in Loja Province, Ecuador, following protocols described in Grijalva et al. 2012³³. DNeasy Blood and Tissue Kit (Qiagen) was used to extract metagenomic DNA. Infected intestinal material of *Panstrongylus geniculatus*, *R. pallescens* and *R. prolixus* from northern Colombia was also collected in previous projects^{34–36}, likewise using DNeasy Blood and Tissue Kit to extract metagenomic DNA. *Panstrongylus geniculatus* specimens from Caracas, Venezuela were collected by the citizen science triatomine collection program (<http://www.chipo.chagas.ucv.ve/vista/index.php>) at Universidad Central de Venezuela. This program has supported various epidemiological studies in the capital district^{37–39}. DNA was extracted from the insect faeces

by isopropanol precipitation. Geographic coordinates and ecotypes (domestic, peri-domestic, or sylvatic) of the sequenced samples are provided in Supplementary Tbl. 1.

T. cruzi epimastigote DNA from reference clones Chile c22 (TcI) Arma18 cl. 1 (TcIII), Saimiri3 cl. 8 (TcIV), Para7 cl. 3 (TcVI), Chaco9 col. 15 (TcVI) and CL Brener (TcVI) was obtained from the London School of Hygiene & Tropical Medicine (LSHTM). DNA extractions at LSHTM followed Messenger et al. 2015⁴⁰.

Uninfected *Rhodnius prolixus* gut tissue samples used for mock infections (see ‘Method development and library preparation’) were also provided by LSHTM. Special thanks to C. Whitehorn and M. Yeo for supervising dissections. Insects were euthanized with CO₂ and hindguts drawn into 5 volumes of RNAlater (Sigma-Aldrich) by pulling the abdominal apex toward the posterior with sterile watchmaker’s forceps.

T. cruzi TcI X10/1 Sylvio reference clone (‘TcI-Sylvio’) epimastigotes used for mock infections and various other stages of method development were obtained from CISEAL. Cryo-preserved cells were returned to log-phase growth in liver infusion tryptose (LIT) and quantified by hemocytometer before pelleting at 25,000 g. Pellets were washed twice in PBS and parasites killed by resuspension in 10 volumes of RNAlater. DNA from these *T. cruzi* cells (and their dilutions with preserved *T. prolixus* intestinal tissue) was extracted by isopropanol precipitation.

Isopropanol precipitation was also used to extract DNA from *T. cruzi* plate clone TBM_2795_CL2. This sample was previously analyzed by WGS¹ and served as a control for GLST method development in this study.

GLST target and primer selection

We began our GLST sequence target selection process by screening single-nucleotide variants previously identified in *T. cruzi* populations from southern Ecuador¹. Briefly, Schwabl et al. sequenced genomic DNA from 45 cloned and 14 non-cloned *T. cruzi* field isolates on the Illumina HiSeq 2500 platform and mapped resultant 125 nt reads to the TcI-Sylvio reference assembly using default settings in BWA-mem v0.7.3⁴¹. Single-nucleotide polymorphisms (SNPs) were summarized by population-based genotype and likelihood assignment in Genome Analysis Toolkit v3.7.0⁴², excluding sites with low cumulative call confidence (QUAL < 1,500) and/or aberrant read-depth (< 10 or > 100) as well as those belonging to clusters of three or more SNPs. A ‘virtual mappability’ mask⁴³ was also applied to avoid SNP inference in areas of high sequence redundancy in the *T. cruzi* genome. Read-mapping and variant exclusion criteria were verified by subjecting TcI-Sylvio Illumina reads from Franzen et al. 2012⁴⁴ to the same pipelines as the Ecuadorian dataset. An additional mask was set around small insertion-deletions suggested to occur in these reads based on the assumption that the reference sample should not present alternate genotypes in high-quality contigs of the assembled genome.

We extracted 160 nt segments from the *T. cruzi* reference genome (.fasta file) whose internal sequence (positions 41 to 120) contained between one and ten of 75,038 SNPs identified in the above WGS dataset.

These 56,428 segments were further filtered for synteny between *T. cruzi* and *Leishmania major* genomes as defined by the OrthoMCL algorithm at TriTrypDB⁴⁵. Such conserved segments may be least prone to repeat-driven nucleotide diversity and as such most amenable to PCR⁴⁶. The 6,259 synteny segments found by OrthoMCL therefore proceeded to primer search with the high-throughput primer design engine BatchPrimer3⁴⁷. As target SNPs did not occur in the outer 40 nt of each synteny segment, these flanking regions provided additional flexibility to identify primers matching the following criteria:

- min. size = 24 nt
- max. size = 35 nt
- optimal size = 24 nt
- min. product size = 120 nt
- max. product size = 160 nt
- optimal product size = 120 nt
- min. melting temperature = 63 °C,
- max. melting temperature = 65 °C,
- optimal melting temperature = 63 °C,
- max. self-complementarity: 4 nt
- max. 3' self-complementarity: 2 nt
- max. length of mononucleotide repeats = 3 nt
- min. GC content = 40%
- max. GC content = 60%

Each of 286 forward primer candidates output by BatchPrimer3 received the additional 5' tag sequence 5'-ACACTGACGACATGGTTCTACA-3' and reverse primer candidates received the 5' tag sequence 5'-TACGGTAGCAGAGACTTGGTCT-3'. These tag sequences enable single-end barcode and Illumina P5/P7 adaptor attachment in second-round PCR. Next, we determined binding energies (ΔG) for all possible primer-pairs using the primer compatibility software MultiPLX v2.1.4. We discarded primers with inter-quartile ranges crossing a threshold of $\Delta G = -12.0$ kcal/mol. Primers with 20 or more interactions showing $\Delta G \leq -12.0$ kcal/mol were also disallowed. The remaining 248 primer-pairs (median $\Delta G = -9.0$) underwent a last filtering step by screening for perfect matches in raw WGS sequence files (.fastq). Low match frequency led to the elimination of 45 additional primer pairs. WGS alignments corresponding to the 203 sequence regions targeted by this final primer set were visualized in Belvu v12.4.3⁴⁸. The 403 SNPs occurring within these sequence regions distributed evenly across individuals in Loja Province. Using the 'nj' function from the 'ape' package v5.0 in R v3.4.1⁴⁹, the 403 SNPs also reproduced neighbor-joining relationships observed based on total polymorphism identified by WGS (Supplementary Fig. 1). These observations lent further support to the suitability of the GLST marker panel for the analysis of genetic differentiation at the landscape-scale. The GLST sequence target selection process described above is summarized in Fig. 1.

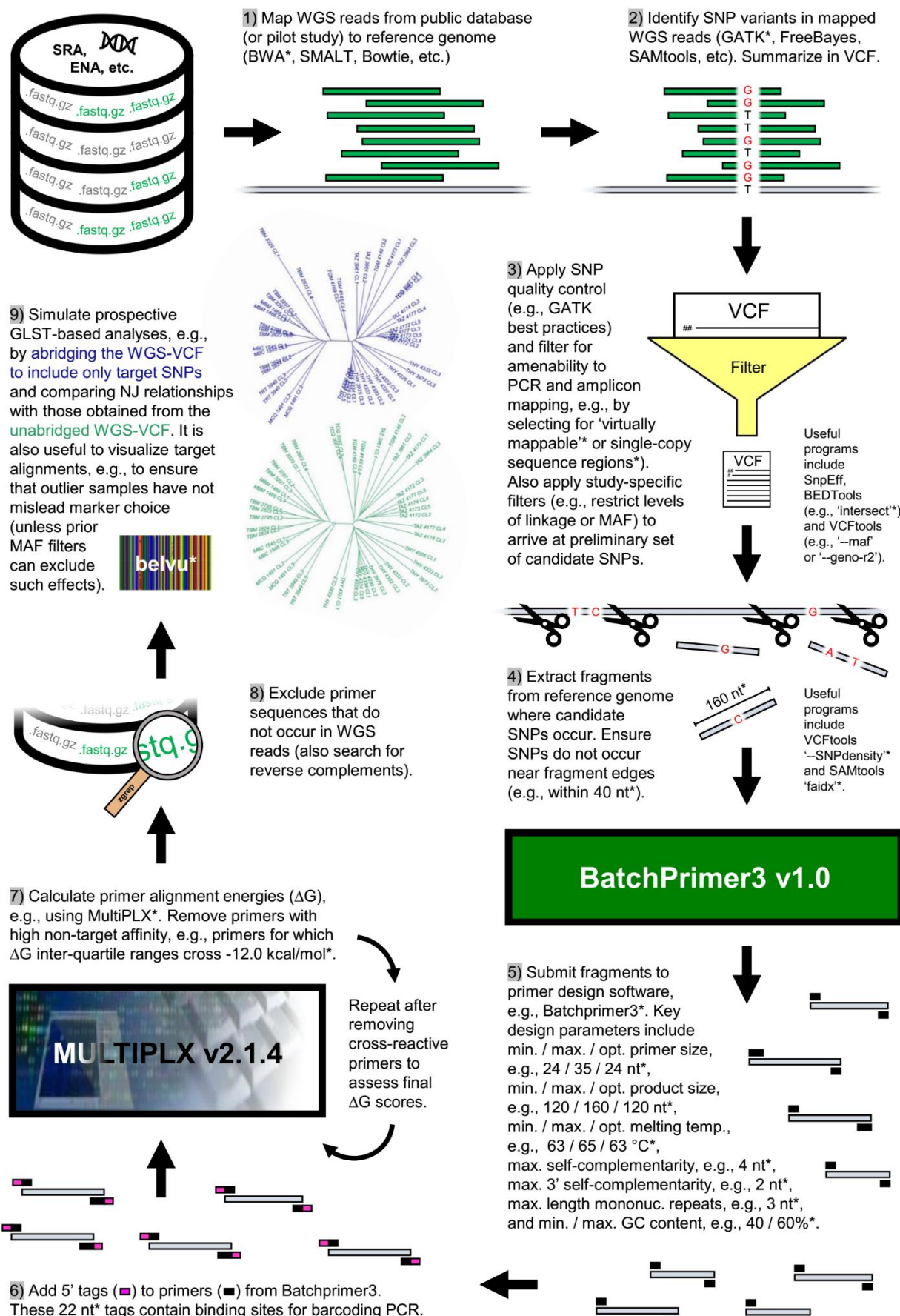


Figure 1 GLST sequence target selection from preliminary genomic data. Nine steps of primer panel construction and validation run clockwise from top left. Various methods and criteria can be applied to complete many of these steps. Those specific to this study are asterisked, e.g., we used BWA in step 1 and GATK in step 2. Abbreviations: SRA (Sequence Read Archive at www.ncbi.nlm.nih.gov/sra); ENA (European Nucleotide Database at www.ebi.ac.uk/ena); WGS (whole-genome sequencing); SNP (single-nucleotide polymorphism); MAF (minor allele frequency); PCR (polymerase chain reaction); VCF (variant call format); NJ (neighbor-joining).

Wet lab method development and library preparation

The 203 primers pairs designed above (Supplementary Tbl. 2) were purchased from Eurofins Genomics (Ebersberg, Germany) at 200 μ M concentration in salt-free, 96-well plate format. Primer pairs were first tested individually to establish cycling conditions for PCR (Supplementary Fig. 2). Optimal target amplification occurred with an initial incubation step at 98 °C (2 min); 30 amplification cycles at 98 °C (10 s), 60 °C (30 s), and 72 °C (45 s); and a final extension step at 72 °C (2 min). The 10 μ l reactions contained 5 μ l Q5 High-Fidelity Master Mix (New England Biolabs), 1 μ l forward primer [10 μ M], 1 μ l reverse primer [10 μ M], and 3 μ l TcI-Sylvio epimastigote DNA. The multiplexed, first-round ‘GLST’ PCR reaction was prepared by combining all 406 primers in equal proportions and diluting the combined mix to 50.75 μ M, resulting in individual primer concentrations of 50.75 μ M / 406 = 125 nM. GLST reactions incorporated 2 μ l of this primer mix rather than two separate 1 μ l forward/reverse primer inputs as above.

We first tested GLST PCR on DNA extracts from mock infections, each consisting of 10^4 , 10^5 or 10^6 TcI-Sylvio epimastigote cells and one uninfected *R. prolixus* intestinal tract (Supplementary Fig. 3). Amplicons from lower concentration epimastigote dilutions gave weaker signals in gel electrophoresis, suggesting lower infection load thresholds at which vector gut DNA becomes unsuitable for GLST. Most vector gut DNA extracts obtained for this study represented donated material of limited quality and infection load, some samples were also without signal in PCR spot tests for the presence of high frequency ‘TcZ’⁵⁰ satellite DNA (commonly targeted to diagnose human *T. cruzi* infections).

We therefore first used qPCR to identify vector gut samples containing *T. cruzi* DNA quantities within ranges successfully visualized from GLST reactions on epimastigote DNA quantified by Qubit fluorometry (Invitrogen) and serially diluted from 1.35 ng/ μ l to 2.50 pg/ μ l in dH₂O (Supplementary Fig. 4). Each 20 μ l qPCR reaction consisted of 10 μ l SensiMix SYBR Low-ROX reagent (Bioline), 1 μ l TcZ forward primer (5'-GCTCTTGCCACAMGGGTGC-3')⁵⁰ [10 μ M], 1 μ l TcZ reverse primer (5'-CCAAGCAGCGGATAGTTCAGG-3')⁵⁰ [10 μ M], 7 μ l dH₂O, and 1 μ l vector gut DNA. Samples were amplified together with a 15-step standard curve containing between 0.30 pg and 4.82 ng *T. cruzi* epimastigote DNA. Reaction conditions consisted of an initial incubation step at 95 °C (10 min) and 40 amplification cycles at 95 °C (15 s), 55 °C (15 s), and 72 °C (15 s). Fluorescence acquisition occurred at the end of each cycle and final product dissociation was measured in 0.5 °C increments between 55 and 95 °C.

Vector gut samples suggested to contain at least 1.0 pg/ μ l *T. cruzi* concentrations based on qPCR proceeded to final library construction (Supplementary. Tbl. 1) alongside DNA from *T. cruzi* clones TBM_2795_cl2 (TcI), Chile c22 (TcI) Arma18 cl. 1 (TcIII), Saimiri3 cl. 8 (TcIV), Para7 cl. 3 (TcV), Chaco9 col. 15 (TcVI) and CL Brener (TcVI). Several samples were processed in 2 – 4 replicates beginning with the first-round GLST PCR reaction step. First-round PCR products were electrophoresed in 0.8% agarose gel to separate target bands (mode =164 nt) from primer polymers quantified with the Agilent Bioanalyzer 2100 System (see 78 nt primer peak in Supplementary Fig. 5). Excised target bands were resolubilized with the PureLink Quick

Gel Extraction Kit (Invitrogen) to create input for subsequent barcoding PCR. This second PCR reaction consisted of an initial incubation step at 98 °C (2 min); 7 amplification cycles at 98 °C (30 s), 60 °C (30 s), and 72 °C (1 min); and a final extension step at 72 °C (3 min). Only 7 amplification cycles were used given polymer ‘daisy-chaining’ observed when cycling at 13 and 18x (Supplementary Fig. 6). The barcoding reaction adds Illumina flow cell and sequencing primer binding sites to each first-round PCR product. A different reverse primer is used for each sample. The reverse primer (5’-CAAGCAGAAGACGGCATACGAGAT*X*TACGGTAGCAGAGACTTGGTCT-3’) contains a 10 nt barcode (*X*) to distinguish reads from different samples during pooled sequencing. It also contains CS2 (sequencing primer binding sites). A single forward primer (5’-AATGATACGGCGACCACCGAGATCTACACTGACGACATGGTTCTA-3’) containing CS1 is used for all samples. Each 20 µl barcoding reaction contained 10 µl Q5 High-Fidelity Master Mix (New England Biolabs), 0.8 µl forward (universal) primer [10 µM], 0.8 µl (barcoded) reverse primer [10 µM], 5.4 µl dH₂O and 3 µl (gel-purified) first-round PCR product. Barcoding primers were purchased from Eurofins Genomics at 100 µM concentration in HPLC-purified, 96-well plate format. Barcoded amplicons (e.g., Supplementary Fig. 7) were quantified by Qubit fluorometry (Thermo Fisher Scientific), and pooled at equimolar concentrations, gel-excised, re-solubilized, and verified by microfluidic electrophoresis (Supplementary Fig. 8) as above.

GLST amplicon sequencing and variant discovery

The GLST pool was sequenced twice on an Illumina MiSeq instrument. We first used the pool to ‘spike’ additional base diversity into a collaborator’s 16S amplicon sequencing run. 16S samples were loaded to achieve 80% sequence output whereas GLST and PhiX DNA⁵¹ were each loaded at 10%. This first run occurred in 500-cycle format using MiSeq Reagent Kit v2. The second run occurred in 300-cycle format using MiSeq Reagent Micro Kit v2 and was dedicated solely to GLST (also no PhiX). Both runs were performed at Glasgow Polyomics using Fluidigm Custom Access Array sequencing primers FL1 (CS1 + CS2) and CS2rc⁵².

Demultiplexed sequence reads were trimmed to 120 nt and mapped to the TcI-Sylvio reference assembly using default settings in BWA-mem v0.7.3. Mapped reads with poor alignment scores (AS < 100) were discarded to decontaminate samples of non-*T. cruzi* sequences sharing barcodes with the GLST dataset. Identical results were achieved using BWA-sw in DeconSeq v0.4.3⁵³ to decontaminate reads. After merging alignment (.bam) files from sequencing runs 1 and 2 with Picard Tools v1.11⁵⁴, single-nucleotide polymorphisms (SNPs) were identified in each sample using the ‘HaplotypeCaller’ algorithm in GATK v3.7.0⁴². Population-based genotype and likelihood assignment followed using ‘GenotypeGVCFs’. We excluded SNP sites with QUAL < 80, D < 10, Mapping Quality (MQ) < 80 and or Fisher Strand Bias (FS) > 10. Individual genotypes were set to missing (./.) if they contained < 10 reads and set to reference (0/0) if they contained only a single alternate read (i.e., if they were classified as heterozygotes based on minor allele frequencies ≤ 10%). These filtering thresholds were cleared by all expected SNPs (i.e., SNPs also found in prior WGS sequencing) but not by all new SNPs found using GLST (e.g., see comparison of QUAL density curves in Supplementary Fig. 9). SNP

calling with GATK was also performed separately for sequencing runs 1 and 2 in order to exclude SNP sites uncommon to both analyses from the merged dataset described above.

GLST repeatability, population genetic and spatial analyses

We used PopART v1.7 to plot genetic differences between samples and sample replicates as a median-joining network, i.e., a minimum spanning tree composed of observed sequences and unobserved (reconstructed) sequence nodes⁵⁵. Genetic differences were measured by applying the ‘vcf-to-tab’ script from VCFtools v0.1.13 to the filtered SNP dataset, concatenating each sample’s output fields and counting the number of mismatching alleles (0, 1 or 2) per site and sample pair. A phylogenetic tree was built by counting the number of non-reference alleles in each genotype with the VCFtools function ‘--012’, summing pairwise Euclidean distances at biallelic sites and plotting neighbor-joining relationships with the ‘nj’ function from the ‘ape’ package v5.0 in R v3.4.1⁴⁹.

Considering only the first replicate of multiply sequenced samples, linkage and neutrality statistics were calculated using VCFtools functions ‘--geno-r2’ (calculates correlation coefficients between genotypes following Purcell et al.⁵⁶), ‘--het’ (calculates inbreeding coefficients using a method of moments⁵⁷) and ‘--hwe’ (filters sites by deviation from Hardy-Weinberg Equilibrium following Wigginton et al.⁵⁸). F_{ST} differentiation was calculated using ARLSUMSTAT v3.5.2^{59,60}.

Correlations between geographic and genetic differences were also calculated from non-reference allele counts in R v3.4.1⁴⁹. The ‘mantel’ function from the ‘vegan’ package v2.4.4⁶¹ was used to test significance of the Mantel statistic by permuting geographic distances and re-measuring correlations to genetic distances 999 times. Again, we used only the first replicate for samples with replicate sets. DTU reference clones were also excluded from analysis. Geographic distances were measured by projecting sample latitude/longitude (WGS 84) coordinates into a common xy plane (EPSG code 3786) selected following Šavrič et al. 2016⁶² (Supplementary Tbl. 1). EPSG 3786 projection was also used to map samples with the Natural Earth quick start kit in QGIS v2.18.4.

Given that missing information in sequence alignment can confound inference on genetic distances between samples⁶³, above repeatability and phylogenetic analyses excluded SNP sites in which genotypes were missing for any individual, and mantel analyses excluded SNP sites in which genotypes were missing in > 10% individuals. These exclusion criteria initially led to significant information loss due to the presence of two outlier samples, ARMA18_CL1_rep2 and COL253, libraries of which had been sequenced despite poor target visibility in gel electrophoresis (i.e., final PCR product banding appeared similar to that of ECU2 in Supplementary Fig. 7). Read-depths for the two samples ended up averaging 1.2 interquartile ranges below the sample set median and precluded genotype assignment at > 25% SNP sites. We therefore decided to exclude them from all analyses.

Results

SNP polymorphism and repeatability

GLST amplicons contained a total of 780 SNP sites, 387 polymorphic among TcI samples and 393 private to non-TcI reference clones (Fig. 2). Median read-depth was 266x across all sites. Of 403 loci targeted from the WGS dataset¹, 97% (391) were recovered by GLST and 82 contained polymorphism outside of Ecuador. GLST recovered 80 of 87 SNPs previously identified in TBM_2795_CL2 using WGS. Minimum parasite DNA concentration successfully genotyped from metagenomic DNA was 3.69 pg/μl (sample ECU36 – see Supplementary Fig. 10).

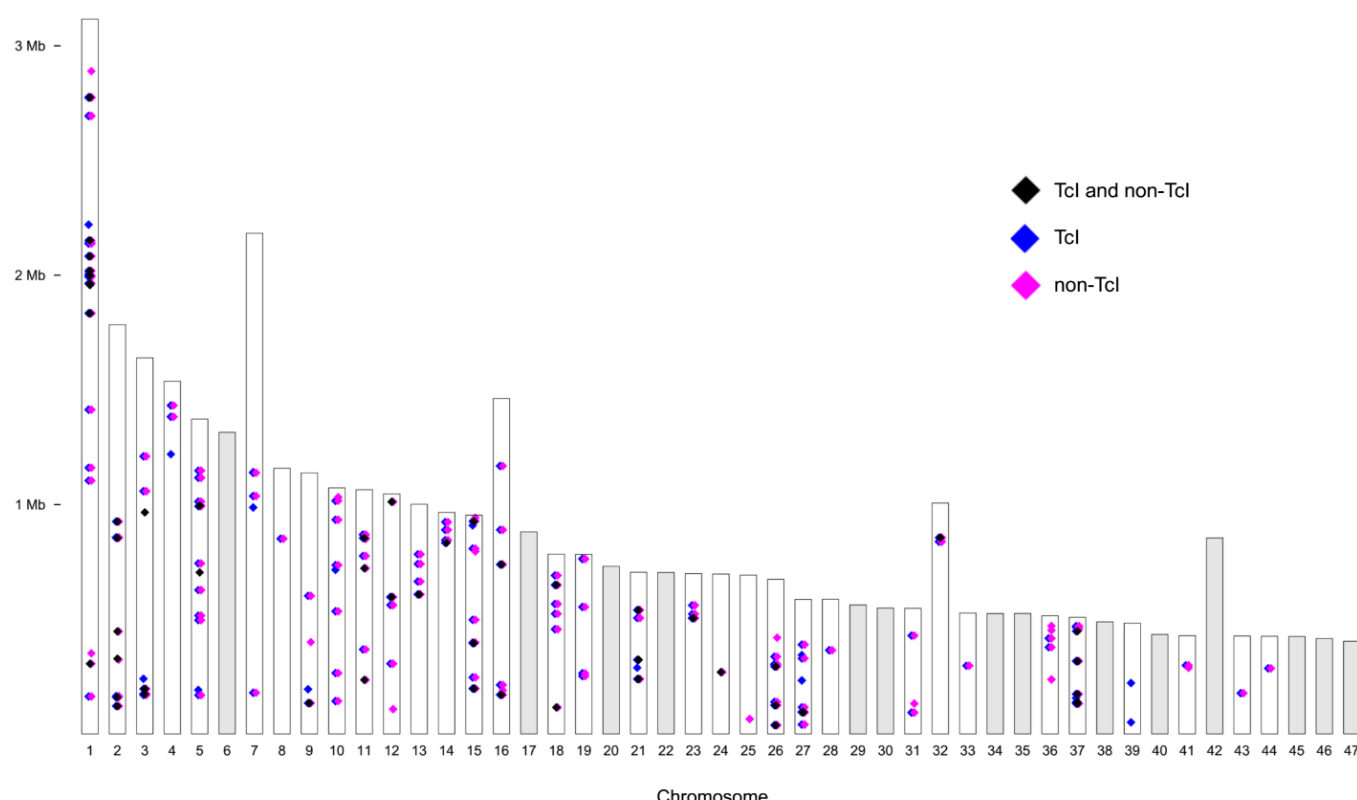


Figure 2 Variant loci detected in *T. cruzi* I samples and reference clones of other sub-lineages. The genome-wide distribution of SNP variants is shown relative to the TcI-Sylvio reference assembly. Each column represents one of 47 putative chromosomes. Pink diamonds comprise 393 variants that occur only in non-TcI samples. The remaining 387 variants are private to (blue) or shared by TcI and other sub-lineages (black). Diamonds representing nearby SNPs (e.g., those occurring on the same GLST target segment) overlap at this scale.

The TBM_2795_CL2 control sample underwent GLST in four replicates. These replicates were identical at all 561 SNP sites for which genotypes were called in all samples of the dataset. Median number of allelic differences (AD = 0, 1 or 2 per site) at non-missing sites between other replicate pairs was 3 (Tbl. 1). Pairwise AD did not correlate to minimum, maximum or difference in mean read-depth between the two replicates ($p < 0.80$).

Read-mapping coverage was inconsistent among replicates but strongly correlated between sequencing runs (Pearson's $r = 0.93$, $p < 0.001$) (Supplementary Figs. 11 – 12). Variant calling was also highly consistent: prior to variant filtration, only 10 SNP sites were called from run1 that were not also called from run 2 (these were excluded from analysis – see Methods).

Differentiation among *T. cruzi* individuals, sampling areas and sub-lineages

Sampling sites in Colombia, Venezuela and Ecuador are plotted in Fig. 3, and a median-joining network of allelic differences among GLST genotypes is shown in Fig. 4. GLST clearly distinguished TcI individuals at common collection sites in Soata (COL466 vs. COL468, AD = 37), Paz de Ariporo (COL133 vs. COL135, AD = 33), Tamara (COL154 vs. COL155 AD = 107) and Lebrija (COL77 vs. COL78, AD = 43) municipalities of Colombia but not in the community of Bramaderos (ECU3 vs. ECU8 vs. ECU10, AD = 0) in Loja Province, Ecuador. Samples from nearby sites within Caracas, Venezuela were also clearly distinguished by GLST (e.g., VZ16816 vs. VZ17114, AD = 43).

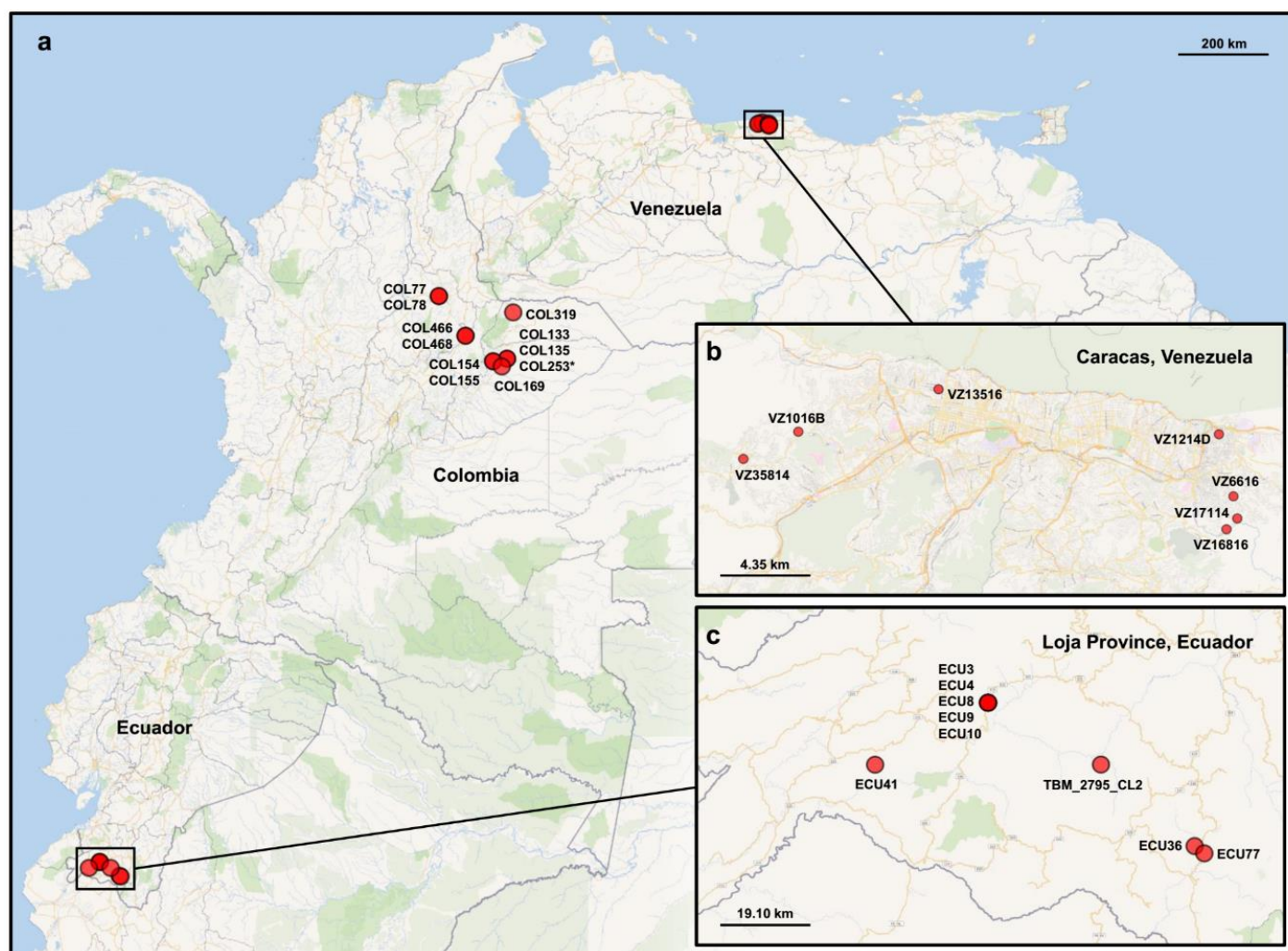


Figure 3 Map of vector sampling sites. **a** Sampling in Colombia involved a larger spatial area than that in Venezuela and Ecuador. *T. cruzi*-infected intestinal material was collected from *Panstrongylus* and *Rhodnius* vectors in Arauca, Casanare, Santander and Boyacá. We asterisk COL253 because low read-depth led to sample exclusion. **b** *P. geniculatus* material from Venezuela was collected within the Metropolitan District of Caracas. **c** *R. ecuadoriensis* and *P. chinai* material from Ecuador was collected in Loja Province. Supplementary Tbl. 1 lists coordinates and other details.

Nucleotide diversity (π = mean pairwise AD) was higher in samples from Caracas (π = 29.0) than in those from Loja Province (π = 22.8) but not in those from Colombia (π = 43.2) (Tbl. 2). Hardy-Weinberg ratios, linkage and inbreeding coefficients are also listed in Tbl. 2.

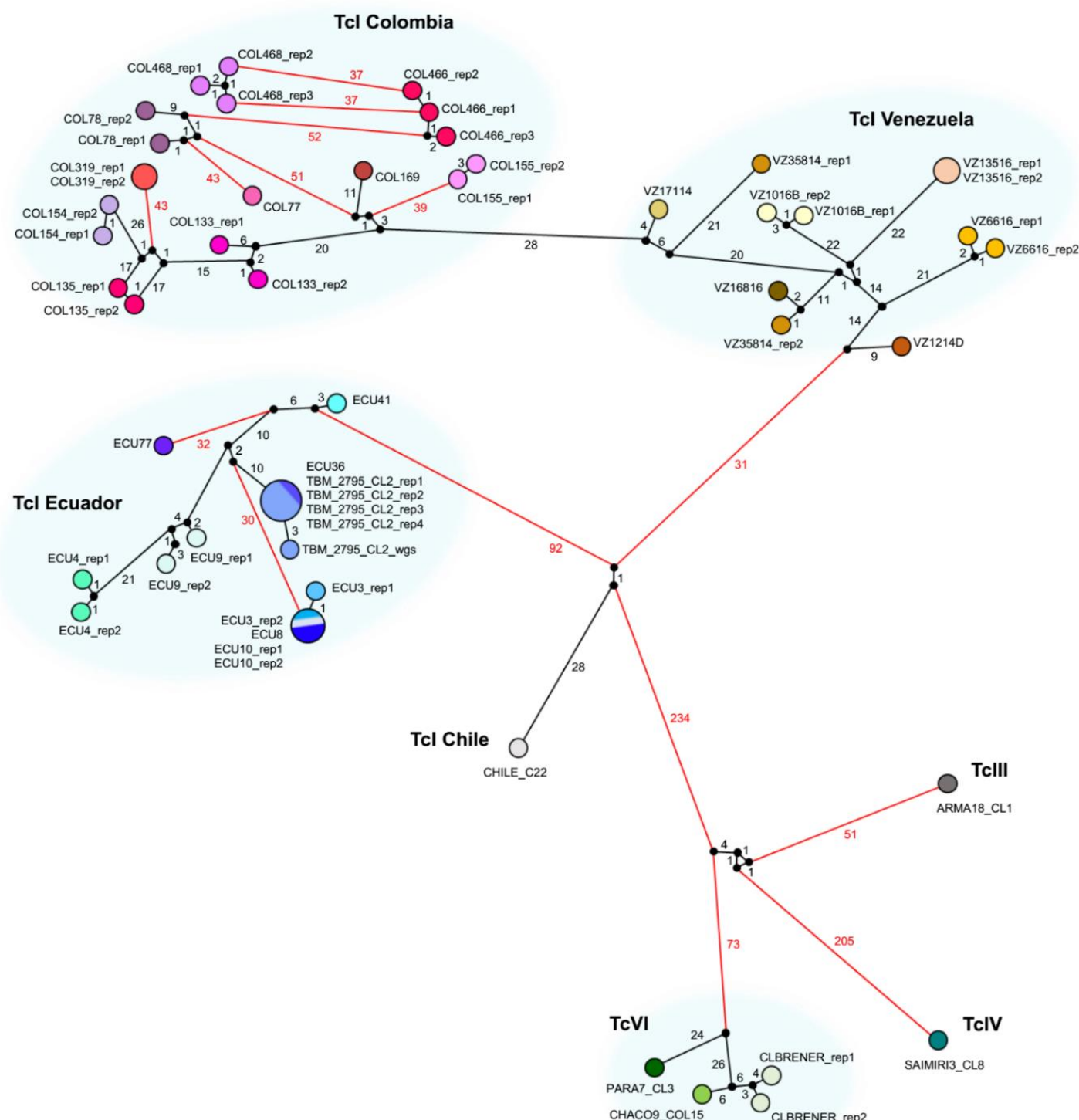


Figure 4 Allelic differences among *T. cruzi* I samples and reference clones of other sub-lineages as a median-joining network. A single SNP locus can differ by 0, 1 or 2 between two individuals (i.e., the individuals match at both, one, or neither allele). The AD measurement indicated on each edge of the network represents the total number of differences across all loci for which genotypes were called in all individuals of the dataset ($n = 561$). Red edges indicate differences of 30 and above. Technical replicates are represented by circles of the same fill color. Larger circles represent the occurrence of identical GLST genotypes. Edge length is not directly proportional to AD.

Table 1 Allelic differences between GLST replicates. Eighteen samples were processed in 2 – 4 replicates after DNA extraction. A single SNP locus can differ by 0, 1 or 2 between two replicates (i.e., replicates can match at both, one, or neither allele). The AD measurement represents the total number of pairwise differences across all loci for which genotypes are called in all individuals (n = 561). The discrepancy between VZ35814 replicates likely represents barcode contamination with VZ16816 (see close similarity in Fig. 3).

Replicate comparison	AD
COL319_rep1 vs. COL319_rep2	0
ECU10_rep1 vs. ECU10_rep2	0
TBM_2795_CL2_rep1 vs. TBM_2795_CL2_rep2	0
TBM_2795_CL2_rep1 vs. TBM_2795_CL2_rep3	0
TBM_2795_CL2_rep1 vs. TBM_2795_CL2_rep4	0
TBM_2795_CL2_rep2 vs. TBM_2795_CL2_rep3	0
TBM_2795_CL2_rep2 vs. TBM_2795_CL2_rep4	0
TBM_2795_CL2_rep3 vs. TBM_2795_CL2_rep4	0
VZ13516_rep1 vs. VZ13516_rep2	0
COL154_rep1 vs. COL154_rep2	1
COL466_rep1 vs. COL466_rep2	1
ECU3_rep1 vs. ECU3_rep2	1
COL135_rep1 vs. COL135_rep2	2
COL468_rep1 vs. COL468_rep2	2
ECU4_rep1 vs. ECU4_rep2	2
COL155_rep1 vs. COL155_rep2	3
COL466_rep1 vs. COL466_rep3	3
COL468_rep1 vs. COL468_rep3	3
COL468_rep2 vs. COL468_rep3	3
VZ6616_rep1 vs. VZ6616_rep2	3
COL466_rep2 vs. COL466_rep3	4
VZ1016B_rep1 vs. VZ1016B_rep2	4
CL_Brener_rep1 vs. CL_Brener_rep2	7
COL133_rep1 vs. COL133_rep2	9
ECU9_rep1 vs. ECU9_rep2	10
COL78_rep1 vs. COL78_rep2	12
VZ35814_rep1 vs. VZ35814_rep2	49

Table 2 Basic diversity statistics for *T. cruzi* I samples from Colombia (COL), Venezuela (VZ) and Ecuador (ECU). Abbreviations: n (sample size); PS (polymorphic sites); HWE (Hardy-Weinberg equilibrium); F_{IS} (inbreeding coefficient), r^2 (linkage coefficient), π (nucleotide diversity), Q (quartile); M (median); F_{ST} (between-group fixation index).

Group (n)	PS	PS in HWE	F_{IS} (Q1, M, Q3)	r^2 (Q1, M, Q3)	π	F_{ST} to COL	F_{ST} to VZ	F_{ST} to ECU
COL (11)	175	169	-0.19, 0.13, 0.24	0.03, 0.07, 0.19	43.2	0.000	0.136	0.595
VZ (7)	147	143	-0.35, -0.19, 0.11	0.02, 0.09, 0.27	29.0	0.136	0.000	0.632
ECU (9)	148	142	-0.20, -0.09, 0.18	0.04, 0.17, 0.36	22.8	0.595	0.632	0.000

Genetic distances increased with spatial distances among samples (Mantel's $r = 0.89$, $p = 0.001$), but the correlation coefficient was largely driven by high F_{ST} between sample sets from Colombia/Venezuela and Ecuador (Tbl. 2 and Fig. 5a): Mantel's r decreased to 0.30 ($p = 0.001$) after restricting analysis to sample pairs separated by < 250 km (Fig. 5b). Within-country IBD appeared to grow stronger for samples separated by < 150 km (Mantel's $r = 0.48$, $p = 0.002$) given a lack of correlation observed at higher distance classes within the Colombian dataset (Fig. 5b).

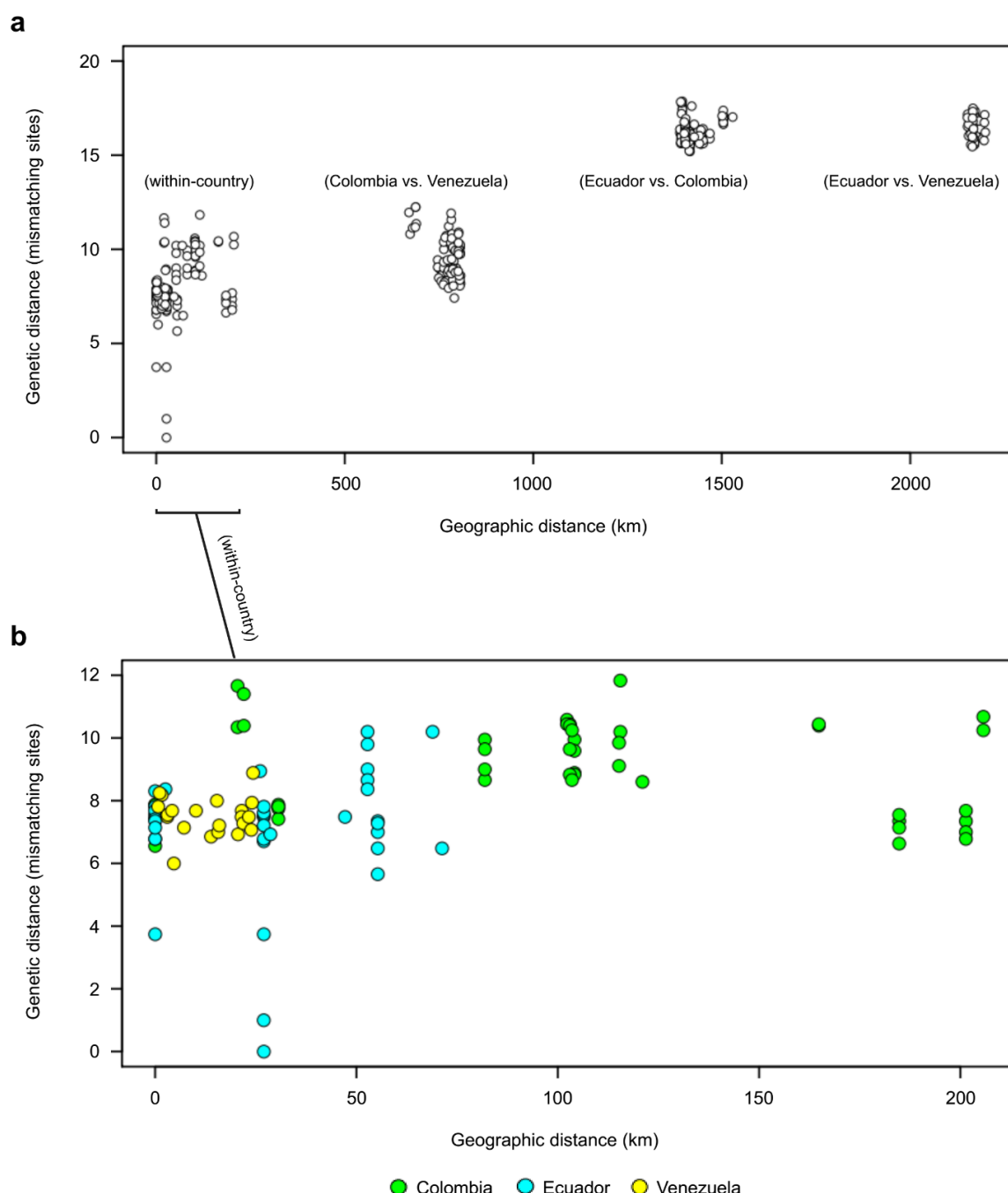


Figure 5 Isolation-by-distance among *T. cruzi* I samples. **a** Each circle represents geographic and genetic distances between two TcI samples. Global isolation-by-distance (IBD) is significant (Mantel's $r = 0.89$, $p = 0.001$) but driven by divergence between Ecuadorian samples and the rest of dataset (see two clusters at top right). **b** Nevertheless, IBD remains significant for within-country comparisons at < 250 km (Mantel's $r = 0.30$, $p = 0.009$) and < 150 km (Mantel's $r = 0.48$, $p = 0.002$). Green, cyan and yellow fill colors represent comparisons within Colombia, Ecuador and Venezuela, respectively. Each of the above Mantel tests remains significant when sample pairs with genetic distances < 2 are removed (see arrows). Only variant sites with $\leq 10\%$ missing genotypes ($n = 285$) are used in analysis. Only the first replicate is used for samples represented by multiple replicates.

Finally, GLST also clearly separated sub-lineages TcI, TcIII, TcIV, and TcVI in network (Fig. 3) and neighbor-joining tree construction (Fig. 6). AD between reference clones of different sub-lineages ranged from 153 (Arma18 cl1 (TcIV) vs. Para7 cl.3 (TcV)) to 472 (Chile c22 (TcI) vs. Saimiri3 cl. 8 (TcIV)).

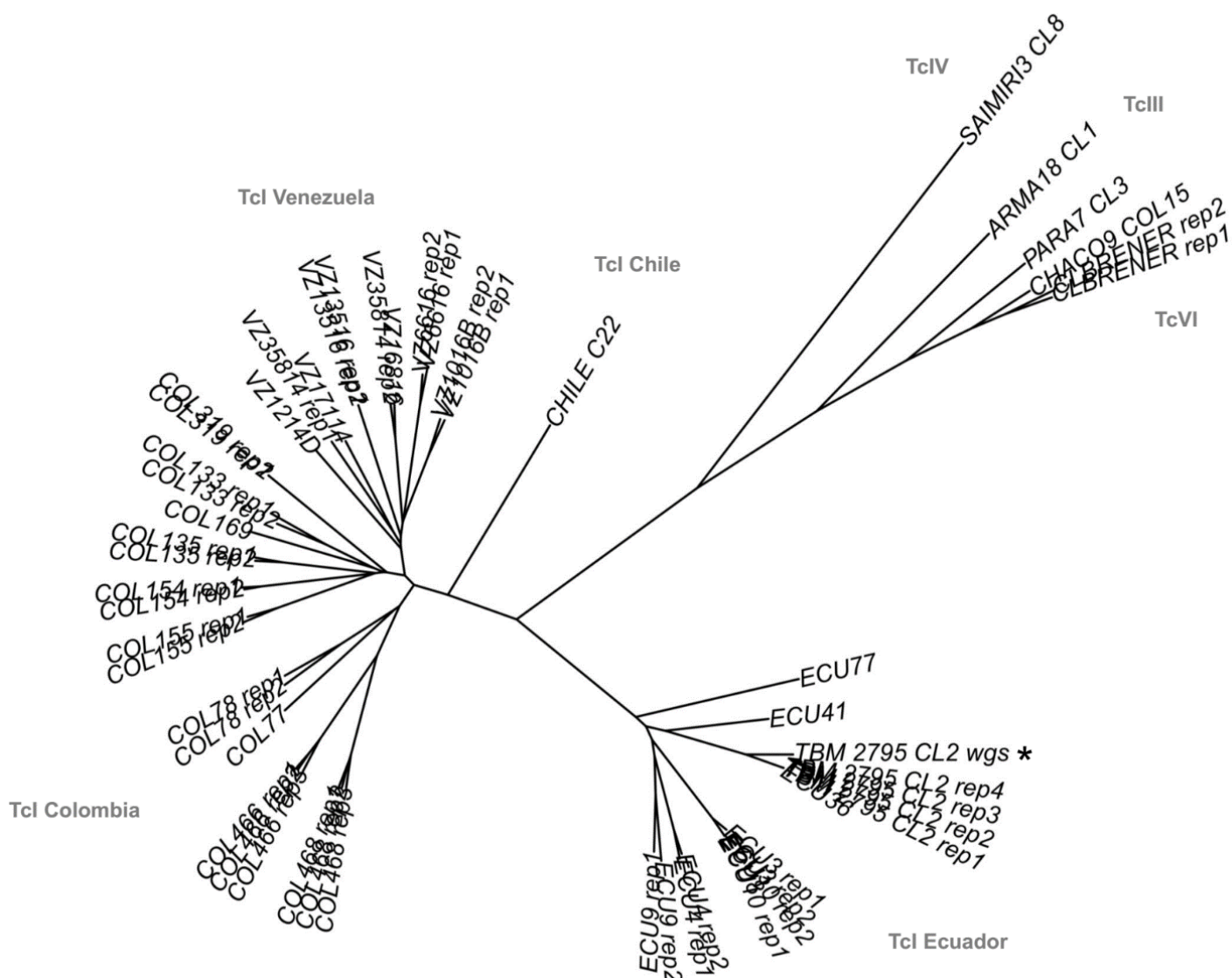


Figure 6 Neighbor-joining relationships among *T. cruzi* I samples and reference clones of other sub-lineages. Genetic distances are based on 556 biallelic SNP sites for which genotypes are called in all individuals. Results indicate high repeatability among most technical replicates (see 'rep1 – 4' suffices) and clearly separate TcI, TcIII, TcIV and TcVI. The tree also contains TBM_2795_CL2_wgs (see asterisk). This control sample was genotyped at the same 556 GLST loci using whole-genome sequencing (Illumina HiSeq) data from Schwabl et al. 2019¹.

Discussion

Principle results

The GLST primer panel design and amplicon sequencing workflow outlined in this study aimed to profile *T. cruzi* genotypes at high resolution directly from infected triatomine intestinal content by simultaneous amplification of 203 genetic target regions that display sequence polymorphism in publicly available WGS reads. Mapped GLST amplicon sequences generated from *T. cruzi* reference clones and from metagenomic intestinal DNA extracts containing a minimum of 3.69 pg/μl *T. cruzi* DNA achieved high target specificity (< 1% off-target mapping) and yield (391 of 403 target SNP sites mapped). Mapping depth variation across target loci was highly repeatable between sequencing runs. 387 SNP sites were identified among *T. cruzi* DTU I samples and 393 SNP sites were identified in non-TcI reference clones. These markers showed low linkage and clearly separated *T. cruzi* individuals within and across DTUs, for the most part also individuals collected at the same or closely separated localities in Colombia, Venezuela, and Ecuador. An increase in pairwise genetic differentiation was observed with increasing geographic distance in analyses within and beyond 150 km.

Cost-effective spatio-genetic analysis

GLST achieved an important resolution benchmark in recovering isolation-by-distance (IBD)⁶⁴ at less than 150 km. These correlations indicate the potential of GLST in spatially explicit epidemiological studies which, for example, aim to identify environmental variables or landscape features that modify IBD²⁷. High spatial sampling effort is typically required by such studies and often limits budget for genotyping tools. GLST appears promising in this context as library preparation costs < 4.00 USD per sample (see cost summary in Supplementary Tbl. 3) and can be completed comfortably in two days. The first-round PCR reaction requires very low primer concentrations (0.125 μM) such that a single GLST panel purchase (0.01 μmol production scale) enables > 100,000 reactions and can be shared by several research groups. Sequencing represents a substantial cost but is highly efficient due to short fragment sizes and few off-target reads. High library complexity also promotes the use of GLST in the role of PhiX, i.e., as a spike-in to enhance read quality in a different sequencing run. Our study easily decontaminated reads from a spiked amplicon pool sharing barcodes with GLST (run 1). Alternatively, i.e, when GLST is sequenced alone (run 2), one Illumina MiSeq run is expected to generate > 70x median genotype depth for 100 samples using Reagent Micro Kit v2 (ca. 1,000 – 1,500 USD, depending on provider; Supplementary Tbl. 3).

GLST in relation to multi-locus microsatellite typing

We consider multi-locus microsatellite typing (MLMT) as the primary alternative for high-resolution *T. cruzi* genotyping directly from metagenomic DNA. MLMT has revolutionized theory on *T. cruzi* ecology and microevolution, for example, on the role of disparate transmission cycles^{65,66}, ecological host-fitting⁶⁷ and ‘cryptic sexuality’⁶⁸ in shaping population genetic structure in TcI. In some cases^{69,70} (but others not^{66,67,71}), the hypervariable, multiallelic nature of microsatellites allows every sample in a dataset to be distinguished

with a different multi-locus genotype (MLG). This depends on panel size and spatial scale but also on local reproductive modes – e.g., sampling from clonal sylvatic vs. non-clonal domestic transmission cycles has correlated with the presence or absence of repeated MLGs⁶⁶. In this study, we found two identical GLST genotypes shared among five samples from southern Ecuador. All other samples appeared unique, including those from Venezuela, where triatomine collection occurred at seven domestic localities within the city of Caracas. The small subset of repeated genotypes found in this study may reflect patchy, transmission cycle-dependent clonal/sexual population structure in southern Ecuador (see Schwabl et al. 2019¹ and Ocaña-Mayorga et al. 2010⁶⁶) but may also represent a weakness in GLST compared to MLMT in tracking individual parasite strains. The use of large MLMT panels, however, is significantly more resource-intensive because each microsatellite marker requires a separate PCR reaction and capillary electrophoresis cannot be highly multiplexed. MLMT data are poorly archivable across studies and may also be less suitable for inter-lineage phylogenetic analyses due to unclear mutational models and artefactual similarity from saturation effects⁷². Although our GLST panel was designed for TcI, its focus on syntenous sequence regions enabled efficient co-amplification of non-TcI DNA. GLST clearly separated TcI samples from all non-TcI reference clones, with highest divergence observed in Saimiri3 cl. 8. Interestingly, large MLMT panels have shown comparatively little differentiation between this sample and TcI, also more generally suggesting that TcIV and TcI represent monophyletic sister clades⁷².

Adjustment and transferability

Considering the great variety of sample types to which studies have successfully applied PCR^{73–77}, we expect that GLST can be applied to metagenomic DNA from many host/vector tissue types, not only from triatomine intestine as shown here. Further tests are required to determine whether low *T. cruzi* DNA concentrations in chronic infections or sparsely infected organs (e.g., liver and heart⁷⁸) are also amenable to GLST. We focused analysis on *T. cruzi* DNA concentrations of at least one picogram per microliter metagenomic DNA (this equates to ca. 30 parasites per microliter in the case of TcI⁷⁹) without heavily investigating options to enhance sensitivity or sensitivity measurement, for example, by additional removal of PCR inhibitors, improved primer purification (e.g., HPLC vs. salt-free), post-PCR probe-hybridization⁸⁰ or barcoding/sequencing of samples with unclear first-round PCR amplicon bands. Even relatively aggressive processing methods may be tolerable given that DNA fragmentation is unlikely to compromise the 120 – 160 nt size range targeted by GLST. Increasing sensitivity by increasing PCR amplification cycles, however, is less advised. PCR error appeared relevant with as little as 30x (+ 7x barcoding) amplification in this study as we observed noise among replicates despite high read-depth and SNP-call overlap between sequencing runs. Rates of error were, however, well within margins expected for methods involving PCR⁸¹. We also note that the exceptional discrepancy between VZ35814 replicates unlikely represents systematic error but barcode contamination with VZ16816. Such error is perhaps less likely if primers are kept in separate vials instead of in the plate format which we have used here.

Wet lab aside, the main objective of this study was to provide a transparent bioinformatic workflow for highly multiplexable primer panel design using freely available softwares and publicly archived WGS reads (e.g., see www.ebi.ac.uk/ena or www.ncbi.nlm.nih.gov/sra). Importantly, we show that knowledge of polymorphic genetic regions in parasite genomes from one small study area (Loja Province, Ecuador) can suffice to guide variant discovery at distant, unassociated sampling sites. Our demonstration using *T. cruzi* should be easily transferable to any other pathogenic species with a published reference genome. Target selection can also be tailored to a variety of objectives. For example, while landscape genetic studies on dispersal often focus on neutral or non-coding sequence variation⁸², experimental (e.g., drug testing) studies may seek to detect single-nucleotide changes in coding regions, perhaps in genes belonging to specific ontology groups or associated with results of high-throughput proteomic screens⁸³. The candidate SNP pool can easily be filtered for such criteria during GLST panel design, e.g., using SnpEff⁸⁴ or BEDTools⁸⁵ and data mining strategies at EuPathDB⁸⁶. Candidate SNP filtering by minor allele frequency (MAF) may also be useful when the target population is closely related to that of the WGS dataset guiding panel design. Placing a minimum threshold on MAF (using VCFtools⁸⁷, etc.), for example, may improve analyses of population structure and genealogy whereas a focus on low-frequency variants may help in tracking individuals or recent gene flow at the landscape scale⁸⁸. It may also be possible to refine panel design towards markers that meet model assumptions in later analysis. Hardy Weinberg Equilibrium (HWE), for example, is a common requirement in demographic modelling^{89–91}, Bayesian clustering⁹², admixture/migration^{93,94} and hybridization tests⁹⁵. Deviation from HWE may occur more frequently in specific genetic regions (e.g., near centromeres⁹⁶), and SNPs in these could be excluded from the target pool. Numerous other filtering options – e.g., based on allele count (to enhance resolution per SNP), distance to insertion-deletions (to improve target alignment), or percent missing information (to avoid poorly mapping regions) – are easily implemented with common analysis tools⁹⁷.

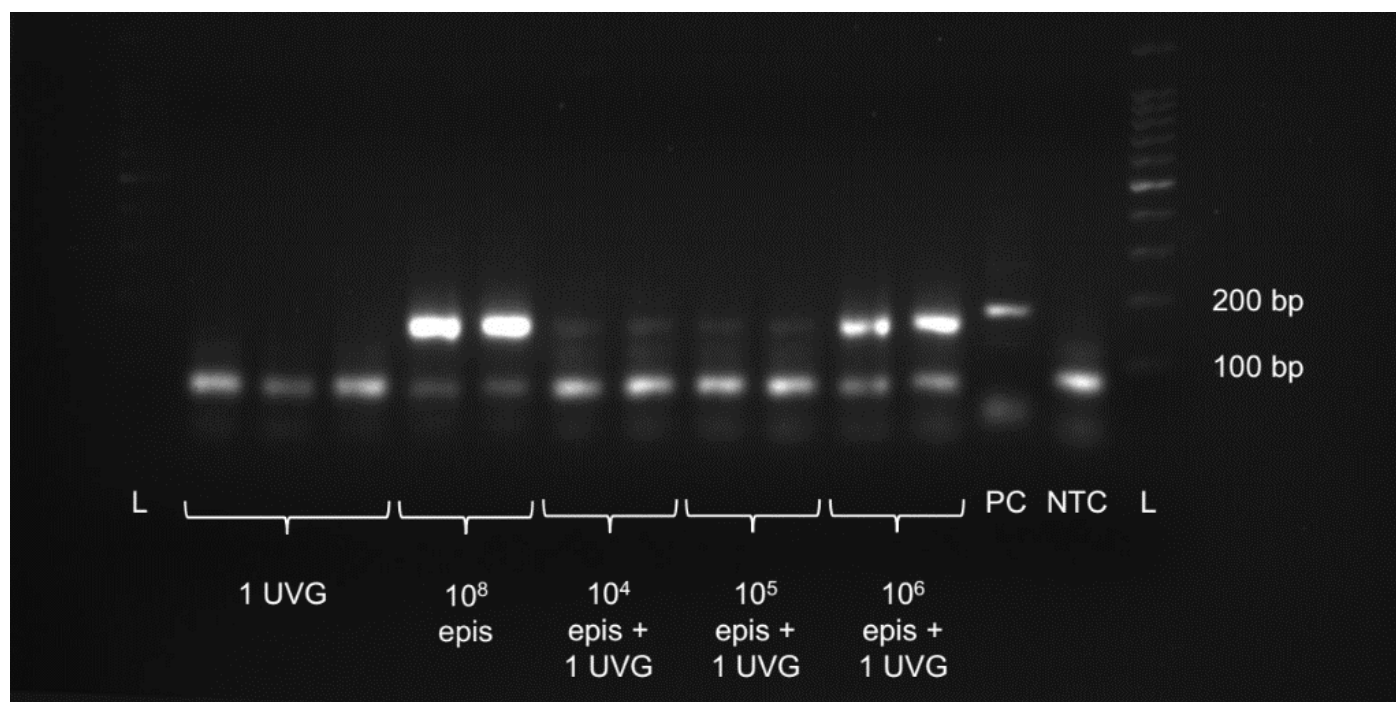
GLST is also highly scalable because increasing panel size does not lead to more laboratory effort or processing time. Sequencing depth requirements and thermodynamic compatibilities among primers are more relevant in limiting panel size. However, it is also possible to divide large GLST panels into two or more PCR multiplexes based on ΔG -based partitioning in MultiPLX⁹⁸. Unintended primer affinities (i.e., polymer formations) can also be removed by gel excision, e.g., as we have done using the PureLink Quick Gel Extraction Kit.

Prospects

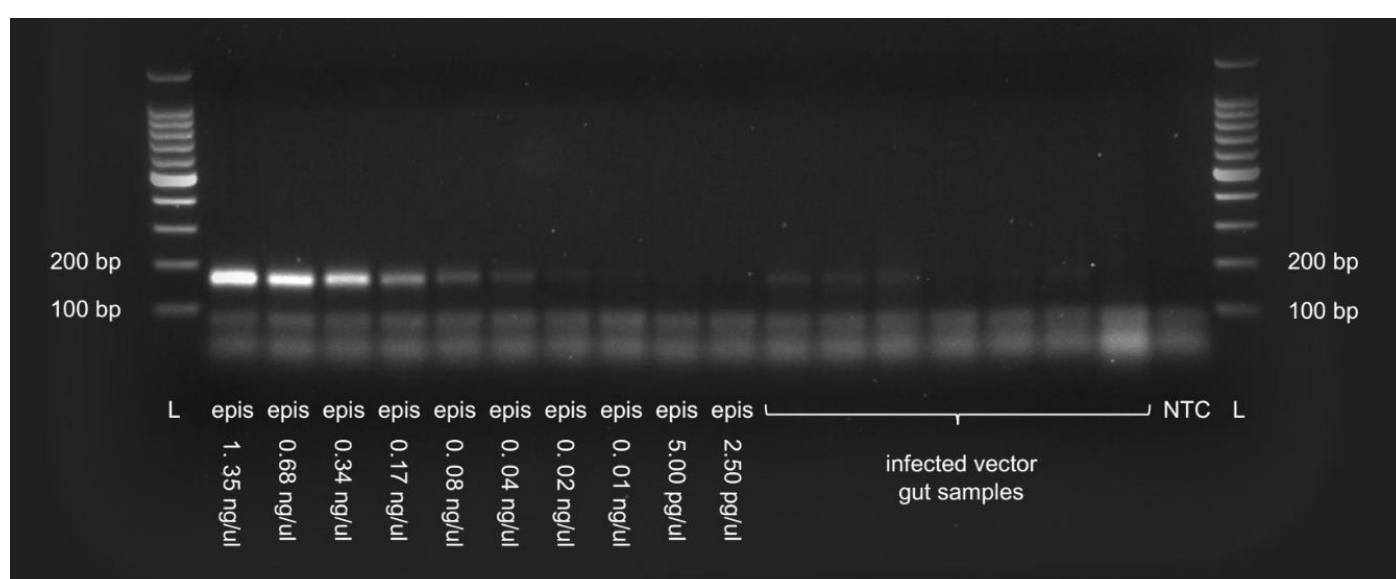
This study sought to provide a framework for various epidemiological research but was restricted in its own ability to make important inferences on *T. cruzi* ecology because only few samples (remainders from different projects) were analyzed. Samples were also aggregated either to domestic or to sylvatic ecotopes (see Supplementary Tbl. 1). More extensive, purposeful sampling could have, for example, helped us explore whether COL468's position deep within the Cordillera Oriental contributes to its strong divergence to samples such as COL135 or COL319, these perhaps more closely related due to lower 'cost-distances'⁹⁹ along the basin range. Fuelling landscape genetic simulators such as CDMetaPOP⁹¹ with high GLST sample sizes is an

especially exciting direction for future research. It would also be interesting, for example, to extend this study's sampling to cover gradients along the perimeter of Caracas and adjacent El Ávila National Park (see Fig. 4b). Sylvatic *P. geniculatus* vector populations appear to be rapidly adapting to habitats within Caracas^{39,100} but parallel changes in the distribution of *T. cruzi* genetic diversity have yet to be tracked. The low cost of GLST also makes it more feasible for studies to simultaneously assess genetic polymorphism in each vector individual from which parasite markers were amplified. Such coupled genotyping would enhance resolution of parasite-vector genetic co-structure and thus, for example, help quantify rates of parasite transmission from domiciliating vectors or determine whether parasite gene flow proxies for (or improves understanding of) dispersal patterns in more slowly evolving vectors or hosts. It would also be interesting to test in how far deep-sequenced GLST libraries could help in detecting (and reconstructing distinct MLGs from) multiclonal *T. cruzi* infections without the use of cloning tools¹⁰¹, e.g., using bioinformatic strategies developed for malaria research^{102–105}. Multiclonality has important implications for public health^{106,107} but its potential prevalence in *T. cruzi* vectors and hosts^{108,101,109} is difficult to describe from cultured cells^{108,110}. Countless other applications are conceivable for GLST. Some research fields, however, will surely be less amenable to the PCR-based approach. Relative amplicon concentrations, for example, appeared to be too stochastic in this study to allow inference of copy number variation or other structural rearrangements based on read-mapping depths. Unintended primer alignment is also likely to occur if PCR targets are located within highly repetitive sequences such as those encoding surface protein families in sub-telomeric regions of the *T. cruzi* genome⁴⁶.

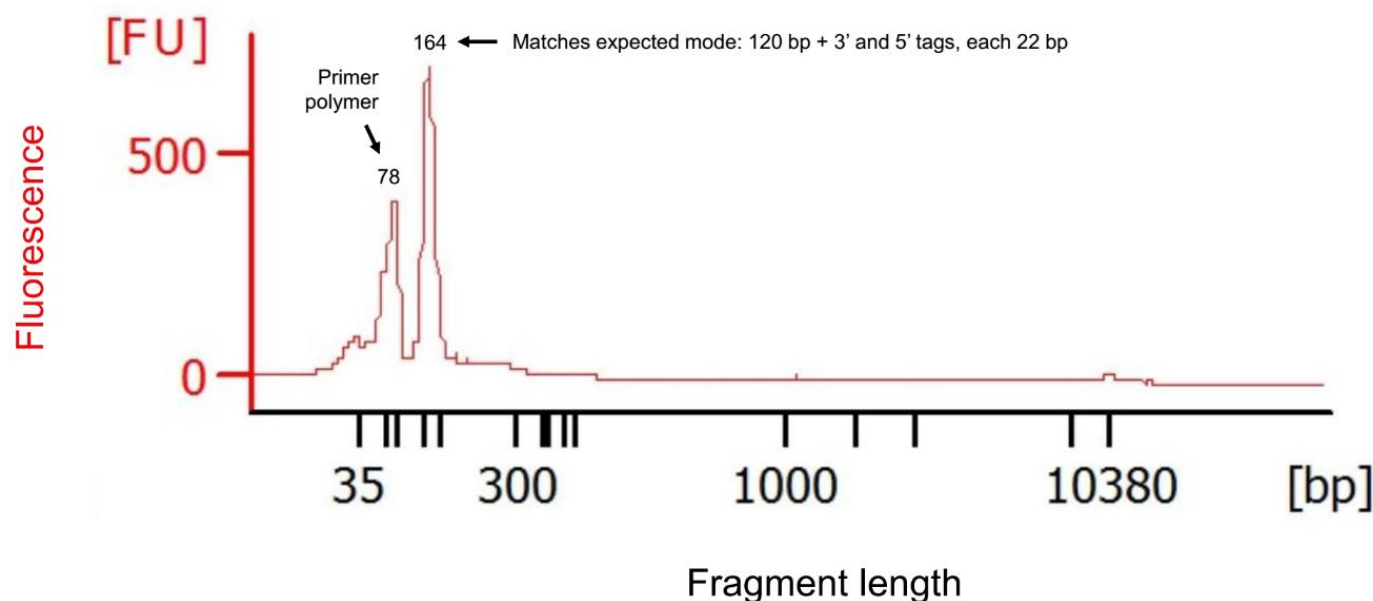
We look forward to seeing GLST approaches in a wide variety of research for which such limitations do not apply. Regarding population and landscape genetic studies, prudent spatial and genetic sampling design is often key to meaningful inference and we hope that the low cost and high flexibility of our pipeline helps researchers achieve all criteria required.



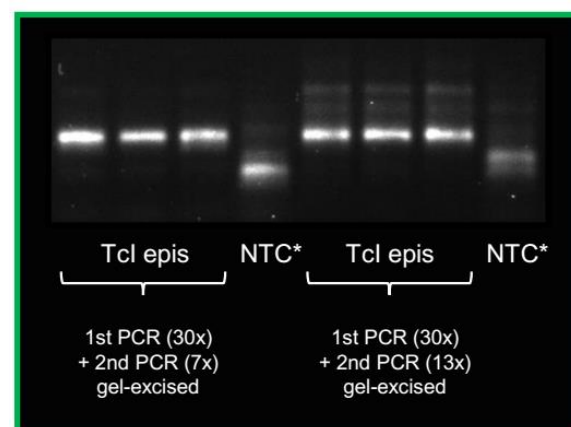
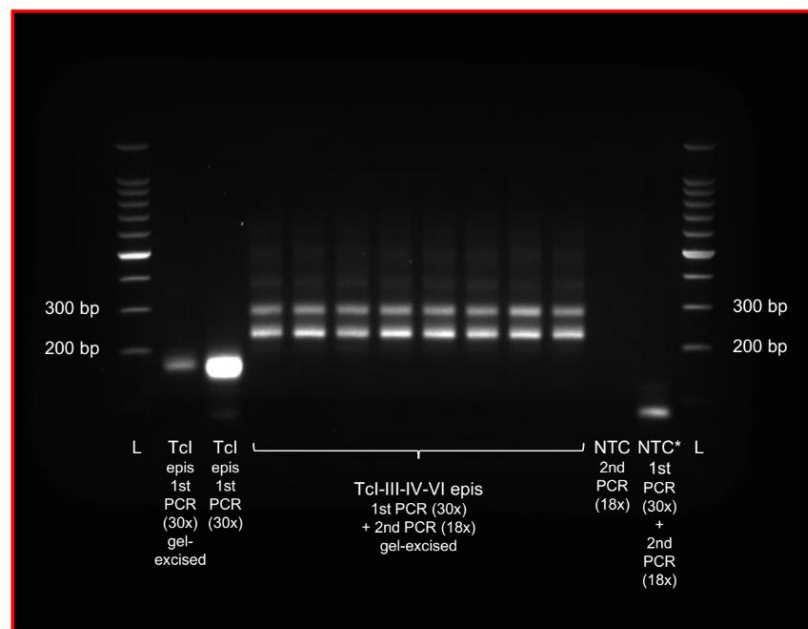
Supplementary Figure 3 Preliminary GLST (multiplex) trials on *T. cruzi* I mock infections. We created mock infections by mixing 10^4 , 10^5 and 10^6 RNA later-preserved TcI-Sylvio epimastigote (epi) cells with uninfected *Rhodnius prolixus* vector gut (UVG). DNA extracted from these mock infections was subjected to the multiplexed, 203-target GLST reaction (using the same cycling conditions as for single-target reactions – see Methods or Supplementary Fig. 2 legend) and products were electrophoresed in 0.8% agarose gel. Fainter banding of GLST products from lower concentration mock infections encouraged follow-up on sensitivity thresholds using additional dilution curves and qPCR. Next to DNA ladder (L) and no-template control (NTC), the gel also contains TcZ primer product from pure TcI epimastigote DNA. TcZ primers provide a highly sensitive positive control (PC) as they target 195 bp satellite DNA repeats that make up ca. 5% of the *T. cruzi* genome.



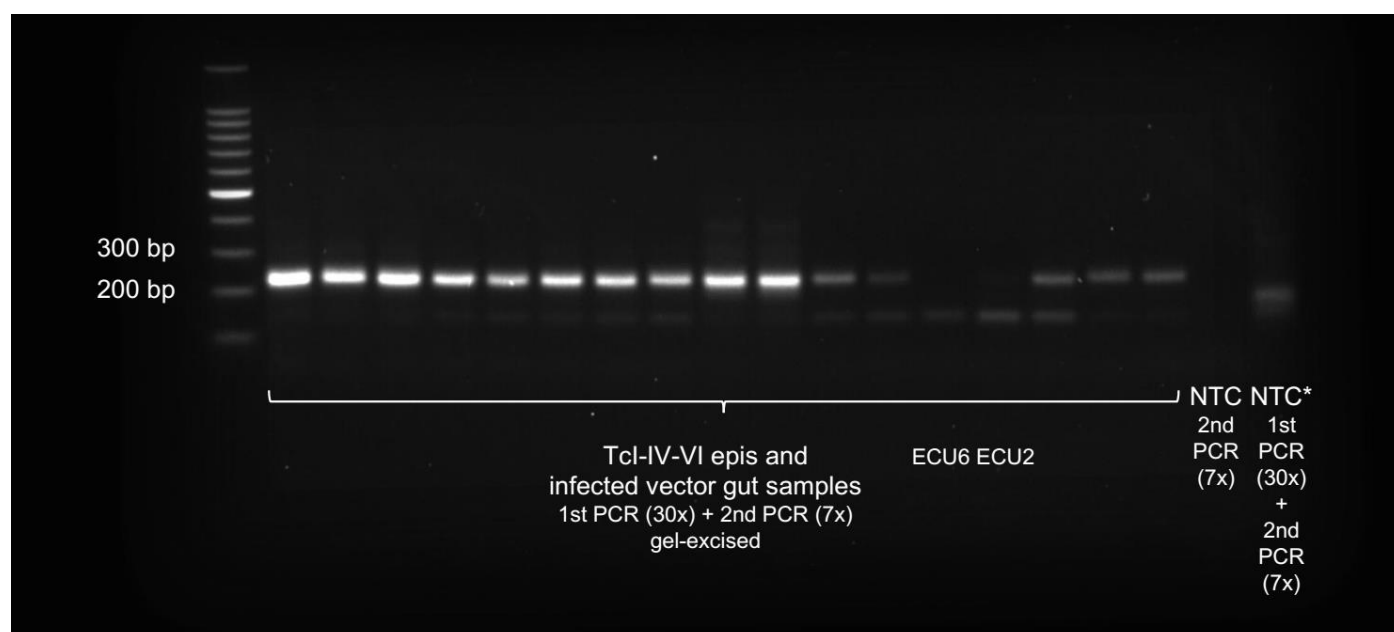
Supplementary Figure 4 *T. cruzi* I DNA dilutions and GLST product visibility in 0.8% agarose gel. The left side shows electrophoresed GLST amplicons generated from 3 μ l pure TcI epimastigote (epi) DNA with concentrations between 1.35 ng/ μ l and 2.50 pg/ μ l (see cycling conditions in Methods or Supplementary Fig. 2 legend). Lanes on the right contain amplicons from seven random metagenomic samples that tested positive for *T. cruzi* satellite DNA (not shown). DNA ladders (L) and no-template control (NTC) are indicated left and right. Poor amplicon visibility occurs at ≤ 60 pg epimastigote DNA input. Gut DNA amplicon visibility is also limited but whether this relates to low *T. cruzi* content or amplification interference is unclear without qPCR.



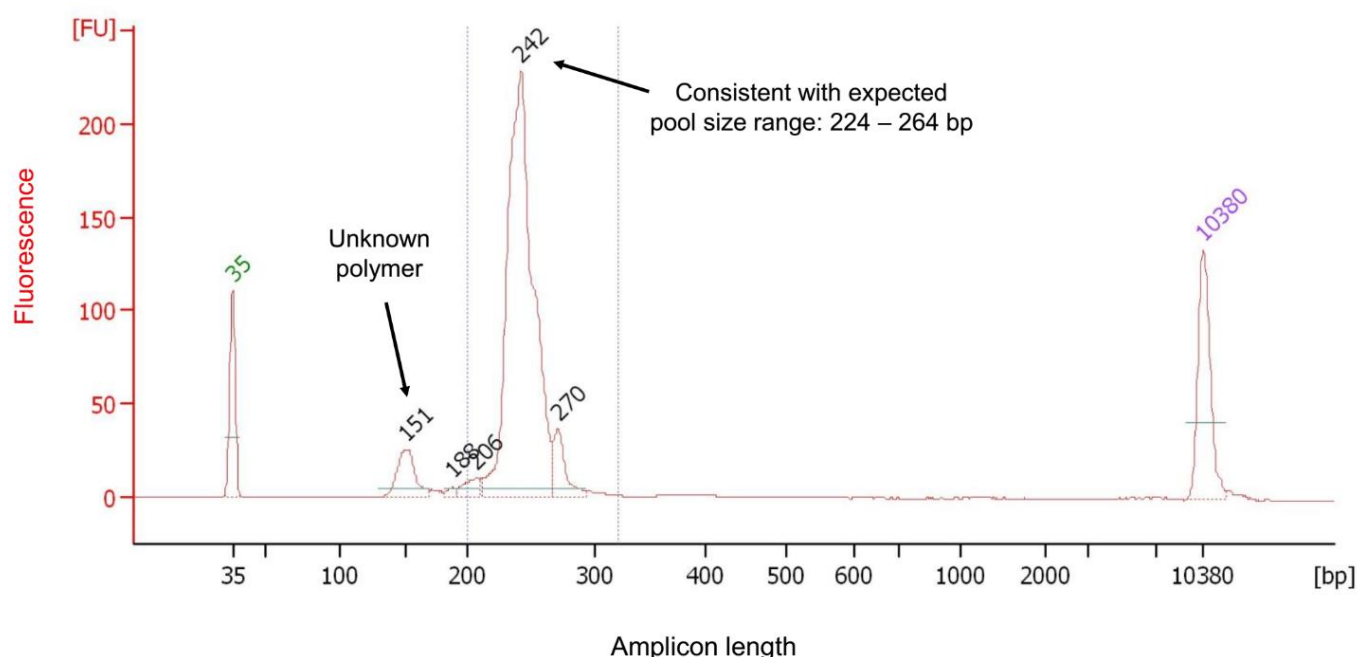
Supplementary Figure 5 First-round (unbarcoded) PCR product size composition measurement using microfluidic electrophoresis. The figure plots fragment sizes (calculated based on migration times relative to those of standards) and fluorescence intensity (FU) of first-round PCR products (see cycling conditions in Methods or Supplementary Fig. 2 legend) measured with the Agilent Bioanalyzer 2100 System. The first peak represents primer polymerization that is removed in subsequent gel excision/re-solubilization steps. The second peak matches expectations for the multi-target GLST product (164 – 204 bp). Special thanks to Craig Lapsley at the Wellcome Centre for Molecular Parasitology in Glasgow for generating this data.



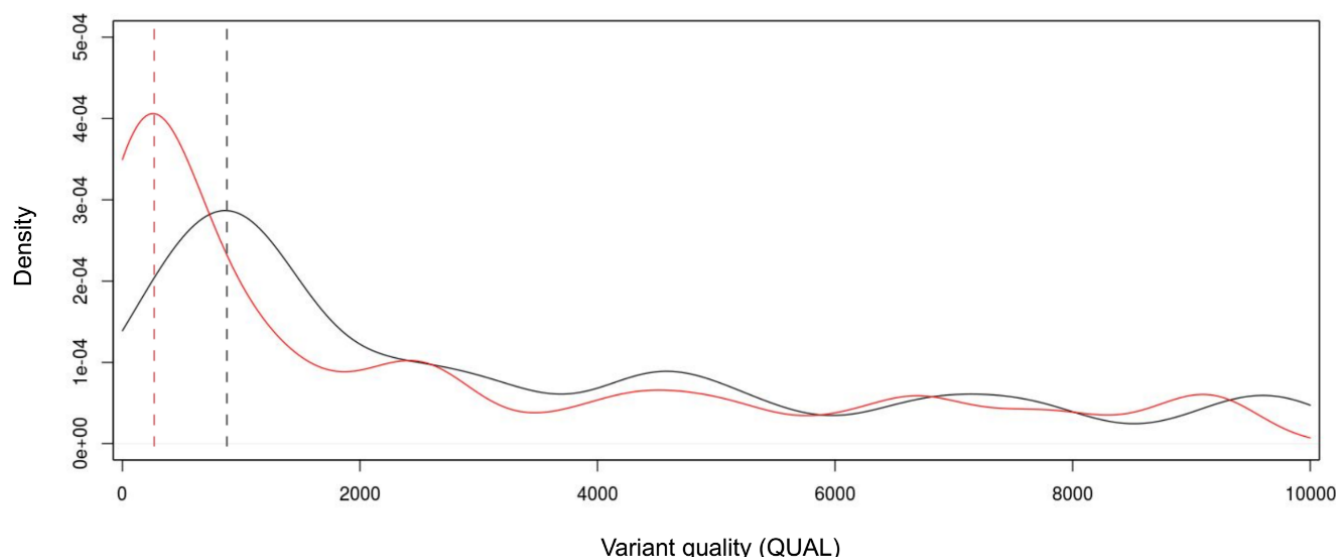
Supplementary Figure 6 Large polymer formation from excessive amplicon barcoding. The second (barcoding) PCR reaction uses an initial incubation step at 98 °C (2 min); 7 amplification cycles at 98 °C (30 s), 60 °C (30 s), and 72 °C (1 min); and a final extension step at 72 °C (3 min). Seven amplification cycles were chosen because unwanted polymers formed at 13 and 18x. The center lanes in the 0.8% agarose gel at left (red border) show electrophoresed GLST products from reference clones after eighteen cycles of barcoding PCR. Large, non-target banding occurs at ≥ 300 bp. Unbarcoded products from Tcl epimastigote (epi) DNA are also shown at left. No template controls from barcoding (NTC) and first-round + barcoding PCR (NTC*) occur next to the DNA ladder (L) on the right side of the gel. The smaller image (green border) to the right shows how unwanted banding becomes less pronounced at 13x and largely disappears at 7x. This 0.8% agarose gel also contains NTC* samples, i.e., negative controls carried through both first and second-round PCR.



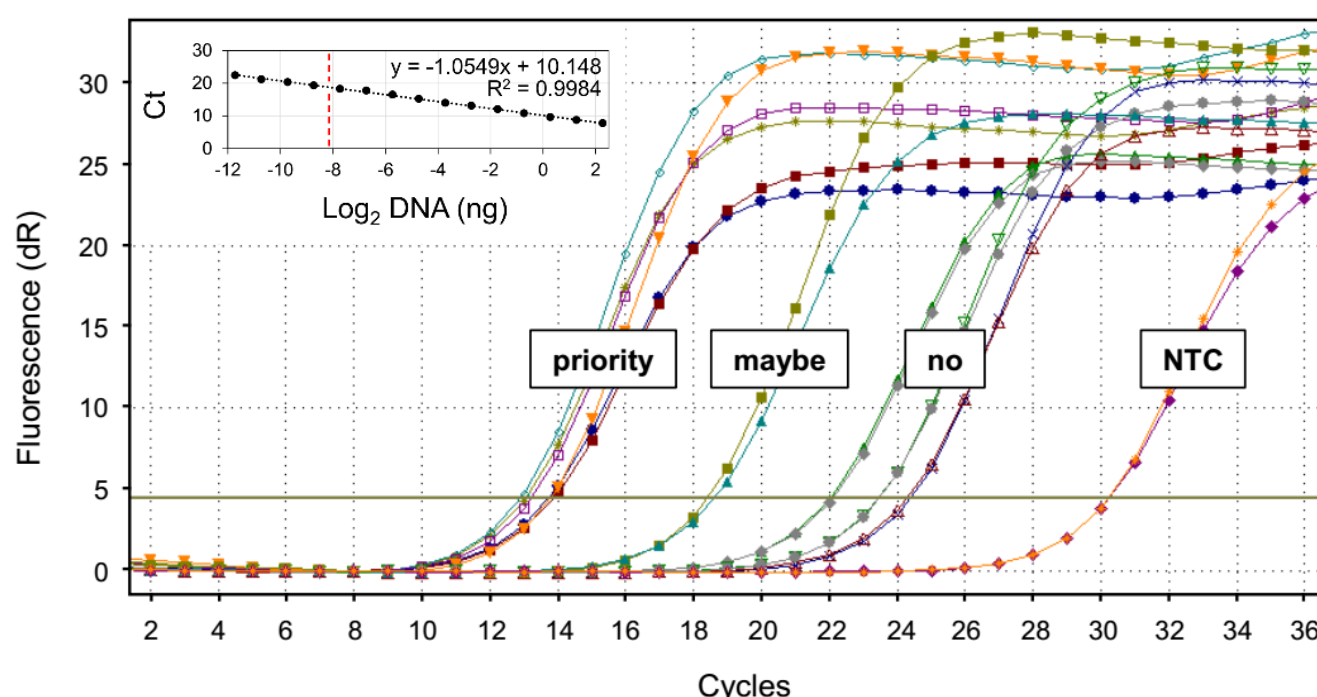
Supplementary Figure 7 Barcoded GLST products ready for final pooling and purification. The 0.8% agarose gel shows a subset of fifteen GLST products from the second-round (barcoding) PCR reaction (see cycling conditions in Methods or Supplementary Fig. 6 legend) prior to equimolar pooling and final gel excision/re-solubilization steps. Products from ECU6 and ECU2 occur in this gel but were not included in the final pool. The gel also contains DNA ladder (L) and no-template controls from barcoding (NTC) and first-round + barcoding PCR (NTC*).



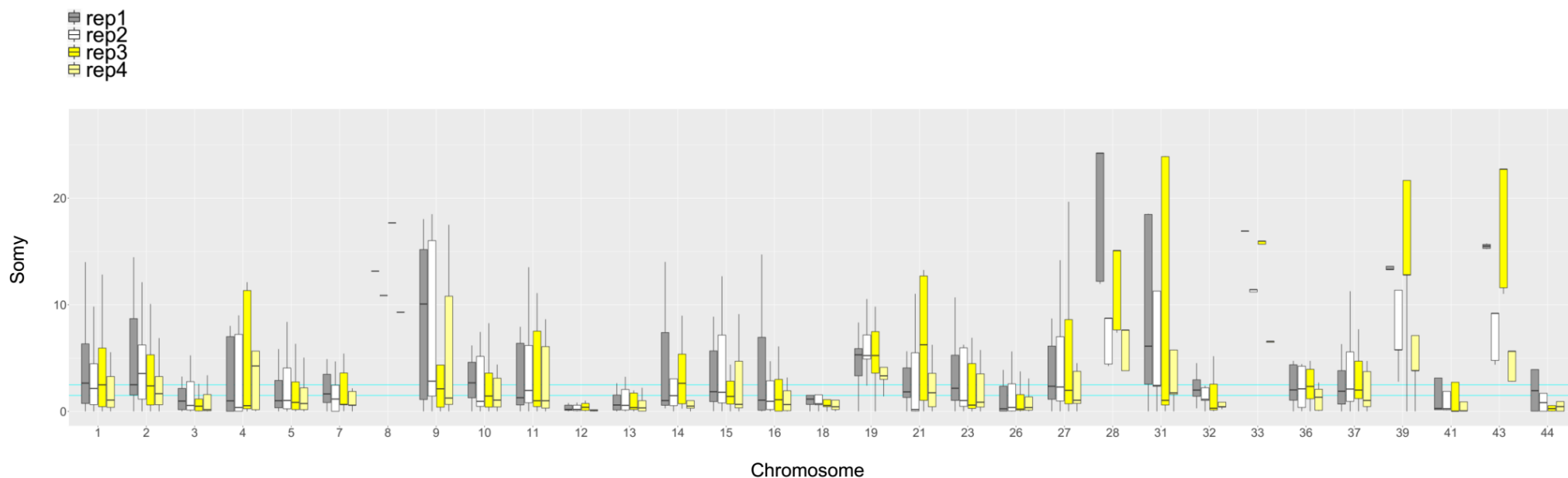
Supplementary Figure 8 Final (barcoded) GLST pool size composition measurement using microfluidic electrophoresis. The figure plots fragment sizes (calculated based on migration times relative to those of standards) and fluorescence intensity (FU) of the final GLST pool measured with the Agilent Bioanalyzer 2100 System. The large peak matches expectations for the multi-target GLST product pool (224 – 264 bp). Left and right peaks labelled in green and purple represent standards of known size. A small non-target peak remaining near 151 bp encourages improvement of prior size selection steps. Special thanks to Julie Galbraith at Glasgow Polyomics for generating this data.



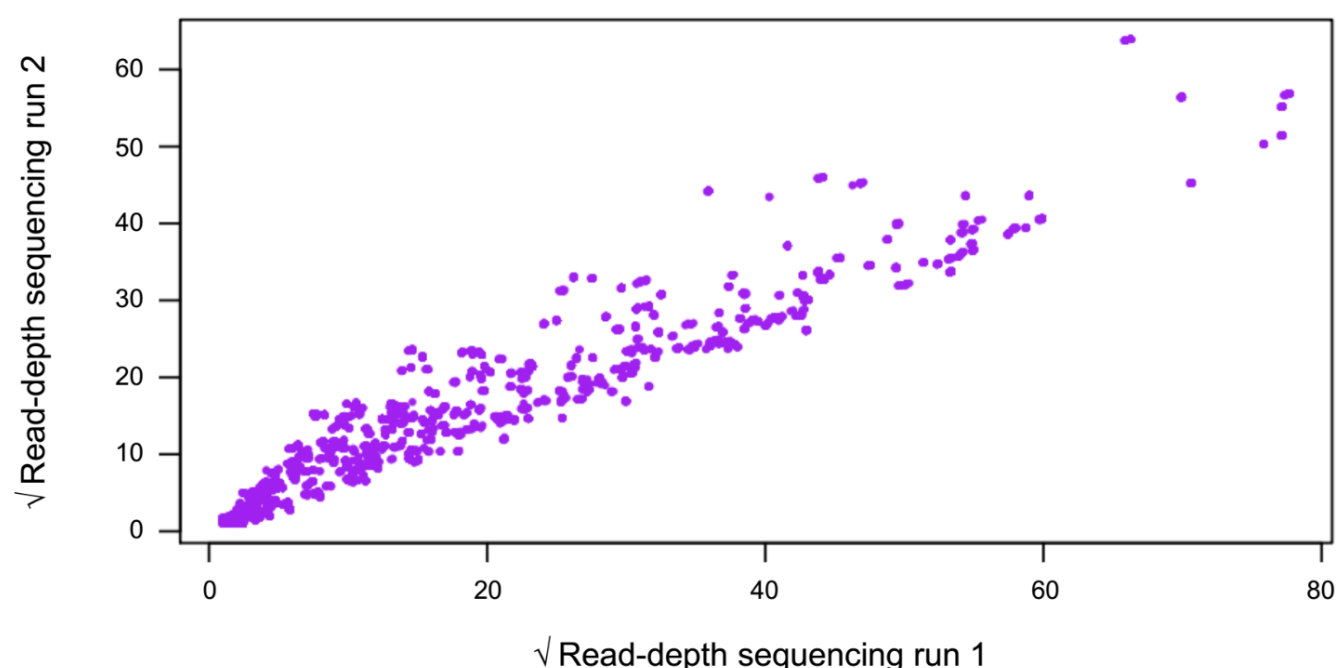
Supplementary Figure 9 Quality scores at previously identified vs. unidentified variant sites. The GLST primer panel was designed based on single-nucleotide polymorphisms (SNPs) in Ecuadorian TcI clones. It was applied, however, to samples from distant geographic locations as well as to non-TcI clones. Additional, previously unidentified SNP sites (PU) were thus expected to be found but we needed to distinguish true PU from PCR and sequencing error. We reasoned that quality statistics (e.g., mapping quality, strand bias, minor allele frequency, etc. – see Methods) at previously identified SNP sites (PI) could help calibrate quality filters applied to the wider dataset. This strategy finds support in the above density plot of QUAL scores computed by Genome Analysis Toolkit⁴². The plot suggests that, prior to variant filtration, lower QUAL scores occur more often at PU (red) than at PI (black). We thus imposed the most stringent filtering criteria possible without losing PI.



Supplementary Figure 10 GLST sample selection and sensitivity estimation via qPCR. We used *T. cruzi* satellite DNA qPCR to identify vector gut samples with *T. cruzi* DNA quantities within ranges successfully visualized in GLST reactions using epimastigote DNA (Supplementary Fig. 4). The qPCR reaction used an initial incubation step at 95 °C (10 min) and 40 amplification cycles at 95 °C (15 s), 55 °C (15 s), and 72 °C (15 s). The plot shows baseline-corrected fluorescence (dR) for seven sample duplicates. Following the regression equation from the standard curve (see inset), the three samples with highest cycle thresholds (Ct values) in this example represent gut extracts with 0.05 to 0.14 ng/μl *T. cruzi* DNA. Such samples with *T. cruzi* DNA concentrations above 0.01 ng/μl were prioritized for GLST and none failed in library construction. ECU36, with a mean Ct value of 18.68 in the plot, was also successfully sequenced. A Ct value of 18.68 represents 3.69 pg/μl *T. cruzi* DNA. Not all samples with concentrations at single-digit picogram levels (per μl) were successful and we did not troubleshoot those with substantially lower concentrations based on qPCR.



Supplementary Figure 11 Target coverage in control replicates confirms expectations that the GLST panel applied in this study is unreliable for copy number estimation. We adapted methods from Schwabl et al. 2019¹ to derive somy estimates for each base position within GLST amplicons. Briefly, we calculated median-read-depth of all target bases for each chromosome. We let the median of these chromosomal medians (the 'inter-chromosomal median') represent expectations for the disomic state, estimating copy number per base position by dividing each position's read-depth by the inter-chromosomal median and multiplying by two. Boxplots show median and interquartile ranges of these site-wise somy estimates for each chromosome in TBM_2975_CL2 control replicates. TBM_2975_CL2 did not show chromosomal amplifications in whole-genome analysis¹. Not unexpectedly for a PCR-based method, somy values estimated from GLST read-depths differ substantially among replicates and are unrealistically high/low on many chromosomes. Estimates on chromosomes with few GLST targets appear especially unreliable – e.g., see chromosomes 8, 28, 33, 39 and 43. These chromosomes contain ≤ 2 GLST targets each. The horizontal lines cyan lines mark $y = 1.5$ and $y = 2.5$.



Supplementary Figure 12 Similar read-depth distribution between separate sequencing runs. We sequenced the same GLST pool in two separate Illumina MiSeq runs. Run 1 involved GLST as a spike to a collaborator's 16S amplicon library, whereby GLST reads were subsequently decontaminated from (barcode-sharing) 16S reads by alignment to the TcI-Sylvio reference genome. Run 2 was dedicated solely to GLST, i.e., no non-GLST libraries were simultaneously sequenced on the flow cell. The plot shows that run 1 and run 2 read-depths at each GLST base position (purple points) are highly correlated (Pearson's $r = 0.93$, $p < 0.001$), and that run 1 had higher sequencing output than run 2. Read-depth values are square-root transformed and represent control sample TBM_2975_CL2_rep1.

Supplementary Table 1 Details on *T. cruzi*-infected metagenomic triatomine gut samples from Colombia (COL), Venezuela (VZ) and Ecuador (ECU). Abbreviations: Dep. (Department); Met. Caracas (Metropolitan District of Caracas); EPSG (European Petroleum Survey Group coordinate system); reps. (technical replicates).

ID	Vector species	Region	Municipality / community	x (EPSG 3786)	y (EPSG 3786)	Ecotope	Year	Reps.
COL77	<i>Rhodnius pallescens</i>	Santander Dep.	Lebrija	-8141577.9370	790936.6092	Sylvatic	2015	1
COL78	<i>Rhodnius sp.</i>	Santander Dep.	Lebrija	-8141577.9370	790936.6092	Sylvatic	2015	2
COL133	<i>Rhodnius prolixus</i>	Casanare Dep.	Paz de Ariporo	-7993997.4220	653950.4247	Domestic	2016	2
COL135	<i>Rhodnius prolixus</i>	Casanare Dep.	Paz de Ariporo	-7993997.4220	653950.4247	Domestic	2016	2
COL154	<i>Rhodnius prolixus</i>	Casanare Dep.	Tamara	-8024081.7980	648298.0468	Domestic	2016	2
COL155	<i>Rhodnius prolixus</i>	Casanare Dep.	Tamara	-8024081.7980	648298.0468	Domestic	2016	2
COL169	<i>Rhodnius prolixus</i>	Casanare Dep.	Pore	-8005271.3760	636869.6421	Domestic	2016	1
COL253	<i>Panstrongylus geniculatus</i>	Casanare Dep.	Paz de Ariporo	-7993997.4220	653950.4247	Domestic	2016	1
COL319	<i>Rhodnius prolixus</i>	Arauca Dep.	Fortul	-7980623.1040	755354.1935	Domestic	2016	2
COL466	<i>Panstrongylus geniculatus</i>	Boyacá Dep.	Soata	-8083880.0490	704231.6027	Unknown	2017	3
COL468	<i>Panstrongylus geniculatus</i>	Boyacá Dep.	Soata	-8083880.0490	704231.6027	Unknown	2017	3
ECU3	<i>Rhodnius ecuadoriensis</i>	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	2
ECU4	<i>Rhodnius ecuadoriensis</i>	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	2
ECU8	<i>Rhodnius ecuadoriensis</i>	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	1
ECU9	<i>Rhodnius ecuadoriensis</i>	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	2
ECU10	<i>Rhodnius ecuadoriensis</i>	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	2
ECU36	<i>Rhodnius ecuadoriensis</i>	Loja Province	Galápagos	-8832711.9860	-483957.8804	Sylvatic	2009	1
ECU41	<i>Rhodnius ecuadoriensis</i>	Loja Province	Guineo	-8899431.9060	-466731.6546	Sylvatic	2009	1
ECU77	<i>Rhodnius ecuadoriensis</i>	Loja Province	Jacapo	-8830688.2360	-485500.9341	Sylvatic	2008	1
TBM_2795_CL2	<i>Panstrongylus chinai</i>	Loja Province	Bella Maria	-8852271.1950	-466705.6350	Domestic	2009	4
VZ1016B	<i>Panstrongylus geniculatus</i>	Met. Caracas	Libertador	-7447967.9080	1167084.6630	Domestic	2016	2
VZ13516	<i>Panstrongylus geniculatus</i>	Met. Caracas	Libertador	-7441110.8420	1169154.1140	Domestic	2016	2
VZ35814	<i>Panstrongylus geniculatus</i>	Met. Caracas	Libertador	-7450655.1580	1165756.5490	Domestic	2014	2
VZ6616	<i>Panstrongylus geniculatus</i>	Met. Caracas	Sucre	-7426686.3980	1163934.1740	Domestic	2016	2
VZ1214D	<i>Panstrongylus geniculatus</i>	Met. Caracas	Sucre	-7427396.8230	1166961.1250	Domestic	2014	1
VZ16816	<i>Panstrongylus geniculatus</i>	Met. Caracas	Sucre	-7427026.2100	1162328.0720	Domestic	2016	1
VZ17114	<i>Panstrongylus geniculatus</i>	Met. Caracas	Sucre	-7426501.1470	1162853.1350	Domestic	2014	1

Supplementary Table 2 GLST primer sequences. The 3' end of each first-round PCR primer is target-specific. The 5' end of each forward primer contains CS1. The 5' end of each reverse primer contains CS2. These sequencing primer binding sites are shown in pink. In subsequent barcoding PCR, the reverse primer consists of 5'-CAAGCAGAAGACGGCATAACGAGAT*X*TACGGTAGCAGAGACTTGGTCT-3', where *X* is a unique 10 nt barcode used to label each sample's sequence reads. The reverse barcoding primer also contains CS2. The forward barcoding primer (5'-AATGATACGGCGACCACCGAGATCTACACTGACGACATGGTTCTA-3') contains CS1 and is the same for all samples.

ID	Target region	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
TC_LOJ_1	chr16:130780-130919	ACACTGACGACATGGTTCTACATGCCAATAACGGTCAAAGTAAACG	TACGGTAGCAGAGACTTGGTCTGCACACGAAGGTACACTCACTTCC
TC_LOJ_2	chr10:534441-534583	ACACTGACGACATGGTTCTACAGAGTTGTGGCATCCTTGTCTTG	TACGGTAGCAGAGACTTGGTCTAAACGCCTTCACCTTACTCAGACA
TC_LOJ_4	chr11:368075-368194	ACACTGACGACATGGTTCTACAGGAGGTGAAACGGATGGTAAAGA	TACGGTAGCAGAGACTTGGTCTTGCGAAGAAGAAGATCAAACCTCTCTC
TC_LOJ_5	chr1:2082456-2082586	ACACTGACGACATGGTTCTACAGCTCAAGGGCTGAAATAGACACA	TACGGTAGCAGAGACTTGGTCTCGTTTAGGCTGGAAAGATGGAAGT
TC_LOJ_6	chr12:1011748-1011869	ACACTGACGACATGGTTCTACCACTCTATCGTCTACGCATCCTC	TACGGTAGCAGAGACTTGGTCTATCATCTTGAGACACATGCCTTGC
TC_LOJ_8	chr5:515822-515951	ACACTGACGACATGGTTCTACAATGGAGATGGAGGATATGAAGCA	TACGGTAGCAGAGACTTGGTCTTTTAGACCTCATGTTCCCGTGTC
TC_LOJ_9	chr1:163164-163296	ACACTGACGACATGGTTCTACACGCTGAGTATCAATTTAAGCGTAGCA	TACGGTAGCAGAGACTTGGTCTACCCATATCCGTCATCCCTATTGT
TC_LOJ_10	chr1:1104374-1104501	ACACTGACGACATGGTTCTACATGCCCTTCACATTTATCCCAAGTA	TACGGTAGCAGAGACTTGGTCTAAATAGCATGGAACCTCAGCCAGAA
TC_LOJ_11	chr5:995176-995297	ACACTGACGACATGGTTCTACAGCAACTCCACAAACGACTCAGAAC	TACGGTAGCAGAGACTTGGTCTGATGCTGCCATTTCTGCTTTACTC
TC_LOJ_12	chr14:833083-833213	ACACTGACGACATGGTTCTACCTTGTTGCTAAGTGTCGGTGTGTC	TACGGTAGCAGAGACTTGGTCTGCCTTTATATTGATCGGCTCCTCT
TC_LOJ_13	chr23:560603-560743	ACACTGACGACATGGTTCTACAGTCTTTGATTCTCGTCCGTACCTT	TACGGTAGCAGAGACTTGGTCTTGCATCTTCTACTTTCTCGGAAGC
TC_LOJ_14	chr19:763581-763703	ACACTGACGACATGGTTCTACAAGATACAAGAGCACGGTACAAAGGA	TACGGTAGCAGAGACTTGGTCTGTGAAGAGGGATGGATCAACATTC
TC_LOJ_15	chr4:1431898-1432017	ACACTGACGACATGGTTCTACAGGACTATGCTCAAGACGGGATCT	TACGGTAGCAGAGACTTGGTCTCATCAAGTGGACACAACAGCAACT
TC_LOJ_16	chr16:1168122-1168248	ACACTGACGACATGGTTCTACATACAAACATCAACGCAGAACATGC	TACGGTAGCAGAGACTTGGTCTCACACATCCCGTAACCTCAATGGTA
TC_LOJ_19	chr43:177414-177556	ACACTGACGACATGGTTCTACACAGTCTCCAGTTCTCCAAGTGAT	TACGGTAGCAGAGACTTGGTCTGAGATTGTTCTCTGTGCCAACG
TC_LOJ_20	chr26:294140-294261	ACACTGACGACATGGTTCTACAGCAAGAACGGGTGTACCTTCTA	TACGGTAGCAGAGACTTGGTCTTGTGTGCGAGGGAATTGATTACTGC
TC_LOJ_23	chr18:690694-690813	ACACTGACGACATGGTTCTACAAAGAACTTCGGGTAGCGACAAC	TACGGTAGCAGAGACTTGGTCTCACCATTCTGCTAGACCACATCC
TC_LOJ_24	chr1:1993894-1994026	ACACTGACGACATGGTTCTACATTCTACACACTCCGCCTTACGTCT	TACGGTAGCAGAGACTTGGTCTGTCTGCAACGACACATAGATTGGA
TC_LOJ_25	chr36:470603-470728	ACACTGACGACATGGTTCTACAGTGCGCTCAGAAGCATGATCGTAAT	TACGGTAGCAGAGACTTGGTCTACCCTTGAGTCTTCGAGTCCTC
TC_LOJ_26	chr13:433737-433859	ACACTGACGACATGGTTCTACAATGGTGATGATGAGGTTAAGCA	TACGGTAGCAGAGACTTGGTCTACGTCCAATACACACAAACACACAG
TC_LOJ_27	chr24:269253-269379	ACACTGACGACATGGTTCTACAGGCGATAAGGAAGAATGGAGAGAA	TACGGTAGCAGAGACTTGGTCTGTCATGTGCTTACGAGAGCCGTAG
TC_LOJ_28	chr27:389665-389794	ACACTGACGACATGGTTCTACACCACTTCACCATTTGTCTGGTATTC	TACGGTAGCAGAGACTTGGTCTTTTAAGATGGCCGCATACAGTGAG
TC_LOJ_29	chr36:451747-451871	ACACTGACGACATGGTTCTACAGTGTGTTTGAGATTGGGCCTGTAT	TACGGTAGCAGAGACTTGGTCTCACATCAAGTACCTCCGTGTACGA
TC_LOJ_30	chr7:1140939-1141071	ACACTGACGACATGGTTCTACAGTTGATCGTCTTTCTTCTTGACC	TACGGTAGCAGAGACTTGGTCTAAATGTTCTGCGTACACCAAGTC
TC_LOJ_32	chr2:120852-120972	ACACTGACGACATGGTTCTACAAATGATGTACTGCCTGAACTGGAA	TACGGTAGCAGAGACTTGGTCTGTTCTCCGCCGTATTCTCTCTAC
TC_LOJ_34	chr16:170448-170597	ACACTGACGACATGGTTCTACAGGAAGAAGGCAGACTAAACAGGATG	TACGGTAGCAGAGACTTGGTCTAGCTTGTCACTGCTCACAGAGTTG

Supplementary Table 2 (continued)

TC_LOJ_35	chr26:125032-125153	ACACTGACGACATGGTTCTACAGTACGCTACACTGCGAGAGGAATG	TACGGTAGCAGAGACTTGGTCTGCACAACCTGAGATTATAGCCAACCTCC
TC_LOJ_36	chr5:1012765-1012911	ACACTGACGACATGGTTCTACATCCGTCCTGTTGTCTTCTCAATA	TACGGTAGCAGAGACTTGGTCTTGAGCAAAGTGCTTATTCTTCAGC
TC_LOJ_37	chr1:2889409-2889535	ACACTGACGACATGGTTCTACACAGAGTTCCACGGATAAGTCGTCA	TACGGTAGCAGAGACTTGGTCTACACACTTCCAGATCACTACGAAGC
TC_LOJ_38	chr21:465093-465213	ACACTGACGACATGGTTCTACATGGTTGTAGTCCGTGATCTCTGGT	TACGGTAGCAGAGACTTGGTCTATAACTGGTTCGGGAAGGAAGAAA
TC_LOJ_39	chr1:1160205-1160334	ACACTGACGACATGGTTCTACACGTCACATTTGTACTGCGAGAGG	TACGGTAGCAGAGACTTGGTCTCCCTTACTTGTCTCCGACTCATTCT
TC_LOJ_40	chr7:1138368-1138496	ACACTGACGACATGGTTCTACAGTCCAAGCCGTTGTCTCTCAATAC	TACGGTAGCAGAGACTTGGTCTTGTTTCGTTGTGGTGGAATGTGTAG
TC_LOJ_41	chr1:2693345-2693466	ACACTGACGACATGGTTCTACATGGCTGGTGCAATGTACTCATATC	TACGGTAGCAGAGACTTGGTCTTAACAAGGTGCCATTGCGTATC
TC_LOJ_42	chr10:1016129-1016269	ACACTGACGACATGGTTCTACATACGACTCCCTTTCCACATACGAC	TACGGTAGCAGAGACTTGGTCTATATTGAGCCGAAACACGAAGTACA
TC_LOJ_43	chr1:1956698-1956821	ACACTGACGACATGGTTCTACAGCTCTCATGGGTGGTAGAAGCTAA	TACGGTAGCAGAGACTTGGTCTCCCACTGTCATTATTCAAAGTCTC
TC_LOJ_44	chr3:173883-174019	ACACTGACGACATGGTTCTACAGTCATCATTTCTCGGAAACAAAGTAGG	TACGGTAGCAGAGACTTGGTCTGTGTCCATCAGCTCTACAATGCAC
TC_LOJ_45	chr3:174152-174277	ACACTGACGACATGGTTCTACAGTACGCCACACGACAGTTTCAGTT	TACGGTAGCAGAGACTTGGTCTTGAGTAGTTGTGCCCTTCGATGTA
TC_LOJ_46	chr1:1833807-1833948	ACACTGACGACATGGTTCTACATTTCGTGTTCATTAGCAGCAGCAAC	TACGGTAGCAGAGACTTGGTCTGACGGTAAATTCTGCGTACACTGC
TC_LOJ_47	chr14:844524-844671	ACACTGACGACATGGTTCTACAGCAATTACGGAGTTCACAGATG	TACGGTAGCAGAGACTTGGTCTAGGAGTACCACAGAAGTCAGAGC
TC_LOJ_48	chr3:1058072-1058196	ACACTGACGACATGGTTCTACAGATAGCACAAACAAGCCAAATGGT	TACGGTAGCAGAGACTTGGTCTGAAAGATACGCCCTTCCAATCATCA
TC_LOJ_51	chr12:596775-596914	ACACTGACGACATGGTTCTACAGATTGACATTACGGCGATTTCAGAG	TACGGTAGCAGAGACTTGGTCTTGTTGGATCTTCTGCCATGATATTG
TC_LOJ_52	chr31:428464-428593	ACACTGACGACATGGTTCTACACCCTCATGGAGACATCTACGAATCT	TACGGTAGCAGAGACTTGGTCTTGAAGAACGAGTGTGCAGGTCATA
TC_LOJ_54	chr2:925727-925855	ACACTGACGACATGGTTCTACAAATGCTAGAGGGCGATAATGAAGAC	TACGGTAGCAGAGACTTGGTCTACCTTTGCCTTGTGTTTACTGCTG
TC_LOJ_55	chr12:306151-306272	ACACTGACGACATGGTTCTACATGGGTCTGCTTGACTGGTTTCTTA	TACGGTAGCAGAGACTTGGTCTGTACGGCGACTCACTTCCAAATAC
TC_LOJ_56	chr21:341510-341636	ACACTGACGACATGGTTCTACATACTCCTCTGCATTACCTCCTG	TACGGTAGCAGAGACTTGGTCTGGTTGGTATAACCGAAGGAAATATGG
TC_LOJ_57	chr37:454539-454662	ACACTGACGACATGGTTCTACAGTACGTGAAACGCCCTGACTTTAC	TACGGTAGCAGAGACTTGGTCTTGATGAACCTCCTTGTAGATGTTG
TC_LOJ_58	chr15:395493-395614	ACACTGACGACATGGTTCTACACTTTGTGACCACCTCCTTGTTATTG	TACGGTAGCAGAGACTTGGTCTAGGTATTTGGCATGTTTGATCTGC
TC_LOJ_59	chr2:856618-856737	ACACTGACGACATGGTTCTACAGCCCGGTTCAACACTTTAGTAGAAA	TACGGTAGCAGAGACTTGGTCTCACCAACACAGCTACGACAACAAC
TC_LOJ_60	chr26:139346-139478	ACACTGACGACATGGTTCTACAGATTATGGTGGTGGTTTCAACACG	TACGGTAGCAGAGACTTGGTCTAAAGTGAATGGCAAATCCTAAGACG
TC_LOJ_61	chr1:1992854-1992995	ACACTGACGACATGGTTCTACATCTGTTGAGGATGACCGAACACT	TACGGTAGCAGAGACTTGGTCTGAGAAATATCGCCGCACCTTCTAC
TC_LOJ_62	chr1:305886-306012	ACACTGACGACATGGTTCTACATACTCAGGCGTAGAAACAGGCTCA	TACGGTAGCAGAGACTTGGTCTTACCTCCGCTTATCAATGTTGTCC
TC_LOJ_63	chr26:303994-304113	ACACTGACGACATGGTTCTACACATGACAAGCATAAATACAGCGAGAG	TACGGTAGCAGAGACTTGGTCTGAAGGTACAAGCAAGGAGCCATCT
TC_LOJ_64	chr14:889253-889389	ACACTGACGACATGGTTCTACACTTCCCAGACTCATCTTTCTGCTG	TACGGTAGCAGAGACTTGGTCTATTCCCAGACTACTTTGGCATGATT
TC_LOJ_67	chr10:143080-143202	ACACTGACGACATGGTTCTACACACTAAGTGGTCAAAGTGTCTTGC	TACGGTAGCAGAGACTTGGTCTAGCAACTGCGGATACTTGGTCTTC
TC_LOJ_69	chr2:446791-446914	ACACTGACGACATGGTTCTACAGGTAGAAGGTACTCTCATCGGTAGCA	TACGGTAGCAGAGACTTGGTCTCAGAAACAGCTCGCCAGAAATAAA
TC_LOJ_70	chr32:839405-839556	ACACTGACGACATGGTTCTACAGGTGCGTACTGTCTTGAAGGTTT	TACGGTAGCAGAGACTTGGTCTGTTGACGATCCACGAAAGATATG
TC_LOJ_71	chr7:179338-179460	ACACTGACGACATGGTTCTACATGGGAGATCGGGAGTACATGAAG	TACGGTAGCAGAGACTTGGTCTTGAAGAGCCAAATGGGACACTAAT

Supplementary Table 2 (continued)

TC_LOJ_74	chr1:1413411-1413530	ACACTGACGACATGGTTCTACA	CAAGATTGTTCCACTGACGAAGACA	TACGGTAGCAGAGACTTGGTCT	TTTTGAGAGCGTGAAGGAGTACACA
TC_LOJ_75	chr23:504383-504519	ACACTGACGACATGGTTCTACA	CTTCATCATCTATGCTCCGACGAC	TACGGTAGCAGAGACTTGGTCT	TCTGAATGACTGGTTGAAAGACGA
TC_LOJ_76	chr23:505516-505635	ACACTGACGACATGGTTCTACA	GTGGACCCAAATGTACTCAGCAAC	TACGGTAGCAGAGACTTGGTCT	GAACCTAAGAAACGAAGAACCCTCA
TC_LOJ_80	chr1:2018618-2018750	ACACTGACGACATGGTTCTACA	AGTGGACATGGTGACGAAGATGAG	TACGGTAGCAGAGACTTGGTCT	GTAGTGCTTCAAACCGCTCAAGAA
TC_LOJ_81	chr37:132370-132499	ACACTGACGACATGGTTCTACA	ACCGGATGTATTCTCTCGTGGTA	TACGGTAGCAGAGACTTGGTCT	CATGCACTTATCGTCGTCACCTTTC
TC_LOJ_82	chr13:741015-741134	ACACTGACGACATGGTTCTACA	CACAAACCGCTTAGACCCTGAAGT	TACGGTAGCAGAGACTTGGTCT	CCAGAAGAAACAATCAATCAACAGC
TC_LOJ_85	chr1:351420-351541	ACACTGACGACATGGTTCTACA	AGACTCAATCGCCTTCACGACATA	TACGGTAGCAGAGACTTGGTCT	CAGAGGTGTTTATGAGCAAGTACCG
TC_LOJ_86	chr18:746701-746824	ACACTGACGACATGGTTCTACA	ACCCACTCCAGTAGCATTTCTTCC	TACGGTAGCAGAGACTTGGTCT	TTAACTATGGCAATGAGGCAGAGC
TC_LOJ_87	chr37:464692-464819	ACACTGACGACATGGTTCTACA	CAGATGCTGCCTTGACAGAGATGTA	TACGGTAGCAGAGACTTGGTCT	ACGAGTGTAGAAGCGAAGATGCTG
TC_LOJ_88	chr16:213322-213477	ACACTGACGACATGGTTCTACA	GTAAATAGACACAAGCCATTCCCATC	TACGGTAGCAGAGACTTGGTCT	TACTATCACTACCGTGGGCGTCAG
TC_LOJ_89	chr2:121560-121715	ACACTGACGACATGGTTCTACA	CTCATACCCTTGCTTTGTATGCT	TACGGTAGCAGAGACTTGGTCT	GTTCCAGGAGACGGACCACTAGGTT
TC_LOJ_91	chr12:107750-107877	ACACTGACGACATGGTTCTACA	GAATGACAACAATGCCCTTTCTTC	TACGGTAGCAGAGACTTGGTCT	GTATCTCCATCCATTTCCCAGTGC
TC_LOJ_93	chr27:329031-329151	ACACTGACGACATGGTTCTACA	TCGTAAAGGTATTGGGCATATTTCG	TACGGTAGCAGAGACTTGGTCT	CCAGGATCATTAGCTTAGTCCAG
TC_LOJ_97	chr26:38201-38343	ACACTGACGACATGGTTCTACA	TTTGAAGAGAAGATGGCCCTGAGT	TACGGTAGCAGAGACTTGGTCT	TTTGAAGAAAGGATCTGCCTCGTAA
TC_LOJ_99	chr33:297174-297306	ACACTGACGACATGGTTCTACA	CAAGTTCCTGTTGGACGTGGTAGT	TACGGTAGCAGAGACTTGGTCT	AATGTACGCAAGGAGCGACTAGAG
TC_LOJ_100	chr26:479107-479233	ACACTGACGACATGGTTCTACA	TATTATTTACGAAACGGCGGAGGA	TACGGTAGCAGAGACTTGGTCT	AGGAGATGGCTCACTCACTTGAAC
TC_LOJ_102	chr11:853646-853766	ACACTGACGACATGGTTCTACA	AGAACAGGAAGTTGTGACGGTTG	TACGGTAGCAGAGACTTGGTCT	ATCACCTCTGAAAGAATCGACTGC
TC_LOJ_103	chr13:783091-783210	ACACTGACGACATGGTTCTACA	GTACACCCGTCCTTGACAGTATGATT	TACGGTAGCAGAGACTTGGTCT	CGCTGAGTTCACGAAGTTATGCTT
TC_LOJ_104	chr15:807734-807870	ACACTGACGACATGGTTCTACA	CAAGTTCGCAATGTAGGAAAGCTG	TACGGTAGCAGAGACTTGGTCT	TATCATGGTGGTTCGATGCTGAATA
TC_LOJ_107	chr2:160058-160182	ACACTGACGACATGGTTCTACA	GTCATACCTTACCAAACGGCACAG	TACGGTAGCAGAGACTTGGTCT	ATGTGAACAACCGTACTGGAGGTG
TC_LOJ_108	chr13:664297-664421	ACACTGACGACATGGTTCTACA	TATCTGTGGTGGCTGTAGATGGTG	TACGGTAGCAGAGACTTGGTCT	CGACGACAACAAGGAAGAAGAGGTA
TC_LOJ_109	chr26:419336-419479	ACACTGACGACATGGTTCTACA	CTTTCGGTGTACGGTGTACTTCAG	TACGGTAGCAGAGACTTGGTCT	TCACTGTTTACAACACGGCCAGA
TC_LOJ_111	chr41:288290-288430	ACACTGACGACATGGTTCTACA	CCACGCCACCAGTAACGATAATAA	TACGGTAGCAGAGACTTGGTCT	GAAGAAGTGGTACTCTCCCGATCC
TC_LOJ_114	chr5:168922-169061	ACACTGACGACATGGTTCTACA	TTAGAAACCGTGAGAGACTTGTCAAC	TACGGTAGCAGAGACTTGGTCT	ATTACCCTGCACCAAGACACATTC
TC_LOJ_116	chr26:336772-336902	ACACTGACGACATGGTTCTACA	GCTGTCTCCAAGAGTCGAGAATA	TACGGTAGCAGAGACTTGGTCT	CATGGATTCTTTCCAGTGCTTTG
TC_LOJ_117	chr3:965641-965793	ACACTGACGACATGGTTCTACA	TCCAATCTCTTATCTTTCAGGAGAACG	TACGGTAGCAGAGACTTGGTCT	CATACTCAAACGAGGCACGAATCT
TC_LOJ_118	chr15:398374-398497	ACACTGACGACATGGTTCTACA	CCACAAGTAGGCTGAACCACAAAT	TACGGTAGCAGAGACTTGGTCT	GTCAAGCCCTTCGTATCCCTGTTA
TC_LOJ_119	chr1:2137512-2137631	ACACTGACGACATGGTTCTACA	GAATCATCAGAGGGTCATTTGCAC	TACGGTAGCAGAGACTTGGTCT	AGTACACAACAAAGTTATCGCGGATG
TC_LOJ_120	chr3:196127-196261	ACACTGACGACATGGTTCTACA	TCATCCTCATCTTCTGGTGGTGAT	TACGGTAGCAGAGACTTGGTCT	TGGACTCTCACTTCTGTATCTACTTTGTTG
TC_LOJ_121	chr27:93351-93474	ACACTGACGACATGGTTCTACA	ACTGCGTTGTATAGCCGAATCACT	TACGGTAGCAGAGACTTGGTCT	GACAGGAACACCAAATGTACTGTGAA
TC_LOJ_122	chr36:377593-377718	ACACTGACGACATGGTTCTACA	CTTTCCTGGGTTCTGTTGGTTAAG	TACGGTAGCAGAGACTTGGTCT	CAGGTGTTCTCGTCAAGCTGTAAT

Supplementary Table 2 (continued)

TC_LOJ_124	chr10:933564-933686	ACACTGACGACATGGTTCTACA	TGCAAATACAGAAGATGAGCTACGC	TACGGTAGCAGAGACTTGGTCT	TGATTATGAGGAGGAGGATGCAGT
TC_LOJ_125	chr21:539837-539959	ACACTGACGACATGGTTCTACA	AAATCTCAGCTACAACAACATCTCTGG	TACGGTAGCAGAGACTTGGTCT	TCATCCTTTCCATCGTTCTCACTT
TC_LOJ_126	chr15:908929-909068	ACACTGACGACATGGTTCTACA	GGCCTTCTCACTAACTGTCTGATCTG	TACGGTAGCAGAGACTTGGTCT	ACCTTCTTATCACGGAAGAGTATCAGG
TC_LOJ_128	chr11:775649-775772	ACACTGACGACATGGTTCTACA	GAAAGAAGCTGAAGAATGGGCAAA	TACGGTAGCAGAGACTTGGTCT	GTTGATCCTGGCAATTACACTCGT
TC_LOJ_129	chr18:115349-115471	ACACTGACGACATGGTTCTACA	GTGACTTGCGGATTATGATTCGTT	TACGGTAGCAGAGACTTGGTCT	CGTTTGTCTTCTCATCCTTCTCG
TC_LOJ_130	chr9:601749-601872	ACACTGACGACATGGTTCTACA	TCCCGTTACATCCAATACATCCAA	TACGGTAGCAGAGACTTGGTCT	TGCATACACAACAGAGCTAAGTGTCG
TC_LOJ_131	chr9:601909-602028	ACACTGACGACATGGTTCTACA	ACAAGCAATCCAATTACAACCACAG	TACGGTAGCAGAGACTTGGTCT	ATTAAAGAAGGTCGCGGCAGTAGA
TC_LOJ_136	chr23:522688-522812	ACACTGACGACATGGTTCTACA	TTTCAAGCTGCGACTTAATCAACG	TACGGTAGCAGAGACTTGGTCT	GATGGAATGCTTCTTGACAGTC
TC_LOJ_137	chr16:889485-889604	ACACTGACGACATGGTTCTACA	CATTTCTGCTGCTTCTTTGAGAA	TACGGTAGCAGAGACTTGGTCT	TCTGATGTTGATCTCTCTTAACTACCG
TC_LOJ_138	chr5:1116604-1116723	ACACTGACGACATGGTTCTACA	CATTTACCAGAAAGTGACAGCAAC	TACGGTAGCAGAGACTTGGTCT	GATGAGGGAGAAGCGAATTTGAAC
TC_LOJ_140	chr19:251999-252118	ACACTGACGACATGGTTCTACA	CCCTCACCTCAATCATATCCACAC	TACGGTAGCAGAGACTTGGTCT	GGGACAAGTACGGGAACAGAATAGA
TC_LOJ_141	chr37:317244-317399	ACACTGACGACATGGTTCTACA	ATTGTGAGAGGATGGGTTCAAATG	TACGGTAGCAGAGACTTGGTCT	CCAGTGCATACTTCTGTGTTATGGTAGA
TC_LOJ_142	chr2:327727-327846	ACACTGACGACATGGTTCTACA	ATGCGGGAGTGTTGTGCATTAGTAT	TACGGTAGCAGAGACTTGGTCT	ACGGAATACGGGTGGAATAAGAAA
TC_LOJ_144	chr11:235518-235637	ACACTGACGACATGGTTCTACA	ACGCAGTTGGTCGAGAATTGTATC	TACGGTAGCAGAGACTTGGTCT	GAAGGAGAGGTGGTGCAGCTTATC
TC_LOJ_145	chr6:23502-23628	ACACTGACGACATGGTTCTACA	TTGGCATAAAGGTACGAATCATGG	TACGGTAGCAGAGACTTGGTCT	GAACTCACGACCCTGAATAAGACG
TC_LOJ_146	chr27:232849-232974	ACACTGACGACATGGTTCTACA	CTCAGTATGAACTCCGCTTCTGT	TACGGTAGCAGAGACTTGGTCT	GGATATGTGCTCAAAGTGCCTTGT
TC_LOJ_147	chr4:1219111-1219233	ACACTGACGACATGGTTCTACA	AAGCTGAATAGATCGCACAAGCTC	TACGGTAGCAGAGACTTGGTCT	TATGCCCTATCCGTGTTCTTACG
TC_LOJ_152	chr19:553417-553540	ACACTGACGACATGGTTCTACA	CATAAGGGCAGTGTCATCAACAAA	TACGGTAGCAGAGACTTGGTCT	GTATTGCTGGTTGGTTCTCTTCCA
TC_LOJ_154	chr37:156377-156496	ACACTGACGACATGGTTCTACA	GTAAGGACCACAAGAGGGAAATGG	TACGGTAGCAGAGACTTGGTCT	GCAGAGTAGACAGCATGGAGTGTG
TC_LOJ_156	chr5:627080-627199	ACACTGACGACATGGTTCTACA	TGGACTACGAGAAGGTTTCATACGAC	TACGGTAGCAGAGACTTGGTCT	GCTGTGGAATGTTGTGATCCTGT
TC_LOJ_157	chr1:1963178-1963304	ACACTGACGACATGGTTCTACA	TAGAAGAGCGTGTGAAGACTGTGG	TACGGTAGCAGAGACTTGGTCT	ATGACAACCGCGTCACTTGAATAC
TC_LOJ_158	chr1:1964699-1964825	ACACTGACGACATGGTTCTACA	CTACACGCATTGTGAGAACTTGG	TACGGTAGCAGAGACTTGGTCT	TGAATTTGTCTGGGATGTGGAAC
TC_LOJ_159	chr1:1998360-1998510	ACACTGACGACATGGTTCTACA	ACCGTGCTACTTTCTTCTTTGGT	TACGGTAGCAGAGACTTGGTCT	AATCTTCTCAATCTCCCTGCTGT
TC_LOJ_160	chr16:738527-738679	ACACTGACGACATGGTTCTACA	CAGCCACTGTTGAGATCCACAAGT	TACGGTAGCAGAGACTTGGTCT	GGCACAAGACCATCAAAGTAGGAC
TC_LOJ_161	chr43:149662-149786	ACACTGACGACATGGTTCTACA	TGTACCTTTCTGCTTTGCTTCTTCC	TACGGTAGCAGAGACTTGGTCT	TGATGACTATCGCTCCATTCTTCC
TC_LOJ_162	chr16:189968-190097	ACACTGACGACATGGTTCTACA	GCTTTGGAGTAGAGCAGATTTGGA	TACGGTAGCAGAGACTTGGTCT	CCGAGTTACATTTCTTTGCCTTG
TC_LOJ_163	chr18:523652-523773	ACACTGACGACATGGTTCTACA	GATCGCGTTGTAAGCAAATTCAG	TACGGTAGCAGAGACTTGGTCT	GGCGTAAAGGGCAACTCAAAGTAT
TC_LOJ_165	chr3:169504-169625	ACACTGACGACATGGTTCTACA	CACGAAAGTCAAACCTCCACAA	TACGGTAGCAGAGACTTGGTCT	GGTAAATACACGTCCACCGACCTT
TC_LOJ_166	chr3:169646-169792	ACACTGACGACATGGTTCTACA	GGCAACGTGGTATGGAATGATAAC	TACGGTAGCAGAGACTTGGTCT	TCTGCTCACACAGGACTGAATCTC
TC_LOJ_168	chr28:364521-364659	ACACTGACGACATGGTTCTACA	CTCGTGGAAGTTAGTGCTGATCG	TACGGTAGCAGAGACTTGGTCT	CGATGATAAAGAAGTCTCCGTACCC
TC_LOJ_169	chr11:721966-722086	ACACTGACGACATGGTTCTACA	ATGAAACACGTATGCACGATATGC	TACGGTAGCAGAGACTTGGTCT	GGCGCTAAATCTGTACGAATACCA

Supplementary Table 2 (continued)

TC_LOJ_170	chr36:416713-416839	ACACTGACGACATGGTTCTACA	GGGAGTACGAGTTTGACAGAGAAGA	TACGGTAGCAGAGACTTGGTCT	AGAGGGTTGACATAAGGATGCAGA
TC_LOJ_171	chr2:854454-854583	ACACTGACGACATGGTTCTACA	AGCAAGGGCAGTCACAAAGTAACA	TACGGTAGCAGAGACTTGGTCT	ACTGTGGGTGATACAGGCAAAGAC
TC_LOJ_173	chr19:264153-264279	ACACTGACGACATGGTTCTACA	CATTGAGAACCACGACTGGCTATT	TACGGTAGCAGAGACTTGGTCT	GGACTATGAGATCGACAAGGAGTTTG
TC_LOJ_174	chr18:456154-456275	ACACTGACGACATGGTTCTACA	ATATCATGGGACTTGCCGGATTAC	TACGGTAGCAGAGACTTGGTCT	CAATGTCTGGTTTGAGGAAGAAG
TC_LOJ_175	chr13:608121-608257	ACACTGACGACATGGTTCTACA	ACTGACATGGATCATAGCCAATCG	TACGGTAGCAGAGACTTGGTCT	CGATAAAGGAACCCAACAAGAACC
TC_LOJ_177	chr7:1112127-1112263	ACACTGACGACATGGTTCTACA	CTTTGAGAGCTTTGCATCCTTCAC	TACGGTAGCAGAGACTTGGTCT	CCGGGACGAGTACACATATACCAA
TC_LOJ_178	chr10:265161-265291	ACACTGACGACATGGTTCTACA	GGTATGAGCATCGCCTTATTGATG	TACGGTAGCAGAGACTTGGTCT	AAGAGAACCAAATCCCTGAGCAAC
TC_LOJ_180	chr8:851024-851146	ACACTGACGACATGGTTCTACA	GACGATGAGGAGTTGGAGGATGTA	TACGGTAGCAGAGACTTGGTCT	AGTGTGGCGATAGGTGATTGTGAT
TC_LOJ_181	chr7:987164-987292	ACACTGACGACATGGTTCTACA	TAGATGTTTGGTCCCATTTGAAGG	TACGGTAGCAGAGACTTGGTCT	TGATACCGTCACTATTACCGCTAGAAA
TC_LOJ_182	chr15:497344-497472	ACACTGACGACATGGTTCTACA	TGTCCAAGACCTTCACATAGTCCA	TACGGTAGCAGAGACTTGGTCT	TGGTTACTTTCCAGACAAGGGATG
TC_LOJ_184	chr37:138690-138820	ACACTGACGACATGGTTCTACA	AGCTTGGCCTTCAACACATCATT	TACGGTAGCAGAGACTTGGTCT	GCGTCATACTCCCTCACATATCCA
TC_LOJ_185	chr27:387192-387314	ACACTGACGACATGGTTCTACA	GGGTGATAGATGCTGTTGCTGAAT	TACGGTAGCAGAGACTTGGTCT	TGAGTTAATGGACCCGAAGGAAC
TC_LOJ_187	chr15:795497-795621	ACACTGACGACATGGTTCTACA	GACAAACATTCGACCTTCATCTTCTG	TACGGTAGCAGAGACTTGGTCT	TGGTATTTGAGGATCATTCCAGTCA
TC_LOJ_188	chr1:2220221-2220341	ACACTGACGACATGGTTCTACA	CCAGGTTGTTGGTTGTTATGTGGT	TACGGTAGCAGAGACTTGGTCT	GCGGAGATTCACGAAATAGAGGAA
TC_LOJ_191	chr5:703969-704096	ACACTGACGACATGGTTCTACA	CTATTGGATGGGAACGTGGTACAG	TACGGTAGCAGAGACTTGGTCT	GCACAATCTCTGTTGTAAGACTAACTCCT
TC_LOJ_192	chr37:447759-447878	ACACTGACGACATGGTTCTACA	CGTATCAAACAGGGCTGGAGACTT	TACGGTAGCAGAGACTTGGTCT	ATCAAGCTGCAAGAAGAGAACATCC
TC_LOJ_195	chr27:40705-40826	ACACTGACGACATGGTTCTACA	ATGTTTCCTTGCATGAGTTTGTGG	TACGGTAGCAGAGACTTGGTCT	GGAGTCGCCGTAGTATTCCCTTATG
TC_LOJ_197	chr41:298702-298834	ACACTGACGACATGGTTCTACA	ATTGGGACGGTAGAGCATGTAAGG	TACGGTAGCAGAGACTTGGTCT	GCCTGAGTTCCCTCCAGTCTTTCTT
TC_LOJ_200	chr37:173415-173536	ACACTGACGACATGGTTCTACA	CACGAAACTGCCAATGATGACTCT	TACGGTAGCAGAGACTTGGTCT	CACCTCCGCTTTTCTCTCCTTCT
TC_LOJ_201	chr32:855499-855637	ACACTGACGACATGGTTCTACA	AAGAGGCGTGTAAGAAAGTATGTGGAG	TACGGTAGCAGAGACTTGGTCT	TGCAAGTAGTCAGCAATGTCCAGT
TC_LOJ_203	chr25:64845-64984	ACACTGACGACATGGTTCTACA	ACGCGGATACTAGGGAACATGAGT	TACGGTAGCAGAGACTTGGTCT	TTGAGCAGAATACCAAAGCAGTTGT
TC_LOJ_204	chr9:194610-194758	ACACTGACGACATGGTTCTACA	CTGTTCAAAGTCCATTGTGCTATCC	TACGGTAGCAGAGACTTGGTCT	ATGACTGCAAGGTATTCCGCTTCT
TC_LOJ_205	chr7:1037003-1037155	ACACTGACGACATGGTTCTACA	ACAGGGCTTCAGGTGGACATTATT	TACGGTAGCAGAGACTTGGTCT	GGTTAAAGGTCGTGGTTGACACAT
TC_LOJ_206	chr19:762223-762346	ACACTGACGACATGGTTCTACA	AGCCTTCCCTTTCTACTGGTGGTA	TACGGTAGCAGAGACTTGGTCT	TCTGATTTCATACAGTTGCTCCTC
TC_LOJ_209	chr1:2005883-2006014	ACACTGACGACATGGTTCTACA	TCTTTGAAGTTCTGGTGTGGTT	TACGGTAGCAGAGACTTGGTCT	TCTCAGGGACGAGGAGACATATAAGA
TC_LOJ_211	chr2:916287-916407	ACACTGACGACATGGTTCTACA	CTTGATAAACTCTGCGGCTTCCTC	TACGGTAGCAGAGACTTGGTCT	CAATGGTACGAACATGATTGACTGTG
TC_LOJ_212	chr44:285730-285879	ACACTGACGACATGGTTCTACA	GCTGTCCATATCCGCATCTTCTAA	TACGGTAGCAGAGACTTGGTCT	ATGTCGTTTCCAAATCAGCACAAAC
TC_LOJ_213	chr32:839358-839478	ACACTGACGACATGGTTCTACA	GGTGACAAACCCATTACAGCTTACA	TACGGTAGCAGAGACTTGGTCT	TACAGCGCAATCAAATCCACTAC
TC_LOJ_214	chr11:849661-849797	ACACTGACGACATGGTTCTACA	TTACTACATTGGTGGCGAGACAAAC	TACGGTAGCAGAGACTTGGTCT	TCAGACGAAACAGATAGCTCGTGA
TC_LOJ_215	chr10:1052122-1052245	ACACTGACGACATGGTTCTACA	CAGAGTTCTACAAGGAAGATCGACAAA	TACGGTAGCAGAGACTTGGTCT	TTAATGATGGGTGGAAGTGAGAGG
TC_LOJ_217	chr1:2773733-2773861	ACACTGACGACATGGTTCTACA	AAACTTATGGCGTACAACAGGGAGT	TACGGTAGCAGAGACTTGGTCT	CGATAACGACGATGAAGATGATGA

Supplementary Table 2 (continued)

TC_LOJ_219	chr26:38066-38187	ACACTGACGACATGGTTCTACA	GTTGATGTGGATAGGCTTGACTACTTTC	TACGGTAGCAGAGACTTGGTCT	TCACCTTCGTAGCACAATACCTTACA
TC_LOJ_220	chr14:923562-923682	ACACTGACGACATGGTTCTACA	TCGGGTAAATGTCTAACGGAGAAA	TACGGTAGCAGAGACTTGGTCT	CCAGATCCAGTGATTTCGTCTTGTT
TC_LOJ_221	chr11:868950-869070	ACACTGACGACATGGTTCTACA	GCTTCACAGCTATCGAGGTGTATTG	TACGGTAGCAGAGACTTGGTCT	CCAGGAGTTTAGTTACAACAGACGAGA
TC_LOJ_223	chr27:96137-96258	ACACTGACGACATGGTTCTACA	CAAGCGCACCCCTAATAAGAAATTG	TACGGTAGCAGAGACTTGGTCT	CAACAAAGAGCTTCAAATGGTGTG
TC_LOJ_224	chr1:2775484-2775623	ACACTGACGACATGGTTCTACA	GGTGTGTACGGATGACTGCTACTTACTT	TACGGTAGCAGAGACTTGGTCT	CAACAAGGACAAAGACAACCACAA
TC_LOJ_225	chr15:246311-246435	ACACTGACGACATGGTTCTACA	CGTGAAAGATACGGCTGACACATA	TACGGTAGCAGAGACTTGGTCT	GTAGTGCGTGTTGCTCCTGTTGTT
TC_LOJ_227	chr27:116142-116263	ACACTGACGACATGGTTCTACA	ATGAGGAGGAGGAGAAATGGAAAC	TACGGTAGCAGAGACTTGGTCT	GTCGATGACACAGTCCAGACACTC
TC_LOJ_228	chr5:1147485-1147616	ACACTGACGACATGGTTCTACA	ACAGTGCAGTCGTAATTCGCAAT	TACGGTAGCAGAGACTTGGTCT	TGTTGACTACTTTGACGGAAATCGT
TC_LOJ_229	chr5:1148049-1148168	ACACTGACGACATGGTTCTACA	AGTGGCTTGGCAGATTTCTTCTGT	TACGGTAGCAGAGACTTGGTCT	TGACAGTTTAGAGAGCGTTGTAGTGAAAG
TC_LOJ_230	chr15:926778-926915	ACACTGACGACATGGTTCTACA	ATTCTGCCTGCGACAGTAGTTCTC	TACGGTAGCAGAGACTTGGTCT	CCATTCTTCGTGAAATTGAGGTTG
TC_LOJ_231	chr1:2138077-2138196	ACACTGACGACATGGTTCTACA	GGCAGACTCCAGATACTGACGAAT	TACGGTAGCAGAGACTTGGTCT	CCACAACCTCCTTGACGACTTTCTT
TC_LOJ_232	chr5:191326-191447	ACACTGACGACATGGTTCTACA	ACATCCTGACCCCTTGCTTTAGAC	TACGGTAGCAGAGACTTGGTCT	GGTTAGAGAGAACATTACGACGGAGA
TC_LOJ_234	chr10:715504-715626	ACACTGACGACATGGTTCTACA	AGTAAGCCTGTTGCTTTGGAACTC	TACGGTAGCAGAGACTTGGTCT	TCAACCCAGACGAAAGTCTAGTGG
TC_LOJ_235	chr15:197505-197635	ACACTGACGACATGGTTCTACA	TCGTCAATTTCCCGTAGGATACTTT	TACGGTAGCAGAGACTTGGTCT	CAGGAGGAGGGTGAACGTATAATG
TC_LOJ_236	chr11:235245-235379	ACACTGACGACATGGTTCTACA	ATCTTTACCATGCACCTCCACAAC	TACGGTAGCAGAGACTTGGTCT	GGTCTCACCACGTATCACGAGAAG
TC_LOJ_237	chr9:134209-134328	ACACTGACGACATGGTTCTACA	CTCTTCACGCCAATACATTCTTG	TACGGTAGCAGAGACTTGGTCT	CCAGCTACAACTGCAAACAAATACAC
TC_LOJ_238	chr21:322787-322911	ACACTGACGACATGGTTCTACA	TCAGGGTAGATTCATCAGGCAGAG	TACGGTAGCAGAGACTTGGTCT	TATCAACAATGCTCGACACCCACT
TC_LOJ_239	chr44:237246-237373	ACACTGACGACATGGTTCTACA	ATTTATGCCCCGAAACCAGATAAC	TACGGTAGCAGAGACTTGGTCT	CGAGGCAATTCGTATAATGTCTTCA
TC_LOJ_242	chr31:92921-93071	ACACTGACGACATGGTTCTACA	ATTGAAGTATCGCCAGAACAGCAT	TACGGTAGCAGAGACTTGGTCT	GTGTTGCTTGAGTAAGGCACTCT
TC_LOJ_243	chr21:288200-288319	ACACTGACGACATGGTTCTACA	CGGTCAGGATCGTTATAGTTTGGTAG	TACGGTAGCAGAGACTTGGTCT	AGACACTTTGTATCGTATGCGTCGT
TC_LOJ_244	chr18:566462-566592	ACACTGACGACATGGTTCTACA	ATTATCTCGTGAGTTTGGCGAAT	TACGGTAGCAGAGACTTGGTCT	CAGAACCGCTTGTCTTCACTTC
TC_LOJ_245	chr3:1209990-1210114	ACACTGACGACATGGTTCTACA	GGATCGACGTATGGGACGTATTTT	TACGGTAGCAGAGACTTGGTCT	TTGAAGGACTGGAGCAAGACAAGT
TC_LOJ_249	chr10:1031977-1032097	ACACTGACGACATGGTTCTACA	AAGCTCAGTGTTCAAAGTGCCATC	TACGGTAGCAGAGACTTGGTCT	TTTCCTTGTTATCGGCTGTGAGAA
TC_LOJ_250	chr21:505080-505199	ACACTGACGACATGGTTCTACA	GTTCTCCGTTACTTTCCGACACAG	TACGGTAGCAGAGACTTGGTCT	TGCCATGTTACCCATAAACCCTT
TC_LOJ_251	chr5:743274-743396	ACACTGACGACATGGTTCTACA	CTAGGGATAGTGCTCAACATTGGCTATAA	TACGGTAGCAGAGACTTGGTCT	CACCTTTAACTTTGAACGAACACG
TC_LOJ_252	chr36:237339-237479	ACACTGACGACATGGTTCTACA	TTAGAGCTTCGTATCGGCATGTTG	TACGGTAGCAGAGACTTGGTCT	CACCTTCATACATTTCTCCAGAGACC
TC_LOJ_253	chr3:240382-240505	ACACTGACGACATGGTTCTACA	CCACTACCATTACCCGTGTCGTTA	TACGGTAGCAGAGACTTGGTCT	CGCAGTCCTTGCTTAACCTCATTT
TC_LOJ_255	chr27:388555-388675	ACACTGACGACATGGTTCTACA	GTTATTTGTATCCGTATCTTGCTGTCG	TACGGTAGCAGAGACTTGGTCT	AGTATCACCTGGAGGACCGTGAAG
TC_LOJ_256	chr39:221720-221854	ACACTGACGACATGGTTCTACA	AACTGACCGGAAGTGAGATTGATG	TACGGTAGCAGAGACTTGGTCT	GGGCGGCGTCGTAGTATAAATAAG
TC_LOJ_257	chr5:992280-992407	ACACTGACGACATGGTTCTACA	CCTTTATTACGCTTCGGCAAGTACA	TACGGTAGCAGAGACTTGGTCT	TTCCACGCAAACAATCAGTATCAG
TC_LOJ_259	chr32:837402-837557	ACACTGACGACATGGTTCTACA	ACTCTACACAAAGGCGTCAGAGATG	TACGGTAGCAGAGACTTGGTCT	CCTGCAAGATCAATAAGGTTTCAGC

Supplementary Table 2 (continued)

TC_LOJ_260	chr4:1353006-1353141	ACACTGACGACATGGTTCTACA	TGGTACTTGTTTCAGCTCGGAAATC	TACGGTAGCAGAGACTTGGTCT	CAAAGGCAGAGGAATGTTCAAAGA
TC_LOJ_262	chr1:2151183-2151303	ACACTGACGACATGGTTCTACA	CCGTAGTTGCGGTACGAATAAGTG	TACGGTAGCAGAGACTTGGTCT	ACTGGGAACGTGTATTAGGTATGGAGT
TC_LOJ_264	chr18:649186-649316	ACACTGACGACATGGTTCTACA	GTGGAGGCGAAGAAGAAGTTTACA	TACGGTAGCAGAGACTTGGTCT	AATAGAAACGGCATTCCATAAGCAC
TC_LOJ_265	chr27:343910-344029	ACACTGACGACATGGTTCTACA	GTGCATCATATTCGATAGGGAGATGT	TACGGTAGCAGAGACTTGGTCT	TATTACAGCATTGACCGTGTCTTCC
TC_LOJ_266	chr1:2205081-2205202	ACACTGACGACATGGTTCTACA	CTACGAAGTGCCTTAACTGCCTCA	TACGGTAGCAGAGACTTGGTCT	ATTCTATGTGCGTTTGGGTTTCAG
TC_LOJ_267	chr26:405401-405543	ACACTGACGACATGGTTCTACA	TTGCTTTTCGATGGAGATAGACCTTT	TACGGTAGCAGAGACTTGGTCT	GCGGAGATGTCTGATTTAGGAATTG
TC_LOJ_268	chr26:302445-302583	ACACTGACGACATGGTTCTACA	CGTAGTCAAACGGACTGAAGTACACA	TACGGTAGCAGAGACTTGGTCT	GAGGAGGCAGTGGAGGTGTTAAAT
TC_LOJ_269	chr5:495739-495879	ACACTGACGACATGGTTCTACA	TCTTTATGACAAGTGCAACCAAAGC	TACGGTAGCAGAGACTTGGTCT	CGTGATACTCCACCGTCTCAATCT
TC_LOJ_271	chr2:323827-323957	ACACTGACGACATGGTTCTACA	GTGGGTTTCATCTCTCGTTTATGC	TACGGTAGCAGAGACTTGGTCT	ACCCTTGTCATGTGTCTTGTAGC
TC_LOJ_273	chr1:2140290-2140430	ACACTGACGACATGGTTCTACA	CAATGGCACCAAGATAATAGTACAGGA	TACGGTAGCAGAGACTTGGTCT	TGCAGAACCATCGTGAGAACTTTA
TC_LOJ_274	chr21:239185-239310	ACACTGACGACATGGTTCTACA	AACAAGGTGAAGAAGAGCCATCAG	TACGGTAGCAGAGACTTGGTCT	AAGGTGGAGGAGTTGAACAGTACG
TC_LOJ_275	chr39:50470-50598	ACACTGACGACATGGTTCTACA	CTGCTCCTGATACTGCACAACTG	TACGGTAGCAGAGACTTGGTCT	GGTGCCTACAATGACTCCGTACAC
TC_LOJ_276	chr1:2694842-2694979	ACACTGACGACATGGTTCTACA	TTACACATTGCAGGGCAGCATATT	TACGGTAGCAGAGACTTGGTCT	GTCTTTGTTTCGTATGTCAGCGTA
TC_LOJ_277	chr4:1382610-1382749	ACACTGACGACATGGTTCTACA	TAGCATCTTAATCAGCTCGGGAGA	TACGGTAGCAGAGACTTGGTCT	GACGAACAAATGGAGAATCAGACG
TC_LOJ_278	chr2:164952-165077	ACACTGACGACATGGTTCTACA	GGTCATTACGCCAGTTCATACAT	TACGGTAGCAGAGACTTGGTCT	ACGGCCTTCTTCATAATCTCCATAA
TC_LOJ_279	chr9:400076-400197	ACACTGACGACATGGTTCTACA	CGAGACAGGGATGGACTCTTCAAT	TACGGTAGCAGAGACTTGGTCT	GTTACGATGGCCTTGAGTGTGAGA
TC_LOJ_280	chr21:505326-505445	ACACTGACGACATGGTTCTACA	GATTGCTACGTGAAGACGTGGAAG	TACGGTAGCAGAGACTTGGTCT	GAGCGTATCGTACAGGCCAAAGTA
TC_LOJ_281	chr10:735827-735952	ACACTGACGACATGGTTCTACA	CAACGCATTTGGATTGCCTACTAA	TACGGTAGCAGAGACTTGGTCT	AAACGTCTTGGTCTGTACGAGGAG
TC_LOJ_282	chr37:470203-470342	ACACTGACGACATGGTTCTACA	CTACTCAAGGAACCAGGCGTATTG	TACGGTAGCAGAGACTTGGTCT	AACGTCCCACCAAGAATAATGAGC
TC_LOJ_283	chr12:561576-561706	ACACTGACGACATGGTTCTACA	CAGAAGGAGAAGACATTGGAAGTCA	TACGGTAGCAGAGACTTGGTCT	TCTTTGCCACTATCAAGCACCAAC
TC_LOJ_285	chr31:132291-132424	ACACTGACGACATGGTTCTACA	CATTGACCTTGCCACAGAAGTGTA	TACGGTAGCAGAGACTTGGTCT	TGGCCTTATTCACATACTCCACAAG
TC_LOJ_286	chr15:941983-942118	ACACTGACGACATGGTTCTACA	GGCGTATCCACCACAAGAGTAGAA	TACGGTAGCAGAGACTTGGTCT	GGATGCCAGATTACGTGAAAGAAA

955

956

957

958

959

960

961

Supplementary Table 3 Summary of GLST library preparation and sequencing costs. Green dots indicate items/costs related to first-round PCR and clean-up. Blue dots indicate items/costs related to barcoding PCR and clean-up. The cost summary does not consider qPCR materials because we applied qPCR only for purposes of method development. It is not essential for GLST. Abbreviations: EUG (Eurofins Genomics); NEB (New England Biolabs); MGRD (median genotype read-depth).

	Item	Availability (quantity / price)	Quantity for 100 samples	Cost for 100 samples	Comment
Library preparation	200 GLST primer primer pairs (EUG) ●	60.90 ml / 1508.88 £	25 pmol	1.26 £	18,861 bases purchased salt-free at 0.08 £ / base; primers delivered at 200 µM in 150 µl
	Q5 High-Fidelity 2X Master Mix (NEB) ●	2.5 ml / 106.75 £	500 µl	21.35 £	
	UltraPure Agarose (Invitrogen) ●	100 g / 124.00 £	15.6 g	19.34 £	13 agarose gels (0.8%) to visualize 100 samples, separated by empty lanes
	100 bp DNA Ladder* (NEB) ●	50 ug / 34.50 £	13 ug	8.97 £	0.5 ug ladder at left and right margins of each gel
	6X Gel Loading Dye (NEB) ●	1 ml comes free with ladder*	226 µl	0.00 £	2 µl dye for each sample/ladder lane
	PureLink Quick Gel Extraction Kit (Invitrogen) ●	3 x 50 units / 143.64 £	100 units	95.76 £	
	SYBR Safe (Invitrogen) ●	400 µl / 62.78 £	60 µl	9.42 £	
	Miscellaneous ●	n/a	n/a	50.00 £	Pipette tips, vials, blades, etc.
	Barcoded reverse primer (EUG) ●	0.02 µmol / 49.95 £	0.8 nmol	2.00 £	Primers purified by manufacturer using high performance liquid chromatography
	Universal forward primer (EUG) ●	0.02 µmol / 49.95 £	0.8 nmol	2.00 £	Primers purified by manufacturer using high performance liquid chromatography
	Q5 High-Fidelity 2X Master Mix (NEB) ●	(see above)	1 ml	42.70 £	
	Nuclease-free dH ₂ O (Qiagen) ●	1000 ml / 35.68 £	540 µl	19.27 £	
	Qubit assay tubes (Invitrogen) ●	500 tubes / 51.50 £	102 tubes	10.51 £	
	Qubit dsDNA HS Assay Kit (Invitrogen) ●	100 assay kit / 66.25 £	100 assays	66.25 £	
	UltraPure Agarose (Invitrogen) ●	(see above)	1.2 g	1.49 £	Only one agarose gel (0.8%) is needed because samples have been pooled
	100 bp DNA Ladder (NEB) ●	(see above)	1 ug	0.69 £	0.5 ug ladder at left and right margins of the gel
	6X Gel Loading Dye (NEB) ●	(see above)	9 µl	0.00 £	7 µl dye for sample (pool) lane, 2 µl for each ladder lane
	PureLink Quick Gel Extraction Kit (Invitrogen) ●	(see above)	1 unit	0.96 £	Only one unit is needed because samples have been pooled
	SYBR Safe (Invitrogen) ●	(see above)	10 µl	1.57 £	
	Miscellaneous ●	n/a	n/a	50.00 £	Pipette tips, vials, blades, etc.
	Total library preparation cost for 100 samples: 256.41 £				~ 3.15 \$ per sample
Sequencing	Item	Availability (quantity / price)	Quantity for 100 samples	Cost for 100 samples	Comment
	Illumina Reagent Kit v2 Micro	1 cartridge / 390.00 £	1 cartridge	390.00 £	As listed at https://emea.illumina.com (March 2020)
	300-cycle Illumina MiSeq	1 run / 40.00 £	1 run	400.00 £	Costs for quality control, data storage, etc. vary considerably among providers
	Total sequencing cost for 100 samples: 790.00 £				~ 9.72 \$ per sample; 70x MGRD expected based on 125x MGRD for 56 samples in run 2

References

1. Schwabl, P. et al. Meiotic sex in Chagas disease parasite *Trypanosoma cruzi*. *Nat. Commun.* **10**, (2019).
2. Guerra-Assunção, J. A. et al. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife* **4**, (2015).
3. Hall, M. D. et al. Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. *eLife* **8**, (2019).
4. Grigg, M. E., Bonnefoy, S., Hehl, A. B., Suzuki, Y. & Boothroyd, J. C. Success and virulence in *Toxoplasma* as the result of sexual recombination between two distinct ancestries. *Science* **294**, 161–165 (2001).
5. Wu, Z. et al. Point mutations in the major outer membrane protein drive hypervirulence of a rapidly expanding clone of *Campylobacter jejuni*. *Proc. Natl. Acad. Sci. USA* **113**, 10690–10695 (2016).
6. Miotto, O. et al. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nat. Genet.* **47**, 226–234 (2015).
7. Auburn, S. et al. Genomic analysis of a pre-elimination Malaysian *Plasmodium vivax* population reveals selective pressures and changing transmission dynamics. *Nat. Commun.* **9**, (2018).
8. Teixeira, D. G. et al. Comparative analyses of whole genome sequences of *Leishmania infantum* isolates from humans and dogs in northeastern Brazil. *Int. J. Parasitol.* **47**, 655–665 (2017).
9. Devera, R., Fernandes, O. & Coura, J. R. Should *Trypanosoma cruzi* be called ‘cruzi’ complex? A review of the parasite diversity and the potential of selecting population after *in vitro* culturing and mice infection. *Mem. Inst. Oswaldo Cruz* **98**, 1–12 (2003).
10. Alves, A. M., De Almeida, D. F. & von Krüger, W. M. Changes in *Trypanosoma cruzi* kinetoplast DNA minicircles induced by environmental conditions and subcloning. *J. Eukaryot. Microbiol.* **41**, 415–419 (1994).
11. Dvorak, J., Hartman, D. & Miles, M. A. *Trypanosoma cruzi*: correlation of growth kinetics to zymodeme type in clones derived from various sources. *J. Eukaryot. Microbiol.* **27**, 472–474 (2007).

- 990 12. Dean, M. P., Jansen, A. M., Mangia, R. H. R., Gonçalves, A. M. & Morel C. M. Are our laboratory
991 ‘strains’ representative samples of *Trypanosoma cruzi* populations that circulate in nature? *Mem.*
992 *Inst. Oswaldo Cruz* **79**, 19–24 (1984).
- 993 13. Lima, F. M. et al. Interclonal Variations in the molecular karyotype of *Trypanosoma cruzi*:
994 chromosome rearrangements in a single cell-derived clone of the G strain. *PLoS One* **8**, e63738
995 (2013).
- 996 14. Reis-Cunha, J. L. et al. Whole genome sequencing of *Trypanosoma cruzi* field isolates reveals
997 extensive genomic variability and complex aneuploidy patterns within TcII DTU. *BMC Genomics*
998 **19**, 816 (2018).
- 999 15. Cuypers, B. et al. Multiplexed spliced-leader sequencing: a high-throughput, selective method for
1000 RNA-seq in trypanosomatids. *Sci. Rep.* **7**, 1–11 (2017).
- 1001 16. Kumar, N. et al. Efficient subtraction of insect rRNA prior to transcriptome analysis of *Wolbachia*-
1002 *Drosophila* lateral gene transfer. *BMC Res. Notes* **5**, 230 (2012).
- 1003 17. Oyola, S. O. et al. Efficient depletion of host DNA contamination in malaria clinical sequencing. *J.*
1004 *Clin. Microbiol.* **51**, 745–751 (2013).
- 1005 18. Feehery, G. R. et al. A method for selectively enriching microbial DNA from contaminating
1006 vertebrate host DNA. *PLoS One* **8**, e76096 (2013).
- 1007 19. Domagalska, M. A. et al. Genomes of intracellular *Leishmania* parasites directly sequenced from
1008 patients. *bioRxiv* 676163 (2019) doi:10.1101/676163.
- 1009 20. Melnikov, A. et al. Hybrid selection for sequencing pathogen genomes from clinical samples.
1010 *Genome Biol.* **12**, R73 (2011).
- 1011 21. Schuenemann, V. J. et al. Genome-wide comparison of medieval and modern *Mycobacterium*
1012 *leprae*. *Science* **341**, 179–183 (2013).
- 1013 22. Metsky, H. C. et al. Zika virus evolution and spread in the Americas. *Nature* **546**, 411–415 (2017).
- 1014 23. Cowell, A. N. et al. Selective whole-genome amplification is a robust method that enables scalable
1015 whole-genome sequencing of *Plasmodium vivax* from unprocessed clinical samples. *mBio* **8**, (2017).
- 1016 24. Hintzsche, J. D., Robinson, W. A. & Tan, A. C. A survey of computational tools to analyze and
1017 interpret whole exome sequencing data. *Int. J. Genomics* **2016**, (2016).

- 1018 25. Gampawar, P. et al. Evaluation of the performance of AmpliSeq and SureSelect exome sequencing
1019 libraries for Ion Proton. *Front. Genet.* **10**, 856 (2019).
- 1020 26. Nag, S. et al. High throughput resistance profiling of *Plasmodium falciparum* infections based on
1021 custom dual indexing and Illumina next generation sequencing-technology. *Sci. Rep.* **7**, (2017).
- 1022 27. Balkenhol, N., Cushman, S., Storfer, A. & Waits, L. *Landscape genetics: concepts, methods,*
1023 *applications.* (John Wiley & Sons, 2015).
- 1024 28. Momčilović, S., Cantacessi, C., Arsić-Arsenijević, V., Otranto, D. & Tasić-Otašević, S. Rapid
1025 diagnosis of parasitic diseases: current scenario and future needs. *Clin. Microbiol. Infect.* **25**, 290–
1026 309 (2019).
- 1027 29. Arias, A. et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission
1028 chains linked to sporadic cases. *Virus Evol.* **2**, vew016 (2016).
- 1029 30. Park, J. et al. Determining genotypic drug resistance by ion semiconductor sequencing with the Ion
1030 AmpliSeq™ TB Panel in multidrug-resistant *Mycobacterium tuberculosis* isolates. *Ann. Lab. Med.*
1031 **38**, 316–323 (2018).
- 1032 31. Ferrario, C. et al. A genome-based identification approach for members of the genus
1033 *Bifidobacterium*. *FEMS Microbiol. Ecol.* **91**, (2015).
- 1034 32. Makowsky, R. et al. Genomic diversity and phylogenetic relationships of human papillomavirus 16
1035 (HPV16) in Nepal. *Infect. Genet. Evol.* **46**, 7–11 (2016).
- 1036 33. Grijalva, M. J., Suarez-Davalos, V., Villacis, A. G., Ocaña-Mayorga, S. & Dangles, O. Ecological
1037 factors related to the widespread distribution of sylvatic *Rhodnius ecuadoriensis* populations in
1038 southern Ecuador. *Parasit. Vectors* **5**, 17 (2012).
- 1039 34. Nascimento, J. D. et al. Taxonomical over splitting in the *Rhodnius prolixus* (Insecta: Hemiptera:
1040 Reduviidae) clade: are *R. taquarussuensis* (da Rosa et al., 2017) and *R. neglectus* (Lent, 1954) the
1041 same species? *PLoS One* **14**, e0211285 (2019).
- 1042 35. Velásquez-Ortiz, N. et al. *Trypanosoma cruzi* infection, discrete typing units and feeding sources
1043 among *Psammolestes arthuri* (Reduviidae: Triatominae) collected in eastern Colombia. *Parasit.*
1044 *Vectors* **12**, 157 (2019).

- 1045 36. Caicedo-Garzón, V. et al. Genetic diversification of *Panstrongylus geniculatus* (Reduviidae:
1046 Triatominae) in northern South America. *PLoS One* **14**, (2019).
- 1047 37. Carrasco, H. J., Torrellas, A., García, C., Segovia, M. & Feliciangeli, M. D. Risk of *Trypanosoma*
1048 *cruzi* I (Kinetoplastida: Trypanosomatidae) transmission by *Panstrongylus geniculatus* (Hemiptera:
1049 Reduviidae) in Caracas (Metropolitan District) and neighboring states, Venezuela. *Int. J. Parasitol.*
1050 **35**, 1379–1384 (2005).
- 1051 38. Carrasco, H. J. et al. Geographical distribution of *Trypanosoma cruzi* genotypes in Venezuela. *PLoS*
1052 *Negl. Trop. Dis.* **6**, (2012).
- 1053 39. Nakad, B. C. C. et al. Genetic variability of *Panstrongylus geniculatus* (Reduviidae: Triatominae) in
1054 the Metropolitan District of Caracas, Venezuela. *Infect. Genet. Evol.* **66**, 236–244 (2018).
- 1055 40. Messenger, L. A., Yeo, M., Lewis, M. D., Llewellyn, M. S. & Miles, M. A. Molecular genotyping
1056 of *Trypanosoma cruzi* for lineage assignment and population genetics. *Methods Mol. Biol.* **1201**,
1057 297–337 (2015).
- 1058 41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
1059 *Bioinformatics* **25**, 1754–1760 (2009).
- 1060 42. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation
1061 DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
- 1062 43. Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS One* **7**, (2012).
- 1063 44. Franzén, O. et al. Comparative genomic analysis of human infective *Trypanosoma cruzi* lineages
1064 with the bat-restricted subspecies *T. cruzi marinkellei*. *BMC Genomics* **13**, 531 (2012).
- 1065 45. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic
1066 genomes. *Genome Res.* **13**, 2178–2189 (2003).
- 1067 46. Talavera-Lopez, C. et al. Repeat-driven generation of antigenic diversity in a major human
1068 pathogen, *Trypanosoma cruzi*. *bioRxiv* 283531 (2018) doi:10.1101/283531.
- 1069 47. You, F. M. et al. BatchPrimer3: a high throughput web application for PCR and sequencing primer
1070 design. *BMC Bioinformatics* **9**, 253 (2008).
- 1071 48. Sonnhammer, E. L. & Hollich, V. Scoredist : a simple and robust protein sequence distance
1072 estimator. *BMC Bioinformatics* **6**, 108 (2005).

- 1073 49. R: the R project for statistical computing. <https://www.r-project.org/>.
- 1074 50. Cummings, K. L. & Tarleton, R. L. Rapid quantitation of *Trypanosoma cruzi* in host tissue by real-
1075 time PCR. *Mol. Biochem. Parasitol.* **129**, 53–59 (2003).
- 1076 51. PhiX Sequencing Control V3. [https://www.illumina.com/products/by-type/sequencing-kits/cluster-](https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html)
1077 [gen-sequencing-reagents/phix-control-v3.html](https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html).
- 1078 52. Access Array System for Illumina Sequencing Systems (user guide).
1079 <https://docplayer.net/78505463-Access-array-system-for-illumina-sequencing-systems.html>.
- 1080 53. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from
1081 genomic and metagenomic datasets. *PloS One* **6**, e17288 (2011).
- 1082 54. Picard Tools. Broad Institute. <http://broadinstitute.github.io/picard/>.
- 1083 55. Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific
1084 phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).
- 1085 56. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage
1086 analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 1087 57. Ritland, K. Inferences about inbreeding depression based on changes of the inbreeding coefficient.
1088 *Evolution* **44**, 1230–1241 (1990).
- 1089 58. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg
1090 equilibrium. *Am. J. Hum. Genet.* **76**, 887–893 (2005).
- 1091 59. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform
1092 population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
- 1093 60. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics*
1094 **139**, 457–462 (1995).
- 1095 61. Oksanen, J. et al. vegan: community ecology package.
- 1096 62. Šavrič, B., Jenny, B. & Jenny, H. Projection wizard – an online map projection selection tool.
1097 *Cartogr. J.* **53**, 177–185 (2016).
- 1098 63. Wiens, J. J. & Morrill, M. C. Missing data in phylogenetic analysis: reconciling results from
1099 simulations and empirical data. *Syst. Biol.* **60**, 719–731 (2011).

- 1100 64. Slatkin, M. Isolation by distance in equilibrium and non-equilibrium populations. *Evol. Int. J. Org.*
1101 *Evol.* **47**, 264–279 (1993).
- 1102 65. Zumaya-Estrada, F. A. et al. North American import? Charting the origins of an enigmatic
1103 *Trypanosoma cruzi* domestic genotype. *Parasit. Vectors* **5**, 226 (2012).
- 1104 66. Ocaña-Mayorga, S., Llewellyn, M. S., Costales, J. A., Miles, M. A. & Grijalva, M. J. Sex,
1105 subdivision, and domestic dispersal of *Trypanosoma cruzi* lineage I in southern Ecuador. *PLoS*
1106 *Negl. Trop. Dis.* **4**, e915 (2010).
- 1107 67. Messenger, L. A. et al. Ecological host fitting of *Trypanosoma cruzi* TcI in Bolivia: mosaic
1108 population structure, hybridization and a role for humans in Andean parasite dispersal. *Mol. Ecol.*
1109 **24**, 2406–2422 (2015).
- 1110 68. Ramírez, J. D. et al. Contemporary cryptic sexuality in *Trypanosoma cruzi*. *Mol. Ecol.* **21**, 4216–
1111 4226 (2012).
- 1112 69. Llewellyn, M. S. et al. *Trypanosoma cruzi* IIc: phylogenetic and phylogeographic insights from
1113 sequence and microsatellite analysis and potential impact on emergent Chagas disease. *PLoS Negl.*
1114 *Trop. Dis.* **3**, e510 (2009).
- 1115 70. Roman, F. et al. Dissecting the phyloepidemiology of *Trypanosoma cruzi* I (TcI) in Brazil by the use
1116 of high resolution genetic markers. *PLoS Negl. Trop. Dis.* **12**, e0006466 (2018).
- 1117 71. Barnabe, C. et al. Putative panmixia in restricted populations of *Trypanosoma cruzi* isolated from
1118 wild *Triatoma infestans* in Bolivia. *PloS One* **8**, e82269 (2013).
- 1119 72. Llewellyn, M. S. The molecular epidemiology of *Trypanosoma cruzi* infection in wild and domestic
1120 transmission cycles with special emphasis on multilocus microsatellite analysis (PhD thesis).
1121 *London School of Hygiene & Tropical Medicine* (2008).
- 1122 73. Shibata, H. et al. The use of PCR in detecting toxoplasma parasites in the blood and brains of mice
1123 experimentally infected with *Toxoplasma gondii*. *Kansenshogaku Zasshi* **69**, 158–163 (1995).
- 1124 74. Yang, H., Golenberg, E. M. & Shoshani, J. Proboscidean DNA from museum and fossil specimens:
1125 an assessment of ancient DNA extraction and amplification techniques. *Biochem. Genet.* **35**, 165–
1126 179 (1997).

- 1127 75. Ramos, R. A. N. et al. Quantification of *Leishmania infantum* DNA in the bone marrow, lymph
1128 node and spleen of dogs. *Rev. Bras. Parasitol. Vet. Braz. J. Vet. Parasitol. Orgao Of. Col. Bras.*
1129 *Parasitol. Vet.* **22**, 346–350 (2013).
- 1130 76. Schubert, G. et al. Targeted detection of mammalian species using carrion fly – derived DNA. *Mol.*
1131 *Ecol. Resour.* **15**, (2014).
- 1132 77. Côté, N. M. L. et al. A new high-throughput approach to genotype ancient human gastrointestinal
1133 parasites. *PLoS One* **11**, (2016).
- 1134 78. Cencig, S., Coltel, N., Truyens, C. & Carlier, Y. Parasitic loads in tissues of mice infected with
1135 *Trypanosoma cruzi* and treated with AmBisome. *PLoS Negl. Trop. Dis.* **5**, (2011).
- 1136 79. Souza, R. T. et al. Genome size, karyotype polymorphism and chromosomal evolution in
1137 *Trypanosoma cruzi*. *PLoS One* **6**, e23042 (2011).
- 1138 80. Reithinger, R., Lambson, B. E., Barker, D. C. & Davies, C. R. Use of PCR to detect *Leishmania*
1139 (*Viannia*) spp. in dog blood and bone marrow. **38**, 5 (2000).
- 1140 81. Wen, C. et al. Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq
1141 platform. *PLoS One* **12**, (2017).
- 1142 82. Storfer, A., Patton, A. & Fraik, A. K. Navigating the interface between landscape genetics and
1143 landscape genomics. *Front. Genet.* **9**, (2018).
- 1144 83. Erben, E. D. High-throughput methods for dissection of trypanosome gene regulatory networks.
1145 *Curr. Genomics* **19**, 78–86 (2018).
- 1146 84. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide
1147 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-
1148 3. *Fly* **6**, 80–92 (2012).
- 1149 85. Quinlan, A. R & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
1150 *Bioinformatics* **26**, 841– 842 (2010).
- 1151 86. Aurrecochea, C. et al. EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic*
1152 *Acids Res.* **45**, D581–D591 (2017).
- 1153 87. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

- 1154 88. Linck, E. & Battey, C. J. Minor allele frequency thresholds strongly affect population structure
1155 inference with genomic data sets. *Mol. Ecol. Resour.* **19**, 639–647 (2019).
- 1156 89. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic
1157 inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
- 1158 90. Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. Inferring species
1159 trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol.*
1160 *Biol. Evol.* **29**, 1917–1932 (2012).
- 1161 91. Landguth, E. L., Bearlin, A., Day, C. C. & Dunham, J. CDMetaPOP: an individual-based, eco-
1162 evolutionary model for spatially explicit simulation of landscape demogenetics. *Methods Ecol. Evol.*
1163 **8**, 4–11 (2017).
- 1164 92. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus
1165 genotype data. *Genetics* **155**, 945–959 (2000).
- 1166 93. Piry, S. et al. GENECLASS2: a software for genetic assignment and first-generation migrant
1167 detection. *J. Hered.* **95**, 536–539 (2004).
- 1168 94. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially
1169 explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224–1228 (2013).
- 1170 95. Anderson, E. C. & Thompson, E. A. A model-based method for identifying species hybrids using
1171 multilocus genetic data. *Genetics* **160**, 1217–1229 (2002).
- 1172 96. Graffelman, J., Jain, D. & Weir, B. A genome-wide study of Hardy–Weinberg equilibrium with next
1173 generation sequence data. *Hum. Genet.* **136**, 727–741 (2017).
- 1174 97. Sefid Dashti, M. J. & Gamielien, J. A practical guide to filtering and prioritizing genetic variants.
1175 *BioTechniques* **62**, 18–30 (2017).
- 1176 98. Kaplinski, L., Andreson, R., Puurand, T. & Remm, M. MultiPLX: automatic grouping and
1177 evaluation of PCR primers. *Bioinformatics* **21**, 1701–1702 (2005).
- 1178 99. Etherington, T. R. Python based GIS tools for landscape genetics: visualising genetic relatedness
1179 and measuring landscape connectivity. *Methods Ecol. Evol.* **2**, 52–55 (2011).

100. Carrasco, H. J. et al. *Panstrongylus geniculatus* and four other species of triatomine bug involved in the *Trypanosoma cruzi* enzootic cycle: high risk factors for Chagas' disease transmission in the Metropolitan District of Caracas, Venezuela. *Parasit. Vectors* **7**, 602 (2014).
101. Valadares, H. M. S. et al. Unequivocal identification of subpopulations in putative multiclonal *Trypanosoma cruzi* strains by FACs single cell sorting and genotyping. *PLoS Negl. Trop. Dis.* **6**, e1722 (2012).
102. Zhu, S. J., Almagro-Garcia, J. & McVean, G. Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data. *Bioinformatics* **34**, 9–15 (2018).
103. Lerch, A. et al. Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. *BMC Genomics* **18**, 864 (2017).
104. Chang, H.-H. et al. The real McCOIL: a method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. *PLoS Comput. Biol.* **13**, (2017).
105. Hathaway, N. J., Parobek, C. M., Juliano, J. J. & Bailey, J. A. SeekDeep: single-base resolution *de novo* clustering for amplicon deep sequencing. *Nucleic Acids Res.* **46**, e21 (2018).
106. Zingales, B. *Trypanosoma cruzi* genetic diversity: something new for something known about Chagas disease manifestations, serodiagnosis and drug sensitivity. *Acta Trop.* **184**, 38–52 (2018).
107. Nunes Maria Carmo Pereira et al. Chagas cardiomyopathy: an update of current clinical knowledge and management: a scientific statement from the American Heart Association. *Circulation* **138**, e169–e209 (2018).
108. Llewellyn, M. S. et al. Extraordinary *Trypanosoma cruzi* diversity within single mammalian reservoir hosts implies a mechanism of diversifying selection. *Int. J. Parasitol.* **41**, 609–614 (2011).
109. Pronovost, H. et al. Deep sequencing reveals multiclonality and new discrete typing units of *Trypanosoma cruzi* in rodents from the southern United States. *J. Microbiol. Immunol. Infect.* (2018).
110. Yeo, M. et al. Resolution of multiclonal infections of *Trypanosoma cruzi* from naturally infected triatomine bugs and from experimentally infected mice by direct plating on a sensitive solid medium. *Int. J. Parasitol.* **37**, 111–120 (2007).