Genomics and spatial surveillance of Chagas disease and American visceral leishmaniasis

Philipp Schwabl

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

Institute of Biodiversity, Animal Health & Comparative Medicine
College of Medical, Veterinary & Life Sciences
University of Glasgow

March 2020

Abstract

The Trypanosomatidae are a family of parasitic protozoa that infect various animals and plants. Several species within the *Trypanosoma* and *Leishmania* genera also pose a major threat to human health. Among these are *Trypanosoma cruzi* and *Leishmania infantum*, aetiological agents of the highly debilitating and often deadly vector-borne zoonoses Chagas disease and American visceral leishmaniasis. Current treatment options are far from safe, only partially effective and rarely available in the impoverished regions of Latin America where these 'neglected tropical diseases' prevail. Wider-reaching, sustainable protection against *T. cruzi* and *L. infantum* might best be achieved by intercepting key routes of zoonotic transmission, but this prophylactic approach requires a better understanding of how these parasites disperse and evolve at various spatiotemporal scales.

This dissertation addresses key questions around trypanosomatid parasite biology and spatial epidemiology based on high-resolution, geo-referenced DNA sequence datasets constructed from disease foci throughout Latin America:

Which forms of genetic exchange occur in *T. cruzi*, and are exchange events frequent enough to significantly alter the distribution of important epidemiological traits? How do demographic histories, for example, the recent invasive expansion of *L. infantum* into the Americas, impact parasite population structure, and do structural changes pose a threat to public health? Can environmental variables predict parasite dispersal patterns at the landscape scale?

Following the first chapter's review of population genetic and genomic approaches in the study of trypanosomatid diseases in Latin America, Chapter 2 describes how reproductive polymorphism segregates *T. cruzi* populations in southern Ecuador. The study is the first to clearly demonstrate meiotic sex in this species, for decades thought to exchange genetic material only very rarely, and only by non-Mendelian means. *T. cruzi* subpopulations from the Ecuadorian study site exhibit all major hallmarks of sexual reproduction, including genome-wide Hardy-Weinberg allele frequencies, rapid decay of linkage disequilibrium with map distance and genealogies that fluctuate among chromosomes. The presence of sex promotes the transfer and transformation of genotypes underlying important epidemiological traits, posing great challenges to disease surveillance and the development of diagnostics and drugs.

Chapter 3 demonstrates that mating events are also pivotal to *L. infantum* population structure in Brazil, where introduction bottlenecks have led to striking genetic discontinuities between

sympatric strains. Genetic hybridization occurs genome-wide, including at a recently identified 'miltefosine sensitivity locus' that appears to be deleted from the majority of Brazilian *L. infantum* genomes. The study combines an array of genomic and phenotypic analyses to determine whether rapid population expansion or strong purifying selection has driven this prominent > 12 kb deletion to high abundance across Brazil. Results expose deletion size differences that covary with phylogenetic structure and suggest that deletion-carrying strains do not form a private monophyletic clade. These observations are inconsistent with the hypothesis that the deletion genotype rose to high prevalence simply as the result of a founder effect. Enzymatic assays show that loss of ecto-3'-nucleotidase gene function within the deleted locus is coupled to increased ecto-ATPase activity, raising the possibility that alternative metabolic strategies enhance *L. infantum* fitness in its introduced range. The study also uses demographic simulation modelling to determine whether *L. infantum* populations in the Americas have expanded from just one or multiple introduction events. Comparison of observed vs. simulated summary statistics using random forests suggests a single introduction from the Old World, but better spatial sampling coverage is required to rule out other demographic scenarios in a pattern-process modelling approach. Further sampling is also necessary to substantiate signs of convergent selection introduced above.

Chapter 4 therefore develops a 'genome-wide locus sequence typing' (GLST) tool to summarize parasite genetic polymorphism at a fraction of genomic sequencing cost. Applied directly to the infection source (e.g., vector or host tissue), the method also avoids bias from cell purification and culturing steps typically involved prior to sequencing of trypanosomatid and other obligate parasite genomes. GLST scans genomic pilot data for hundreds of polymorphic sequence fragments whose thermodynamic properties permit simultaneous PCR amplification in a single reaction tube. For proof of principle, GLST is applied to metagenomic DNA extracts from various Chagas disease vector species collected in Colombia, Venezuela, and Ecuador. Epimastigote DNA from several *T. cruzi* reference clones is also analyzed. The method distinguishes 387 single-nucleotide polymorphisms (SNPs) in *T. cruzi* sub-lineage TcI and an additional 393 SNPs in non-TcI clones. Genetic distances calculated from these SNPs correlate with geographic distances among samples but also distinguish parasites from triatomines collected at common collection sites. The method thereby appears suitable for agent-based spatio-genetic (simulation) analyses left wanted by Chapter 3 – and further formulated in Chapter 5.

The potential to survey parasite genetic diversity abundantly across landscapes compels deeper, more systematic exploration of how environmental variables influence the spread of disease. As environmental context is only marginally considered in the population genetic analyses of Chapters 2 – 4, Chapter 5 proposes a new, spatially explicit modelling framework

to predict vector-borne parasite gene flow through heterogeneous environment. In this framework, remotely sensed environmental raster values are re-coded and merged into a composite 'resistance surface' that summarizes hypothesized effects of landscape features on parasite transmission among vectors and hosts. Parasite population genetic differentiation is then simulated on this surface and fitted to observed diversity patterns in order to evaluate original hypotheses on how environmental variables modulate parasite gene flow. The chapter thereby makes a maiden step from standard population genetic to 'landscape genomic' approaches in understanding the ecology and evolution of vector-borne disease.

In summary, this dissertation first demonstrates the power of population genetics and genomics to understand fundamental biological properties of important protist parasites, then identifies areas where analytical tools are missing and creates new technical and conceptual frameworks to help fill these gaps. The general discussion (Chapter 6) also outlines several follow-up projects on the key finding of meiotic genetic signatures in *T. cruzi*. Exploiting recently developed *T. cruzi* genome-editing systems for the detection of meiotic gene expression and heterozygosis will help understand why and in which life cycle stage some parasite populations use sex and others do not. Long-read sequencing of parental and recombinant genomes will help understand the extent to which sex is diversifying *T. cruzi* phenotypes, especially virulence and drug resistance properties conferred by surface molecules with repetitive genetic bases intractable to short-read analysis. Chapter 6 also provides follow-up plans for all other research chapters. Emphasis is placed on advancing the complementarity, transferability and public health benefit of the many different methods and concepts employed in this work.

Table of contents

List of main figures

List of supplementary figures

List of supplementary figures (continued)

List of main tables and boxes

List of supplementary tables

List of abbreviations

°C	degree(s) Celsius
Δx	change in variable x
µ	mutation rate
µg	microgram
µl	microliter
µM	micromolar
3'-AMP	adenosine 3'-monophosphate
AAFM	alternate allele frequency mean
ABCRF	Approximate Bayesian Computation via Random Forests
AD	allelic differences
Ae.	*Aedes*
AIDS	acquired immune deficiency syndrome
AM	Amazonas (state of Brazil) **or** ancient migration (model)
AMbot	ancient migration with bottleneck (model)
An.	*Anopheles*
ANCOVA	analysis of covariance
AR	Ardanza (community in Loja Province, Ecuador)
ATP	adenosine 5'-triphosphate
BA	Bahia (state of Brazil)
BEAST	Bayesian evolutionary analysis by sampling trees
BIC	Bayesian information criterion
BM	Bella Maria (community in Loja Province, Ecuador)
b.	*brucei*
bp	base pair
BS_n	non-recombinant control data simulated with BAMSurgeon
BWA	Burrows-Wheeler Aligner
C.	*Cerdocyon*
ca.	*circa* (approximately)
chr	chromosome (most often used for chr31, i.e., chromosome 31)
chr31 deletion	large (> 12 kb) deletion of interest on *L. infantum* chromosome 31
CISeAL	Centro de Investigación para la Salud en América Latina
cl.	clone
CLIOC	Coleção de *Leishmania* do Instituto Oswaldo Cruz
CNV	copy number variation
CO_2	carbon dioxide
coA	coenzyme A
COII	cytochrome oxidase subunit II
COL	Colombia
crv	cross-validation error
CS	common sequence
Ct	cycle threshold
CV	classification vote (in ABCRF)
D.	*Didelphis*
DAPC	discriminant analysis of principle components
Del	*L. infantum* sample with homozygous chr31 deletion
dep.	department
DF	Distrito Federal (state of Brazil)
DGF	dispersed gene family
dH_2O	distilled water
DNA	deoxyribonucleic acid
DTU	discrete typing unit
e.g.	*exempli gratia* (for example)
et al.	*et alia* (and others)
ECU	Ecuador
EH	El Huayco (community in Loja Province, Ecuador)
ENA	European Nucleotide Archive
ENM	environmental niche modelling
epis	epimastigotes
EPSG	European Petroleum Survey Group

ES	Espírito Santo (state of Brazil)
etc.	*et cetera* (and other similar things)
EUG	Eurofins Genomics
F_1	first filial generation
F_2	second filial generation
FCS	fetal calf serum
FDR	false discovery rate
FIOCRUZ	Fundação Oswaldo Cruz
F_{IS}	inbreeding coefficient
FOU	bottleneck size (fastsimcoal2 simulation parameter)
FSC_n	non-recombinant control data simulated with fastsimcoal2
FSC_r	recombinant control data simulated with fastsimcoal2
F_{ST}	fixation index
FU	fluorescence units
g	gram
GATK	Genome Analysis Toolkit
GC	guanine-cytosine
GE	Gerinoma (community in Loja Province, Ecuador)
GEA	genotype-by-environment association
GIS	geographic information systems
GLST	genome-wide locus sequence typing
GP63	63-kilodalton glycoprotein (often termed leishmanolysin for *Leishmania*)
GPI	glucose-6-phosphate isomerase
GTP	guanosine 5'-triphosphate
GTR	general time-reversible
h	hour
H_2O	water
HCl	hydrochloric acid
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
het.	heterozygosity
HIV	human immunodeficiency virus
HPLC	high-performance liquid chromatography
HS	heterozygous sites
HTZ	*L. infantum* sample with heterozygous chr31 deletion
HWE	Hardy-Weinberg equilibrium
i.e.	*id est* (in other words)
IBD	isolation-by-distance
IBR	isolation-by-resistance
ID	identifier
IM	isolation with migration (model)
IMbot	isolation with migration with bottleneck (model)
IM_{change}	isolation with change in migration (model)
ITM Antwerp	Institute of Tropical Medicine Antwerp
INDEL	insertion-deletion
JPCM5	*L. infantum* reference strain MCAN/ES/98/LLM-724
k	number of groups
K	mean number of alleles over loci
kb	kilobase
KCl	potassium chloride
kDNA	kinetoplastid DNA
KI	Karolinska Institutet
km	kilometer
L	liter **or** DNA ladder
L.	*Leishmania*
LdMT	*L. donovani* miltefosine transporter
LIT	liver infusion tryptose
log	natural logarithm (unless a different base is specified)
$\log_x$	logarithm (base x)
LSSP	low-stringency single-specific primer
LSHTM	London School of Hygiene & Tropical Medicine
LSU rDNA	DNA encoding the large subunit of the ribosome (e.g., eukaryotic 28S)
Lu.	*Lutzomyia*
M	molar

MA	Maranhão (state of Brazil)
MAF	minor allele frequency
MASP	mucin-associated surface protein
max.	maximum
Mb	megabase
MCMC	Markov chain Monte Carlo
MDC	Metropolitan District of Caracas (Venezuela)
met.	metropolitan
MG	Minas Gerais (state of Brazil)
Mg^{2+}	divalent magnesium cation
$MgCl_2$	magnesium chloride
MGRD	median genotype read-depth
$MIG_{x>>>y}$	migration rate from x to y (fastsimcoal2 simulation parameter)
min	minute
min.	minimum
MIX	*L. infantum* sample with both Del- and NonDel-associated PCR amplicons
ml	milliliter
MLEE	multi-locus enzyme electrophoresis
MLG	multi-locus genotype
MLMT	multi-locus microsatellite typing
MLST	multi-locus sequence typing
mM	millimolar
MRD	mitochondrial read-depth
MRDM	multiple regression on distance matrices
MS	Mato Grosso do Sul (state of Brazil)
MT	Mato Grosso (state of Brazil)
MWU	Mann-Whitney U
n	sample size
N	population size
NA	not applicable
NaCl	sodium chloride
ND	not determined
ND1	NADH dehydrogenase subunit I
N_{draws}	number of parameter draws simulated by fastsimcoal2
NEB	New England Biolabs
ng	nanogram
NJ	neighbor-joining
nm	nanometer
nM	nanomolar
nmol	nanomole
NNN	Novy-MacNeal-Nicolle
NonDel	*L. infantum* sample without chr31 deletion
NRD	nuclear read-depth
nt	nucleotide
NTC	no-template control
NW	New World
OW	Old World
P.	*Panstrongylus* **or** *Plasmodium*
PacBio	Pacific Biosciences
PBS	phosphate-buffered saline
PC	positive control
PCA	principle component analysis
PCE	predominant clonal evolution
PCoA	principle coordinate analysis (metric multidimensional scaling)
PCR	polymerase chain reaction
PE	Pernambuco (state of Brazil)
pg	picogram
P_i	inorganic phosphate
PI	Piauí (state of Brazil) **or** previously identified
PM	peritrophic matrix
pmol	picomole
pop.	population
pos.	position

PP	posterior probability (approximated via ABCRF)
PRS	private sites
PS	polymorphic sites
qPCR	quantitative (real-time) PCR
QTL	quantitative trait locus
r	recombination rate
R.	*Rhodnius*
R_0	basic reproductive number
r^2	squared correlation coefficient between genotypes at two SNP loci
RAPD	random amplification of polymorphic DNA
rc	reverse complement
rDNA	DNA encoding ribosomal RNA
RDP	Russian doll patterns
reps.	technical replicates
RFLP	restriction fragment length polymorphism
RJ	Rio de Janeiro (state of Brazil)
RN	Rio Grande do Norte (state of Brazil)
RNA	ribonucleic acid
RPMI	Roswell Park Memorial Institute
RS	Rio Grande do Sul (state of Brazil)
s.d.	standard deviation
s	haploid somy estimate **or** second
S	genome size
S.	*Sciurus*
SA	Salvador (state of Brazil)
SC	Santa Catarina (state of Brazil) **or** secondary contact (model)
SC-Bol	Santa Cruz (department of Bolivia)
$SCbot_{nomig}$	secondary contact with bottleneck without hard admixture (model)
SC_{nomig}	secondary contact without hard admixture (model)
SE	Sergipe (state of Brazil)
SI	strict isolation (model)
SIbot	strict isolation with bottleneck (model)
SL-IR	spliced leader intergenic region
SNP	single-nucleotide polymorphism
SP	São Paulo (state of Brazil)
SR	Santa Rita (community in Loja Province, Ecuador)
SRA	Sequence Read Archive **or** serum resistance-associated
SS	singleton sites
SSU rDNA	DNA encoding the small subunit of the ribosome (e.g., bacterial 16S)
SV	stomodeal valve
T.	*Trypanosoma* **or** *Triatoma*
TcBat	bat-associated *T. cruzi* sublineage
TcI – TcVI	*T. cruzi* discrete typing units I – VI
TcIa – TcId	*T. cruzi* I subtypes suggested by some studies (TcIa is also termed TcI_{DOM})
TcI-Sylvio	*T. cruzi* I reference strain Sylvio X10/1
TR	trypanothione reductase
Twisst	topology weighting by iterative sampling of sub-trees
UI	previously unidentified
UPGMA	unweighted pair group method with arithmetic mean
UVG	uninfected vector gut
VCF	variant call format
v	version
vs.	*versus* (against)
VL	visceral leishmaniasis
VZ	Venezuela
GWAS	genome-wide association study
WGS	whole-genome sequencing
θ	median Watterson estimator
π	median nucleotide diversity
ρ	population recombination parameter
χ^2	chi-squared statistic

The abbreviations above are generally also defined upon first use within each chapter.

Acknowledgments

A great many people and institutions were essential to this dissertation.

Thank you, Martin Llewellyn, for your extraordinary ideas, optimism and compassion in science and life. You opened my mind and built my confidence in ways I cannot summarize in a sentence or two. Great thanks to my co-supervisor Roman Biek for your special insight, always swift and powerful advice. Thank you to my annual assessors Barbara Mable, Poppy Lamberton and Richard Reeve. Your guidance and reinforcement were key to completing this PhD. Many thanks also to Dan Haydon for your special dedication to the institute. Your diverse understanding and vision as director are an inspiration to students and staff.

Thank you to all the sparkling members of the Llewellyn lab. Alessandro Busetti, Michele De Noia and Raminta Kazlauskaitė, it was a special privilege to learn from your mastery in the molecular and microbiological arts. Great thanks, Luis Hernández, for frequent guidance in methods and logistics but also in life beyond the GK. Thank you, Joey Humble, for your careful watch over 202. Thank you, Marnie Davidson and Antonella Bacigalupo, for your tremendous efforts to sustain the malaria project. Thank you, Bachar Cheaib, for your support in bioinformatics and such infective enthusiasm for history, politics and science. Thank you, Leonie Maier and Salem Sueto, for your help in deep sequencing analysis. Thank you, Elle Lindsay, for being so thoughtful and for all the organizational and media talent you put to work for the lab. Thank you, Toni Dwyer and Patrick Schaal, for invigorating our research with your humor and charm. Thank you, Chloe Heys, Lissa Cruz, Maikell Segovia, Maria Augusta Dario, Marianthi Varka, Márlon Grégori, Pei Yang and William Perry, for all the expertise and energy you brought to the North, and thank you, Felix Warren, for bringing cheese intelligence into trypanosomatid research.

Great thanks also to Office 306. Agnes Olin, Diana Meza, Julio Benavides, Kirstyn Brunker, Laura Bergner, Louise Riotte-Lambert and Maude Jacquot, you made the daily grind more pleasurable and gave very valuable advice.

Thanks to so many of you at El Centro de Investigación para la Salud en América Latina. Ailín Blasco-Zúñiga, Álvaro Lara, Ana León, Andrea Vela, Anita Villacís, Bruno Oury, Camila Cilveti, Carolina Crespo, Carolina Portero, César Yumiseva, Christopher García, Claudia Vera, Fabián Sáenz, Fernando Marín, Isabel Tamayo, Jaime Costales, Jalil Maiguashca, Josue Pinto, María de Lourdes Torres, Mario Grijalva, Michelle Del Salto, Patricia Mora, Rossana Romo, Sebasthian Real, Sofía Ocaña, Cynthia Gordón and Verito Aguayza, thank you for making me feel so at home at the institute.

Special thanks to Jaime Costales and Jalil Maiguashca for rescuing me and my research projects a number of times. You took me in as a laboratory novice and completed parasite cloning when I could not. I am also so grateful for all the parasite subcloning, DNA extractions and sample export logistics. Thank you, Camila Cilveti, Christopher García and Sofía Ocaña, for your help around the laboratory and especially during parasite cloning and cryopreservation work. Thank you to Anita Villacís and the countless others behind the tremendous sampling efforts in Loja Province and other parts of Ecuador. Thank you, Mario Grijalva, for establishing these projects and more generally for your great leadership and passion toward protecting public health. Thank you, Josue Pinto, for your kindness and for fast and furious rides to Nayón. Thank you, Rossana Romo, for your vibrance and trust.

Many thanks to Jean-Claude Dujardin, Hideo Imamura, Frederik Van den Broeck, Franck Dumetz, Aya Hefnawy and Bart Cuypers at the Institute of Tropical Medicine in Antwerp. My visit to the ITM was pivotal to this PhD. Thank you so much, Hideo, for your friendship and for greatly expanding my abilities in pathogen genomics.

Thank you, Erin Landguth, for your continuous support in landscape genetics. I am very grateful for all your flexibility and patience surrounding my research exchange to Montana. Thanks also to Casey Day, Dave Holley and Zachary Holden for your support around Skaggs.

Great thanks to Michael Miles for sharing your wisdom in epidemiology and population genetics. Thank you also to Michael Lewis for preparing parasite samples and organizing vector dissections with Matthew Yeo and Cheryl Whitehorn at the London School of Hygiene & Tropical Medicine. And thank you, Louisa Messenger, for your guidance in parasite cloning and microsatellite analysis.

Thank you, Elisa Cupolillo and Mariana Boité, for the strong partnership between Glasgow and the Oswaldo Cruz Institute in Brazil. Thanks also to Gérald Späth and Giovanni Bussotti at Institut Pasteur in France for your collaboration in the *Leishmania* project.

Thank you, Arne Jacobs, for teaching me about coalescent simulation modelling.

Thanks to Bjorn Andersson and Hamid Darban for sequencing work at Karolinska Institutet.

Also many thanks for extensive sequencing work by Julie Galbraith and David McGuinness at Glasgow Polyomics. I am sorry I was often in a scramble and cannot thank you enough for always being so clutch.

Thank you, Iain Sim, for managing the Getafix server, the main workhorse of this PhD.

Thank you also, Umer Ijaz, for the power of Orion, and for all the time you spent troubleshooting my codes.

Many thanks to the members of our vector-borne disease network group, including Alberto Paniz, Belkisyolé Alarcón de Noya, Heather Ferguson, Hernán Carrasco, James Crainey, Jorge Moreno, Juan Vicente Hernández, Maria Eugenia Grillet, Leonardo Ortega, Oscar Noya and Sergio Luz. I have learned so much from your expertise and united effort to improve disease control where it is difficult but needed most.

Many thanks to Andrey Yurchenko, Annette MacLeod, Carlos Talavera-López, Christina Faust, Craig Lapsley, Daniel Streicker, Francesco Baldini, James Cotton, Juan David Ramírez, Katie Hampson, Mafalda Viana, Oscar Franzén, Richard McCulloch, Thomas Otto, Tutku Aykanat and Willie Weir for your assistance with various methods and in improving grant proposals, presentations and manuscripts.

I am also very grateful for the funding I received from the College of Medical, Veterinary & Life Sciences of the University of Glasgow, the Scottish Universities Life Sciences Alliance, the British Society for Parasitology, the Wellcome Trust and the National Institutes of Health in the United States.

Finally, and above all, I thank my family. Mom, Dad and Hannah, thank you for always being there for me, for all your love, patience and understanding.

Hannah, this dissertation is dedicated to you.

Author's declaration

The material presented in this dissertation is the result of research conducted between January 2016 and March 2020.

All research was conducted under the primary supervision of Martin Llewellyn and has not been submitted as part of any other degree. Co-supervision was provided by Roman Biek.

All research represents my own work unless otherwise indicated in the text.

Specifically, in Chapter 2, I established all solid-phase *T. cruzi* plate cultures but did not re-expand a subset of monoclonal microcolonies back into liquid culture for DNA extraction. Jaime Costales and Jalil Maiguashca took over during the re-expansion process (see Section 2.3.1) and also established subclones by limiting dilution as described in Section 2.3.7 and Supplementary Tbl. 2.4. Sequencing occurred at SciLifeLab in Sweden and at Glasgow Polyomics. I performed all bioinformatic analyses on the resultant data. Hideo Imamura provided guidance and Frederik Van den Broeck helped with linkage decay plots. Chapter 2 also incorporated ideas and advice from Michael Miles, Björn Andersson, Mario Grijalva and Martin Llewellyn. All of these contributors therefore represent co-authors with affiliations listed on the title page of Chapter 2.

In Chapter 3, Mariana Boité and Otacilio Moreira performed flow-cytometry, DNA extractions, conventional and quantitative qPCR (Sections 3.3.3 and 3.3.4). Anita Freitas-Mesquita and José Roberto Meyer-Fernandes performed enzymatic assays (Section 3.3.5). Sequencing occurred at SciLifeLab in Sweden and at Glasgow Polyomics. I performed all bioinformatic analyses on the resultant data. Arne Jacobs provided guidance with coalescent simulation modelling. Chapter 3 also incorporated ideas and advice from Björn Andersson, Gerald Späth, Elisa Cupolillo and Martin Llewellyn. All of these contributors therefore represent co-authors with affiliations listed on the title page of Chapter 3.

Chapter 4 involved metagenomic extracts provided through collaborations with Jaime Costales, Jalil Maiguashca, Sofía Ocaña, Carolina Hernández, Juan David Ramírez, Maikell Segovia and Hernán Carrasco. I also used *T. cruzi* epimastigote pellets prepared by Jalil Maiguashca and Michael Lewis (Section 4.3.1). I performed all other laboratory procedures prior to sequencing at Glasgow Polyomics and performed all bioinformatic analyses with the resultant data. Chapter 4 also incorporated ideas and advice from Mario Grijalva and Martin Llewellyn. All of these contributors therefore represent co-authors with affiliations listed on the title page of Chapter 4.

I devised Chapter 5's framework and created all figures and text. However, the chapter was facilitated by preliminary discussions with Martin Llewellyn, Erin Landguth, Björn Andersson, Uriel Kitron, Jaime Costales, Sofía Ocaña and Mario Grijalva. Also, Figs. 5.1 and 5.2 incorporate maps previously created by Erin Landguth. All of these contributors therefore represent co-authors with affiliations listed on the title page of Chapter 5.

Given the collaborative nature of my PhD research as described above, I chose to narrate using the plural 'we' throughout the text.

Chapter 1

General introduction and literature review: population genetics and genomics in the study of Chagas disease and American visceral leishmaniasis

1.1 Population genomics as a tool to address *Trypanosoma cruzi* and *Leishmania infantum* epidemiology

Trypanosomatid parasites pose grave threat to life and livelihood in Latin America. Human infection by *Trypanosoma cruzi* leads to a wide range of cardiac and gastrointestinal disorders known as Chagas disease[1]. Severe forms claim an estimated 12,000 lives and $1.2 billion in lost productivity per year[2,3]. Up to 100 million people stand at risk of infection, especially the rural poor[2].

The related trypanosomatid *Leishmania infantum* also threatens impoverished communities with deadly visceral disease, parasitizing internal tissues and organs such as the bone marrow, liver and spleen[4]. Up to 6,800 cases of visceral leishmaniasis are estimated to occur each year in Latin America, mainly in Brazil, and case fatality rates lie between 10 and 20%[5].

Despite this major socioeconomic impact, the eco-epidemiology of *T. cruzi* and *L. infantum* is only very partially understood. Even some of the most basic biological properties, for example, the rate of sexual versus clonal reproduction, remain relatively obscure[6]. Such gaps in understanding come as no surprise. The elaborate life cycles of these vector-borne parasites are very difficult to study. Their microscopic size, inaccessible sites of (intra-cellular) infection and sensitivity to culture, for example, often inhibit direct observation of dispersal, biomedical properties and other critical life-history traits[7,8]. Unbiased sampling is not ethically possible in humans and also challenging in other cryptic, elusive or asymptomatic vectors and hosts.

Fortunately, however, information on unobserved trypanosomatid behavior is not forever lost. Wherever it goes, a lineage keeps a diary of its experiences in heritable genetic code. This allelic repertoire, when analyzed within and among individuals over space and time, unveils traces of population structure and its antecedents, for example, paths and barriers of dispersal, mechanisms and frequency of mating, local adaptation, or historical bottleneck and radiation events[7,9–11]. This concept lies at the core of population genetics, which has come to form a primary theoretical framework for trypanosomatid disease surveillance and control[12]. Unfortunately, however, blunt and unstandardized molecular and statistical tools have kept this framework far from complete[13–15].

1.2 Literature review synopsis

This dissertation exploits novel sequencing and computational approaches to help resolve major open questions about the ecology and evolution of important human parasites such as *T. cruzi* and *L. infantum* in Latin America. The following literature review first provides a brief description of Chagas disease and its burden to public health (Section 1.3.1), then highlights cornerstones of past *T. cruzi* population genetic research. Current understanding of intra-specific subdivision and lineage-associated geographic distributions, disease phenotypes and transmission ecologies are discussed (Sections 1.3.2 to 1.3.6). Section 1.3.7 describes theories about reproductive mechanisms and the frequency of genetic exchange in *T. cruzi*, a subject of ongoing debate[6]. This species has for decades been considered a paradigm of 'predominant clonal evolution'[16], but largely based on low-resolution genetic marker systems and questionable sampling designs[6]. Sections 1.4.1 and 1.4.2 then introduce visceral leishmaniasis and its non-endemic distribution in the New World (*L. infantum* is thought to have been introduced to the American continent during European colonization ca. 500 years ago[17–19]). Section 1.4.3 highlights gene and chromosomal copy number variation as key drivers of evolutionary adaptation in *Leishmania* parasites[20,21]. Section 1.4.4 describes genetic hybridization as another important source of genetic and phenotypic change in the genus (unlike in *T. cruzi*, meiotic sex has been experimentally proven to occur within and between *Leishmania* spp., but the relevance of genetic exchange to diversity patterns in natural *Leishmania* populations remains poorly described[22]). The aim of the literature review is also to highlight long-standing challenges in trypanosomatid population genetic inference. A number of these challenges are summarized in the penultimate section, after key advantages and prospects of whole-genome sequencing (WGS) studies have been highlighted in Section 1.5. These include the prospect of applying 'landscape genomics'[23] approaches to trypanosomatid research. Arthropod-borne parasite dispersal is sensitive to environmental heterogeneity[24,25], and a landscape genomics framework may contribute to the design of intervention strategies by clarifying how environmental conditions promote or inhibit the spread of disease into human populations from the wild. The final section recapitulates key knowledge gaps and research opportunities evident from the preceding review. These topics become the focus of (research) Chapters 2 – 5.

1.3.1 Chagas disease – a public health burden

Chagas disease has been considered the most important parasitic infection in Latin America[26]. Recent estimates indicate that ca. 10 million people are infected[27] by its etiological agent, *Trypanosoma cruzi*, a zoonotic kinetoplastid protozoan pathogen transmitted by more than 100 species of hematophagous triatomine vectors among an even greater number of domestic

and sylvatic mammalian hosts[28,29]. Also known as American trypanosomiasis, this disease begins when infective metacyclic trypomastigotes from infected triatomine feces enter host mucosal membranes, conjunctivae or abraded skin[30]. The full life cycle is described in Fig. 1.1. Oral outbreaks, congenital transmission and blood transfusions are important secondary, vector-independent routes of infection[15].

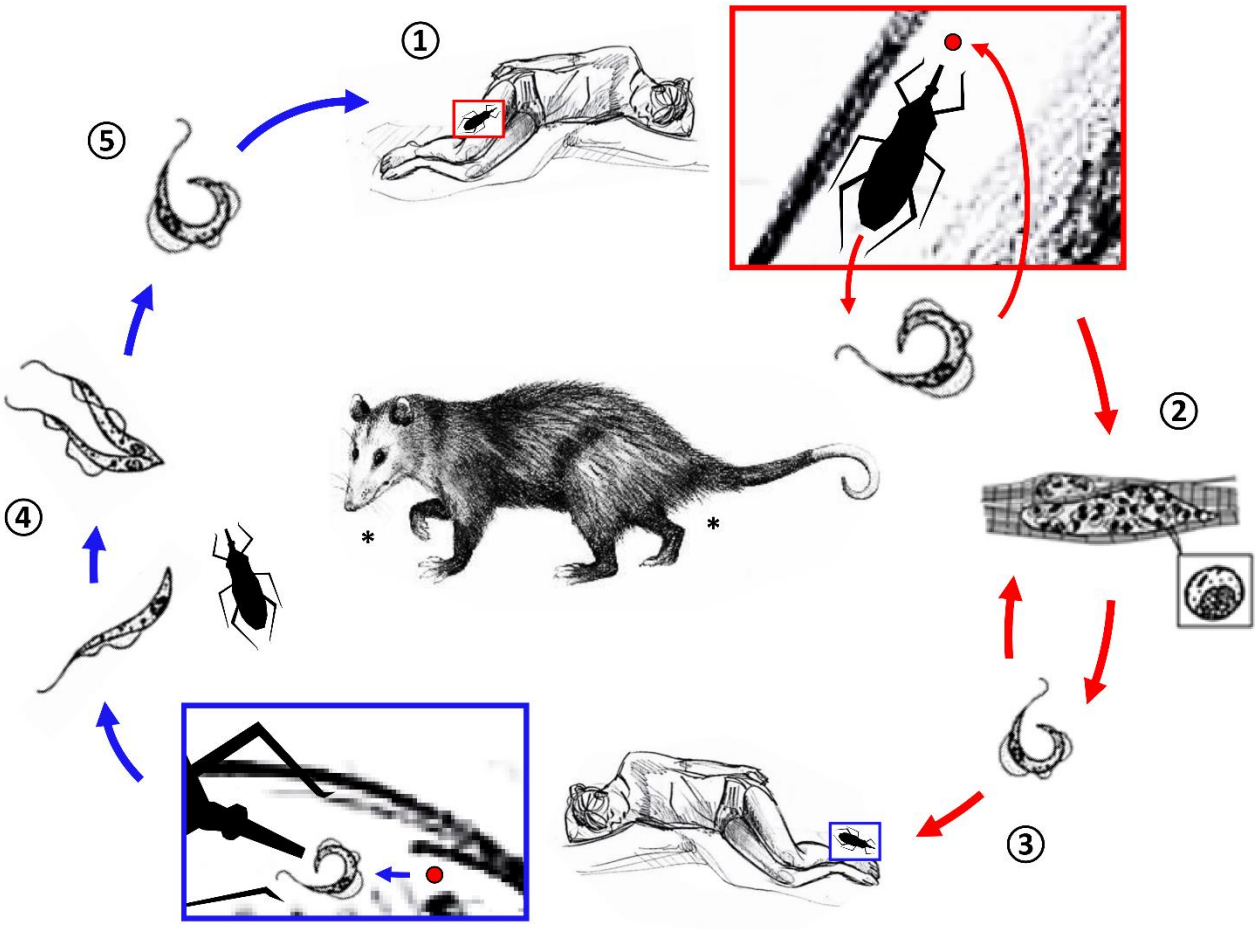


Figure 1.1 The *T. cruzi* life cycle – image modified from www.cdc.gov and descriptions based on Nagajyothi et al. (2012)[31]. 1) An infected triatomine releases metacyclic trypomastigotes in its urine and feces while feeding on host blood. These flagellated parasite stages enter the host at the site of the bloodmeal or through intact mucosa, especially the eyes. 2) Inside the host, the metacyclic trypomastigotes invade local cells and differentiate into amastigotes. These intra-cellular stages undergo multiple rounds of multiplication by binary fission before differentiating into trypomastigotes. The cell eventually ruptures and bloodstream trypomastigotes are released into circulation to infect new tissues, preferentially muscle and reticuloendothelial cells. 3) Bloodstream trypomastigotes can also be ingested during triatomine blood meal. 4) The ingested trypomastigotes differentiate into epimastigotes in the triatomine midgut and multiply. 5) A proportion of parasite cells periodically differentiates into metacyclic trypomastigotes in the hindgut and is released with the feces. The triatomine generally remains infective for life[32]. Vector stages are outlined in blue. Host stages are outlined in red. Blood meals are shown on a human host but countless other mammals can be infected[33]. Opossums, for example, are frequently infected by TcI (see intra-specific taxonomy in Section 1.3.2 and ecological relationships in Section 1.3.6). *Infection by predation also occurs in many species as well as via anal scent glands in opossums[33,34].

Following the acute (often subclinical) phase during which *T. cruzi* circulates at highest levels in the blood for up to three months, 30 – 40% of cases develop chronic infections characterized by various irreversible, potentially fatal cardiac, gastrointestinal and/or neurological syndromes[35]. Chagas disease accounts annually not only for ca. 12,000 deaths, but for at least 800,000 disability-associated life years and a global economic toll of ca. 1.2 billion USD[2,3,36]. Vaccinations do not yet exist and only two drugs, benznidazole and nifurtimox, are available for treatment[37]. Both of these nitroheterocyclic compounds can involve severe side effects (e.g., neurological disorders[38,39]) and typically fail in the chronic phase (success rates below 20%)[37]. Treatment is more effective in acute and/or pediatric patients (ca. 80% success rate) but is often inaccessible to the poor and rural communities where human infection prevails[37,40]. Chagas disease is for such reasons considered a 'neglected tropical disease' whose intervention requires much stronger support[41]. Investment decisions must recognize the widespread, poorly defined endemicity of *T. cruzi* in the wild and its multifarious routes to human infection[12,42]. In the face of global change, these features recommend heavy resource allocation to research on the transmission ecology and evolution of Chagas disease. If the biological and environmental variables shaping transmission pathways and/or promoting the emergence and transformation of epidemiologically relevant phenotypes (e.g., more drug-resistant and virulent parasite populations) are identified, this likely indelible zoonosis may become more anticipatable and therefore fencible in its spread. Population genetics, genomics and landscape genomics (i.e., the formal unison of landscape ecology and population genomics) are core to reaching the understanding required.

1.3.2 Genetic Subdivision within *T. cruzi*

T. cruzi population genetics is an area of active research. Basic ancestral relationships[43,44] remain disputed and early dogma[45] about the (in-) frequency of genetic exchange might soon need to be replaced[6,14,46,47]. There is broad consensus around the subdivision of *T. cruzi* into six distinct lineages, so-called discrete typing units (DTUs), now numbered TcI through TcVI, and referred to as TcI, TcIIb, TcIIc, TcIIa, TcIId and TcIIe, respectively, prior to 2012[13]. These DTUs are defined as 'sets of stocks that are genetically more related to each other than to any other stock and that are identifiable by common genetic, molecular or immunological markers'[48].

Numerous genetic markers have been applied to define genetic diversity in *T. cruzi* (Tbl. 1.1). Analogous results from disparate typing methods generally substantiate the six-DTU subdivision, but there is a lack of consensus around what markers to use and many typing assays are not DTU-specific (e.g., differentiating across, but not within, zymodemes) and/or do not yield repeatable results. A very commonly used mini-exon size polymorphism typing

assay, for example, does not distinguish among TcII, TcV and TcVI and only sometimes distinguishes TcII from TcIV[49–51]. An equally popular size polymorphism assay at the 24Sα rRNA locus does not distinguish, e.g., TcI from TcIII or TcII from TcVI, and amplicons for TcIII and TcIV appear to vary based on geographic origin[50,52,53]. There is also suspicion that additional lineages (e.g., a more recently identified bat-associated lineage known as TcBat[54]) have been neglected (e.g., classified as TcI) by sparse marker sets of the past[55].

Table 1.1 Molecular markers used to distinguish *T. cruzi* DTUs – based on Messenger et al. (2015)[15].

DTU-assignment method	Example of genetic markers
Multi-locus enzyme electrophoresis	ASAT, ALAT, PGM, ACON, MPI, ADH, MDH, ICD, 6PGD, G6PD, PEP, GPI
Restriction fragment polymorphism analysis	HSP60, GPI, COII, GP72, 1F8, Histone H3, ITS, TcSC5D, mHVR
Multi-locus sequence typing of nuclear genes	TcMSH2, DHFR-TS, TR, LYT1, Met-II, Met-III, TcAPX, TcGPX, TcMPX, GPI, HMCOAR, PDH, GTP, STTP2, RHO1
Multi-locus sequence typing of kinetoplast genes	12S rRNA, 9S rRNA, cytB, MURF1, ND1, COII, ND4, ND5, ND7, mHVR
Fluorescent fragment length barcoding	28Sα rRNA, 18S rRNA
Multi-locus microsatellite typing	10101$(CA)_a$, 11283$(TA)_b$, 7093$(TA)_b$, TcUn4, mclf10, 10359(CA), 10187(TTA)
High-resolution melting analysis	SL-IR, 24Sα rRNA
Single-stranded conformation polymorphism analysis	18S rRNA, cruzipain, P7-P8
Amplicon sequencing	TcGP63, ND5, 18S rRNA, SL-IR
Size polymorphism analysis of multi-copy genetic markers and minicircle sequences*	SL-IR, 24Sα rRNA, 18S rRNA, A10, P7-P8, mHVR*
Pulsed field gel electrophoresis (and hybridization to labelled probes)	Chromosomal bands (1F8, cruzipain, FFAg6, Tc2, P19)
Random amplification of polymorphic DNA	None (no prior sequence info. needed)

Several authors have therefore sifted through previous typing systems to work out multi-step, multi-marker PCR-based protocols or identify targets for Sanger sequencing that most efficiently and accurately discriminate all DTUs. These typically suggest the analysis of multiple different single-copy genes (to limit disorientation by hybrid and repeat-rich genotypes), e.g., a size polymorphism triple-assay of heat shock protein 60, glucose-6-phosphate isomerase (GPI) and LSU rDNA gene fragments or multi-locus sequence typing (MLST) of house-keeping genes such as C-5 sterol desaturase, rho-like GTP-binding protein,

mitochondrial peroxidase, 3-hydroxy-3-methylglutaryl-CoA reductase and GPI[56–58]. MLST is advantageous also because results are easily shared in online archives and can be extended for high-resolution intra-lineage analysis otherwise performed via multi-locus microsatellite typing (MLMT)[56,58], results of which cannot be shared systematically among labs.

1.3.3 Geographic distribution of DTUs

Heavy past emphasis on lineage assignment based on the six-DTU framework have helped chart *T. cruzi* phylogeography across much of the American continent. A recent meta-analysis shows that over six thousand strains (from 137 different publications) have been classified to the DTU level[55]. TcI clearly appears to be the most widely dispersed DTU. It is detected throughout the range of its six most important vector genera (see Section 1.3.6), as far north as California and south into northern Chile and Argentina. Human TcI infections appear to predominate in Central and northern South America but are less frequently detected south of the Amazon basin. TcII is most often reported from the southern and central regions of South America, extending northward primarily along the Atlantic Forest of Brazil[59–61]. TcIII appears to be most common in Bolivia, Paraguay and Brazil. It is rarely found in human hosts. TcIV accompanies TcIII in Amazonia and is found with TcI in northern South America as well as in the southern United States. Distributions of TcV and TcVI are also thought to overlap considerably in the Gran Chaco and spread into the Southern Cone, though TcV may reach farther southwest and TcVI may take more into southern Brazil[13,55,62]. TcV and TcVI are also occasionally reported from the North of South America (e.g., Colombia), most likely due to long-range anthropogenic importation events[63].

Our understanding of DTU phylogeography is incomplete (especially for TcII) due to limited and highly patchy geographic sampling coverage (e.g., just two DTU-assigned strains from Panama vs. hundreds from Colombia in meta-analysis by Brenière et al. (2016)[55]) and various other forms of bias, e.g., less frequent sampling from sylvatic ecotopes (more than two thirds of DTU-assigned strains appear to represent domestic or peri-domestic environments[55]) or from elusive non-human hosts. Differential tissue tropism and mixed-strain infections (see Sections 1.3.5 and 1.6) also jeopardize representative sampling and culture within and across DTUs.

1.3.4 Phylogenetic ancestry among DTUs

Enzymatic, genetic and – as far as available – genomic analyses (see current reference assemblies at https://www.ncbi.nlm.nih.gov/genome/genomes) consistently suggest that TcI and TcII represent ancestral lineages, i.e., are not derived from other DTUs[52,64–68]. A number of individual (GPI, COII-ND1, TR)[69,70] and concatenated marker sets[68,71] have yielded similar

divergence time estimates based on Bayesian evolutionary analysis by sampling trees (BEAST), suggesting that the common ancestor of TcI and TcII occurred less than four million years ago. It has also become widely accepted that TcV and TcVI are the hybrid progeny of the ancestors of TcII and TcIII[43,69,71–76]. TcV and TcVI are the most similar of all DTUs and the TcVI ('CL Brener') reference genome comprises two divergent haplotypes, one highly similar to TcII (i.e., the 'Esmeraldo-like' haplotype) and the other highly similar to TcIII[77]. Despite their high genotypic similarity, TcV and TcVI most likely derive from two separate ancestral TcII/TcIII hybridization events because Esmeraldo-like and non-Esmeraldo-like haplotypes of TcV typically cluster closer to TcII and TcIII (respectively) than to correspondent haplotypes of TcVI[70,71,69,76]. While these points on DTU ancestry attract relatively little debate, details surrounding the speciation of TcIII and TcIV have been subject to controversy for several years. A central question has been whether an early hybridization between the ancestors of TcI and TcII produced a lineage that diverged into TcIII and TcIV prior to the hybridization(s) of TcII and TcIII that produced TcV and TcVI (the 'two-hybridization' model[43]) as illustrated in Fig. 1.2a. The alternative 'three-ancestor' model[44] (see Fig. 1.2b) suggests that no other nuclear hybridizations occurred prior to those that most recently led to TcV and TcVI. A related model by Tomasini and Diosque[76] (see Fig. 1.2c) elaborates that ancestral TcIV diverged into separate North ($TcIV_N$) and South American ($TcIV_S$) lineages during the Great American Interchange[78] and that mitochondrial introgression occurred several times from ancestors of $TcIV_S$ to ancestors of TcIII. Tomasini and Diosque also suggest that TcI, TcIII, and TcIV form a private monophyletic group[76]. The three-ancestor model, by contrast, does not define TcIV ancestry and depicts TcIII as an independent (i.e., non-nested) ancestral strain. Proponents of the two-hybridization model have emphasized the presence of mosaic markers in TcIII and TcIV, i.e., mutations shared between TcI and TcIII or TcIV and between TcII and TcIII or TcIV at different positions of the same allele[43,79]. Mosaicism has also been observed across markers, i.e., phylogenies built from some markers showing TcIII cluster with TcI[43,80–82] and those built from other markers showing TcIII cluster with TcII[43,66,67,74]. The separation of TcIII and/or TcIV from TcII towards TcI, however, predominates in analyses based on concatenated marker sets and was also inferred from recent genomic assembly of TcIII ('231') and comparisons to TcVI and TcI-Sylvio genomes[68]. Furthermore, analysis of the TcVI genome finds evidence for mosaicism in less than one percent of core regions in the non-Esmeraldo-like haplotype[83,84]. Proponents of the three-ancestor model[44] and that of Tomasini and Diosque[76] suggest that intermittent base or sequence similarities between TcI and TcII have too rarely been confirmed as synapomorphies based on outgroups that help identify ancestral states[71,76], with Tomasini and Diosque also having illustrated how inference changes when *T. c. marinkellei*

sequences are added to alignments previously used to establish the two-hybridization model by Westenberger et al. (2005)[76]. With the additional outgroup included, nucleotide positions where TcI-like TcIII or TcIV sequences become TcII-like appear to be homoplasies in TcI rather than symptoms of mixed ancestry due to previous hybridization between TcI and TcII[76]. The model by Tomasini and Diosque, however, is not supported by recent sequence analysis of satellite DNA[85] and other authors suggest that GPI sequences from TcIII/TcIV resemble TcII/III mosaics also with outgroups included in analysis[69]. Like the three-ancestor model, which did not include any samples of TcIV, the model by Tomasini and Diosque cannot be considered stable without better representation of *T. cruzi* diversity within genomes and DTUs, especially TcIV. The TcIV genome has yet to be sequenced or assembled and most studies have used three or less reference sequences to represent this DTU[43,70,76]. Limited character and taxon sampling is well known to mislead phylogenetic inference[86–88] and therefore must be ameliorated to clarify theory. It may also be helpful to reconsider the use of standard (bifurcating) tree construction when speciation is thought to involve introgression and genome-wide hybridization events[70,76,89]. Modifications to classical phylogenetic analyses (e.g., network models[90,91]) may help resolve this issue pending more regular use of WGS methods that explicitly account for lineage sorting, e.g., by comprehensively quantifying ancestry contributions (i.e., for each base in the genome) or sliding-window network construction across chromosomes[92,93].

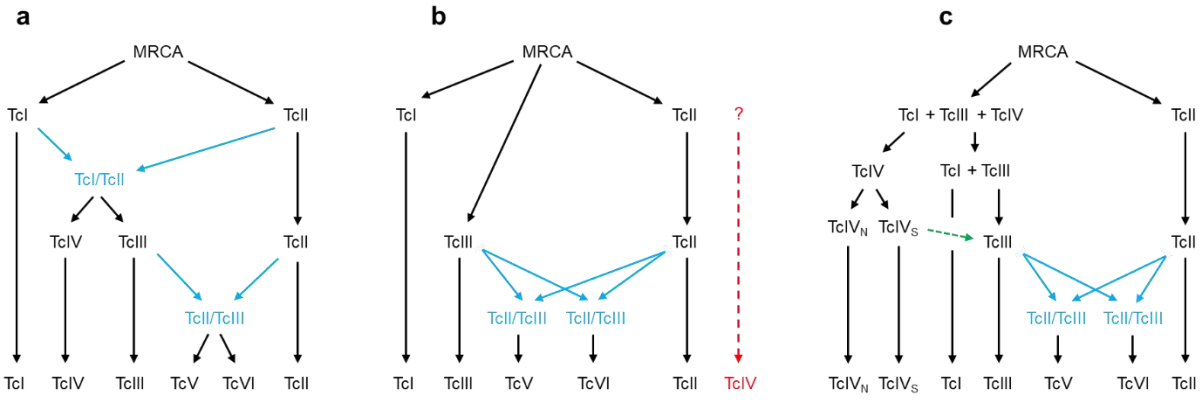


Figure 1.2 Three major models of DTU speciation. **a** The two-hybridization model[43] suggests one ancient genetic exchange event between ancestors of TcI and TcII (and subsequent loss of heterozygosity) leading to TcIII and TcIV as well as a more recent hybridization event between ancestors of TcII and TcIII leading to TcV and TcVI. **b** The three-ancestor model[44] suggests two recent hybridization events between ancestors of TcII and TcIII leading to TcV and TcVI without participation of TcI[13]. **c** A variation of the three-ancestor model by Tomasini and Diosque shows TcII diverging from all other DTUs before these diverged from TcI[76]. Tomasini and Diosque also specify recurrent mitochondrial introgression from ancestors of TcIV$_S$ to those of TcIII (green arrow). MRCA abbreviates most recent common ancestor.

1.3.5 DTU-specific pathologies

A comprehensive meta-analysis conducted by Messenger et al. in 2015 suggests that clear evidence for an association between *T. cruzi* genotype and key disease phenotypes (chronic morbidity, risk of reactivation, congenital or oral transmission) does not yet exist[15]. The review details how variable methods and quality of clinical characterization, stage classification and lineage typing have encumbered progress on linking genotypes with phenotypes to enhance the relevance of the six-DTU framework in the medical field. For example, studies have often examined only cardiac (not gastrointestinal) tissues and rarely assessed whether indeterminate infections later turned symptomatic. Prior to 2002, *T. cruzi* lineages were often typed using unstandardized multi-locus enzyme electrophoresis (MLEE) protocols without validation from other markers and various mistakes in nomenclature have occurred (see references in Messenger et al. (2015)[15]). Conventional *T. cruzi* sampling methods are also very prone to selection bias. Only few tissue types are assessable via biopsy, and parasites are typically isolated by hemoculture or xenodiagnosis. Clones isolated from the blood are often genetically different from those sequestered in tissues[94–96] and these differences may be non-random, i.e., reflect differential tropism within and among DTUs (e.g., TcI in the esophagus and TcII in the heart of a single patient[97]), host genetics, or immune state. It is possible that distinct subpopulations or constellations of subpopulations govern disease outcomes and these are unlikely to be represented accurately in the blood. Parasite genotypes have also been shown to vary across sequential blood samples[98] and xenodiagnosis is affected by the permissivity of the vector individuals or species applied[99–101]. Selection bias continues when isolated parasites are brought to the laboratory to enrich or separate ('clone') cells for further study due to variable growth rates, starting concentrations or sensitivity to culture and handling[102]. Diversity typically decreases over time[103].

Despite the above caveats and little transition so far towards culture-free genotyping techniques, some general associations between DTUs and disease phenotypes have been advanced in the literature. These associations largely track the geographic distributions of the different DTUs (Section 1.3.3) and thus might be argued to affirm the importance of parasite genetic variation in determining clinical outcomes because no ethnic or human genetic patterns are apparent across this range[104]. TcII is considered the primary agent of severe acute and chronic Chagas disease in central and southern Brazil, where it is also frequently associated with megacolon and megaesophagus without the detection of other DTUs[51,105,106]. TcII is also involved in human infections in Bolivia and the Southern Cone, but patients often appear co-infected with TcV or TcVI, and these DTUs are frequently detected alone (without TcII) in severe cases of disease[107–110]. Interestingly, TcV is also linked with congenital transmission in Argentina, Bolivia and southern Brazil because rates of congenital

transmission reach up to twelve percent in areas where TcV predominates but remain below one percent in areas associated to TcII[104,111–113]. Similar argument is also used to suggest that TcI does not cause digestive syndromes as these are rarely found in the Amazon and in northern South America where this DTU prevails[114]. Instead, TcI is associated with chagasic cardiomyopathy and, although long considered more benign than other DTUs, is increasingly associated with severe forms of disease in Venezuela and Colombia, occasionally also in the Southern Cone[96,115–118]. In contrast to TcI, TcII, TcV and TcVI, relatively little has been proposed about the clinical associations of TcIII and TcIV. These DTUs appear to be most common in sylvatic ecotypes (see next section) and may thus be less relevant to human disease. Nevertheless, TcIV has been involved in severe (including lethal) cases of foodborne transmission in Colombia and Brazil[119–121]. It is unclear whether these events reflect an intrinsic propensity toward oral transmission and/or acute symptomology by the parasite or food and living practices in rural areas where this DTU occurs. Fatal cases are also known from TcI[122].

Much remains to be done to verify the above associations and explain why different DTUs might cause different forms of disease. As recently reviewed by Jiménez et al. (2019)[123], a number of studies point to DTU-specific recognition by the immune system, and therefore, DTU-specific (dysregulation of the) inflammatory response. It will be key to pursue such hypotheses with more standardized clinical descriptions and methods that better apprehend multiple (tissue-specific) genotypes occurring within single hosts. Previous success in genotyping *T. cruzi* directly from infected tissue (e.g., via low-stringency single-specific primer (LSSP) PCR fingerprinting[106], rDNA qPCR[105], kDNA restriction fragment length polymorphism (RFLP)[124] or nested microsatellite analysis[95]) has generally involved a tradeoff in which high sensitivity at a small set of markers is favored over amenability to further sequence analysis within and across DTUs. Creating efficient sequence-based approaches like MLST for use on uncultured samples would help detail and more systematically document relationships between infection diversity and disease phenotypes.

1.3.6 DTU-specific transmission cycles

Numerous ecological specificities such as vector/host species, climate and vegetation type (even stratum, e.g., arboreal vs. terrestrial) have been designated to the DTUs. Notably, TcII, TcV and TcVI seem to rarely occur in sylvatic transmission cycles and predominate in *Triatoma infestans*, whereas TcI features in domestic and wild cycles (in the lowland tropics as well as in arid environments), at least six genera of triatomine vector (*Rhodnius*, *Triatoma*, *Panstrongylus* and *Eratyrus*, *Mepraia* and *Dipetalogaster*) and dozens of genera of mammalian hosts[13,15,62]. TcI is very common in Didelphimorphia (e.g., 262 of 509 cases

summarized by Brenière et al. (2016)[55]), especially in *Didelphis marsupialis*, and range overlap with the anthropophilic vector *Rhodnius prolixus* is considered an important driver of human disease in northern South America[125]. Didelphid opossums have been shown to tolerate high parasitaemia by TcI but rapidly suppress infections by TcII[126]. TcI is also frequently reported from rodents, bats, carnivores and primates[55]. Infections in Artiodactyla, Pilosa and Xenarthra are only sporadically found[55]. TcIII and TcIV share a variety of hosts with TcI but are much more rarely detected in domestic cycles[15,127–129]. TcIII appears to thrive in areas where terrestrial vectors (e.g., *Panstrongylus* and *Triatoma* spp.) interact with fossorial hosts. *Dasypus noveminctus* infection is especially common and armadillos have been proposed to have facilitated the emergence and spread of ancestral TcII/TcIII hybrids into domestic cycles of the Southern Cone[51,129]. Other TcIII hosts include caviomorph rodents, bats, coatis, opossums and carnivores[33,130]. TcIV is also found frequently in Cingulata as is TcIII but has also been linked to arboreal cycles and palm-associated vectors such as *Rhodnius robustus*, *R. pictipes* and *R. brethesi* in the Amazon[121]. Hosts include arboreal (e.g., howler monkeys, *Marmosa* opossums, rodents such as *Oecomys mamorae*) bus also terrestrial (armadillos, rodents such as *Proechimys* spp., opossums such as *D. brevicaudata*) and semi-terrestrial (e.g., coatis, *Philander* opossums, various bats) mammals in diverse biomes, e.g., Pantanal, Caatinga, Atlantic Forest of Brazil[33,130]. In the United States, $TcIV_N$ is reported from raccoons and domestic dogs[131,132].

Several studies have also attempted to define associations between genotypes and transmission cycles at the sub-lineage level, particularly within TcI, the most ecologically eclectic and genetically diverse DTU. Again, a key focus has been placed on diversity in sylvatic vs. domestic groups[14,56,133–141]. Parasite population genetic differentiation between these environments is of applied interest because it illuminates rates of parasite domiciliation from the wild (e.g., before/after intervention measures and awareness-building) or parasite genetic traits and vector associations that increase fitness in the domestic niche. Like at the inter-DTU level, intra-DTU genetic discontinuity between sylvatic and domestic populations may also reflect ancient divergence into different transmission cycles and/or co-evolution with associated vectors and hosts. Similar isoenzyme profiles and phenotypes were noted early among widely dispersed (i.e., > 100 km) domestic and peri-domestic strains (e.g., see Widmer et al. (1985)[133] and Saravia et al. (1987)[134]), and significant genetic follow-up studies began in 2007. Herrera et al.[135] detected a domestic 'haplotype 1' (later referred to as TcIa) in distantly separated Colombian departments (Magdalena, Caquetá, and Boyacá) based on single-nucleotide and insertion-deletion polymorphism in the non-transcribed spacer region of the mini-exon gene (SL-IR). This haplotype was associated with the domestic cycle and the vector *R. prolixus*. The study also found a 'haplotype 2 (later TcIb) associated with

domestic and peri-domestic cycles and the vector *Triatoma dimidiata*, a 'haplotype 3' (later TcIc) associated with the peri-domestic cycle, and 'haplotype 4' (later TcId) associated with sylvatic transmission. Haplotypes TcIa and TcId were also identified beyond the borders of Colombia when Cura et al. (2010)[137] expanded SL-IR analysis to 105 isolates from eleven countries between the United States and Argentina, but TcId showed no clear affinity to specific ecotopes, occurring in various transmission settings in Colombia and Argentina, sylvatic cycles in Brazil and human patients from French Guiana, Venezuela and Panama. TcIa, however, showed a very clear pattern, remaining strictly associated to domestic cycles throughout South America and becoming closely linked to sylvatic cycles at Central and North American sites. This study was published just after that of Llewellyn at al. in 2009[56] which also examined TcI diversity throughout the endemic range. The 48-marker microsatellite panel applied to 135 samples in this study exposed extraordinary parasite genetic diversity across South America and differentiation that correlated with geographic distance, but one important exception was observed. Domestic samples from eleven different states of Venezuela appeared highly similar to another and were clearly more closely related to Central and North American than to Venezuelan sylvatic strains. Several high-resolution mitochondrial and nuclear MLST/MLMT studies[14,139,142] followed to show that this 'VEN_{Dom}' group corresponded to the previously indicated TcIa SL-IR genotype and Zumaya-Estrada et al. (2012)[139] advanced the hypothesis that this lineage (renamed TcI_{DOM}) likely broke through an ancient transmission bottleneck in North America and accompanied human migration into South America within the last 23,000 years. Highly inefficient stercorarian transmission (i.e., > 900 bloodmeals before successful human infection[143]) relative to rates of congenital transmission (e.g., 58% in BALB/c mice[144]) and long-distance anthropogenic dispersal were suggested to have helped TcI_{DOM} perpetuate in domestic settings with little admixture from sylvatic parasite diversity even in areas where infected triatomines frequently enter from the wild[56]. Nevertheless, previous hypotheses of TcI_{DOM} emergence due to adaptive changes in epidemiologically relevant genes[136,138] deserve further study as some important biological differences have been observed relative to sylvatic genotypes. For example, Cruz et al. (2015)[145] observed lower levels of histopathological damage by TcI_{DOM} than by sympatric sylvatic strains in mice and several studies have suggested higher bloodstream parasitaemia by TcI_{DOM} in chronic cases of human disease[96,116,146].

Apart from the domestic-sylvatic interface, a number of high-resolution multi-marker studies have also focused on possible (mechanisms of) substructure within sylvatic TcI[147–151]. Notable among these was a powerful MLST/MLMT analysis by Messenger et al. (2015)[149] from Bolivia that described limited TcI gene flow between nearby arboreal and terrestrial transmission cycles in contrast to low genetic subdivision (F_{ST}) among parasites from similar

ecotopes at much more distant sampling sites. These results supported ecological host fitting as a predominant mechanism of *T. cruzi* diversification, not only within TcI but also in regard to niche-associated inter-lineage speciation patterns described above. Ecological fitting describes a 'process whereby organisms colonize and persist in novel environments, use novel resources or form novel associations with other species' by re-tooling existing trait repertoires rather than through *de novo* adaptation (positive selection) after contact. This process also been proposed to explain host ranges among different trypanosomatid species[152] and even to have facilitated first transitions to parasitism in the free-living (bodonid) relatives of the Trypanosomatidae[153].

1.3.7 Reproduction

Parasite reproductive mode is central to epidemiology because it determines how parasite diversity distributes and changes in space and time. Genetic exchange can create important new genetic combinations or transfer these among divergent strains, for example, it has been shown to increase vector transmissibility, parasitaemia and phenotypic plasticity in *Leishmania*[154,155] and to confer human infectivity to previously non-infective subspecies of *Trypanosoma brucei*[156]. Genetic exchange also accelerates diversification, and can thereby help parasites evade the immune system[157], adapt to environmental change[158] or outpace drug design[159]. Clonality, on the other hand, implies that population genetic subdivisions are stable and that genomes decay over time. Rates of divergence and dispersal become predictable and simple marker systems may suffice to track outbreaks or guide treatment of human disease[160]. Although *T. cruzi* primarily uses clonal reproduction, the holocenic expansion of virulent inter-lineage hybrids (TcV and TcVI) into domestic cycles[69] makes it clear that genetic exchange is also pivotal to its speciation and the long-term evolution of Chagas disease. Where, how, and how much genetic exchange occurs in contemporary populations, however, remains incompletely understood and has attracted decades of debate[6].

For many years genetic exchange was considered too rare to be relevant to contemporary variation in *T. cruzi* diversity and population structure. A theory known as 'predominant clonal evolution' (PCE) came to dominate the literature as first reports of strong linkage disequilibrium at multi-locus enzyme electrophoresis (MLEE) loci from the 1980's[45,161] were substantiated by linkage among independent markers sets ('criterion g' in Tibayrenc et al. (1990)[162]), e.g., between MLEE and randomly amplified polymorphic DNA (RAPD)[64,67], between microsatellites and rDNA RFLP[163] and among MLST, MLEE and RAPD[164]. The perseverance of the DTU framework was also emphasized as evidence that recombination is only meaningful at the macroevolutionary scale[165], and it was proposed that stable inter-lineage divisions are mirrored within each DTU[166]. Evidence for these so-called 'Russian doll

patterns' (RDP)[166], however, remains scarce. The model was introduced[166] (and remains[16]) based primarily on dispersed TcI substructures TcI_{DOM} and TcId (see Section 1.3.6)[135,137,167] without taking effects of ancient bottlenecks through domestic/sylvatic subdivisions or convergent host fitting processes into account. The authors also do not address reticulate phylogenetic structures in data suggested to evince RDP in Ramirez et al. (2012)[16,166,167] and are not deterred by various studies[140,142,147,168] that contradict the model in TcI[16].

A number of authors have considered past observations inadequate to quantify the relevance of genetic exchange within DTUs, calling for genetic marker coverage to be extended and spatial sampling designs corrected (e.g., to avoid Wahlund effects[169]) for an accurate representation of *T. cruzi*'s reproductive mode or modes[6,14,170]. This view is also inspired by laboratory work by Gaunt et al. in 2003[171] which demonstrates that *T. cruzi* has an extant capacity for genetic exchange. The authors transfected putative parental TcI isolates from Carrasco et al. (1996)[172] with recombinant plasmids conferring resistance to either neomycin or hygromycin B and then co-passaged these through mammalian (Vero) cell cultures and *in vivo* in mice and triatomines. Six clones from Vero cell culture (but none from mice or triatomines) survived double drug selection and were confirmed to be intra-lineage recombinants by MLEE, karyotyping, microsatellite analysis and nucleotide sequencing of housekeeping genes. Surprisingly, however, the recombinants had inherited both parental alleles at most nuclear loci in what appeared to have been a non-meiotic genome fusion event. The authors suggested a mechanism similar to that known from pathogenic fungi whereby diploid genomes fuse to form tetraploid offspring and concerted chromosome loss gradually brings these tetraploids back to the diploid state[173]. Consistent with this hypothesis, follow-up flow cytometric analyses in by Lewis et al.[174] showed that nuclear DNA content in the six hybrids had reduced by ca. 15% by 2009, and this sub-tetraploid state was shown to remain stable during various forms of stress. The authors also examined DNA content in the natural hybrids TcV and TcVI and both appeared to be fully diploid. Lewis et al. (2019) noted the possibility of complete erosion of tetraploidy but also that heterozygosity patterns in TcV and TcVI are more consistent with meiotic than with parasexual origin because random post-fusion chromosome losses are expected to generate non-recombinant (homozygous) genotypes in approximately one third of the genome[174].

Various authors have therefore set out in search of reproductive phenomena and further evidence for/against parasexuality in the field, many also shifting study focus to finer spatial scales. Evidence for nuclear genetic exchange is accumulating from such studies in the form of local Hardy-Weinberg allele frequencies, linkage equilibrium between loci and a lack of repeated multi-locus genotypes[140,147,168,175,176]. Messenger et al. (2015) also point to dissimilar heterozygosity estimates between TcI populations in Bolivia as a possible indicator of recent

hybrid origin in some strains or different mating systems in different ecotopes[149]. Furthermore, several studies demonstrate phylogenetic incongruence between nuclear and maxicircle sequences, suggesting that genetic exchange can (additionally or exclusively) involve the transfer of mitochondrial DNA[14,139,142,148,149,177]. Some authors demonstrate that mitochondrial introgression can be very frequent (e.g., Ramirez et al. (2012) detected 17 introgression events among 100 clones[14]), perhaps even significantly more common than nuclear genetic exchange[170]. Ploidy or allele frequency patterns consistent with genome fusion as in Gaunt et al. (2003)[171], however, have not surfaced in TcI populations from the field. Recent genomic analysis did find extensive aneuploidy in TcII isolates but with no further evidence as to whether karyotypes reflected non-meiotic reproductive histories or mitotic amplifications from stress[178]. The latter is not uncommon in eukaryotic microbes, e.g., in *Saccharomyces*[179] or *Leishmania* spp.[180].

In light of growing evidence of contemporary genetic exchange, the PCE model has been refitted several times since its first announcement in 1986[45,10,166,16]. Tibayrenc et al. (2015)[165] recently suggested, for example, that 'it is quite possible that genetically related strains undergo more genetic exchange than clonal propagation'. Nevertheless, these authors have remained relatively hostile towards most new evidence of intra-lineage recombination (e.g., see exchanges with Ramirez and Llewellyn[6,46,47] or response to work on *T. congolense*[181–184]) and frequently discard evidence of Hardy-Weinberg equilibrium as type II error (i.e., the inability to reject the null hypothesis of panmixia)[165] or suggest that mito-nuclear incongruences reflect disparate evolutionary pressures and/or mutation rates[166].

Strategic, high-intensity surveys of genome-wide (mitochondrial and nuclear) polymorphism among sympatric *T. cruzi* individuals are therefore key to resolving this debate. As sympatry is not a simple concept in this species, it will be important to design these surveys such that the possibility of recombination can be examined not only between isolates from different vector/host individuals but also between parasite clones from the same infection source. It may also be helpful to target 'potential hybridization zones'[6,149] and generally to return to places where genetic exchange has already been suggested to occur, e.g., in rural areas of Loja Province, Ecuador[140] or in undisturbed enzootic cycles of the Amazon, the approach taken by Gaunt et al. (2003)[171,172].

1.4.1 Visceral leishmaniasis – a public health burden

Visceral leishmaniasis follows malaria as the world's second deadliest parasitic infection[185] and its global economic impact ranks near that of Chagas disease[186]. Prevalence is highest in East Africa and on the Indian subcontinent but is also significant in Brazil, where over 50,000 cases have been recorded since 2001[187]. Less than 10% of cases appear to occur in other Latin American countries, but gaps in surveillance and reporting across the continent keep true rates of infection unclear. This vector-borne zoonosis is caused by the trypanosomatid parasite *Leishmania donovani* in Asia and East Africa and by its closely-related congener *L. infantum* in the Americas, North Africa and Europe. Infection occurs when *Phlebotomus* (Old World) or *Lutzomyia* (New World) sandflies feed on vertebrate blood and infective (promastigote) parasite stages within the saliva invade and replicate (as amastigotes) in host macrophages and other mononuclear phagocytic cells, especially in the bone marrow, liver and spleen (see life cycle in Fig. 1.3). Although symptoms are not always overt and the incubation period can last from weeks to several months, the human host generally dies within two years of infection without treatment[4]. It is therefore all the more cruel that visceral leishmaniasis, like Chagas disease, prevails in regions where disease awareness is limited and public health infrastructure is absent or frail[188]. Even when acknowledged and accessible, anti-leishmanial drugs are expensive (costs often exceed household income[189]) and not consistently effective (due also to the evolution of drug resistance[190]) or safe (systemic antimonial treatment, for example, can have lethal side effects (severe nephro- and cardiotoxicity, etc.) but remains a drug of choice in Latin America due to higher costs of less toxic liposomal amphotericin B[191]). The zoonotic nature of visceral leishmaniasis caused by *L. infantum* complicates the situation. Unlike the anthroponotic transmission cycles typical of *L. donovani*, the transmission of *L. infantum* is thought to rely heavily on intermediate hosts, particularly on domestic dogs, the only primary reservoir confirmed for Brazil[192]. Human treatment alone is therefore unlikely to protect public health unless an economic (mass-administrable) vaccine is found. No human vaccine has yet been approved. A number of canine vaccines, however, are becoming available and are widely recommended over dog culling approaches used to date in Brazil[193,194]. Future design of vaccines and drugs needs to consider how parasite diversity and abundance is spatially distributed and changes over time. Without such population genetic understanding, vaccines may confer incomplete immunity (i.e., only against a subset of genotypes) or drugs may fail when parasites show unexpected polymorphism or exploit alternative metabolic paths.

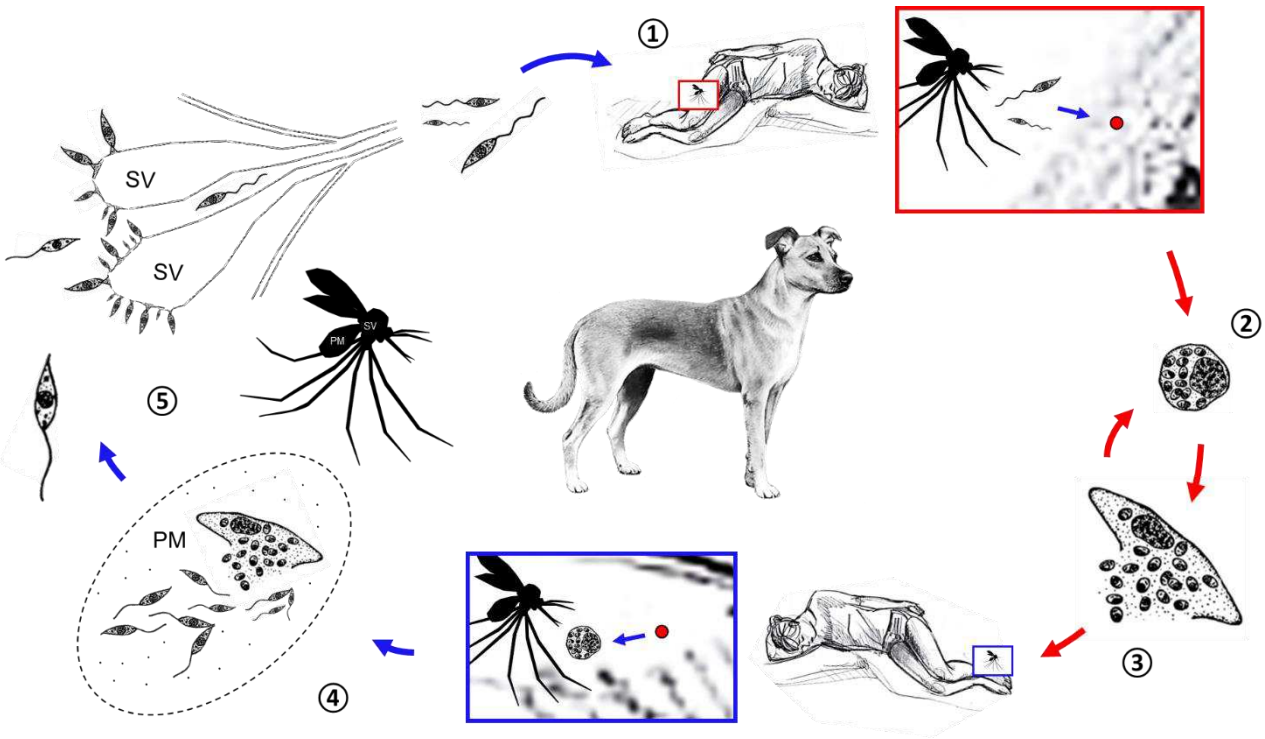


Figure 1.3 The *L. infantum* life cycle – image modified from www.cdc.gov and descriptions based on Sadlova et al. (2017)[195]. 1) An infected sand fly releases metacyclic promastigotes through the proboscis while feeding on host blood. 2) The promastigotes are phagocytized by macrophages and other types of mononuclear phagocytic cells. 2) They then differentiate into amastigotes and multiply by binary fission within the phagolysosome. 3) The cell eventually ruptures, and the amastigotes are released into circulation. They become phagocytized by other cells and begin to accumulate in the deep organs of the reticuloendothelial system, e.g., in the lymph nodes, bone marrow, liver and spleen. Infected macrophages can also be ingested when a sand fly take a blood meal. 4) Amastigotes within the ingested macrophages differentiate into procyclic promastigotes and multiply within the peritrophic matrix (PM), a chitinous envelope secreted by midgut epithelial cells. This envelope degenerates within three days and the parasites are released as nectomonads. 5) The nectomonads migrate towards the thoracic midgut, where they multiply as leptomonads and later haptomonads that colonize the stomodeal valve (SV). Metacyclic promastigotes also form. Blocking and damage to the valve by the haptomonads facilitates the release of the metacyclic promastigotes through the proboscis into the host. The sand fly generally remains infective for life[196]. Vector stages are outlined in blue. Host stages are outlined in red. Blood meals are illustrated on a human host but the domestic dog is the primary reservoir of *L. infantum* in the Americas[197].

Vaccine/drug specificity and diversity also affect the risk of resistance evolution and can only be chosen correctly if parasite genetic distribution and gene flow are well understood[198,199]. Unfortunately, the required population genetic insight has often come too late. Retrospective population genomic analyses, for example, now distinguish the molecular bases of widespread antimonial resistance on the Indian subcontinent, showing how fully resistant groups emerged independently from pre-adapted subpopulations, some also transferring key resistance mutations by genetic exchange[200]. Apart from predicting antimonial resistance, such high-resolution population genetic studies could have prioritized vector control, e.g., towards areas where basic reproductive number (R_0) is high or where gene flow is expected between resistant and susceptible parasite groups.

1.4.2 Visceral leishmaniasis as a non-endemic, imported disease

The first case of visceral leishmaniasis in the Americas was diagnosed in 1913, the adult patient having fallen ill with severe malaria-like symptoms three years earlier while working in railroad construction near Corumbá, western Brazil[201]. No other cases were reported from the country until 1934, when Penna detected intra-cellular *Leishmania* parasites in autopsies of patients thought to have died from yellow fever in rural regions of the Southeast, North and Northeast[202]. Few new cases emerged in the next two decades, the disease considered more of a medical curiosity than any true threat to public health[203]. This view changed in the early 1950's when reports of visceral leishmaniasis rapidly multiplied in the rural Northeast – first in the state of Ceará, where Deane and Deane diagnosed 188 cases (more than four times as many as previously reported countrywide) over the course of a few months around the town of Sobral[204]. The burden of visceral leishmaniasis continued to grow across the Northeast in the following thirty years, with new foci emerging in Bahia, Pernambuco and Piauí[203]. Infections most often occurred in young children[204] and rarely in urban areas or in states outside of Northeast Brazil, for example, in Goiás, Minas Gerais, Mato Grosso and Mato Grosso do Sul[203]. As the country rapidly industrialized during these thirty years, millions of people left the countryside for economic opportunities only to end up in the periphery of fast-growing cities where sanitation and other infrastructure could not be maintained[203]. Probably due in large part to this uncontrolled urban growth, the early 1980's saw visceral leishmaniasis begin its own process of urbanization and expansion across Brazil. The first major outbreak took place in Teresina, the capital of Piauí[205], where > 1,000 cases occurred in just six months. Large urban outbreaks followed in many major cities, for example, in the state capitals São Luís (Maranhão)[206], Natal (Pernambuco)[207], Rio de Janeiro (Rio de Janeiro)[208], Belo Horizonte (Minas Gerais)[209], and Campo Grande (Mato Grosso do Sul)[210]. By 1990, 53,480 cases of visceral leishmaniasis had been reported in the country[211]. More than 50,000 cases were also recorded between 1990 and 2006[212] as well as between 2001 and 2017[187], the Northeast now accounting for about half of national cases compared to ca. 90% in earlier decades[211,213]. This rapid urbanization and expansion of visceral leishmaniasis does not appear to have occurred elsewhere on the continent. In 2012, less than five percent of cases occurred outside the twenty states affected by visceral leishmaniasis in Brazil[213,214]. These cases include human and canine infections from northern Argentina, Uruguay, Paraguay, Bolivia, Guyana, Venezuela, Colombia, Costa Rica, Nicaragua, Honduras, El Salvador, Guatemala, Mexico and southern USA[214–217]. This distribution of *L. infantum* across the Americas coincides with that of *Lutzomyia longipalpis*, the parasite's most important New World vector based on decades of field-based and experimental research[218]. A number of alternative vector species may occur but appear to have lower vectorial capacity and a more restricted geographic range.

Lu. evansi, for example, has been recorded in parts of Mexico, Central America, Venezuela and Colombia, sometimes where *Lu. longipalpis* is less abundant or does not occur, such as near Colombia's Caribbean coast[219]. Several cases of natural infection by *L. infantum* have been reported[219–222] but infections appear to be less successful than in *Lu. longipalpis*. *Lu. cruzi*, *Lu. intermedia* and *Lu. whitmani* have also been suggested as significant vectors of *L. infantum* in parts of Mato Grosso[223], Mato Grosso do Sul[224], Goiás[225] and Minas Gerais[226], but arguments rest mainly on the low abundance of *Lu. longipalpis* and less on evidence that these congeners can maintain the transmission of disease.

Despite the vast geographic range in which American visceral leishmaniasis occurs, genetic[227,228,17] and enzymatic diversity[229,230] in New World *L. infantum* populations is far lower than that in *L. infantum* populations from the Old World[17]. Genetic divergence between these populations is also very limited, often indistinguishable using classic marker-based analyses such as RAPD[228] or RFLP[227]. For this reason, it has long been hypothesized that *L. infantum* was introduced to the Neotropics from Mediterranean Europe or North Africa within the last 500 years[227,17]. Some authors have argued against such recent, post-Columbian introduction, proposing that a distinct species, *L. chagasi*, entered South America with ancient canids upon the formation of the Isthmus of Panama ca. 3 million years ago[231,232]. This argument centered on the detection of benign infections in wild New World mammals (primarily the crab-eating fox, *Cerdocyon thous*[231], and to a lesser extent, *Didelphis albiventris*[233] *and D. marsupialis*[234]) and the premise that adaptation to *Lutzomyia*, the New World vector genus, could not have occurred in so little time[235]. Duration of host-parasite association, however, does not necessarily correlate with virulence[236,237] and sampling bias towards healthy individuals is likely to occur in surveys of wild mammal hosts[197]. A relatively narrow host spectrum also does not accord well with millions of years of coexistence with the exceptional mammalian diversity known of the New World. Lainson and colleagues, for example, examined 2,637 animals for *L. chagasi* infection, including marsupials, procyonids, rodents, canids, monkeys and edentates from Amazonian Brazil. Infection was found only in *C. thous*[231]. Prevalence of infection has been high in a number of other studies on the crab-eating fox[238–240], but infectiousness and therefore, R_0, in this species may be very low[197,241]. The argument that adaptation to a new vector genus is not possible within a few hundred years is also easily dismissed. *Lutzomyia longipalpis* has been shown to be as susceptible to European *L. infantum* parasites as is *Phlebotomus ariasi*, one of many different *Phlebotomus* vector species exploited by *L. infantum* in the Old World[242]. All these points are consistent with the arrival of *L. infantum* after Columbus and make a weak case for an anciently endemic *L. chagasi* parasite, a case perhaps fully closed following higher-resolution genetic comparisons of New World and Old World parasite populations based on MLMT[18,19].

Applying highly polymorphic markers to exceptionally large sample sizes (e.g., 406 *L. infantum* strains from seven countries of the New World and thirteen countries of the Old World in Kuhls et al. (2011)[18]), these studies demonstrated that low parasite genetic diversity and divergence in the New World are very unlikely artefacts of previous resolution limits or spatial focus. The interspersed phylogenetic positions of New World MLMT genotypes within a wider Old World clade also reinforced the idea of multiple post-Columbian introduction events. Multiple introductions more simply explain the widespread occurrence of *L. infantum* in the Americas[18] than does ancient dispersal (without diversification) across this range[17].

Range expansion can precipitate strong natural selection and/or neutral population genetic change, e.g., when pioneering species encounter novel environmental conditions, escape native competition, or expand from small founding groups with high sensitivity to genetic drift[243]. Hybridization, an important source of novel diversity in *Leishmania*[154,155,244–246] (see Section 1.4.4) is also possible when previously isolated populations meet due to multiple introduction events[247]. It is therefore surprising that, although most authors now recognize American visceral leishmaniasis as an introduced disease, relatively little effort has been made to distinguish or disentangle selective and demographic processes contributing to parasite genetic divergence, and ultimately, clinical variation, in the New World. Microsatellite-based approaches, for example, have described genetically divergent *L. infantum* subpopulations in the West of Brazil[18,248], but none have followed up on (vague) hypotheses that some sort of unique selection pressure (e.g., a distinct vector species) is operating near the Pantanal, or alternatively, that this divergence stems from a separate bottleneck and/or introduction event. Meanwhile, in other areas of the New World where clinical outcomes vary but genetic subdivision appears absent or weak, it has been concluded that *L. infantum* diversity is too low to account for differences in pathogenicity or response to drugs[249]. This lack of association between parasite genotype and disease phenotype could be true when multiple marker systems do not differentiate strains with highly contrasting clinical profiles, e.g., strains that cause non-ulcerating cutaneous lesions vs. strains that cause the expected, visceral form of disease[249–251]. In such cases, properties of the vector (e.g., biochemical characteristics of the saliva[252]) and host (e.g., age[253], nutritional status[254], or presence of co-infections such as with HIV[255]) may predict disease outcome better than do parasite genetic traits[251]. Nevertheless, it seems unwise to generalize that *L. infantum* genetic diversity is too low to help determine disease outcomes or identify genetic bases of pathogenicity anywhere in the New World range. Large microsatellite-based surveys (e.g., 15 microsatellites genotyped in 132 isolates) of *L. donovani* diversity, for example, also showed no link between drug resistance and genotype[256] where WGS later pinpointed resistance mechanisms and

independent waves of purifying selection in low-diversity, yet cryptically diverging parasite groups[200]. While the first WGS study on American *L. infantum* strains did not find any association between individual sequence variants and clinical outcome or host type[257], the second (and only other WGS study on American *L. infantum* to date) found a strong association between the presence of a 'miltefosine sensitivity locus' and positive response to treatment with miltefosine, an important anti-leishmanial drug[258]. This relationship was recently substantiated with experimental evidence that locus knockout induces miltefosine resistance *in vitro* (findings presented at the British Society for Parasitology's March 2020 Trypanosomiasis and Leishmaniasis Seminar[259]). The locus is expected to occur in at least four copies within each cell given its position on chromosome 31, the only chromosome that consistently shows tetra- or pentasomy in *L. infantum* and various other *Leishmania* genomes[244,257,260]. All four copies were often found to be deleted in *L. infantum* samples from different states of Brazil, most often in those isolated from patients that relapsed after treatment[258]. When, why or where this deletion arose and how it confers resistance to miltefosine remains unknown. Gene and chromosomal copy number variation is thought to constitute a primary adaptive strategy in *Leishmania* (see Section 1.4.3) but the genes that occur within the deleted locus (ecto-3′-nucleotidase/nuclease, ecto-3'-nucleotidase precursor, helicase-like protein and 3,2-trans-enoyl-CoA isomerase) show no obvious relationship to the metabolism of miltefosine within the parasite cell[261]. It is also possible that the deletion itself is non-adaptive but linked to an unnoticed complex of selected traits. Another possibility is that genomes containing the deletion have proliferated in the absence of any true fitness advantage, as could have occurred if the mutation arose early on an expanding wave front and/or happened to survive a significant bottleneck event[243].

With so many questions opening up upon closer analysis of *L. infantum* diversity in the New World, the simplification that these populations were bottlenecked and therefore now too homogenous to cause variable disease outcomes does not seem useful for future research. Much work lies ahead to uncover underappreciated population genetic structure, its ecological and evolutionary precedents, and relationships to variation in disease phenotypes. Major human demographic changes within Brazil and elsewhere in the Americas will greatly complicate this task. Frequent internal migrations, for example, make it difficult to distinguish autochthony or obscure other, less recent demographic events, and changes in the prevalence of different diseases (e.g., AIDS[255]) are known to affect the transmission and pathogenicity of *Leishmania* parasites. Climate change is also rapidly changing the geographic distributions and ecological associations of vector-borne diseases, confusing what little is known so far, for example, about the epidemiological roles of vector and host species other than *Lu. longipalpis* and the domestic dog.

1.4.3 Molecular mechanisms of divergence and adaptation within *Leishmania* spp.

Unlike *T. cruzi* lineages and subspecies that diverge strongly in repetitive surface gene families at expanded telomeres[262,263], the *Leishmania* genus shows strikingly little variation in coding sequence composition or genome size[264]. Coding sequence nucleotide identity between *L. infantum* and *L. major*, for example, is approximately 94%, and both species appear to have 32 Mb genomes composed of 36 chromosomes[264]. At the intra-specific level within the *T. cruzi* complex, by contrast, coding sequence nucleotide identity between *T. cruzi* I and *T. c. marinkellei* is less than 93%, and the genome of *T. c. marinkellei* is 11% smaller than that of *T. cruzi* I[265]. Despite little apparent sequence variation within the *Leishmania* genus, phenotypic variation is remarkably high. *L. donovani* and *L. infantum* typically cause visceral leishmaniasis, *L. braziliensis*, *L. guyanensis* and *L. panamensis* can create highly disfiguring, mucocutaneous lesions and *L. mexicana*, *L. major* and *L. tropica*, among several other species, most often lead to less destructive, cutaneous disease[191]. These associations, however, do not always occur as such, with highly contrasting clinical outcomes observed within a single species and even a single focus of transmission. Dermotropic *L. infantum* strains, for example, sometimes circulate among the expected, visceralizing forms[249], and *L. braziliensis* is known to cause both cutaneous and mucocutaneous disease[191]. Complex influences of host/vector genotype and environment may contribute substantially to such variation within species but do not explain the general discrepancy observed between low genetic differentiation and vastly different phenotypes in the *Leishmania* genus. Different programming of gene expression is likely key to generating diversity, including short- and long-term adaptation, divergence and, ultimately, speciation in these parasites[180,266].

Like all other trypanosomatids, however, *Leishmania* species do not use monocistronic transcription, constitutionally transcribing genes in long polycistronic units instead of individually from distinct promoter motifs[267]. Aside from modifying transcript abundance by deadenylation and cap removal as do other Trypanosomatidae[268–270], the *Leishmania* genus exploits an exceptional tolerance for karyotypic plasticity to modulate and diversify gene expression[271,272]. This karyotypic plasticity occurs in the form of gene copy number variation and chromosomal aneuploidy and extends the parasites' ability to regulate expression levels, possibly also the complexity of pleiotropic interactions among genes[271,272].

Gene copy number variation is enabled by abundant intergenic repeat sequences, in large part degenerate retroposons, that alter DNA replication patterns in several ways. Homologous repeats of similar orientation (i.e., positioned head-to-tail) can anneal by loop formation of the intervening sequence, and these loops can separate as circular amplicons (Fig. 1.4a)[273]. Such extrachromosomal amplicon formation can be conservative or nonconservative,

whereby the locus between the repeats is deleted from the original template strand[273]. Alternatively, head-to-tail repeats (Fig. 1.4b) can lead to intra-chromosomal tandem duplications by unequal sister chromatid exchange[273]. Yet another form of copy number variation occurs when inverted (tail-to-head) repeats trigger DNA strand breaks and hairpin structures that result in the formation of extrachromosomal linear amplicons[273]. Patterns of copy number variation appear to differ between strains and species and may in some cases distinguish populations from different transmission cycles more clearly than do SNPs[21,264,274,275]. A variety of associations between copy numbers and phenotypes, e.g., tissue tropism[276], stress response or drug resistance[272,277,278], have also been observed, although the mechanisms behind these associations often remain unclear[20].

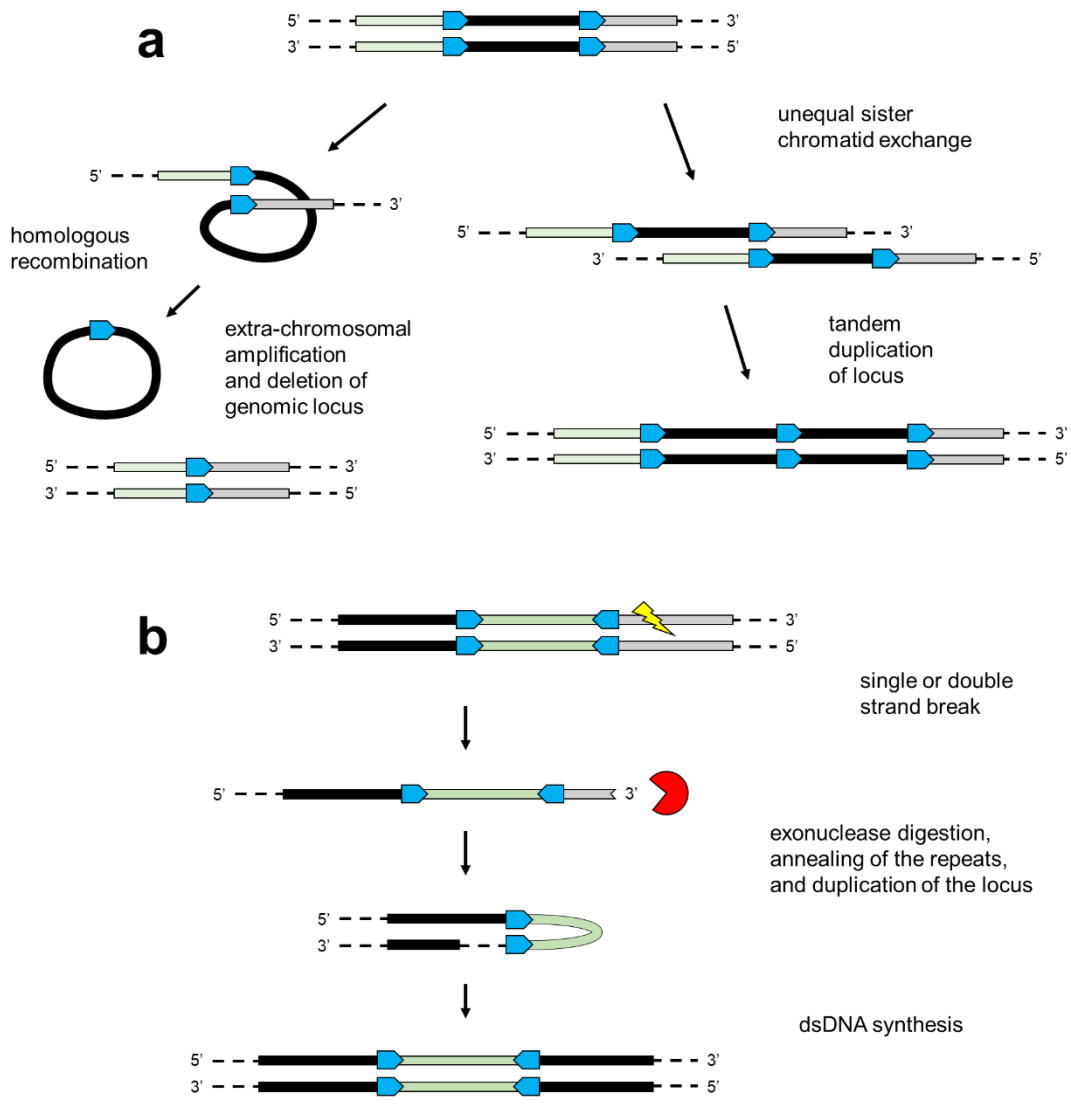


Figure 1.4 Mechanisms of gene amplification in *Leishmania* – based on Ubeda et al. (2014)[273]. **a** Homologous recombination between direct repeats can lead to circular amplification or to tandem duplication by unequal sister chromatid exchange. **b** Linear amplification can also occur in the presence of DNA strand breaks near inverted repeats. Broken ends are digested (by MRE11 exonuclease) and hairpin formation enables duplication of the locus. See Ubeda et al. for details[273].

Aneuploidy in *Leishmania* is thought to occur due to unusually high rates of asymmetric chromosomal allotment in mitotically dividing cells[279]. Like gene copy number variation, baseline ploidy levels and amplification programs appear to be species-[21,274] and strain-specific[274] and enable rapid, reversible adaptation to changing environments, for example, during transition between the vector midgut and the host phagolysosome[260,280]. Also as with gene copy number changes, correlations between altered somy levels and drug resistance, e.g., to antimony[272,281] or methotrexate[271], frequently occur, but mechanisms (i.e., which specific genes within a set of amplified chromosomes promote resistance) remain largely unsolved[20]. Another interesting aspect of chromosomal copy number variation in *Leishmania* relates to the frequent presence of mosaic aneuploidy within strains[279]. Mosaic aneuploidy occurs when cells within a single cell population do not all have identical karyotypes but comprise a diversity of subpopulations, each with a different karyotype, and subpopulations with karyotypes advantageous to the present environmental conditions may thrive over others until conditions change again, conceivably increasing the fitness of the strain as a whole. A recent study also exposed that chromosomes with higher mutation rates may be more prone to amplification and facilitate haplotype selection (i.e., deletion of less advantageous chromosomes) within this mosaicism to accelerate adaptation in *Leishmania* parasites[282].

Although relatively limited sequence diversity within the *Leishmania* genus has directed much research interest towards gene dosage and mechanisms of post-transcriptional control (reviewed elsewhere[266]), SNP and insertion-deletion (INDEL) mutations clearly also contribute to speciation and phenotypic change. Comparing *L. infantum*, *L. major* and *L. braziliensis*, for example, the relative frequency of non-synonymous vs. synonymous SNPs and INDELs (i.e., the possible occurrence of positive selection) differs in approximately eight percent of syntenic genes, and many of these genes relate to core metabolic processes linked to pathogenic traits[264]. SNPs and INDELs also drive pseudogene formation that contributes to divergence among the three genomes[264]. Next to comparative genomics, experimental and field studies have found a handful of point mutations that alter tissue tropism (e.g., SNPs in a ras-like RagC GTPase enzyme appear to attenuate visceralization by *L. donovani*[276]) or predict susceptibility to drugs such as pentavalent antimonials[283–285,200] and miltefosine[286,287]. Associations established from field and laboratory settings, however, often fail to corroborate one another's results. SNPs in the miltefosine transporter LdMT, for example, were recurrently associated to miltefosine resistance *in vitro*[287–290] but were not observed in any of thirty *L. donovani* isolates taken from patients that relapsed after treatment with miltefosine in India and Nepal[291], suggesting that multiple routes to resistance occur in the field and/or that *in vitro* conditions do not accurately model the natural environment. Relatively limited success to date in identifying SNP and INDEL variants responsible for important phenotypic

changes may also be attributable to the frequent use of targeted approaches rather than full genomic scans[292]. While targeted approaches (e.g., microarrays or PCR) are commendable in representing hypothesis-driven science, they are unlikely to detect mutations that transform phenotypes through pleiotropic or cumulative effects and are generally biased towards annotated or previously studied genes.

Increasingly accessible WGS data will improve theory on genotype-phenotype interactions as they are unrestricted by *a priori* knowledge and can assess multiple adaptive mechanisms (i.e., gene copy number variation, aneuploidy, SNPs and INDELs) as well as demographic patterns across many samples at once (see Section 1.5). Various precautions, however, are relevant, both in regard to the use of classic population genetic approaches on trypanosomatid parasites and in handling the massive amounts of data that lay ahead (Section 1.6).

1.4.4 Hybridization in *Leishmania* spp.

Leishmania parasites can also generate new diversity through genetic exchange. Intra-specific *L. donovani* and *L. major* crosses have between achieved in the sand fly vector by co-infecting cell lines carrying different fluorescent or drug resistance genes[244,245,293]. The drug selection approach used in both *L. major* studies allowed the hybrid progeny to be isolated and characterized. Heterozygosity and ploidy levels were consistent with classical meiosis in most offspring but several cases of genome-wide triploidy and, less frequently, tetraploidy, also suggested that the hybridizing parents occasionally failed to initiate or complete meiotic division[244,245]. Similar rates of diploid and aneuploid hybrid offspring were also produced in inter-specific crosses between *L. infantum* and *L. major* by Romano et al. (2014)[294]. The authors went on to infect mice with the hybrid offspring, revealing clear differences in their abilities to produce dermal lesions or to disseminate and grow in the liver and spleen relative to parental strains.

Evidence of hybridization and its major impact on *Leishmania* phenotypes and epidemiology is also found in the field. Natural hybridization between *L. braziliensis* and *L. peruviana*, for example, has been implicated in the emergence of especially destructive forms of mucocutaneous disease[295], and natural *L. infantum*/*L. major* hybrids, appear to have gained the ability to infect *Phlebotomus papatasi*, a widespread Old World sand fly species previously considered permissive only to *L. major*[154,296]. Intra-specific hybridization also appears to have preceded the widespread expansion of a fixed heterozygous *L. tropica* genotype through much of Asia, including Turkey, India and the Middle East[297].

Only two studies, however, have used WGS to better understand the demographic histories behind hybrid *Leishmania* genomes. The first study by Rogers et al. (2014) examined eleven

vector-isolated *Leishmania* strains from a focus of cutaneous leishmaniasis in Turkey, and genome-wide patterns of patchy heterozygosity could be clearly traced back to a single outcrossing event and a low frequency of inbreeding ($1.3 \cdot 10^{-5}$ meioses per mitosis) among offspring genotypes such that initial full-chromosome heterozygosity fragmented into shorter blocks of mixed and non-mixed ancestry over time[298]. Phasing haplotypes over heterozygous loci could also specify *L. infantum* and an *L. donovani*-like species as parental strains. The second study by Cotton et al. (2019) described a more complex history of hybridization within *L. donovani* populations of northern Ethiopia, where mixed-strain sand fly infections may be more common than in the Turkish locality above[246]. Patterns of inheritance indicate that extant Ethiopian *L. donovani* hybrids originate from multiple separate initial crossing events, and that these events were also followed by backcrossing to parents and/or with other hybrid lines.

As the above WGS studies targeted aberrant and/or putatively hybrid populations based on previous MLST and MLMT[299,300], further WGS surveys are required to clarify how common and/or influential hybridization events are to genetic structure and diversity at other disease foci, and in which ecological or demographic circumstances hybridization is most likely to occur. Many previous marker-based studies have emphasized the presence of high homozygosity due to endogamic mating (i.e., selfing or inbreeding) and that predominant clonal evolution governs the population structure of *Leishmania* strains. Clonality is undoubtedly the most frequent form of *Leishmania* reproduction, but above examples demonstrate that much less common exogamic genetic exchange can be pivotal to parasite diversity and fitness. Sparse marker panels used in the majority of past studies may have very often missed other such examples when these did occur and would definitely have been powerless to distinguish complex hybridization patterns like those described using WGS in Cotton et al. (2019)[246].

Introduced American *L. infantum* populations appear especially deserving of WGS studies with attention to causes and consequences of genetic exchange. Hybridization is often linked to range expansion in other species, either because it enhances or facilitates survival in the new environment[301] or because demographic restructuring during the expansion process connects populations that otherwise rarely meet[302,303]. Surprisingly, the first and only two WGS studies carried out on American *L. infantum* populations to date[257,258] chose not to examine genome-wide heterozygosity distributions or even to construct phylogenetic trees in their efforts to find reasons behind phenotypic differences among strains. As highlighted above, genetic exchange can transform phenotypes at various levels (permissiveness to vectors, tissue tropism in hosts), and such analyses would have been first steps into investigating this possibility in the New World.

1.5 Advantages and prospects of the genomic age

The potential of WGS to enhance various lines of population genetic research on trypanosomatid parasites has been touched on throughout this review. Three (intertwined) advantages can be summarized. These relate to 1) innovative inference from 'comprehensive' genotyping, 2) extraordinary (and non-focal) resolving power and 3) potential for cross-disciplinary integration (highlighting the 'landscape genomics' approach).

First, comprehensive genotyping, i.e., genotyping at all genomic loci as opposed to discontinuous genotyping at selected markers without information on adjacent sequences (as in MLST, MLMT, microarrays, etc.), offers unprecedented opportunity to reconstruct and quantify demographic processes behind parasite diversification and extant population structure, e.g., the frequency and mechanisms of genetic exchange[304]. One of the best examples was just referenced in Section 1.4.4. The landmark study by Rogers et al. (2014)[298] visualized genome-wide mutation patterns to infer the series of mating events leading to aberrant *L. infantum* genomes in south-central Turkey, showing how intermittent blocks of heterozygosity derived from a single hybridization event followed by inbreeding or selfing among outcrossed strains. The authors then used information relating to the size and frequency of these blocks and estimates of genome-wide mutational diversity to infer the relative rates of meiotic and mitotic cell division in the population. The low frequency of meiosis in *Leishmania* would have been very difficult to measure in the laboratory or by any marker-based technique. Another fascinating example is given by Weir et al. (2016)[305]. The authors used comprehensive sequence information to demonstrate strict asexuality in *Trypanosoma brucei gambiense*, the genomes of which showed linkage disequilibrium across all chromosomes, i.e., each genome formed a single linkage group. Accumulation of mutations on separate, co-evolving haplotypes also showed the Meselson Effect like few studies in any species have ever achieved, and long tracts of homozygosity suggested gene conversion as a possible compensatory effect. Apart from continuous genome-wide information on point mutations, WGS can also distinguish structural rearrangements as a key source of novelty in parasite genomes. Talavera-Lopez et al. (2018), for example, used long-read PacBio sequencing to demonstrate that *T. cruzi* uses radical, inter-chromosomal translocations to transform its antigenic repertoire[306]. The authors also used genome-wide linkage scans to identify selective sweeps in important surface molecule gene arrays in human-isolated parasite genomes (vs. balancing selection in these arrays in vector-isolated parasite genomes). Scans across continuous genomic sequence can also, for example, quantify the strength of selection at individual loci[307,308] or detect important (e.g., virulence-associated[309]) introgressive events.

The second key virtue relates simply to the enormous sensitivity and resolution of the data-driven approach. Decreasing costs of WGS enable large sample sizes to be measured at high sequence read coverage, with no *a priori* target selection required. The advantages are manifold. Deep sequencing, for example, creates unprecedented power to detect rare variants, e.g., deleterious mutations in important parasite genes[310], minimizing past biases toward positively selected traits[304]. Next to rare genotypes, WGS also has the power to detect rare genomes. Trypanosomatid infections, especially those of *T. cruzi*, often comprise multiple clones, not just a single monoclonal strain[311–313], and bioinformatic pipelines (e.g., from malaria research[314]) applied to high-depth sequencing data can potentially deconvolute component genomes. Beyond just distinguishing (rare) variants, high depth WGS also facilitates the precise measurement of variant allele frequencies. Pattern-process modelling of the site frequency spectrum or its summary statistics can be used to reconstruct various demographic processes, e.g., past admixture, bottleneck or expansion events[315,316]. While approaches to reconstruct demographic history often require information on neutral sequence variation, other studies may need to filter out such genetic structure, e.g., to identify loci under selection in sample genomes[317]. WGS allows for various kinds of data separation *post hoc*, e.g., after distinguishing synonymous and non-synonymous mutations based on annotated codons and genes[307]. Finally, high sequence read coverage also enables detection of aneuploidy and gene copy number variation simultaneously with SNPs and INDELs, the importance of which has been elaborated in Section 1.4.3. Chromosome-wide vs. local copy number changes cannot be differentiated in such detail using other molecular techniques, e.g., fluorescence *in situ* hybridization, relatively sensitive, but very prone to artefacts (Hideo Imamura, pers. comm.).

There is also great prospect in integrating WGS with other 'omics' approaches (transcriptomics, proteomics, metabolomics, etc.), e.g., to better understand how diverse phenotypes arise from relatively low genetic diversity and an absence of monocistronic transcription control, but this integration is only beginning to take form[318]. The integrated analysis of high-resolution genetic and spatial data, however, is further along and has been formalized under the term 'landscape genetics' – or 'landscape genomics' when WGS technologies are applied. Landscape genetics/genomics is a research field that aims to explicitly quantify the effects of environmental composition and configuration on genetic variation with novel spatial statistics[319,320]. As these effects are tested at either the ecological or the evolutionary scale, distinct data models and dimensions of genetic structure are drawn into analysis. The ecological focus, set to test landscape effects on dispersal and resultant demography, assesses genome-wide, neutral genetic structure. Many landscape genetic studies assess correlations between pairwise measures of genetic dissimilarity and distance-

based landscape data describing the intervening matrix between sampling sites. This 'link-level' analysis thus often summarizes genetic and spatial data in distance matrices, whereby associations among component vectors can be evaluated by Mantel statistics, partial multivariate regression of distance matrices (when multiple explanatory variables are of interest) or other matrix-based statistical tests[321]. The evolutionary focus, by contrast, generally aims to elucidate genotype-by-environment associations (i.e., to detect selection) in non-neutral regions of the genome[322]. Methods are therefore often 'node-based', assessing correlations between local environmental metrics and allele frequencies without reference to landscape that intervenes sampling sites[321]. Multivariate statistical methods such as ordination by redundancy and canonical correspondence analysis are commonly applied[323]. Some landscape genetic approaches, however, combine both link- and node-based perspectives to predict environmentally driven changes to neutral and adaptive genetic structure over time. This includes landscape genetic simulation modelling, a spatially explicit modelling technique in which population genetic structure is simulated over a raster of different environmental conditions. Each individual begins simulation in the raster with a defined genotype and moves from cell to cell according to hypothesized effects of local and adjacent cell conditions on survival, dispersal, mating, mutation, etc. The raster can code for multiple environmental conditions in a landscape of interest (e.g., using remote-sensing data (elevation, temperature, vegetation cover, etc.) or estimates of host and vector abundance based on environmental niche models) such that the comparison of simulated population genetic structure to observed population genetic structure helps test the landscape resistance or selection hypotheses applied. This pattern-process modelling approach is being pioneered for conservation purposes (e.g., to predict the effect of reintroductions and hydroelectric infrastructure on fish diversity and dispersal[324]) using simulators such as CDPOP[325] and CDMetaPOP[326] but could also help in the understanding and management of parasitic disease. The idea to summarize hypotheses about environmental effects on genetic structure into a digital 'resistance raster' is especially intriguing for ancient endemic parasites such as *T. cruzi* that are known to disperse through a wide range of environments yet with transmission cycles tuned by ecological host-fitting[149] and with various conceivable barriers to dispersal or development (e.g., high altitudes[327], rivers[328], desert[329]) as well as anthropogenic influences such as insecticide-spraying[330], deforestation[331] and long-distance transportation of infected vectors and hosts[149,332,333].

1.6 Challenges in trypanosomatid population genetics, genomics, and spatial genomics

Built on relatively simple mathematical formulae such as the Hardy-Weinberg Law (which states the expected heterozygote and homozygote genotype frequencies in a randomly mating population[334]), it comes to no surprise that a model-based system of inference pervades population genetic theory and application. Reference to idealized null distributions perpetuates through all stages of analysis, e.g., from first generation of summary metrics, to their algorithmic implementation, to data transformation, to the manner in which final results are instinctively interpreted. For example, Nei's D, a basic metric of genetic distance, refers to constant mutation rates among loci. The algorithm behind STRUCTURE[335], one of the most heavily used methods of population assignment, assumes Hardy Weinberg and linkage equilibrium within clusters as do the popular programs GeneClass, BATWING and BAPS[336]. Simulation approaches are also inherently model-based. Current CDMetaPOP code, for example, applies mating as Mendelian sex[326]. Principal component analysis (PCA), universally applied for dimension reduction of (genomic) information, also assumes non-independence (i.e., no linkage) among data points. And perhaps most critical among these examples – the ever-present Hardy-Weinberg law equates mating to blending inheritance that restores equilibrium allele frequencies, i.e., HWE.

Although underlying assumptions are not meant to be met at all times, continual violation makes for trouble. Trypanosomatid parasites such as *T. cruzi* and *L. infantum*, however, seem to break the rules very often, definitely much more than pea plants do. An apparent mix of clonality and (perhaps unorthodox) sex make strong linkage and Hardy-Weinberg disequilibrium pervasive[10]. Hardy-Weinberg disequilibrium and linkage in particular risk distorting population genetic inference because many applications require the use of neutral loci inherited according to Mendelian laws[337]. Yet there are ways to manage. First, one may proceed more heuristically and seek methods based on fewer or different assumptions. Discriminant analysis of principle components[338], for example, offers a non-model based alternative to STRUCTURE, and recent modifications to classical multivariate ordination apply linkage information to handle non-independence among markers[339]. In a second (more ideal) approach, one may develop new analyses based on the models of demography and evolution that most likely apply. Regarding the frequency and mechanism of sex in *T. cruzi*, these *a priori* hypotheses for analyses may soon become more tenable as resolutions from genomic sequencing and further experimental studies clarify theory on reproductive mode. In a third approach, also promoted by today's sequencing power, the effects of aberrant genetic properties on existing metrics and models may be quantified through rigorous comparative analyses to explicitly recalibrate past and present inference. Lastly, genomic data sets may be partitioned and filtered *ad hoc* to accommodate statistical assumptions. For example,

measures or multi-collinearity or eigen analysis may be used to omit loci, and DNA segments may be screened individually with tests for HWE[340–342].

As previously mentioned, multi-clonality (co-infection by multiple intra-specific strains) presents another biological feature of *T. cruzi*[15,95,149,311,343] (and to a lesser extent of *L. infantum*[344]) that can severely mislead inference if overlooked. When WGS reads from a multiclonal sample are mistaken to represent those from a single clone, point mutations in the genome of this supposed clone may appear to be abundant whereas structural diversity may appear to be low (signs of trisomy in the form of unbalanced (33% and 67%) allele counts, multiallelism or chromosome-wide elevation in read depth, for example, could be obscured by mosaic aneuploidy (see Section 1.4.3) among cells). Questions relating to individual genotypes (e.g., relationships between multi-locus genotypes and environment) or interactions between individual clones (e.g., genetic exchange) will thus be difficult to solve. Fortunately, as mixed infections are ubiquitous among micro-pathogen taxa[345], a variety of statistics established in other study systems[314,346] can help disentangle component genotypes from multi-clonal infections by trypanosomatid parasites. These bioinformatic solutions, however, involve a margin of error and will not suffice for all objectives, e.g., to provide definitive proof of genetic exchange among individual clones. Fortunately, *T. cruzi* and *L. infantum* are relatively amenable to long-term culture (in contrast to, e.g., *Plasmodium vivax*[347]). Incorporating methods such as fluorescence-activated cell sorting (FACS), single-cell microfluidic partitioning (e.g., 10x Genomics), biological cloning by limiting dilution or plating on solid media[58,312], individual cells or monoclonal strains can be separated from multiclonal samples prior to sequencing. It is important to consider, however, that all forms of parasite culture and micromanipulation risk selection bias. The best course of action, i.e., how much laboratory handling vs. bioinformatic sequence separation is best applied will depend on the study objective and the sensitivity of analyses to multiclonality or representative sampling. Some studies might even require culture-free approaches, e.g., using probe-based target enrichment or selective whole-genome amplification from the infection source. Some such methods (e.g., based on spliced leader trapping[348] or SureSelect technology[349]) have been established for *Leishmania* but are not yet described in *T. cruzi* research.

Another important challenge in WGS-based (trypanosomatid) studies relates to read-mapping error and thus, artefactual variance in sequence composition and depth. No matter whether based on hash tables (e.g., Stampy[350] or SMALT[351]) or suffix arrays (e.g., Bowtie[352] or BWA[353]), alignment programs cannot correctly map short (i.e., Illumina) reads when these represent substrings of sequences that occur in many similar homologs throughout a reference genome. Such sequences are highly abundant in *T. cruzi*, especially in its surface molecule-

encoding tandem gene arrays[77,306]. Mis-mapping in these areas leads to artefactual point mutations, and these can confound metrics of linkage, and ratios of coding vs. noncoding mutation or purifying vs. diversifying selection, etc. Spikes in read depth also occur and can be misinterpreted as local copy number change. Unless long-read (e.g, PacBio or Oxford Nanopore) technologies are applied, this mapping problem can only be circumnavigated by omitting ('masking') unreliable regions from analysis. Identification of these regions is possible by self-blasting and virtual read alignment strategies or by identifying genetic areas where different mapping and variant-calling programs produce inconsistent results. Regions where all sample genomes show unexpected sequence or structural aberrations may also reflect systematic error. Ideally, sequences from a control sample (e.g., a reference strain such as TcI-Sylvio or JPCM5) can be obtained to confirm masking decisions and calibrate settings in various other bioinformatic steps.

Spurious associations are another major concern in data-heavy WGS, especially when analyses integrate additional data types, e.g., in search of correlations between gene dosage and phenotype (i.e., GWAS) or between population genetic differentiation and environmental variation measured using high-resolution, remote-sensing techniques. Various statistical methods help correct for extreme multiplicity in testing[354], reduce collinearities[355], control for neutral structure or detect outlier effects[337,342,356–360] but other issues are not so easily cleared *post hoc*. In landscape genomic studies, for example, spatial change in environmental variables of interest can coincide with demographic movements (e.g., altitudinal or humidity gradients can coincide with expansion axes of an introduced species) and contemporary (observed) population structure may be governed by historic (unmeasured) conditions and events (e.g., past land-use change, vector intervention, species introduction, etc.)[361,362]. High prudence in scientific approach and sampling design is therefore at least as important as are later decisions on data filtering and statistical controls. Although studies using next-generation sequencing/sensing technologies are in part so powerful because no *a priori* target selection is required, this release from hypothesis-driven target selection should not encourage a departure from hypothesis-driven science. It is important to formulate expectations before beginning any high-throughput analysis, and also long prior to the computational stage. Deliberate study site selection, spatial configuration and intensity of sampling is essential for unbiased, meaningful inference[362–366] and must base on sound hypotheses or knowledge of the study environment (e.g., historic disturbances, cryptic barriers and patterns of environmental values – linear, modal, random, etc.) and ecology of the study organism[362]. Regarding vector-borne parasites, this latter condition is not easily met, as parasite gene flow depends not only on the intrinsic biological properties of the parasite (e.g., reproductive mechanism, ability to infect certain taxa, virulence, etc.) but on a factorial of host and vector

traits (abundance, lifespan, dispersal patterns, etc.)[367]. This trait space determines the degree of contact and transmissivity among hosts and vectors and therefore modulates parasite population structure and genetic connectivity in the landscape. Parasite population structure may range from highly segregated and metapopulational, with little or no gene flow to (or absence of infections at) nearby sampling locations to relatively continuously distributed, with genetic similarity fading as a function of geographic distance[368]. Like most vector-borne parasite species, *T. cruzi* populations conceivably place towards the metapopulational end of this spectrum given that stercorarian host infection is highly inefficient[143] and ecological host-fitting (e.g., separate terrestrial and arboreal niches) is observed at the landscape scale[149]. Nevertheless, transmission cycles can contain a high abundance of hosts[130] and may increasingly overlap[369] if interactions between generalist hosts and vectors increase (e.g., in areas disturbed by deforestation or climate change). Genetic connectivity may also be enhanced by non-vectorial[370] transmission and long-range synanthropic dispersal routes[149]. Population structure is thus likely less patchy than that of *L. infantum* in sylvatic or rural landscapes of the New World. *L. infantum* host diversity appears to be much less extensive and only domestic dogs are considered primary reservoir hosts[197]. Genetic connectivity may be high within urban regions but not in other environments or at larger scales. Effects of recent parasite bottlenecks and expansions[17,18], also human migrations[203,371], have also likely been pivotal to *L. infantum* population structure in the non-endemic range. It is important that such hypotheses contribute to spatial study design.

1.7 Research chapter synopsis

Several fundamentals of *T. cruzi* and *L. infantum* biology and epidemiology described in the above literature review have yet to be solved. The extent of genetic recombination occurring within natural *T. cruzi* infections (see Section 1.3.7), for example, remains unknown. Mating by polyploidization has been observed *in vitro* but does not reconcile with allele frequency and somy patterns observed in the field. Inference from the field, however, often remains inconclusive due to low-resolution genotyping of uncloned, potentially mixed-strain isolates sampled at inappropriate scales, e.g., across disparate transmission cycles or from different points in time. Chapter 2 therefore uses plate-cloning to establish monoclonal *T. cruzi* cultures from recent vector/host captures at a single transmission focus in southern Ecuador, then examines nucleotide and copy number variation in the sequenced genomes to identify reproductive mechanisms and quantify possible events of genetic exchange. Clones are also subcloned and re-sequenced after cryopreservation to assess karyotypic plasticity and mosaicism as evidence for/against initial hypotheses of parasexual aneuploidy in the dataset.

Another major open question relates to much larger spatial patterns – the distribution of *L. infantum* diversity throughout Brazil and relationships to source populations in the Old World (see Section 1.4.2). American *L. infantum* populations are likely to have undergone significant macrogeographic restructuring in the course of recent importation and expansion into the New World. Distinct transmission ecology (e.g., use of *Lutzomyia* vectors, more restricted host range, etc.) may also have elicited significant adaptive genetic change. Microsatellite approaches have adumbrated complex population structure in these populations but are of little help in clarifying drivers of non-neutral genetic variation and important clinical features of disease (e.g., miltefosine resistance) observed in Brazil. Chapter 3 therefore uses WGS reads from 126 New and Old World *L. infantum* strains to reconstruct invasion history and possible adaptive processes occurring in the introduced range. A wide variety of genomic methods (copy number analyses, simulation modelling, etc.) as well as phenotypic tests are employed. Special emphasis is placed on hypotheses of neutral vs. selected copy number variation at a recently identified miltefosine sensitivity locus, associated enzymatic activity, and alternative metabolic paths. Sample size and distribution represented limiting factors in both Chapters 2 and 3 because inefficient parasite 'isolation-by-culture' restrained the extent to which hypothesis-driven spatial sampling could be optimized (see Section 1.6).

Chapter 4 therefore develops a 'genome-wide locus sequence typing' (GLST) tool to summarize parasite genetic polymorphism at low cost and without cell purification and culturing steps. Loss of parasite diversity *in vitro* is a significant concern in trypanosomatid research but few such methods have been developed to extract genome-wide trypanosomatid sequence information from uncultured sample types.

Inspired in part by the prospect of rapidly surveying parasite diversity across landscapes using tools like GLST, Chapter 5 constructs a new landscape genomic framework for the prediction and prevention of vector-borne disease. The framework proposes landscape genetic simulation modelling (see Section 1.5) on a composite resistance raster that integrates hypothesized effects of host and vector activity on parasite dispersal pathways in the landscape. Chapter 2's Chagas disease study system in Ecuador is used to walk readers through different principles and methodological steps.

Key findings, limitations and possibilities of follow-up to the four research chapters are discussed in Chapter 6.

Chapter 2

Meiotic sex in Chagas disease parasite *Trypanosoma cruzi*

Philipp Schwabl[a], Hideo Imamura[b], Frederik Van den Broeck[b], Jaime A. Costales[c], Jalil Maiguashca[c], Michael A. Miles[d], Björn Andersson[e], Mario J. Grijalva[c,f] and Martin S. Llewellyn[a]

[a]Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK

[b]Unit of Molecular Parasitology, Institute of Tropical Medicine Antwerp, 155 Nationalestraat, 2000, Antwerp, Belgium

[c]Center for Research on Health in Latin America, School of Biological Sciences, Pontifical Catholic University of Ecuador, Quito, Ecuador

[d]London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

[e]Department of Cell and Molecular Biology, Science for Life Laboratory, Karolinska Institutet, Biomedicum 9C, 171 77 Stockholm, Sweden

[f]Infectious and Tropical Disease Institute, Biomedical Sciences Department, Heritage College of Osteopathic Medicine, Ohio University, 45701 Athens, OH, USA

This chapter has led to a publication in Nature Communications (2019),
available at doi: 10.1038/s41467-019-11771-z.

2.1 Abstract

Genetic exchange enables parasites to rapidly transform disease phenotypes and exploit new host populations. *Trypanosoma cruzi,* the parasitic agent of Chagas disease and a public health concern throughout Latin America, has for decades been presumed to exchange genetic material rarely and without classic meiotic sex. We present compelling evidence from 45 genomes sequenced from southern Ecuador that *T. cruzi* in fact maintains truly sexual, panmictic groups that can occur alongside others that remain highly clonal after past hybridization events. These groups with divergent reproductive strategies appear genetically isolated despite possible co-occurrence in vectors and hosts. We propose biological explanations for the fine-scale disconnectivity we observe and discuss the epidemiological consequences of flexible reproductive modes. Our study reinvigorates the hunt for the site of genetic exchange in the *T. cruzi* life cycle, provides new tools to define the genetic determinants of parasite virulence, and reforms longstanding theory on clonality in trypanosomatid parasites.

2.2 Introduction

Trypanosoma cruzi is a kinetoplastid parasite and the causative agent of Chagas disease in Latin America, where ca. six million people are currently infected[1]. Mucosal or abrasion contact with the infected feces of hematophagous triatomines constitutes the primary mode of *T. cruzi* transmission. Infection with *T. cruzi* results in chronic Chagas disease in 30 – 40% of cases, characterized by a spectrum of fatal cardiac and intestinal pathologies. Early-stage acute Chagas disease can also be fatal, especially among infants and in orally transmitted outbreaks of the disease[372]. *T. cruzi* transmission is a zoonosis maintained by numerous species of triatomine insects and hundreds of different species of mammals[373].

The Trypanosomatidae, the family to which *T. cruzi* belongs, is a monophyletic group of obligate parasites and includes several species of medical and veterinary importance – e.g., *Trypanosoma brucei* ssp., *Leishmania* spp., *Trypanosoma vivax* and *Trypanosoma congolense*[374]. The Trypanosomatidae are early branching eukaryotes in evolutionary terms and share many biological characteristics, including the process of U-indel RNA editing in the kinetoplast[375] and polycistronic transcription control[376]. Despite their basal status, the Trypanosomatidae possess much of the core meiotic machinery of higher eukaryotes[377]. However, the extent to which such machinery might actually support genetic exchange within trypanosomatid species has been slow to come to light[6]. Establishing the occurrence of regular meiotic recombination in *T. b. brucei* has taken decades of laboratory and field research; not until 2014 was haploid gamete production (coincident to peak meiosis-specific gene expression) confirmed by fluorescence microscopy as a normal phase of development

in the vector's salivary gland[378–380]. More recently, genome-scale signatures of meiosis have also been detected in *T. congolense*[181,183]. In contrast, robust genomic evidence now suggests that the human-infective *T. b. gambiense* subspecies is completely asexual[305]. Life histories in *Leishmania* seem no less complex. Despite a clear propensity for mitotic clonality, sporadic sexual hybrid formation appears to underlie important diversification events both within and between species[295,381], and meiotic offspring are readily produced in laboratory crosses[244,382]. An alternation of clonal and sexual, endogamic reproduction has also been proposed to define population genetic structure in the *Viannia* complex[383].

T. cruzi is the last of the Tritryps (*Leishmania* spp., *T. brucei* ssp. and *T. cruzi*) for which the extent and mechanism of genetic exchange remains to be fully elucidated. Limited evidence for genetic recombination has been observed in the field[170,176] although inappropriate study designs, genetic marker systems of insufficient resolution, and low genetic diversity in study populations have all hampered interpretation of the data[6]. Furthermore, the parasexual mechanism of genetic exchange proposed for *T. cruzi* based on a single experimental cross – one of whole-genome fusion followed by stochastic chromosomal decay and return to diploidy[171] – has been irreconcilable with patterns of somy and genetic diversity observed in natural populations[170,174,384]. This lack of clarity has lead some to propose *T. cruzi* as a paradigm for Predominant Clonal Evolution (PCE)[16,162] in parasitic protozoa – an idea which may not reflect biological reality.

To address this fundamental knowledge gap in the biology of trypanosomatids, in this study we generate whole-genome sequence data from 45 *T. cruzi* Discrete Typing Unit I clones, as well as several non-cloned *T. cruzi* strains, collected from triatomine vectors and mammalian hosts in an endemic transmission focus in Loja Province, southern Ecuador. After mapping sequences against a recent PacBio sequence assembly[306], we explore patterns of population structure and genetic recombination. Our data reveal that *T. cruzi* does indeed reproduce sexually at high frequency via a mechanism consistent with classic meiosis. However, we demonstrate that parasite groups with radically distinct reproductive modes also co-occur at the same transmission focus. As the last medically important trypanosome for which meiosis has not yet been demonstrated in lab or field, our data on *T. cruzi* make a significant contribution towards the consolidation of current theories around genetic exchange in the Trypanosomatidae.

2.3 Methods

2.3.1 Parasite collection and cloning

Trypanosomes were isolated from triatomines (*Rhodnius ecuadoriensis*, *Panstrongylus chinai*, *P. rufotuberculatus*, and *Triatoma carrioni*), rodents (*Rhipidomys leucodactylus*, *Sciurus stramineus*) and bats (*Artibeus fraterculus*) captured between 2011 and 2015 in eastern Loja Province, Ecuador. Capture coordinates, dates and ecotypes (i.e., domestic, peri-domestic or sylvatic) are provided in Supplementary Tbl. 2.1 and associated protocols are detailed in previous studies led by the Center for Research on Health in Latin America (CISeAL)[385]. Individual parasite cells were cloned on solid medium to derive single-strain colonies following Yeo et al. (2007)[386]. Briefly, aliquots of $10^2 – 10^3$ epimastigote cells were mixed with 36 °C (molten) low melting point agarose and distributed over supplemented blood agar for stationary colony formation on petri dishes with the addition of 5% CO_2 at 28 °C for ca. three months. Successful microcolonies were then expanded in biphasic Novy-MacNeal-Nicolle (NNN) and liver infusion tryptose (LIT) media. Complementary to 19 non-cloned primary cultures, this process yielded 64 axenic monocultures for subsequent DNA extraction and sequencing.

2.3.2 DNA sequencing and variant discovery

Genomic DNA was extracted from 83 *T. cruzi* cultures by isopropanol precipitation (great thanks to Jalil Maiguashca for completing this step). DNA was sonicated and size-selected (median insert size = 198 nt; median absolute deviation = 69 nt) by covalent binding prior to paired-end sequencing on the Illumina HiSeq 2500 platform. To guide variant discovery from resultant 125 nt sequence reads, we optimized reference-mapping and SNP-calling pipelines using paired-end Illumina reads (kindly provided by Carlos Talavera-López, SciLifeLab, Sweden) for *T. cruzi* TcI X10/1 (termed TcI-Sylvio elsewhere in the text) against the newly available PacBio sequence for the same reference strain[306]. Based on comparisons with TcI-Sylvio mapping results from various configurations in SMALT v0.7.4[351] (we tested 12 – 14 kmer hash indexes and 2 – 8 base skip sizes), we chose to map samples using default settings (gap-open penalty = 6 and mismatch penalty = 4) in BWA-mem v0.7.3[353]. We then sorted alignments with SAMtools v0.1.18[387], marked PCR-duplicates with Picard v1.85[388] and identified single-nucleotide polymorphisms (SNPs) by local re-assembly with Genome Analysis Toolkit (GATK) v3.7.0[389] (also benchmarked for *L. donovani*[200]). Individual records produced by the HaplotypeCaller algorithm were subsequently merged for population-based genotype and likelihood assignment (GATK GenotypeGVCFs). Next, we calibrated variant filters by incrementally tightening thresholds for genotype quality (Q), read-depth (D) and local polymorphism density (C) until non-

reference homozygous SNP-calls for TcI-Sylvio reached asymptotic decay. We then applied a virtual mappability (V) mask to exclude variant-calls in unreliable mapping areas of the reference genome. Specifically, we generated synthetic, non-overlapping 125 nt sequence reads from the PacBio assembly and mapped these back to itself with the Genomic Multi-tool software suite[390]. Only variants from areas with perfect, i.e., singleton (V = 1), synthetic mapping coverage were kept for analysis. These regions represented areas of low sequence complexity and/or redundancy and made up large fractions of all reference chromosomes. With the above filters in place (Q > 1,500; 10 > D < 100; C < 3 SNPs per 10 nt; V = 1), samples retained tens of thousands of homozygous variant loci, whereas TcI-Sylvio Illumina vs. TcI-Sylvio PacBio showed just 58. Nevertheless, the guide-sample presented ca. 20,000 small insertions and ca. 1,000 small deletions relative to the reference. We placed an additional mask ±3 nt around these positions to avoid potential faults in the published genome. Final masking thus disqualified a total of 24 Mb (including all of chromosomes 17, 40 and 47) from polymorphism analysis. This highly conservative, diagnostic variant-screening approach also led us to exclude 24 low-depth samples for which genotypes could not be assigned at more than 40% variant sites. The final set of SNPs (in 59 samples) were annotated with snpEff v4.3t[391] using the TcI-Sylvio annotation file at TriTrypDB (http://tritrypdb.org/common/downloads/release-34/TcruziSylvioX10-1/gff/data).

2.3.3 Computational phasing of heterozygous SNP sites

Heterozygous SNP sites were phased over 30 iterations in BEAGLE v4.1[392]. The algorithm also imputes missing genotypes from identity-by-state segments found in the data. For haplotype co-ancestry and general comparative analysis, we restricted imputation to sites containing information for > 60% samples. Later, in windowed phylogenetic comparison, however, we refrained from genotype imputation, i.e., used only sites with genotypes called in all individuals of the dataset.

2.3.4 Detection of population genetic substructure

We used the Neighbor-Net algorithm in SplitsTree v4[91] to visualize genome-wide phylogenetic relationships among samples in split network representation. Neighbor-Net extends Satou and Nei's neighbor-joining algorithm to accommodate evolutionary processes such as recombination and hybridization that lead to non-treelike patterns of inheritance. We also optimized a general time-reversible (GTR) substitution model with ascertainment bias correction (for accurate branch lengths in the absence of constant sites) to construct phylogenies from proportions of non-shared alleles, i.e., considering two haplotypes per variant site. Haplotype concatenations were also used to derive a minimum-spanning network, the set of edges that links nodes (individuals) by the shortest possible cumulative

distance (i.e., maximum-parsimony). We inferred genetic subdivisions in the sample-set by unsupervised k-means clustering and discriminant analysis of principle components (DAPC)[338]. These analyses applied genetic distances as the proportion of non-shared genotypes at all variant loci (i.e., considering variants at the genotypic level), as did Neighbor-Net and subsequent measurements of F_{ST}. After phasing heterozygous SNP sites (see above), we used fineSTRUCTURE v2.0.4[339] to recover traces of identity-by-descent in similar haplotypes. This program was recently used to expose hybridization events in congeneric *T. congolense*[181], as well as to disentangle reticulate ancestries in the closely-related *L. donovani* complex[200]. Its Chromopainter algorithm constructs a semi-parametric summarization of co-ancestry among all pairs of individuals based on variable rates of haplotype-sharing and linkage disequilibrium across sample genomes. We applied fineSTRUCTURE over a uniform recombination map, running $6 \cdot 10^5$ Markov chain Monte Carlo (MCMC) iterations ($1 \cdot 10^5$ iterations burn-in) and $4 \cdot 10^5$ maximization steps in the final tree-building step. Following indications of mosaic inheritance in these analyses, we assessed phylogenetic (dis)continuity by comparing genotype-trees built for individual chromosomes using neighbor-joining as implemented in the 'ape' package v5.0[393] in R v3.4.1[394]. We also built distance matrices based on haplotypes phased without imputation (see previous section) to quantify changes in genetic similarity between windows within chromosomes.

2.3.5 Analyses of population genetic diversity and linkage

To assess group-level genetic diversity, we calculated site-wise nucleotide diversity (π), Watterson's theta (θ) and F_{IS} using the 'hierfstat' package v0.04-22[395] in R v3.4.1[394]. F_{IS} values rate heterozygosity observed within and between individuals, varying between -1 (all loci heterozygous for the same alleles) and 1 (all loci homozygous for different alleles). Values at 0 indicate Hardy-Weinberg equilibrium. We also measured rates of shared and private allele use (e.g., proportions of fixed heterozygous and singleton sites), assessed variant neutrality based on Tajima's D, quantified haplotype diversity by counting unique haplotypes per 10 – 100 kb, and scanned for long runs of homozygosity using VCFtools v0.1.13[396]. To determine linkage patterns within chromosomes 1, 5, 21 and 26 (the genome's best-mappable chromosomes) we recoded sample genotypes with values of 0, 1 or 2 to represent the number of non-reference alleles at each variant site. After filtering out all SNP-pairs separated by masked sequence (in effect, confining analysis to sites separated by < 100 kb), we measured linkage (r^2) as the correlation between genotypic allele counts and then binned r^2 into distance classes (from 0 to 100 kb in increments of 2 kb) to visualize relationships between map distance and linkage disequilibrium in R v3.4.1[394]. These analyses were also run separately on core sequence areas, as defined by areas of synteny

among TcI-Sylvio, *T. b. brucei* and *L. major* annotated at http://tritrypdb.org. Intra-haplotypic recombination is unlikely to accompany meiotic crossover events in these areas of the genome[306]. Furthermore, we considered the extent to which our multiple-clone sampling strategy (chosen to avoid underrepresentation of SNP linkage (or diversity) within infections might affect sample independence and variance-based statistical results. Linkage decay plots and other diversity metrics above were therefore also repeated using only one clone per infection source.

2.3.6 Estimation of meiotic vs. mitotic division

Following methods established to quantify complex microbial life cycles[397], we inferred the frequency of sex and clonality in *T. cruzi* isolates by comparing two different estimates of effective population size. The first estimate, N_ρ, is based on recombinational diversity observed in the sample. N_ρ represents the number of cells derived from mating, i.e., the number of zygotes present in the population, and is calculated as $\rho / 4r (1 - F)$, where ρ denotes nucleotide covariation between sites, r denotes rate of recombination per bp per generation, and F represents Wright's inbreeding coefficient. The second estimate, N_θ, is based on mutational diversity observed in the sample. N_θ represents the total population size, i.e., the number of cells irrespective of sexual or mitotic origin, and is calculated as $\theta (1 + F) / 4\mu$, where θ denotes nucleotide variation at single sites and μ denotes the rate of mutation per bp per generation. N_ρ / N_θ thus quantifies the frequency of meiotic reproduction in the population. To estimate this quotient from our sample, we derived θ from Watterson's estimator at non-coding sites and derived ρ based on reversible-jump MCMC likelihood curves generated by the interval program in LDhat v2.1[398]. We used $1 \cdot 10^7$ MCMC iterations with 2,000 updates between samples and block penalties set to five. We estimated r from the equation $r = 0.043 \cdot S^{-1.310}$ and μ from the equation $\mu = 2.5866 \cdot 10^{10} \cdot S^{0.584}$. These regression models were developed in Rogers et al. (2014)[298] based on the observation that genome size (S) correlates strongly to rates of recombination and mutation in unicellular eukaryotes. We validated ρ estimates by simulating input for LDhat in two ways. First, we created sequence alignment maps for ten non-recombinant individuals based on observed genotypes using BAMSurgeon v1.0.0[399]. Maps were set up for each individual by inserting fixed polymorphisms from the true sample set into TcI-Sylvio sequence reads, then spiking in random mutations at rates corresponding to the average number of pairwise differences in the observed data. Individual SNP records for the ten mutant alignment files were then compiled and merged in GATK as outlined above. In the second approach, we used fastsimcoal2 v2.5.2[315] to simulate ten non-recombinant and ten recombinant genotypes, applying r and μ from above equations to an effective population of 100,000 diploid individuals under a finite-sites model of evolution for chromosome 1. We also visualized

linkage patterns by measuring taxon topology weightings in windowed analysis. Taxon topology weightings provide a means to clarify phylogenetic structure by summarizing the extent to which tree topologies for a subset of samples contribute to the topology of the full tree[400]. We applied this concept to neighbor-joining trees constructed for overlapping 50 kb sequence windows in PhyML v3.1[401]. Topology weightings were calculated and plotted across chromosomes with loess smoothing (span = 0.125) using scripts provided at GitHub repository https://github.com/simonhmartin/twisst. These analyses prompted further sequence visualizations with Artemis v16.0.0[402] genome browser tool.

2.3.7 Chromosomal somy analysis

To estimate somy levels for each sample, we first measured mean-read-depth for successive 1 kb windows spanning each chromosome using default options of the 'depth' function from SAMtools v0.1.18[387]. We then calculated the median of these windowed-depth-means (m), i.e., a median-of-means (M_m), for each chromosome. After testing at various distribution points, we let the 30th percentile (p30) of (skewed) M_m values represent expectations for the disomic state, estimating copy number for each chromosome by dividing its M_m by the sample's p30 value and multiplying by two. This procedure produced estimates of disomy for all chromosomes of the TcI-Sylvio guide-sample and outperformed techniques based on different window-sizes as well as those refined according to sequence annotation (e.g., only single-copy genes) or mapping quality (data not shown). We validated cases of chromosomal copy number variation by plotting kernel densities of window-based somy estimates (i.e., density distributions of $2 \cdot m / p30$ of M_m calculated from each window), as well as by assessing raw depth and alternate allele frequencies across variant sites. True, whole-chromosomal trisomy, for example, should translate to chromosome-wide elevations in read-depth and reductions in minor allele contributions to ca. 33% (i.e., one 'A' and two 'B' alleles – and, in cases of tri-allelism, one of each 'A', 'B' and 'C' alleles) at all heterozygous (i.e., 'A/B/B' or 'A/B/C') sites. Intra-chromosomal amplification, in contrast, should create local shifts in read-depth and allelic composition within chromosomes. In follow-up assessment of temporal and sub-clonal ploidy variation, we re-sequenced three clones and derivative subclones on the Illumina NextSeq 500 platform. Subclones were obtained using the limiting dilution method as described in Messenger et al. (2015) (section 3.2.3)[58]. Briefly, logarithmic phase cell cultures were diluted to 50 parasites/ml in Roswell Park Memorial Institute (RPMI) 1640 medium, then divided into 200 µl aliquots across multiple 96-microwell plates. Wells presenting individual cells were incubated at 28 °C for ca. 6 weeks and further expanded in LIT. Subcloning work was performed by Jaime Costales and Jalil Maiguashca at CISeAL.

2.4 Results

2.4.1 Extensive genetic divergence between sympatric parasites

Paired-end sequence reads from 45 single-clone and 14 non-cloned *T. cruzi* cultures aligned to the reference assembly (*T. cruzi* TcI X10/1 Sylvio) at a mean depth of 27x, ranging between 13 and 64x (Supplementary Tbl. 2.1). The *T. cruzi* genome is highly repetitive, especially in the sub-telomeric regions[306]. Extensive optimization of variant filtration and masking was therefore undertaken before a total of 206,619 SNP sites could be robustly identified against the reference (see Methods). Including only single-clone *T. cruzi* cultures founded from individual parasites in the laboratory, 130,996 SNP sites were identified that clearly separated our samples into two highly distinct phylogenetic clusters within the small study area (Fig. 2.1, Supplementary Fig. 2.1, Supplementary Tbl. 2.1). Cluster 1 contained 15 of 17 clones isolated from triatomine vectors and mammal hosts captured in the community of Bella Maria. Cluster 2 contained 2 clones from Bella Maria, 11 clones from nearby Ardanza (ca. 7 km south), as well as 3 clones from Gerinoma and 12 from El Huayco study sites ca. 35 km northwest of Bella Maria. Two clones from Santa Rita (near El Huayco) associated to Cluster 1. Unsupervised k-means clustering further confirmed two major clusters (i.e., k = 2) among the samples, although mild improvements to model fit continued through to k = 6 (Supplementary Fig. 2.2).

To further detail parasite population genetic substructure within and across potentially multiclonal infections (multiple clones were often sampled from a single vector/host individual – see clone ID prefixes), we reconstructed each phased genome as a mosaic of haplo-segments sharing ancestry with other samples of the dataset[339]. In the resultant co-ancestry matrix (Fig. 2.2), which also includes isolates that had not been subject to solid-phase cloning, intensity of haplotype-sharing (see color scale) increased within both clusters relative to the spatial origin of each clone, with the exception of TCQ_3087 (sampled in Bella Maria but associated to Cluster 2) and TRT_3949 clones (sampled near El Huayco, but associated to Cluster 1).

Importantly, four non-cloned samples (TBM_3131_MIX, TBR_4307_MIX and TRT_4082_MIX, cultured from the triatomine species *Rhodnius ecuadoriensis*, and MBC_1529_MIX, cultured from the rodent species *Sciurus stramineus*) showed shared ancestry across Clusters 1 and 2. Clones derived from the same strains did not show shared ancestry. These data may indicate the presence of multiclonal infections in which parasites from these distinct groups co-occur in the same vectors and hosts (Supplementary Tbl. 2.1).

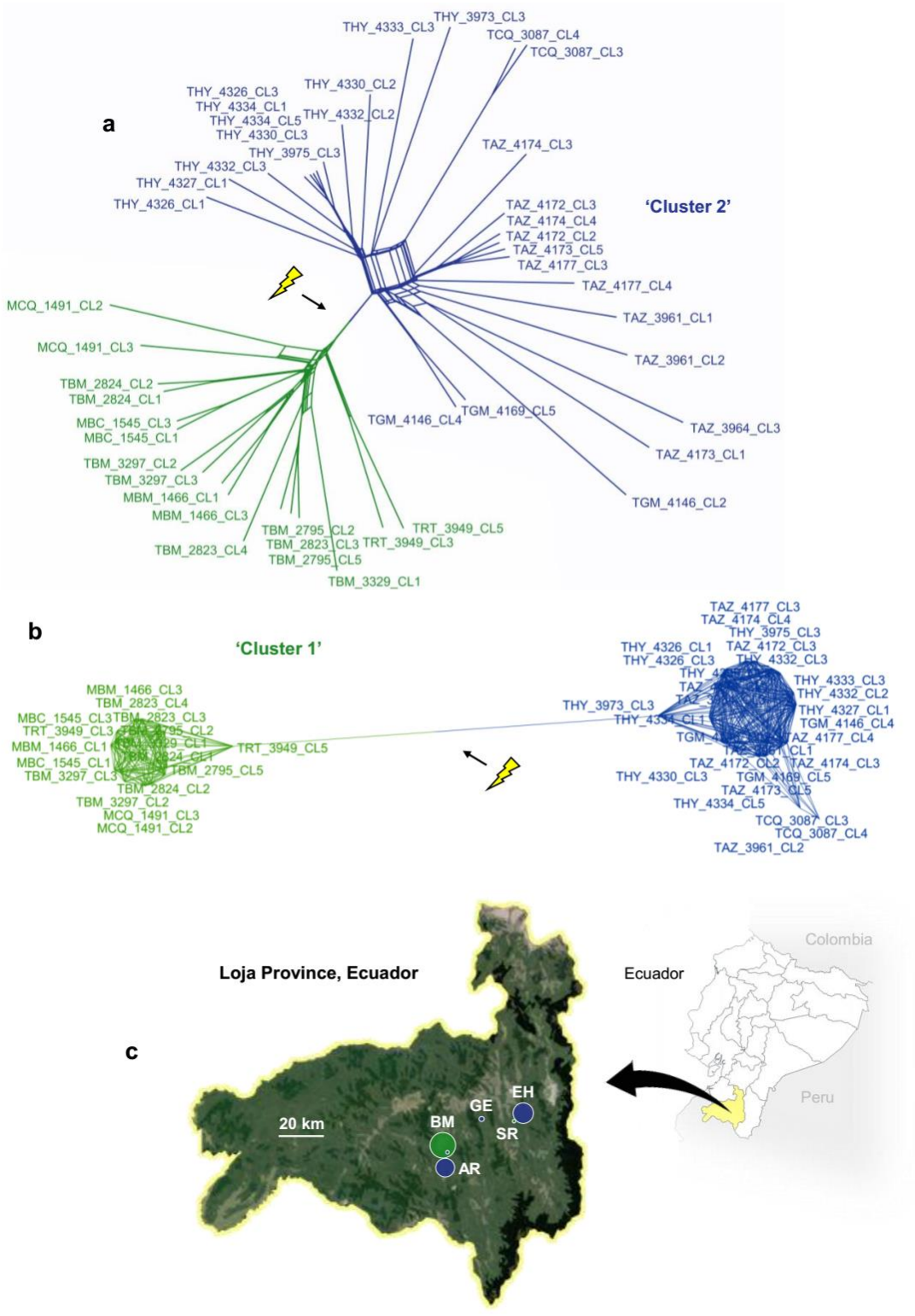


Figure 2.1 Phylogenomic relationships among *T. cruzi* I clones from southern Ecuador. **a** Data are represented as a split network by the Neighbor-Net algorithm[91]. Pairwise genetic distances are defined as the proportion of non-shared genotypes across all biallelic SNP sites for which genotypes are called in > 40 individuals (n = 68,449). Arrow (and flash) indicate a strong, unambiguous break in gene flow between two reticulate assemblages, Cluster 1 (green) and Cluster 2 (blue). Though non-treelike phylogenetic models are better suited to the data, a maximum-likelihood tree is also provided for comparison in Supplementary Fig. 2.1. **b** A minimum-spanning network[403] further illustrates the genetic disconnectivity between Clusters 1 and 2. Multi-furcating nodes are arranged such that cumulative edge distance is minimized among samples. Pairwise genetic distances are haplotype-based, defined as the proportion of non-shared alleles across all SNP sites for which genotypes are called for all individuals (n = 7,392). **c** Sampling regions in Loja Province, Ecuador, are abbreviated as BM (Bella Maria), AR (Ardanza), EH (El Huayco), SR (Santa Rita) and GE (Gerinoma). Point sizes correspond to sample sizes and colors correspond to cluster membership.

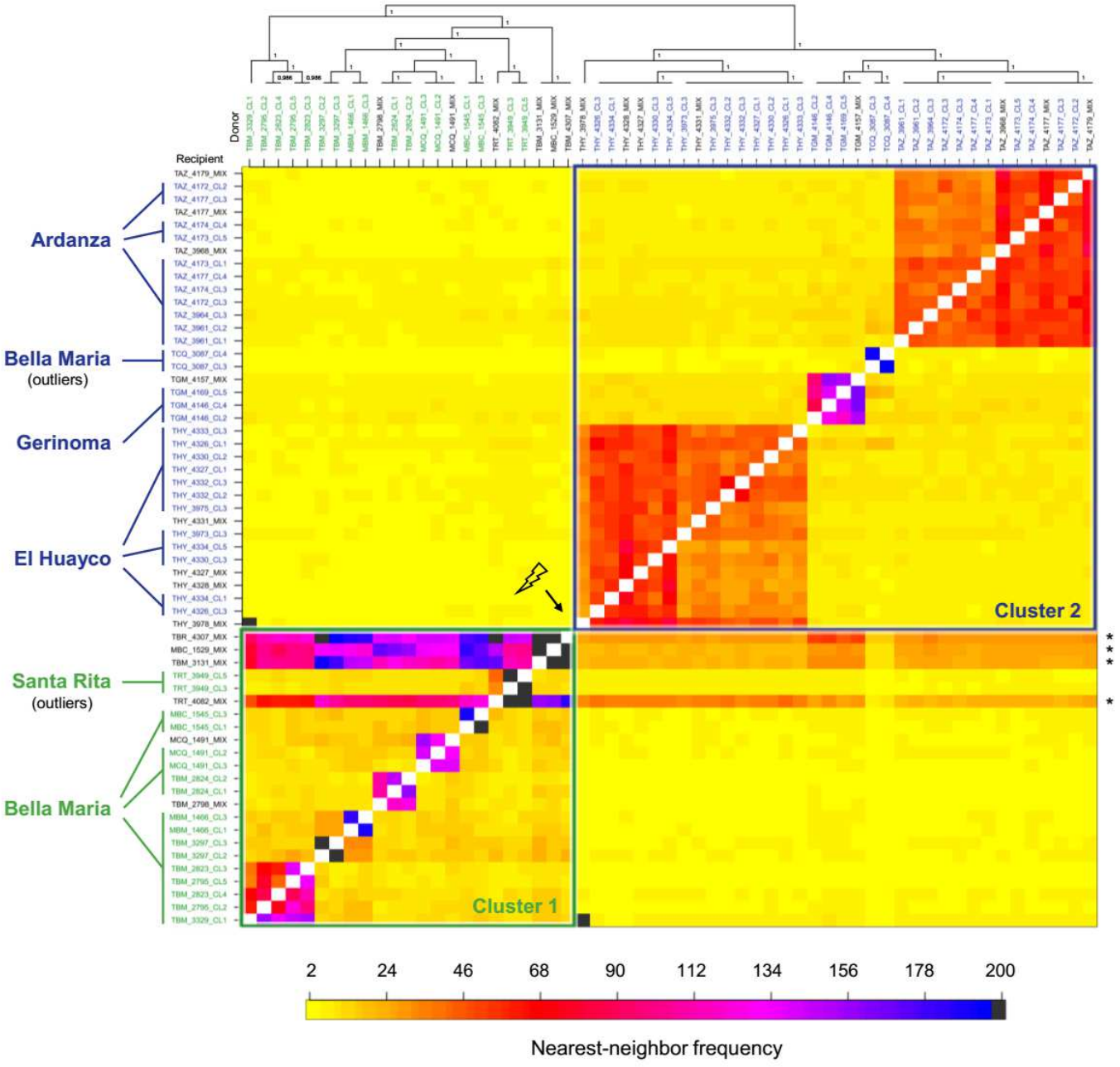


Figure 2.2 Haplotype co-ancestry among *T. cruzi* I clones from southern Ecuador. The heatmap of co-ancestry is based on a sorted haplotype co-ancestry matrix x_{ij}, which estimates the number of discrete segments of genome i that are most closely related to the corresponding segment of genome j. These nearest-neighbor relationships from fineSTRUCTURE[339] analysis are sorted such that samples clustered along the diagonal are those that most share recent genealogical events and pairwise comparisons outside of the diagonal indicate levels of genetic connectivity among these clusters. The matrix also includes 'genomes' of non-cloned *T. cruzi* cultures. Strong horizontal banding points to the accumulation of diversity from throughout the dataset in four of these original infections. Cell color represents the frequency of nearest-neighbor relationships for each sample pair, increasing from yellow (2) through red (68) and pink (134) to black (200). Four anomalous (outlier) samples are described further on in main text. Analysis uses 110,326 phased SNP sites.

2.4.2 Sympatric Mendelian and non-Mendelian genetic traits

To explore eco-evolutionary processes potentially underpinning sympatric divergence in *T. cruzi*, we established key metrics of population genetic structure at different sites. Among the 15 Bella Maria clones of Cluster 1, allele frequencies at variable loci matched those predicted for random mating, with estimated inbreeding coefficients predominantly near zero ($\overline{x}$ = -0.11, σ = 0.38; Supplementary Fig. 2.3) and 87,600 of 96,691 (91%) variant loci meeting expectations for Hardy-Weinberg equilibrium (Tbl. 2.1). Heterozygosity was unevenly distributed across each chromosome (see below), fixed at only 4% (2,134 / 58,102) polymorphic sites (Tbl. 2.1) and often interrupted by long runs (> 100 kb) of homozygosity (Supplementary Tbl. 2.2). Patterns of allelic diversity in Cluster 2 groups were highly distinct to those observed in Cluster 1. In El Huayco and Ardanza, departures from Hardy-Weinberg equilibrium were noted at 42% and 46% of total polymorphic sites (Tbl. 2.1). High levels of heterozygosity (Supplementary Fig. 2.3) extended continuously across all chromosomes (see below). Seventy-six per cent (44,945 / 58,980) of heterozygous loci occurred as fixed SNPs within El Huayco and 78% (45,287 / 58,392) occurred as such in Ardanza. Unlike in Bella Maria, long runs of homozygosity occurred in just two of 23 samples (1 instance each) in El Huayco and Ardanza (Supplementary Tbl. 2.2). Analysis repeated with only one random clone per vector/host showed the same strong contrasts between Clusters 1 and 2, but low sample sizes restricted significance tests (Supplementary Tbl. 2.3, Supplementary Fig. 2.4).

Table 2.1 Population genetic descriptive metrics for *T. cruzi* I clones from Bella Maria (Cluster 1), El Huayco and Ardanza (Cluster 2). Please see Supplementary Tbl. 2.3 for analogous results from analysis repeated with only one parasite clone per vector/host. Abbreviations: PS (polymorphic sites); π (median nucleotide diversity, per site); θ (median Watterson estimator, per site); MAF (within-group minor allele frequency); PRS (private sites); SS (singleton sites); HWE (Hardy-Weinberg equilibrium); HS (heterozygous sites).

Group (n)	PS	π	θ	PS at MAF > 0.05	PRS (vs. BM / EH / AR)	SS	PS in HWE	HS	Fixed HS
Bella Maria (15)	96691	0.09	0.001	48%	0 / 40177 / 40262	14013	87500	58102	2134
El Huayco (12)	80052	0.15	0.001	70%	23538 / 0 / 18016	4525	33980	58980	44945
Ardanza (11)	78325	0.16	0.001	71%	21896 / 16289 / 0	6064	35799	58392	45287

As well as extreme differences in the frequency and genomic distribution of heterozygous sites, other features of allelic diversity also diverged starkly among our sympatric study groups. Sliding window analyses of haplotype-sharing among individuals revealed, on average, much larger contiguous blocks of shared identity among samples from El Huayco and Ardanza (Cluster 2) than Bella Maria (Cluster 1) (Supplementary Fig. 2.5) despite lower

nucleotide diversity (π) in the latter group (Tbl. 2.1). Short blocks of shared identity among samples could be consistent with meiotic recombination in Bella Maria and we undertook further analyses to establish if this was the case.

2.4.3 Linkage decay and rates of meiotic recombination

In sexually recombining organisms, pairwise SNP-associations (r^2) are predicted to decay with map distance due to crossover that occurs between homologous chromosomes during meiosis. We plotted r^2 against pairwise map distance for all diagnostic SNP loci identified at Bella Maria. Fig. 2.3a depicts results for chromosome 1, with linkage declining sharply in the first few kilobases, then more gradually and approaching zero near 60 kb. Linkage decay was apparent on other chromosomes examined (chromosomes 5, 21 and 26 (Fig. 2.3b)). These chromosomes were selected based on their superior mapping quality, avoiding those with extensive masking (Supplementary Fig. 2.6). Decay curves were robust to reduction of the dataset to include only one clone per infection (Supplementary Fig. 2.7a) and also emerged in analysis restricted to core sequence regions (genes syntenous to *T. b. brucei* and *L. major*) (Supplementary Fig. 2.7b). In contrast to clones from Bella Maria, analyses of linkage decay for clones from El Huayco and Ardanza showed no relationship between r^2 and map distance. Rather, complete and intermediate linkage, as well as an abundance of random variant-associations, featured continuously through all distance classes on the same chromosomes surveyed in Bella Maria – e.g., chromosome 1 (Figs. 2.3c-d).

We estimated the frequency of meiosis (N_ρ / N_θ) in our study groups by comparing two different estimates of effective population size. The first estimate, N_ρ, is based on recombinational diversity observed in the sample and represents the number of cells derived from mating. The second, N_θ, is based on mutational diversity and represents the total number of cells, irrespective of sexual or mitotic origin (see Methods). As in linkage decay analysis, we considered the best-mapping chromosomes 1, 5, 21 and 26. Values of ρ for Bella Maria suggested ca. 3 meioses per 1,000 mitotic events in this group. In contrast, all approximations of ρ for El Huayco and Ardanza fell within confidence limits of the simulated, non-recombinant FSC_n control. These limits also contained $\rho = 0$ (Tbl. 2.2).

The intra-chromosomal recombination detected for Bella Maria was further explored by aligning individual windowed alternate allele frequency means (AAFM) among clones (Figs. 2.4a-b). As indicated previously (Supplementary Tbl. 2.2), sample genomes in Bella Maria presented intermittent patches of high homozygosity (where AAFM approaches 1), and these patches were often shared by variable subsets of clones (see windows with red fill

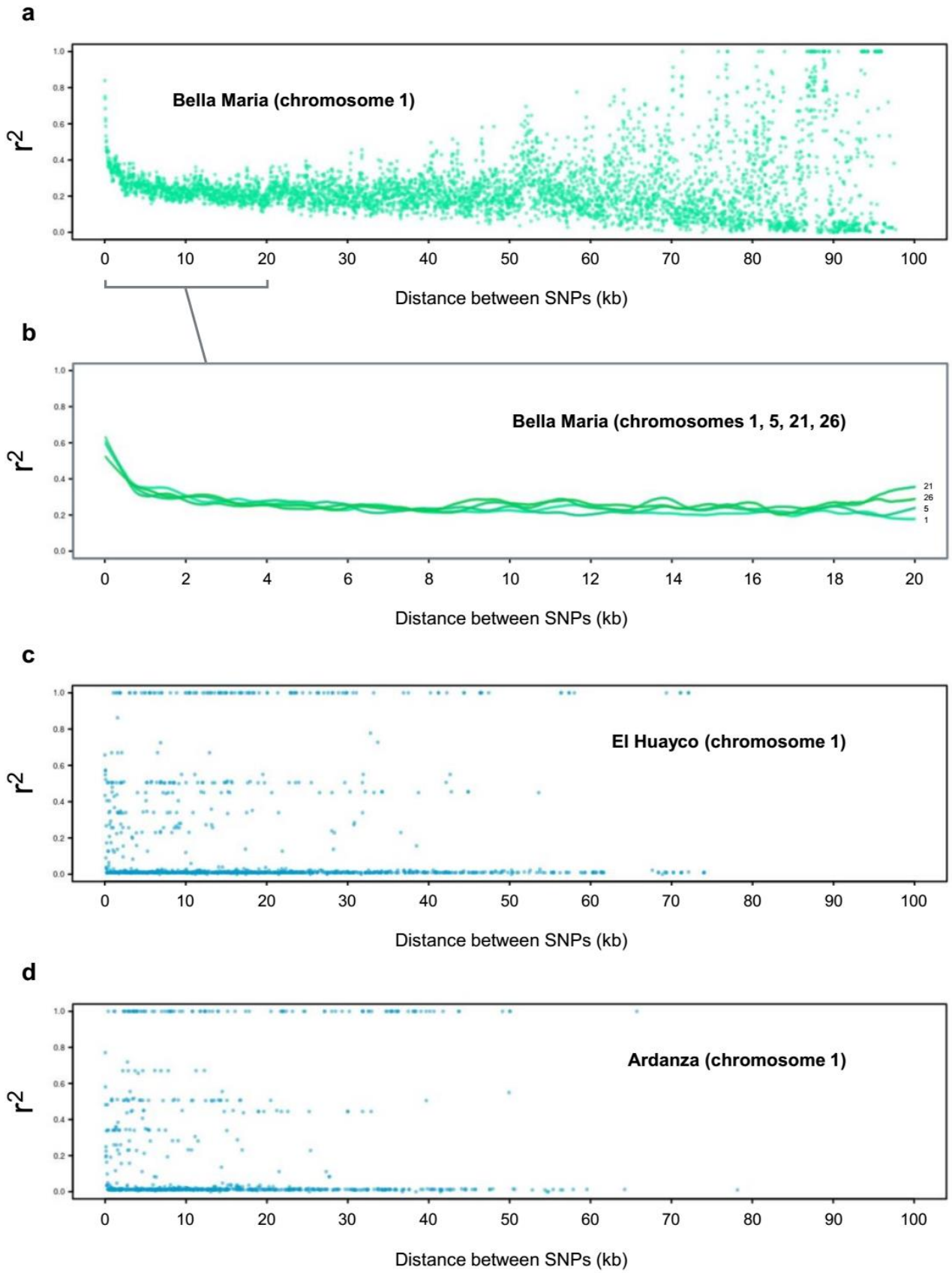


Figure 2.3 Linkage decay and different rates of recombination in *T. cruzi* I groups. **a** Decay of linkage disequilibrium on chromosome 1 for *T. cruzi* I clones from Bella Maria. Average pairwise linkage values (r^2) among SNP sites present in at least 90% individuals (n = 5,373) are plotted for map distance classes between 0 and 100 kb. **b** Local regression curves for the decay of linkage disequilibrium on chromosomes 1, 5, 21 and 26 for *T. cruzi* I clones from Bella Maria. **c-d** Lack of linkage decay on chromosome 1 for *T. cruzi* I clones from El Huayco (4,093 SNPs) and Ardanza (3,306 SNPs). Linkage values are plotted against genetic map distance as for Bella Maria above.

Table 2.2 Composite-likelihood approximation of the population recombination parameter ρ. Positive approximations of ρ for *T. cruzi* I isolates from Bella Maria differ from estimates derived for synthetic non-recombinant controls. The FSC_n control represents ten 3.1 Mb chromosomes simulated without recombination in fastsimcoal2[315]. The confidence interval around ρ estimates for FSC_n overlaps zero. It also overlaps estimates for El Huayco, Ardanza and BS_n, a second synthetic non-recombinant dataset generated by BamSurgeon[399] simulation approach (see Methods). Results from chromosome simulation with the recombination rate r set to $3.2 \cdot 10^{-4}$ (FSC_r) demonstrate the sensitivity of the LDhat[398] interval program applied to 100,000 diploid individuals under a finite-sites model of evolution.

Region	Group (n)	Median ρ (Morgans · kb^{-1})	95% Confidence Interval
Chr. 1	Bella Maria (15)	0.424	0.370 – 0.562
Chr. 5	Bella Maria (15)	0.549	0.400 – 0.647
Chr. 21	Bella Maria (15)	0.534	0.514 – 0.560
Chr. 26	Bella Maria (15)	0.357	0.338 – 0.392
Chr. 1	El Huayco (12)	0.004	0.004 – 0.004
Chr. 5	El Huayco (12)	0.002	0.001 – 0.004
Chr. 21	El Huayco (12)	0.002	0.001 – 0.003
Chr. 26	El Huayco (12)	0.005	0.002 – 0.016
Chr. 1	Ardanza (11)	0.005	0.005 – 0.005
Chr. 5	Ardanza (11)	0.003	0.002 – 0.003
Chr. 21	Ardanza (11)	0.002	0.000 – 0.004
Chr. 26	Ardanza (11)	0.002	0.001 – 0.002
Chr. 1, simulated	FSC_r (10)	78.886	77.023 – 80.739
Chr. 1, simulated	FSC_n (10)	0.001	0.000 – 0.007
Chr. 1, simulated	BS_n (10)	0.000	0.000 – 0.000

color in Figs. 2.4a-b). Given that SNP polymorphism was predominantly bi-allelic (< 1.5% sites with > 2 alleles) in Bella Maria as well as in Cluster 2, these patches corresponded directly to abrupt segmental increases in sequence similarity between clones (see SNP alignment in Supplementary Fig. 2.8, expanded in Supplementary Figs. 2.9 (chr. 1) and 2.10 (genome-wide)). Mosaic patterns of recombination between Bella Maria clones were confirmed by fluctuating intra-chromosomal genealogies established using sliding-window neighbor-joining topology weighting in Twisst[400]. Fig. 2.4c shows how strong support for various different tree topologies emerges sporadically throughout chromosome 1. Such mosaicism occurred genome-wide for most samples from Bella Maria (Supplementary Figs. 2.10 and 2.11b), but very infrequently in Cluster 2 (Fig. 2.4d, Supplementary Figs. 2.10 and 2.11).

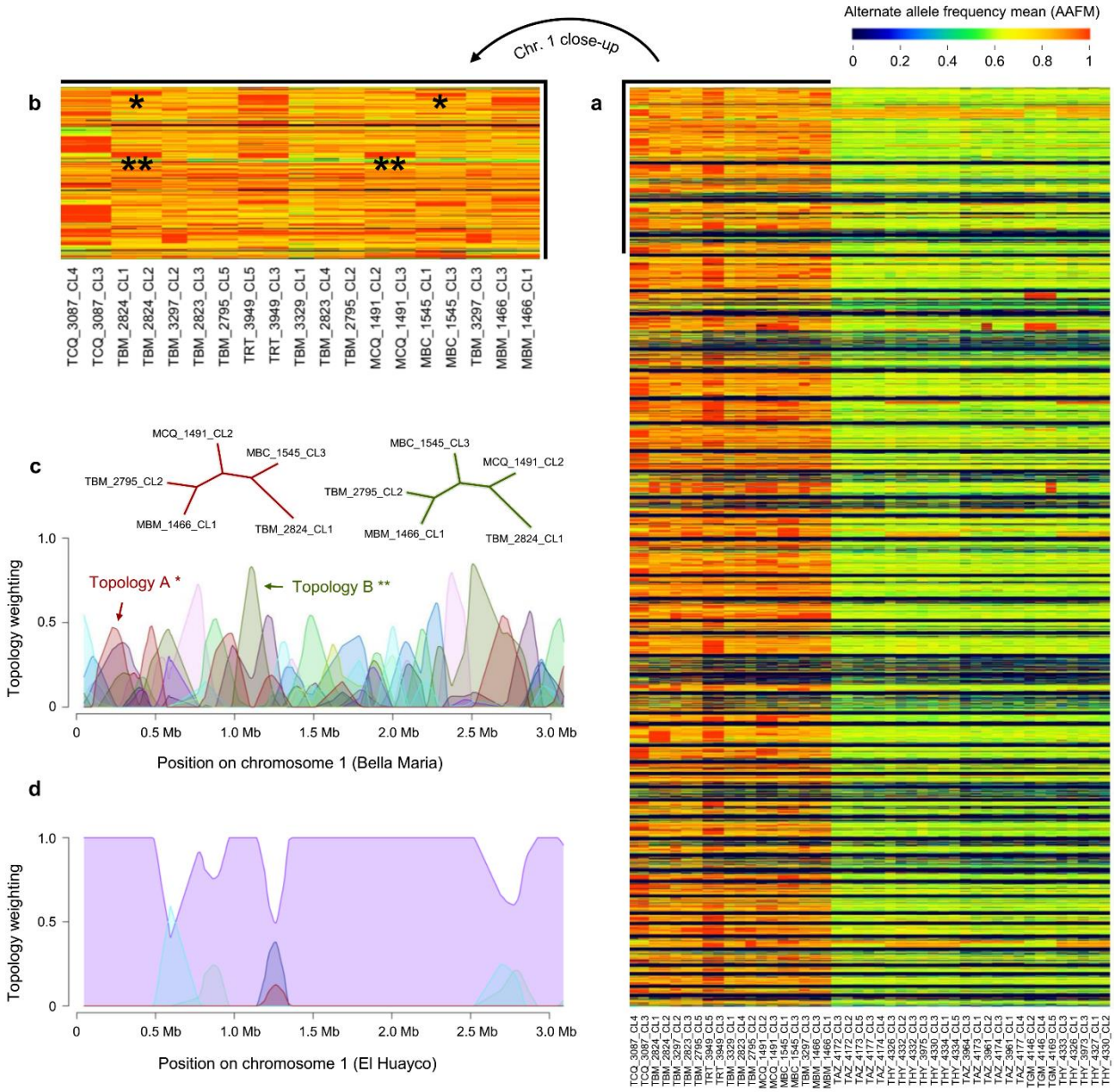


Figure 2.4 Genome-wide heterozygosity patterns and intra-chromosomal mosaics in *T. cruzi* I clones. In **a**, each column represents the genome of one clone, considering the dataset's total 130,996 SNPs. Rows within each column represent consecutive 5 kb sequence bins. Alternate allele frequency means (AAFM) determine the color of each bin – blue (0) through green (0.5) to red (1). Clones from Bella Maria tend to carry patchy homozygosity while those of Cluster 2 appear highly heterozygous throughout the genome. Isolated tracts of high homozygosity (i.e., red patches) shared between pairs of Bella Maria clones imply sudden sequence similarity and fluctuating phylogenetic relationships inconsistent with divergence through drift. **b** provides a close-up on chromosome 1. **c** and **d** demonstrate the impact of this intra-chromosomal mosaicism on the topology of phylogenetic trees derived in a sliding window across chromosome 1. Multiple incongruent topologies are present in Bella Maria (**c**), consistent with widespread genetic exchange. Only a single topology dominates for samples of El Huayco (**d**), consistent with limited genetic exchange in Cluster 2. An example of how AAFM heatmaps correspond to topology analyses is indicated in the heatmap close-up (**b**) and tree topologies in **c**: a shared red patch between TBM_2824_CL1 and MBC_1545_CL3 corresponds to neighbor-joining tree topology A in **c**. Later, near ca. 1,100 kb, a shared patch of high AAFM between TBM_2824_CL1 and MCQ_1491_CL2 begins. This patch occurs where tree topology B best describes phylogenetic relationships in Bella Maria. Topology B is identical to topology A except for the replacement of MBC_1545_CL3 by MCQ_1491_CL2 as nearest neighbor to TBM_2824_CL1.

2.4.4 Evidence of independent chromosomal ancestries in all groups

Apart from disrupting sequence patterns within chromosomes, sexual reproduction breaks up associations between chromosomes within the genome. Given sufficient population diversity, therefore, incongruent phylogenies are expected depending on the chromosome used to construct them. As one might expect given estimated rates of meiotic sex in this group, we encountered many such incongruences among Bella Maria clones belonging to Cluster 1 (Fig. 2.5). Intriguingly, ancestries among several clones from Cluster 2 also showed signs of incongruence at the chromosomal level (Fig. 2.5).

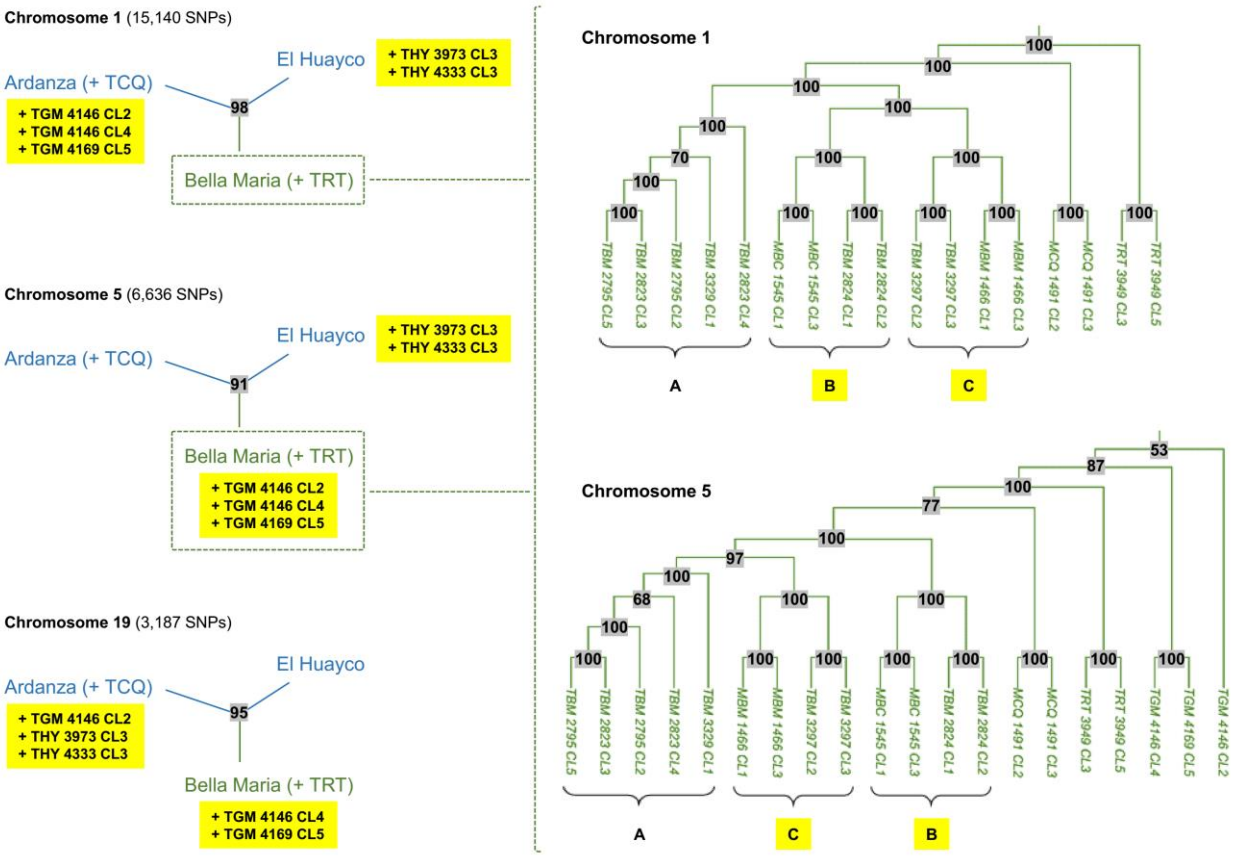


Figure 2.5 Incongruent trees exemplify independent chromosomal ancestries among *T. cruzi* I clones. Within individual sample genomes from Cluster 1 and Cluster 2, different chromosomes present different phylogenetic ancestries. For example, when neighbor-joining trees are constructed separately for chromosomes 1, 5 and 19, Gerinoma clones (prefix TGM) cluster with those from Ardanza on chromosome 1. On chromosomes 5 and 19, they cluster with clones from Bella Maria. El Huayco clones THY 3973 CL3 and THY 4333 CL3 also join the Ardanza clade on chromosome 19. Within cluster 1 (right panel), chromosome 1 presents a monophyletic clade composed of MBC_1545 + TBM_2824 (labelled B) and MBM_1466 + TBM_3297 clones (C). TBM_2795 + TBM_2823 + TBM_3329 clones (A) form an outgroup. These clades rearrange on chromosome 5, where A changes places with C. The A+B clade occurs again on chromosome 19, while the B+C group makes appearances on chromosomes 9 and 16, etc. Discrepant phylogenies such as those highlighted here occur in various chromosomal comparisons throughout the genome. Nodes are labelled in grey with support values from 100 bootstrap replicates. Green denotes the Cluster 1 clade. Blue denotes Cluster 2. Yellow highlights unstable phylogenetic positions among different chromosomes. Branch lengths are not proportional to genetic distance.

We also recognized these varying affinities among El Huayco, Ardanza and Gerinoma clones in discriminant analysis results for higher k-means solutions (e.g., see individual membership probabilities for $k = 5$ in Supplementary Fig. 2.2b) and noted occasional shifts to common homozygosity unrelated to coding vs. non-coding sequence annotation in painted genomes (e.g., see chromosomes 6, 14 and 41 in Supplementary Fig. 2.10). Whilst subtle, such segmental changes argued against divergence in strict isolation among Cluster 2 clones: if not classic chromosomal reassortment, some form of introgression appears to have occurred in this group.

2.4.5 Signatures of hybridization in highly heterozygous genomes

Of 80,052 SNP sites that differed from the TcI-Sylvio reference genome in El Huayco, 62,036 also differed in Ardanza, and > 50% of this polymorphism occurred as fixed heterozygous loci across the two groups. These observations, supported by population genetic statistics (see Tbl. 2.1) and phylogenetic similarity (Figs. 2.1 and 2.2), provided indications of potential shared ancestry across clones of Cluster 2, and possibly a hybrid origin of this group.

To further explore potential hybrid origins of Cluster 2 clones, we first expanded our previous within-group windowed haplotype analyses to include comparisons of *T. cruzi* clones between El Huayco and Ardanza groups (Supplementary Fig. 2.12a). These between-group pairwise comparisons of phased SNPs exposed the frequent co-occurrence of haplotype polymorphism in clones from Ardanza and El Huayco (but not clones from Bella Maria). Reaching up to 180 kb, shared haplotype segments appeared similar in size to those found in pairwise comparisons within Bella Maria (Supplementary Fig. 2.12b) and suggest recent genetic connectivity throughout Cluster 2. This between-group connectivity is also apparent upon careful examination of the co-ancestry matrix in Fig. 2.2.

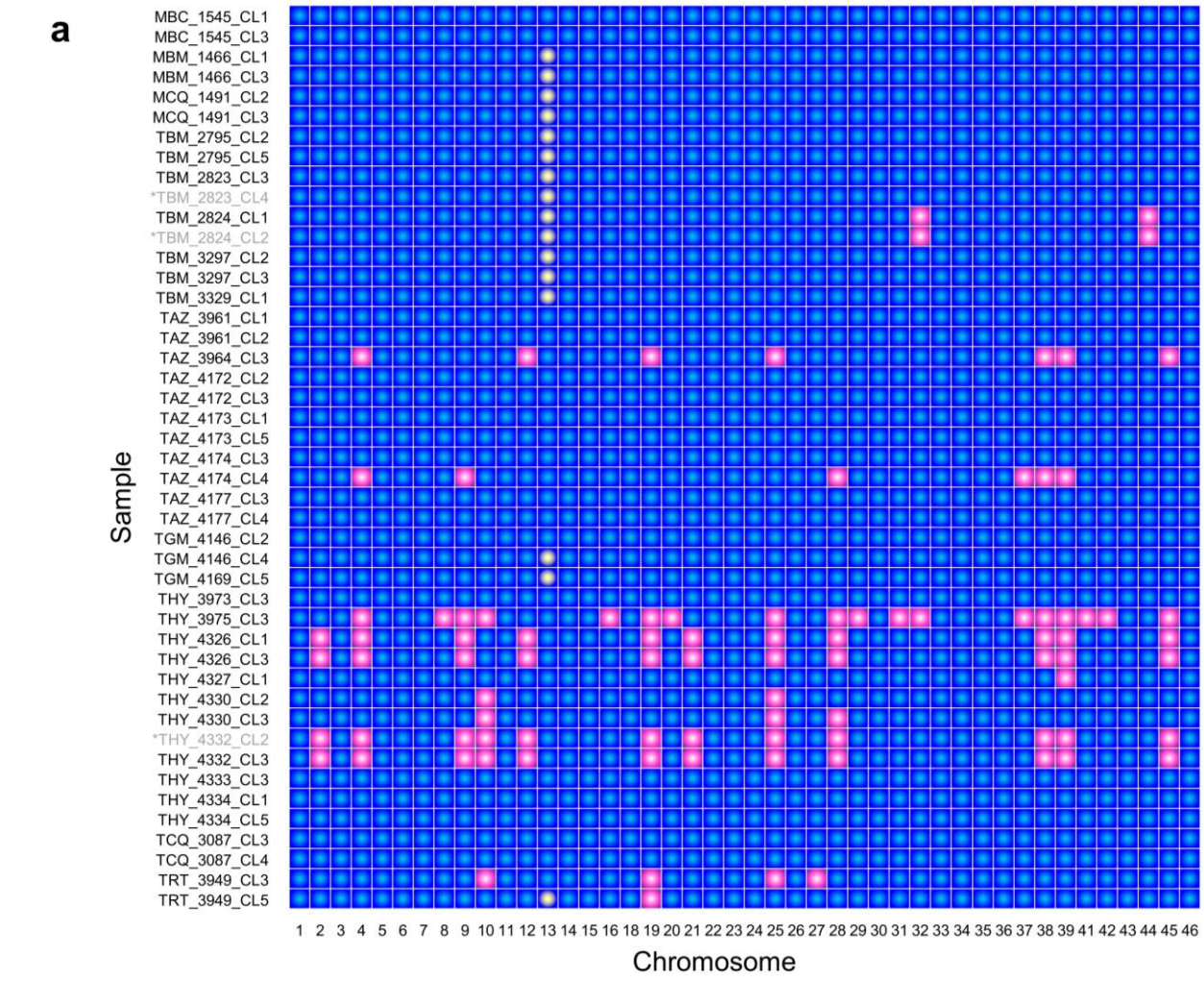
Patches of low differentiation observed in pairwise comparisons within El Huayco (Supplementary Fig. 2.12c) and within Ardanza (Supplementary Fig. 2.12d) often involved both haplotypes. Unlike cases of haplotype-sharing described above, these were stretches of similar or identical heterozygous diplotypes (i.e., phased regions in which haplotype A occurs in samples 1 and 2, and also haplotype B occurs in samples 1 and 2) interrupting otherwise dissimilar heterozygous sequence between two clones. Such diplotype sharing within groups of Cluster 2 extended far beyond 180 kb, often to the end of the chromosome. The same phenomenon was rarely observed in comparisons of clones within Bella Maria (Supplementary Fig. 2.12f) and could point to the passage of Cluster 2 clones through an ancestral polyploid state.

To characterize ploidy variation in Cluster 2, somy analysis was undertaken (Fig. 2.6). Chromosome-wide deviations in variant allele fraction and total read-depth suggested full-chromosome trisomies in ten samples (Figs. 2.6a-b), with highest rates in THY_3975, THY_4326 and THY_4332 clones (> 10 trisomies each). Of 21 chromosomes with apparent trisomy, ten appeared trisomic in ≥ 5 samples, with similar biases apparent in El Huayco and Ardanza (e.g., chromosomes 19, 25 and 39). To explore the intra-clonal stability of somy variation over time, we re-sequenced three aneuploid clones after sample cryo-preservation and re-expansion in liquid culture (results are denoted with T2 suffix). While inferred karyotypes of THY_4326_CL1_T2 and TAZ_4174_CL4_T2 matched initial results (Supplementary Figs. 2.13a-b), several aneuploid chromosomes in THY_4332_CL3 appeared to have reverted to the disomic state by time T2 (Supplementary Fig. 2.13c). We also examined ploidy in subclones of each re-passaged clone. No significant variation occurred among the three subclones obtained from THY_4332_CL3_T2 nor between the two obtained from THY_4326_CL1_T2, each with a karyotype matching that of the parental clone (Supplementary Figs. 2.13b-c). Somy estimates for the single subclone obtained from TAZ_4174_CL4_T2, however, were inconsistent to the progenitor karyotype (Supplementary Fig. 2.13a).

In contrast to karyotypic variation in Cluster 2, we found minimal rates of aneuploidy in Cluster 1 (Bella Maria). With the exception of TBM_2824 clones (trisomic for chromosomes 32 and 44), no Bella Maria clones showed increased somy despite similar levels of intra-chromosomal read-depth variation as clones from Cluster 2. Interestingly, most Bella Maria genomes showed severe reductions in sequencing coverage over chromosome 13. Such reductions did not occur in El Huayco or Ardanza (Fig. 2.6a). Somy plots for all initial samples are provided in Supplementary Fig. 2.14.

2.4.6 Mysterious migrants imply further forms of genetic exchange

Two samples in the dataset stand out as clear migrants with idiosyncratic genomic features that indicate the possibility of further genetic exchange events. TRT_3949 (sampled near El Huayco but associated to Cluster 1) and TCQ_3087 clones (sampled in Bella Maria but associated to Cluster 2) were the only samples for which geographic and nuclear phylogenetic neighbors did not match (Fig. 2.1, Supplementary Fig. 2.1, Supplementary Tbl. 2.1). These clones also provided the dataset's only cases of discordant nuclear vs. mitochondrial phylogenies: TRT_3949 clones carried a maxicircle genotype otherwise found only in Cluster 2 and TCQ_3087 clones carried a maxicircle genotype highly divergent to any other observed in the study area (Supplementary Fig.15a; see also *cytochrome b* alignment in Supplementary Fig. 2.15b).



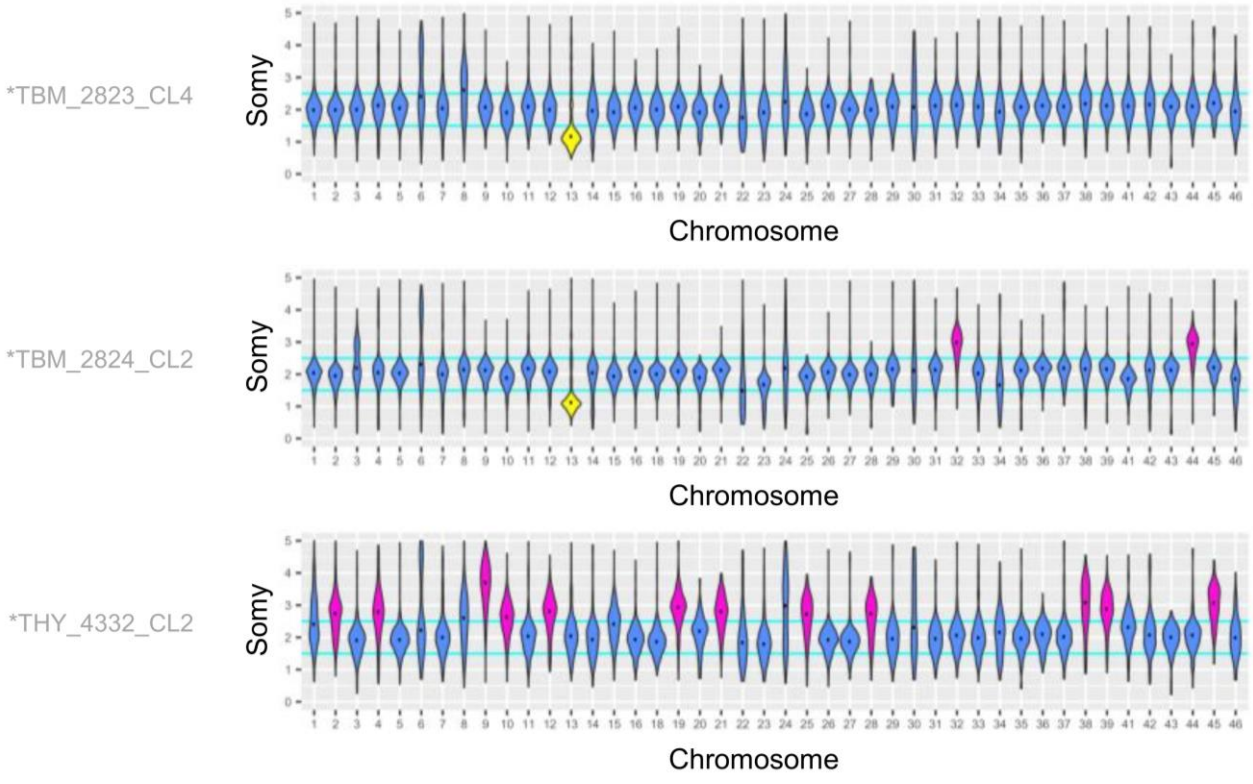


Figure 2.6 (continues on next page)

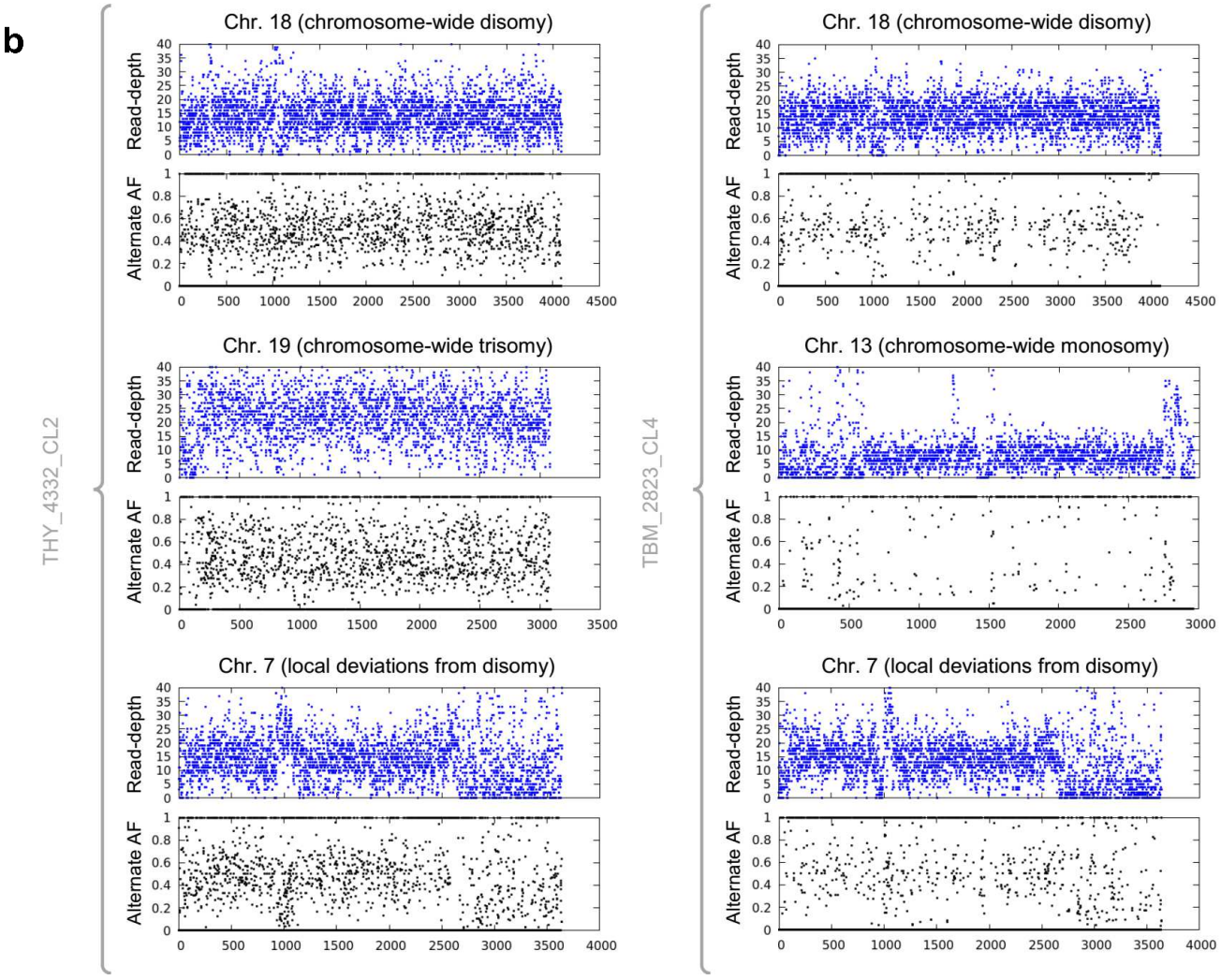


Figure 2.6 Group-level aneuploidy among *T. cruzi* I clones. **a** We distinguished chromosomal and intra-chromosomal copy number variation by evaluating kernel density distributions of window-based somy estimates (see Methods). These distributions suggest multiple cases of whole-chromosome somy elevation (highlighted in pink) for El Huayco clone THY_4332_CL2 (bottom violin plot). Several clones from El Huayco and Ardanza present similar patterns (see Supplementary Fig. 2.14 for more violin plots), as summarized in the heatmap. Read-depth densities suggest few cases of whole-chromosome somy elevation for clones from Bella Maria (e.g., see violin plots for TBM_2823_CL4 and TBM_2824_CL2). However, mapping coverage drops dramatically (yellow) on chromosome 13 in most clones of this group. **b** Chromosome-wide shifts in sequence read-depth (blue) and alternate allele frequency (AF, black) support whole-chromosome aneuploidies inferred from density distributions above. In El Huayco clone THY_4332_CL2 (left column), for example, read-depth is elevated over the entirety of trisomic chromosome 19 (sequence positions are plotted on the x-axis). Alternate allele frequencies at heterozygous sites also distribute around values of 0.33 and 0.67 on this chromosome (as compared to frequencies around 0.50 on disomic chromosome 18). Cases of intra-chromosomal copy number variation for sample THY_4332_CL2 are marked by local shifts in read-depth and alternate allele frequency on chromosome 7. Comprehensive read-depth reduction on chromosome 13 is exemplified for Bella Maria clone TBM_2823_CL4 (right column). Alternate allele frequency values of 0 (indicative of the reference allele) predominate on this chromosome. Patterns on chromosomes 7 and 18 also point to intra-chromosomal copy number variation and stable disomy, respectively, for the TBM_2823_CL4 clone.

These apparent migrants were also exceptional in nuclear sequence alignment: within a single individual, some chromosome segments appeared to derive from Cluster 1, others from Cluster 2. For example, on chromosome 1, TCQ_3087 shared a heterozygous patch with Ardanza clones between approximately 785 and 920 kb. At ca. 1,117 kb, sequences were similar to those of Gerinoma (TGM) clones and then, at ca. 1,122 kb, similar to El Huayco clones. A long stretch of similarity to Cluster 1 ensued at ca. 1,285 kb (Supplementary Fig. 2.9). TCQ_3087 and TRT_3949 clones were also the only samples for which homozygosity was widespread throughout the nuclear genome (Fig. 2.4a). Making up just 10% total polymorphic loci, bi-allelic SNPs were found restricted to scattered patches. High levels of overall homozygosity observed in these clones could not be attributed to certain chromosomes or to deviations in read-depth.

2.5 Discussion

2.5.1 Principle findings

Our comparative genomic analysis of 45 biological clones from an area of endemic transmission supports the remarkable conclusion that a) *T. cruzi* undergoes meiosis and b) that grossly disparate reproductive strategies and rates of genetic exchange occur simultaneously at a single disease focus.

In a subsection of the region (Bella Maria), signs of regular meiotic sex are markedly clear. Genome-wide allele frequencies occur at Hardy-Weinberg equilibrium and ancestries among individuals fluctuate from chromosome to chromosome. Parasite genotypes on individual chromosomes appear equally mosaic: linkage between polymorphisms clearly correlates with map distance, disequilibrium plummeting within just a few hundreds of bp. We gauge that the meiosis driving these patterns of diversity occurs more than once every 1,000 reproductive events in Bella Maria. In nearby El Huayco, Ardanza and Gerinoma groups, meiosis appears essentially absent. Instead, these groups exhibit high levels of heterozygosity across the entire genome. We do detect discordant chromosomal phylogenies among these parasites, but recombination estimates within chromosomes match those for simulated, non-recombining controls and there are no signs of intra-chromosomal linkage decay. Alongside excess heterozygosity, several El Huayco and Ardanza clones also present extensive aneuploidy as well as long blocks of near-identical diplotypes.

2.5.2 General discussion

The strong signatures of meiotic sex we report from Bella Maria redefine our understanding of *T. cruzi* biology and, alongside data from *T. b. brucei*[380] and *Leishmania*[244], indicate that this mode of genetic exchange is ancestral among medically important trypanosomatids[380].

We previously advised caution to those applying generalized theories of clonal evolution (e.g., PCE[162]) to parasitic protozoa[6]. Our revelations around *T. cruzi* population genomic structure in this study broadly support our case. Nonetheless, meiotic sex has never been observed in the laboratory and multiple aspects of meiosis in *T. cruzi* remain obscure[171]. The site of genetic exchange (vector or host) in *T. cruzi* is still not known, for example, nor is it understood from which parasite life cycle stage gametes might develop. In contrast, *T. b. brucei* gametes have been characterized in the salivary glands of tsetse flies and a mechanism for subsequent cytoplasmic fusion described[380]. Clearly much basic research remains to be done.

The distribution of genetic diversity we describe in Cluster 2 suggests that meiosis is largely absent among these strains. Patterns of heterozygosity recently observed in *T. b. gambiense* were attributed to the Meselson Effect[305,404], whereby mutations accumulate in the absence of recombination between homologous chromosomes during long-term clonality. The high levels of heterozygosity we observe in Cluster 2 differs in important ways from the *T. b. gambiense* dataset and from predictions of the Meselson Effect[305,404]. For example, discontinuities in genetic differentiation among individuals, instead of occurring as stretches of absolute homozygosity on disomic chromosomes as they did in *T. gambiense*[305], occur in our dataset as shared patches of heterozygosity among geographically distinct groups (e.g., El Huayco and Ardanza). Furthermore, we see no evidence of accumulation of private heterozygous sites within individuals as one might expect during long term asexual propagation – rather, over 50% of heterozygous sites are shared among samples in Cluster 2. If long-term asexuality is a poor explanation for heterozygosity in our dataset, an ancestral outcrossing event could perhaps have played a role.

In Cluster 2, we observed incongruent phylogenies between different chromosomes, but no evidence for linkage decay within individual chromosomes. In the only genetic exchange event observed experimentally in *T. cruzi* to date[171], parental genomes fused to tetraploid hybrids and then began erosion back toward the disomic state. This fusion-then-loss process resembles that in parasexual pathogenic fungi (*Candida* spp.) and allows for independent chromosomal ancestries without intra-chromosomal linkage decay[405]. Moreover, gene conversion in tetraploids can produce long tracts of increased identity on both homologs (i.e., the diplotype-sharing we refer to in our results) without loss of heterozygosity upon reduction to the disomic state[406]. This is especially true when genome erosion is biased against the retention of similar homologs[173], a condition that aligns with our results (e.g., we observed elevations to average homozygosity in just two chromosomes (Supplementary Fig. 2.16), not in fifteen (33%) as would be expected in the case of random chromosome loss). Aside from parasexual mating, however, polyploidization via failed meiotic division might

also explain aneuploidy levels observed in El Huayco and Ardanza. Given that failed chromosome segregation typically involves failed crossover[407], this explanation also reconciles a lack of linkage decay in Cluster 2. A third possibility, high levels of aneuploidy via frequent asymmetric chromosome allotment in mitotically dividing nuclei[279], also finds direct support in this dataset. Unlike occasional accounts of stable aneuploidy in the *Trypanosoma* genus[174,408,409], we detected short-term somy reductions in one of three re-sequenced aneuploid clones and also found evidence for sub-clonal ploidy variation, often termed mosaic aneuploidy in *Leishmania* research[279]. Congruent aneuploidies observed in closely-related Cluster 2 genotypes may thus reflect strain-specific or pre-adapted amplification programs as in *Leishmania* spp.[200,274].

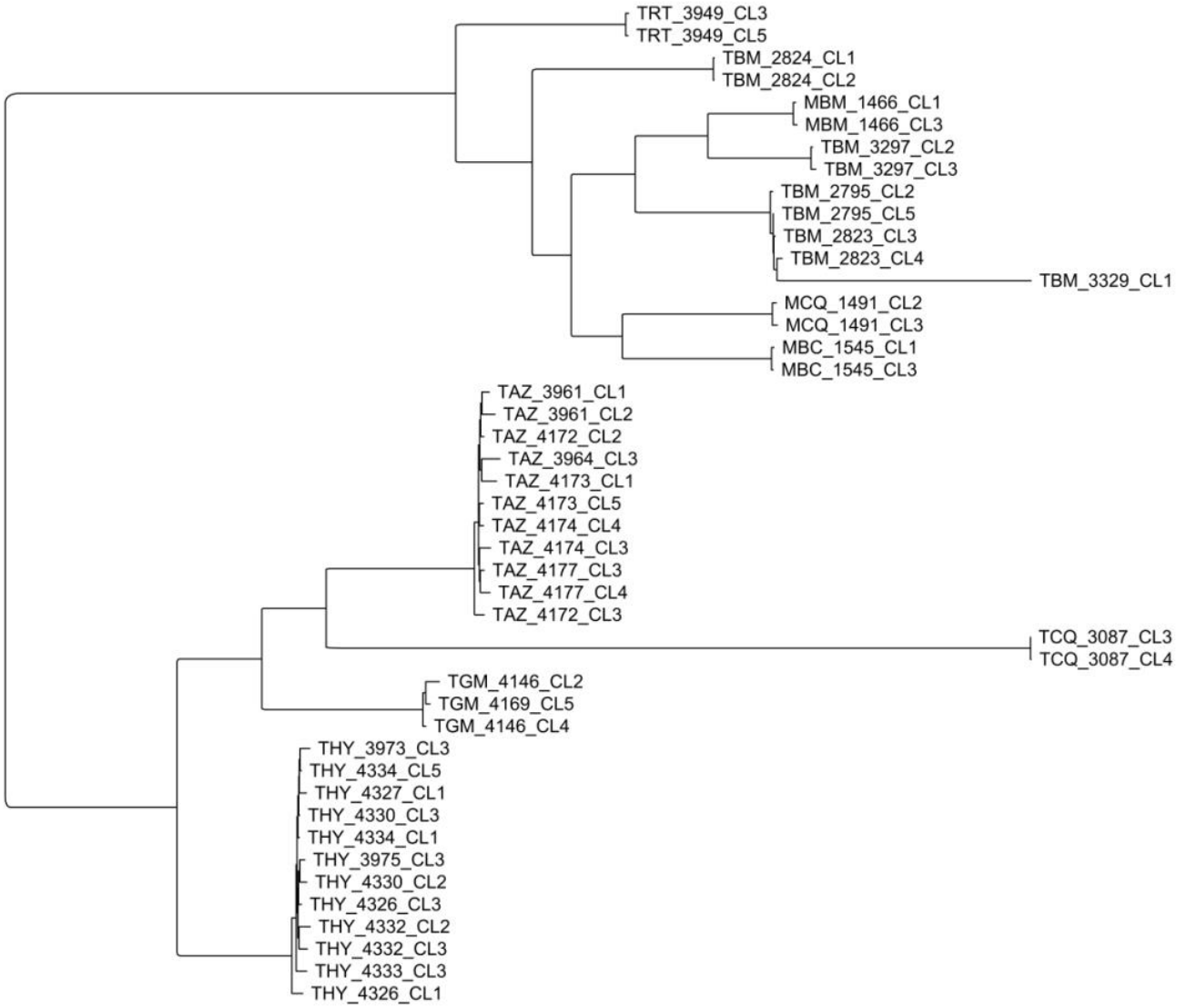
The ecological and evolutionary drivers of distinct but sympatric reproductive modes in *T. cruzi* are not clear. While *T. cruzi* is able to infect a remarkable variety of insects and vertebrates, its stercorarian transmission route is highly inefficient. *T. cruzi*'s vectors and hosts vary immensely in transmission competence and availability and occupy an array of disparate niches (including the domestic-sylvatic interface)[410–412]. The parasite's life cycle thus likely represents a continuum of bottlenecks linked to frequent local extinction and recolonization events that increase levels of genetic drift and identity by descent (IBD). It may thus come to less surprise that observations of diffuse hybrid clonality around a restricted focus of sex in Bella Maria resemble spatio-temporal patterns of heterogony demonstrated in various other metapopulation systems[413–416]. Facultative sex often coincides with strong metapopulation structure, in which sexual variants are predicted to occupy core habitat (where population subdivision and inbreeding depression are minimized) while asexual variants disperse more freely without fitness costs from high IBD during frequent founding events[417]. Extensive asexual dispersal eventually brings divergent lineages into contact, creates potential to mate, form heterotic offspring, and reset clonal decay. Divergent homologs, however, may impair canonical sex when F_1 hybrids mate[245,418,419]. We noted mass elevation of Tajima's D in Cluster 2 of this study (Supplementary Fig. 2.17) and this offers further support for both hybridization and bottlenecked clonal propagation in generating an excess of intermediate-frequency variants over El Huayco and Ardanza[420,421]. Such excess, however, can also arise in simple (e.g., island model) demographic scenarios when mating becomes very scarce, whereupon the influence of demographic changes on the site-frequency spectrum becomes difficult to disentangle by current methods of inference[422]. Nevertheless, large patches of low differentiation observed in this study suggest a relatively recent contribution of hybridization to allelic divergence in El Huayco and Ardanza. Spatially correlated genetic substructure and low effective population sizes further attest the role of metapopulation dynamics in structuring genetic diversity in these groups.

Our genotype- and haplotype-based summaries of co-ancestry indicate that the meiotic parasite group in Bella Maria is genetically segregated from others with distinct reproductive histories in nearby El Huayco and Ardanza. Genetic discontinuity occurs consistently for samples collected within a few kilometers distance and despite evidence for vector/host co-infection and migration between divergent groups. Putative migrants, possibly the progeny of these divergent groups, exhibit extensive (nuclear) homozygosity and, in the case of TCQ_3087 clones, extreme maxicircle divergence and very high maxicircle read-depth. Such observations are reminiscent of *L. major* crosses formed in non-native vectors[245] and of irregular, biparental mitochondrial inheritance in *T. b. brucei*[423].

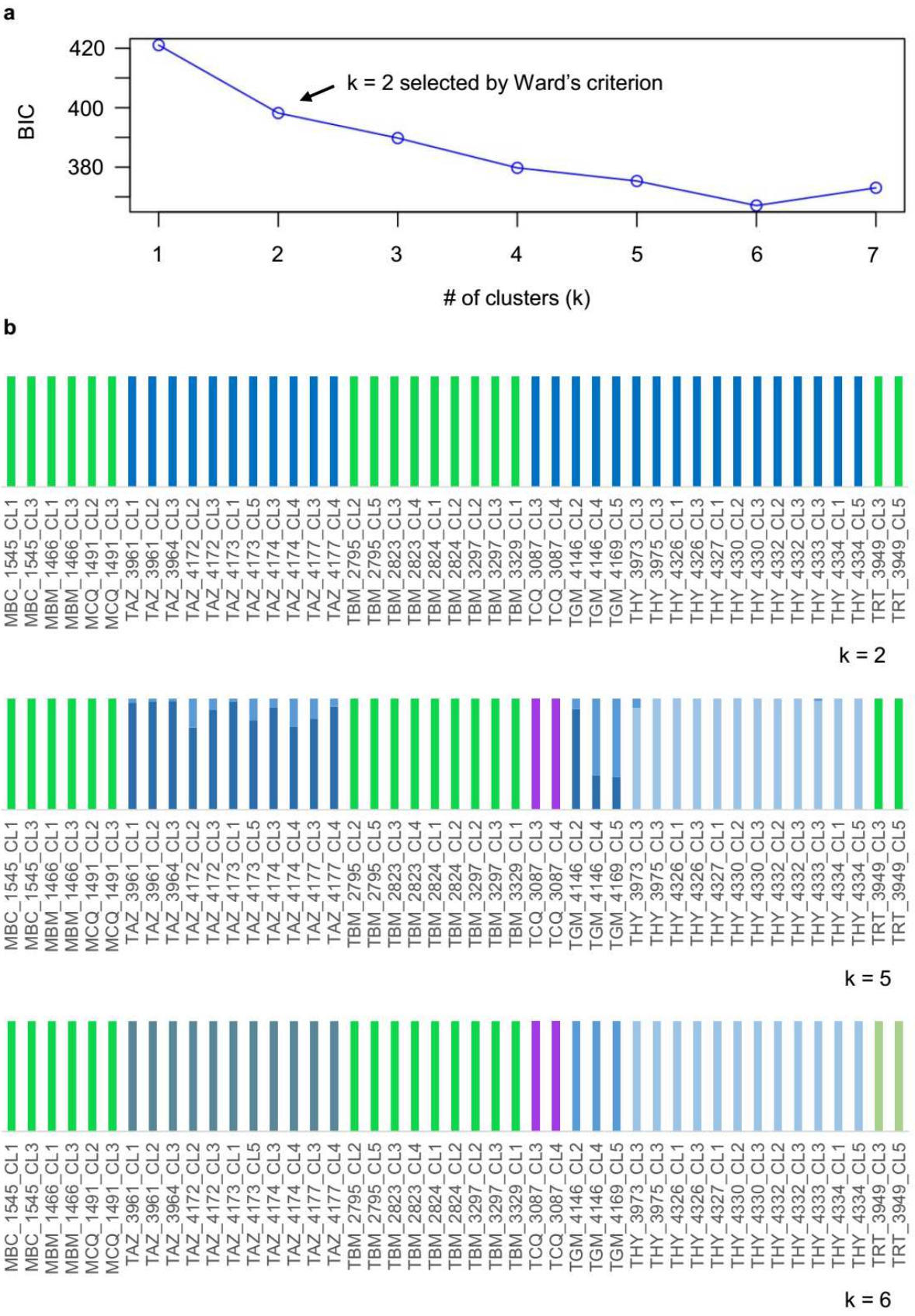
Unexpected and poorly repeatable hybrid genomes have arisen on a number of occasions in experimental Tritryps research[171,244,424]. Sensitivity to cryptic biochemical cues is clearly high, but the molecular signals that incite recombination and control mating compatibilities within these species remain essentially unknown[425]. Our observations from the field do not identify such mechanisms but provide many relevant questions to explore. For instance, do ploidy barriers segregate transmission cycles in *T. cruzi*? Is certain monosomy (e.g., recall chromosome 13 in Bella Maria clones) associated with mating locus activation and sex? Is high homozygosity a direct result of improper mating or a subsequent effect (gene conversion, selfing, etc.)? What are the adaptive processes that underpin switching between different reproductive modes?

Our work presents hard evidence for meiotic sex in *T. cruzi*, as well as evidence for widespread clonal expansion, after episodic hybridization events. Recent evidence for sex obtained from Arequipa, Peru, in contrast, cannot be reliably distinguished from complex patterns of gene conversion in a fully clonal population[176]. Complex mating structures are of acute relevance to Chagas disease control. Recombination implies that important epidemiological traits are transferable, not locked into stable subdivisions in space and time (for case in point, consider, e.g., SRA gene transfer from *T. b. rhodesiense* to *T. b. brucei*[156]). Recombination has driven major changes in *T. cruzi* transmission in the past, including adaptation to the domestic niche[69,70]. Our data suggest that recombination may continue to transform contemporary disease cycles, as suggested for *Toxoplasma gondii*[426] and in *Leishmania* spp.[154,295,381]. The proven presence of a sexual cycle in *T. cruzi* should now reinvigorate the hunt for the site of genetic exchange within the host or vector, as well as its cytological mechanism. An *in vitro* model for meiotic genetic exchange in *T. cruzi* will dramatically improve our ability to distinguish the genetic bases of virulence, drug resistance and other epidemiologically relevant phenotypes. Determination of such traits may underpin future efforts to treat and control Chagas disease.

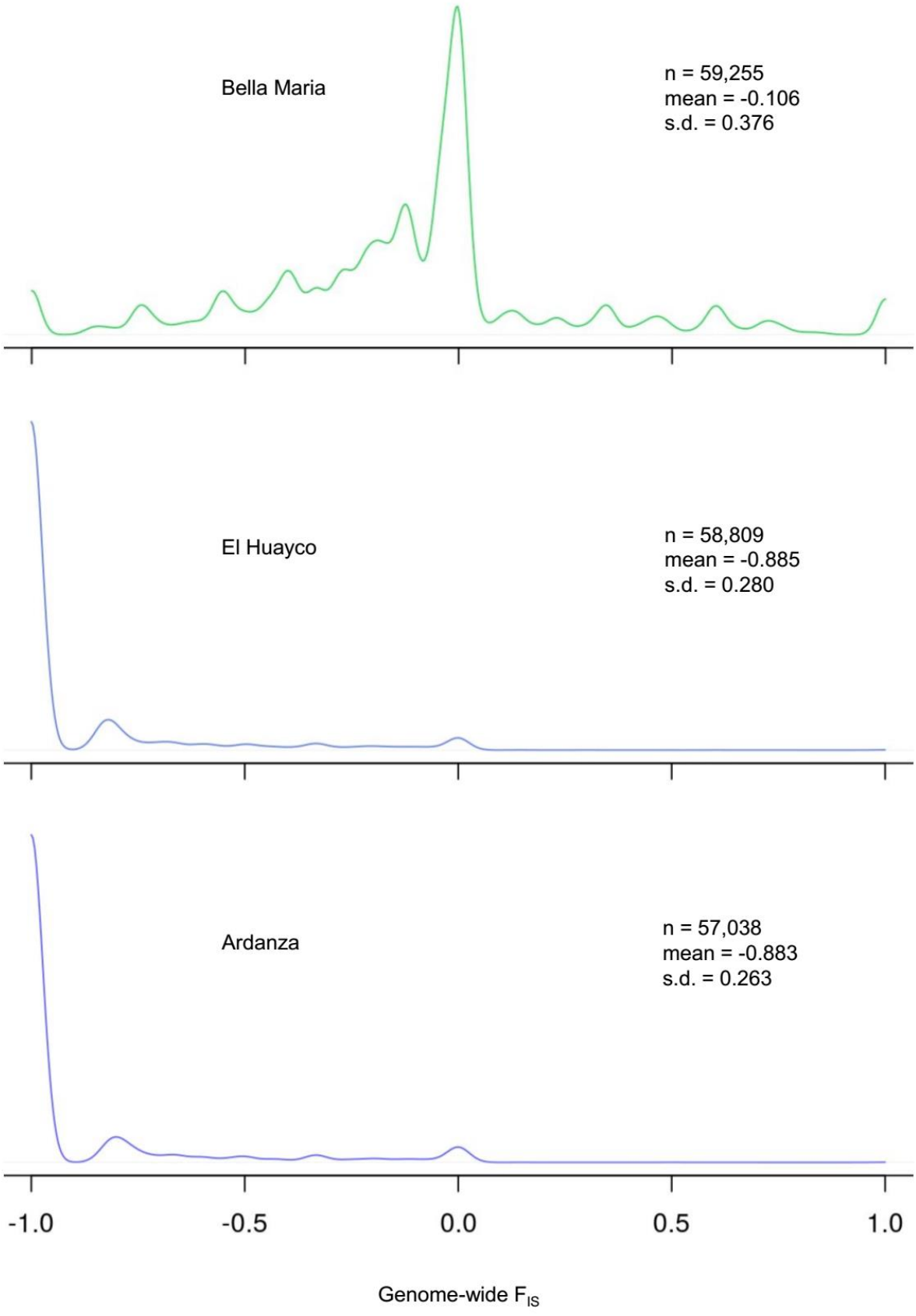
2.6 Supplementary figures and tables



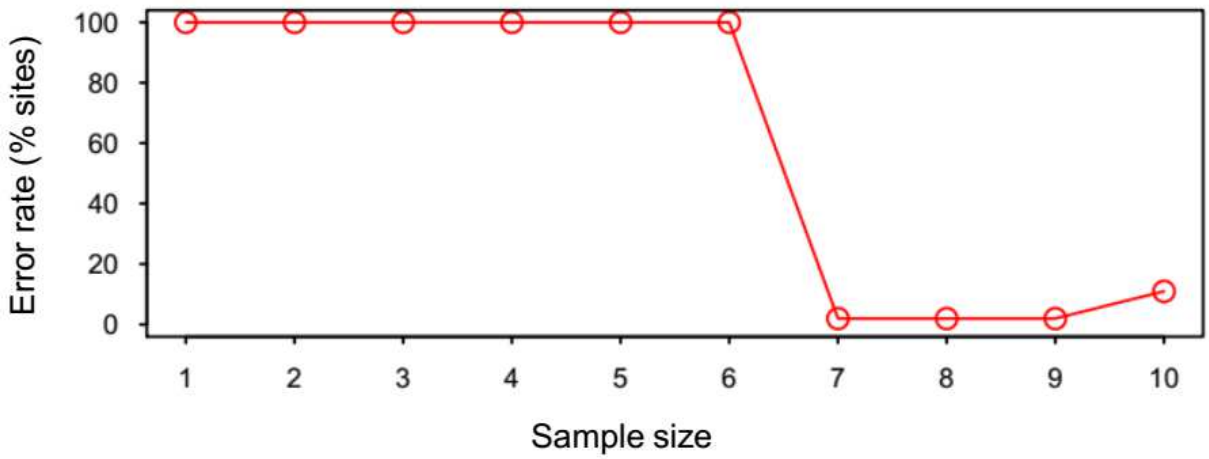
Supplementary Figure 2.1 Maximum-likelihood phylogenetic relationships among *T. cruzi* I clones**.** Pairwise genetic distances are haplotype-based, defined as the proportion of non-shared alleles across all SNP sites for which genotypes are called for all individuals (n = 7,392). The tree follows a general time-reversible (GTR) substitution model with ascertainment bias correction.



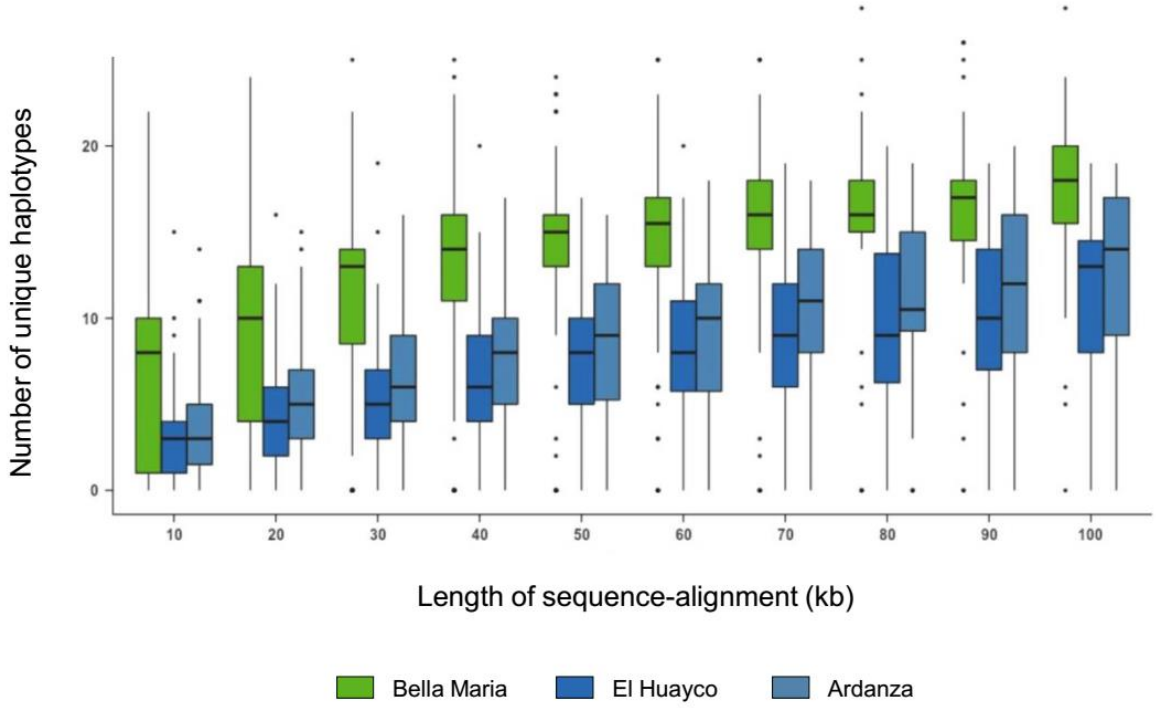
Supplementary Figure 2.2 Nonparametric population clustering of *T. cruzi* I clones. **a** Bayesian Information Criterion (BIC) scores of k-means clustering solutions for population assignment of *T. cruzi* I clones, based on 68,449 biallelic sites. Ward's criterion[427] is used for objective selection of k. **b** Discriminant analysis of principle components (DAPC) membership probabilities for k = 2, k = 5 and k = 6. Latter k-means solutions allow for additional partitioning of genetic diversity but do not necessarily imply true population subdivision. Colors represent different population assignments.



Supplementary Figure 2.3 Rates of homozygosity relative to Hardy-Weinberg expectations in *T. cruzi* I groups. Genome-wide density distributions of Wright's inbreeding coefficient F_{IS} are plotted for *T. cruzi* I clones from Bella Maria, El Huayco and Ardanza. F_{IS} sample size, mean and standard deviation are also given for each group, based on the dataset's total 130,996 SNPs.

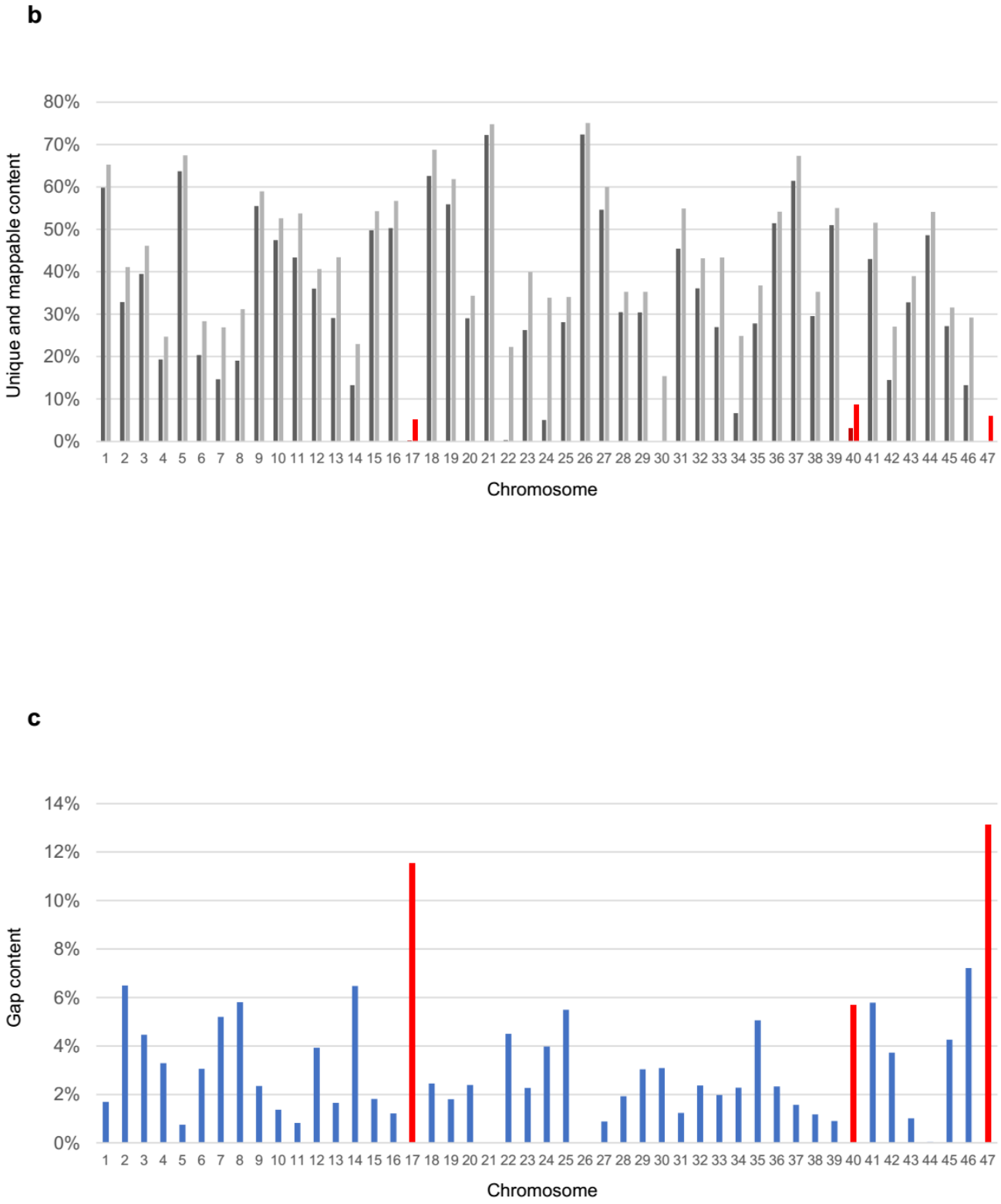


Supplementary Figure 2.4 Power to reject Hardy-Weinberg equilibrium in asexual genomes. We measured the proportion of SNP sites for which the '–hwe' function in VCFtools[396] incorrectly accepts a null hypothesis of Hardy-Weinberg equilibrium (i.e., $p > 0.05$) in sets of 1 – 10 non-recombinant *T. cruzi* genomes (22,475 SNPs simulated with BamSurgeon[399]; see Methods). Type II error predominates when the simulated data is reduced to < 7 individuals, as occurs when observations from Loja are restricted to one clone per vector/host (see Ardanza in Supplementary Tbl. 2.2).

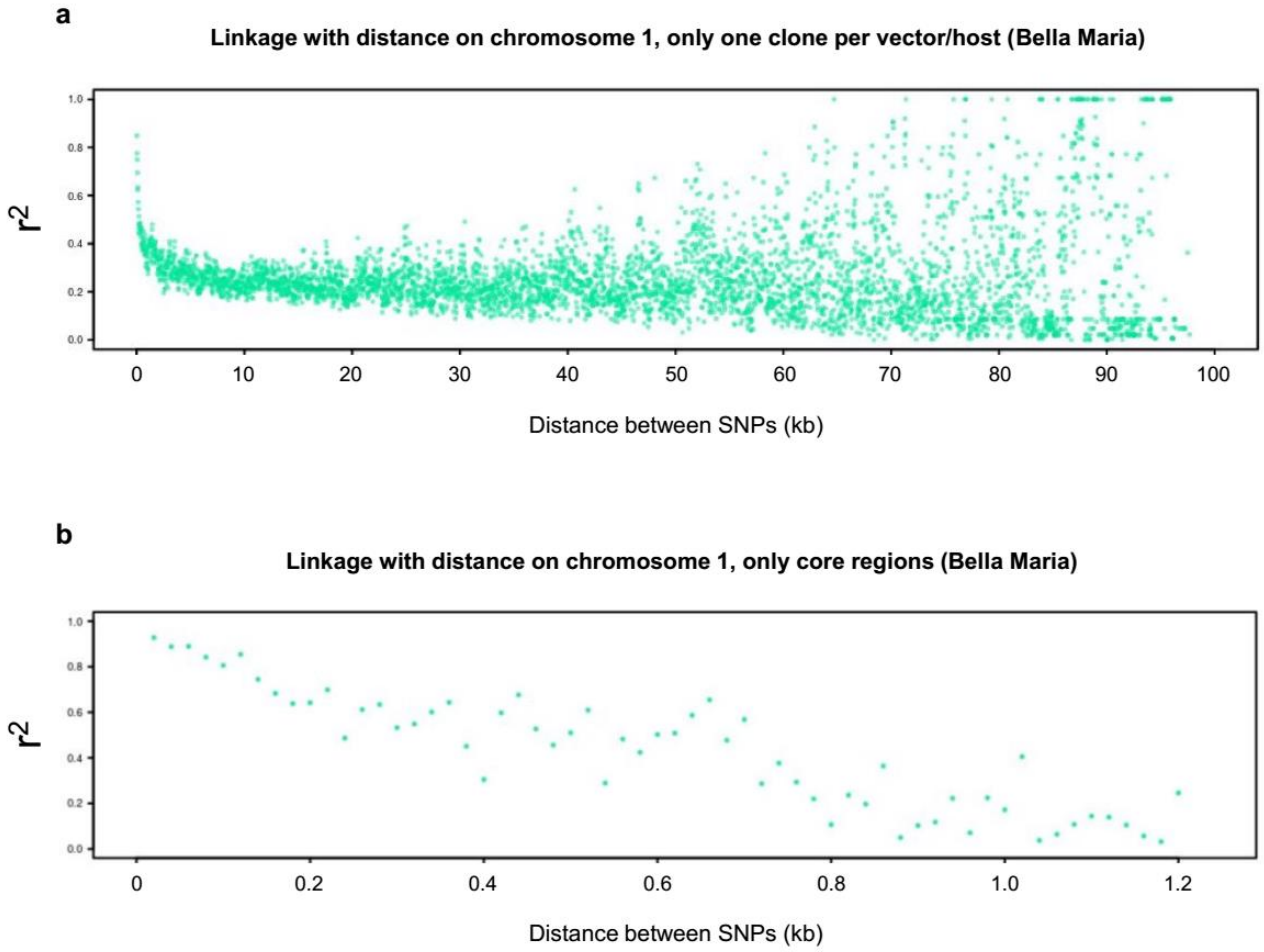


Supplementary Figure 2.5 Rates of haplotype differentiation relative to sequence length in *T. cruzi* I groups. Boxplots show median and interquartile range for the number of distinct haplotypes found in phased sequence alignment (n = 70,306 SNPs) at window sizes between 0 and 100 kb for Bella Maria, El Huayco and Ardanza groups.

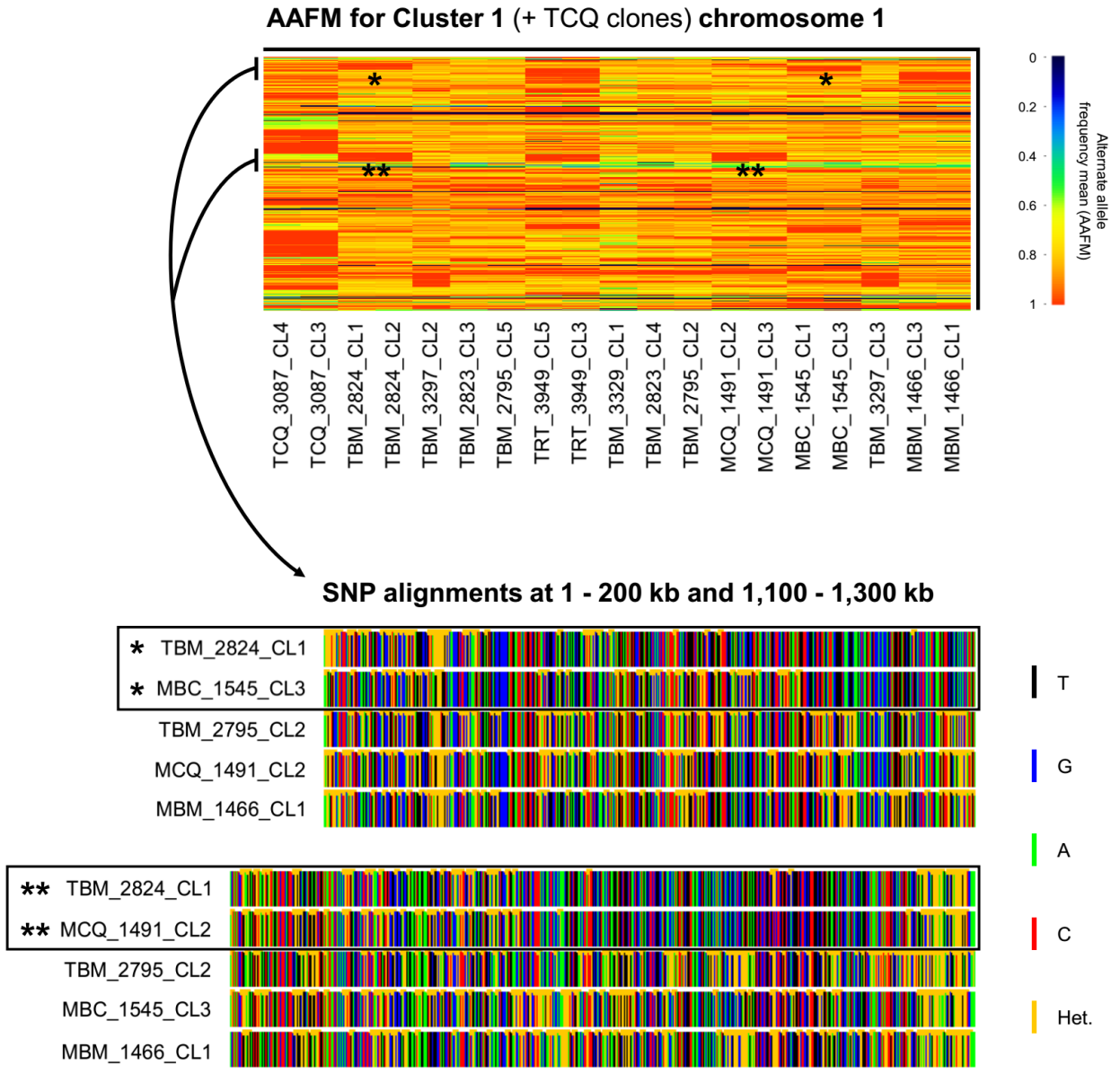
Supplementary Figure 2.6 (continues with legend on next page)



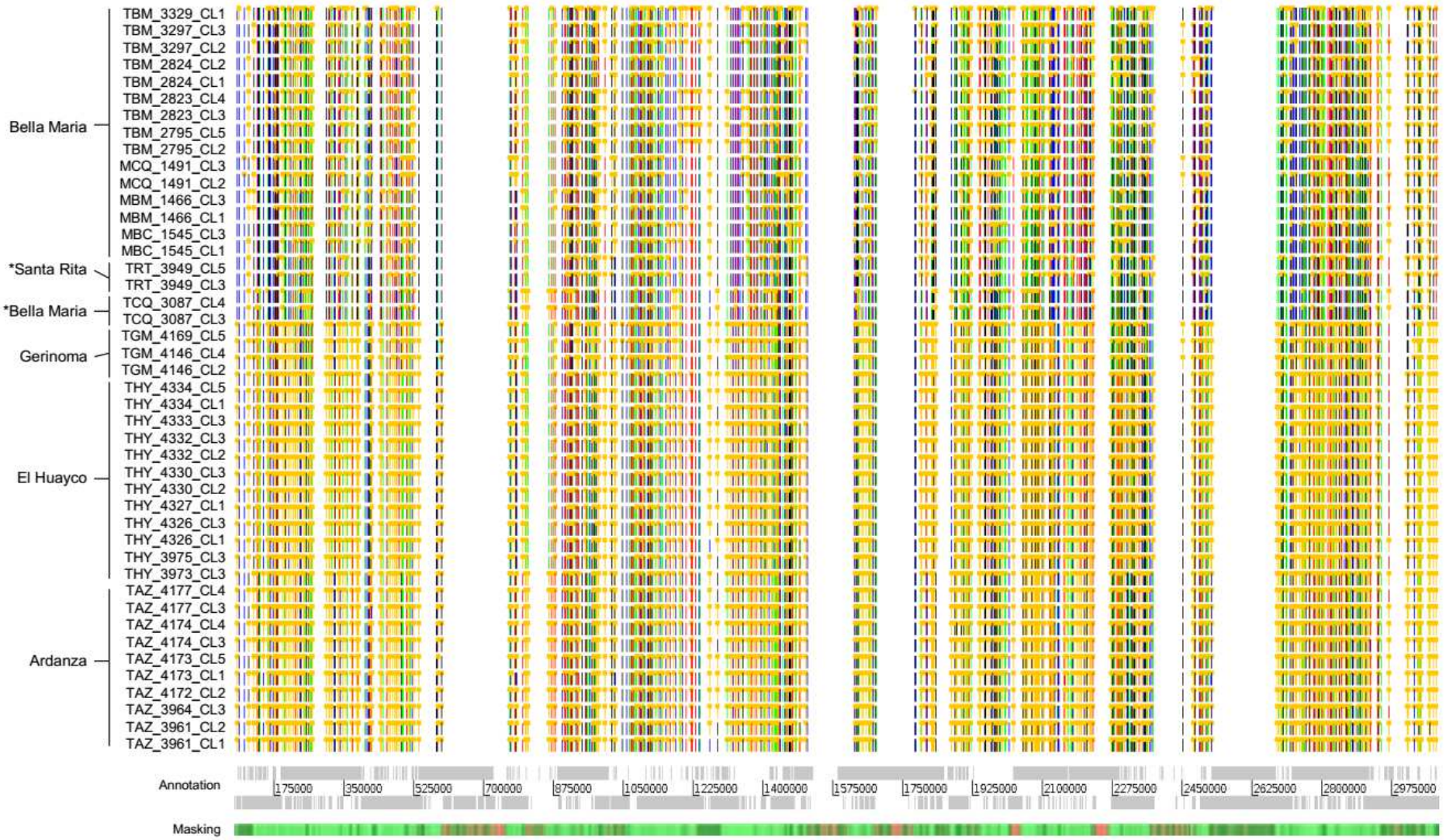
Supplementary Figure 2.6 TcI-Sylvio reference evaluation and masking. **a** Masks applied to the TcI-Sylvio reference genome based primarily on virtual mappability[390]. Final masking (red) disqualified a total of 24 Mb (including entire chromosomes 17, 40 and 47) of 42 Mb from polymorphism analysis. Annotated genes are marked in black. **b-c** Proportions of mappable, unique (determined by self-blasting) and gap content on TcI-Sylvio reference chromosomes are indicated in light grey, dark grey and blue, respectively. Red bars distinguish chromosomes excluded from analysis based on these metrics.



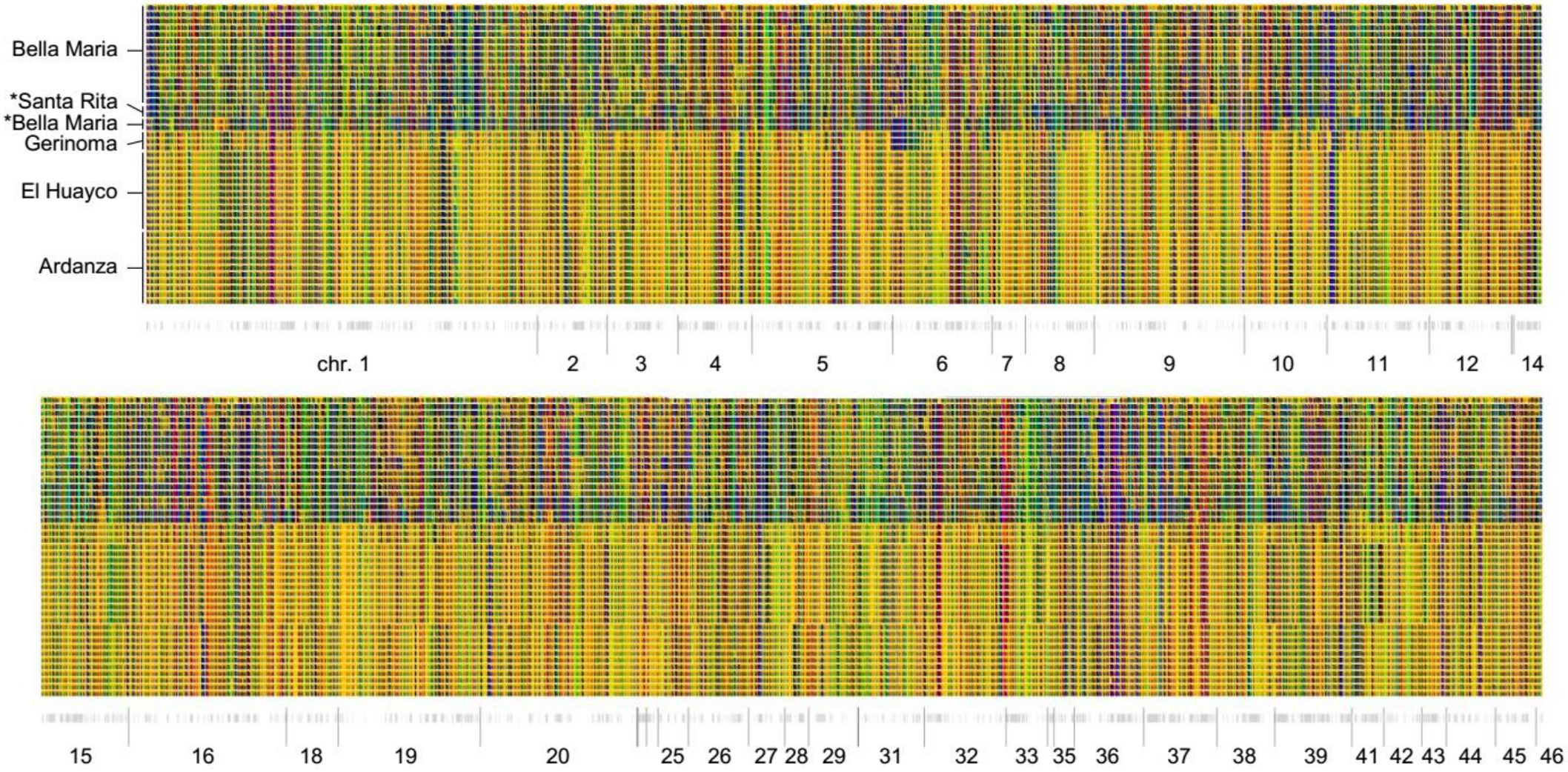
Supplementary Figure 2.7 Linkage decay in *T. cruzi* I clones from Bella Maria, after subsampling. Linkage decay on chromosome 1 remains when analysis is restricted to **a** one random clone per host/vector (n = 4,670 SNP sites) or to **b** core sequence regions, defined as areas of synteny among TcI-Sylvio, *T. b. brucei* and *L. major* reference genomes. The latter reduction in sample size to 1,178 sites limits analysis to short map distance classes (0 – 1.2 kb). Presentation is otherwise analogous to that in Fig. 2.3a.



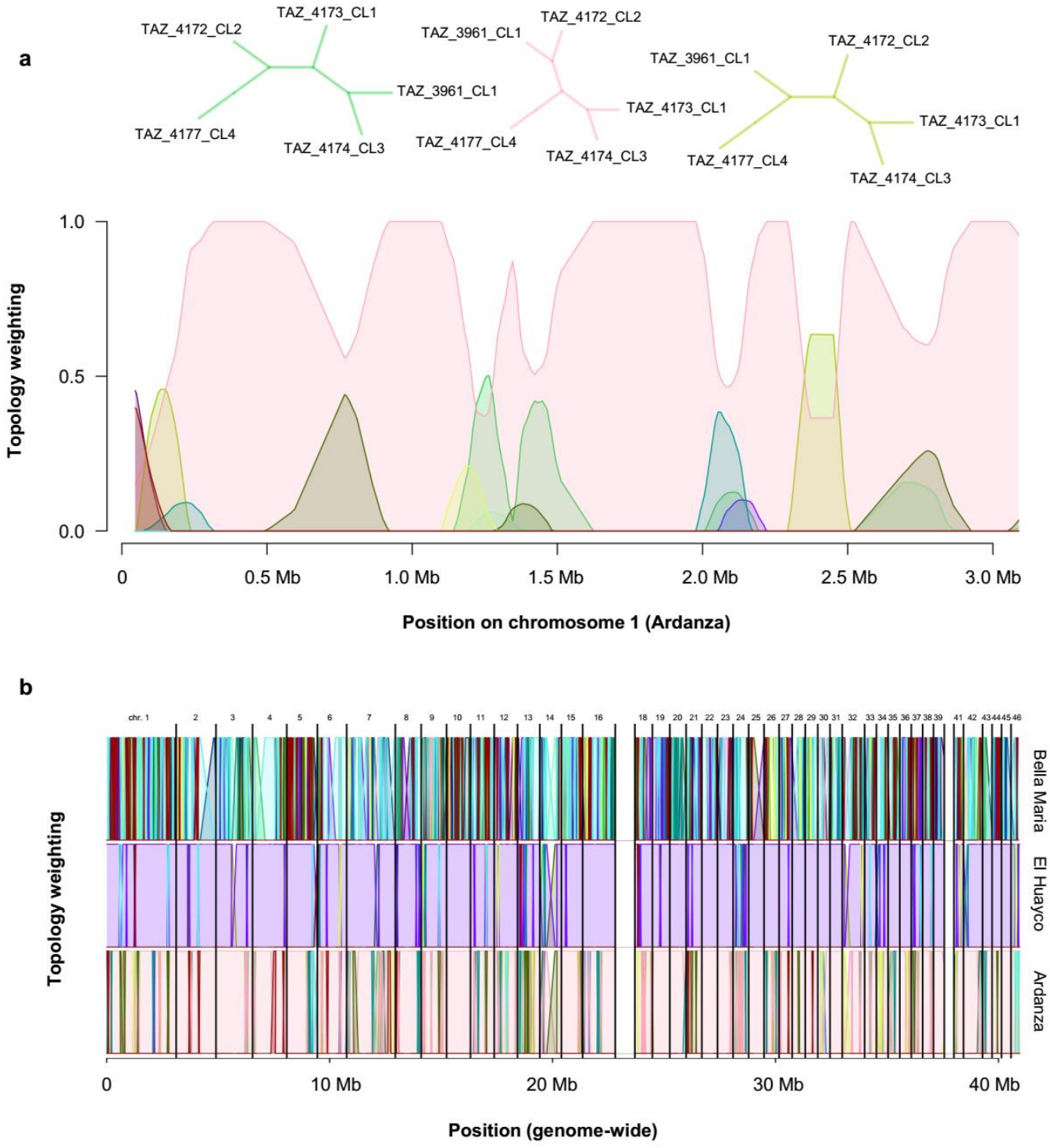
Supplementary Figure 2.8 Patchy homozygosity and SNP-sharing suggests recombination among *T. cruzi* I clones. In the top plot, each column represents the first chromosome of one clone. Rows within each column represent consecutive 5 kb sequence bins. Alternate allele frequency means (AAFM) determine the color of each bin – blue (0) through green (0.5) to red (1). Long tracts of high AAFM (i.e., large red patches) expose abrupt segmental increases in sequence similarity between different pairs of clones, as exemplified in the SNP concatenations below. Homozygous SNPs are colored according to base identity – black (T), blue (G), green (A) and red (C). Heterozygous SNPs are colored yellow. Single-asterisked AAFM patches reflect high sequence similarity between TBM_2824_CL1 and MBC_1545_CL3 near the start of the chromosome 1. Double-asterisked patches at ca. 1,200 kb reflect a sudden shift in pairwise similarity. Here, SNP identities in TBM_2824_CL1 and MCQ_1491_CL2 begin to align.



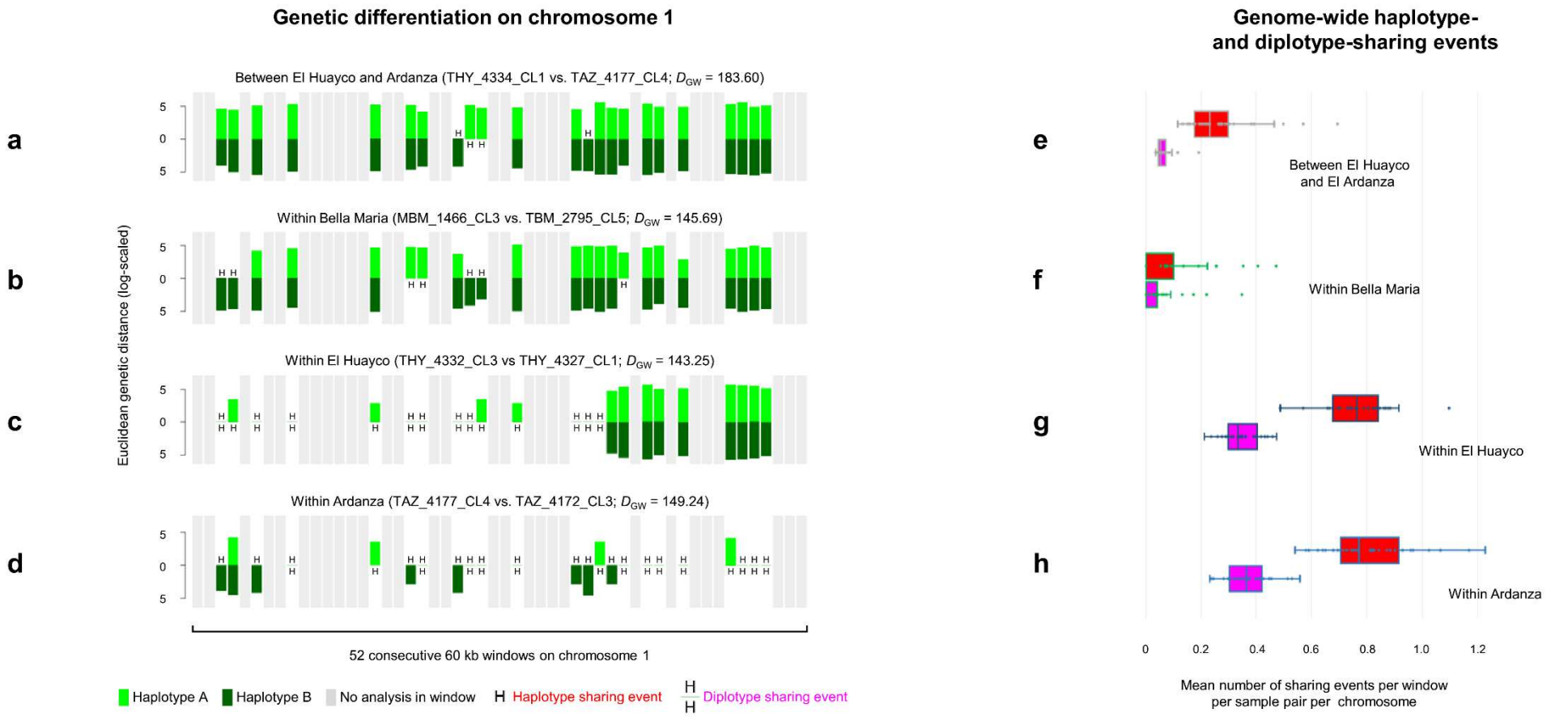
Supplementary Figure 2.9 SNP alignment across chromosome 1 for all *T. cruzi* I clones. Homozygous SNPs are colored according to base identity – black (T), blue (G), green (A) and red (C). Heterozygous SNPs are colored yellow. Colors overlap where SNP density is high. Only sites without any missing genotypes are shown. White spaces in grey bars below alignment represent coding regions on forward (top) and reverse (bottom) strands. The third bar indicates masked sequence regions in red, unmasked regions in green.



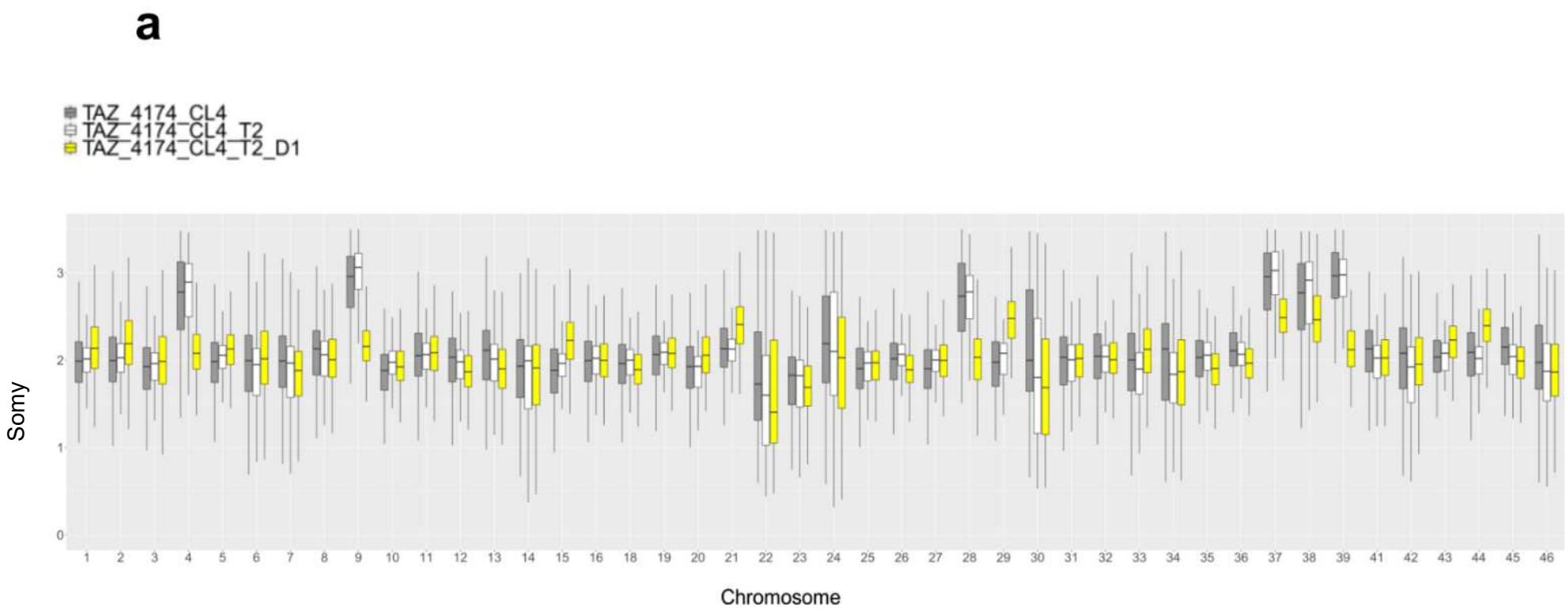
Supplementary Figure 2.10 Genome-wide SNP alignment for all *T. cruzi* I clones. Homozygous SNPs are colored according to base identity – black (T), blue (G), green (A) and red (C). Heterozygous SNPs are colored yellow. Only variable sites without any missing genotypes are shown. Grey bars below alignment represent SNPs in coding regions. Asterisks denote outlier samples from Santa Rita (TRT_3949 clones) and Bella Maria (TCQ_3087 clones). Occasional patches of shared homozygosity (e.g., see chromosome 36) in Cluster 2 were not associated to coding vs. non-coding sequence annotation (χ^2 = 0.089, df = 1, p-value = 0.764). Sample order matches that in Supplementary Fig. 2.9.



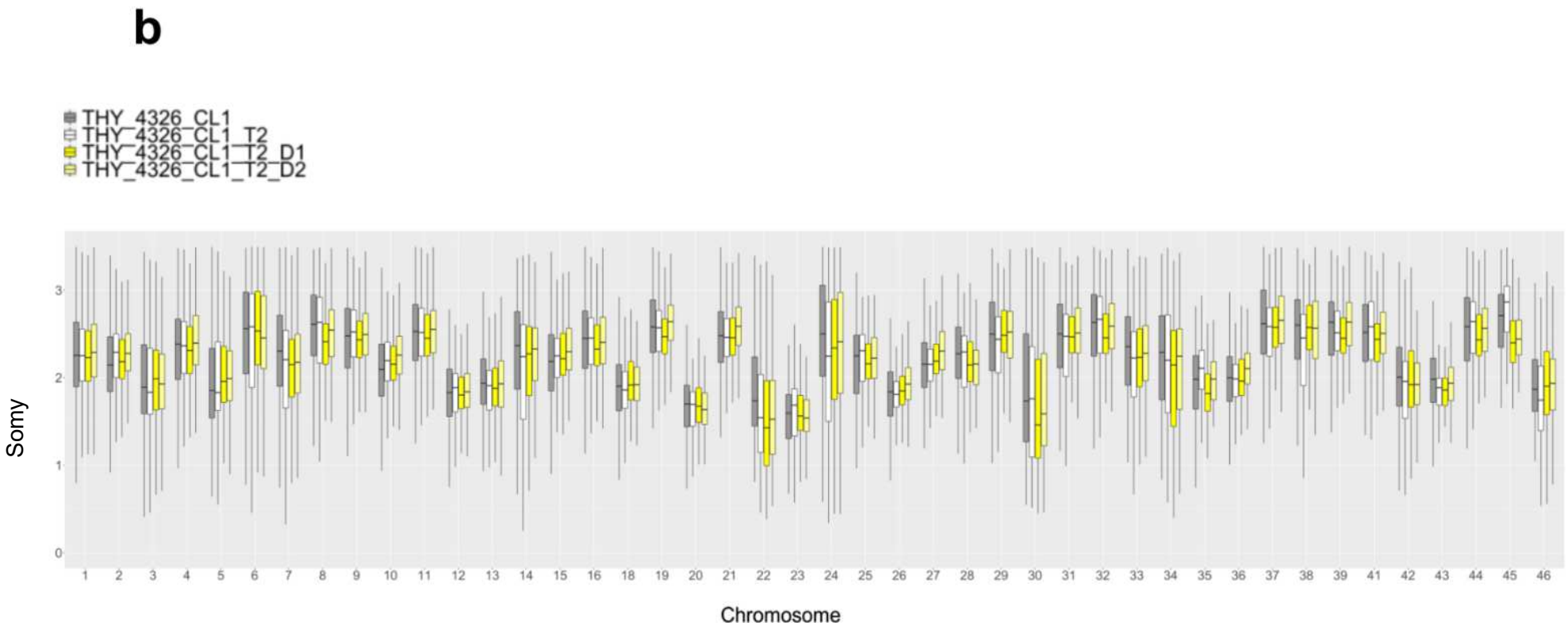
Supplementary Figure 2.11 Intra-chromosomal phylogenetic relationships among *T. cruzi* I clones. **a** Theoretically, fifteen neighbor-joining (NJ) topologies can be drawn to describe relationships among five samples within a larger phylogenetic tree. When NJ trees are constructed in 50 kb sliding-window analysis (step size = 10 kb), a single topology dominates across chromosome 1 for a five-sample subset from Ardanza. Similar is true for El Huayco (see Fig. 2.4d). Topology weightings (the relative abundances of the different five-sample topologies after iterative sampling of sub-trees[400]) are plotted (with loess smoothing; span = 0.125) for each window across the chromosome. **b** Mosaic (Bella Maria) vs. stable (Cluster 2) genealogies occur as such genome-wide. Colors represent different tree topologies. Poorly mapping chromosomes 17, 40 and 47 are excluded from analysis.



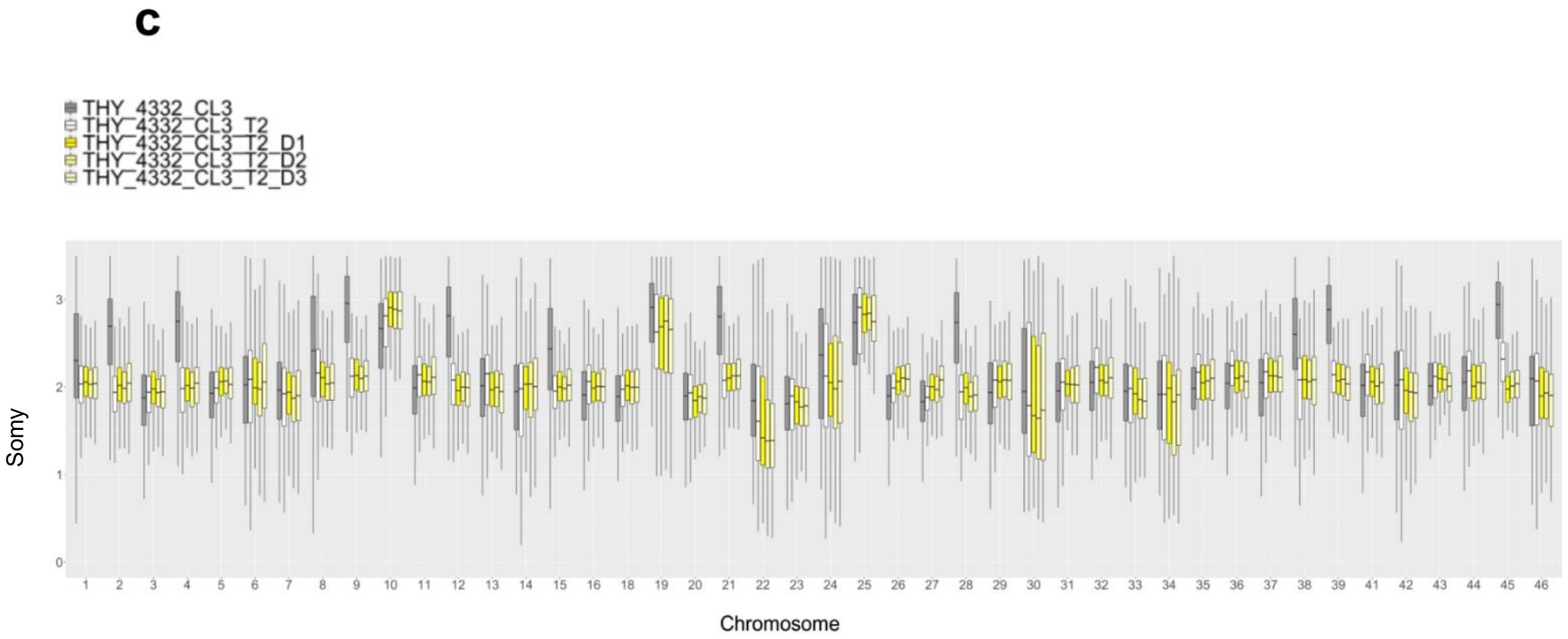
Supplementary Figure 2.12 Pairwise haplotype and diplotype sharing within and between *T. cruzi* I groups. In plots **a-d**, light green bars indicate genetic distances for pairs of samples in consecutive 60 kb sequence windows along phased haplotype A on chromosome 1 (996 SNP sites). Opposite bars (dark green) quantify distances for haplotype B. Windows of low and/or masked polymorphism (< 20 SNP sites per 60 kb) are shown in grey. These windows are excluded from analysis. **a** compares sample THY 4334 CL1 (El Huayco) to sample TAZ 4177 CL4 (Ardanza) and exemplifies between-group haplotype sharing (marked by the letter H) observed in Cluster 2. Several windows present matching 60 kb haplotypes, i.e., zero differentiation on haplotype A or B. Light or dark green bars therefore do not appear in these windows. **b** shows similar results from a pairwise comparison representative of haplotype differentiation within the Bella Maria group. Pairwise haplotype differentiation within El Huayco (**c**) and within Ardanza (**d**) is different. Shared haplotypes are much more abundant and many windows also present diplotype sharing (marked by two H's), i.e., identical SNP calls on both homologous haplotype segments in both *T. cruzi* clones. Plots **e-h** demonstrate how haplotype (red) and diplotype sharing (pink) events depicted in windowed bar plots for chromosome 1 also occur frequently on other chromosomes and in all possible pairwise comparisons within El Huayco (n = 66 pairwise comparisons) and Ardanza (n = 55). They occur less frequently within the Bella Maria group (*n* = 105). Each point represents the mean number of sharing events per window per sample pair for one of 37 chromosomes (7,299 sites). Chromosomes 13, 17, 22, 23, 24, 30, 34, 40, 46 and 47 are excluded from analysis due to low polymorphism and/or heavy masking. Vertical bars in boxes and at box edges mark medians and interquartile ranges. D_{GW} is the genome-wide Euclidean genetic distance.



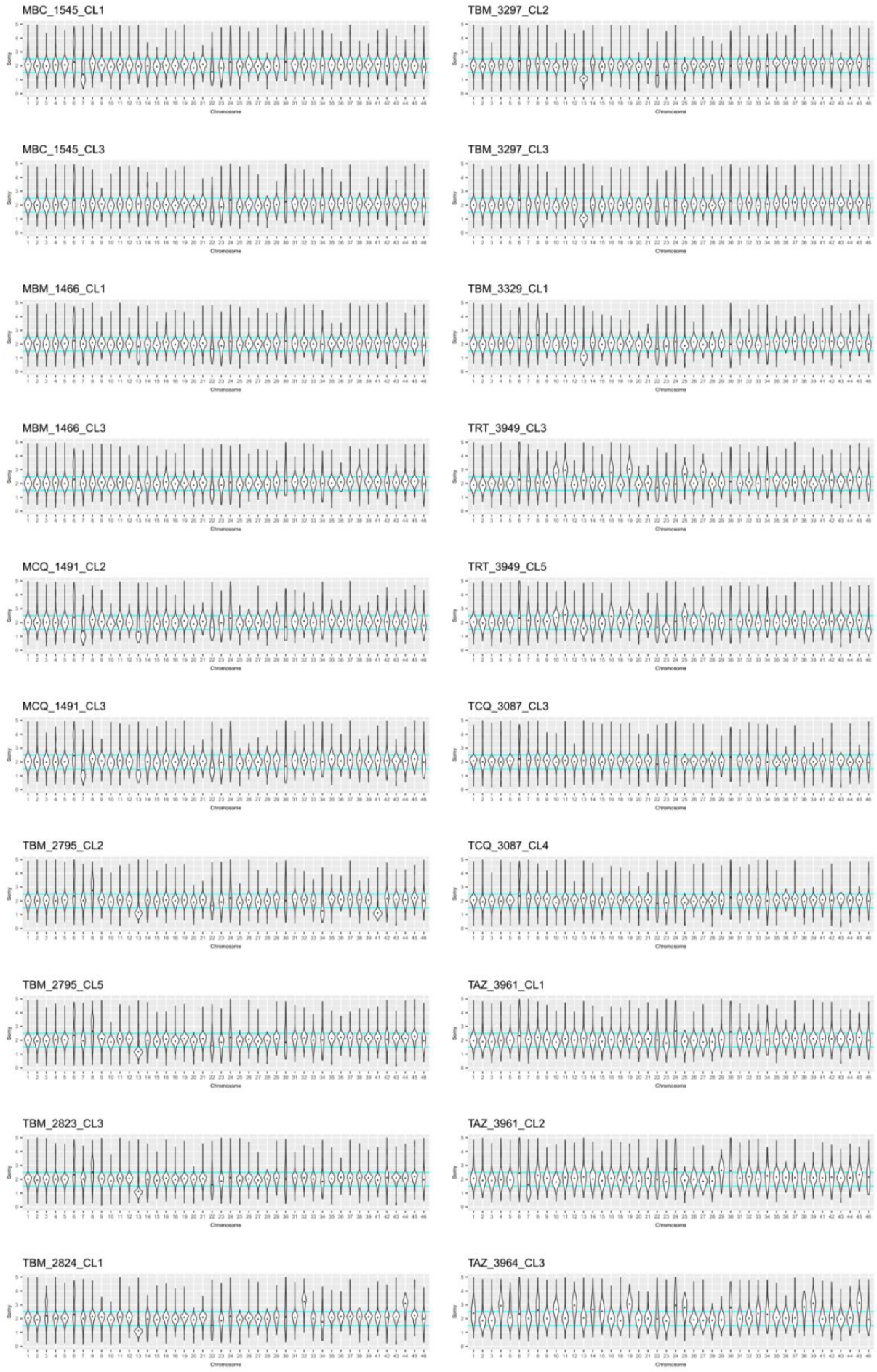
Supplementary Figure 2.13 (continues on next page)



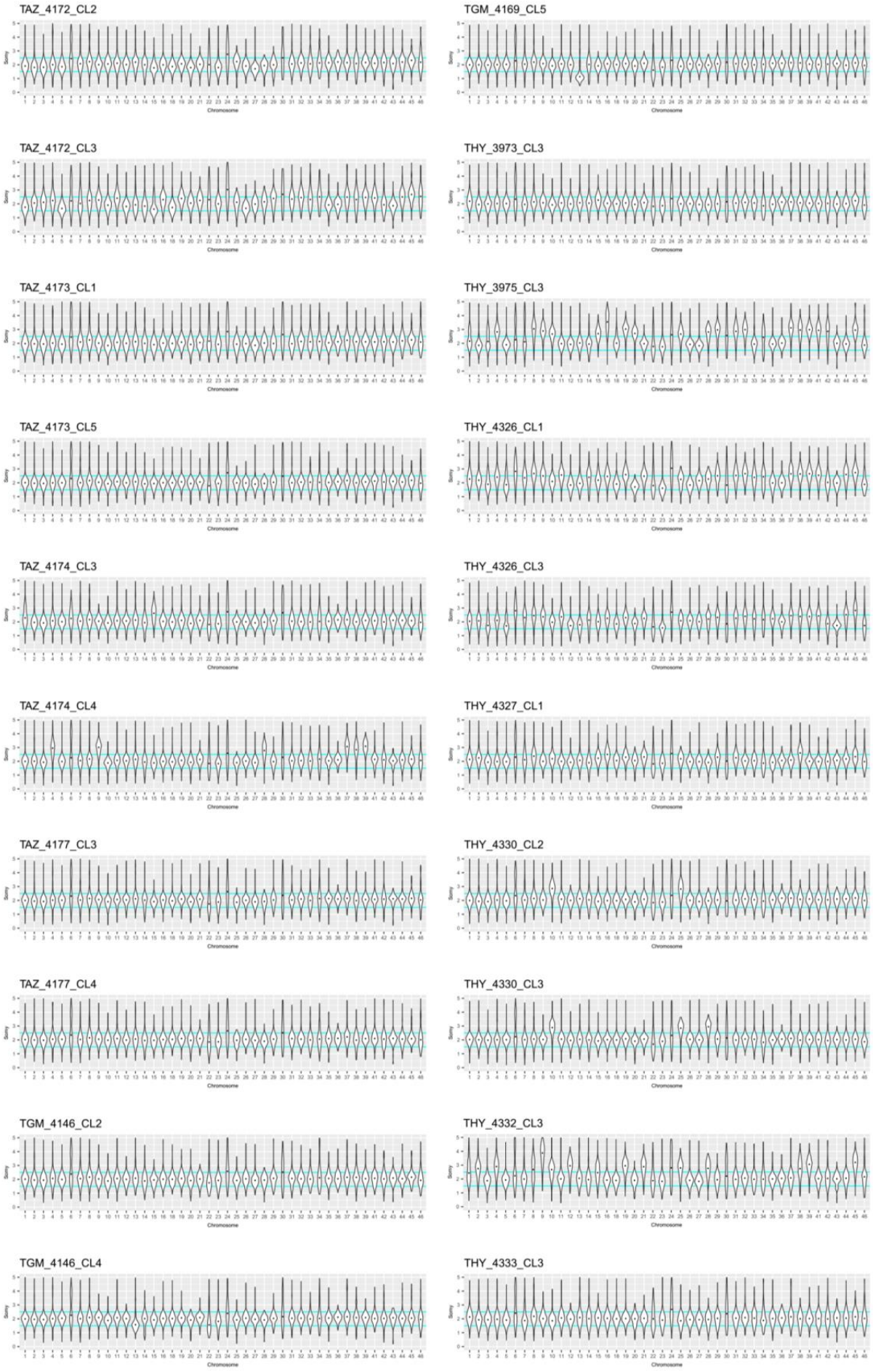
Supplementary Figure 2.13 (continues on next page)



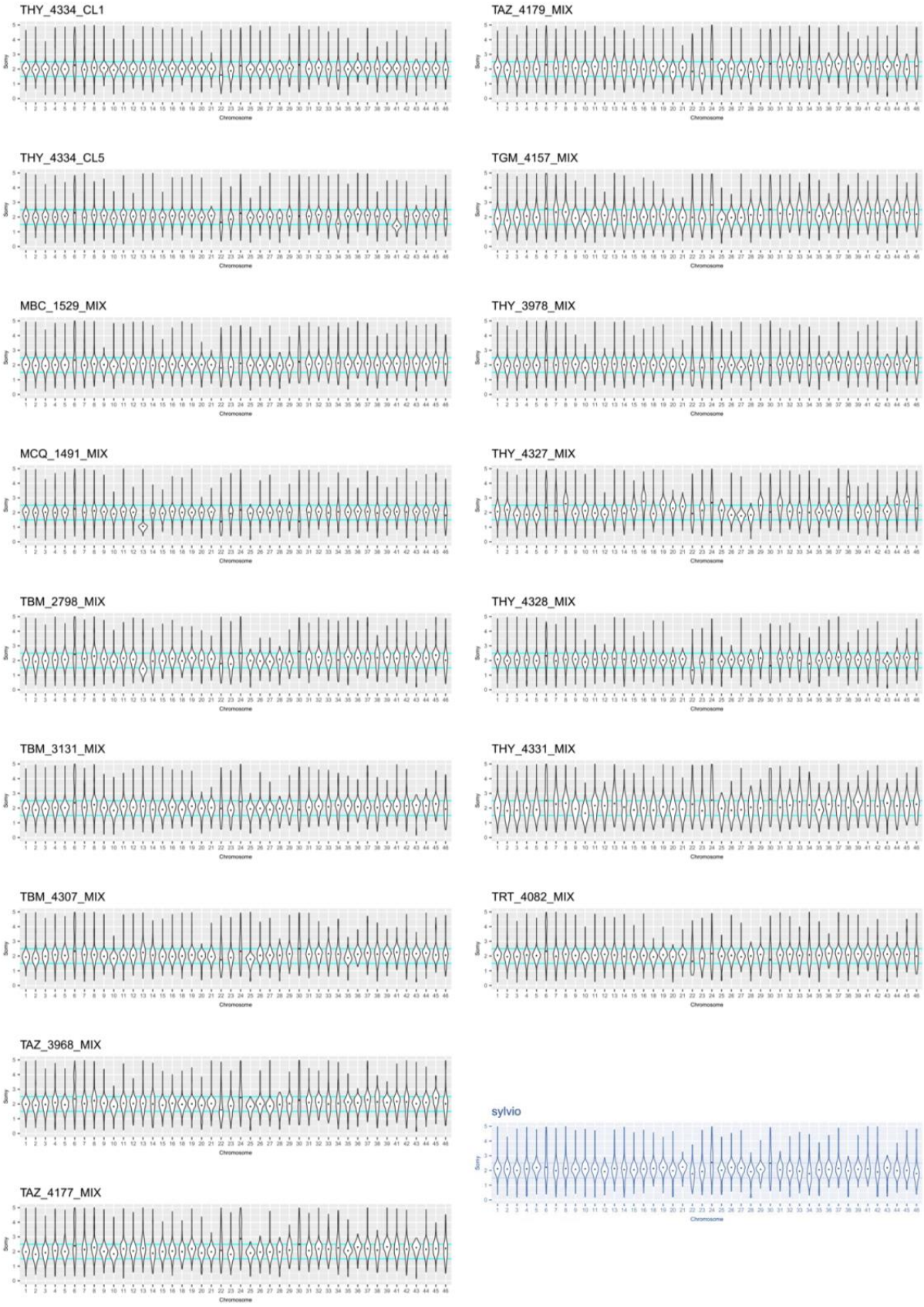
Supplementary Figure 2.13 Temporal and sub-clonal somy variation for selected *T. cruzi* I clones. a Following first sequencing and sample cryopreservation, TAZ_4174_CL4 was thawed, re-expanded in Liver Infusion Tryptose (LIT) medium (no additional passages) and sequenced for a second time. One subclone obtained from the re-cultured sample was also sequenced. Boxplots show median and interquartile range of site-wise somy estimates ($2 \cdot m$ / p30 of M_m) for each chromosome (see Methods). While the 'parent' clone karyotype appeared unchanged at time of second sequencing (T2), results for subclone T2_D1 suggest sub-clonal chromosomal copy number variation (e.g., see white vs. yellow boxplots for chromosomes 4 and 39). b THY_4326_CL1 was also re-sequenced but showed no evidence of somy differences between subclones (n = 2) or over time. The sample was passaged three times post-cryopreservation. c THY_4332_CL3 appeared to have reduced somy levels between first and second sequencing (four passages), but no sub-clonal variation was observed (n = 3).



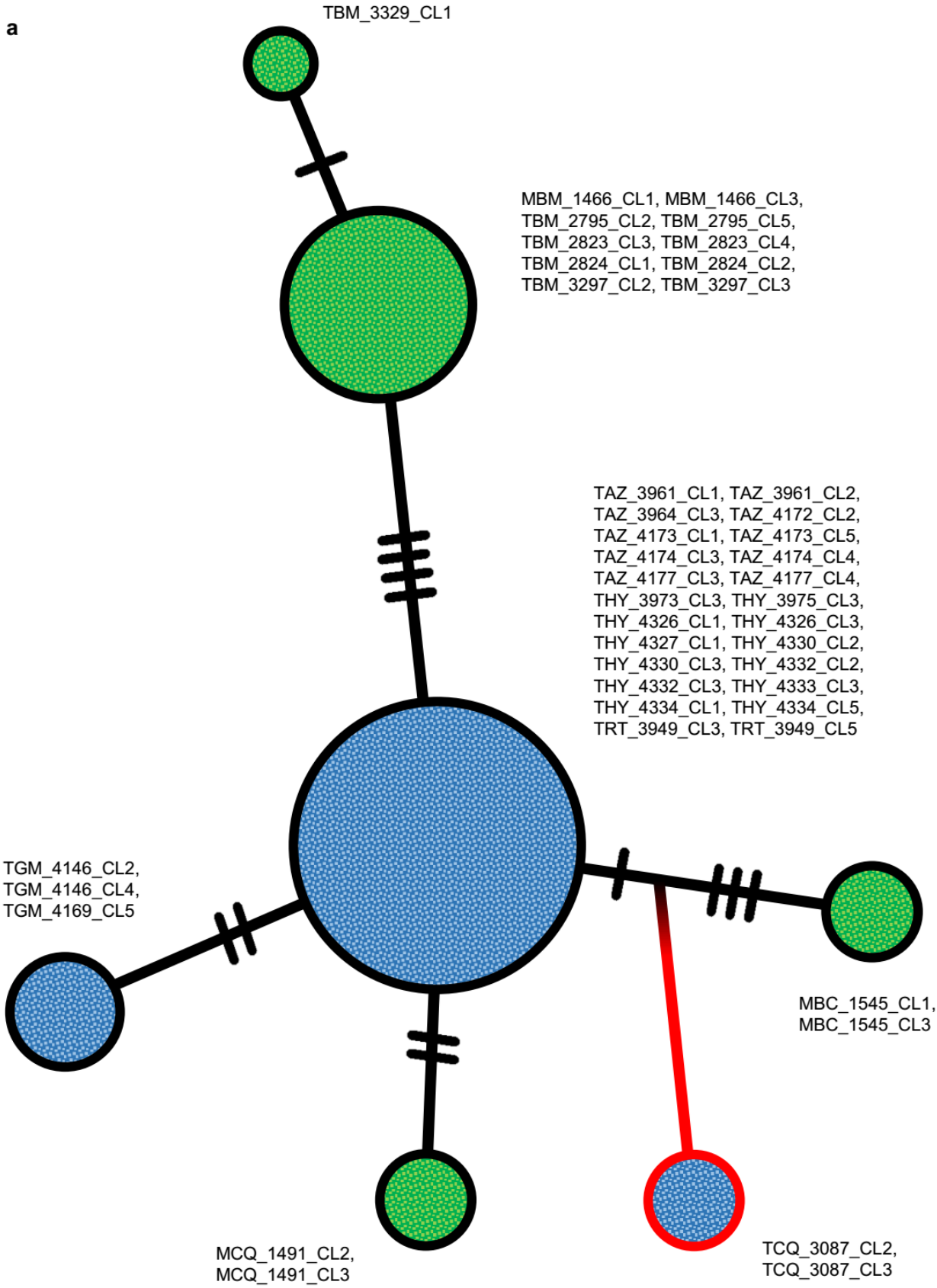
Supplementary Figure 2.14 (continues on next page)



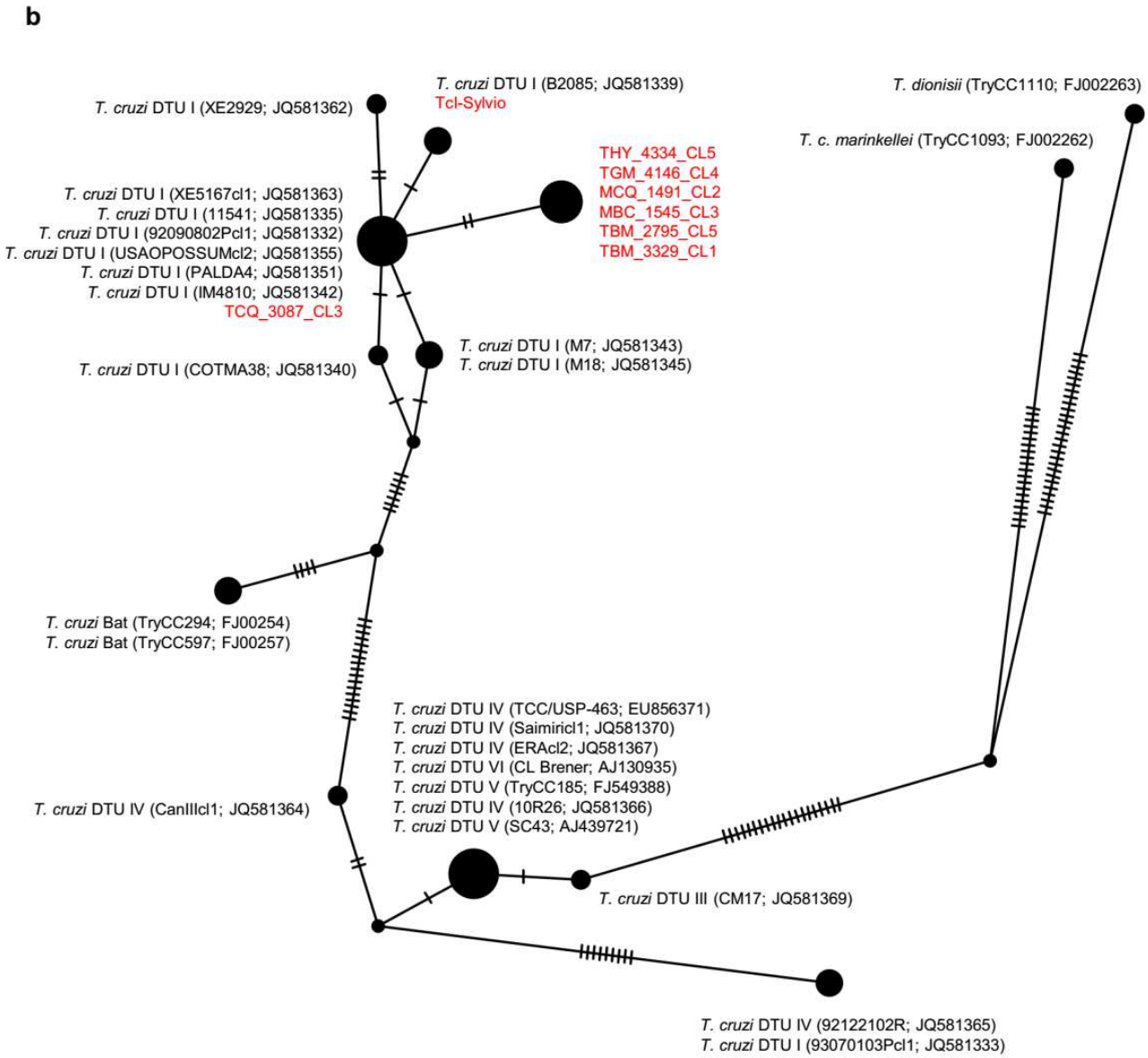
Supplementary Figure 2.14 (continues on next page)



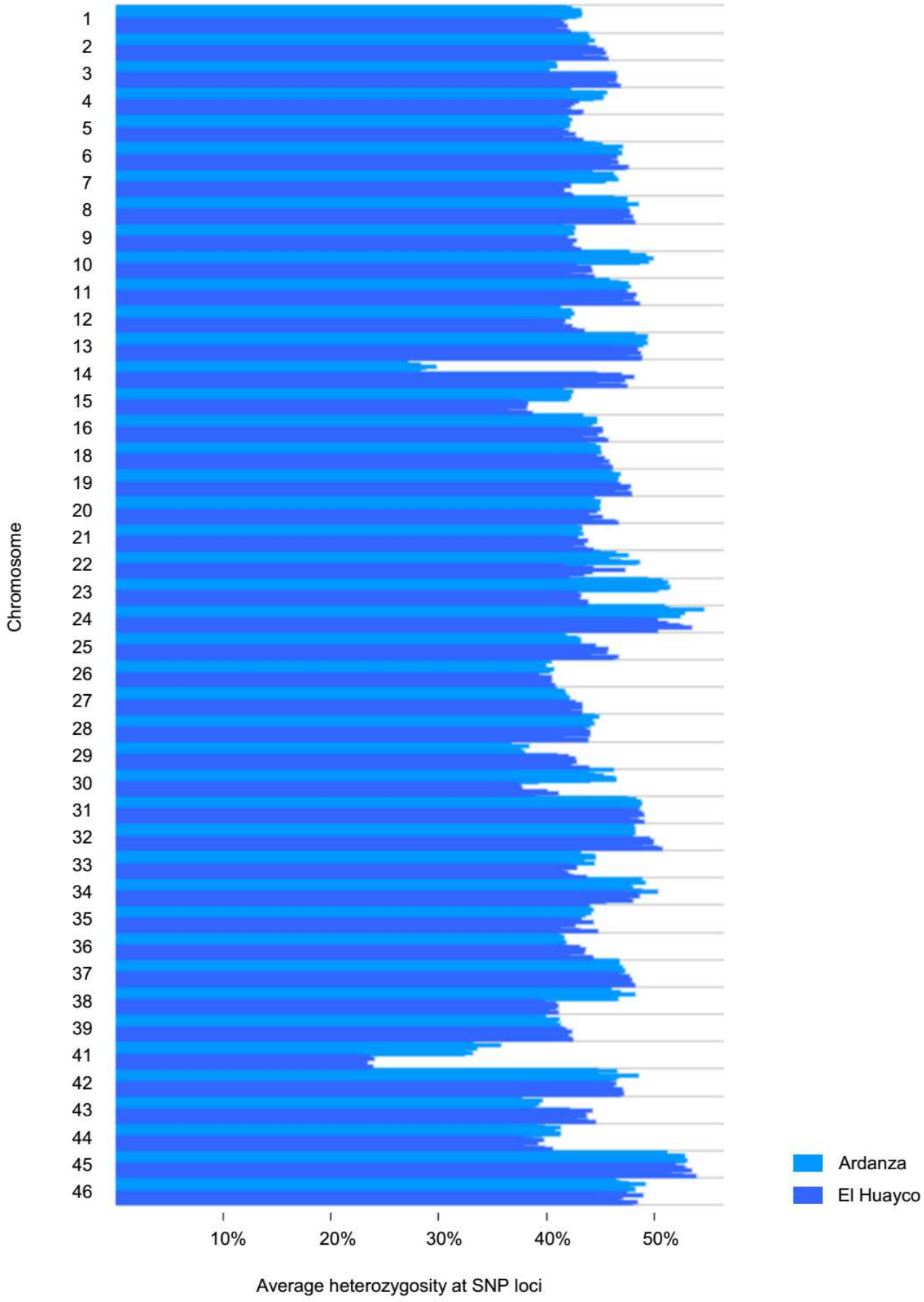
Supplementary Figure 2.14 Somy estimates for cloned and non-cloned *T. cruzi* samples. Violin plots show density distributions of window-based somy estimates for each chromosome (see Methods). Dots indicate medians. TcI-Sylvio was used to validate calculations. Results for non-cloned, low-diversity infections (e.g., MCQ_1491_MIX or THY_4327_MIX) suggest that aneuploidies in clones are not consequences of stress from plate-cloning in the lab. Plots for TBM_2823_CL4, TBM_2824_CL2 and THY_4332_CL2 are excluded because they are already provided in Fig. 2.6a.



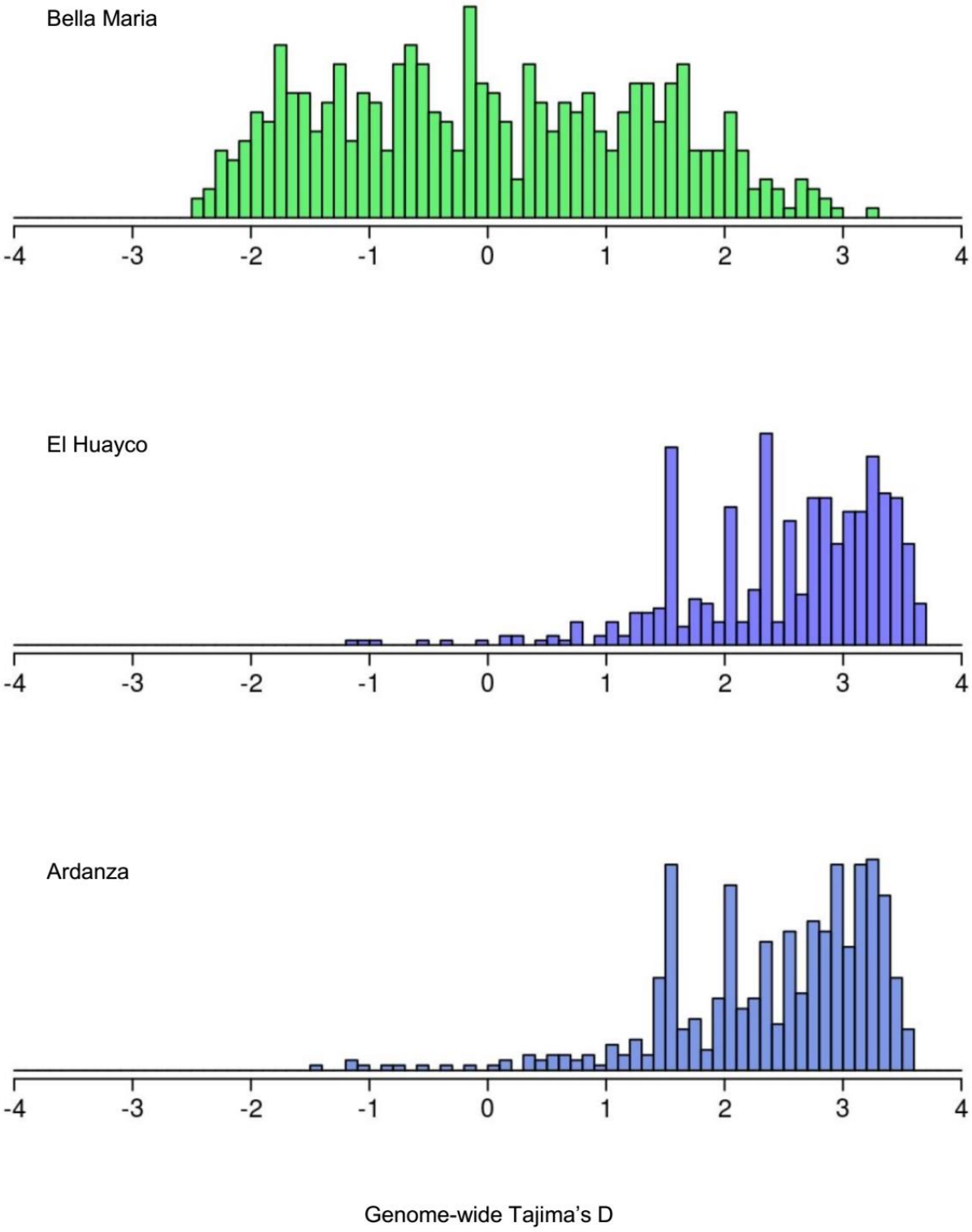
Supplementary Figure 2.15 (continues on next page)



Supplementary Figure 2.15 Mitochondrial phylogenies for *T. cruzi* I clones. **a** Maxicircle sequence variation among all samples except TAZ_4172_CL3 (missing information at 44% SNP sites) represented as a TCS network[428]. Black tick marks between nodes indicate the number of mutations between genotypes. Node sizes correspond to the number of samples represented by the particular maxicircle variant. Green nodes contain members of Cluster 1, as defined in nuclear phylogenetic analysis (Fig. 2.1). Blue nodes contain members of Cluster 2, but also TRT_3949 clones. TCQ_3087 clones appear divergent, with 668 diagnostic SNP differences relative to other clones of Cluster 2. **b** TCS network from cytochrome b alignment (617 bp), for which sequences from all 7 *T. cruzi* sub-lineages (including TcBat) and other congeners are available for comparison. These sequences are detailed in Marcili et al. (2009)[54] and Messenger et al. (2012)[142]. Tick marks indicate number of mutations between genotypes. Samples from this study (one representative per maxicircle variant) are shown in red.



Supplementary Figure 2.16 Heterozygosity per chromosome in *T. cruzi* I clones from El Huayco and Ardanza. Average heterozygosity values fall between 40 and 50% for most chromosomes. Only chromosomes 14 (in Ardanza clones) and 41 (in both Ardanza and El Huayco clones) show substantial increases in homozygosity.



Supplementary Figure 2.17 SNP variation relative to neutral expectations in T. *cruzi* I groups. Histograms plot variation in Tajima's D values over 50 kb sequence bins in genomes from Bella Maria (96,691 SNPs), El Huayco (80,052 SNPs) and Ardanza (78,325 SNPs). Empty bins (i.e., windows lacking polymorphism within the group) do not enter analysis.

Supplementary Table 2.1 Host/vector sampling sites and T. cruzi I genomic sequencing coverage. Abbreviations: NRD (average nuclear read-depth); MRD (average maxicircle read-depth).

ID	Longitude (°)	Latitude (°)	Altitude (m)	Region	Ecotype	Host	Year	NRD	MRD
MBC_1545_CL1	-79.6039	-4.2134	1143	Bella Maria	sylvatic	*Artibeus fraterculus*	2012	23.3	75.4
MBC_1545_CL3	-79.6039	-4.2134	1143	Bella Maria	sylvatic	*Artibeus fraterculus*	2012	24.8	114.0
MBM_1466_CL1	-79.6166	-4.2185	1376	Bella Maria	sylvatic	*Rhipidomys leucodactylus*	2012	22.4	92.2
MBM_1466_CL3	-79.6166	-4.2185	1376	Bella Maria	sylvatic	*Rhipidomys leucodactylus*	2012	20.5	110.0
MCQ_1491_CL2	-79.5995	-4.2241	1272	Bella Maria	sylvatic	*Sciurus stramineus*	2012	16.4	167.4
MCQ_1491_CL3	-79.5995	-4.2241	1272	Bella Maria	sylvatic	*Sciurus stramineus*	2012	18.8	112.6
TAZ_3961_CL1	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	15.1	34.3
TAZ_3961_CL2	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	15.1	47.1
TAZ_3964_CL3	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	13.6	50.5
TAZ_4172_CL2	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	29.8	26.8
TAZ_4172_CL3	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	24.7	6.0
TAZ_4173_CL1	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	14.0	23.5
TAZ_4173_CL5	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	23.0	57.7
TAZ_4174_CL3	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	16.0	59.5
TAZ_4174_CL4	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	22.8	116.3
TAZ_4177_CL3	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	21.4	47.0
TAZ_4177_CL4	-79.5969	-4.2878	1311	Ardanza	domestic	*Panstrongylus rufotuberculatus*	2015	16.1	65.9
TBM_2795_CL2	-79.6063	-4.2115	1132	Bella Maria	domestic	*Panstrongylus chinai*	2011	25.9	149.2
TBM_2795_CL5	-79.6063	-4.2115	1132	Bella Maria	domestic	*Panstrongylus chinai*	2011	55.1	406.7
TBM_2823_CL3	-79.6063	-4.2115	1132	Bella Maria	domestic	*Panstrongylus chinai*	2011	53.5	185.3
TBM_2823_CL4	-79.6063	-4.2115	1132	Bella Maria	domestic	*Panstrongylus chinai*	2011	17.3	71.6
TBM_2824_CL1	-79.6063	-4.2115	1132	Bella Maria	domestic	*Panstrongylus chinai*	2011	50.9	296.7
TBM_2824_CL2	-79.6063	-4.2115	1132	Bella Maria	domestic	*Panstrongylus chinai*	2011	53.8	180.0
TBM_3297_CL2	-79.5985	-4.2285	1265	Bella Maria	sylvatic	*Rhodnius ecuadoriensis*	2013	64.0	474.7
TBM_3297_CL3	-79.5985	-4.2285	1265	Bella Maria	sylvatic	*Rhodnius ecuadoriensis*	2013	21.7	154.2
TBM_3329_CL1	-79.6170	-4.2085	1379	Bella Maria	sylvatic	*Rhodnius ecuadoriensis*	2013	18.5	61.3
TCQ_3087_CL3	-79.5972	-4.2258	1203	Bella Maria	sylvatic	*Rhodnius ecuadoriensis*	2012	54.6	580.8
TCQ_3087_CL4	-79.5972	-4.2258	1203	Bella Maria	sylvatic	*Rhodnius ecuadoriensis*	2012	45.7	301.4
TGM_4146_CL2	-79.4635	-4.0952	1801	Gerinoma	peri-domestic	*Triatoma carrioni*	2015	12.7	53.3

Supplementary Table 2.1 (continued)

TGM_4146_CL4	-79.4635	-4.0952	1801	Gerinoma	peri-domestic	*Triatoma carrioni*	2015	19.5	86.5
TGM_4169_CL5	-79.4635	-4.0952	1801	Gerinoma	peri-domestic	*Triatoma carrioni*	2015	19.5	67.4
THY_3973_CL3	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	13.6	55.1
THY_3975_CL3	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	25.2	129.9
THY_4326_CL1	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	16.9	34.3
THY_4326_CL3	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	36.9	82.3
THY_4327_CL1	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	16.7	82.3
THY_4330_CL2	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	16.3	70.7
THY_4330_CL3	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	25.8	125.5
THY_4332_CL2	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	20.8	28.0
THY_4332_CL3	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	22.8	12.4
THY_4333_CL3	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	15.1	68.1
THY_4334_CL1	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	27.2	118.7
THY_4334_CL5	-79.3188	-4.0906	1375	El Huayco	sylvatic	*Rhodnius ecuadoriensis*	2015	36.2	255.8
TRT_3949_CL3	-79.3475	-4.1125	1278	Santa Rita	domestic	*Panstrongylus chinai*	2015	20.2	88.1
TRT_3949_CL5	-79.3475	-4.1125	1278	Santa Rita	domestic	*Panstrongylus chinai*	2015	18.9	68.8

Supplementary Table 2.2 Examples of long tracts of homozygosity found in *T. cruzi* I genomes. This table summarizes long tracts of homozygosity found in MBC_1545_CL1 and MCQ_1491_CL2, two clones that typify homozygosity patterns in the Bella Maria group. Clones from El Huayco and Ardanza except THY_4330_CL3 and TAZ_3961_CL2 (one single occurrence each) entirely lack these tracts.

ID	Chromosome	Start position (bp)	End position (bp)	Number of variants	Number of mismatches	Region	Cluster
MBC_1545_CL1	1	2527616	2735662	377	1	Bella Maria	1
MBC_1545_CL1	3	688123	992700	193	0	Bella Maria	1
MBC_1545_CL1	3	688123	992700	204	0	Bella Maria	1
MBC_1545_CL1	4	1078068	1217217	98	3	Bella Maria	1
MBC_1545_CL1	5	244489	367798	226	0	Bella Maria	1
MBC_1545_CL1	5	642558	754816	540	1	Bella Maria	1
MBC_1545_CL1	7	330951	911251	282	2	Bella Maria	1
MBC_1545_CL1	7	934166	1174673	793	0	Bella Maria	1
MBC_1545_CL1	10	74632	241049	910	1	Bella Maria	1
MBC_1545_CL1	10	553233	691425	89	1	Bella Maria	1
MBC_1545_CL1	10	752929	1034308	766	3	Bella Maria	1
MBC_1545_CL1	15	106823	318213	1304	3	Bella Maria	1
MBC_1545_CL1	15	510886	782403	925	1	Bella Maria	1
MBC_1545_CL1	15	784193	934142	659	0	Bella Maria	1
MBC_1545_CL1	16	206821	502967	544	1	Bella Maria	1
MBC_1545_CL1	19	235735	342459	289	5	Bella Maria	1
MBC_1545_CL1	19	625331	769267	826	1	Bella Maria	1
MBC_1545_CL1	31	323674	456229	766	7	Bella Maria	1
MBC_1545_CL1	35	213894	501036	803	0	Bella Maria	1
MBC_1545_CL1	36	99405	228968	448	0	Bella Maria	1
MBC_1545_CL1	36	99405	228968	894	0	Bella Maria	1
MBC_1545_CL1	41	135171	239973	513	0	Bella Maria	1
MBC_1545_CL1	41	135171	239973	1023	0	Bella Maria	1
MBC_1545_CL1	42	394695	618542	513	5	Bella Maria	1
MCQ_1491_CL2	1	590428	761231	10	0	Bella Maria	1
MCQ_1491_CL2	1	1494729	1618252	64	3	Bella Maria	1
MCQ_1491_CL2	1	1494729	1618252	68	0	Bella Maria	1

Supplementary Table 2.2 (continued)

MCQ_1491_CL2	1	1494729	1618252	69	0	Bella Maria	1
MCQ_1491_CL2	1	2527893	2702894	123	4	Bella Maria	1
MCQ_1491_CL2	3	1190725	1409940	625	5	Bella Maria	1
MCQ_1491_CL2	4	1097976	1222344	27	0	Bella Maria	1
MCQ_1491_CL2	7	934166	1196685	262	0	Bella Maria	1
MCQ_1491_CL2	10	351472	525227	182	0	Bella Maria	1
MCQ_1491_CL2	10	750071	1059591	799	2	Bella Maria	1
MCQ_1491_CL2	12	615126	831143	32	0	Bella Maria	1
MCQ_1491_CL2	12	615126	831143	62	0	Bella Maria	1
MCQ_1491_CL2	13	380831	571325	672	2	Bella Maria	1
MCQ_1491_CL2	13	380831	571325	694	0	Bella Maria	1
MCQ_1491_CL2	13	590920	891408	285	1	Bella Maria	1
MCQ_1491_CL2	13	590920	891408	290	0	Bella Maria	1
MCQ_1491_CL2	14	768946	873091	352	0	Bella Maria	1
MCQ_1491_CL2	16	923925	1076491	615	0	Bella Maria	1
MCQ_1491_CL2	25	693	273362	1039	4	Bella Maria	1
MCQ_1491_CL2	27	46321	205232	627	2	Bella Maria	1
MCQ_1491_CL2	28	405150	507611	139	2	Bella Maria	1
TAZ_3961_CL2	7	925153	1113043	177	3	Ardanza	2
THY_4330_CL3	35	21625	153853	22	0	El Huayco	2

Supplementary Table 2.3 Recalculation of population genetic descriptive metrics using only one random *T. cruzi* I clone per vector/host. We reduced the dataset to identify biases related to multiple- vs. single-clone sampling per infection. While overall inference is similar, single-clone sampling can raise estimates of nucleotide diversity and rates of type II error in Hardy-Weinberg equilibrium null hypothesis testing (see power analysis in Supplementary Fig. 2.4). Abbreviations: PS (polymorphic sites); π (median nucleotide diversity, per site); θ (median Watterson estimator, per site); MAF (within-group minor allele frequency); PRS (private sites); SS (singleton sites); HWE (Hardy-Weinberg equilibrium); HS (heterozygous sites).

Group (n)	PS	π	θ	PS at MAF > 0.05	PRS (vs. BM / EH / AR)	SS	PS in HWE	HS	Fixed HS
Bella Maria (8)	95313	0.13	0.001	59%	0 / 41270 / 41063	22344	90461	55,571	2848
El Huayco (8)	76889	0.53	0.001	71%	22846 / 0 / 17681	6855	44911	56016	45792
Ardanza (6)	75709	0.55	0.001	71%	21459 / 16501 / 0	9844	72968	55638	47761

Supplementary Table 2.4 Re-sequencing of clones and subclones for additional ploidy analyses. Having entered cryopreservation (-150 °C) immediately after the first epimastigote DNA extraction (Dec. 2016), three clones were re-expanded into liquid culture and further subcloned by limiting dilution starting Dec. 2018. These clones and subclones underwent ≤ 4 passages in liver infusion tryptose (LIT) medium prior to epimastigote DNA extraction in Mar. 2019. Huge thanks to Jaime Costales and Jalil Maiguashca for preparing these samples. Abbreviations: RL (read-length); NRD (average nuclear read-depth).

ID	Type	Number of passages in LIT	RL (bp)	NRD
TAZ_4174_CL4_T2	Clone	0	2 x 75	135
THY_4326_CL1_T2	Clone	3	2 x 75	131
THY_4332_CL3_T2	Clone	4	2 x 75	180
TAZ_4174_CL4_T2_D1	Subclone	0	2 x 150	31
THY_4326_CL1_T2_D1	Subclone	0	2 x 150	49
THY_4326_CL1_T2_D2	Subclone	0	2 x 150	26
THY_4332_CL3_T2_D1	Subclone	0	2 x 150	44
THY_4332_CL3_T2_D2	Subclone	0	2 x 150	59
THY_4332_CL3_T2_D3	Subclone	0	2 x 150	41

Chapter 3

Hidden diversification during range expansion by *Leishmania infantum*, parasitic agent American visceral leishmaniasis

Philipp Schwabl[a], Mariana C. Boité[b], Arne Jacobs[c], Björn Andersson[d], Otacilio Moreira[e], Anita Freitas-Mesquita[f], José Roberto Meyer-Fernandes[f], Gerald F. Späth[g], Elisa Cupolillo[b] and Martin S. Llewellyn[a]

[a]Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK

[b]Laboratório de Pesquisa em Leishmaniose, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

[c]Department of Natural Resources, 02 Fernow Hall, Cornell University, Ithaca, NY, USA

[d]Department of Cell and Molecular Biology, Science for Life Laboratory, Karolinska Institutet, Biomedicum 9C, 171 77 Stockholm, Sweden

[e]Laboratório de Biologia Molecular e Doenças Endêmicas, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

[f]Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

[g]Hub de Bioinformatique et Biostatistique, Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France

Work presented in this chapter is currently under peer review at Communications Biology.

3.1 Abstract

Leishmania infantum causes American visceral leishmaniasis, a deadly vector-borne disease introduced to the New World during European colonization less than five hundred years ago. Within this short time period, the parasite has established widespread endemic transmission cycles using non-native vectors and human infection has become a major concern to public health, especially in Brazil. A multi-kilobase deletion occurs frequently in Brazilian *L. infantum* genomes and this deletion has been associated with resistance to miltefosine, an important anti-leishmanial drug. We apply multiple phenotypic and phylodynamic analyses to 126 *L. infantum* genomes to determine how demographic and selective consequences of recent invasive history have contributed to the emergence of this genotype and other epidemiological variability across Brazil. We revise geographical associations and describe deletion size differences with phylogenetic signals consistent with the occurrence of convergent deletion events in multiple clades. The deleted locus encodes ecto-3'-nucleotidase, and we show that loss of function in this important metabolic enzyme is coupled to ecto-ATPase upregulation, possibly creating a fitness advantage because both enzymes enable purine salvage, but they differ in antigenic traits. We also demonstrate altered phenotypes in heterozygous, 'half-deletion' genomes and prove that these represent recent genome-wide hybridizations between deletion-carrying and non-deletion isolates. The intricate and alarmingly labile population genetic structures we expose herein must be precisely monitored to guide future disease control.

3.2 Introduction

Species invasion creates a unique opportunity for extreme evolutionary transformation. Small founding populations face unfamiliar selection pressures and sampling effects that drive genetic drift. Rapid changes in genetic makeup may occur and can dictate long-term population genetic structure throughout the invasive range[429]. Subsequent secondary introductions into the same area can also reshape diversity patterns in the population, for example, by promoting introgressive hybridization events between ancestrally allopatric groups[430]. One medically relevant but little explored example of species invasion is represented by the introduction of *Leishmania infantum*, parasitic agent of visceral leishmaniasis (VL), into the New World during European colonization of the Americas ca. five hundred years ago[18,19]. Population structure and genetic change in *Leishmania* populations are of major concern to public health, as intra-specific genetic variation within this genus is associated with major differences in pathology[431–433], drug resistance[200,434] and other eco-epidemiological traits[435,436]. Driven by karyotypic plasticity[282,437], *Leishmania* parasites are capable of rapid adaptation and epidemic expansion after environmental change

and/or bottleneck events[434]. Genetic recombination among *L. infantum* populations is another potential source of phenotypic diversity. Hybridization between divergent *Leishmania* isolates and species that cause distinct forms of disease[294] can impact pathogenicity[155,244,294,438,439] as well as facilitate vector[154] and geographic range expansion[297].

In the Americas, VL is a zoonosis transmitted by *Lutzomyia* sandflies which have evolved in isolation of *Phlebotomus*, the Old World vector genus, for ca. two hundred million years[440]. Domestic dogs represent the principal reservoir hosts. The New World distribution of *L. infantum* now extends from the southern United States to northern Argentina[441] and Uruguay[442], but prevalence and/or reporting varies considerably across this range. Over one thousand VL cases have been recorded yearly in Brazil since the 1980's, first limited to the Northeast[443] but now increasingly dispersed, including in urban areas such as those in Mato Grosso, Minas Gerais and São Paulo state. VL infections are significantly less common elsewhere on the continent compared to Brazil[444]. Atypical cases, e.g., involving dermotropic or, more rarely, drug resistant *L. infantum* isolates, are also sporadically observed in the New World[445,446,250], but direct links between changes in disease progression and specific host or parasite factors are rarely established. A recently published genome-wide association study (GWAS)[258], however, reports that *L. infantum* populations from Piauí, Maranhão and Minas Gerais (Brazil) show resistance to miltefosine, an important anti-leishmanial drug, and associates this resistance to a large (> 12 kb) deletion said to increase in prevalence from north- to southeastern Brazil (e.g., 5% in Rio Grande do Norte and 95% in Minas Gerais). The deletion is homozygous, spanning across all four copies of tetrasomic chromosome 31 (chr31). It covers four open reading frames: LinJ.31.2370 (ecto-3′-nucleotidase/nuclease), LinJ.31.2380 (ecto-3'-nucleotidase precursor), LinJ.31.2390 (helicase-like protein) and LinJ.31.2400 (3,2-trans-enoyl-CoA isomerase). Ecto-3'-nucleotidases take part in purine salvage, macrophage infection and escape from neutrophil extracellular traps[447–449]. Helicases are essential to DNA replication and 3,2-trans-enoyl-CoA isomerase contributes to fatty acid oxidation, a critical component of gluconeogenesis in amastigote parasite forms[450]. The simultaneous deletion of these four genes likely occurs through non-conservative homologous recombination between repetitive elements shown to border the deletion site[258,20]. Carnielli et al. also very recently substantiated the statistical association between chr31 deletion and miltefosine treatment outcome[258] by demonstrating that locus knockout induces miltefosine resistance *in vitro* (findings presented at the British Society for Parasitology's March 2020 Trypanosomiasis and Leishmaniasis Seminar[259]). The mechanisms by which the chr31 deletion has come to occur in multiple different areas of Brazil, however, remain completely unknown. Its abundance and geographic patterns are also only rudimentarily described[258]. Selection pressure by miltefosine is unlikely to be

involved because the drug was not used in Brazil until 2005[451] and its very high costs have kept treatment with antimonials and/or amphotericin B far more common since then[452,453]. Analyses of demographic history, epidemiological phenotypes and genetic covariation in deletion-carrying isolates are urgently required to clarify the emergence of the deletion genotype and quantify its spread and implications for disease control.

The present study first extends survey for the chr31 deletion into nine additional states of Brazil, including also isolated localities in Bolivia, Honduras and Panama. Deletion-carrying isolates appear to dominate in most states (also in Rio Grande do Norte, in contrast to descriptions by Carnielli et. al (2018)[258]), yet with notable discontinuities within Piauí and between Mato Grosso and Mato Grosso do Sul. Our whole-genome and amplicon-based analysis of 201 *L. infantum* isolates then goes on to search for the deletion's origin and mechanisms of its proliferation in the context of invasive parasite expansion into the New World. We describe deletion size differences and phylogenetic relationships that are not symptomatic of an early ancestral mutation having risen to high prevalence simply by founder effect. Instead, multiple independent deletion events may have occurred and expanded into various clades. We demonstrate loss of ecto-3'-nucleotidase function coupled to increased ecto-ATPase activity in deletion-carrying isolates, suggesting the possibility that alternative metabolic strategies enhance *L. infantum* fitness in the introduced range. We also demonstrate altered phenotypes in highly heterozygous, 'half-deletion' parasite genomes. These are clearly the product of hybridization events between deletion-carrying and non-deletion isolates, also involving a highly divergent population from Mato Grosso do Sul. The distribution of *L. infantum* genetic and phenotypic diversity we summarize herein must be precisely monitored to guide future visceral leishmaniasis control.

3.3 Methods

3.3.1 Parasite samples and whole-genome sequencing

All 201 *L. infantum* samples assessed in this study are listed in Supplementary Tbl. 3.1, which also provides information on alternative nomenclatures, geographic origin, chr31 read-depth profile (i.e., whether or not isolate carry the chr31 deletion described by Carnielli et al. (2018)[258]) and analysis type (i.e., whole-genome sequencing (WGS) analysis or PCR product electrophoresis). All parasites sequenced in this study were obtained from the Coleção de *Leishmania* do Instituto Oswaldo Cruz (CLIOC) and were cultured in biphasic (Novy-MacNeal-Nicolle (NNN) + Schneider's) medium prior to genomic DNA extraction (DNeasy Blood & Tissue Kit (Qiagen). Mariana Boité performed all parasite handling and DNA extraction procedures above. Fragmented DNA (mean insert size = 377 nt) was sequenced using Illumina NextSeq 500 and HiSeq 2500 instruments and mapped to the

MCAN/ES/98/LLM-724 (termed JPCM5 elsewhere in the text) reference assembly available at https://tritrypdb.org/common/downloads/release-33/LinfantumJPCM5/fasta/ using default settings for BWA-mem v0.7.3[353]. Publicly archived and/or previously published *L. infantum* reads were mapped using the same conditions as the newly generated WGS data (see mapping coverage per sample in Supplementary Tbl. 3.1). For enzymatic assays (further described in Section 3.3.5), parasites were cultivated in flasks containing Schneider's medium with 20% fetal calf serum (FCS) and 2% filtered urine until late log-phase expansion. Growth curves were obtained to rule out samples with possible confounding differences in replication rate. All parasites used in the experiments showed similar replication rates. These parasites had been kept in culture between 10 and 20 passages after isolation and cryopreservation by CLIOC.

3.3.2 Phylogenetic, demographic modelling and selection analyses

We visualized genome-wide phylogenetic relationships among samples by maximum-likelihood tree construction in IQ-Tree v1.5.4[454], optimizing a general time-reversible substitution model based on single-nucleotide differences at polymorphic sites. *L. donovani* isolate MHOM/NP/03/BPK282/0 was temporarily included as an outgroup in order to identify an *L. infantum* sample to subsequently root the tree. Euclidean dissimilarities among genotypes were visualized by metric multidimensional scaling (PCoA)[455] using the base 'stats' package v3.4.1 in R v3.4.1[394]. Ancestry estimation was performed using ADMIXTURE v1.3[456] and putative first-generation (F_1) hybrid genotypes simulated from observed data by calculating allele frequencies of two parental populations, then drawing gametes following a multinomial distribution in the R package 'adegenet'[457]. Second-generation (F_2) hybrids were simulated by iterating the same process but with parental populations comprising the prior F_1 genotypes, and neighbor-joining (NJ) relationships among the simulated and observed data were plotted with the 'ape' package v5.0[393] in R v3.4.1[394]. For haplotype-based NJ trees, heterozygous single-nucleotide polymorphisms (SNPs) were first phased over 30 iterations using BEAGLE v4.1[392]. No genotype imputation was performed. We tested for admixture events in populations showing poor fit (high residuals) in tree-based phylogenies by searching non-treelike (graph) structures for higher maximum-likelihood in TreeMix v1.13[458]. The program also implements F_4-statistics to test significance of the improved fit.

Demographic histories inferred from phylogenetic analyses above were further tested by simulating ten different scenarios of pairwise divergence (ancient migration; ancient migration with bottleneck, isolation with (constant) migration; isolation with (constant) migration and bottleneck; isolation with change in migration; secondary contact; secondary

contact without hard admixture; secondary contact without hard admixture with bottleneck; strict isolation; and strict isolation with bottleneck) and associated genome-wide SNP polymorphism in fastsimcoal2 v2.5.2[315]. For each of > 100,000 random parameter sets simulated per divergence model, twelve summary statistics (total number of polymorphic sites; mean total heterozygosity; number of segregating sites per population; number of private sites per population; number of pairwise differences per population; mean and standard deviation of segregating sites over populations; and mean and standard deviation of pairwise differences over populations) were computed in ARLSUMSTAT v3.5.2[459]. Model selection and parameter estimations followed by Approximate Bayesian Computation via Random Forests (ABCRF) using 1,000-tree regression forests in the 'abcrf' package v1.7[316] in R v3.4.1[394].

Selection analyses between predefined groups (deletion-carrying and non-deletion type isolates) were performed by assessing site-wise F_{ST} neutrality with BayeScan v2.1[460]. We set prior odds for the neutral model to 100 and retained loci with $\log_{10}$ q-values less than -2, where false discovery rate is expected to fall below 1%. Results were then filtered for coding regions and SNP and insertion-deletion (INDEL) effects predicted with SNPEff v3t[391] using the JPCM5 annotation file available at https://tritrypdb.org/common/downloads/release-33/LinfantumJPCM5/gff/data/.

All above analyses were applied to SNPs and INDELs identified by local re-assembly, population-based genotype and likelihood assignment with Genome Analysis Toolkit (GATK) v3.7.0[389]. After testing various filtering criteria on the re-sequenced (paired-end 2 x 150 nt Illumina NextSeq) JPCM5 isolate, we chose to exclude variants occurring in tight clusters (i.e., more than three variants within ten bases) as well as those achieving less than 1,500 phred-scaled call quality (i.e., the variant-call-format QUAL field) as calculated by GATK. We also excluded variants assigned in non-unique mapping positions of the reference genome. Specifically, we generated synthetic, non-overlapping 125 nt sequence reads from the JPCM5 reference assembly (excluding unassigned contigs) and mapped these reads back to this same assembly using the 'mappability' program in the Genomic Multi-tool software suite v1.376[390,461]. Only variants from areas with perfect, i.e., singleton, synthetic mapping coverage were kept for SNP and INDEL analysis.

3.3.3 Chromosomal and gene copy number analyses

To estimate chromosomal somy, we calculated mean-read-depth (m) for successive 1 kb windows using SAMtools v0.1.18[387] 'depth' (default options) and then calculated a 'median-of-means' (M_m) for each chromosome. We let the 40th percentile (p40) of M_m values represent expectations for the disomic state, estimating copy number for each chromosome

by dividing its M_m by the sample's p40 value and multiplying by two. Copy numbers were then visualized with the 'heatmap.2' function in the 'gplots' package v3.0.1.2[462] in R v3.4.1[394]. Samples were organized in the heatmap based on UPGMA clustering of Bray-Curtis dissimilarities measured using the 'vegdist' function in the 'vegan' package v2.4.4[463].

Gene copy number analyses were guided by Hideo Imamura using scripts from Imamura et al. (2016)[200]. Briefly, we calculated median read-depth for each coding region (c) in the JPCM5 annotation file and then divided each c value by the median of c values across the chromosome to obtain a normalized copy number estimate (s) for each coding region of each sample. We then averaged s values from corresponding coding regions across samples within each of two predefined groups (deletion-carrying and non-deletion type isolates). Coding regions for which group means differed by more than 0.3 were selected for Mann-Whitney U (MWU) significance tests using SciPy v1.3.1[464]. Following Bonferroni correction (i.e., dividing the standard p-value cut-off of 0.05 by the number of coding regions submitted to MWU), we generated a heatmap of s values at coding regions which showed significant differences between the two groups, organizing samples by UPGMA clustering of Bray-Curtis similarities as in chromosomal somy visualization above. Coding regions with significant MWU results were also reassessed by analysis of covariance (ANCOVA) using the 'car' package v3.0.2[465] in R v3.4.1[394] to determine whether p-values remained significant after controlling for sample geographic origin. Isolates from Teixeira et al. (2017)[257] (see Supplementary Tbl. 3.1) were excluded from gene copy number analyses as these had not been made available as complete read-pairs in public sequence archives.

3.3.4 Monoclonal subcultures and qPCR

Single cell sorting was performed on a MoFLO ASTRIOS Cell Sorter (Beckman Coulter) by Mariana Boité at the Oswaldo Cruz Institute in Rio de Janeiro, Brazil. *L. infantum* isolates IOCL 2949 and IOCL 3134 entered cell sorting at $1 \cdot 10^6$ cells/µl and individual cells were collected in a 96-well plate, each well containing 200 µl Schneider's medium supplemented with 2% FCS. Wells were inspected five days later using an inverted microscope and liquid from those containing single parasites transferred to separate tubes of NNN. Parasites were pelleted three days later at 1,200 g for 15 min and DNA extracted with DNeasy Blood and Tissue Kit (Qiagen). Primer sequences 5'-ACGATCGGCCTCAAAACACT-3' (forward) and 5'- GGTGAAGTCTTCGTCCGTGT-3' (reverse) were designed to target LinJ.31.2380 (within the chr31 deletion site), and primer sequences 5'-CGAACCTTGGAGCTTCCCTT-3' (forward) and 5'-TCAAGGTTGTGTCCGTCGAG-3' (reverse) were designed to target LinJ.31.2330 (downstream of the chr31 deletion site). IOCL 2666 was used as a reference sample to calibrate the ΔΔCt method described by Livak & Schmittgen (2001)[466]. Briefly,

qPCR cycle thresholds (Ct values) for both chr31 sequence targets were determined for the samples of interest (IOCL 2949 and 3134 and their monoclonal subcultures) and for IOCL 2666. Ct values for the LinJ.2330 target were assumed to be equivalent between the sample of interest and the reference in the case of equal quantities of input DNA. Deviations from the 1:1 ratio for the LinJ.31.2330 target were used to normalize Ct ratios for the LinJ.31.2380 target between the sample of interest and the reference. The normalized ratios were considered to represent a fold change estimate of gene dose within the deletion site relative to that within downstream sequence. The qPCR reaction used 0.2 nM primer input and 1x SYBR Green Master Mix with 40 amplification cycles and an annealing temperature of 62 °C. Three experiments were performed per sample, each in technical triplicate. The same fold change estimation protocol was performed in follow-up analysis of monoclonal subcultures 2949 B2 and 2949 G1 using the parental culture IOCL 2949 as the reference. All above qPCR experiments were completed by Mariana Boité and Otacilio Moreira.

3.3.5 Ecto-3'-nucleotidase and ecto-ATPase activity measurement

Ecto-3′-nucleotidase activity was quantified by measuring inorganic phosphate (P_i) release during adenosine 3'-monophosphate (3'-AMP) hydrolysis as in Freitas-Mesquita et al. (2016)[447]. Briefly, *L. infantum* promastigotes ($1.0 \cdot 10^7$ cells/ml) were incubated at 25 °C for 1 h in 0.5 ml reaction mixture containing 16.0 mM NaCl, 5.4 mM KCl, 5.5 mM D-glucose, 50.0 mM HEPES (pH 7.4) and 3.0 mM 3'-AMP. Reactions were terminated by adding 1.0 ml ice-cold 25% charcoal in 0.1 M HCl and centrifuged at 1,500 g for 15 min to remove nonhydrolyzed 3'-AMP. Equal volumes of supernatant and Fiske & Subbarow reagent (0.1 ml each) were mixed to effect the (phosphate-dependent) reduction of ammonium molybdate to phosphomolybdate and absorbance at 660 nm in samples and P_i standards measured after 30 min to derive sample P_i. Ecto-ATPase activity was measured with the same protocol except replacing 3'-AMP with 1.0 mM adenosine 5'-triphosphate (ATP) and 1.0 mM $MgCl_2$. Experiments were performed in technical triplicates using IOCL 2664, 2666, 2972, 3598 and 3634 and monoclonal subcultures 2949 B2 and 3134 B1. All above procedures were completed by Anita Freitas-Mesquita and José Roberto Meyer-Fernandes.

3.4 Results

3.4.1 High prevalence of multi-kilobase deletion on chr31

Comparative analysis of 126 New World and Old World *L. infantum* genomes against the JPCM5 reference assembly confirmed the occurrence of a > 12 kb homozygous deletion on tetraploid chr31 (see somy values in Supplementary Fig. 3.1), previously described as a miltefosine sensitivity locus[258]. The deletion occurred in 73 sequenced New World genomes and in 55 of 75 additional New World samples screened via qPCR. Sequenced deletion-carrying isolates (hereafter referred to as 'Del') originated from Brazil (71 of n = 91) and Honduras (2 of n = 2) but none were found in the Old World (0 of n = 19). Thirty-eight non-deletion ('NonDel') isolates were found in restricted regions of Brazil (concentrated primarily in Piauí and Mato Grosso do Sul), but also in Panama (2 of n = 2) (Supplementary Tbl. 3.1, Fig. 3.1) and in the Old World (19 of n = 19). The deleted region spans base pair positions 1,122,848 to 1,135,161 in most Del samples (but see variability in deletion start/stop sites in Supplementary Tbl. 3.2) and comprises genes encoding for ecto-3'-nucleotidase (LinJ.31.2370), ecto-3'-nucleotidase precursor (LinJ 31.2380), helicase-like protein (LinJ 31.2390), and 3-2-trans-enoyl-CoA isomerase (LinJ.31.2400). Apart from the relatively large deletion, an additional 391 fixed INDELs (small insertions or deletions up to 30 nt) were found in Del isolates. Of these, however, 260 were also fixed in New World NonDel isolates and 98% occurred in non-coding sequence regions. The two INDELs found in coding regions and fixed only in Del isolates (Supplementary Tbl. 3.3) affect hypothetical proteins LinJ.25.0280 and LinJ.27.0140 without further annotation on TriTrypDB. Forty-one SNPs occurring in coding regions and fixed only in Del isolates affected annotated proteins (Supplementary Tbl. 3.3) but none deviated from neutrality in site-wise F_{ST} differentiation tests (see BayeScan and alternative selection analyses (including also non-fixed variants) in Supplementary Fig. 3.2 and Supplementary Tbl. 3.4). Forty-two coding regions showed significant copy number variation (CNV) between Del and New World NonDel groups in haploid somy estimate (s) comparison using Mann-Whitney U tests (Supplementary Tbl. 3.5), but reassessment by ANCOVA suggested that most of these differences are driven by population structure, i.e., common descent. Supplementary Fig. 3.3 illustrates how CNV profiles cluster by geographic origin, and geographic origin correlates to chr31 read-depth profile. The five coding regions for which s remained significantly differentiated between Del and New World NonDel groups after controlling for geographic origin encode amastin-like protein, nucleoside transporter and paraflagellar rod protein paralogs (see asterisked columns in Supplementary Fig. 3.3). Effect size, however, is low ($0.317 \leq |\Delta s| \leq 0.552$) (Supplementary Tbl. 3.5).

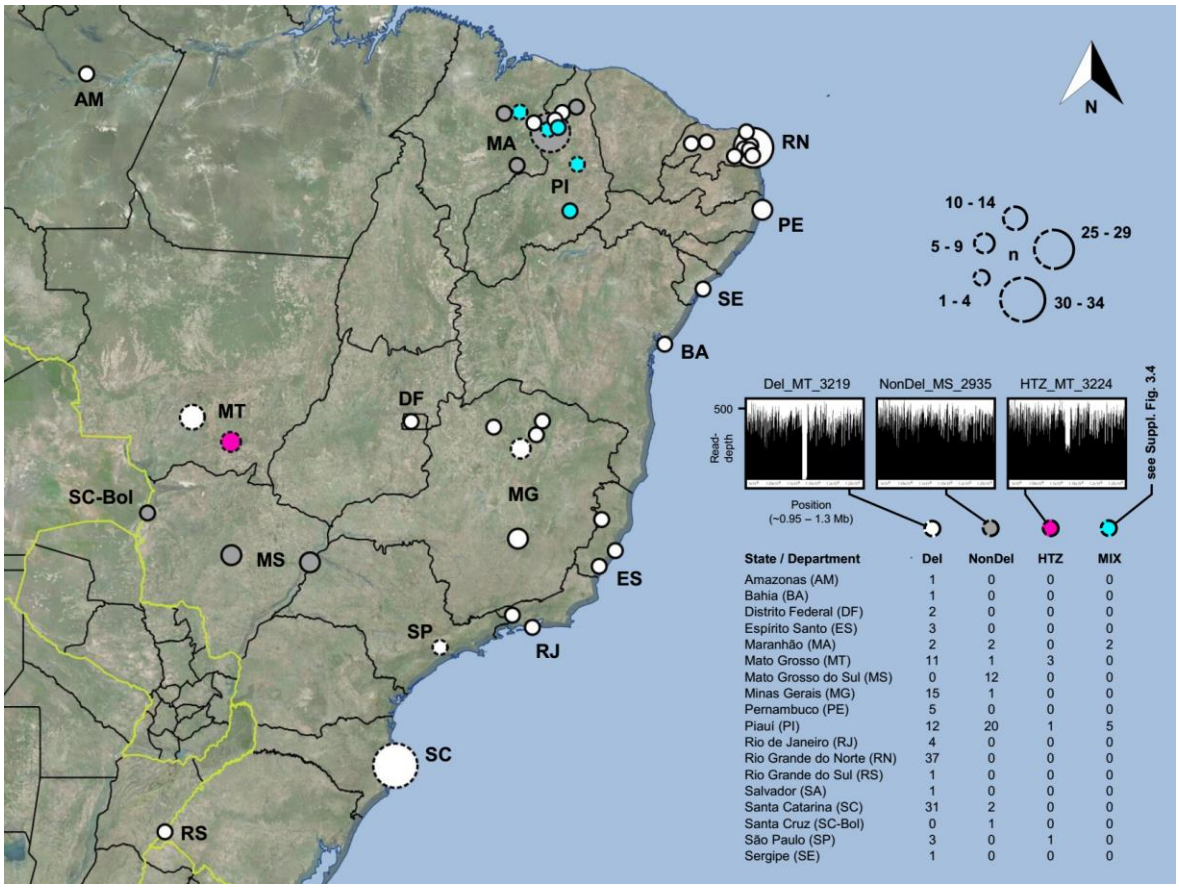


Figure 3.1 Different read-depth profiles found in *L. infantum* isolates from Brazil. Del isolates contain a > 12 kb deletion between 1.122 Mb and 1.135 Mb on chr31 (e.g., Del_MT_3219 in the left graph). NonDel isolates do not contain the deletion, showing full read-depth at the locus (center graph). HTZ isolates are heterozygous for the deletion, with read-depth dropping to ca. 50% (right graph). Quantitative PCR confirmed heterozygosity at the deletion locus in monoclonal HTZ subcultures. MIX isolates appear to contain a mixture of NonDel and Del or HTZ profiles based on subclone PCR by Carnielli et al. (2018)[258]. However, full read-depth is observed at the deletion locus in all MIX isolates except in MIX_PI_05A and MIX_PI_08A (showing ca. 75% read-depth – see Supplementary Fig. 3.4). This suggests that NonDel cells are more abundant than Del and/or HTZ cells within MIX isolates. Circle radius indicates the number of isolates (each from a different canine or human host) representing the study site. Dotted circles represent study sites where multiple read-depth profiles occur (see table inset). Fill color indicates the majority read-depth profile at such study sites.

3.4.2 Partial deletion genotypes occur in sympatry with Del and NonDel isolates

Six *L. infantum* samples sequenced in this study had an intermediate read-depth profile within the chr31 deletion site (Supplementary Tbl. 3.1). In such genotypes, sequences mapped to the deletion site achieve approximately 50% read coverage relative to the rest of the chromosome (Supplementary Fig. 3.4), suggesting one of two scenarios: an abundance of balanced heterozygous cells or a mixed population of Del and NonDel isolates. We therefore extracted DNA from eleven monoclonal subcultures established from two isolates representing putative heterozygotes (IOCL 2949 and 3134) and measured relative abundance of the deletion target by qPCR. Results from ten monoclonal subcultures showed a reduction of ca. 50% in the abundance of the amplified target sequence relative to the NonDel

representative NonDel_MS_2666 (Fig. 3.2), confirming the presence of cells heterozygous at the deletion locus as opposed to a mix of (homozygous) Del and NonDel genotypes. Clone 2949 G1 showed 25% relative target amplification (Fig. 3.2b), suggesting the presence of three chromosome copies with the deletion, and one copy without. Subpopulations with different levels of heterozygosity appear to occur but 'equivalent' heterozygotes – i.e., cells in which two copies of chr31 carry the deletion, and two copies do not – appear most abundant based on read-depths of DNA sequenced from the parental culture (Supplementary Fig. 3.4). Apart from these six isolates (hereafter termed 'HTZ'), seven isolates sequenced by Carnielli et al. (2018) simultaneously showed Del and NonDel deletion site PCR amplicons[258] but ca. 75 – 100% read-depth within the deletion site (Supplementary Fig. 3.4). The authors were not conclusive about whether these samples represented single or mixed isolates; we therefore refer to them hereafter as 'MIX'.

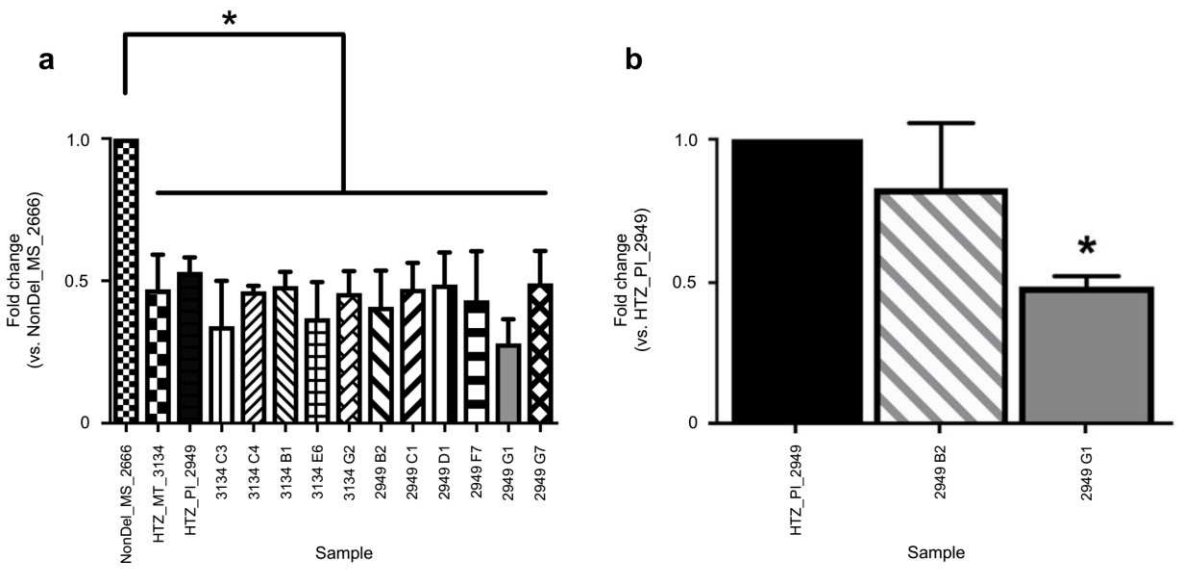


Figure 3.2 Quantitative PCR confirms that intermediate read-depth profiles represent heterozygous deletions in *L. infantum* clones. **a** HTZ_PI_2949 and HTZ_MT_3134 were selected as representatives of isolates for which read-depth drops to ca. 50% between 1.122 Mb and 1.135 Mb on chr31 (see Supplementary Fig. 4). DNA from monoclonal subcultures established from these two isolates was analyzed in qPCR targeting LinJ.31.2380 (within the chr31 deletion site) and LinJ.31.2330 (downstream of the chr31 deletion site). Differences in Ct values for LinJ.31.2330 between each HTZ sample and the NonDel reference (NonDel_MS_2666) were used to normalize a fold change estimate at LinJ.31.2380 based on the ΔΔCt method by Livak and Schmittgen (2001)[466]. Student's t-test was applied to test whether fold change estimates obtained from triplicate reactions differed significantly from the 1:1 ratio represented by the reference sample. Results were considered significant at $p < 0.05$ (*) and indicate that intermediate read-depth profiles represent abundant heterozygous deletions as opposed to mixtures of deletion-carrying and non-deletion type cells within isolates. **b** Fold change was calculated the same way for monoclonal HTZ subcultures using the parental isolate as the reference. Results indicate that 'unbalanced' heterozygotes also occur, e.g., 2949 G1 appears to contain three chromosome copies with the chr31 deletion, and one copy without.

3.4.3 HTZ isolates represent the hybrid offspring of Del and NonDel isolates

Given the vast geographic range occupied by Del isolates (Fig. 3.1) and its broad overlap with that of the parasite's new vector species, *Lu. longipalpis*[467,218], we considered the possibility of independent deletion emergence as an adaptive process recurring frequently across the American continent. We hypothesized that HTZ isolates represent former NonDel isolates currently undergoing a step-wise deletion process – as in the non-equivalent HTZ clone 2949 G1, and that these later give rise to homozygous variants by biased mitotic replication, i.e., haplotype selection[282]. Following ADMIXTURE analysis (Supplementary Fig. 3.5), however, in which HTZ_MT_3134, HTZ_MT_3135, HTZ_MT_3137, HTZ_MT_3224 and HTZ_SP_3254 (i.e., all HTZ samples except HTZ_PI_2949) received simultaneous Del + NonDel group assignment, we also considered the alternate hypothesis that HTZ isolates represent hybrid offspring forming at contact zones between Del and NonDel groups (Fig. 3.1). Support for this alternate hypothesis quickly accumulated through several analyses and metrics.

HTZ samples showed marked, statistically significant reductions in total homozygosity and F_{IS} values (which describe the extent to which individual heterozygosity is reduced by inbreeding) relative to Del and to New World NonDel isolates (Fig. 3.3a, Supplementary Tbl. 3.6). Median F_{IS} was lowest in HTZs (relative to Del, New World and Old World NonDel groups) in 30 of 36 chromosomes (Fig. 3.3b). Except for HTZ_PI_2949, HTZs occurred in peripheral positions relative to monophyletic Del subclades in maximum-likelihood phylogeny (Fig. 3.4) and showed intermediate axis positions in PCoA (Fig. 3.5a). We also constructed neighbor-joining trees from phased chromosomes (Supplementary Fig. 3.6), and homologous haplotypes of HTZ isolates divided between Del and Mato Grosso Do Sul NonDel clades, consistent with genome fusion or a Mendelian mechanism of genetic exchange with back-crossing or inter-breeding among hybrid isolates. F_{ST} differentiation to Mato Grosso do Sul samples also fluctuated among HTZ chromosomes, consistent with chromosomal reassortment as a result of mating between Del and NonDel isolates (Supplementary Fig. 3.7). We further examined a potential hybrid origin by comparing the phylogenetic positions of HTZ isolates from Mato Grosso with those generated by simulated sexual mating between populations from Mato Grosso and nearby Mato Grosso do Sul. Phylogenetic positions for simulated hybrids corresponded to those observed for HTZ isolates (Fig. 3.5b). In these simulations, we also hypothesized the presence of second-generation (F_2) hybrids, that is, we simulated back-crossing and hybrid inter-crossing to account for the origin of Mato Grosso samples Del_3223 and NonDel 3210 (respectively).

These two samples are not heterozygous for the deletion on chr31 but show aberrant genome-wide heterozygosity and F_{IS} (Fig. 3.3). Phylogenetic positions of the simulated F_2 hybrids matched positions of Del_3223 and NonDel_3210. Similar F_2 hybridization events may also explain the outlying phylogenetic positions of samples such as NonDel_MG_14A or NonDel_MS_2688 (Figs. 3.4 and 3.5a, and Supplementary Fig. 3.6).

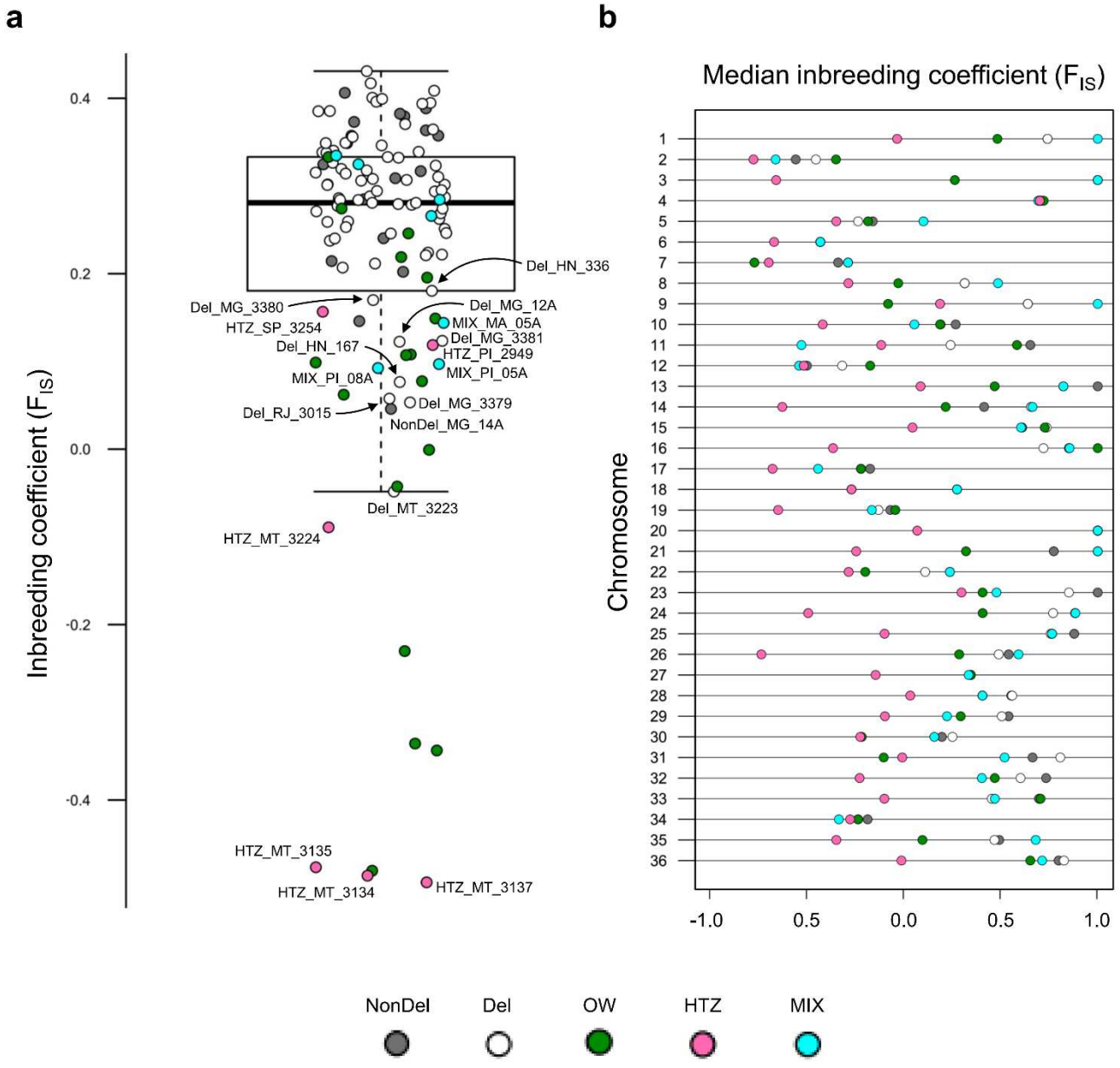


Figure 3.3 Homozygosity relative to Hardy-Weinberg expectations in New and Old World *L. infantum* isolates. **a** The box plot shows median and interquartile ranges of genome-wide inbreeding coefficients (F_{IS}). Values are generally high for New World isolates. Values for HTZ isolates, however, all occur below the second quartile and strong excess heterozygosity is suggested in HTZ_MT_3134, HTZ_MT_3135, and HTZ_MT_3137. **b** Relatively low genome-wide F_{IS} in HTZ isolates is not driven by values from a subset of chromosomes. Values appear low throughout the genome. Circle fill color indicates New vs. Old World origin and read-depth profile on chr31.

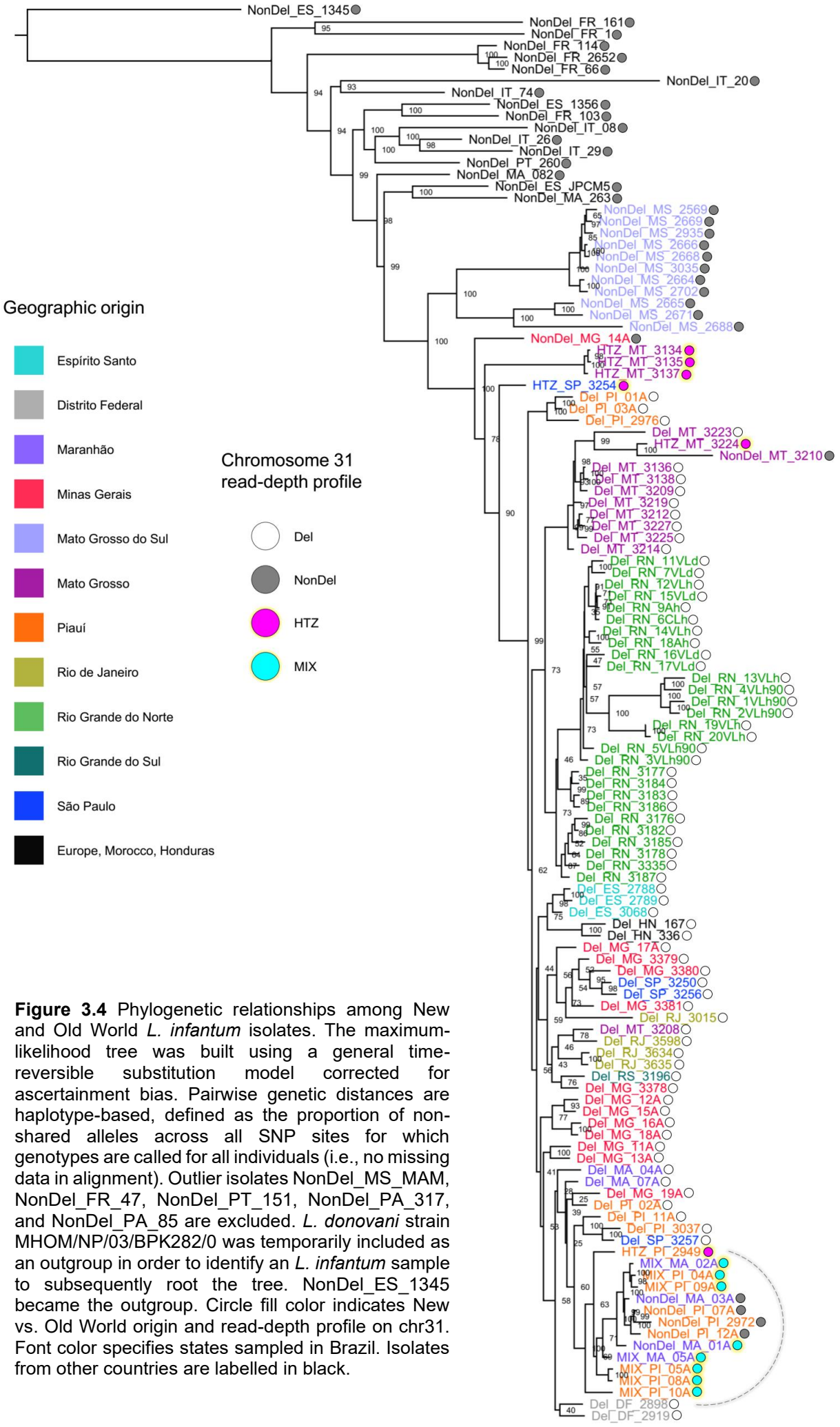


Figure 3.4 Phylogenetic relationships among New and Old World *L. infantum* isolates. The maximum-likelihood tree was built using a general time-reversible substitution model corrected for ascertainment bias. Pairwise genetic distances are haplotype-based, defined as the proportion of non-shared alleles across all SNP sites for which genotypes are called for all individuals (i.e., no missing data in alignment). Outlier isolates NonDel_MS_MAM, NonDel_FR_47, NonDel_PT_151, NonDel_PA_317, and NonDel_PA_85 are excluded. *L. donovani* strain MHOM/NP/03/BPK282/0 was temporarily included as an outgroup in order to identify an *L. infantum* sample to subsequently root the tree. NonDel_ES_1345 became the outgroup. Circle fill color indicates New vs. Old World origin and read-depth profile on chr31. Font color specifies states sampled in Brazil. Isolates from other countries are labelled in black.

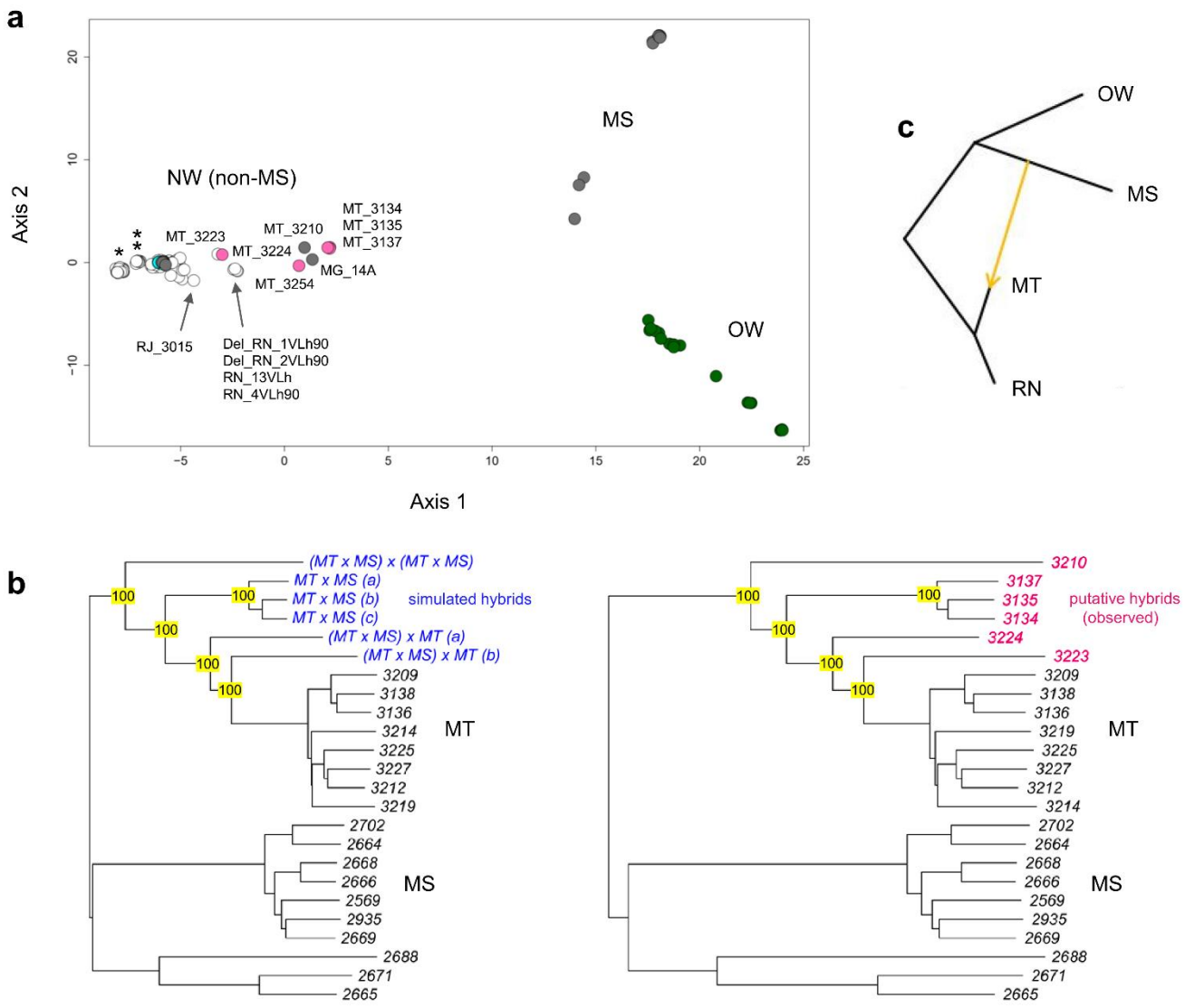


Figure 3.5 Metric multidimensional scaling, simulated mating and tree-to-graph conversion suggest admixture and hybridization between Del and NonDel *L. infantum* groups. **a** Metric multidimensional scaling clearly separates New and Old World (NW and OW) isolates on two axes of variation (goodness-of-fit = 0.40). NonDel isolates from Mato Grosso do Sul (MS) and Del isolates from Rio Grande do Norte (RN, see asterisk) and Mato Grosso (MT, see double-asterisk) position at opposite ends of axis 1, the primary axis of divergence within and between NW populations. HTZ isolates occur at intermediate positions (see pink circles) between these dissimilar groups. Othe r isolates with such intermediate positions are labelled and may also represent mating events between dissimilar groups. Grey, white and cyan fill colors, respectively, indicate NonDel, Del and MIX read-depth profiles found in the NW. Circles for OW (NonDel) isolates are green. Five outlier isolates are excluded as in Fig. 3.4. **b** Neighbor-joining positions of simulated hybrids (blue font, left tree) correspond to those of observed HTZ isolates (pink font, right tree) from MT. Hybrids were simulated in two steps. Random 50% haplotype contributions were first drawn from Del and NonDel isolates observed in MT and MS. The resultant offspring genotypes were then either let diversify through random mutation or subjected to a second round of Mendelian recombination as before. The same tree topology resulted in each of 100 simulation replicates. **c** Given that mating can create non-treelike divergence patterns within species, TreeMix[458] was used to search iteratively for up to five migration edges that improve the fit of a maximum-likelihood tree built based on Gaussian approximation of genetic drift among isolates from MT, MS, RN and OW groups. This input tree (black edges) suggests dichotomous differentiation into MT/RN and MS/OW clades and has a log-likelihood of 84.9206. Tree-to-graph conversion by addition of a migration edge from MS to MT increases log-likelihood to 84.9775. No other edges further increase the fit of the input tree. A four-population test[468] also supports post-split admixture between MS and MT or RN because differences in allele frequencies between MT and RN isolates correlate with those within the other population pair (F_4-statistic = $5 \cdot 10^{-5}$, Z-score = 3.51).

3.4.4 Possible Del paraphyly and phenotypic consequences of the chr31 deletion

Although the above analyses indicate that HTZs in this sample set represent hybrids, not intermediate forms within a process of stepwise mitotic deletion, they do not exclude that such mitotic deletion is recurrently creating Del genotypes (via intermediate forms) throughout Brazil. In SNP-based phylogeny (Fig. 3.4), New World isolates branch out from within the Old World clade into two main clusters, one containing divergent isolates from Mato Grosso do Sul, the second containing all other sample genomes. Within this second cluster, Del isolates do not form a private clade (see phylogenetic positions of NonDel isolates from Maranhão and Piauí, as well as HTZ_PI_2949). This paraphyly contradicts the hypothesis that the chr31 deletion represents a rare ancestral mutation whose present abundance mimics the results of frequent selection but is actually a consequence of founder effect, i.e., success due to emergence in small populations before or during early phases of range expansion. We also find that the deletion locus start/stop coordinates differ among Del subgroups (Supplementary Tbl. 3.2), and stop coordinates statistically correlate with sample phylogenetic positions (Blomberg's $K = 3.192$, $p = 0.001$)[469], further evidence against ancestral deletion and common descent. Proliferation of Del genotypes against different genetic backgrounds suggests recurrent selection, which implies that locus deletion alters phenotype. We performed an assay for ecto-3'-nucleotidase activity on Del, NonDel and HTZ isolates from different geographic regions (Fig. 3.6a). Results demonstrate complete loss of function in Del isolates by comparison to HTZ and NonDel parasites ($p < 0.05$). Significant inter-individual variation also occurs in ecto-3'-nucleotidase activity between NonDel isolates NonDel_MS_2664 and NonDel_MS_2666 ($p < 0.05$). We also measured ecto-ATPase activity (Fig. 3.6b), thought to be involved in purine salvage pathways alternative to those of ecto-3'-nucleotidase[470,471]. We found greater ecto-ATPase activity among Del vs. NonDel isolates ($p < 0.05$).

3.4.5 Pattern-process modelling and the biogeography of *L. infantum* diversity

Parasite isolates from Mato Grosso do Sul sequenced in this study stand out in their complete lack of Del genotypes and their basal phylogenetic positions in Fig. 3.4. Compared to the rest of the New World sample set, this outgroup also showed higher nucleotide diversity (π) per site (0.046 vs. 0.061, respectively), more than twice as many private SNP sites per sample (15.3 vs. 31.8) and lower F_{ST}-differentiation from Old World isolates (0.413 vs. 0.303) (Tbl. 3.1). We therefore hypothesized (H_1) that Mato Grosso do Sul isolates represent a separate, more recent introduction to Brazil or that (H_2) they stem from the same introduction event as other New World isolates of the sample set but experienced distinct

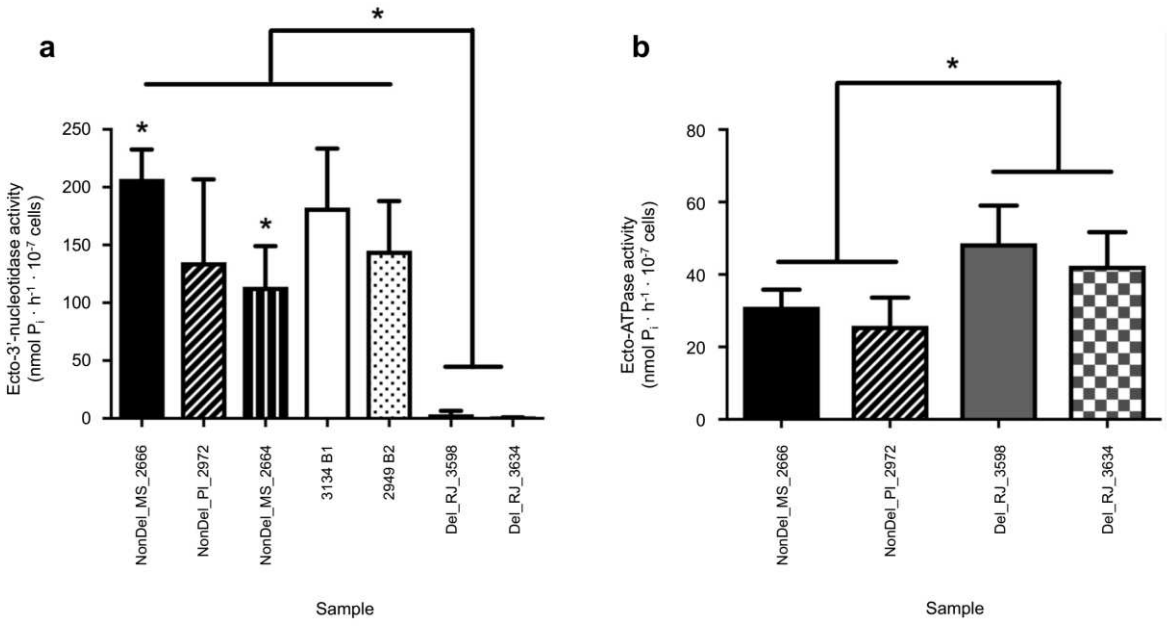


Figure 3.6 Ecto-3'-nucleotidase and ecto-ATPase activity correlates to read-depth profiles on chr31. **a** Ecto-3′-nucleotidase activity was quantified by measuring the rate of inorganic phosphate (P_i) release during adenosine 3'-AMP hydrolysis as described in Freitas-Mesquita et al. (2016)[447]. Bar plots show mean and standard error for three replicate assays. Student's t-test was applied to test for statistical significance between pairs of samples at $p < 0.05$ (*). Results indicate complete loss of function in Del isolates and no significant differences between NonDel isolates and monoclonal HTZ subcultures. **b** Ecto-ATPase activity was quantified with the same protocol except replacing 3'-AMP with equimolar ATP and Mg^{2+}. T-tests between NonDel and Del isolates suggest higher ecto-ATPase activity in Del than in NonDel isolates, but larger samples sizes are required to substantiate the effect.

demographic processes in subsequent dispersal. In either case, considering overall New World monophyly and 164 (of 2,383) SNP sites fixed across all New World samples, but not also fixed across all Old World samples, both Mato Grosso do Sul and non-Mato Grosso do Sul populations likely originate from a common, unsampled Old World region. Treating Mato Grosso isolates as representative of the wider non-Mato Grosso do Sul clade, we used pattern-process modelling to test demographic histories implied by hypotheses H_1 and H_2. H_1 implies a 'secondary contact' (SC) model of divergence, whereby gene flow between Mato Grosso and Mato Grosso do Sul fully ceases but is later reestablished, as would be the case if there had been distinct introduction events of *L. infantum* into the Americas. H_2 implies an 'isolation with migration' (IM) model of divergence, whereby contact never completely ceases, as would likely occur if samples from Mato Grosso do Sul and Mato Grosso had diverged following a single introduction event to Brazil. For both SC and IM models, we simulated individual genome-wide SNP diversity in three variations relating to bottleneck (yes/no in Mato Grosso founder population) and admixture type (hard

Table 3.1 Population genetic descriptive metrics for New World and Old World *L. infantum* groups. HTZ and MIX genotypes are not used in this analysis. Abbreviations: MS (Mato Grosso Do Sul); non-MS (New World, excluding MS); n (sample size); K (mean number of alleles, per locus); Het. (mean heterozygosity); PS (total polymorphic sites); PRS (private sites, per sample); π (nucleotide diversity); F_{ST} (between-group fixation index).

Group (n)	K	Het.	PS	PRS	π	F_{ST} to OW	F_{ST} to MS	F_{ST} to non-MS
non-MS (80)	2.01	0.122	1782	15.3	0.046	0.419	0.495	0.000
MS (11)	2.00	0.324	903	31.8	0.061	0.304	0.000	0.495
Old World (17)	2.00	0.195	3069	149.1	0.125	0.000	0.304	0.419

introgression and/or permanent migration vs. temporary genetic exchange). We also ran simulations for two implausible models of Mato Grosso – Mato Grosso do Sul divergence, 'strict isolation' (SI, i.e., no contact between populations) and 'ancient migration' (AM, i.e., no contact after an early period of contact between populations). These served as controls for the ABCRF[316] method, which uses random forests to rank the fit of observed vs. simulated summary statistics. Simulations for Mato Grosso – Old World and Mato Grosso do Sul – Old World population pairs, both assumed to follow an 'ancient migration' with bottleneck (AM_{bot}) model of divergence, provided additional method control (see fastsimcoal2[315] template files and model illustrations in Supplementary Tbl. 3.7 and Supplementary Fig. 3.8). Following expectations, the AM_{bot} model achieved highest posterior probability support for both Mato Grosso – Old World and Mato Grosso do Sul – Old World divergence, with Mato Grosso experiencing a tighter bottleneck than Mato Grosso do Sul (-80% vs. -71%) at the time of separation from the ancestral Old World group (Tbl. 3.2). Also as expected, AM and SI models received lowest (near zero) support for the Mato Grosso – Mato Grosso do Sul population pair, but support values of the remaining IM and SC models did not clearly favor H_1 or H_2 (Tbl. 3.2). The top-ranked IM base model (without explicit bottlenecking in the early Mato Grosso population) scored only slightly higher than SC_{nomig} (no explicit bottleneck in the early Mato Grosso population, and no complete migration when contact is later reestablished with Mato Grosso do Sul). Neither achieved high posterior probability support given competition from highly similar variations of each model (Tbl. 3.2). Parameterization of the top-ranked IM model indicated unbalanced gene flow (MIG), predominantly to Mato Grosso from Mato Grosso do Sul ($MIG_{\text{Mato Grosso do Sul – Mato Grosso}} = 0.025$ vs. $MIG_{\text{Mato Grosso – Mato Grosso do Sul}} = 0.004$) (Tbl. 3.2). Gene flow in this direction also significantly increased likelihood in phylogenetic tree optimization by graph conversion using TreeMix[458], complemented by F_4 statistics support (Fig. 3.5c). Definitive evidence for a single or multiple *L. infantum* invasion events into Brazil, however, could not be found.

Table 3.2 Demographic simulation in fastsimcoal2[315] and model selection by Approximate Bayesian Computation via Random Forests (ABCRF)[316]. In fastsimcoal2 simulation, values for past and present population sizes were drawn randomly from a uniform distribution between 100 and 10^6 individuals. Values for time of secondary contact were drawn randomly from a uniform distribution between 0 and $2 \cdot 10^4$ generations before present. Values for relative migration rates between populations were drawn randomly from a log-uniform distribution between 10^{-10} and 0.1. Values for bottleneck size were drawn randomly from a uniform distribution between 0.05 and 0.5. The mutation rate was fixed at $1.99 \cdot 10^{-9}$ mutations per bp on all chromosomes. The ten different demographic models are illustrated in Supplementary Fig. 3.8 and template file content is provided is provided in Supplementary Tbl. 3.7. Abbreviations: CV (classification vote, i.e., the number of times a model is selected in a forest of 1,000 trees (the model with the most votes corresponds to the model best suited to the dataset)); PP (ABCRF approximation of the posterior probability of the selected model); FOU (bottleneck size, i.e., the fraction of prior population size at the end of the bottleneck); N_{draws} (number of parameter draws simulated by fastsimcoal2 as input for ABCRF); $MIG_{x>>>y}$ (migration rate from x to y); Pop. (population); MT (Mato Grosso); MS (Mato Grosso do Sul); OW (Old World).

Models of divergence between MT and MS *L. infantum* groups

Model	Pop. 1 / Pop. 2	CV	N_{draws}	
AM	MT / MS	0.020	474177	
IMbot	MT / MS	0.151	452533	
IM_{change}	MT / MS	0.089	476483	
IM	MT / MS	0.215	474263	*selected model
SC	MT / MS	0.125	473082	
$SCbot_{nomig}$	MT / MS	0.187	427249	
SC_{nomig}	MT / MS	0.201	474782	
SI	MT / MS	0.012	466136	

PP = 0.282
$MIG_{MT>>>MS}$ = 0.004
$MIG_{MS>>>MT}$ = 0.025

Models of divergence between MT and OW *L. infantum* groups

Model	Pop. 1 / Pop. 2	CV	N_{draws}	
AMbot	MT / OW	0.304	432323	*selected model
AM	MT / OW	0.186	458125	
IM_{change}	MT / OW	0.106	470330	
IM	MT / OW	0.215	459566	
SC	MT / OW	0.161	421405	
SC_{nomig}	MT / OW	0.013	464907	
SIbot	MT / OW	0.003	385170	
SI	MT / OW	0.012	409244	

PP = 0.485
FOU = 0.204

Models of divergence between MS and OW *L. infantum* groups

Model	Pop. 1 / Pop. 2	CV	N_{draws}	
AMbot	MS / OW	0.385	413704	*selected model
AM	MS / OW	0.161	472457	
IM_{change}	MS / OW	0.145	473388	
IM	MS / OW	0.170	471073	
SC	MS / OW	0.025	471677	
SC_{nomig}	MS / OW	0.031	472251	
SIbot	MS / OW	0.035	463789	
SI	MS / OW	0.048	457084	

PP = 0.521
FOU = 0.292

3.5 Discussion

3.5.1 Principal findings

Our results reveal the widespread distribution of a major genetic alteration found in New World *L. infantum* isolates, clarifying that a multi-kilobase, loss-of-function deletion on chr31 predominates in the South(east), East as well as in the Northeast of Brazil. Additional point sampling also detected the deletion in distant Amazonas and as far north as Honduras, but not in Panama. Our observations do not suggest a continuum in deletion rate as previously proposed, showing instead an intermittent preeminence of sometimes closely related, other times highly divergent NonDel isolates, particularly in Piauí and Mato Grosso do Sul. Deletion size and phylogenetic variation suggest that recurrent evolution may have led to the widespread, yet discontinuous preponderance of Del genotypes we observe. The New World parasite population has, however, undergone recent invasive expansion, which, in the most parsimonious case, involved just a single, not multiple, introduction events. Confirming paraphyletic deletion origins is especially complicated by the frequent outcrossing and inbreeding events we expose among deletion-carrying and non-deletion parasite genomes.

3.5.2 General discussion

In microbial ecology, but also at various other scales, species invasion or population expansion creates the unique opportunity for rare, non-adaptive mutations to spread rapidly across new territories by riding expanding wave fronts[472–475], where population density is low and growth rate is high[476]. Genetic diversity patterns in this study, however, are not clearly symptomatic of such so-called 'allele-surfing' effect. Rather, we find evidence for recurrent independent deletion events based on subtle, yet significant deletion stop site differences among geographically separated parasite groups and phylogenetic nesting of NonDel genotypes (from Piauí and Maranhão) within what would otherwise appear as a monophyletic deletion-carrying clade. Results from our phenotypic assays also suggest that alternate, compensatory metabolic pathways may exist to counteract the elimination of ecto-3'-nucleotidase, for which Del parasites showed complete loss of function despite the presence of JPCM5-like (but apparently pseudogenic) paralogs on chromosome 12. Ecto-3'-nucleotidases participate in purine salvage essential to trypanosomatid survival[477,478] but also act as virulence factors during infection of mammalian hosts[470,479]. It would be interesting to test for virulence differences in relation to the increase in ecto-ATPase levels we observe and whether reduced virulence or antigenicity might confer positive fitness effects within vertebrate and invertebrate phases of the life cycle. Natural selection may favor reduced

virulence in chronic infections with low transmission rates[480], as is the case for VL in Brazil[197,481], where symptomatic hosts are also targeted by disease control[193].

Considering the possible compensatory elevation in ecto-ATPase activity we measured at the phenotypic level, we also scanned for possible compensatory gene CNV in Del isolates. CNV-based UPGMA placed NonDel samples from Mato Grosso do Sul and Minas Gerais (NonDel_MG_14A) into separate basal clades while keeping NonDel isolates from Piauí and Maranhão clustered together with Del isolates from the same northern states. This geographically correlated CNV topology mirrors that of the SNP-based phylogeny and thus suggests that (baseline) gene copy numbers or deletion/amplification programs triggered *in vitro* are conserved among related isolates. The latter phenomenon was also recently proposed in Bussotti et al. (2018)[274]. Our results do not suggest that a single CNV regime underlies enzymatic changes (e.g., ecto-ATPase upregulation) that might be occurring to compensate loss of function at the deleted locus on chr31. Such compensation may occur through unique (i.e., sample-specific) CNV solutions or by various other epigenetic, post-transcriptional or post-translational effects. The five copy number differences showing statistical significance in our ANCOVA analyses do nevertheless deserve further investigation. Effect sizes were small but the transport (LinJ.08.0700, LinJ.15.1240, LinJ.15.1250) and cytoskeletal (LinJ.29.1880, LinJ.29.1890) proteins involved carry out vital cell functions, variation in which has also been linked to drug resistance in previous research[482–484].

No coding region SNPs and only one coding region INDEL variant (affecting hypothetical protein LinJ.25.0280) was found to differ among Del and New World NonDel groups. As was the case with CNV, the statistical association of this 15 base-pair inframe deletion on chromosome 25 was driven by common descent, not by the presence of deletion on chr31. This INDEL variant did not occur in NonDel isolates from Mato Grosso do Sul but appeared fixed across the Del + nested NonDel clade and thus most likely represents a mutation that arose soon after the major population subdivision that defines this dataset began.

This major subdivision of samples from Mato Grosso do Sul relative to all others analyzed in the study raises the question as to whether two separate events could have introduced *L. infantum* into Brazil, possibly one arriving via Spanish territories to the West, the other arriving at Portuguese ports in the Northeast of the continent. In our pattern-process modelling, however, the highest ranked and most parsimonious model proposes divergence from a single introduction event, with distinct bottleneck intensities in the diverging populations explaining different levels of nucleotide diversity and genetic distance to Old World isolates. An abundance of fixed polymorphisms shared between these divergent New

World groups, but unfixed in Old World isolates, further supports the model. To fully confirm the occurrence of a single introduction, however, further *L. infantum* sampling from Iberia, coastal Africa, western Brazil and its neighboring states is likely required. It is also possible that Old World *L. infantum* infections were contracted in trading or departure areas shared among distinct colonizing groups, such that even high spatial sample effort could fail to distinguish a single or separate introduction events. Another possibility is that New World isolates originate from Old World regions where the disease has since been eradicated. In such cases, spatially explicit (e.g., landscape genetic simulation) modelling methods[80] within the New World could become useful, e.g., by testing for differences in dispersal directionality between parasite populations in western vs. eastern Brazil. Another fruitful approach might consist in assessing epidemiological phenotypes in the divergent subpopulation from Mato Grosso do Sul. If these NonDel genotypes do not represent a distinct introduction source, perhaps they have diverged so markedly due to unique selection pressures in this part of Brazil. Previous microsatellite-based studies which also detected strong divergence in Mato Grosso do Sul parasites hypothesized that the presence of an alternative VL vector, *Lu. cruzi*, might substantially modify *L. infantum* genetic diversity in the region, especially near Corumbá[18,248]. Comparing *Lu. cruzi* infection and transmission success by local NonDel vs. other (Del) *L. infantum* genotypes are interesting next steps.

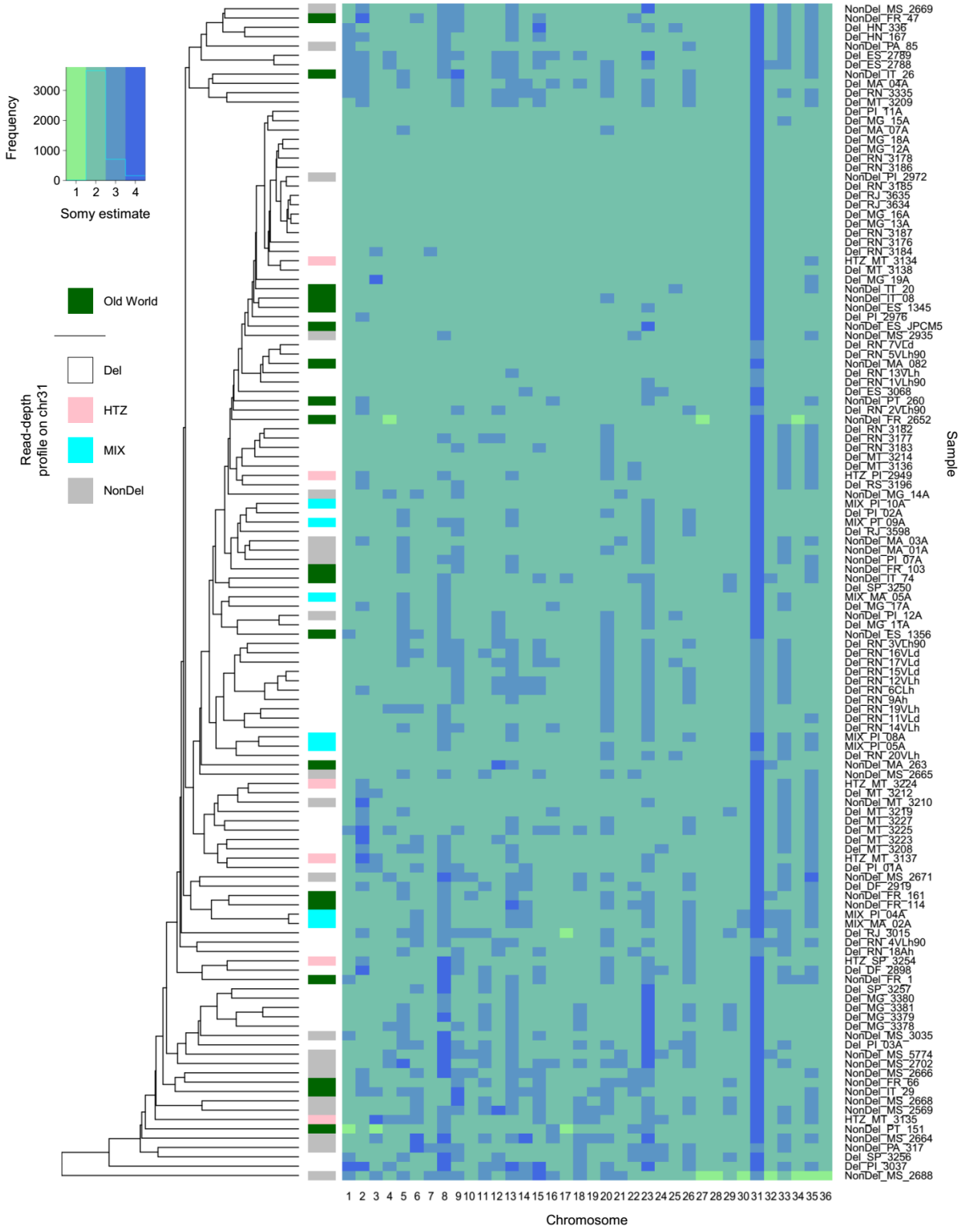
Finally, this study also presents the extraordinary finding that mating events are occurring abundantly in areas of contact between divergent *L. infantum* isolates, specifically in the city of Rondonópolis, located along the interstate highway between Cuiabá (Mato Grosso) and Campo Grande (Mato Grosso do Sul). The evidence for hybridization here is unmistakable: highly heterozygous genomes with half-deletion profiles on chr31 and phased haplotypes that divide between putative, significantly less heterozygous parental groups. These HTZ isolates occupy long-branching outgroup positions in phylogenetic trees and present intermediate PCoA axis values relative to Del isolates from Mato Grosso and NonDel isolates from Mato Grosso do Sul. They do not show higher aneuploidy rates, though cytometric measurement of total DNA content was not performed.

The particular case of hybridization we present here provides a compelling example of the evolutionary importance of genetic exchange between divergent lineages. For a strictly clonal parasite, the homozygous deletion mutation central to this study would be absolutely irreversible: zero alleles remain at the locus in Del isolates. Through hybridization, however, even these completely deleted alleles can reemerge – and with immediate effect on offspring phenotype: we report a return to function in ecto-3'-nucleotidase, a potential virulence factor, and more importantly, changes to miltefosine sensitivity are directly implied[258,259].

Our results closely mirror recent work by Cotton et. al (2019) [246], one of only two studies to analyze putative hybrid field isolates using WGS. Similar to our hypotheses on second-generation hybrids (e.g., NonDel_MT_3210 and Del_MT_3223) occurring beside F_1-like crosses such as HTZ_MT_3134, Cotton et al. describe asymmetrical ancestry components in *L. donovani* that point to intercrossing and backcrossing subsequent to a 'founding' hybridization event. Like theirs from Ethiopia, our dataset also appears to contain parental genotypes and covers areas where distinct vector and parasite populations are prone to connect. In contrast to the divergent crossing events our study highlights from such areas, evidence for less conspicuous, endogamic forms of mating is only weak. Moderate homozygosity deficits (median F_{IS} = -0.251) in Rio Grande do Norte, our largest sympatric group, do raise the possibility that inbreeding is slowing heterozygosity accumulation in Brazilian *L. infantum* genomes, but this process is unlikely occurring as frequently as in other species of the Viannia subgenus[22,383,485]. We should not discard the possibility of cryptic mating and its potential to mislead phylogenetic inference, including signs of Del paraphyly observed in Fig. 3.4. Further sampling is therefore necessary to confirm relationships between Del and closely-related NonDel isolates from, e.g., Piauí and Maranhão. NonDel isolates from these two states do not diverge strongly from the surrounding clade (Supplementary Fig. 3.7), including at 'Del-distinctive' sites (Supplementary Fig. 3.9), but is it nevertheless possible that their nested positions (see dotted circle in Fig. 3.4) are a consequence of cryptic backcrossing events?

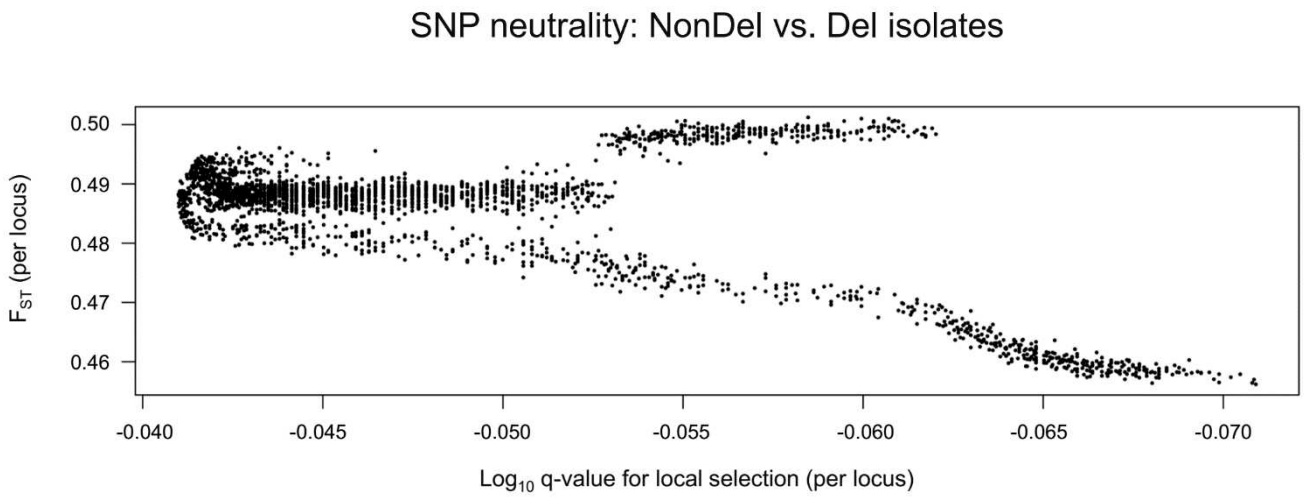
Many open questions remain regarding the chr31 deletion anomaly, likewise about the precise Old World origins of complex *L. infantum* genetic diversity in Brazil. Taken together, this study clearly demonstrates how much *L. infantum* research can learn from scrutinizing the country's underappreciated parasite diversity and the demographic processes that contribute to strong population structure and hybridization at divergent contact zones. Recognizing strong, yet changeable population structure is critical to VL control. Left ignored or unobserved, it can confound covariation measured between parasite genetic and phenotypic traits or lead to failure in the application of diagnostics and drugs.

3.6 Supplementary figures and tables

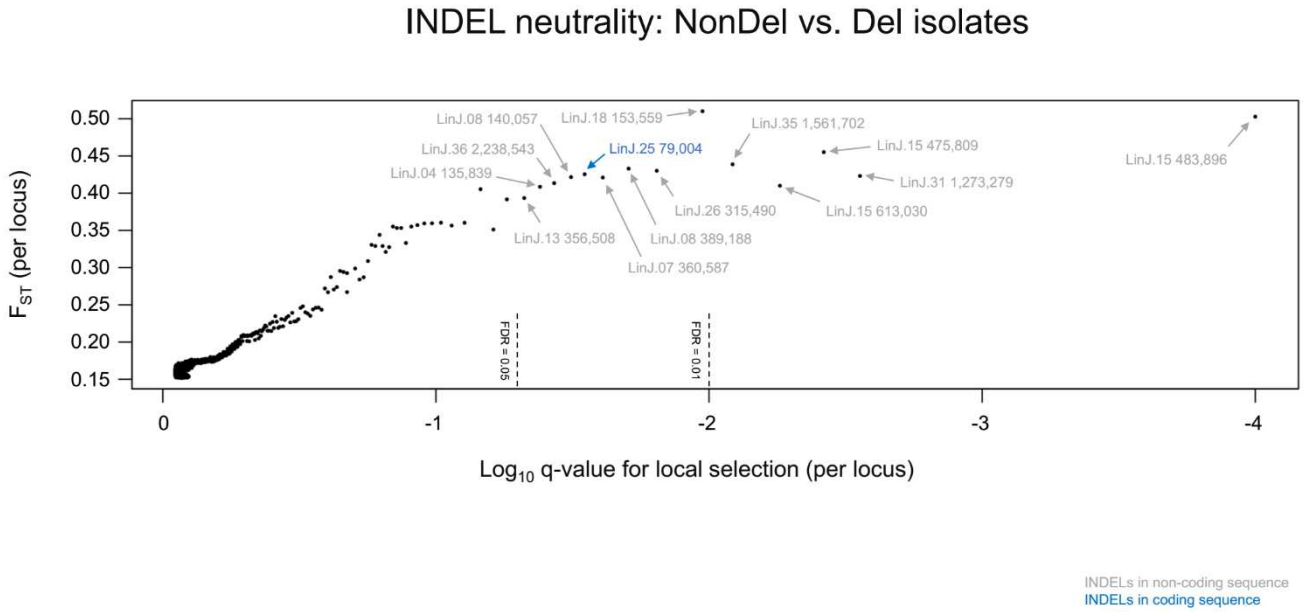


Supplementary Figure 3.1 Chromosomal copy number variation in New and Old World *L. infantum* isolates. To estimate chromosomal copy numbers for each sample, we calculated mean read-depths for successive 1 kb windows on each chromosome. We then calculated the median of these window means on each chromosome and let the 40th percentile (p40) of the sample's 36 chromosomal medians represent expectations for the disomic state. Somy estimates for each chromosome by median normalization to p40 are plotted in the heatmap. Isolates are ordered on the y-axis by UPGMA clustering of Bray-Curtis dissimilarities. The adjacent column indicates read-depth profiles on (tetrasomic) chr31 according to the color key at left. No correlation to somy is observed.

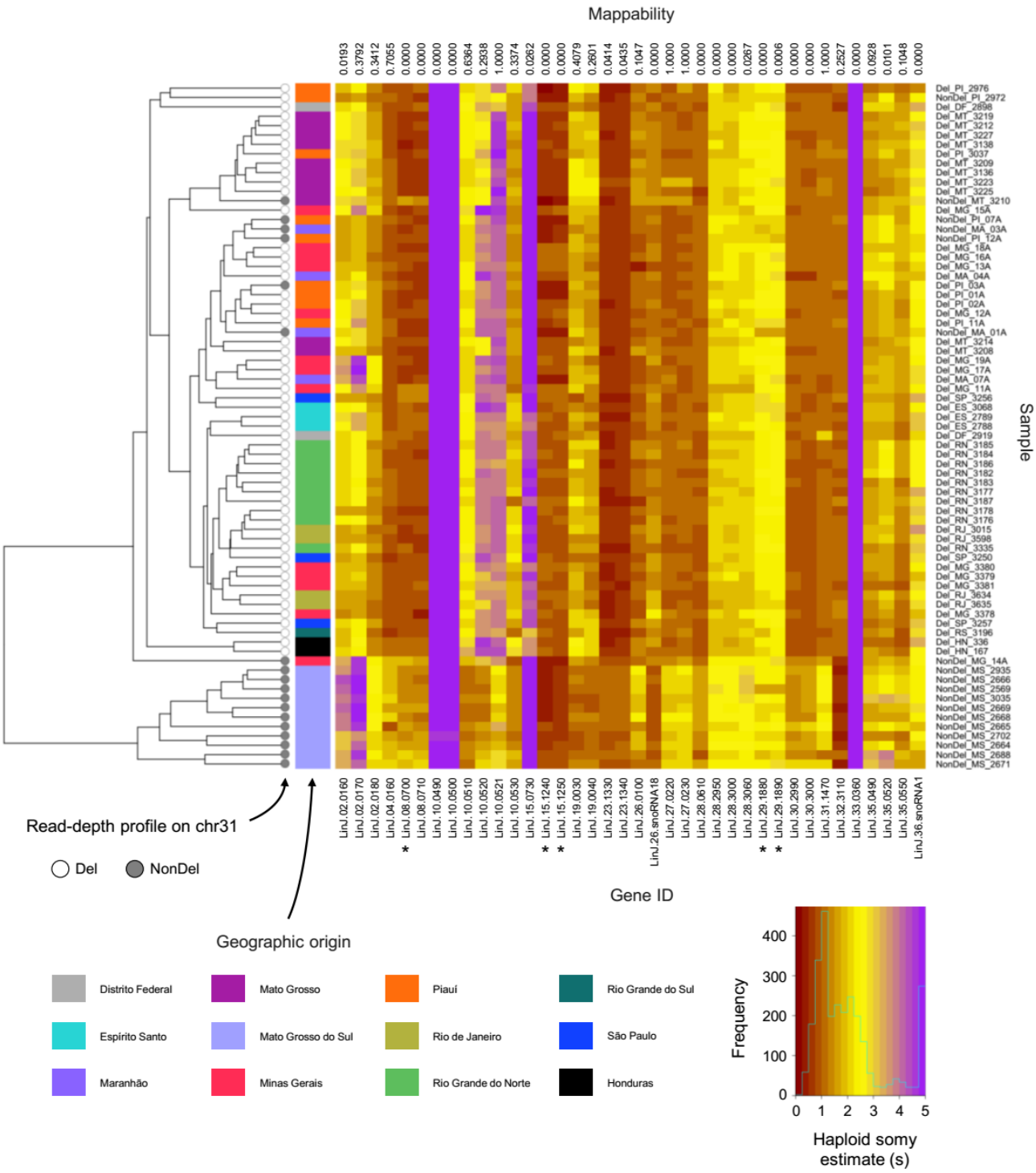
a



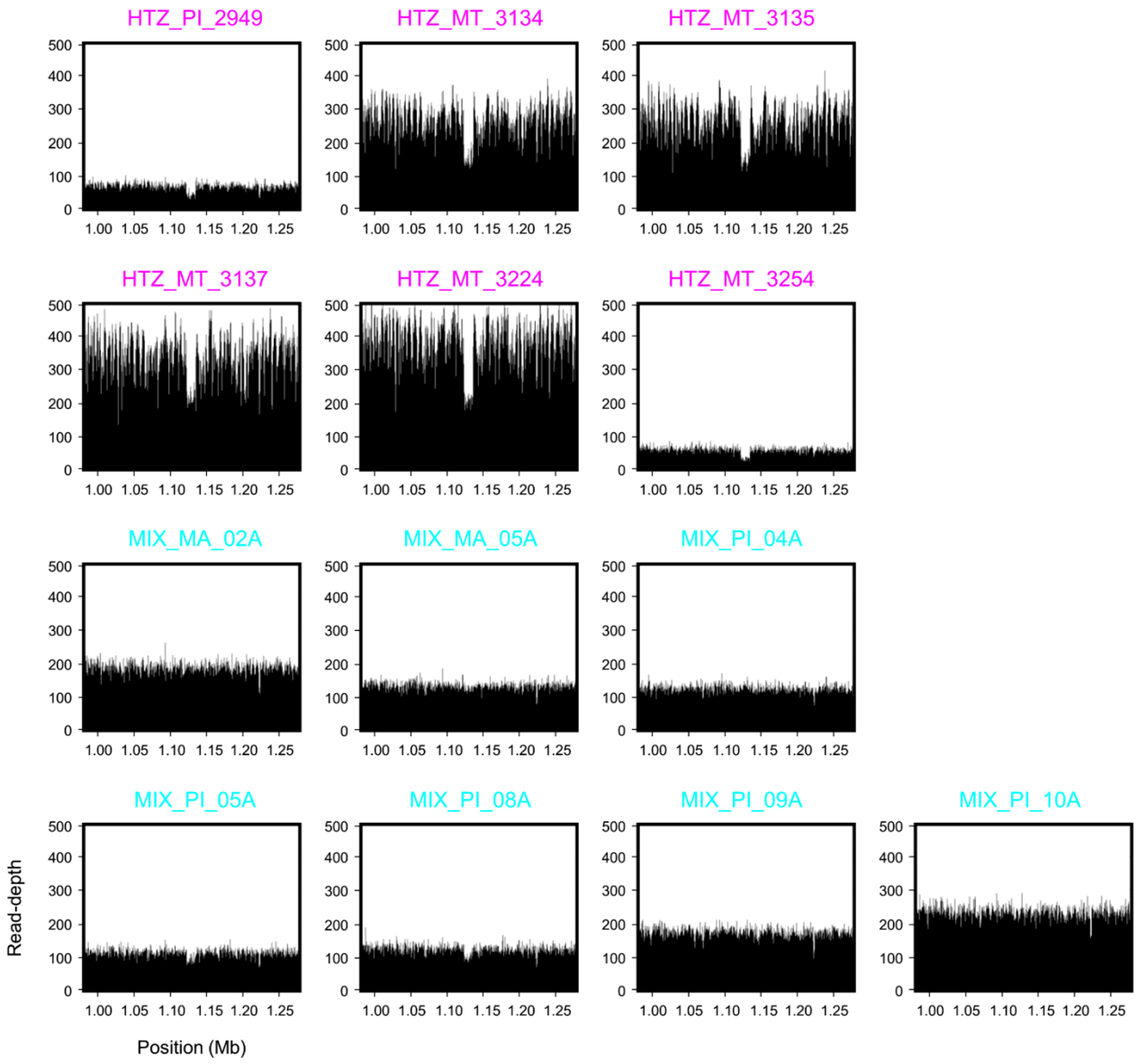
b



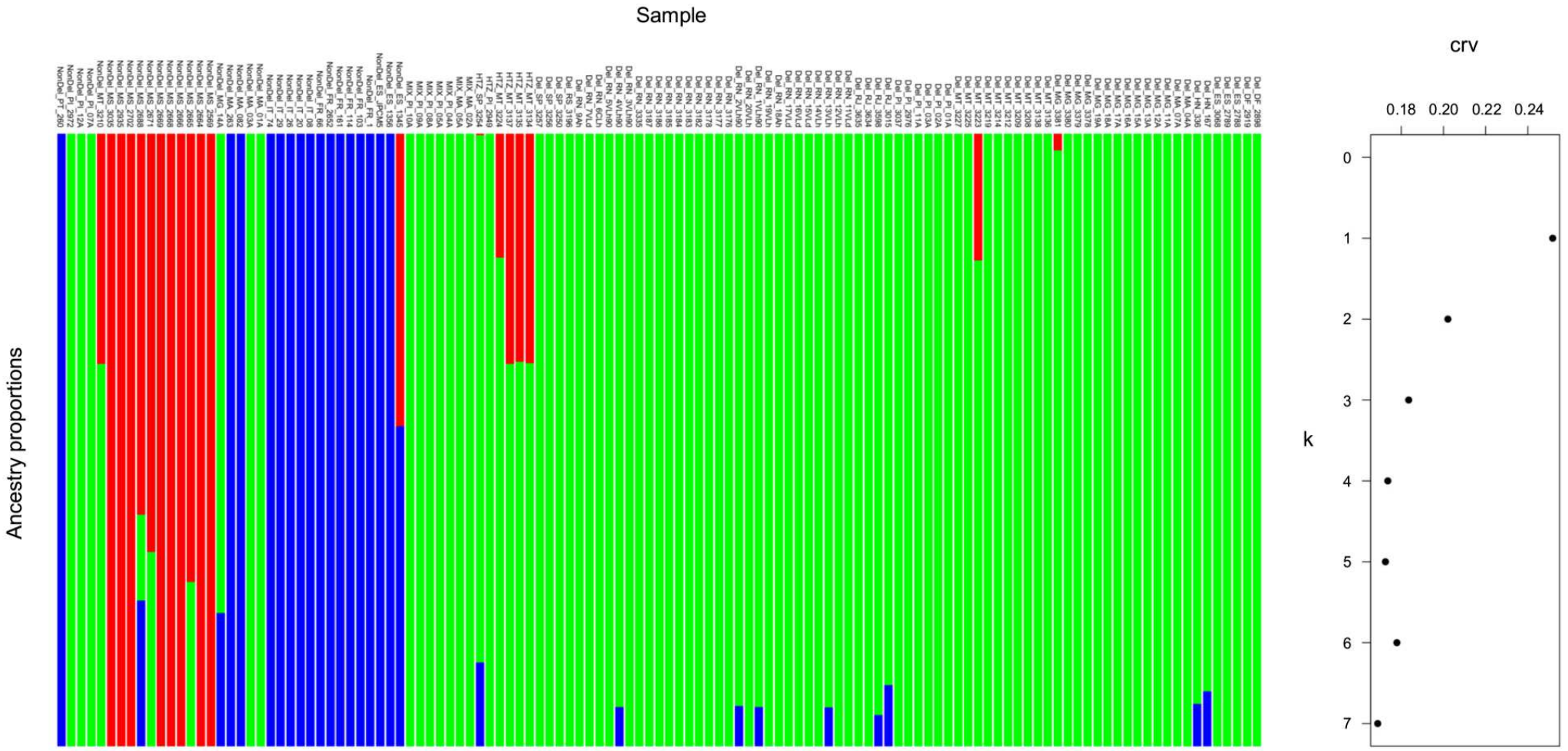
Supplementary Figure 3.2 Low inter-locus variance in F_{ST} differentiation between Del and NonDel *L. infantum* isolates. **a** F_{ST} values at genome-wide SNP loci range between 0.456 and 0.501. Log_{10} q-values (x-axis) indicate the level of support for selection at each locus. None are significant based on a false discovery rate (FDR) of 5%. SNP sites with genotypes missing in > 50% individuals are excluded from analysis. **b** F_{ST} values at genome-wide INDEL loci range between 0.152 and 0.511. Fourteen outlier loci show significant support for selection at FDR = 5%. Only one of these outliers occurs within coding sequence (blue font) and represents a disruptive inframe deletion in LinJ.25.280. This gene encodes a protein of unknown function on chromosome 25. INDEL sites with genotypes missing in > 50% individuals are excluded from analysis.



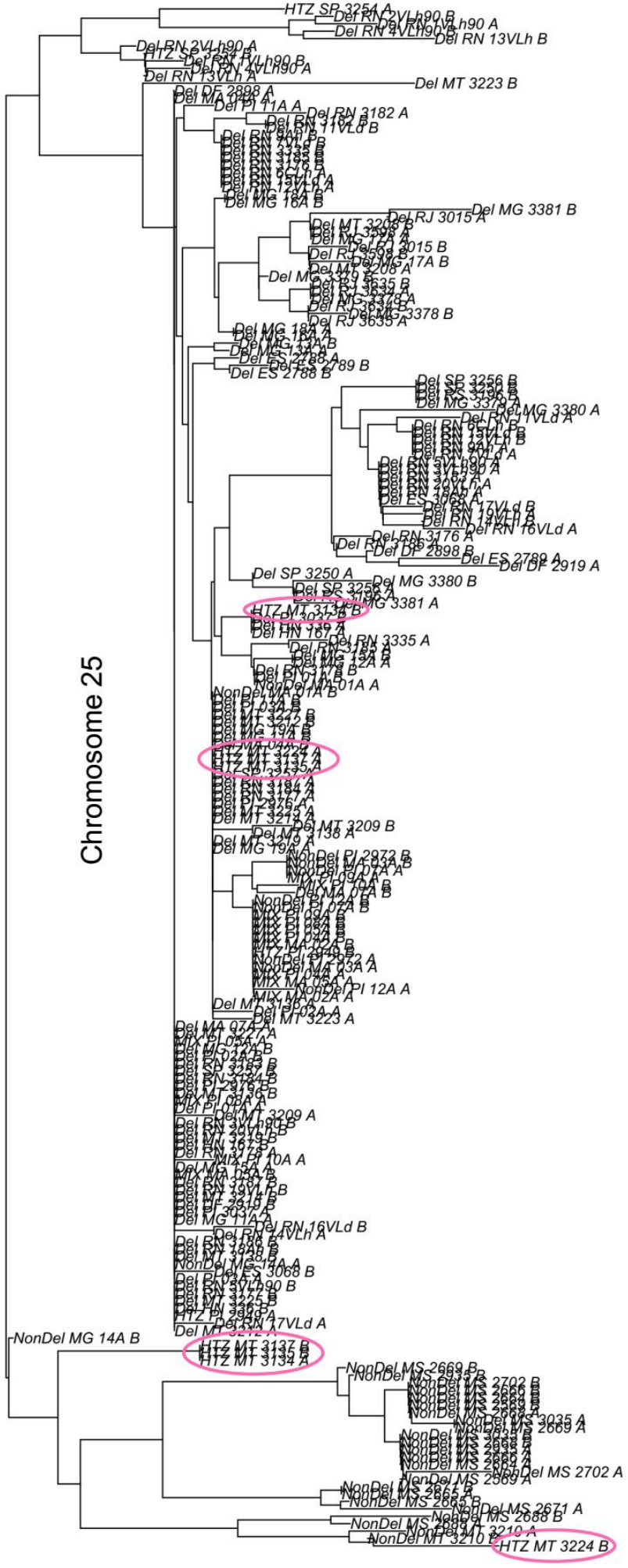
Supplementary Figure 3.3 Gene copy number variation in Del and NonDel *L. infantum* isolates from the New World. We obtained haploid somy estimates (s) by dividing each coding region's median read-depth (c) by the median of all c values across the chromosome. The heatmap plots s values for all coding regions that differed significantly between Del and NonDel isolates in the New World (see Mann-Whitney U statistics in Supplementary Tbl. 3.5). Isolates are ordered on the y-axis by UPGMA clustering of Bray-Curtis dissimilarities. Circles at the tips of the tree indicate read-depth profiles on chr31. The adjacent column indicates geographic origin according to the color key below the heatmap. Somy profiles cluster predominantly by geographic origin and only indirectly by chr31 read-depth profile. Only five coding regions differ significantly between Del and NonDel groups after controlling for geographic origin by analysis of covariance (see asterisks). Numbers above the heatmap columns indicate the proportion of uniquely mapping nucleotides within each coding region (see Methods). Poor mappability represents an intrinsic property of genes occurring in multiple paralogs and may explain instances where s > 4 occurs in many samples (i.e., purple columns). Isolates from Teixeira et al. (2017)[257] (see Supplementary Tbl. 3.1) were excluded from gene copy number analyses because reverse reads were not made available in public sequence archives.



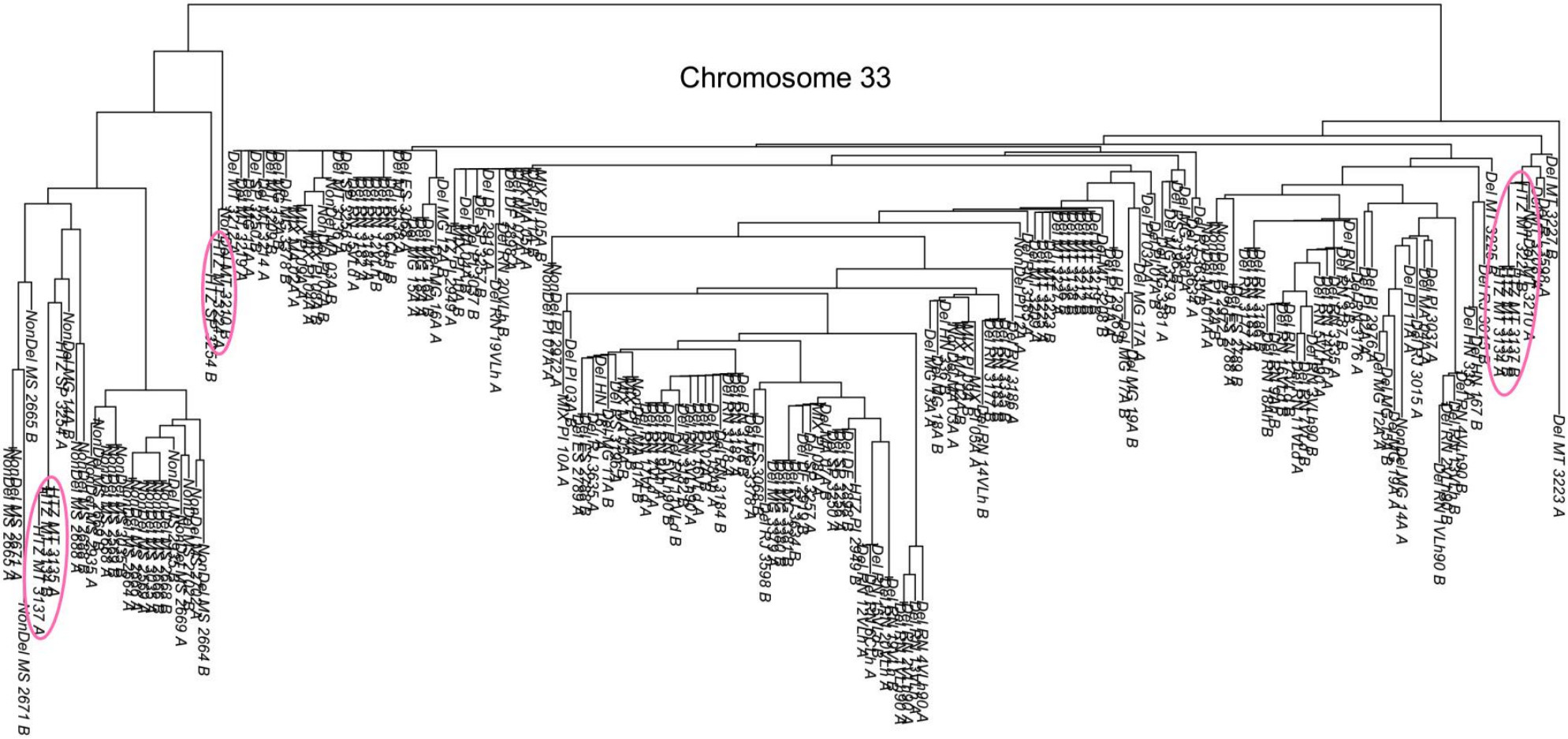
Supplementary Figure 3.4 Sequence read-depth profiles on chr31 in MIX and HTZ *L. infantum* isolates. Read-depth drops to ca. 50% between 1.122 Mb and 1.135 Mb in all six HTZ isolates. Quantitative PCR confirmed partial deletion at this locus in HTZ cultures derived from single cells. MIX isolates, on the other hand, appear to contain a mixture of NonDel and Del or HTZ profiles based on subclone PCR by Carnielli et al. (2018)[258]. NonDel cells likely predominate in these mixtures given that full read-depth occurs in all but MIX_PI_05A and MIX_PI_08A. Del and HTZ cells may occur more frequently in the latter two isolates.



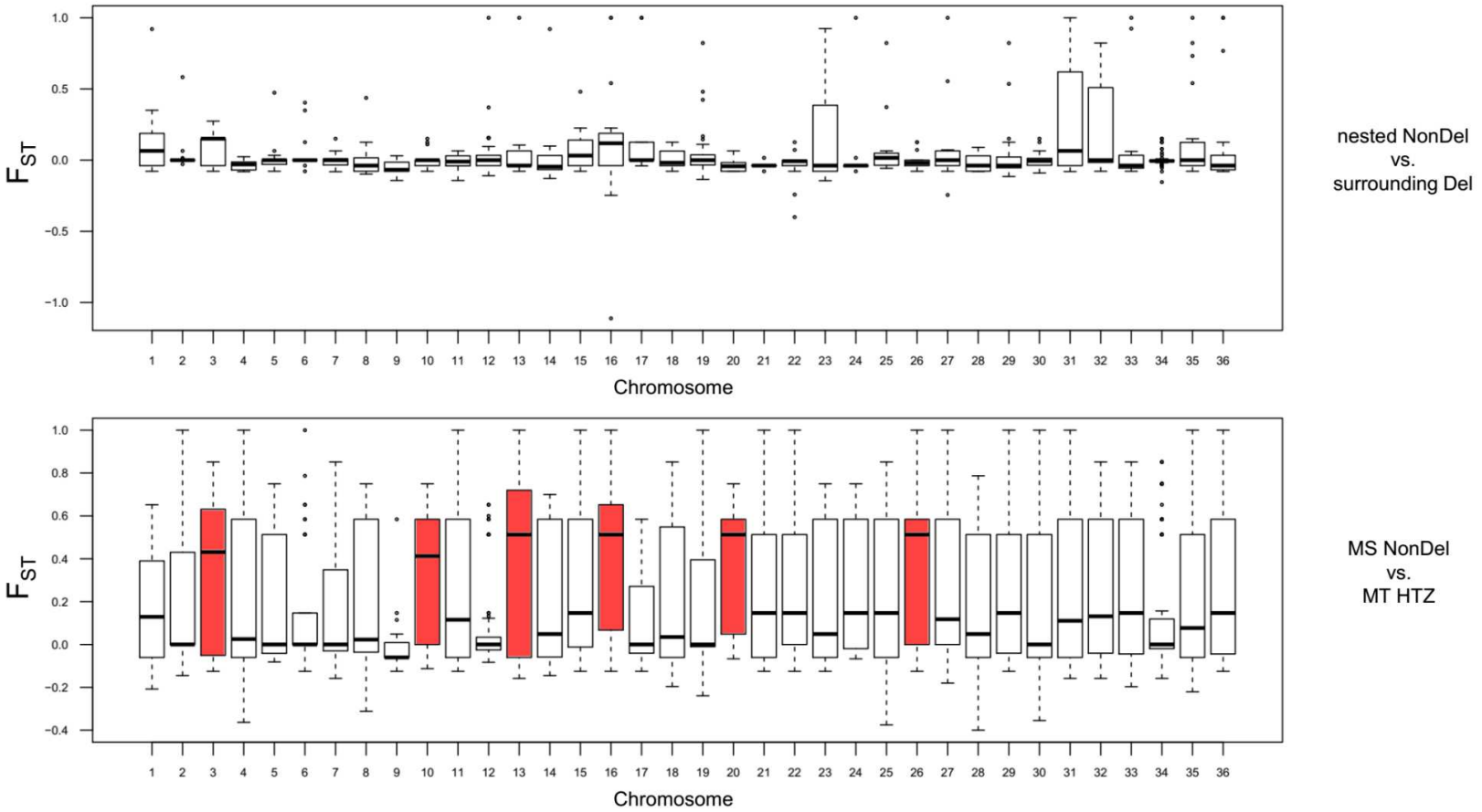
Supplementary Figure 3.5 Estimated ancestry proportions in New and Old World *L. infantum* isolates. The ADMIXTURE program[6] estimates individual ancestry proportions by optimizing likelihood in the parametric admixture model introduced by Pritchard et al. (2000)[332]. In this model, the number of subpopulations (k) is defined *a priori* and alleles within each individual represent binomial samples from the allele frequencies specific to each subpopulation. Individual ancestry proportions and subpopulation allele frequencies are jointly estimated during model optimization to the genotypes observed. The bar plot summarizes ADMIXTURE results for the *L. infantum* SNP dataset at k = 3. We chose k = 3 based on complementary PCoA analyses (see subsequent figures and text) and because cross-validation error (crv) is low without overfitting the data (right plot). Each column represents one *L. infantum* isolate and relative fill color quantities indicate estimated ancestry proportions. Blue predominates in isolates from the Old World while green and red correspond to New World subpopulations. Several isolates from Mato Grosso and Mato Grosso do Sul show substantial quantities of both red and green, symptomatic of admixture between dissimilar subpopulations in these adjacent states of Brazil. Analyses used all SNP loci for which genotypes are called for all individuals.



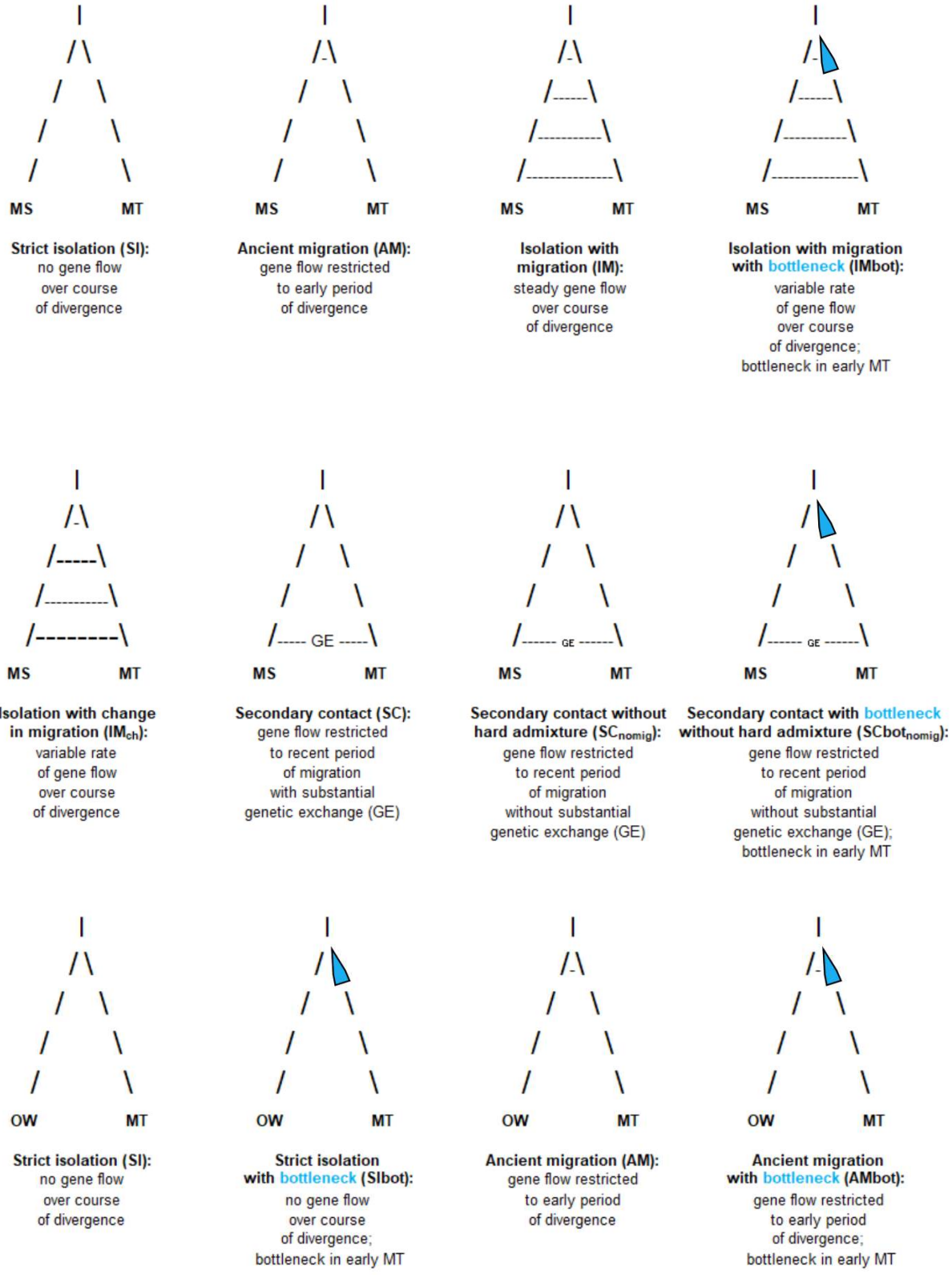
Supplementary Figure 3.6 (continues on next page)



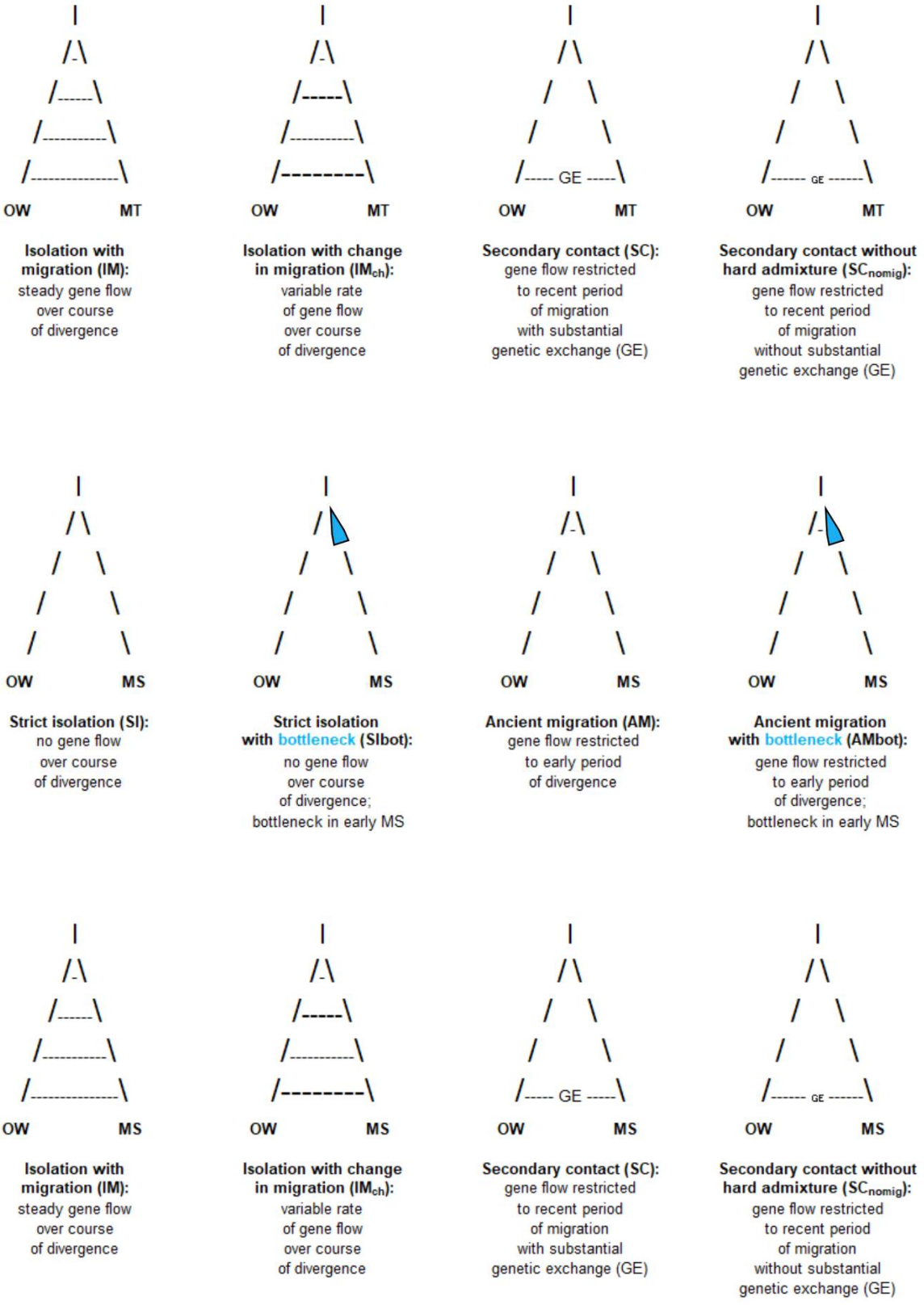
Supplementary Figure 3.6 Neighbor-joining trees built from phased chromosomes suggest that heterozygous loci in HTZ isolates originate from genetic exchange between divergent haplotypes. Examples are shown from chromosomes 25 and 33. Each tree contains two phased haplotypes per isolate (see label suffixes 'A' and 'B'). Both A and B haplotypes of NonDel isolates from Mato Grosso do Sul (MS) occur within a clade that splits away basally from that containing most other haplotypes. HTZ isolates from the neighboring state of Mato Grosso (MT) often show one haplotype clustering towards this divergent MS clade and the other haplotype showing similarity to Del haplotypes found in MT and other parts of Brazil (see pink circles). A similar trend is observed for isolates such as Del_MT_3223 and NonDel_MT_3210, possibly the result of hybridization with unsampled lineages or secondary hybridizations involving progenitors of this study's sample set.



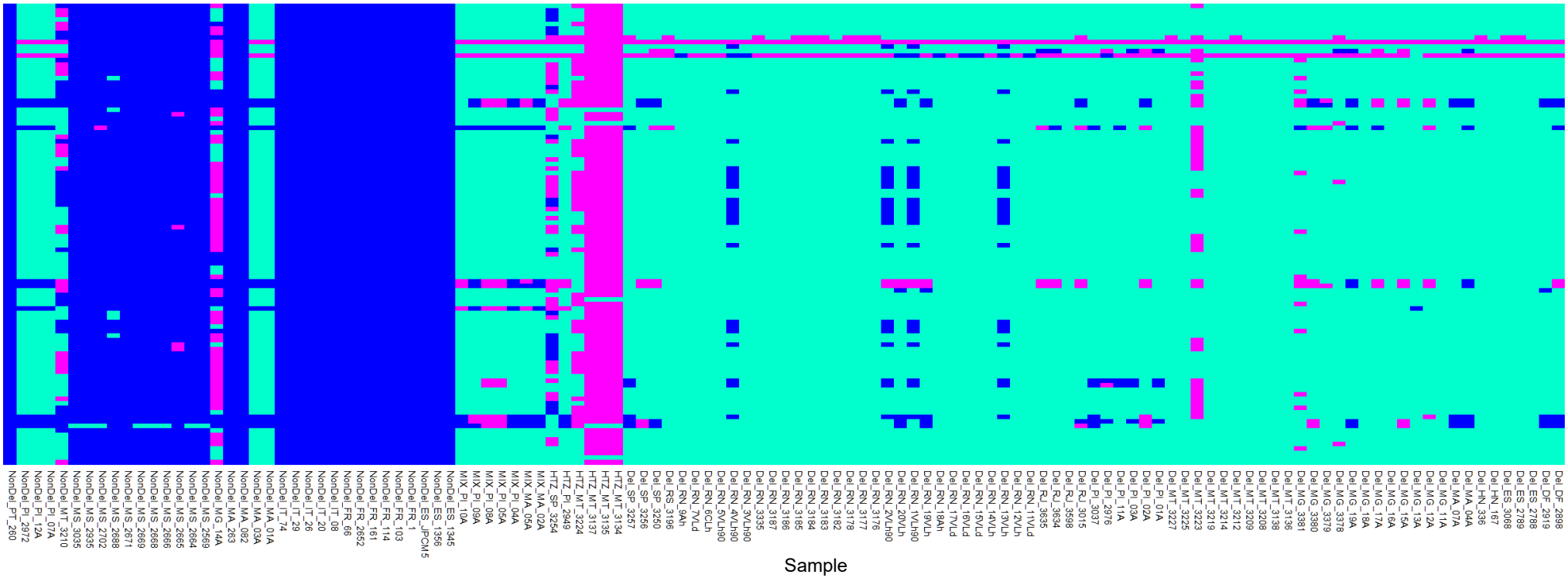
Supplementary Figure 3.7 NonDel *L. infantum* isolates from Piauí and Maranhão show low divergence to Del isolates on all chromosomes. The top panel shows F_{ST} (Weir and Cockerham) for NonDel_MA_01A, NonDel_MA_03A, NonDel_PI_07A, NonDel_PI_12A and NonDel_PI_2972 relative to Del isolates that surround this nested NonDel group in the phylogenetic tree provided in Fig. 3.4. These are Del_MA_04A, Del_MA_07A, Del_MG_19A, Del_PI_02A, Del_PI_11A, Del_PI_3037, Del_DF_2898, Del_DF_2919 and Del_SP_3257. Boxplots indicate low F_{ST} medians (bold horizontal bars) on all chromosomes, inconsistent with the hypothesis that NonDel isolates observed in Piauí and Maranhão represent backcrossed hybrid genotypes. Patterns of F_{ST} for putative hybrids (HTZs) from Mato Grosso (MT) relative to NonDel isolates from Mato Grosso do Sul (MS) are distinct. Values are less stable among chromosomes and several medians exceed 0.4 (see red fill).



Supplementary Figure 3.8 (continues on next page)



Supplementary Figure 3.8 Ten scenarios of pairwise divergence simulated using fastsimcoal2[315]. The first set of scenarios depicts divergence between *L. infantum* populations from Mato Grosso (MT) and Mato Grosso do Sul (MS). The second set depicts divergence between populations from MT and the Old World (OW). The third set depicts divergence between populations from MS and OW. Corresponding input syntax is provided in Supplementary Tbl. 3.7. Dashes horizontal lines indicate gene flow. Blue shapes indicate bottleneck events. Time runs from top (ancestral events) to bottom (present).



Supplementary Figure 3.9 Visualization of biallelic single-nucleotide polymorphisms (SNPs) prevalent (> 80%) in Del but uncommon (< 50%) in New World NonDel *L. infantum* isolates. These 102 'Del-distinctive' variants are listed in ascending order, i.e., the top rows representing SNP sites on chromosome 2 and the bottom rows representing sites on chromosome 36. Each column represents the concatenated genotypes of one *L. infantum* isolate. Heterozygous (0/1) genotypes are colored in pink, homozygous reference (0/0) genotypes in blue and homozygous non-reference (1/1) genotypes in green. This format helps visualize patterns of fixed (perhaps ancestral) vs. non-fixed (perhaps convergently evolving) sequence variation in New World *L. infantum* groups. It also emphasizes continuous genome-wide heterozygosity in HTZ isolates.

Supplementary Table 3.1 *L. infantum* isolates analyzed by whole-genome sequencing or polymerase chain reaction product electrophoresis. Abbreviations: IOCL (Coleção de Leishmania do Instituto Oswaldo Cruz); NRD (average nuclear read-depth); PCR (polymerase chain reaction testing for presence/absence of the chr31 deletion locus); WGS (whole-genome sequencing); ENA (European Nucleotide Archive); NA (not applicable); ND (not determined).

ID	IOCL code	International code	Geographic origin	Deletion?	Data type	NRD	WGS source
NonDel_BO_3756	3756	MCAN/BO/2017/PQ102P	Puerto Quijarro, Bolivia	NO	PCR	NA	NA
Del_DF_2898	2898	MHOM_BR_2006_NMT-HUB402982MO	Distrito Federal, Brazil	YES	WGS	37.3	ENA
Del_BA_579	579	MHOM/BR/1974/PP75	Bahia, Brazil	YES	PCR	NA	NA
Del_DF_2919	2919	MCAN_BR_2005_NMT-DF544MO	Distrito Federal, Brazil	YES	WGS	23.4	ENA
Del_ES_2788	2788	MHOM_BR_2005_DRD	Espírito Santo, Brazil	YES	WGS	47.8	This study
Del_ES_2789	2789	MHOM_BR_2005_HRNS-1	Espírito Santo, Brazil	YES	WGS	50.5	This study
Del_ES_3068	3068	MCAN_BR_2008_CP-18	Espírito Santo, Brazil	YES	WGS	48.2	This study
Del_HN_167	NA	MHOM_HN_1989_167	Orocuina, Honduras	YES	WGS	32.7	ENA
Del_HN_336	NA	MHOM_HN_1993_336	San Juan Bautista, Honduras	YES	WGS	51.8	ENA
Del_PE_2647	2647	MHOM/BR/2003/ACS	Pernambuco, Brazil	YES	PCR	NA	NA
Del_PE_3434	3434	MHOM/BR/2012/EJC	Pernambuco, Brazil	YES	PCR	NA	NA
Del_PE_2769	2769	MCAN/BR/2005/SPACK	Pernambuco, Brazil	YES	PCR	NA	NA
Del_PE_3053	3053	MHOM/BR/2008/RJS	Pernambuco, Brazil	YES	PCR	NA	NA
Del_PE_2933	2933	MCAN/BR/2006/MAIKE	Pernambuco, Brazil	YES	PCR	NA	NA
Del_MA_04A	NA	MHOM_BR_06_MA04A	Maranhão, Brazil	YES	WGS	65.4	Carnielli et al. (2018)
Del_MA_07A	NA	MHOM_BR_06_MA07A	Maranhão, Brazil	YES	WGS	36.1	Carnielli et al. (2018)
Del_MG_11A	NA	MHOM_BR_05_MG11A	Minas Gerais, Brazil	YES	WGS	113.1	Carnielli et al. (2018)
Del_MG_12A	NA	MHOM_BR_05_MG12A	Minas Gerais, Brazil	YES	WGS	62.5	Carnielli et al. (2018)
NonDel_PI_3680	3680	MHOM/BR/2006/6909	Piauí, Brazil	NO	PCR	NA	NA
NonDel_PI_3681	3681	MHOM/BR/2006/6912	Piauí, Brazil	NO	PCR	NA	NA
NonDel_PI_3682	3682	MHOM/BR/2006/P112A 4P NMV	Piauí, Brazil	NO	PCR	NA	NA
NonDel_PI_3683	3683	MHOM/BR/2006/JSMG 3P	Piauí, Brazil	NO	PCR	NA	NA
Del_PI_3684	3684	MHOM/BR/2006/ARS 3P	Piauí, Brazil	YES	PCR	NA	NA
NonDel_PI_3685	3685	MHOM/BR/2006/MA03-A 4P (GMS)	Piauí, Brazil	NO	PCR	NA	NA
Del_PI_3684	3686	MHOM/BR/2006/MA04A 4P JNN	Piauí, Brazil	YES	PCR	NA	NA
NonDel_PI_3687	3687	MHOM/BR/2006/FFS 3P PI-05	Piauí, Brazil	NO	PCR	NA	NA
NonDel_PI_3688	3688	MHOM/BR/2006/P109A 4P JAS	Piauí, Brazil	NO	PCR	NA	NA
NonDel_PI_3692	3692	MHOM/BR/2006/891	Piauí, Brazil	NO	PCR	NA	NA

Supplementary Table 3.1 (continued)

NonDel_PI_3693	3693	MHOM/BR/2006/930	Piauí, Brazil	NO	PCR	NA	NA
NonDel_PI_3694	3694	MHOM/BR/2006/890	Piauí, Brazil	NO	PCR	NA	NA
Del_PI_3695	3695	MHOM/BR/2006/867	Piauí, Brazil	YES	PCR	NA	NA
NonDel_PI_3696	3696	MHOM/BR/2006/6889	Piauí, Brazil	NO	PCR	NA	NA
NonDel_PI_3697	3697	MHOM/BR/2006/6893	Piauí, Brazil	NO	PCR	NA	NA
Del_PI_3698	3698	MHOM/BR/2006/791	Piauí, Brazil	YES	PCR	NA	NA
NonDel_PI_3699	3699	MHOM/BR/2006/6905	Piauí, Brazil	NO	PCR	NA	NA
NonDel_PI_3700	3700	MHOM/BR/2006/6914	Piauí, Brazil	NO	PCR	NA	NA
NonDel_PI_3702	3702	MHOM/BR/2006/Pi 04A 4P DHRi	Piauí, Brazil	NO	PCR	NA	NA
Del_PI_3703	3703	MHOM/BR/2006/AAS 3P	Piauí, Brazil	YES	PCR	NA	NA
NonDel_PI_3704	3704	MHOM/BR/2006/AAS 4P	Piauí, Brazil	NO	PCR	NA	NA
Del_PI_3705	3705	MHOM/BR/2006/Pi 11A 4P MBS	Piauí, Brazil	YES	PCR	NA	NA
NonDel_PI_3707	3707	MHOM/BR/2006/MA01-A 4P	Piauí, Brazil	NO	PCR	NA	NA
Del_MG_13A	NA	MHOM_BR_05_MG13A	Minas Gerais, Brazil	YES	WGS	69.3	Carnielli et al. (2018)
Del_MG_15A	NA	MHOM_BR_05_MG15A	Minas Gerais, Brazil	YES	WGS	100.9	Carnielli et al. (2018)
Del_MG_16A	NA	MHOM_BR_05_MG16A	Minas Gerais, Brazil	YES	WGS	137.5	Carnielli et al. (2018)
Del_MG_17A	NA	MHOM_BR_05_MG17A	Minas Gerais, Brazil	YES	WGS	97.5	Carnielli et al. (2018)
Del_MG_18A	NA	MHOM_BR_05_MG18A	Minas Gerais, Brazil	YES	WGS	66.3	Carnielli et al. (2018)
Del_MG_19A	NA	MHOM_BR_05_MG19A	Minas Gerais, Brazil	YES	WGS	80.4	Carnielli et al. (2018)
Del_MG_3378	3378	MCAN_BR_2010_CA1	Minas Gerais, Brazil	YES	WGS	23.6	This study
Del_MG_3379	3379	MCAN_BR_2010_CA2	Minas Gerais, Brazil	YES	WGS	30.4	This study
Del_MG_3380	3380	MCAN_BR_2010_CA3	Minas Gerais, Brazil	YES	WGS	34.5	This study
Del_MG_3381	3381	MCAN_BR_2010_CA4	Minas Gerais, Brazil	YES	WGS	21.0	This study
Del_MT_3136	3136	MCAN_BR_2009_PATETA	Mato Grosso, Brazil	YES	WGS	129.8	This study
Del_SC_3716	3716	MCAN/BR/2010/IRIS	Santa Catarina, Brazil	YES	PCR	163.3	This study
Del_SC_3717	3717	MCAN/BR/2014/BOB	Santa Catarina, Brazil	YES	PCR	135.4	This study
Del_SC_3718	3718	MCAN/BR/2015/SNOOPY	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3719	3719	MCAN/BR/2015/PACO	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3733	3733	MCAN/BR/2015/LAIKA	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3720	3720	MCAN/BR/2014/FAISCA	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3721	3721	MCAN/BR/2010/CHOCOLATE	Santa Catarina, Brazil	YES	PCR	NA	NA

Supplementary Table 3.1 (continued)

Del_SC_3722	3722	MCAN/BR/2010/LOBA	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3724	3724	MCAN/BR/2014/NINNA	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3725	3725	MCAN/BR/2015/TWISTY	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3726	3726	MCAN/BR/2014/PINGO 2	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3734	3734	MCAN/BR/2015/LUMA	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3727	3727	MCAN/BR/2015/KIARA 2	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3728	3728	MHOM/BR/2017/PC	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3729	3729	MCAN/BR/2016/PRETO	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3730	3730	MCAN/BR/2014/DORA	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3731	3731	MCAN/BR/2010/ZEUS	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3735	3735	MCAN/BR/2010/PONGO	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3736	3736	MCAN/BR/2014/AMIGO	Santa Catarina, Brazil	YES	PCR	NA	NA
NonDel_SC_3737	3737	MCAN/BR/2017/BETHOVEN	Santa Catarina, Brazil	NO	PCR	NA	NA
Del_SC_3738	3738	MCAN/BR/2016/NICK	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3739	3739	MCAN/BR/2015/BOMBOM	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3740	3740	MCAN/BR/2010/BAROLO	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3741	3741	MCAN/BR/2015/SCOOBY	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3742	3742	MCAN/BR/2010/CÃO	Santa Catarina, Brazil	YES	PCR	NA	NA
NonDel_SC_3743	3743	MCAN/BR/2010/CHANEL	Santa Catarina, Brazil	NO	PCR	NA	NA
Del_SC_3744	3744	MCAN/BR/2015/BONECA	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3750	3750	MCAN/BR/2015/GORKIM	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3751	3751	MCAN/BR/2015/HANNA	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3752	3752	MCAN/BR/2017/JUJU	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3753	3753	MCAN/BR/2015/MAGRELO	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3754	3754	MCAN/BR/2016/BOLA	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SC_3755	3755	MCAN/BR/2015/NICO	Santa Catarina, Brazil	YES	PCR	NA	NA
Del_SE_3595	3595	MHOM/BR/2014/LVH5	Sergipe, Brazil	YES	PCR	NA	NA
Del_AM_3118	3118	MHOM/BR/2009/BLVD	Amazonas, Brazil	YES	PCR	NA	NA
Del_MT_3138	3138	MCAN_BR_2009_BRONCRIS	Mato Grosso, Brazil	YES	WGS	135.4	This study
Del_MT_3208	3208	MCAN_BR_2010_ZEUS	Mato Grosso, Brazil	YES	GENOME	194.1	This study
Del_MT_3209	3209	MCAN_BR_2010_ROBI	Mato Grosso, Brazil	YES	GENOME	166.9	This study

Supplementary Table 3.1 (continued)

Del_MT_3212	3212	MCAN_BR_2010_DIMY I	Mato Grosso, Brazil	YES	GENOME	186.9	This study
Del_MT_3214	3214	MCAN_BR_2010_MEG	Mato Grosso, Brazil	YES	GENOME	126.9	This study
Del_MT_3219	3219	MCAN_BR_2010_TITÃ I	Mato Grosso, Brazil	YES	GENOME	174.0	This study
Del_MT_3223	3223	MCAN_BR_2010_BOLINHA III	Mato Grosso, Brazil	YES	GENOME	166.9	This study
Del_MT_3225	3225	MCAN_BR_2010_CHITARA I	Mato Grosso, Brazil	YES	GENOME	180.1	This study
Del_MT_3227	3227	MCAN_BR_2010_MAGRÃO	Mato Grosso, Brazil	YES	GENOME	179.8	This study
Del_PI_02A	NA	MHOM/BR/06/PI02A	Piauí, Brazil	YES	WGS	89.5	Carnielli et al. (2018)
Del_PI_03A	NA	MHOM/BR/06/PI03A	Piauí, Brazil	YES	WGS	97.5	Carnielli et al. (2018)
Del_PI_01A	NA	MHOM/BR/06/PI01A	Piauí, Brazil	YES	WGS	152.0	Carnielli et al. (2018)
Del_PI_11A	NA	MHOM/BR/06/PI11A	Piauí, Brazil	YES	WGS	61.5	Carnielli et al. (2018)
Del_PI_2976	2976	MHOM_BR_2006_1406MBS	Piauí, Brazil	YES	WGS	5.6	Carnielli et al. (2018)
Del_PI_3037	3037	MCAN_BR_2007_LIBPI-50	Piauí, Brazil	YES	WGS	180.1	Carnielli et al. (2018)
Del_RJ_3015	3015	MHOM_BR_2007_WC	Rio de Janeiro, Brazil	YES	WGS	95.0	ENA
Del_RJ_3598	3598	MCAN_BR_2015_TUBO9	Rio de Janeiro, Brazil	YES	WGS	74.1	This study
Del_RJ_3634	3634	MCAN_BR_2016_90	Rio de Janeiro, Brazil	YES	WGS	29.0	This study
Del_MG_2906	2906	MHOM/BR/2002/LPC-RPV	Minas Gerais, Brazil	YES	PCR	NA	NA
Del_MG_3368	3368	MHOM/BR/2011/AD1	Minas Gerais, Brazil	YES	PCR	NA	NA
Del_MG_3370	3370	MHOM/BR/2011/CR1	Minas Gerais, Brazil	YES	PCR	NA	NA
Del_RJ_3635	3635	MCAN_BR_2016_89	Rio de Janeiro, Brazil	YES	WGS	22.3	This study
Del_RN_11VLd	NA	11VLd	Rio Grande do Norte, Brazil	YES	WGS	97.4	Teixeira et al. (2017)
Del_RN_12VLh	NA	12VLh	Rio Grande do Norte, Brazil	YES	WGS	97.2	Teixeira et al. (2017)
Del_RN_13VLh	NA	13VLh	Rio Grande do Norte, Brazil	YES	WGS	108.0	Teixeira et al. (2017)
Del_RN_14VLh	NA	14VLh	Rio Grande do Norte, Brazil	YES	WGS	105.2	Teixeira et al. (2017)
Del_RN_15VLd	NA	15VLd	Rio Grande do Norte, Brazil	YES	WGS	103.5	Teixeira et al. (2017)
Del_RN_16VLd	NA	16VLd	Rio Grande do Norte, Brazil	YES	WGS	105.6	Teixeira et al. (2017)
Del_RN_17VLd	NA	17VLd	Rio Grande do Norte, Brazil	YES	WGS	113.7	Teixeira et al. (2017)
Del_RN_18Ah	NA	18Ah	Rio Grande do Norte, Brazil	YES	WGS	108.2	Teixeira et al. (2017)
Del_RN_19VLh	NA	19VLh	Rio Grande do Norte, Brazil	YES	WGS	115.0	Teixeira et al. (2017)
Del_RN_1VLh90	NA	1VLh90	Rio Grande do Norte, Brazil	YES	WGS	114.0	Teixeira et al. (2017)
Del_RN_20VLh	NA	20VLh	Rio Grande do Norte, Brazil	YES	WGS	103.5	Teixeira et al. (2017)
Del_RN_2VLh90	NA	2VLh90	Rio Grande do Norte, Brazil	YES	WGS	109.5	Teixeira et al. (2017)

Supplementary Table 3.1 (continued)

Del_RN_3176	3176	MCAN_BR_2010_PV63	Rio Grande do Norte, Brazil	YES	WGS	68.9	This study
Del_RN_3177	3177	MCAN_BR_2010_PV64	Rio Grande do Norte, Brazil	YES	WGS	43.3	This study
Del_RN_3178	3178	MCAN_BR_2010_PV65	Rio Grande do Norte, Brazil	YES	WGS	73.6	This study
Del_RN_3182	3182	MCAN_BR_2010_PV69	Rio Grande do Norte, Brazil	YES	WGS	80.1	This study
Del_RN_3183	3183	MCAN_BR_2010_PV71	Rio Grande do Norte, Brazil	YES	WGS	93.6	This study
Del_RN_3184	3184	MCAN_BR_2010_PV72	Rio Grande do Norte, Brazil	YES	WGS	45.3	This study
Del_RN_3185	3185	MCAN_BR_2010_PV73	Rio Grande do Norte, Brazil	YES	WGS	154.6	This study
Del_RN_3186	3186	MCAN_BR_2010_PV74	Rio Grande do Norte, Brazil	YES	WGS	44.2	This study
Del_RN_3187	3187	MCAN_BR_2010_PV75	Rio Grande do Norte, Brazil	YES	WGS	24.5	This study
Del_RN_3335	3335	MCAN_BR_2011_PV 128	Rio Grande do Norte, Brazil	YES	WGS	33.8	This study
Del_RN_3VLh90	NA	3VLh90	Rio Grande do Norte, Brazil	YES	WGS	110.1	Teixeira et al. (2017)
Del_RN_4VLh90	NA	4VLh90	Rio Grande do Norte, Brazil	YES	WGS	110.6	Teixeira et al. (2017)
Del_RN_5VLh90	NA	5VLh90	Rio Grande do Norte, Brazil	YES	WGS	113.8	Teixeira et al. (2017)
Del_RN_6CLh	NA	6CLh	Rio Grande do Norte, Brazil	YES	WGS	103.1	Teixeira et al. (2017)
Del_RN_7VLd	NA	7VLd	Rio Grande do Norte, Brazil	YES	WGS	112.6	Teixeira et al. (2017)
Del_RN_9Ah	NA	9Ah	Rio Grande do Norte, Brazil	YES	WGS	110.2	Teixeira et al. (2017)
Del_RS_3196	3196	MCAN_BR_2010_LUNA II	Rio Grande do Sul, Brazil	YES	WGS	24.1	This study
Del_SP_3250	3250	MCAN_BR_2009_CLV9	São Paulo, Brazil	YES	WGS	26.7	This study
Del_SP_3256	3256	MCAN_BR_2009_CLV22	São Paulo, Brazil	YES	WGS	27.8	This study
Del_SP_3257	3257	MCAN_BR_2011_IMTS-14	São Paulo, Brazil	YES	WGS	39.8	This study
MIX_MA_02A	NA	MHOM_BR_05_MA02A	Maranhão, Brazil	NO	WGS	103.7	Carnielli et al. (2018)
MIX_MA_05A	NA	MHOM_BR_05_MA05A	Maranhão, Brazil	NO	WGS	60.1	Carnielli et al. (2018)
HTZ_MT_3134	3134	MCAN_BR_2009_GRANDÃO I	Mato Grosso, Brazil	ca. 50%	WGS	144.1	This study
HTZ_MT_3135	3135	MCAN_BR_2009_GRANDÃO II	Mato Grosso, Brazil	ca. 50%	WGS	160.8	This study
HTZ_MT_3137	3137	MCAN_BR_2009_SOL	Mato Grosso, Brazil	ca. 50%	WGS	163.3	This study
HTZ_MT_3224	3224	MCAN_BR_2010_GUG	Mato Grosso, Brazil	ca. 50%	WGS	194.3	This study
MIX_PI_04A	NA	MHOM/BR/05/PI04A	Piauí, Brazil	NO	WGS	67.8	Carnielli et al. (2018)
MIX_PI_05A	NA	MHOM/BR/06/PI05A	Piauí, Brazil	ca. 75%	WGS	59.5	Carnielli et al. (2018)
Del_RN_3330	3330	MHOM/BR/2011/Diag 1367	Rio Grande do Norte, Brazil	YES	PCR	NA	NA
Del_RN_3336	3336	MHOM/BR/2011/TC 03	Rio Grande do Norte, Brazil	YES	PCR	NA	NA
Del_RN_3339	3339	MHOM/BR/2011/TC 18	Rio Grande do Norte, Brazil	YES	PCR	NA	NA

Supplementary Table 3.1 (continued)

Del_RN_3340	3340	MHOM/BR/2011/TC 28	Rio Grande do Norte, Brazil	YES	PCR	NA	NA
Del_RN_3341	3341	MHOM/BR/2011/TC 50	Rio Grande do Norte, Brazil	YES	PCR	NA	NA
Del_RN_3342	3342	MHOM/BR/2011/TC 65	Rio Grande do Norte, Brazil	YES	PCR	NA	NA
Del_RN_3343	3343	MHOM/BR/2011/TC 95	Rio Grande do Norte, Brazil	YES	PCR	NA	NA
MIX_PI_08A	NA	MHOM/BR/05/PI08A	Piauí, Brazil	ca. 75%	WGS	66.4	Carnielli et al. (2018)
MIX_PI_09A	NA	MHOM/BR/05/PI09A	Piauí, Brazil	NO	WGS	88.7	Carnielli et al. (2018)
MIX_PI_10A	NA	MHOM/BR/06/PI10A	Piauí, Brazil	NO	WGS	118.0	Carnielli et al. (2018)
HTZ_PI_2949	2949	MCAN_BR_2004_LIBPI-18	Piauí, Brazil	ca. 50%	WGS	33.4	This study
HTZ_SP_3254	3254	MCAN_BR_2009_CLV17	São Paulo, Brazil	ca. 50%	WGS	29.5	This study
NonDel_ES_1345	NA	NA	ND, Spain	NO	WGS	110.5	This study
NonDel_ES_1356	NA	NA	ND, Spain	NO	WGS	131.8	This study
NonDel_ES_JPCM5	NA	MCAN/ES/98/LLM-724	ND, Spain	NO	WGS	128.7	This study
NonDel_FR_1	NA	MCAN_FR_1987_RM1	Marseille, France	NO	WGS	62.0	ENA
NonDel_FR_103	NA	MHOM_FR_1999_CRE103	Corsica, France	NO	WGS	99.1	ENA
NonDel_FR_114	NA	MHOM_FR_1995_LPN114	Côte d'Azur, France	NO	WGS	80.1	ENA
NonDel_FR_161	NA	MHOM_FR_1996_LPM161	Provence, France	NO	WGS	72.5	ENA
NonDel_FR_2652	NA	MHOM_FR_1993_LEM2652	Pyrenees Orientales, France	NO	WGS	86.4	ENA
NonDel_FR_47	NA	MHOM_FR_1962_LRC-L47	ND, France	NO	WGS	59.5	ENA
NonDel_FR_66	NA	MHOM_FR_1990_LPN66	Côte d'Azur, France	NO	WGS	93.7	ENA
NonDel_IT_08	NA	MHOM_IT_2002_ISS2508	Piemonte/Lombardia, Italy	NO	WGS	62.0	ENA
NonDel_IT_20	NA	MCAN_IT_2002_ISS2420	Sicily, Italy	NO	WGS	76.4	ENA
NonDel_IT_26	NA	MHOM_IT_2002_ISS2426	Campania, Italy	NO	WGS	60.1	ENA
NonDel_IT_29	NA	MHOM_IT_2002_ISS2429	Campania, Italy	NO	WGS	84.5	ENA
NonDel_IT_74	NA	IPRF_IT_85_ISS174	Abruzzo, Italy	NO	WGS	55.4	ENA
NonDel_MA_01A	NA	MHOM_BR_06_MA01A	Maranhão, Brazil	NO	WGS	104.5	Carnielli et al. (2018)
NonDel_MA_03A	NA	MHOM_BR_06_MA03A	Maranhão, Brazil	NO	WGS	151.6	Carnielli et al. (2018)
NonDel_MA_082	NA	MHOM_MA_67_ITMAP26-sc-1866082	ND, Morocco	NO	WGS	38.4	ENA
NonDel_MA_263	NA	MHOM_MA_67_ITMAP-263	ND, Morocco	NO	WGS	53.4	ENA
NonDel_MG_14A	NA	MHOM_BR_05_MG14A	Minas Gerais, Brazil	NO	WGS	84.2	Carnielli et al. (2018)
NonDel_MS_MAM	NA	MHOM_BR_2003_MAM	Mato Grosso do Sul, Brazil	NO	WGS	73.1	This study
NonDel_MS_2569	2569	MCAN_BR_2003_HUGER	Mato Grosso do Sul, Brazil	NO	WGS	158.3	This study

Supplementary Table 3.1 (continued)

NonDel_MS_2664	2664	MCAN_BR_2002_LVV-135	Mato Grosso do Sul, Brazil	NO	WGS	89.6	This study
NonDel_MS_2665	2665	MCAN_BR_2002_LVV-136	Mato Grosso do Sul, Brazil	NO	WGS	152.5	This study
NonDel_MS_2666	2666	MCAN_BR_2002_LVV-137	Mato Grosso do Sul, Brazil	NO	WGS	155.3	This study
NonDel_MS_2668	2668	MCAN_BR_2002_LVV-139	Mato Grosso do Sul, Brazil	NO	WGS	176.3	This study
NonDel_MS_2669	2669	MCAN_BR_2002_LVV-140	Mato Grosso do Sul, Brazil	NO	WGS	119.4	This study
NonDel_MS_2671	2671	MCAN_BR_2002_LVV-145	Mato Grosso do Sul, Brazil	NO	WGS	35.3	This study
NonDel_MS_2688	2688	MCAN_BR_2002_JACK_CUSTEAU	Mato Grosso do Sul, Brazil	NO	WGS	79.1	This study
NonDel_MS_2702	2702	MHOM_BR_2003_phufms-155	Mato Grosso do Sul, Brazil	NO	WGS	26.8	This study
NonDel_MS_2935	2935	MHOM_BR_2007_ARL	Mato Grosso do Sul, Brazil	NO	WGS	198.5	This study
NonDel_MS_3035	3035	MCAN_BR_2007_CG-2	Mato Grosso do Sul, Brazil	NO	WGS	194.4	This study
NonDel_MT_3210	3210	MCAN_BR_2010_CALSITO I	Mato Grosso, Brazil	NO	WGS	140.7	This study
NonDel_PA_317	NA	MHOM_PA_1979_WR317	ND, Panama	NO	WGS	87.5	ENA
NonDel_PA_85	NA	MHOM_PA_1978_WR285	ND, Panama	NO	WGS	28.9	ENA
NonDel_PI_07A	NA	MHOM/BR/06/PI07A	Piauí, Brazil	NO	WGS	65.0	Carnielli et al. (2018)
NonDel_PI_12A	NA	MHOM/BR/05/PI12A	Piauí, Brazil	NO	WGS	86.0	Carnielli et al. (2018)
NonDel_PI_2972	2972	MHOM_BR_2005_742EMS	Piauí, Brazil	NO	WGS	13.8	This study
NonDel_PT_151	NA	MHOM_PT_1988_IMT151	Lisbon, Portugal	NO	WGS	33.4	ENA
NonDel_PT_260	NA	MHOM_PT_2000_IMT260	Lisbon, Portugal	NO	WGS	74.8	ENA

Supplementary Table 3.2 Boundaries of the > 12 kb deletion on chr31. Start and stop sites of the deletion locus were determined by identifying base positions of the JPCM5 reference assembly where read-depth increases from a continuous stretch of zero read-depth observed on chr31. They have not yet been confirmed by amplicon analysis, e.g., sequencing across breakpoints of homologous recombination. Samples are listed in order of ascending stop sites.

ID	Start site (bp)	Stop site (bp)
Del_PI_3037	1122848	1135079
Del_SP_3257	1122842	1135089
Del_MT_3225	1122817	1135149
Del_MT_3136	1122835	1135150
Del_MT_3227	1122847	1135150
Del_RN_3186	1122848	1135152
Del_RN_3183	1122847	1135155
Del_MT_3219	1122826	1135156
Del_RN_3185	1122848	1135158
Del_RJ_3634	1122751	1135158
Del_PI_03A	1122841	1135160
Del_PI_2976	1122758	1135161
Del_MT_3138	1122843	1135161
Del_RN_3176	1122836	1135161
Del_RN_3178	1122846	1135161
Del_RN_3182	1122847	1135161
Del_MT_3208	1122834	1135161
Del_MT_3209	1122843	1135161
Del_MT_3212	1122848	1135161
Del_MT_3214	1122846	1135161
Del_MT_3223	1122847	1135161
Del_SP_3250	1122815	1135161
Del_RJ_3598	1122847	1135161
Del_MG_11A	1122848	1135161
Del_MG_16A	1122841	1135161
Del_MG_17A	1122842	1135161
Del_MG_18A	1122857	1135161
Del_MG_19A	1122840	1135161
Del_MA_04A	1122847	1135161
Del_RJ_3015	1122834	1135161
Del_HN_336	1122840	1135161
Del_PI_01A	1122847	1135162
Del_RN_3177	1122846	1135163
Del_DF_2898	1122848	1135164
Del_HN_167	1122841	1135164
Del_MA_07A	1122841	1135166
Del_ES_3068	1122847	1135167
Del_MG_15A	1122847	1135167
Del_MG_3379	1122842	1135168
Del_MG_3381	1122805	1135168
Del_RJ_3635	1122842	1135168
Del_RN_3335	1122815	1135169
Del_MG_3378	1122827	1135169
Del_MG_3380	1122846	1135169
Del_MG_12A	1122841	1135169
Del_MG_13A	1122838	1135169
Del_PI_02A	1122848	1135169

Supplementary Table 3.2 (continued)

Del_RS_3196	1122837	1135174
Del_PI_11A	1122841	1135175
Del_ES_2789	1122839	1135180
Del_SP_3256	1122846	1135181
Del_RN_3184	1122848	1135182
Del_RN_3187	1122845	1135182
Del_DF_2919	1122847	1135197
Del_ES_2788	1122849	1135215
Del_RN_11VLd	1122848	1135346
Del_RN_12VLh	1122848	1135346
Del_RN_13VLh	1122847	1135346
Del_RN_14VLh	1122848	1135346
Del_RN_15VLd	1122856	1135346
Del_RN_16VLd	1122847	1135346
Del_RN_17VLd	1122847	1135346
Del_RN_18Ah	1122848	1135346
Del_RN_19VLh	1122848	1135346
Del_RN_1VLh90	1122847	1135346
Del_RN_20VLh	1122847	1135346
Del_RN_2VLh90	1122883	1135346
Del_RN_3VLh90	1122848	1135346
Del_RN_4VLh90	1122868	1135346
Del_RN_5VLh90	1122848	1135346
Del_RN_6CLh	1122848	1135346
Del_RN_7VLd	1122848	1135346
Del_RN_9Ah	1122847	1135346

Supplementary Table 3.3 Short insertion-deletion and single-nucleotide variants fixed in Del but not fixed in NonDel *L.* infantum isolates of the New World. Variant effect and impact was determined by SNPEff[391] using the JPCM5 annotation file available at https://tritrypdb.org/common/downloads/release-33/LinfantumJPCM5/gff/data/. Abbreviations: chr. (chromosome); pos. (position); single-nucleotide polymorphism (SNP); insertion-deletion variant (INDEL).

Chr.	Pos.	Type	Effect	Impact	Affected gene ID	Product Description
25	79004	INDEL	disruptive inframe deletion	moderate	LinJ.25.0280	hypothetical protein - conserved
27	36142	INDEL	frameshift variant	low	LinJ.27.0140	hypothetical protein - conserved
2	284601	SNP	missense variant	moderate	LinJ.02.0580	hypothetical protein - conserved
8	76409	SNP	missense variant	moderate	LinJ.08.0210	inositol phosphosphingolipid phospholipase C-Like
10	213490	SNP	missense variant	moderate	LinJ.10.0510	GP63 - leishmanolysin
13	560852	SNP	missense variant	moderate	LinJ.13.1480	hypothetical protein - conserved
14	27920	SNP	missense variant	moderate	LinJ.14.0100	hypothetical protein - conserved
14	448465	SNP	missense variant	moderate	LinJ.14.1130	dynein heavy chain - putative
14	532704	SNP	missense variant	moderate	LinJ.14.1300	hypothetical protein - conserved
16	132446	SNP	synonymous variant	low	LinJ.16.0370	hypothetical protein - conserved
17	57639	SNP	synonymous variant	low	LinJ.17.0140	receptor-type adenylate cyclase - putative
18	46363	SNP	missense variant	moderate	LinJ.18.0140	Domain of unknown function DUF21 - putative
18	246206	SNP	missense variant	moderate	LinJ.18.0610	Zn-finger in Ran binding protein and others/FYVE zinc finger containing protein - putative
19	126972	SNP	missense variant	moderate	LinJ.19.0330	intraflagellar transport protein 80 - putative
22	295409	SNP	synonymous variant	low	LinJ.22.0660	5'a2rel-related protein
22	295589	SNP	stop gained	high	LinJ.22.0660	5'a2rel-related protein
22	570468	SNP	synonymous variant	low	LinJ.22.1350	ATP-dependent DEAD/H RNA helicase - putative
26	923932	SNP	synonymous variant	low	LinJ.26.2400	Fibronectin type III domain containing protein - putative
27	461406	SNP	synonymous variant	low	LinJ.27.0950	hypothetical protein - conserved
28	276690	SNP	missense variant	moderate	LinJ.28.0770	Sas10/Utp3/C1D family/Sas10 C-terminal domain containing protein - putative
29	117964	SNP	missense variant	moderate	LinJ.29.0360	hypothetical protein - conserved
29	259409	SNP	missense variant	moderate	LinJ.29.0730	hypothetical protein - conserved
29	432730	SNP	missense variant	moderate	LinJ.29.1200	hypothetical protein - conserved
30	337551	SNP	synonymous variant	low	LinJ.30.1080	hypothetical protein - conserved
31	815200	SNP	missense variant	moderate	LinJ.31.1790	hypothetical protein - conserved
32	166359	SNP	missense variant	moderate	LinJ.32.0430	Kinesin motor domain/MORN repeat - putative
32	753608	SNP	synonymous variant	low	LinJ.32.2030	Vitamin B6 photo-protection and homoeostasis - putative
33	277519	SNP	missense variant	moderate	LinJ.33.0840	NLI interacting factor-like phosphatase - putative

Supplementary Table 3.3 (continued)

33	804784	SNP	missense variant	moderate	LinJ.33.2150	hypothetical protein - conserved
33	1228999	SNP	missense variant	moderate	LinJ.33.3040	hypothetical protein - conserved
33	1310102	SNP	synonymous variant	low	LinJ.33.3130	hypothetical protein - conserved
33	1329486	SNP	missense variant	moderate	LinJ.33.3160	hypothetical protein - conserved
34	8751	SNP	synonymous variant	low	LinJ.34.0020	hypothetical protein - conserved
34	1149856	SNP	missense variant	moderate	LinJ.34.2680	regulatory subunit of protein kinase a-like protein
35	578928	SNP	synonymous variant	low	LinJ.35.1390	mitochondrial processing peptidase - beta subunit - putative
35	779019	SNP	synonymous variant	low	LinJ.35.1980	U3 small nucleolar RNA-associated protein 6 - putative
35	872261	SNP	missense variant	moderate	LinJ.35.2180	hypothetical protein - conserved
36	1595079	SNP	missense variant	moderate	LinJ.36.4410	Arrestin (or S-antigen) - N-terminal domain/Arrestin N terminal like - putative
36	2194829	SNP	synonymous variant	low	LinJ.36.6000	tRNA pseudouridine synthase TruD - putative
36	2239961	SNP	missense variant	moderate	LinJ.36.6120	a44l protein-like protein
36	2318978	SNP	missense variant	moderate	LinJ.36.6300	COG (conserved oligomeric Golgi) complex component - COG2/Domain of unknown function (DUF3510) - putative
36	2353499	SNP	missense variant	moderate	LinJ.36.6400	hypothetical protein - conserved
36	2389221	SNP	missense variant	moderate	LinJ.36.6500	hypothetical protein - conserved

Supplementary Table 3.4 Short insertion-deletion and single-nucleotide variants prevalent (> 70%) in Del but uncommon (< 50%) in NonDel isolates of the New World. Variant effect and impact was determined by SNPEff[391] using the JPCM5 annotation file available at https://tritrypdb.org/common/downloads/release-33/LinfantumJPCM5/gff/data/. Abbreviations: chr. (chromosome); pos. (position); single-nucleotide polymorphism (SNP); insertion-deletion variant (INDEL).

Chr.	Pos.	Type	Fixed in Del?	Effect	Impact	Affected gene ID	Product Description
2	284601	SNP	yes	missense variant	moderate	LinJ.02.0580	hypothetical protein - conserved
8	76409	SNP	yes	missense variant	moderate	LinJ.08.0210	inositol phosphosphingolipid phospholipase C-Like
10	213490	SNP	yes	missense variant	moderate	LinJ.10.0510	GP63 - leishmanolysin
13	560852	SNP	yes	missense variant	moderate	LinJ.13.1480	hypothetical protein - conserved
14	27920	SNP	yes	missense variant	moderate	LinJ.14.0100	hypothetical protein - conserved
14	448465	SNP	yes	missense variant	moderate	LinJ.14.1130	dynein heavy chain - putative
14	532704	SNP	yes	missense variant	moderate	LinJ.14.1300	hypothetical protein - conserved
16	132446	SNP	yes	synonymous variant	low	LinJ.16.0370	hypothetical protein - conserved
18	46363	SNP	yes	missense variant	moderate	LinJ.18.0140	Domain of unknown function DUF21 - putative
18	246206	SNP	yes	missense variant	moderate	LinJ.18.0610	Zn-finger in Ran binding protein and others/FYVE zinc finger containing protein - putative
19	126972	SNP	yes	missense variant	moderate	LinJ.19.0330	intraflagellar transport protein 80 - putative
22	570468	SNP	yes	synonymous variant	low	LinJ.22.1350	ATP-dependent DEAD/H RNA helicase - putative
26	923932	SNP	yes	synonymous variant	low	LinJ.26.2400	Fibronectin type III domain containing protein - putative
27	461406	SNP	yes	synonymous variant	low	LinJ.27.0950	hypothetical protein - conserved
28	276690	SNP	yes	missense variant	moderate	LinJ.28.0770	Sas10/Utp3/C1D family/Sas10 C-terminal domain containing protein - putative
29	117964	SNP	yes	missense variant	moderate	LinJ.29.0360	hypothetical protein - conserved
29	259409	SNP	yes	missense variant	moderate	LinJ.29.0730	hypothetical protein - conserved
29	432730	SNP	yes	missense variant	moderate	LinJ.29.1200	hypothetical protein - conserved
30	337551	SNP	yes	synonymous variant	low	LinJ.30.1080	hypothetical protein - conserved
31	815200	SNP	yes	missense variant	moderate	LinJ.31.1790	hypothetical protein - conserved
32	753608	SNP	yes	synonymous variant	low	LinJ.32.2030	Vitamin B6 photo-protection and homoeostasis - putative
33	277519	SNP	yes	missense variant	moderate	LinJ.33.0840	NLI interacting factor-like phosphatase - putative
33	804784	SNP	yes	missense variant	moderate	LinJ.33.2150	hypothetical protein - conserved
33	1228999	SNP	yes	missense variant	moderate	LinJ.33.3040	hypothetical protein - conserved
33	1310102	SNP	yes	synonymous variant	low	LinJ.33.3130	hypothetical protein - conserved
33	1329486	SNP	yes	missense variant	moderate	LinJ.33.3160	hypothetical protein - conserved
34	8751	SNP	yes	synonymous variant	low	LinJ.34.0020	hypothetical protein - conserved
34	1149856	SNP	yes	missense variant	moderate	LinJ.34.2680	regulatory subunit of protein kinase a-like protein

Supplementary Table 3.4 (continued)

35	578928	SNP	yes	synonymous variant	low	LinJ.35.1390	mitochondrial processing peptidase - beta subunit - putative
35	779019	SNP	yes	synonymous variant	low	LinJ.35.1980	U3 small nucleolar RNA-associated protein 6 - putative
35	872261	SNP	yes	missense variant	moderate	LinJ.35.2180	hypothetical protein - conserved
36	1595079	SNP	yes	missense variant	moderate	LinJ.36.4410	Arrestin (or S-antigen) - N-terminal domain/Arrestin N terminal like - putative
36	2194829	SNP	yes	synonymous variant	low	LinJ.36.6000	tRNA pseudouridine synthase TruD - putative
36	2239961	SNP	yes	missense variant	moderate	LinJ.36.6120	a44l protein-like protein
36	2318978	SNP	yes	missense variant	moderate	LinJ.36.6300	COG (conserved oligomeric Golgi) complex component - COG2/Domain of unknown function (DUF3510) - putative
36	2353499	SNP	yes	missense variant	moderate	LinJ.36.6400	hypothetical protein - conserved
36	2389221	SNP	yes	missense variant	moderate	LinJ.36.6500	hypothetical protein - conserved
10	527755	SNP	no	missense variant	moderate	LinJ.10.1420	Protein of unknown function (DUF1861) - putative
11	512816	SNP	no	missense variant	moderate	LinJ.11.1240	ATP-binding cassette protein subfamily A, member 4, putative
15	361793	SNP	no	missense variant	moderate	LinJ.15.0870	Putative methyltransferase/GDP dissociation inhibitor - putative
16	215250	SNP	no	synonymous variant	low	LinJ.16.0590	carbamoyl-phosphate synthase - putative
18	461497	SNP	no	synonymous variant	low	LinJ.18.1120	Multisite-specific tRNA:(cytosine-C(5))-methyltransferase - putative
19	250811	SNP	no	synonymous variant	low	LinJ.19.0590	protein kinase - putative
24	60769	SNP	no	missense variant	moderate	LinJ.24.0250	hypothetical protein - conserved
24	288610	SNP	no	synonymous variant	low	LinJ.24.0810	hypothetical protein - conserved
24	375050	SNP	no	synonymous variant	low	LinJ.24.1080	AKAP7 2'5' RNA ligase-like domain containing protein - putative
24	396569	SNP	no	missense variant	moderate	LinJ.24.1130	formin-like protein
25	460586	SNP	no	missense variant	moderate	LinJ.25.1230	modification methylase-like protein
25	760791	SNP	no	missense variant	moderate	LinJ.25.2090	2 -4-dihydroxyhept-2-ene-1 -7-dioic acid aldolase - putative
26	158632	SNP	no	missense variant	moderate	LinJ.26.0560	spliced leader RNA PSE-promoter transcription factor - putative
31	640893	SNP	no	missense variant	moderate	LinJ.31.1450	hypothetical protein - conserved
31	1254441	SNP	no	missense variant	moderate	LinJ.31.2680	RNA polymerase ii largest subunit
31	1380994	SNP	no	missense variant	moderate	LinJ.31.3080	acetyl-CoA carboxylase - putative
32	439198	SNP	no	synonymous variant	low	LinJ.32.1160	exportin 1 - putative
32	1040642	SNP	no	missense variant	moderate	LinJ.32.2790	SpoU rRNA Methylase family - putative
32	1077100	SNP	no	missense variant	moderate	LinJ.32.2890	Guanine nucleotide exchange factor in Golgi transport N-terminal/Sec7 domain containing protein - putative
34	461911	SNP	no	missense variant	moderate	LinJ.34.1110	uracil phosphoribosyltransferase - putative
35	1558774	SNP	no	synonymous variant	low	LinJ.35.4020	predicted zinc finger protein
35	1885700	SNP	no	synonymous variant	low	LinJ.35.4960	hypothetical protein - conserved

Supplementary Table 3.4 (continued)

35	1941935	SNP	no	missense variant	moderate	LinJ.35.5080	SPRY domain/HECT-domain (ubiquitin-transferase)
25	79004	INDEL	yes	disruptive inframe deletion	moderate	LinJ.25.0280	hypothetical protein - conserved
27	36142	INDEL	yes	frameshift variant	high	LinJ.27.0140	hypothetical protein - conserved
36	643759	INDEL	no	disruptive inframe deletion	moderate	LinJ.36.1720	universal minicircle sequence binding protein - putative

Supplementary Table 3.5 Gene copy number variation between Del and NonDel *L. infantum* isolates of the New World. Haploid somy estimates (s) in 89 coding regions differed by more than 0.3 between Del and NonDel groups. This table describes results from the 42 of these 90 regions that appear statistically significant in Mann-Whitney U (MWU) analysis using a Bonferroni-corrected p-value cut-off of 0.05 / 89 = 0.000562. A Bray-Curtis distance matrix calculated from the s values of these significantly differentiated regions was used to cluster samples in a heatmap in Supplementary Fig. 3.3. The heatmap excludes values for the four genes within the chr31 deletion locus (grey font) in order to assess whether other copy number changes correlate with this trait. The heatmap exposes a strong correlation between geographic origin and s. We therefore reassessed all 42 regions by analysis of covariance (ANCOVA) with geographic origin applied as a covariate to Del vs. NonDel chr31 read-depth profile. Only nine coding regions remain significant (bold font) with the additional covariate applied. The last column of this table also indicates the proportion of uniquely mapping nucleotides within each coding region (see Methods). Poor mappability is likely for genes occurring in multiple paralogs and may explain instances where s > 4 in Supplementary Fig. 3.3. Product descriptions were obtained from the JPCM5 annotation file available at https://tritrypdb.org/common/downloads/release-24/LinfantumJPCM5/gff/data/. Additional abbreviations: n (sample size); ∆s (mean s in Del minus mean s in NonDel isolates).

Gene ID and product description	MWU p-value	Mean s in Del (n = 55)	Mean s in NonDel (n = 18)	∆s	ANCOVA p-value	Uniquely mappable (%)
LinJ.02.0160:LinJ.02:80373:82337:phosphoglycan+beta+1_3+galactosyltransferase+_SCGR4_	0.0004	2.2180	3.0810	-0.8630	0.3839	1.93
LinJ.02.0170:LinJ.02:89052:91370:phosphoglycan+beta+1_3+galactosyltransferase+_SCGR3_	0.0004	2.6400	3.7940	-1.1540	0.4152	37.92
LinJ.02.0180:LinJ.02:93807:96398:phosphoglycan+beta+1_3+galactosyltransferase+_SCGR2_	0.0003	1.5740	2.0740	-0.5000	0.5721	34.12
LinJ.04.0160:LinJ.04:50165:51751:hypothetical+protein_+conserved+in+leishmania	< 0.0001	0.9600	1.5960	-0.6370	0.0125	70.55
LinJ.08.0700:LinJ.08:301452:302054:amastin-like+protein	< 0.0001	0.8250	1.3770	-0.5520	**0.0001**	0.00
LinJ.08.0710:LinJ.08:306082:306684:amastin-like+protein	0.0001	0.8570	1.3030	-0.4470	0.0079	0.00
LinJ.10.0490:LinJ.10:206839:208638:GP63_+leishmanolysin+_GP63-1_	0.0001	11.2940	8.3450	2.9490	0.4722	0.00
LinJ.10.0500:LinJ.10:209886:211685:GP63_+leishmanolysin_metallo-peptidase_+Clan+MA_M__+Family+M8+_GP63-2_	0.0001	11.0570	8.1720	2.8850	0.3610	0.00
LinJ.10.0510:LinJ.10:212954:214879:GP63_+leishmanolysin_metallo-peptidase_+Clan+MA_M__+Family+M8+_GP63-3_	< 0.0001	2.3750	1.6110	0.7650	0.1451	63.64
LinJ.10.0520:LinJ.10:216085:218376:GP63_+leishmanolysin_metallo-peptidase_+Clan+MA_M__+Family+M8+_GP63-3_	< 0.0001	3.6430	2.2390	1.4040	0.0642	29.38
LinJ.10.0521:LinJ.10:220045:220938:hypothetical+protein_+unknown+function	0.0001	4.0800	2.9580	1.1220	0.0615	100.00
LinJ.10.0530:LinJ.10:222401:224197:GP63_+leishmanolysin_metallo-peptidase_+Clan+MA_M__+Family+M8+_GP63-4_	0.0002	1.7590	1.4110	0.3480	0.1868	33.74
LinJ.15.0730:LinJ.15:282530:286768:hypothetical+protein	< 0.0001	4.8650	6.6360	-1.7710	0.1775	2.62
LinJ.15.1240:LinJ.15:489392:490867:nucleoside+transporter+1_+putative	< 0.0001	0.8970	0.5210	0.3760	**< 0.0001**	0.00
LinJ.15.1250:LinJ.15:493117:494592:nucleoside+transporter+1_+putative	< 0.0001	0.9040	0.5850	0.3190	**< 0.0001**	0.00
LinJ.19.0030:LinJ.19:7832:8239:histone+H2B	0.0002	1.8520	1.4110	0.4400	0.9973	40.79
LinJ.19.0040:LinJ.19:8691:9014:histone+H2B	0.0001	1.6770	1.3340	0.3430	0.5829	26.01
LinJ.23.1330:LinJ.23:531423:533402:hypothetical+protein_+unknown+function	< 0.0001	0.5280	0.8580	-0.3300	0.5873	4.14
LinJ.23.1340:LinJ.23:534939:536918:hypothetical+protein_+unknown+function	< 0.0001	0.5710	0.8800	-0.3090	0.5916	4.35
LinJ.26.0100:LinJ.26:24208:25278:hypothetical+protein_+conserved	0.0005	1.1170	1.5740	-0.4560	0.6186	10.47
LinJ.26.snoRNA18:LinJ.26:695914:695973:LM26Cs2H2	< 0.0001	1.3840	1.0630	0.3210	0.3624	0.00

Supplementary Table 3.5 (continued)

LinJ.27.0220:LinJ.27:52294:53238:hypothetical+protein_+conserved	< 0.0001	1.0570	1.6950	-0.6380	0.0342	100.00
LinJ.27.0230:LinJ.27:54073:54636:hypothetical+protein_+conserved	0.0001	1.0820	1.7550	-0.6730	0.0384	100.00
LinJ.28.0610:LinJ.28:211929:213629:major+surface+protease+gp63_+putative_leishmanolysin_+putative	< 0.0001	1.0310	1.3660	-0.3350	0.7427	0.00
LinJ.28.2950:LinJ.28:1063717:1065579:heat-shock+protein+hsp70_+putative	< 0.0001	2.0600	2.3980	-0.3380	0.1473	0.00
LinJ.28.3000:LinJ.28:1078481:1080445:heat-shock+protein+hsp70_+putative	< 0.0001	2.0390	2.3490	-0.3100	0.3610	0.00
LinJ.28.3060:LinJ.28:1094771:1096981:heat-shock+protein+hsp70_+putative	< 0.0001	2.1350	2.5150	-0.3790	0.3960	2.67
LinJ.29.1880:LinJ.29:803599:805374:paraflagellar+rod+protein+1D_+putative	< 0.0001	2.4860	2.1690	0.3170	**< 0.0001**	0.00
LinJ.29.1890:LinJ.29:807645:809432:paraflagellar+rod+protein+1D_+putative	0.0001	2.5870	2.2600	0.3270	**0.0001**	0.06
LinJ.30.2990:LinJ.30:1081801:1082886:glyceraldehyde+3-phosphate+dehydrogenase_+glycosomal	< 0.0001	1.0080	1.3920	-0.3830	0.0022	0.00
LinJ.30.3000:LinJ.30:1083609:1084694:glyceraldehyde+3-phosphate+dehydrogenase_+glycosomal	< 0.0001	0.9980	1.4510	-0.4530	0.0221	0.00
LinJ.31.1470:LinJ.31:655800:657122:hypothetical+protein_+unknown+function	< 0.0001	1.1220	1.8260	-0.7040	0.1780	100.00
LinJ.31.2370:LinJ.31:1123458:1124504:3%27-nucleotidase%2Fnuclease_+putative	< 0.0001	0.0010	0.9850	-0.9850	**< 0.0001**	97.23
LinJ.31.2380:LinJ.31:1126381:1127517:3%27-nucleotidase%2Fnuclease+precursor_+putative	< 0.0001	0.0000	0.9900	-0.9900	**< 0.0001**	97.45
LinJ.31.2390:LinJ.31:1128003:1130729:helicase-like+protein	< 0.0001	0.0000	1.0160	-1.0160	**< 0.0001**	100.00
LinJ.31.2400:LinJ.31:1133533:1134582:3_2-trans-enoyl-CoA+isomerase_+mitochondrial+precursor_+putative	< 0.0001	0.0120	0.9780	-0.9660	**< 0.0001**	33.46
LinJ.32.3110:LinJ.32:1163891:1164346:nucleoside+diphosphate+kinase+b	< 0.0001	1.1180	0.7250	0.3930	0.0159	25.27
LinJ.33.0360:LinJ.33:117668:119770:heat+shock+protein+83+_HSP83-2_	0.0004	5.9000	6.2400	-0.3390	0.5122	0.00
LinJ.35.0490:LinJ.35:150339:168242:proteophosphoglycan+ppg4	0.0005	1.6420	2.0680	-0.4260	0.3840	9.28
LinJ.35.0520:LinJ.35:206553:214586:proteophosphoglycan+ppg4	0.0005	1.8450	2.3600	-0.5150	0.1763	1.01
LinJ.35.0550:LinJ.35:239069:241789:proteophosphoglycan+ppg1	0.0002	1.4800	1.8120	-0.3320	0.5309	10.48
LinJ.36.snoRNA1:LinJ.36:933685:933707:LM36C3C3	0.0005	2.6750	2.3050	0.3700	0.9473	0.00

Supplementary Table 3.6 Significant heterozygosity increases in HTZ and Old World *L. infantum* groups. The Kruskall-Wallis rank sum test indicates that genome-wide inbreeding coefficients (F_{IS} values) differ among Del, HTZ, MIX, New World (NW) and Old World (OW) NonDel groups (p-value < 0.001). This table lists F_{IS} medians and p-values from post-hoc pairwise comparisons using the Tukey and Kramer (Nemenyi) test. Results indicate significant F_{IS} reductions in HTZ and Old World NonDel groups. Hyphens replace redundant comparisons. Medians for raw counts of heterozygous loci (Het.) are also shown. Het. values produce analogous values in Kruskal-Wallis and Nemenyi tests (not shown).

Group	Median F_{IS}	Median Het.	vs. Del	vs. HTZ	vs. MIX	vs. NW NonDel
Del	0.293	277	-	-	-	-
HTZ	-0.284	499	**< 0.001**	-	-	-
MIX	0.265	299	0.65635	0.24284	-	-
NW NonDel	0.336	265.5	0.86028	**< 0.001**	0.38029	-
OW NonDel	0.098	379	**< 0.001**	0.92008	0.46149	**< 0.001**

Supplementary Table 3.7 Demographic simulation model input. Template (.tpl) files describe demographic models and parameters of interest in fastsimcoal2[315]. File content unique to each of the ten models (bold font) simulated in this study is listed below. Data type descriptions (e.g., contig numbers and sizes, recombination and mutation rates) common to all model templates occur at the end each .tpl file. This information is shown after the asterisked rows at the bottom of the table. Each model is further outlined in Supplementary Fig. 3.8.

```
//AMbot parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_OW
N_MT or N_MS
//sampless sizes and samples age
17
11 or 15, respectively
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices : 0 implies no migration between demes
2
//migration matrix 0
0 0
0 0
//migration matrix 1
0 MIG12
MIG21 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
3 historical events
TMIG 0 0 0 1 0 1
TBOT 1 1 0 FOU 0 1
TDIV 1 0 1 1 0 0
//number of independent loci (chromosomes)
36 1
```

Supplementary Table 3.7 (continued)

//**AM** parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_pop1
N_pop2
//sampless sizes and samples age
n_pop1
n_pop2
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices : 0 implies no migration between demes
2
//migration matrix 0
0 0
0 0
//migration matrix 1
0 MIG12
MIG21 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
2 historical events
TMIG 0 0 0 1 0 1
TDIV 0 1 1 1 0 0
//number of independent loci (chromosomes)
36 1

//**IMbot** parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_MT
N_MS
//sampless sizes and samples age
15
11
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices : 0 implies no migration between demes
2
//migration matrix 0
0 MIG12
MIG21 0
//migration matrix 1
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
2 historical events
TBOT 0 0 0 FOU 0 0
TDIV 0 1 1 1 0 1
//number of independent loci (chromosomes)
36 1

Supplementary Table 3.7 (continued)

//**IM$_{change}$** parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_pop1
N_pop2
//sampless sizes and samples age
n_pop1
n_pop2
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices : 0 implies no migration between demes
3
//migration matrix 0
0 mig12
mig21 0
//migration matrix 1
0 MIG12
MIG21 0
//migration matrix 2
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
2 historical events
TMIG 0 0 0 1 0 1
TDIV 0 1 1 1 0 2
//number of independent loci (chromosomes)
36 1

//**IM** parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_pop1
N_pop2
//sampless sizes and samples age
n_pop1
n_pop2
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices : 0 implies no migration between demes
2
//migration matrix 0
0 MIG12
MIG21 0
//migration matrix 1
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
1 historical event
TDIV 1 0 1 1 0 1
//number of independent loci (chromosomes)
36 1

Supplementary Table 3.7 (continued)

//**SC** parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_pop1
N_pop2
//sampless sizes and samples age
n_pop1
n_pop2
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices : 0 implies no migration between demes
2
//migration matrix 0
0 MIG12
MIG21 0
//migration matrix 1
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
3 historical events
TSC 0 1 ADM01 1 0 1
TSC 1 0 ADM10 1 0 1
TDIV 0 1 1 1 0 1
//number of independent loci (chromosomes)
36 1

//**SCbot**$_{\textbf{nomig}}$ parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_MT
N_MS
//sampless sizes and samples age
15
11
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices : 0 implies no migration between demes
2
//migration matrix 0
0 MIG12
MIG21 0
//migration matrix 1
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
3 historical events
TSC 0 0 0 1 0 1
TBOT 0 0 0 FOU 0 1
TDIV 0 1 1 1 0 1
//number of independent loci (chromosomes)
36 1

Supplementary Table 3.7 (continued)

//**SC_{nomig}** parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_pop1
N_pop2
//sampless sizes and samples age
n_pop1
n_pop2
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices : 0 implies no migration between demes
2
//migration matrix 0
0 MIG12
MIG21 0
//migration matrix 1
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
2 historical events
TSC 0 0 0 1 0 1
TDIV 0 1 1 1 0 1
//number of independent loci (chromosomes)
36 1

//**SIbot** parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_OW
N_MT or N_MS
//sampless sizes and samples age
17
11 or 15, respectively
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices: 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
2 historical events
TBOT 1 1 0 FOU 0 0
TDIV 1 0 1 1 0 0
//number of independent loci (chromosomes)
36 1

Supplementary Table 3.7 (continued)

//**SI** parameters for the coalescence simulation program fsc252.exe

2 samples to simulate
//population effective sizes (number of genes)
N_OW
N_MT or N_MS
//sampless sizes and samples age
17
11 or 15, respectively
//growth rates: negative growth implies population expansion
0
0
//number of migration matrices: 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
1 historical event
TDIV 1 0 1 1 0 0
//number of independent loci (chromosomes)
36 1

****

****

****

\\number of contiguous locus blocks on chromosome 1
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 277951 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 2
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 334113 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 3
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 382367 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 4
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 475338 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 5
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 449024 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 6
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 523352 0 1.99e-9 OUTEXP

Supplementary Table 3.7 (continued)

\\number of contiguous locus blocks on chromosome 7
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 592382 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 8
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 495393 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 9
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 572115 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 10
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 547235 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 11
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 575792 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 12
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 568477 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 13
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 645761 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 14
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 639279 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 15
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 617636 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 16
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 698903 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 17
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 667340 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 18
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 720194 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 19
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 742501 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 20
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 732590 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 21
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 759899 0 1.99e-9 OUTEXP

Supplementary Table 3.7 (continued)

\\number of contiguous locus blocks on chromosome 22
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 659512 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 23
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 774004 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 24
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 867075 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 25
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 886912 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 26
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 1050165 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 27
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 1043947 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 28
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 1163438 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 29
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 1221905 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 30
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 1365115 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 31
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 1468864 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 32
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 1547509 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 33
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 1448148 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 34
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 1668239 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 35
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 2068523 0 1.99e-9 OUTEXP
\\number of contiguous locus blocks on chromosome 36
1
\\Per block: number of loci, recombination rate to the right-side locus, plus optional parameters
DNA 2673956 0 1.99e-9 OUTEXP

Chapter 4

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

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

[a]Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK

[b]Center for Research on Health in Latin America, School of Biological Sciences, Pontifical Catholic University of Ecuador, Quito, Ecuador

[c]Laboratorio de Biología Molecular de Protozoarios, Instituto de Medicina Tropical, Universidad Central de Venezuela, Caracas, Venezuela

[d]Grupo de Investigaciones Microbiológicas, Programa de Biología, Universidad del Rosario, Bogotá, Colombia

[e]London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

[f]Infectious and Tropical Disease Institute, Biomedical Sciences Department, Heritage College of Osteopathic Medicine, Ohio University, 45701 Athens, OH, USA

Work presented in this chapter is currently under peer review at PLoS Genetics.

4.1 Abstract

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

4.2 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* (Chapter 2)), reconstruct past[486] and present transmission networks (e.g., *Staphylococcus* infections within hospitals)[487] or identify the genetic bases of virulence[426,488] and resistance to drugs (see examples from *Plasmodium* spp.[489,490]). 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 hosts[257]. 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[102,491,492], and gradual reductions to genetic diversity are often observed over time[103]. Karyotypic changes are also known to arise during *T. cruzi* micromanipulation and axenic growth[178,493].

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 pathogen[348], vector[494] or host[495,496]. Others describe the use of biotinylated hybridization probes[349,497–499] or selective whole-genome amplification, e.g., based on the strand displacement function of phi29 DNA polymerase[500]. 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'[501], 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 Scientific[502], 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 markers[503] or to genotype neutral SNP variation for landscape genetic techniques[23]. 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 common[504], 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 ebola[505] and tuberculosis[506] but the use of custom panels for other pathogen species (e.g., *Bifidobacterium*[507] or human papilloma virus[508]) 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.

4.3 Methods

4.3.1 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 Center for Research on Health in Latin America (CISeAL) in Loja Province, Ecuador, following protocols described in Grijalva et al. (2012)[509]. 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[510–512], 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[513–515]. 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. 4.1.

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 and Tropical Medicine (LSHTM). DNA extractions at LSHTM followed Messenger et al. (2015)[58].

Uninfected *Rhodnius prolixus* gut tissue samples used for mock infections (see 'Method development and library preparation') were also provided by LSHTM. Special thanks to M. Lewis and M. Yeo for supervising dissections. Insects were euthanized with CO_2 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 the Center for Research on Health in Latin America (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 WGS (see Chapter 2) and served as a control for GLST method development in this study.

4.3.2 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 Ecuador (Chapter 2). Briefly, Chapter 2 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 v0.7.3[353]. Single-nucleotide polymorphisms (SNPs) were summarized by population-based genotype and likelihood assignment in Genome Analysis Toolkit v3.7.0[389], 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' mask[390] 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. (2012)[265] 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 TriTrypDB[516]. Such conserved segments may be least prone to repeat-driven nucleotide diversity and as such most amenable to PCR[306]. The 6,259 synteny segments found by OrthoMCL therefore proceeded to primer search with the high-throughput primer design engine BatchPrimer3[517]. 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 ΔG = -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 ΔG = -9.0) underwent a last filtering step by screening for perfect matches in Chapter 2's 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 v12.4.3[518]. 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 v3.4.1[394], the 403 SNPs also reproduced neighbor-joining relationships observed based on total polymorphism identified by WGS (Supplementary Fig. 4.1). 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. 4.1.

4.3.3 Wet lab method development and library preparation

The 203 primers pairs designed above (Supplementary Tbl. 4.2) were purchased from Eurofins Genomics (Ebersberg, Germany) at 200 µM concentration in salt-free, 96-well plate format. Primer pairs were first tested individually to establish cycling conditions for PCR (Supplementary Fig. 4.2). 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 µl reactions contained 5 µl Q5 High-Fidelity Master Mix (New England Biolabs), 1 µl forward primer [10 µM], 1 µl reverse primer [10 µM], and 3 µl 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 µM, resulting in individual primer concentrations of 50.75 µM / 406 = 125 nM. GLST reactions incorporated 2 µl of this primer mix rather than two separate 1 µl 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. 4.3). 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'[519] satellite DNA (commonly targeted to diagnose human *T. cruzi* infections).

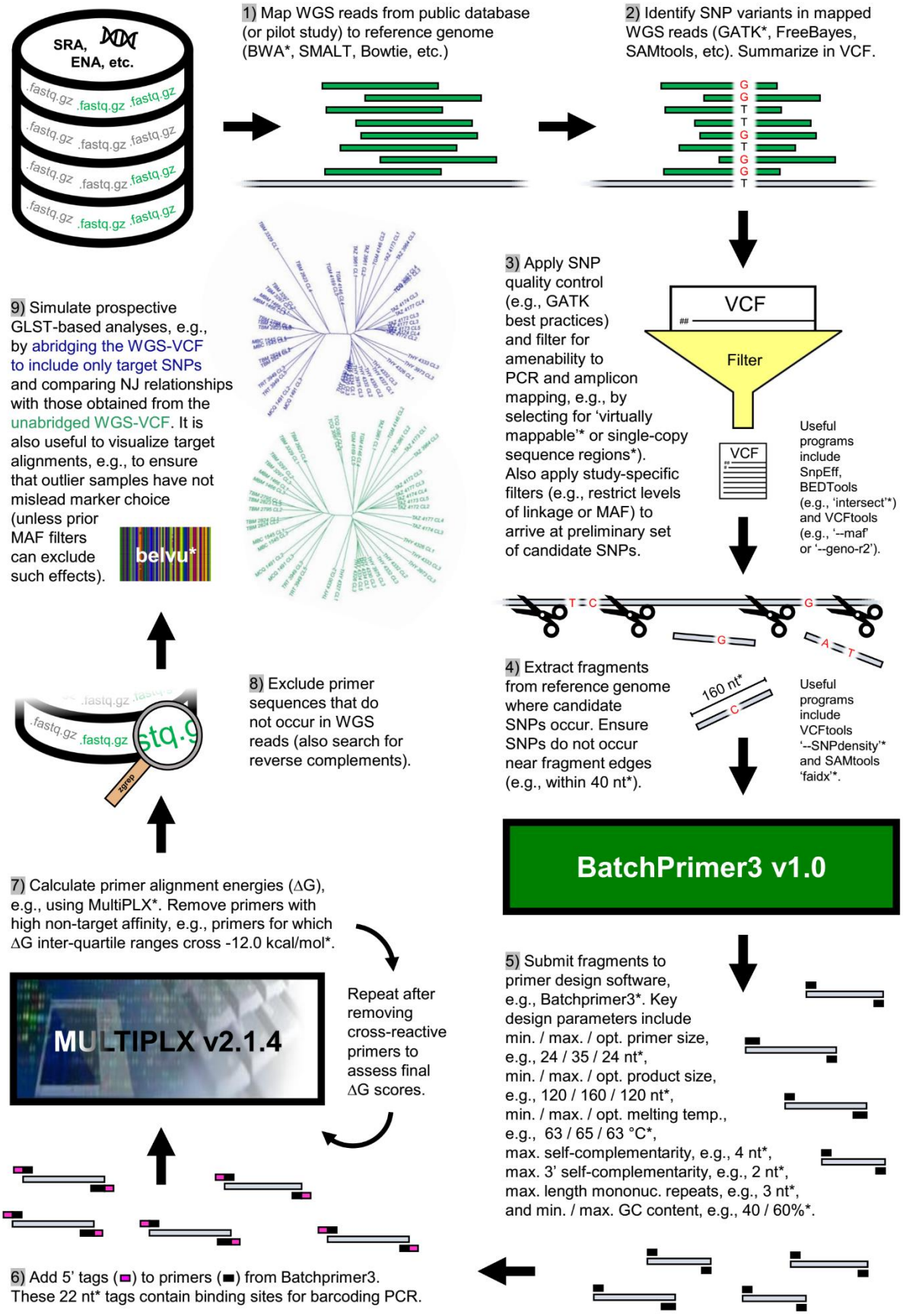


Figure 4.1 GLST sequence target selection from preliminary genomic data. Nine steps of primer panel construction and validation run clockwise from top left. Various methods and criteria can be applied to complete many of these steps. Those specific to this study are asterisked, e.g., we used BWA in step 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).

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/µl to 2.50 pg/µl in dH_2O (Supplementary Fig. 4.4). Each 20 µl qPCR reaction consisted of 10 µl SensiMix SYBR Low-ROX reagent (Bioline), 1 µl TcZ forward primer (5'-GCTCTTGCCCACAMGGGTGC-3')[519] [10 µM], 1 µl TcZ reverse primer (5'-CCAAGCAGCGGATAGTTCAGG-3')[519] [10 µM], 7 µl dH_2O, and 1 µl 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/µl *T. cruzi* concentrations based on qPCR proceeded to final library construction (Supplementary. Tbl. 4.1) 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. 4.5). 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. 4.6). 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 dH_2O 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. 4.7) were quantified by Qubit fluorometry (Thermo Fisher Scientific), and pooled at equimolar concentrations, gel-excised, re-solubilized, and verified by microfluidic electrophoresis (Supplementary Fig. 4.8) as above.

4.3.4 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 DNA[520] 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 CS2rc[521].

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 v0.4.3[522] to decontaminate reads. After merging alignment (.bam) files from sequencing runs 1 and 2 with Picard Tools v1.11[388], single-nucleotide polymorphisms (SNPs) were identified in each sample using the 'HaplotypeCaller' algorithm in GATK v3.7.0[389]. 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. 4.9). 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.

4.3.5 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 nodes[523]. 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 v3.4.1[394].

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.[524]), '--het' (calculates inbreeding coefficients using a method of moments[525]) and '--hwe' (filters sites by deviation from Hardy-Weinberg Equilibrium following Wigginton et al.[526]). F_{ST} differentiation was calculated using ARLSUMSTAT v3.5.2[459,527].

Correlations between geographic and genetic differences were also calculated from non-reference allele counts in R v3.4.1[394]. The 'mantel' function from the 'vegan' package v2.4.4[463] 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. (2016)[528] (Supplementary Tbl. 4.1). 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 samples[529], 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. 4.7). 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.

4.4 Results

4.4.1 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. 4.2). Median read-depth was 266x across all sites. Of 403 loci targeted from Chapter 2's WGS dataset, 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. 4.10).

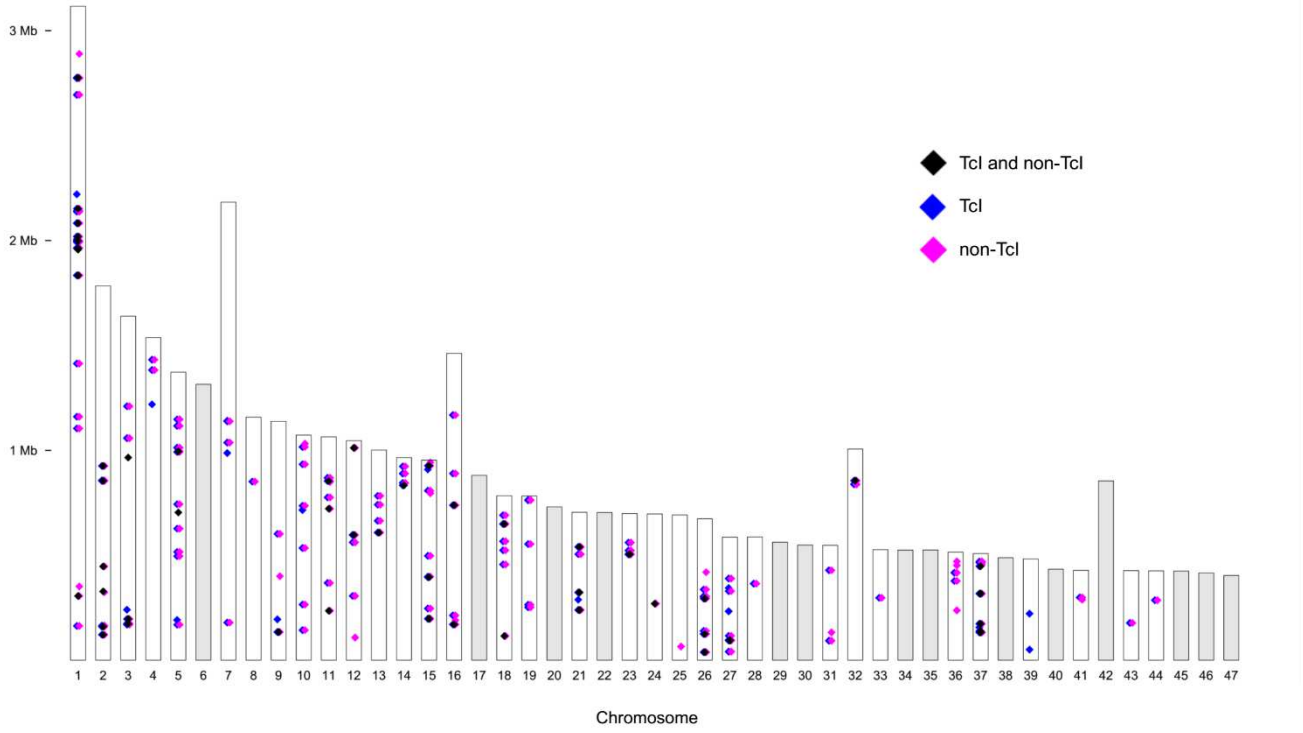


Figure 4.2 Variant loci detected in *T. cruzi* I samples and reference clones of other sub-lineages. The genome-wide distribution of SNP variants is shown relative to the TcI-Sylvio reference assembly. Each column represents one of 47 putative chromosomes. Pink diamonds comprise 393 variants that occur 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. 4.1). Pairwise AD did not correlate to minimum, maximum or difference in mean read-depth between the two replicates.

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).

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

Sampling sites in Colombia, Venezuela and Ecuador are plotted in Fig. 4.3, and a median-joining network of allelic differences among GLST genotypes is shown in Fig. 4.4. 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).

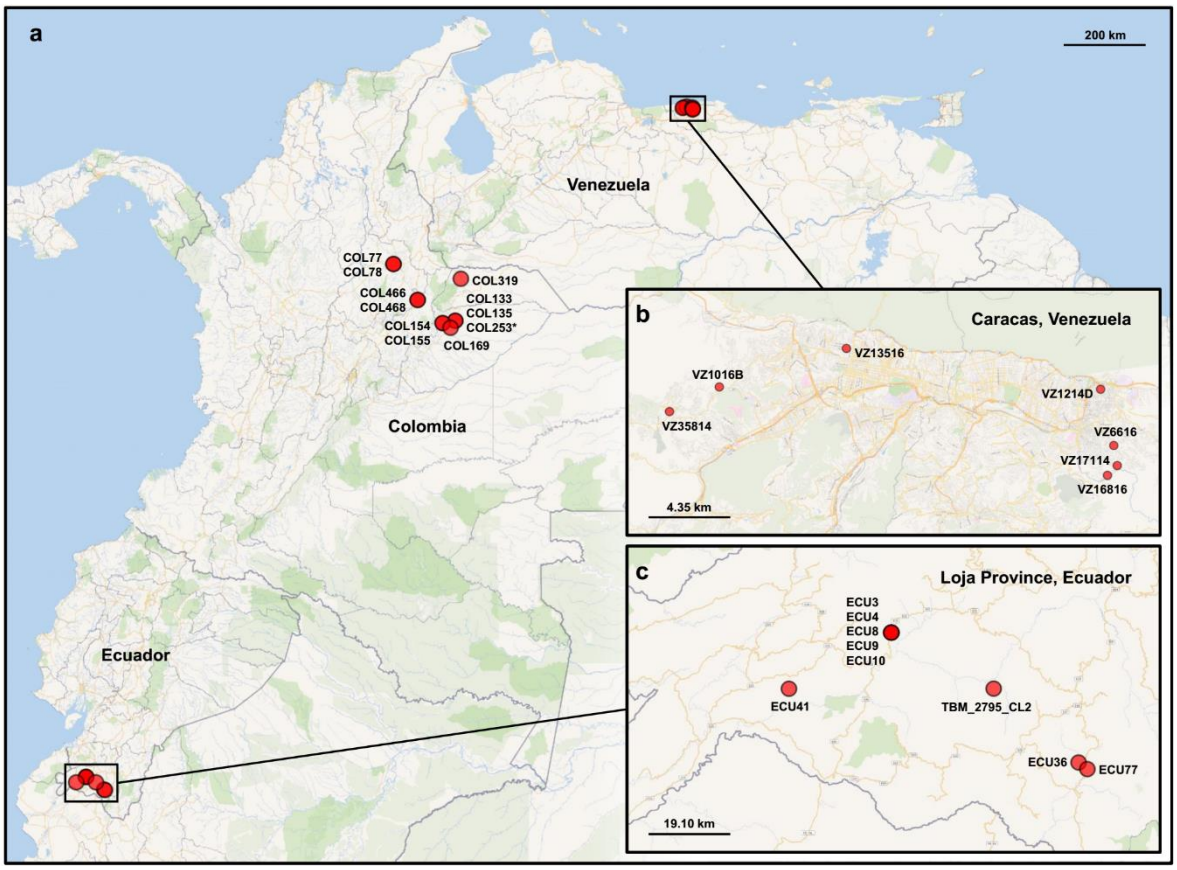


Figure 4.3 Map of vector sampling sites. **a** Sampling in Colombia involved a larger spatial area than that in Venezuela and Ecuador. *T. cruzi*-infected intestinal material was collected from *Panstrongylus* and *Rhodnius* vectors in Arauca, Casanare, Santander and Boyacá. We asterisk COL253 because 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. 4.1 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. 4.2). Hardy-Weinberg ratios, linkage and inbreeding coefficients are also listed in Tbl. 4.2.

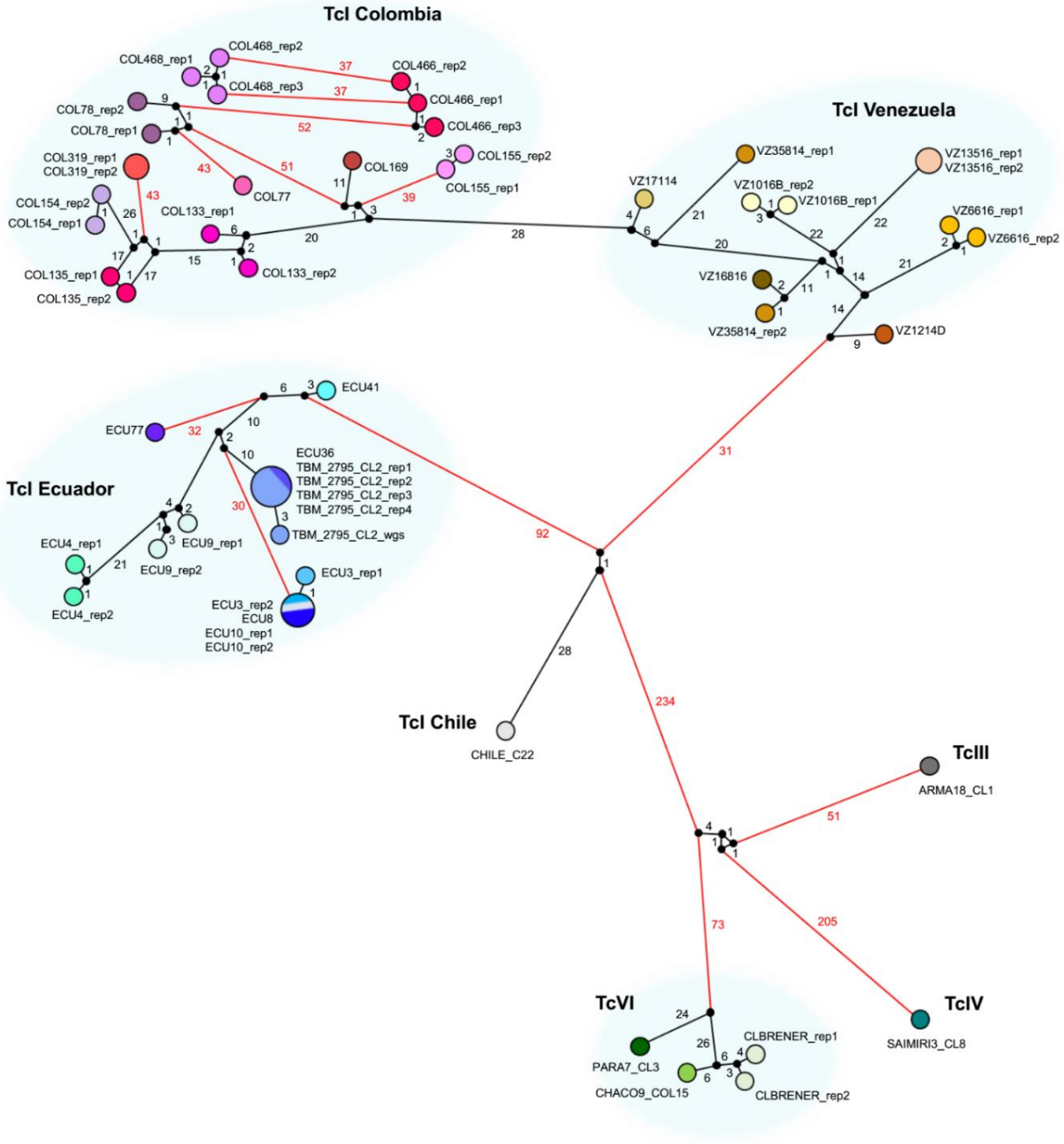


Figure 4.4 Allelic differences among *T. cruzi* I samples and reference clones of other sub-lineages as a median-joining network. A single SNP locus can differ by 0, 1 or 2 between two individuals (i.e., the individuals match at both, one, or neither allele). The AD measurement indicated on each edge 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 4.1 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. 4.3).

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 4.2 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. 4.2 and Fig. 4.5a): Mantel's r decreased to 0.30 (p = 0.001) after restricting analysis to sample pairs separated by < 250 km (Fig. 4.5b).

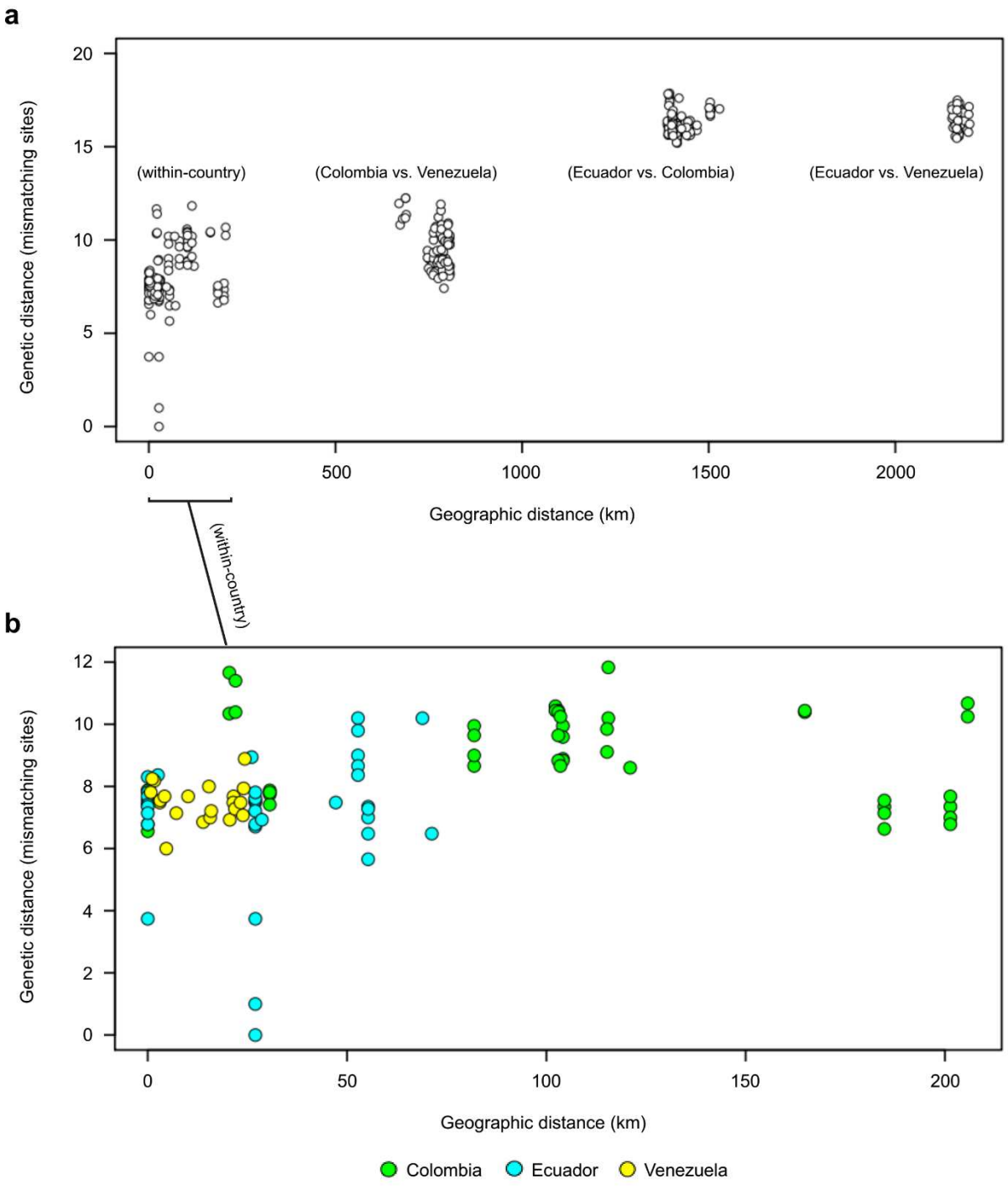


Figure 4.5 Isolation-by-distance among *T. cruzi* I samples. **a** Each circle represents geographic and genetic distances between two TcI samples. Global IBD is significant (Mantel's r = 0.89, p = 0.001) but driven by divergence between Ecuadorian samples and the rest of dataset (see two clusters at 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 ≤ 10% missing genotypes (n = 285) are used in analysis. Only the first replicate is used for samples represented by multiple replicates.

Within-country isolation by distance (IBD)[530] 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. 4.5b).

Finally, GLST also clearly separated sub-lineages TcI, TcIII, TcIV, and TcVI in network (Fig. 4.3) and neighbor-joining tree construction (Fig. 4.6). 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)).

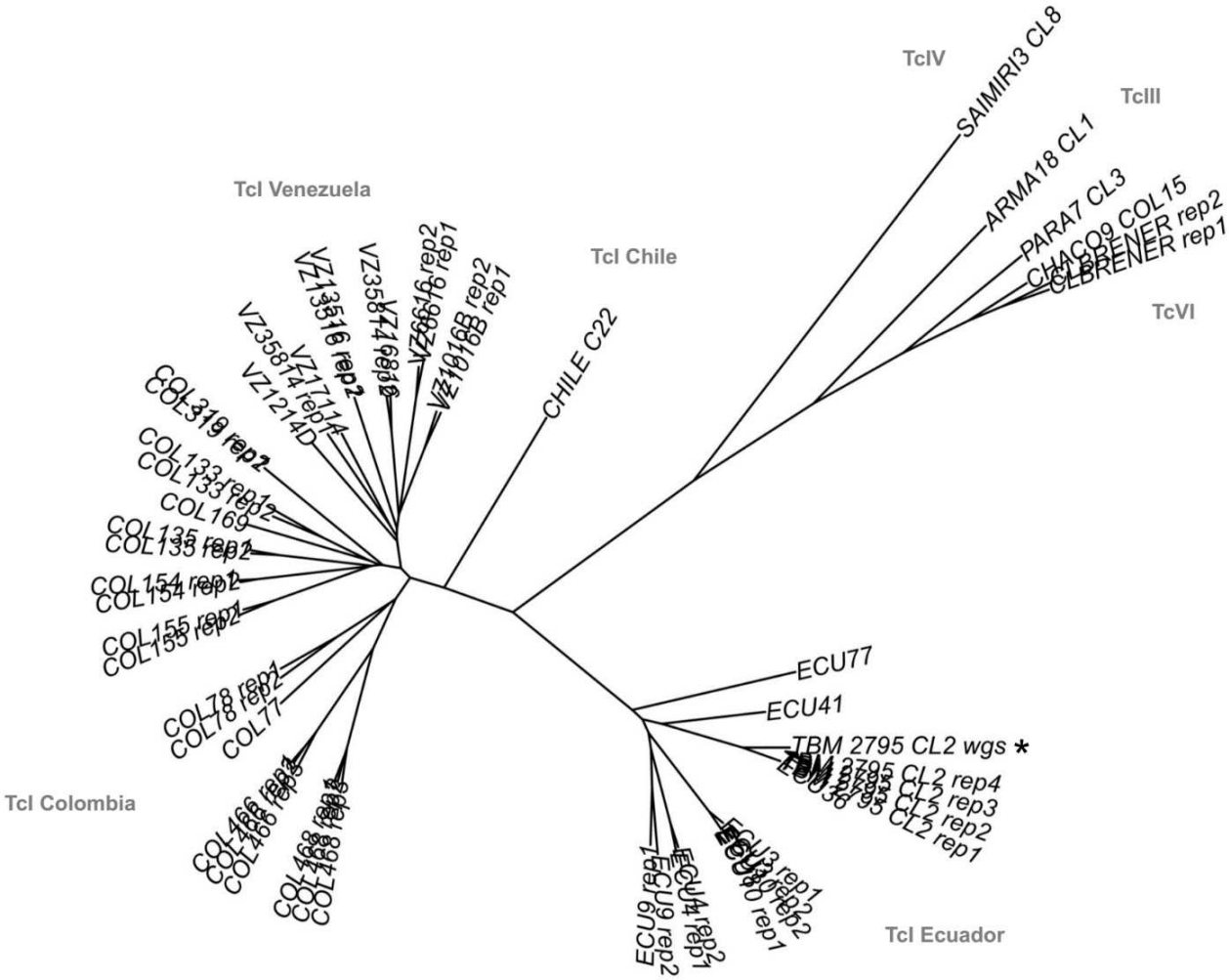


Figure 4.6 Neighbor-joining relationships among *T. cruzi* I samples and reference clones of other sub-lineages. Genetic distances are based on 556 biallelic SNP sites for which genotypes are called in all individuals. Results indicate high repeatability among most technical replicates (see 'rep1 – 4' 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 Chapter 2.

4.5 Discussion

4.5.1 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.

4.5.2 Cost-effective spatio-genetic analysis

GLST achieved an important resolution benchmark in recovering IBD 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 IBD[23]. 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. 4.3) 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. 4.3).

4.5.3 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[139,140], ecological host-fitting[149] and 'cryptic sexuality'[14] in shaping population genetic structure in TcI. In some cases[129,151] (but others not[140,147,149]), 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 MLGs[140]. 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 Chapter 2 and Ocaña-Mayorga et al. (2010)[140]) 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 effects[531]. 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 clades[531].

4.5.4 Adjustment and transferability

Considering the great variety of sample types to which studies have successfully applied PCR[532–536], 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 heart[537]) 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 TcI[538])

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-hybridization[539] 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 or error were, however, well within margins expected for methods involving PCR[540]. 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 variation[541], 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 screens[542]. The candidate SNP pool can easily be filtered for such criteria during GLST panel design, e.g., using SnpEff[391] or BEDTools[543] and data mining strategies at EuPathDB[544]. 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 VCFtools[396], 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 scale[545]. 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[315,326,546], Bayesian clustering[335], admixture/migration[547,548] and hybridization tests[549]. Deviation from HWE may occur more frequently in specific genetic regions (e.g., near centromeres[550]), 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 tools[551].

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 MultiPLX[552]. 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.

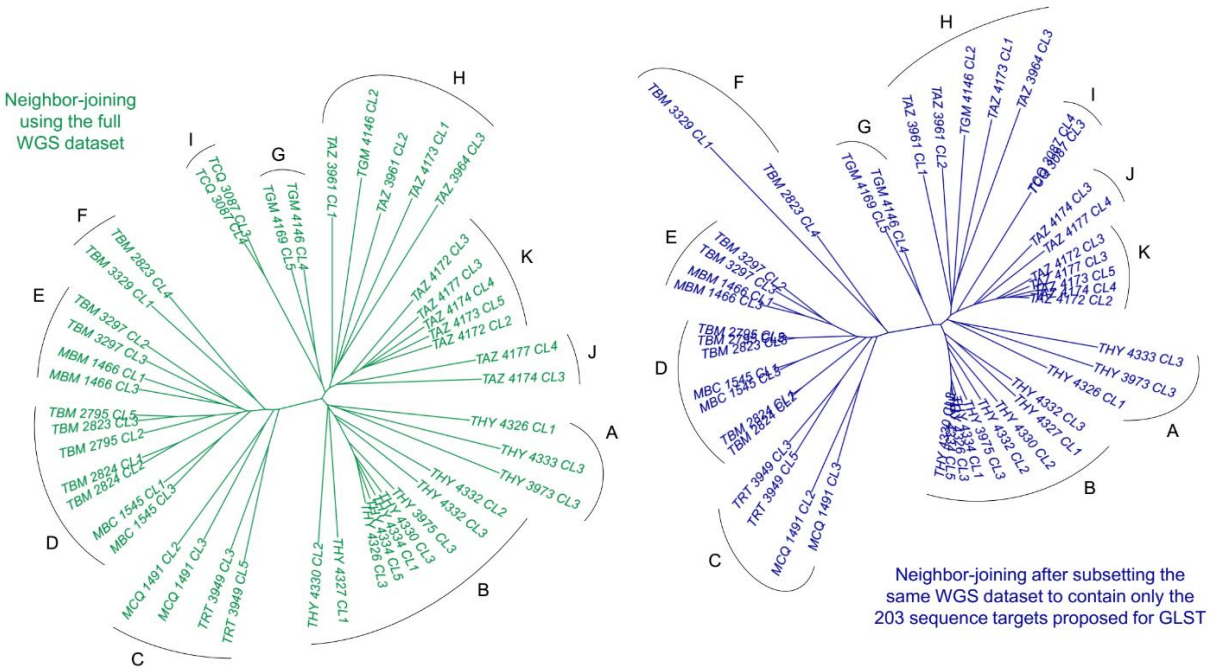
4.5.5 Prospects

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. 4.1). 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' (as opposed to geographical distances – see Chapter 5's glossary of landscape genetic terms (Box 5.3))[553] along the basin range. Fuelling landscape genetic simulators such as CDMetaPOP[326] 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. 4.4b). Sylvatic *P. geniculatus* vector populations appear to be rapidly adapting to habitats within Caracas[515,554] 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 tools[312], e.g., using bioinformatic strategies developed for malaria

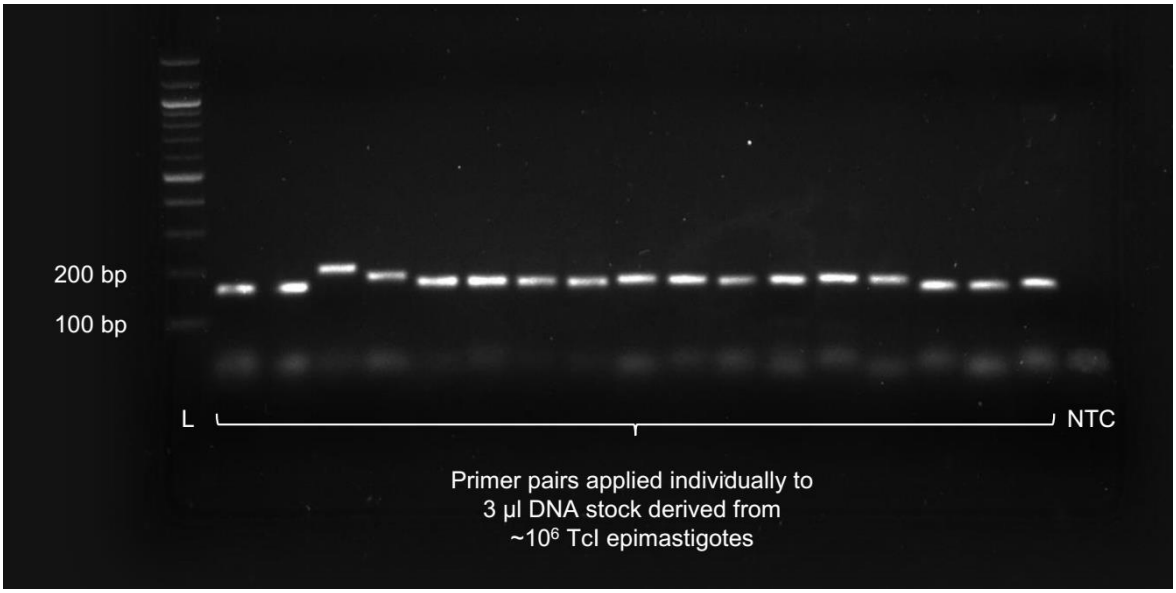
research[314,555–557]. Multiclonality has important implications for public health[558,559] but its potential prevalence in *T. cruzi* vectors and hosts[311–313] is difficult to describe from cultured cells[311,386]. Countless other applications are conceivable for GLST. Some research fields, however, 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* genome[306].

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.

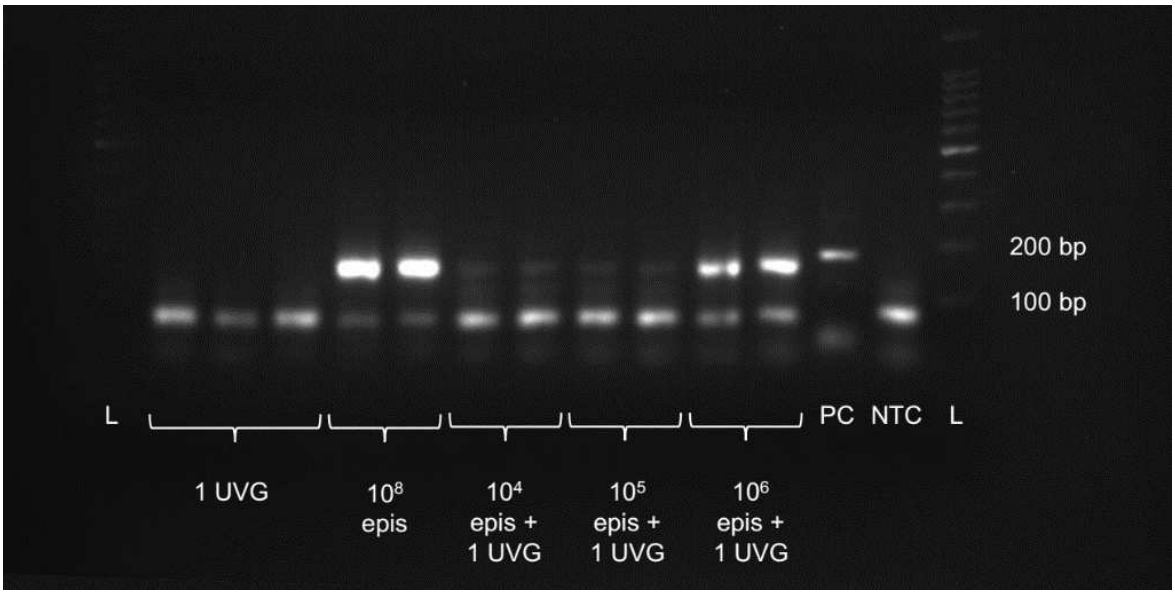
4.6 Supplementary figures and tables



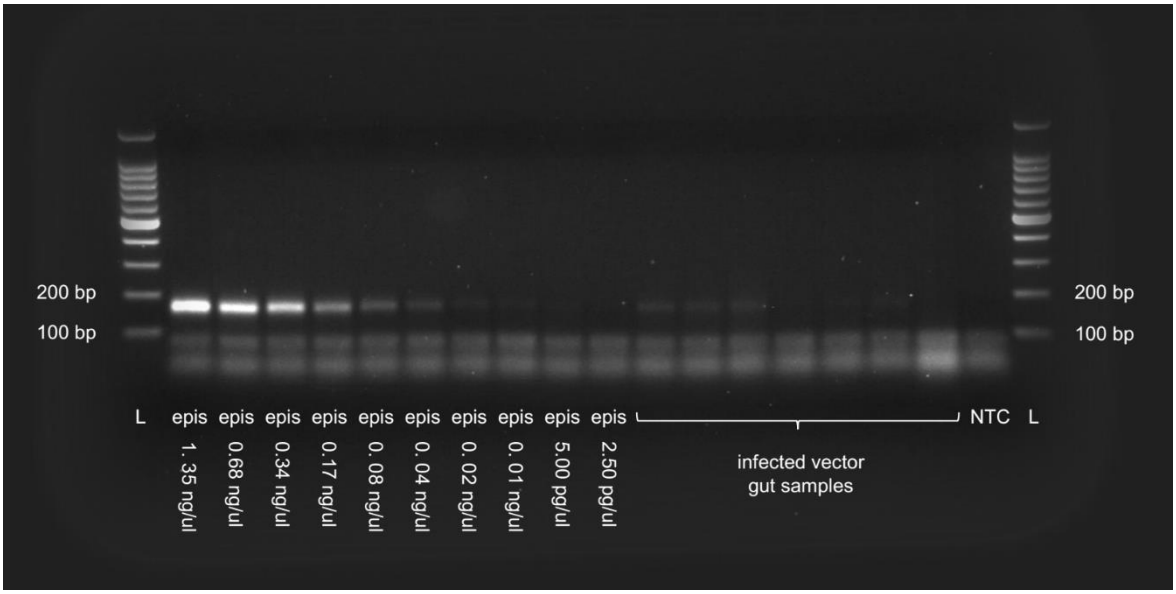
Supplementary Figure 4.1 Phylogenetic resolution at GLST loci *in silico*. The green tree shows neighbor-joining (NJ) relationships calculated from 106,007 SNP sites identified from whole-genome sequencing (WGS) of 45 TcI clones in southern Ecuador (Chapter 2). Sites missing genotypes in ≥ 10% individuals are excluded. Less than 45 km separate the most distant sampling sites within the study region. Several pairs of clones also represent the same host/vector individual (see first seven characters of IDs). NJ was repeated after abridging the WGS dataset to contain only SNPs within the 203 sequence targets proposed by GLST (also excluding sites missing ≥ 10% genotypes). This resultant tree (blue, at right) uses 391 SNP sites and recreates clusters A-K observed in WGS.



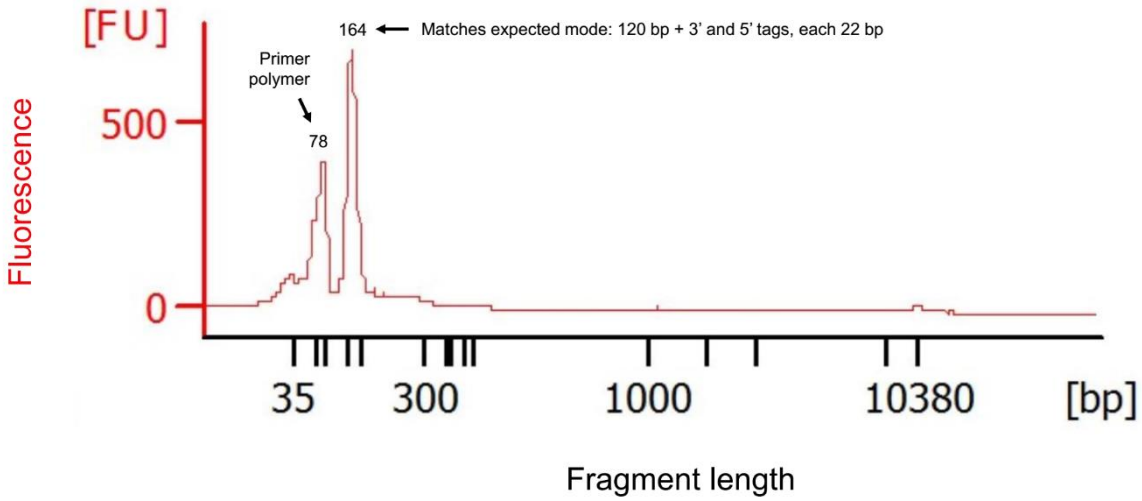
Supplementary Figure 4.2 Individual primer pair validation. Primer pairs were first applied individually to pure TcI epimastigote DNA to confirm product amplification within the expected size range (164 – 204 bp). The figure shows the electrophoresed products of 17 different primer pairs in 0.8% agarose gel as well as DNA ladder (L) and no-template control (NTC). All other primer pairs achieved similar results using 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).



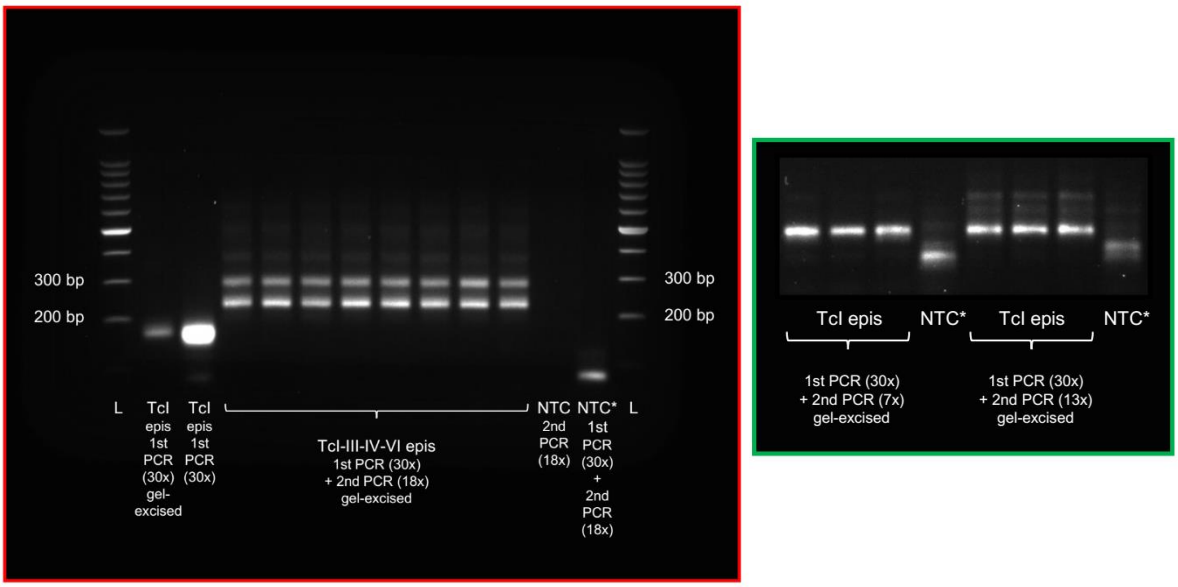
Supplementary Figure 4.3 Preliminary GLST (multiplex) trials on *T. cruzi* I mock infections. We created mock infections by mixing 10^4, 10^5 and 10^6 RNAlater-preserved TcI-Sylvio epimastigote (epi) cells with uninfected *Rhodnius prolixus* vector gut (UVG). DNA extracted from these mock infections was subjected to the multiplexed, 203-target GLST reaction (using the same cycling conditions as for single-target reactions – see Methods or Supplementary Fig. 4.2 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.



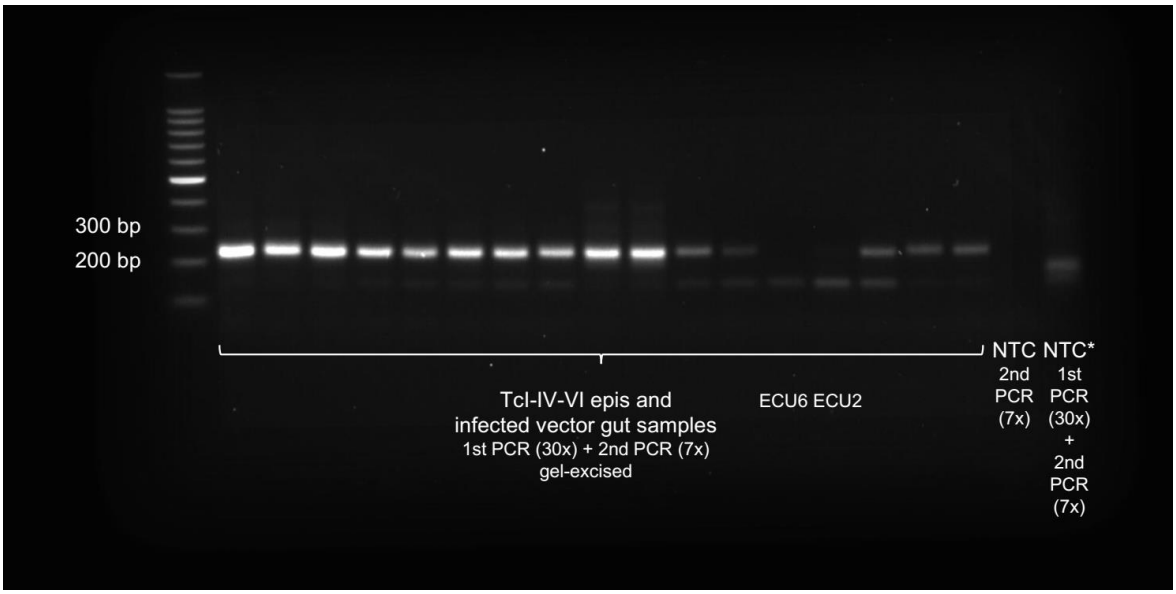
Supplementary Figure 4.4 *T. cruzi* I DNA dilutions and GLST product visibility in 0.8% agarose gel. The left side shows electrophoresed GLST amplicons generated from 3 µl pure TcI epimastigote (epi) DNA with concentrations between 1.35 ng/µl and 2.50 pg/µl (see cycling conditions in Methods or Supplementary Fig. 4.2 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.



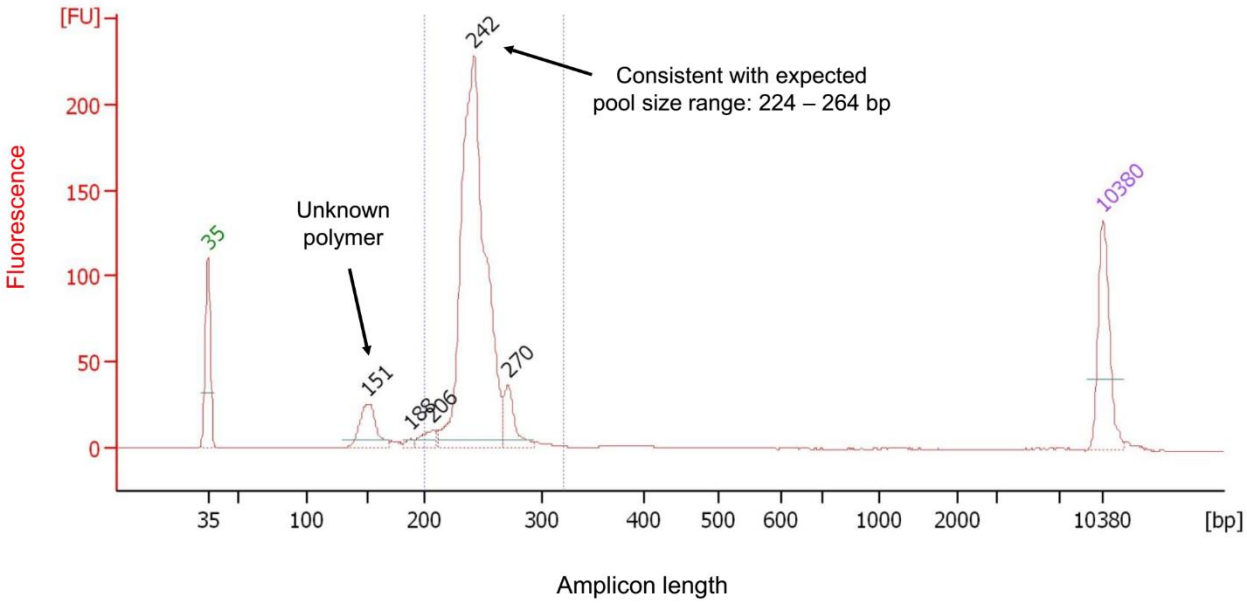
Supplementary Figure 4.5 First-round (unbarcoded) PCR product size composition measurement using microfluidic electrophoresis. The figure plots fragment sizes (calculated based on migration times relative to those of standards) and fluorescence intensity (FU) of first-round PCR products (see cycling conditions in Methods or Supplementary Fig. 4.2 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.



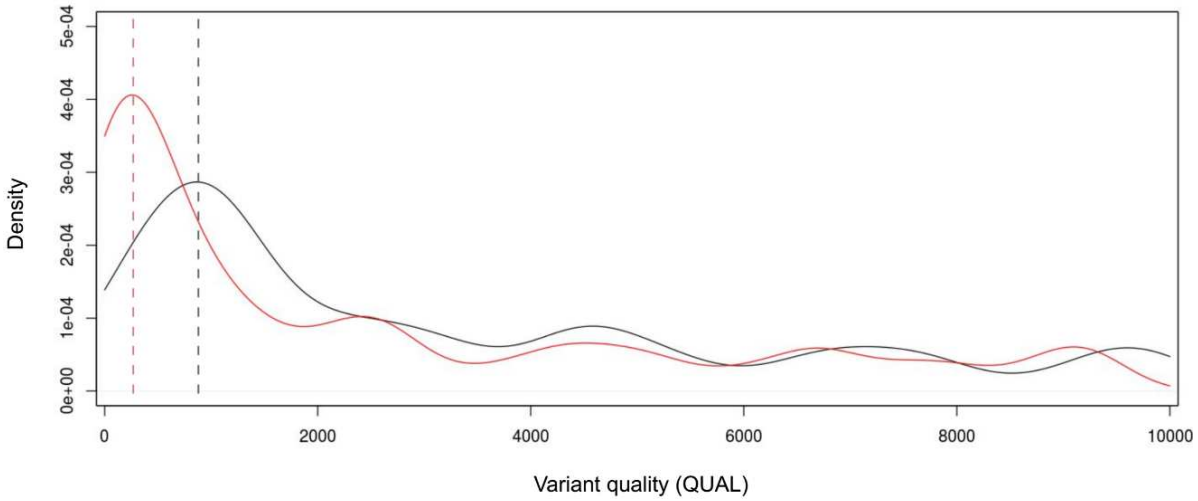
Supplementary Figure 4.6 Large polymer formation from excessive amplicon barcoding. The second (barcoding) PCR reaction uses an initial incubation step at 98 °C (2 min); 7 amplification cycles at 98 °C (30 s), 60 °C (30 s), and 72 °C (1 min); and a final extension step at 72 °C (3 min). Seven amplification 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 TcI 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.



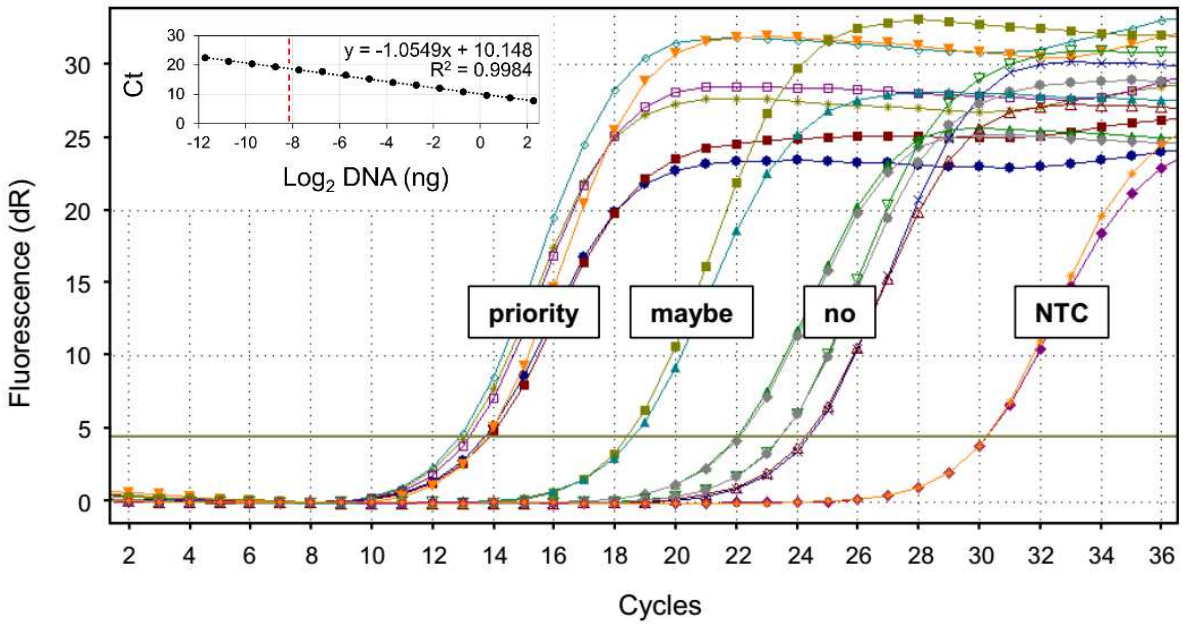
Supplementary Figure 4.7 Barcoded GLST products ready for final pooling and purification. The 0.8% agarose gel shows a subset of fifteen GLST products from the second-round (barcoding) PCR reaction (see cycling conditions in Methods or Supplementary Fig. 4.6 legend) prior to equimolar pooling and final 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*).



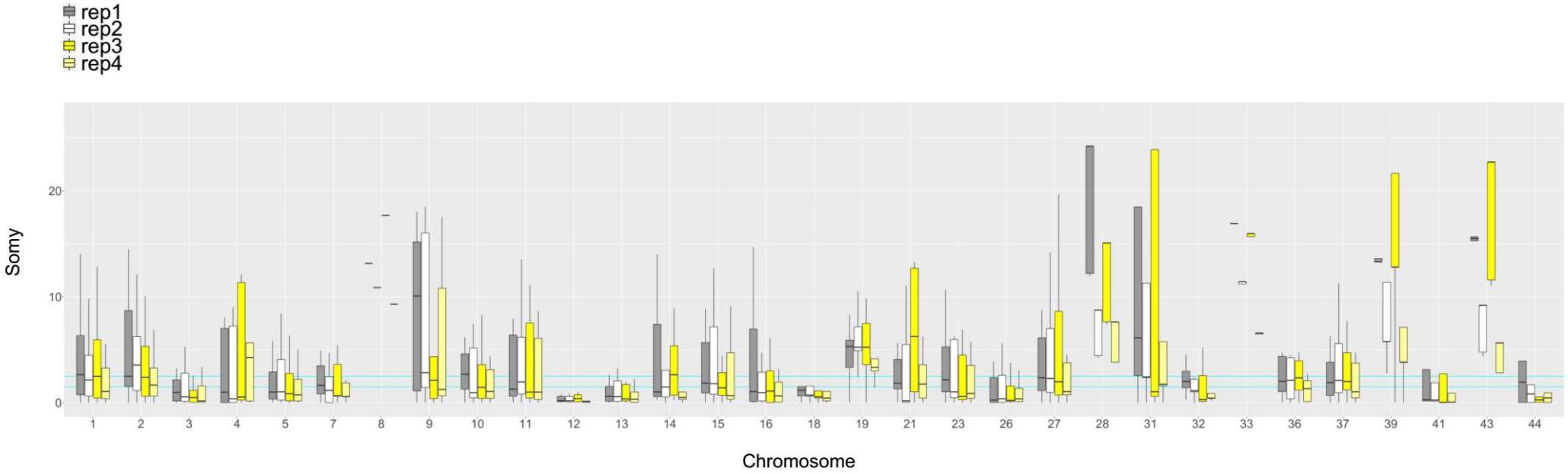
Supplementary Figure 4.8 Final (barcoded) GLST pool size composition measurement using microfluidic electrophoresis. The figure plots fragment sizes (calculated based on migration times relative to those of standards) and fluorescence intensity (FU) of the final GLST pool measured with the Agilent Bioanalyzer 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.



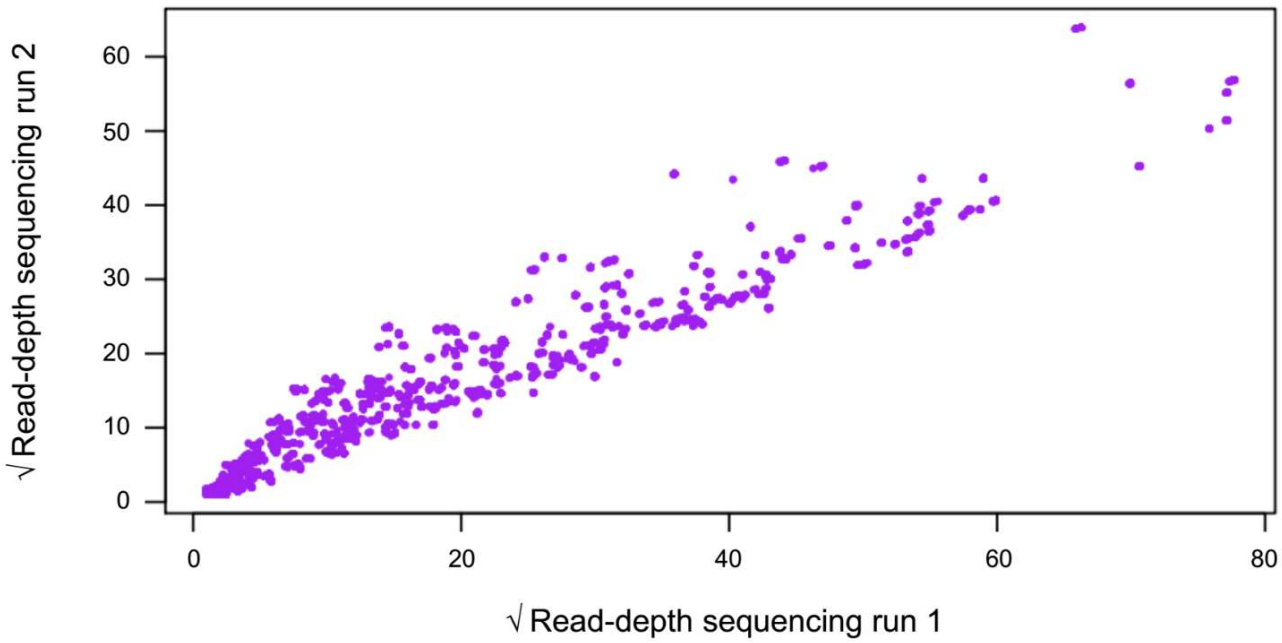
Supplementary Figure 4.9 Quality scores at previously identified vs. unidentified variant sites. The GLST primer panel was designed based on single-nucleotide polymorphisms (SNPs) in Ecuadorian TcI clones. It was applied, however, to samples from distant geographic locations as well as to non-TcI clones. 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 Toolkit[389]. 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 4.10 Real-time PCR for GLST sample selection and sensitivity estimation. 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. 4.4). 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, showing 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 in qPCR.



Supplementary Figure 4.11 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 Chapters 2 and 3 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_2795_CL2 did not show chromosomal amplifications in whole-genome analysis (see Chapter 2). 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 4.12 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 Tcl-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 4.1 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	*Rhodnius pallescens*	Santander Dep.	Lebrija	-8141577.9370	790936.6092	Sylvatic	2015	1
COL78	*Rhodnius sp.*	Santander Dep.	Lebrija	-8141577.9370	790936.6092	Sylvatic	2015	2
COL133	*Rhodnius prolixus*	Casanare Dep.	Paz de Ariporo	-7993997.4220	653950.4247	Domestic	2016	2
COL135	*Rhodnius prolixus*	Casanare Dep.	Paz de Ariporo	-7993997.4220	653950.4247	Domestic	2016	2
COL154	*Rhodnius prolixus*	Casanare Dep.	Tamara	-8024081.7980	648298.0468	Domestic	2016	2
COL155	*Rhodnius prolixus*	Casanare Dep.	Tamara	-8024081.7980	648298.0468	Domestic	2016	2
COL169	*Rhodnius prolixus*	Casanare Dep.	Pore	-8005271.3760	636869.6421	Domestic	2016	1
COL253	*Panstrongylus geniculatus*	Casanare Dep.	Paz de Ariporo	-7993997.4220	653950.4247	Domestic	2016	1
COL319	*Rhodnius prolixus*	Arauca Dep.	Fortul	-7980623.1040	755354.1935	Domestic	2016	2
COL466	*Panstrongylus geniculatus*	Boyacá Dep.	Soata	-8083880.0490	704231.6027	Unknown	2017	3
COL468	*Panstrongylus geniculatus*	Boyacá Dep.	Soata	-8083880.0490	704231.6027	Unknown	2017	3
ECU3	*Rhodnius ecuadoriensis*	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	2
ECU4	*Rhodnius ecuadoriensis*	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	2
ECU8	*Rhodnius ecuadoriensis*	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	1
ECU9	*Rhodnius ecuadoriensis*	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	2
ECU10	*Rhodnius ecuadoriensis*	Loja Province	Bramaderos	-8875849.2150	-453603.4112	Sylvatic	2009	2
ECU36	*Rhodnius ecuadoriensis*	Loja Province	Galápagos	-8832711.9860	-483957.8804	Sylvatic	2009	1
ECU41	*Rhodnius ecuadoriensis*	Loja Province	Guineo	-8899431.9060	-466731.6546	Sylvatic	2009	1
ECU77	*Rhodnius ecuadoriensis*	Loja Province	Jacapo	-8830688.2360	-485500.9341	Sylvatic	2008	1
TBM_2795_CL2	*Panstrongylus chinai*	Loja Province	Bella Maria	-8852271.1950	-466705.6350	Domestic	2009	4
VZ1016B	*Panstrongylus geniculatus*	Met. Caracas	Libertador	-7447967.9080	1167084.6630	Domestic	2016	2
VZ13516	*Panstrongylus geniculatus*	Met. Caracas	Libertador	-7441110.8420	1169154.1140	Domestic	2016	2
VZ35814	*Panstrongylus geniculatus*	Met. Caracas	Libertador	-7450655.1580	1165756.5490	Domestic	2014	2
VZ6616	*Panstrongylus geniculatus*	Met. Caracas	Sucre	-7426686.3980	1163934.1740	Domestic	2016	2
VZ1214D	*Panstrongylus geniculatus*	Met. Caracas	Sucre	-7427396.8230	1166961.1250	Domestic	2014	1
VZ16816	*Panstrongylus geniculatus*	Met. Caracas	Sucre	-7427026.2100	1162328.0720	Domestic	2016	1
VZ17114	*Panstrongylus geniculatus*	Met. Caracas	Sucre	-7426501.1470	1162853.1350	Domestic	2014	1

Supplementary Table 4.2 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'-CAAGCAGAAGACGGCATACGAGAT*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	ACACTGACGACATGGTTCTACAAGAGTTGTGGCATCCTTGTTCTTG	TACGGTAGCAGAGACTTGGTCTAAACGCCTTCACCTTACTCAGACA
TC_LOJ_4	chr11:368075-368194	ACACTGACGACATGGTTCTACAAGGAGGTGAAACGGATGGTAAAGA	TACGGTAGCAGAGACTTGGTCTTGCGAAGAAGAAGATCAAACTCTCTC
TC_LOJ_5	chr1:2082456-2082586	ACACTGACGACATGGTTCTACAAGCTCAAGGGCTGAAATAGACACA	TACGGTAGCAGAGACTTGGTCTCGTTTAGGCTGGAAAGATGGAAGT
TC_LOJ_6	chr12:1011748-1011869	ACACTGACGACATGGTTCTACACCACTCTATCGTCTACGCATCCTC	TACGGTAGCAGAGACTTGGTCTATCATCTTGAGACACATGCCTTGC
TC_LOJ_8	chr5:515822-515951	ACACTGACGACATGGTTCTACAAATGGAGATGGAGGATATGAAGCA	TACGGTAGCAGAGACTTGGTCTTTTAGACCTCATGTTTCCCGTGTC
TC_LOJ_9	chr1:163164-163296	ACACTGACGACATGGTTCTACACGCTGAGTATCAATTTAAGCGTAGCA	TACGGTAGCAGAGACTTGGTCTACCCATATCCGTCATCCCTATTGT
TC_LOJ_10	chr1:1104374-1104501	ACACTGACGACATGGTTCTACATGCCCTTCACATTTATCCCAAGTA	TACGGTAGCAGAGACTTGGTCTAAATAGCATGGAACTCAGCCAGAA
TC_LOJ_11	chr5:995176-995297	ACACTGACGACATGGTTCTACAGCAACTCCACAAACGACTCAGAAC	TACGGTAGCAGAGACTTGGTCTGATGCTGCCATTTCGTCTTTACTC
TC_LOJ_12	chr14:833083-833213	ACACTGACGACATGGTTCTACACTTGTTGCTAAGTGTCCGTGTGTC	TACGGTAGCAGAGACTTGGTCTGCCTTTATATTGATCGGCTCCTCT
TC_LOJ_13	chr23:560603-560743	ACACTGACGACATGGTTCTACAGTCTTTGATTTCTCGTCCGTACCTT	TACGGTAGCAGAGACTTGGTCTTGCATCTTCTACTTTCTCGGAAGC
TC_LOJ_14	chr19:763581-763703	ACACTGACGACATGGTTCTACAAAGATACAAGAGCACGGTACAAAGGA	TACGGTAGCAGAGACTTGGTCTGTGAAGAGGGATGGATCAACATTC
TC_LOJ_15	chr4:1431898-1432017	ACACTGACGACATGGTTCTACAAGGACTATGCTCAAGACGGGATCT	TACGGTAGCAGAGACTTGGTCTCATCAAGTGGACACAACAGCAACT
TC_LOJ_16	chr16:1168122-1168248	ACACTGACGACATGGTTCTACATACAAACATCAACGCAGAACATGC	TACGGTAGCAGAGACTTGGTCTCACACATCCCGTAACTCAATGGTA
TC_LOJ_19	chr43:177414-177556	ACACTGACGACATGGTTCTACACAGTCCTCCAGTTCTCCAAGTGAT	TACGGTAGCAGAGACTTGGTCTGAGATTGTTCTCTCTGTCCCAACG
TC_LOJ_20	chr26:294140-294261	ACACTGACGACATGGTTCTACAGCACAAGAACGGGTGTACCTTCTA	TACGGTAGCAGAGACTTGGTCTTGTGTCGAGGGAATTGATTACTGC
TC_LOJ_23	chr18:690694-690813	ACACTGACGACATGGTTCTACAAAAGAAACTTCGGGTAGCGACAAC	TACGGTAGCAGAGACTTGGTCTCACCACTTCTGCTAGACCACATCC
TC_LOJ_24	chr1:1993894-1994026	ACACTGACGACATGGTTCTACATTCTACACACTCCGCCTTACGTCT	TACGGTAGCAGAGACTTGGTCTGTCTGCAACGACACATAGATTGGA
TC_LOJ_25	chr36:470603-470728	ACACTGACGACATGGTTCTACAGTGGCTCAGAAGCATGATCGTAAT	TACGGTAGCAGAGACTTGGTCTACCCTTGTAGTCTTCGCAGTCCTC
TC_LOJ_26	chr13:433737-433859	ACACTGACGACATGGTTCTACACAATGGTGATGATGAGGTTAAGCA	TACGGTAGCAGAGACTTGGTCTACGTCCAATACACACAAACACACAG
TC_LOJ_27	chr24:269253-269379	ACACTGACGACATGGTTCTACAGGCGATAAGGAAGAATGGAGAGAA	TACGGTAGCAGAGACTTGGTCTGTCATGTGCTTACGAGAGCCGTAG
TC_LOJ_28	chr27:389665-389794	ACACTGACGACATGGTTCTACAACCACTTCACCATTTGTCTGGTATTC	TACGGTAGCAGAGACTTGGTCTTTTAAGATGGCCGCATACAGTGAG
TC_LOJ_29	chr36:451747-451871	ACACTGACGACATGGTTCTACAGTGTGTTTGAGATTGGGCCTGTAT	TACGGTAGCAGAGACTTGGTCTCACATCAAGTACCTCCGTGTACGA
TC_LOJ_30	chr7:1140939-1141071	ACACTGACGACATGGTTCTACAAGTTGATCGTCTTTCTTCCTTGACC	TACGGTAGCAGAGACTTGGTCTAAATGTTCCTGCGTACACCAAGTC
TC_LOJ_32	chr2:120852-120972	ACACTGACGACATGGTTCTACAAAATGATGTACTGCCTGAACTGGAA	TACGGTAGCAGAGACTTGGTCTGTTCTCCGCCGTATTCTCCTCTAC
TC_LOJ_34	chr16:170448-170597	ACACTGACGACATGGTTCTACAGGAAGAAGGCAGACTAAACAGGATG	TACGGTAGCAGAGACTTGGTCTAGCTTGTCACTGCTCACAGAGTTG

Supplementary Table 4.2 (continued)

TC_LOJ_35	chr26:125032-125153	ACACTGACGACATGGTTCTACAGTACGCTACACTGCGAGAGGAATG	TACGGTAGCAGAGACTTGGTCTGCACAACTGAGATTATAGCCAACTCC
TC_LOJ_36	chr5:1012765-1012911	ACACTGACGACATGGTTCTACATCCGTCCCTGTTGTCTTCTCAATA	TACGGTAGCAGAGACTTGGTCTTGAGCAAAGTGTCCTTATTCTTCAGC
TC_LOJ_37	chr1:2889409-2889535	ACACTGACGACATGGTTCTACACAGAGTTCCACGGATAAGTCGTCA	TACGGTAGCAGAGACTTGGTCTACACACTTCCAGATCACTACGAAGC
TC_LOJ_38	chr21:465093-465213	ACACTGACGACATGGTTCTACATGGTTGTAGTCCGTGATCTCTGGT	TACGGTAGCAGAGACTTGGTCTATAACTGGTTCGGGAAGGAAGAAA
TC_LOJ_39	chr1:1160205-1160334	ACACTGACGACATGGTTCTACAACGTCACATTTGTACTGCGAGAGG	TACGGTAGCAGAGACTTGGTCTCCCTTACTTGTCTCCGACTCATTCT
TC_LOJ_40	chr7:1138368-1138496	ACACTGACGACATGGTTCTACAGTCCAAGCCGTTGTCTCTCAATAC	TACGGTAGCAGAGACTTGGTCTTGTTCGTTGTGGTGGAATGTGTAG
TC_LOJ_41	chr1:2693345-2693466	ACACTGACGACATGGTTCTACATGGCTGGTGCAAATGTACTCATATC	TACGGTAGCAGAGACTTGGTCTTAAACAAGTGTGCCATTGCGTATC
TC_LOJ_42	chr10:1016129-1016269	ACACTGACGACATGGTTCTACATACGACTCCCTTTCCACATACGAC	TACGGTAGCAGAGACTTGGTCTATATTGAGCCGAAACACGAAGTACA
TC_LOJ_43	chr1:1956698-1956821	ACACTGACGACATGGTTCTACAGCTCTCATGGGTGGTAGAAGCTAA	TACGGTAGCAGAGACTTGGTCTCCCACTGTCATTATTCAAACTGCTC
TC_LOJ_44	chr3:173883-174019	ACACTGACGACATGGTTCTACAGTCATCATTCTCGGAAACAAAGTAGG	TACGGTAGCAGAGACTTGGTCTGTGTCCATCAGCTCTACAATGCAC
TC_LOJ_45	chr3:174152-174277	ACACTGACGACATGGTTCTACAAGTACGCCACACGACAGTTCAGTT	TACGGTAGCAGAGACTTGGTCTTGAGTAGTTGTGCCCTTCGATGTA
TC_LOJ_46	chr1:1833807-1833948	ACACTGACGACATGGTTCTACAATTCGTGTCATTAGCAGCAGCAAC	TACGGTAGCAGAGACTTGGTCTGACGGTAAATTCTGCGTACACTGC
TC_LOJ_47	chr14:844524-844671	ACACTGACGACATGGTTCTACAAGCAATTCACGGAGTTCACAGATG	TACGGTAGCAGAGACTTGGTCTAGGAGTCACCACAGAAGTCAGAGC
TC_LOJ_48	chr3:1058072-1058196	ACACTGACGACATGGTTCTACAGATAGCACAAACAAGCCAAATGGT	TACGGTAGCAGAGACTTGGTCTGAAAGATACGCCTTCCAATCATCA
TC_LOJ_51	chr12:596775-596914	ACACTGACGACATGGTTCTACAGATTGACATTACGGCGATTCAGAG	TACGGTAGCAGAGACTTGGTCTTGTGGATCTTCTGCCATGATATTG
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	ACACTGACGACATGGTTCTACAATACTCCTCTGCATTCACCTCCTG	TACGGTAGCAGAGACTTGGTCTGGTTGGTATAACCGAAGGAAATATGG
TC_LOJ_57	chr37:454539-454662	ACACTGACGACATGGTTCTACAGTACGTGAAACGCCCTGACTTTAC	TACGGTAGCAGAGACTTGGTCTTGGATGAACCTCCTTGTAGATGTTG
TC_LOJ_58	chr15:395493-395614	ACACTGACGACATGGTTCTACACTTTGTGACCACCTCCTTGTTATTG	TACGGTAGCAGAGACTTGGTCTAGGTATTTGGCATGTTTGATCTGC
TC_LOJ_59	chr2:856618-856737	ACACTGACGACATGGTTCTACAGCCCGGTTCACAACTTTAGTAGAAA	TACGGTAGCAGAGACTTGGTCTCACCAACACAGCTACGACAACAAC
TC_LOJ_60	chr26:139346-139478	ACACTGACGACATGGTTCTACAGATTATGGTGGTGGTTTCAACACG	TACGGTAGCAGAGACTTGGTCTAAAGTGAATGGCAAATCCTAAGACG
TC_LOJ_61	chr1:1992854-1992995	ACACTGACGACATGGTTCTACAATCTGTTGAGGATGACCGAACACT	TACGGTAGCAGAGACTTGGTCTGAGAAATATCGCCGCACCTTCTAC
TC_LOJ_62	chr1:305886-306012	ACACTGACGACATGGTTCTACATACTCAGGCGTAGAAACAGGCTCA	TACGGTAGCAGAGACTTGGTCTTACCTCCGCTTATCAATGTTGTCC
TC_LOJ_63	chr26:303994-304113	ACACTGACGACATGGTTCTACACATGACAAGCATAAATACAGCGAGAG	TACGGTAGCAGAGACTTGGTCTGAAGGTACAAGCAAGGAGCCATCT
TC_LOJ_64	chr14:889253-889389	ACACTGACGACATGGTTCTACACTTCCCAGACTCATCTTTCTGCTG	TACGGTAGCAGAGACTTGGTCTATTCCCGACTACTTTGGCATGATT
TC_LOJ_67	chr10:143080-143202	ACACTGACGACATGGTTCTACACACTAACTGGGTCAAAGTGTTCTTGC	TACGGTAGCAGAGACTTGGTCTAGCAACTGCGGATACTTGGTCTTC
TC_LOJ_69	chr2:446791-446914	ACACTGACGACATGGTTCTACAGGTAGAAGGTACTCTCATCGGTAGCA	TACGGTAGCAGAGACTTGGTCTCAGAAACAGCTCGCCAGAAATAAA
TC_LOJ_70	chr32:839405-839556	ACACTGACGACATGGTTCTACAGGTGCGTACTGTCTTGGAAGGTTT	TACGGTAGCAGAGACTTGGTCTGTTGACGATCCACGGAAAGATATG
TC_LOJ_71	chr7:179338-179460	ACACTGACGACATGGTTCTACAATGGGAGATCGGGAGTACATGAAG	TACGGTAGCAGAGACTTGGTCTTGAAGAGCCAAATGGGACACTAAT

Supplementary Table 4.2 (continued)

TC_LOJ_74	chr1:1413411-1413530	ACACTGACGACATGGTTCTACACAAGATTGTTCCACTGACGAAGACA	TACGGTAGCAGAGACTTGGTCTTTTGAGAGCGTGAAGGAGTACACA
TC_LOJ_75	chr23:504383-504519	ACACTGACGACATGGTTCTACACTTCATCATCTATGCTCCGACGAC	TACGGTAGCAGAGACTTGGTCTTCTGAATGACTGGTTGAAAGACGA
TC_LOJ_76	chr23:505516-505635	ACACTGACGACATGGTTCTACAGTGGACCCAAATGTACTCAGCAAC	TACGGTAGCAGAGACTTGGTCTGAACCTAAGAAACGAAGAACCCTCA
TC_LOJ_80	chr1:2018618-2018750	ACACTGACGACATGGTTCTACAAGTGGACATGGTGACGAAGATGAG	TACGGTAGCAGAGACTTGGTCTGTAGTGCTTCAAACCGCTCAAGAA
TC_LOJ_81	chr37:132370-132499	ACACTGACGACATGGTTCTACAACCGGATGTATTCCTCTCGTGGTA	TACGGTAGCAGAGACTTGGTCTCATGCACTTATCGTCGTCACTTTC
TC_LOJ_82	chr13:741015-741134	ACACTGACGACATGGTTCTACACACAAACCGCTTAGACCCTGAAGT	TACGGTAGCAGAGACTTGGTCTCCAGAAGAAACAATCAATCAACAGC
TC_LOJ_85	chr1:351420-351541	ACACTGACGACATGGTTCTACAAGACTCAATCGCCTTCACGACATA	TACGGTAGCAGAGACTTGGTCTCAGAGGTGTTTATGAGCAAGTACCG
TC_LOJ_86	chr18:746701-746824	ACACTGACGACATGGTTCTACAACCCACTCCAGTAGCATTTCTTCC	TACGGTAGCAGAGACTTGGTCTTTAACTATGGCAATGAGGCAGAGC
TC_LOJ_87	chr37:464692-464819	ACACTGACGACATGGTTCTACACAGATGCTGCCTTGACAGAGATGTA	TACGGTAGCAGAGACTTGGTCTACGAGTGTAGAAGCGAAGATGCTG
TC_LOJ_88	chr16:213322-213477	ACACTGACGACATGGTTCTACAGTAAATAGACACAAGCCATTCCCATC	TACGGTAGCAGAGACTTGGTCTTACTATCACTACCGTGGGCGTCAG
TC_LOJ_89	chr2:121560-121715	ACACTGACGACATGGTTCTACACTCATACCCTTGCTTTGTCATGCT	TACGGTAGCAGAGACTTGGTCTGTTCAGGAGACGGACCACTAGGTT
TC_LOJ_91	chr12:107750-107877	ACACTGACGACATGGTTCTACAGAATGACAACAATGCCCTTTCTTC	TACGGTAGCAGAGACTTGGTCTGTATCTCCATCCATTTCCCAGTGC
TC_LOJ_93	chr27:329031-329151	ACACTGACGACATGGTTCTACATCGTAAAGGTATTGGGCATATTCG	TACGGTAGCAGAGACTTGGTCTCCAGGATCATTCAGCTTAGTCCAG
TC_LOJ_97	chr26:38201-38343	ACACTGACGACATGGTTCTACATTTGAAGAGAAGATGGCCCTGAGT	TACGGTAGCAGAGACTTGGTCTTTGAAGAAAGGATCTGCCTCGTAA
TC_LOJ_99	chr33:297174-297306	ACACTGACGACATGGTTCTACACAAGTTCCTGTTGGACGTGGTAGT	TACGGTAGCAGAGACTTGGTCTAATGTACGCAAGGAGCGACTAGAG
TC_LOJ_100	chr26:479107-479233	ACACTGACGACATGGTTCTACATATTATTTACGAAACGGCGGAGGA	TACGGTAGCAGAGACTTGGTCTAGGAGATGGCTCACTCACTTGAAC
TC_LOJ_102	chr11:853646-853766	ACACTGACGACATGGTTCTACAAGAACAGGAAGTTTGTGACGGTTG	TACGGTAGCAGAGACTTGGTCTATCACCTCTGAAAGAATCGACTGC
TC_LOJ_103	chr13:783091-783210	ACACTGACGACATGGTTCTACAGTACACCCGTCCTTGCAGTATGATT	TACGGTAGCAGAGACTTGGTCTCGCTGAGTTCACGAAGTTATGCTT
TC_LOJ_104	chr15:807734-807870	ACACTGACGACATGGTTCTACACAAGTTCGCAATGTAGGAAAGCTG	TACGGTAGCAGAGACTTGGTCTTATCATGGTGGTCGATGCTGAATA
TC_LOJ_107	chr2:160058-160182	ACACTGACGACATGGTTCTACAGTCATACCTTACCAAACGGCACAG	TACGGTAGCAGAGACTTGGTCTATGTGAACAACCGTACTGGAGGTG
TC_LOJ_108	chr13:664297-664421	ACACTGACGACATGGTTCTACATATCTGTGGTGGCTGTAGATGGTG	TACGGTAGCAGAGACTTGGTCTCGACGACAACAAGGAAGAAGAGGTA
TC_LOJ_109	chr26:419336-419479	ACACTGACGACATGGTTCTACACTTTCGGTGTTACGGTGTACTTCAG	TACGGTAGCAGAGACTTGGTCTTCACTGTTTACAACTACGGCCAGA
TC_LOJ_111	chr41:288290-288430	ACACTGACGACATGGTTCTACACCACGCCACCAGTAACGATAATAA	TACGGTAGCAGAGACTTGGTCTGAAGAAGTGGTACTCTCCCGATCC
TC_LOJ_114	chr5:168922-169061	ACACTGACGACATGGTTCTACATTAGAAACCGTGTAGAGACTTGTCAGC	TACGGTAGCAGAGACTTGGTCTATTACCCTGCACCAAGACACATTC
TC_LOJ_116	chr26:336772-336902	ACACTGACGACATGGTTCTACAGCTGTCTCCAAGAGTCGCAGAATA	TACGGTAGCAGAGACTTGGTCTCATGGATTTCTTTCCAGTGCTTTG
TC_LOJ_117	chr3:965641-965793	ACACTGACGACATGGTTCTACATCCAATCTCTTATCTTTCAGGAGAACG	TACGGTAGCAGAGACTTGGTCTCATACTCAAACGAGGCACGAATCT
TC_LOJ_118	chr15:398374-398497	ACACTGACGACATGGTTCTACACCACAAGTAGGCTGAACCACAAAT	TACGGTAGCAGAGACTTGGTCTGTCAAGCCCTTCGTATCCCTGTTA
TC_LOJ_119	chr1:2137512-2137631	ACACTGACGACATGGTTCTACAGAATCATCAGAGGGTCATTTGCAC	TACGGTAGCAGAGACTTGGTCTAGTACACAACAAAGTTATCGCGGATG
TC_LOJ_120	chr3:196127-196261	ACACTGACGACATGGTTCTACATCATCCTCATCTTCTGGTGGTGAT	TACGGTAGCAGAGACTTGGTCTTGGACTCTCACTTCTGTATCTACTTTGTTG
TC_LOJ_121	chr27:93351-93474	ACACTGACGACATGGTTCTACAACTGCGTTGTATAGCCGAATCACT	TACGGTAGCAGAGACTTGGTCTGACAGGAACACCAAATGTACTGTGAA
TC_LOJ_122	chr36:377593-377718	ACACTGACGACATGGTTCTACACTTTCCTGGGTTCGTTGGTTTAAG	TACGGTAGCAGAGACTTGGTCTCAGGTGTTCCTCGTCAAGCTGTAAT

Supplementary Table 4.2 (continued)

TC_LOJ_124	chr10:933564-933686	ACACTGACGACATGGTTCTACATGCAAATACAGAAGATGAGCTACGC	TACGGTAGCAGAGACTTGGTCTTGATTATGAGGAGGAGGATGCAGT
TC_LOJ_125	chr21:539837-539959	ACACTGACGACATGGTTCTACAAAATCTCAGCTACAACAACATCTCTGG	TACGGTAGCAGAGACTTGGTCTTCATCCTTTCCATCGTTCTCACTT
TC_LOJ_126	chr15:908929-909068	ACACTGACGACATGGTTCTACAGGCCTTCTCACTAACTGTCGATCTG	TACGGTAGCAGAGACTTGGTCTACCTTCTTATCACGGAAGAGTATCAGG
TC_LOJ_128	chr11:775649-775772	ACACTGACGACATGGTTCTACAGAAAGAAGCTGAAGAATGGGCAAA	TACGGTAGCAGAGACTTGGTCTGTTGATCCTGGCAATTACACTCGT
TC_LOJ_129	chr18:115349-115471	ACACTGACGACATGGTTCTACAGTGACTTGGCGATTATGATTCGTT	TACGGTAGCAGAGACTTGGTCTCGTTTGTCTTCTCATCCTTCTTCG
TC_LOJ_130	chr9:601749-601872	ACACTGACGACATGGTTCTACATCCCGTTACATCCAATACATCCAA	TACGGTAGCAGAGACTTGGTCTTGCATACACAACAGAGCTAAGTGTCG
TC_LOJ_131	chr9:601909-602028	ACACTGACGACATGGTTCTACAACAAGCAATCCAATTACAACCACAG	TACGGTAGCAGAGACTTGGTCTATTAAAGAAGGTCGCGGCAGTAGA
TC_LOJ_136	chr23:522688-522812	ACACTGACGACATGGTTCTACATTTCAAGCTGCGACTTAATCAACG	TACGGTAGCAGAGACTTGGTCTGATGGAAATGCTTCTTGCACAGTC
TC_LOJ_137	chr16:889485-889604	ACACTGACGACATGGTTCTACACATTTCTGCTGCTTCCTTTGAGAA	TACGGTAGCAGAGACTTGGTCTTCTGATGTTGATCTCTCTTTAACCTACCG
TC_LOJ_138	chr5:1116604-1116723	ACACTGACGACATGGTTCTACACATTTCACCAGAAGTGACAGCAAC	TACGGTAGCAGAGACTTGGTCTGATGAGGGAGAAGCGAATTTGAAC
TC_LOJ_140	chr19:251999-252118	ACACTGACGACATGGTTCTACACCCTCACCTCAATCATATCCACAC	TACGGTAGCAGAGACTTGGTCTGGGACAAGTACGGGAACAGAATAGA
TC_LOJ_141	chr37:317244-317399	ACACTGACGACATGGTTCTACAATTGTGAGAGGATGGGTTCAAATG	TACGGTAGCAGAGACTTGGTCTCCAGTGCATACTTCTGTGTTATGGTAGA
TC_LOJ_142	chr2:327727-327846	ACACTGACGACATGGTTCTACAATGCGGGAGTGTTGTGCATTAGTAT	TACGGTAGCAGAGACTTGGTCTACGGAATACGGGTGGAATAAGAAA
TC_LOJ_144	chr11:235518-235637	ACACTGACGACATGGTTCTACAACGCAGTTGGTCGAGAATTGTATC	TACGGTAGCAGAGACTTGGTCTGAAGGAGAGGTGGTGCAGCTTATC
TC_LOJ_145	chr6:23502-23628	ACACTGACGACATGGTTCTACATTGGCATAAAGGTACGAATCATGG	TACGGTAGCAGAGACTTGGTCTGAACTCACGACCCTGAATAAGACG
TC_LOJ_146	chr27:232849-232974	ACACTGACGACATGGTTCTACACTCAGTATGAACTCCGCTTCCTGT	TACGGTAGCAGAGACTTGGTCTGGATATGTGCTCAAAGTGCCTTGT
TC_LOJ_147	chr4:1219111-1219233	ACACTGACGACATGGTTCTACAAAGCTGAATAGATCGCACAAGCTC	TACGGTAGCAGAGACTTGGTCTTATGCCCTATCCGTGTTTCTTACG
TC_LOJ_152	chr19:553417-553540	ACACTGACGACATGGTTCTACACATAAGGGCAGTGTCATCAACAAA	TACGGTAGCAGAGACTTGGTCTGTATTGCTGGTTGGTTCTCTTCCA
TC_LOJ_154	chr37:156377-156496	ACACTGACGACATGGTTCTACAGTAAGGACCACAAGAGGGAAATGG	TACGGTAGCAGAGACTTGGTCTGCAGAGTAGACAGCATGGAGTGTG
TC_LOJ_156	chr5:627080-627199	ACACTGACGACATGGTTCTACATGGACTACGAGAAGGTTTCATACGAC	TACGGTAGCAGAGACTTGGTCTGCTGTGGAAATGTTGTGATCCTGT
TC_LOJ_157	chr1:1963178-1963304	ACACTGACGACATGGTTCTACATAGAAGAGCGTGTGAAGACTGTGG	TACGGTAGCAGAGACTTGGTCTATGACAACCGCGTCACTTGAATAC
TC_LOJ_158	chr1:1964699-1964825	ACACTGACGACATGGTTCTACACTACACGCATTGTGAGAAACTTGG	TACGGTAGCAGAGACTTGGTCTTGAATTTGTCTGGGATGTGGAAAC
TC_LOJ_159	chr1:1998360-1998510	ACACTGACGACATGGTTCTACAACCGTGCTACTTTCTTCCTTTGGT	TACGGTAGCAGAGACTTGGTCTAATCTTCCTCAATCTCCCTGCTGT
TC_LOJ_160	chr16:738527-738679	ACACTGACGACATGGTTCTACACAGCCACTGTTCAGATCCACAAGT	TACGGTAGCAGAGACTTGGTCTGGCACAAGACCATCAAAGTAGGAC
TC_LOJ_161	chr43:149662-149786	ACACTGACGACATGGTTCTACATGTACCTTTCTGCTTTGTCTTCTTCC	TACGGTAGCAGAGACTTGGTCTTGATGACTATCGCTCCATTCTTCC
TC_LOJ_162	chr16:189968-190097	ACACTGACGACATGGTTCTACAGCTTTGGAGTAGAGCAGATTTGGA	TACGGTAGCAGAGACTTGGTCTCCGAGTTACATTTCTTTGCCTTTG
TC_LOJ_163	chr18:523652-523773	ACACTGACGACATGGTTCTACAGATCGCGTTGTAAGCAAATTCAAG	TACGGTAGCAGAGACTTGGTCTGGCGTAAAGGGCAACTCAAAGTAT
TC_LOJ_165	chr3:169504-169625	ACACTGACGACATGGTTCTACACACGAAAGTCAAACTCCTCCACAA	TACGGTAGCAGAGACTTGGTCTGGTAAATACACGTCCACCGACCTT
TC_LOJ_166	chr3:169646-169792	ACACTGACGACATGGTTCTACAGGCAACGTGGTATGGAATGATAAC	TACGGTAGCAGAGACTTGGTCTTCTGCTCACACAGGACTGAATCTC
TC_LOJ_168	chr28:364521-364659	ACACTGACGACATGGTTCTACACTCGTGGAAGTTTAGTGCTGATCG	TACGGTAGCAGAGACTTGGTCTCGATGATAAAGAAGTCTCCGTACCC
TC_LOJ_169	chr11:721966-722086	ACACTGACGACATGGTTCTACAATGAAACACGTATGCACGATATGC	TACGGTAGCAGAGACTTGGTCTGGCGCTAAATCTGTACGAATACCA

Supplementary Table 4.2 (continued)

TC_LOJ_170	chr36:416713-416839	ACACTGACGACATGGTTCTACAGGGAGTACGAGTTTGCAGAGAAGA	TACGGTAGCAGAGACTTGGTCTAGAGGGTTGACATAAGGATGCAGA
TC_LOJ_171	chr2:854454-854583	ACACTGACGACATGGTTCTACAAGCAAGGGCAGTCACAAAGTAACA	TACGGTAGCAGAGACTTGGTCTACTGTGGGTGATACAGGCAAAGAC
TC_LOJ_173	chr19:264153-264279	ACACTGACGACATGGTTCTACACATTGAGAACCACGACTGGCTATT	TACGGTAGCAGAGACTTGGTCTGGACTATGAGATCGACAAGGAGTTTG
TC_LOJ_174	chr18:456154-456275	ACACTGACGACATGGTTCTACAATATCATGGGACTTGCCGGATTAC	TACGGTAGCAGAGACTTGGTCTCAATGTCTGGTTTGGAGGAAGAAG
TC_LOJ_175	chr13:608121-608257	ACACTGACGACATGGTTCTACAACTGACATGGATCATAGCCAATCG	TACGGTAGCAGAGACTTGGTCTCGATAAAGGAACCCAACAAGAACC
TC_LOJ_177	chr7:1112127-1112263	ACACTGACGACATGGTTCTACACTTTGAGAGCTTTGCATCCTTCAC	TACGGTAGCAGAGACTTGGTCTCCGGGACGAGTACACATATACCAA
TC_LOJ_178	chr10:265161-265291	ACACTGACGACATGGTTCTACAGGTATGAGCATCGCCTTATTGATG	TACGGTAGCAGAGACTTGGTCTAAGAGAACCAAATCCCTGAGCAAC
TC_LOJ_180	chr8:851024-851146	ACACTGACGACATGGTTCTACAGACGATGAGGAGTTGGAGGATGTA	TACGGTAGCAGAGACTTGGTCTAGTGTGGCGATAGGTGATTGTGAT
TC_LOJ_181	chr7:987164-987292	ACACTGACGACATGGTTCTACATAGATGTTTGGTCCCATTTGAAGG	TACGGTAGCAGAGACTTGGTCTTGATACCGTCACTATTACCGCTAGAAA
TC_LOJ_182	chr15:497344-497472	ACACTGACGACATGGTTCTACATGTCCAAGACCTTCACATAGTCCA	TACGGTAGCAGAGACTTGGTCTTGGTTACTTTCCAGACAAGGGATG
TC_LOJ_184	chr37:138690-138820	ACACTGACGACATGGTTCTACAAGCTTGGCCTTCAACACATCATTA	TACGGTAGCAGAGACTTGGTCTGCGTCATACTCCCTCACATATCCA
TC_LOJ_185	chr27:387192-387314	ACACTGACGACATGGTTCTACAGGGTGATAGATGCTGTTGCTGAAT	TACGGTAGCAGAGACTTGGTCTTGAGTTTAATGGACCCGAAGGAAC
TC_LOJ_187	chr15:795497-795621	ACACTGACGACATGGTTCTACAGACAAACATTCGACCTTCATCTTCTG	TACGGTAGCAGAGACTTGGTCTTGGTATTTGAGGATCATTCCAGTCA
TC_LOJ_188	chr1:2220221-2220341	ACACTGACGACATGGTTCTACACCAGGTTGTTGGTTGTTATGTGGT	TACGGTAGCAGAGACTTGGTCTGCGGAGATTCACGAAATAGAGGAA
TC_LOJ_191	chr5:703969-704096	ACACTGACGACATGGTTCTACACTATTGGATGGGAACGTGGTACAG	TACGGTAGCAGAGACTTGGTCTGCACAATCTCTGTTGTAAGACTAAACTCCT
TC_LOJ_192	chr37:447759-447878	ACACTGACGACATGGTTCTACACGTATCAAACAGGGCTGGAGACTT	TACGGTAGCAGAGACTTGGTCTATCAAGCTGCAAGAAGAGAACATCC
TC_LOJ_195	chr27:40705-40826	ACACTGACGACATGGTTCTACAATGTTTCCTTGCATGAGTTTGTGG	TACGGTAGCAGAGACTTGGTCTGGAGTCGCCGTAGTATTCCCTTATG
TC_LOJ_197	chr41:298702-298834	ACACTGACGACATGGTTCTACAATTGGGACGGTAGAGCATGTAAGG	TACGGTAGCAGAGACTTGGTCTGCCTGAGTTCCTCCAGTCTTTCTT
TC_LOJ_200	chr37:173415-173536	ACACTGACGACATGGTTCTACACACGAAACTGCCAATGATGACTCT	TACGGTAGCAGAGACTTGGTCTCACCTCCGTCTTTCTTCTCCTTCT
TC_LOJ_201	chr32:855499-855637	ACACTGACGACATGGTTCTACAAAGAGGCGTGTAAGAAGTATGTGGAG	TACGGTAGCAGAGACTTGGTCTTGCAAGTAGTCAGCAATGTCCAGT
TC_LOJ_203	chr25:64845-64984	ACACTGACGACATGGTTCTACAACGCGGATACTAGGGAACATGAGT	TACGGTAGCAGAGACTTGGTCTTTGAGCAGAATACCAAAGCAGTTGT
TC_LOJ_204	chr9:194610-194758	ACACTGACGACATGGTTCTACACTGTTCAAAGTCCATTGTGCTATCC	TACGGTAGCAGAGACTTGGTCTATGACTGCAAGGTATTCCGCTTCT
TC_LOJ_205	chr7:1037003-1037155	ACACTGACGACATGGTTCTACAACAGGGCTTCAGGTGGACATTATT	TACGGTAGCAGAGACTTGGTCTGGTTAAAGGTCGTGGTTGACACAT
TC_LOJ_206	chr19:762223-762346	ACACTGACGACATGGTTCTACAAGCCTTCCCTTTCTACTGGTGGTA	TACGGTAGCAGAGACTTGGTCTTCTGATTTCATACACGTTGCTCCTC
TC_LOJ_209	chr1:2005883-2006014	ACACTGACGACATGGTTCTACATCTTTGAAGGTTCTGGTGTTGGTT	TACGGTAGCAGAGACTTGGTCTTCTCAGGGACGAGGAGACATATAAGA
TC_LOJ_211	chr2:916287-916407	ACACTGACGACATGGTTCTACACTTGATAAACTCTGCGGCTTCCTC	TACGGTAGCAGAGACTTGGTCTCAATGGTACGAACATGATTGACTGTG
TC_LOJ_212	chr44:285730-285879	ACACTGACGACATGGTTCTACAGCTGTCCATATCCGCATCTTCTAA	TACGGTAGCAGAGACTTGGTCTATGTCGTTTCCAAATCAGCACAAC
TC_LOJ_213	chr32:839358-839478	ACACTGACGACATGGTTCTACAGGTGACAAACCCATTCAGCTTACA	TACGGTAGCAGAGACTTGGTCTTACAGCGCCAATCAAATCCACTAC
TC_LOJ_214	chr11:849661-849797	ACACTGACGACATGGTTCTACATTACTACATTGGTGGCGAGACAAAC	TACGGTAGCAGAGACTTGGTCTTCAGACGAAACAGATAGCTCGTGA
TC_LOJ_215	chr10:1052122-1052245	ACACTGACGACATGGTTCTACACAGAGTTCTACAAGGAAGATCGACAAA	TACGGTAGCAGAGACTTGGTCTTTAATGATGGGTGGAAGTGAGAGG
TC_LOJ_217	chr1:2773733-2773861	ACACTGACGACATGGTTCTACAAAACTTATGGCGTACAACAGGGAGT	TACGGTAGCAGAGACTTGGTCTCGATAACGACGATGAAGATGATGA

Supplementary Table 4.2 (continued)

TC_LOJ_219	chr26:38066-38187	ACACTGACGACATGGTTCTACAGTTGATGTGGATAGGCTTGACTACTTTC	TACGGTAGCAGAGACTTGGTCTTCACCTTCGTAGCACAATACCTTACA
TC_LOJ_220	chr14:923562-923682	ACACTGACGACATGGTTCTACATCGGGTAAATGTCTAACGGAGAAA	TACGGTAGCAGAGACTTGGTCTCCAGATCCAGTGATTCGTCTTGTT
TC_LOJ_221	chr11:868950-869070	ACACTGACGACATGGTTCTACAGCTTCACAGCTATCGAGGTGTATTG	TACGGTAGCAGAGACTTGGTCTCCAGGAGTTTAGTTACAACAGACGAGA
TC_LOJ_223	chr27:96137-96258	ACACTGACGACATGGTTCTACACAAGCGCACCCTAATAAGAAATTG	TACGGTAGCAGAGACTTGGTCTCAACAAAGAGCTTCAAATGGTGTG
TC_LOJ_224	chr1:2775484-2775623	ACACTGACGACATGGTTCTACAGGTGTGTACGGATGACTGCTACTTACTT	TACGGTAGCAGAGACTTGGTCTCAACAAGGACAAAGACAACCACAA
TC_LOJ_225	chr15:246311-246435	ACACTGACGACATGGTTCTACACGTGAAAGATACGGCTGACACATA	TACGGTAGCAGAGACTTGGTCTGTAGTGCGTGTTGCTCCTGTTGTT
TC_LOJ_227	chr27:116142-116263	ACACTGACGACATGGTTCTACAATGAGGAGGAGGAGAAATGGAAAC	TACGGTAGCAGAGACTTGGTCTGTCGATGACACAGTCCAGACACTC
TC_LOJ_228	chr5:1147485-1147616	ACACTGACGACATGGTTCTACAACAGTGCAGTCGTACTTTCGCATT	TACGGTAGCAGAGACTTGGTCTTGTTGACTACTTTGACGGAAATCGT
TC_LOJ_229	chr5:1148049-1148168	ACACTGACGACATGGTTCTACAAGTGGCTTGGCAGATTTCTTCTGT	TACGGTAGCAGAGACTTGGTCTTGACAGTTTAGAGAGCGTTGTAGTGAAAG
TC_LOJ_230	chr15:926778-926915	ACACTGACGACATGGTTCTACAATTCTGCCTGCGACAGTAGTTCTC	TACGGTAGCAGAGACTTGGTCTCCATTCTTCGTGAAATTGAGGTTG
TC_LOJ_231	chr1:2138077-2138196	ACACTGACGACATGGTTCTACAGGCAGACTCCAGATACTGACGAAT	TACGGTAGCAGAGACTTGGTCTCCACAACTCCTTGACGACTTTCTT
TC_LOJ_232	chr5:191326-191447	ACACTGACGACATGGTTCTACAACATCCTGACCCTTGGCTTTAGAC	TACGGTAGCAGAGACTTGGTCTGGTTAGAGAGAACATTACGACGGAGA
TC_LOJ_234	chr10:715504-715626	ACACTGACGACATGGTTCTACAAGTAAGCCTGTTGCTTTGGAAACTC	TACGGTAGCAGAGACTTGGTCTTCAACCCAGACGAAAGTCTAGTGG
TC_LOJ_235	chr15:197505-197635	ACACTGACGACATGGTTCTACATCGTCAATTTCCCGTAGGATACTTT	TACGGTAGCAGAGACTTGGTCTCAGGAGGAGGGTGAACTGATAATG
TC_LOJ_236	chr11:235245-235379	ACACTGACGACATGGTTCTACAATCTTTACCATGCACCTCCACAAC	TACGGTAGCAGAGACTTGGTCTGGTCTCACCACGTATCACGAGAAG
TC_LOJ_237	chr9:134209-134328	ACACTGACGACATGGTTCTACACTCTTCACGCCAATACATTCCTTG	TACGGTAGCAGAGACTTGGTCTCCAGCTACAACTGCAAACAAATACAC
TC_LOJ_238	chr21:322787-322911	ACACTGACGACATGGTTCTACATCAGGGTAGATTCATCAGGCAGAG	TACGGTAGCAGAGACTTGGTCTTATCAACAATGCTCGACACCCACT
TC_LOJ_239	chr44:237246-237373	ACACTGACGACATGGTTCTACAATTTATGCCCGCAAACCAGATAAC	TACGGTAGCAGAGACTTGGTCTCGAGGCAATTCGTATAATGTCTTCA
TC_LOJ_242	chr31:92921-93071	ACACTGACGACATGGTTCTACAATTGAAGTATCGCCAGAACAGCAT	TACGGTAGCAGAGACTTGGTCTGTGTTGCTTGGAGTAAGGCACTCT
TC_LOJ_243	chr21:288200-288319	ACACTGACGACATGGTTCTACACGGTCAGGATCGTTATAGTTTGGTAG	TACGGTAGCAGAGACTTGGTCTAGACACTTTGTATCGTATGCGTCGT
TC_LOJ_244	chr18:566462-566592	ACACTGACGACATGGTTCTACAATTATCTCGTGAGTTTGGCGGAAT	TACGGTAGCAGAGACTTGGTCTCAGAACCGTCTTGTCCTTCACTTC
TC_LOJ_245	chr3:1209990-1210114	ACACTGACGACATGGTTCTACAGGATCGACGTATGGGACGTATTTC	TACGGTAGCAGAGACTTGGTCTTTGAAGGACTGGAGCAAGACAAGT
TC_LOJ_249	chr10:1031977-1032097	ACACTGACGACATGGTTCTACAAAGCTCAGTGTTCAAAGTGCCATC	TACGGTAGCAGAGACTTGGTCTTTTCCTTGTTATCGGCTGTGAGAA
TC_LOJ_250	chr21:505080-505199	ACACTGACGACATGGTTCTACAGTTCTCCGTTACTTTCCGACACAG	TACGGTAGCAGAGACTTGGTCTTGCCATGTTACCCATAAACCACTT
TC_LOJ_251	chr5:743274-743396	ACACTGACGACATGGTTCTACACTAGGGATAGTGTCTCAACATTGGCTATAA	TACGGTAGCAGAGACTTGGTCTCACCCTTTAACTTTGAACGAACACG
TC_LOJ_252	chr36:237339-237479	ACACTGACGACATGGTTCTACATTAGAGCTTCGTATCGGCATGTTG	TACGGTAGCAGAGACTTGGTCTCACTTCATACATTTCCTCCAGAGACC
TC_LOJ_253	chr3:240382-240505	ACACTGACGACATGGTTCTACACCACTACCATTACCCGTGTCGTTA	TACGGTAGCAGAGACTTGGTCTCGCAGTCCTTGCTTAACCTCATTT
TC_LOJ_255	chr27:388555-388675	ACACTGACGACATGGTTCTACAGTTATTTGTATCCGTATCTTGCTGTCG	TACGGTAGCAGAGACTTGGTCTAGTATCACCTGGAGGACCGTGAAG
TC_LOJ_256	chr39:221720-221854	ACACTGACGACATGGTTCTACAAACTGACCGGAAGTGAGATTGATG	TACGGTAGCAGAGACTTGGTCTGGGCGGCGTCGTAGTATAAATAAG
TC_LOJ_257	chr5:992280-992407	ACACTGACGACATGGTTCTACACCTTTATTACGCTTCGGCAAGTACA	TACGGTAGCAGAGACTTGGTCTTTCCACGCAAACAATCAGTATCAG
TC_LOJ_259	chr32:837402-837557	ACACTGACGACATGGTTCTACAACTCTACACAAAGGCGTCAGAGATG	TACGGTAGCAGAGACTTGGTCTCCTGCAAGATCAATAAGGTTCAGC

Supplementary Table 4.2 (continued)

TC_LOJ_260	chr4:1353006-1353141	ACACTGACGACATGGTTCTACATGGTACTTGTTCAGCTCGGAAATC	TACGGTAGCAGAGACTTGGTCTCAAAGGCAGAGGAATGTTCAAAGA
TC_LOJ_262	chr1:2151183-2151303	ACACTGACGACATGGTTCTACACCGTAGTTGCGGTACGAATAAGTG	TACGGTAGCAGAGACTTGGTCTACTGGGAACGTGTATTAGGTATGGAGT
TC_LOJ_264	chr18:649186-649316	ACACTGACGACATGGTTCTACAGTGGAGGCGAAGAAGAAGTTTACA	TACGGTAGCAGAGACTTGGTCTAATAGAAACGGCATTCCATAAGCAC
TC_LOJ_265	chr27:343910-344029	ACACTGACGACATGGTTCTACAGTGCATCATATTCGATAGGGAGATGT	TACGGTAGCAGAGACTTGGTCTTATTACAGCATTGACCGTGTCTTCC
TC_LOJ_266	chr1:2205081-2205202	ACACTGACGACATGGTTCTACACTACGAAGTGCCTTAACTGCCTCA	TACGGTAGCAGAGACTTGGTCTATTCTATGTGCGTTTGGGTTTCAG
TC_LOJ_267	chr26:405401-405543	ACACTGACGACATGGTTCTACATTGCTTTCGATGGAGATAGACCTTT	TACGGTAGCAGAGACTTGGTCTGCGGAGATGTCTGATTTAGGAATTG
TC_LOJ_268	chr26:302445-302583	ACACTGACGACATGGTTCTACACGTAGTCAAACGGACTGAAGTACACA	TACGGTAGCAGAGACTTGGTCTGAGGAGGCAGTGGAGGTGTTAAAT
TC_LOJ_269	chr5:495739-495879	ACACTGACGACATGGTTCTACATCTTTATGACAAGTGCAACCAAAGC	TACGGTAGCAGAGACTTGGTCTCGTGATACTCCACCGTCTCAATCT
TC_LOJ_271	chr2:323827-323957	ACACTGACGACATGGTTCTACAGTGGGTTTCATCTCTCGTTTATGC	TACGGTAGCAGAGACTTGGTCTACCCTTGTCCATGTGTCTTGTAGC
TC_LOJ_273	chr1:2140290-2140430	ACACTGACGACATGGTTCTACACAATGGCACCAAGATAATAGTACAGGA	TACGGTAGCAGAGACTTGGTCTTGCAGAACCATCGTGAGAACTTTA
TC_LOJ_274	chr21:239185-239310	ACACTGACGACATGGTTCTACAAACAAGGTGAAGAAGAGCCATCAG	TACGGTAGCAGAGACTTGGTCTAAGGTGGAGGAGTTTGAACAGTACG
TC_LOJ_275	chr39:50470-50598	ACACTGACGACATGGTTCTACACTGCTCCTGATACTGCACAAACTG	TACGGTAGCAGAGACTTGGTCTGGTGCCTACAATGACTCCGTACAC
TC_LOJ_276	chr1:2694842-2694979	ACACTGACGACATGGTTCTACATTACACATTGCAGGGCAGCATATT	TACGGTAGCAGAGACTTGGTCTGTCTTTGTTCGTCATGTCAGCGTA
TC_LOJ_277	chr4:1382610-1382749	ACACTGACGACATGGTTCTACATAGCATCTTAATCAGCTCGGGAGA	TACGGTAGCAGAGACTTGGTCTGACGAACAAATGGAGAATCAGACG
TC_LOJ_278	chr2:164952-165077	ACACTGACGACATGGTTCTACAGGTCATTCACGCCAGTTCATACAT	TACGGTAGCAGAGACTTGGTCTACGGCCTTCTTCATAATCTCCATAA
TC_LOJ_279	chr9:400076-400197	ACACTGACGACATGGTTCTACACGAGACAGGGATGGACTCTTCAAT	TACGGTAGCAGAGACTTGGTCTGTTACGATGGCCTTGAGTGTGAGA
TC_LOJ_280	chr21:505326-505445	ACACTGACGACATGGTTCTACAGATTGCTACGTGAAGACGTGGAAG	TACGGTAGCAGAGACTTGGTCTGAGCGTATCGTACAGGCCAAAGTA
TC_LOJ_281	chr10:735827-735952	ACACTGACGACATGGTTCTACACAACGCATTTGGATTGCCTACTAA	TACGGTAGCAGAGACTTGGTCTAAACGTCTTGGTCTGTACGAGGAG
TC_LOJ_282	chr37:470203-470342	ACACTGACGACATGGTTCTACACTACTCAAGGAACCAGGCGTATTG	TACGGTAGCAGAGACTTGGTCTAACGTCCCACCAAGAATAATGAGC
TC_LOJ_283	chr12:561576-561706	ACACTGACGACATGGTTCTACACAGAAGGAGAAGACATTGGAACTCA	TACGGTAGCAGAGACTTGGTCTTCTTTGCCACTATCAAGCACCAAC
TC_LOJ_285	chr31:132291-132424	ACACTGACGACATGGTTCTACACATTGACCTTGCCACAGAAGTGTA	TACGGTAGCAGAGACTTGGTCTTGGCCTTATTCACATACTCCACAAG
TC_LOJ_286	chr15:941983-942118	ACACTGACGACATGGTTCTACAGGCGTATCCACCACAAGAGTAGAA	TACGGTAGCAGAGACTTGGTCTGGATGCCAGATTACGTGAAAGAAA

Supplementary Table 4.3 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_2O (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

	Item	Availability (quantity / price)	Quantity for 100 samples	Cost for 100 samples	Comment
Sequencing	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

Chapter 5

Prediction and prevention of parasitic diseases using a landscape genomics framework

Philipp Schwabl[a], Martin S. Llewellyn[a], Erin L. Landguth[b], Björn Andersson[c], Uriel Kitron[d] Jaime A. Costales[e], Sofía Ocaña[e] and Mario J. Grijalva[e,f]

[a]Institute of Biodiversity, Animal Health & Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK

[b]Division of Biological Sciences, University of Montana, 32 Campus Drive, Missoula, Montana, 59812, USA.

[c]Department of Cell and Molecular Biology, Science for Life Laboratory, Karolinska Institutet, Berzelius väg 35, 171 77 Stockholm, Sweden

[d]Department of Environmental Sciences, Emory University, Atlanta, GA, USA

[e]Center for Research on Health in Latin America, School of Biological Sciences, Pontifical Catholic University of Ecuador, Quito, Ecuador

[f]Infectious and Tropical Disease Institute, Biomedical Sciences Department, Heritage College of Osteopathic Medicine, Ohio University, 45701 Athens, OH, USA

This chapter has led to a publication in Trends in Parasitology (2017),
available at doi: 10.1016/j.pt.2016.10.008.

5.1 Abstract

Substantial heterogeneity exists in the dispersal, distribution and transmission of parasitic species. Understanding and predicting how such features are governed by the ecological variation of landscape they inhabit is the central goal of spatial epidemiology. Genetic data can further inform functional connectivity among parasite, host and vector populations in a landscape. Gene flow correlates with the spread of epidemiologically relevant phenotypes among parasite and vector populations (e.g., virulence, drug and pesticide resistance), as well as invasion and re-invasion risk where parasite transmission is absent due to current or past intervention measures. However, the formal integration of spatial and genetic data ('landscape genetics') is scarcely ever applied to parasites. Here, we discuss the specific challenges and practical prospects for the use of landscape genetics and genomics to understand the biology and control of parasitic disease and present a practical framework for doing so.

5.2 Introduction

5.2.1 Parasites, genes, and landscapes

Individual parasite species around the world are distributed across different ecological settings, spanning rural, peri-urban and urban areas. For widely distributed parasitic diseases, 'patchy' geographic distribution of cases frequently occurs, where parasite, vector and host-related factors conspire to promote intense local transmission[560]. Understanding how abiotic and biotic environment features affect the movement of parasites, their hosts and vector species, is critical for disease control.

Spatial or landscape epidemiologists aim to exploit prior knowledge about environmental heterogeneity, often to the level of communities and households, to map current parasite distributions and develop models to predict future disease incidence (e.g., Vazquez-Prokopec et al. (2012)[561]). In addition to using spatial information to predict the presence and abundance of parasitic agents, it is also vital to establish the extent to which environmental features impact genetic connectivity between individuals and populations. The spatial distribution of genetic diversity directs the co-evolutionary outcome of host-vector-parasite interactions when selection is spatially heterogeneous[562]. Gene flow modifies this genetic distribution and therefore not only correlates to the spread of epidemiologically relevant traits (e.g., drug resistance[563] or virulence[309]) but also regulates local adaptation, the emergence of novel phenotypes and their invasion of areas free of parasite transmission (e.g., Fitzpatrick et al. (2008)[564]), including those subjected to past or current intervention measures. However, while models of parasitic disease spread are becoming spatially explicit

(e.g., Vazquez-Prokopec et al. (2012)[561]), these still rarely incorporate genetic data. Studies on host, vector and parasite population genetics also abound, but these, in turn, too seldom incorporate spatial data. We believe that a framework for the formal integration of parasite genetic connectivity with host-vector dynamics in heterogeneous space is needed to bridge these gaps.

Landscape genetics, a body of theory aimed at combining landscape ecology and population genetics, is now approaching twenty years old[565]. Over this period, landscape genetic approaches have primarily examined the impact of habitat fragmentation on genetic differentiation (e.g., Cushman et al. (2012)[566]), land use and environmental change on the genetic diversity of threatened species (e.g., Wasserman et al. (2013)[567]), as well as the sustainable management and commercial exploitation of others (reviewed in Sommer et al. (2013)[568]). The spread of parasitic disease, however, has drawn only limited attention from the field. Pioneered by work on rabies[569] and chronic wasting disease[570], research has targeted a handful of viruses (reviewed by Biek and Real (2010)[571]; see also Dellicour et al. (2016)[572]) and microbes (notably *Batrachochytrium dendrobatidis*[573]), helminths with direct life cycles[574,575] and their hosts. Systems involving vector-borne pathogens[576–579] or several intermediate hosts[580] have been mostly spared from investigation. We believe the application of landscape genetics to vector-borne disease agents, especially including **landscape genetic simulation modelling**[581] (see Glossary), has significant, underappreciated potential to inform targeted disease control strategies.

In this opinion piece, we highlight the need for landscape genetic and genomic tools to study parasitic disease and present a framework for how they might be implemented. In doing so, we first provide an overview of landscape genetics/genomics, the role of landscapes in driving genome-wide adaptation in parasites, and discuss challenges and prospects for the use of landscape genomics to understand the biology and control of parasitic disease. We often refer to Chagas disease, recently ranked as the highest parasitic disease burden in the Western hemisphere[582]. In the absence of vaccine or cure, intervention strategies against such neglected zoonoses may profit most from the landscape genomic approach.

5.2.2 What is landscape genetics?

A primary goal of landscape genetics is to understand how landscape features influence observed spatial genetic (neutral or selection-driven) structure[583]. Key concepts in landscape genetics involve correlating genetic data with geographic data through individual-based measurements of dissimilarity. For example, genetic distances (i.e., dissimilarity matrices) can be quantified using individual-based metrics, such as proportion of shared alleles D_{ps}[584] or Rousset's A[585]. In all but the simplest models (i.e., **isolation-by-distance** or isolation-by-

barrier), geographic distance is typically replaced by **cost-distance**[586], which reflects both the geographic distance between individuals and the degree to which the intervening landscape is hypothesized to impede gene flow and underlying dispersal movements (i.e., **isolation-by-resistance**[587]). Cost-distance is calculated across a **resistance surface** wherein each cell in a geographic information systems (GIS) raster map is assigned a value based on a hypothesized species-specific resistance to traversing the landscape feature the cell represents[588].

In a typical landscape genetic approach, cost-distances among individuals are calculated based on multiple, competing resistance hypotheses. These cost-distances are then evaluated (i.e., correlated) against empirical genetic distances among the same individuals, primarily using the Mantel test and its derivatives (e.g., multiple regression on distance matrices (MRDM) or partial Mantel tests within a causal modelling framework[589]). Although techniques such as distance-based redundancy analysis[590] are increasingly applied to test landscape resistance on gene flow (e.g., Geffen et al. (2004)[591]), Mantel-based approaches are still the mainstay of landscape genetic analyses.

Landscape genomics extends landscape genetics to the exploration of genome-wide data (the two terms are applied accordingly herein), often in search of patterns of covariation between allele frequencies and environmental conditions (i.e., **genotype-by-environment associations** (GEAs) – see Box 5.1). As these signs of selection may point to the role of local adaptation in structuring populations, their study necessarily fuses into the framework we introduce in the next section.

5.3 Landscape genomics to study parasitic disease

With the exception of recent theoretical work in the context of Lyme´s disease[579], essentially all landscape genetic studies applied to parasitic diseases to date have considered a single level of transmission, focusing primarily on landscape resistance hypotheses that influence movement processes and thus, gene flow, of principal reservoir hosts. For complex, multi-species disease systems, we find that today's landscape genomic methods warrant a more inclusive, multi-level approach. In particular, we recognize resistance surface construction, a precursor to several landscape genetic applications, as a convenient analytical step during which interactions among host, vector and parasite can be formally integrated for further analysis. In Fig. 5.1, we outline a multi-species landscape genomic approach to predicting disease spread in host-vector-parasite systems. Fig. 5.2 breaks down the key translational step, resistance surface construction, by example of Chagas disease. In brief, host distribution (i.e., all spaces that permit host movement) is abridged by vector distribution relative to host movement rate (parasite transmission remains viable where the two

distributions do not coincide so long as movement rate allows the host to re-enter areas of overlap within the infective period). Likewise, vector distribution is abridged by host distribution relative to vector movement rate. Added together, these effective distributions determine parasite distribution. First, the values of different potential environmental influences on principal local host and vector species movement through the landscape are mapped. These landscape data become the primary sources for studying parasite spread. Host and vector conductivity-to-movement surfaces are then calibrated based on transmission competence and merged to create a parasite resistance-to-movement surface. Parasite dispersal and resultant population genetic structure over this composite surface are modelled directly using landscape genetic simulation software. Finally, simulated and empirical parasite population genetic structure are compared to evaluate hypothesized landscape effects at the multiple transmission levels that take part in the spread of disease. Crucially, this approach does not rely on any assumption of genetic co-structure between vector or host and parasite, a phenomenon rarely ever observed in natural systems[592].

Box 5.1 Landscape genomics and genotype-by-environment associations of parasitic disease

Landscape genomics scans genome-wide, high-density marker datasets to elucidate GEAs[593]. As specialized regression methods (e.g., mixed models that control for demographic history and drift[594]) identify environment-related clines in allele frequencies, possible targets of selection are not exposed *per se.* Better yet perhaps, these emerge from regression as correlations to environmental predictors, i.e., coupled to possible cause. Central ecological proxies such as temperature present intuitive starting points in the search for these GEAs. Yet, depending on the system and objective, exploration may venture far beyond classic considerations. In exploring the 'landscapes' of parasites, for example, hosts and vectors often bear environmental variables of primary interest (and relevant values might be retrieved from auxiliary sampling – e.g., clinical observations or genetic data from the vector source). Here, the genetic bases of a certain phenotype (e.g., virulence) may stand at question, such that putative ecological pressures (e.g., host density, coinfection[595]) on this particular trait are chosen to be scanned for responsive loci. In time, as countless cases of heterogeneity enter regression and ever more GEAs unfold, landscape genomics promises to pass on a kaleidoscope of potential gene function for follow-up experimental studies to explore.

GEAs are also essential to downstream analyses within the same field, e.g., to incorporate selective forces in spatially explicit simulations of population genetic change (e.g., CDPOP[325]). Analyses of this type may expose fundamental adaptive constraints that limit parasite range expansion and response to climate change. Apart from such implementation, GEAs also enhance interpretation of independent results. The upscaling of analysis to many thousands of markers vastly improves power to unmask intricate demographic and evolutionary structures – gradients of selection, incipient speciation, cryptic niches, etc. This enhanced resolution, however, also requires enhanced approaches to interpretation and often calls on GEAs. For example, novel spatio-genetic visualization tools (e.g., MEMGENE[596]) may expose instances where gene flow deviates from consistent patterns of isolation-by-resistance. These deviations may issue from any number of selective processes. Local adaptation is one such process and a topic of ongoing discussion in the study of parasites. While the presence of locally adapted residents may impede genetic introgression (e.g., selection against hybrids), it may just as well take opposite effect (e.g., frequency-dependent selection of rare variants)[597]. Complementary information on GEAs provides critical guidance in navigating the many possibilities and understanding how gene flow, drift and selection mosaics interact to drive parasite local adaptation (see Gandon and Nuismer (2009)[598]).

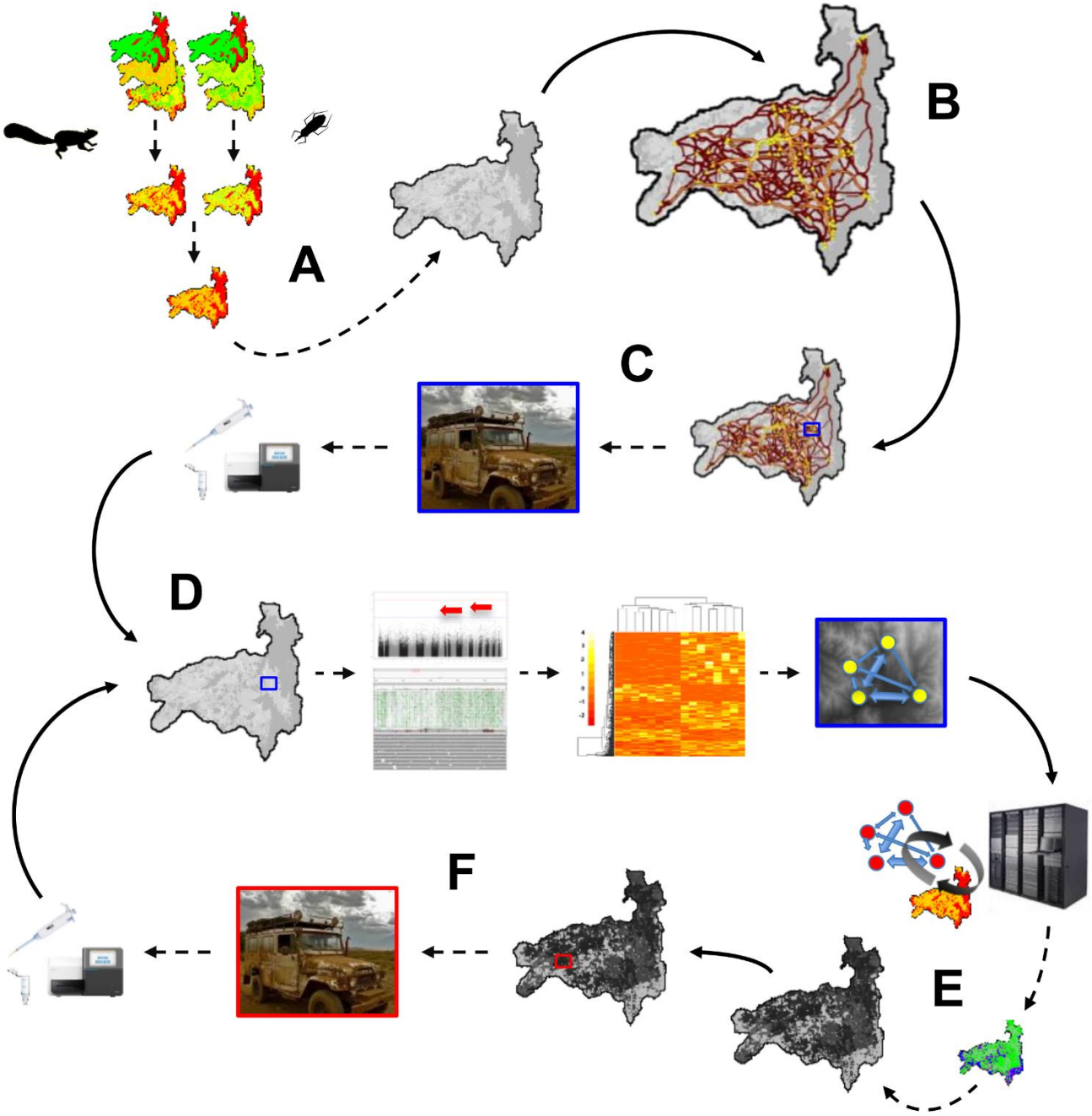


Figure 5.1 Exploiting landscape genomics to predict parasite dispersal in heterogeneous landscapes. The construction of a predictive map of parasite dispersal from high-resolution landscape and genetic data is outlined in six steps (A-F). Step A is further detailed in Fig. 5.2 by example of *Trypanosoma cruzi* transmission in southern Ecuador.

A. Host/vector resistance surface construction. Informed by biological and ecological data, principal host and vector species are specified and the landscape variables underlying their movement are mapped. Landscape features are assigned levels according to their putative impact on host/vector movement and merged to create a landscape conductivity-to-movement surface. Surfaces generated for both host and vector are then weighted, merged and converted to a composite resistance-to-movement surface. If additional, host/vector-independent variables extrinsic to parasite survival and development are hypothesized, the resistance surface may be further updated to incorporate these requirements.

B. Landscape connectivity analysis. A landscape connectivity model (e.g., least-cost path analysis or circuit theory) is generated using programs such as PATHMATRIX[599] or CIRCUITSCAPE[600]. While least-cost models specify single optimal paths of movement between sites on a resistance surface, circuit theory considers multiple pairwise connections[587] and may enhance prediction of passive, multi-dependent dispersal systems in landscapes of continuous resistance.

Figure 5.1 (legend continued)

C. Study site identification and phase one genetic data collection. Guided by path analysis results, phase one sampling locations are selected to encompass heterogeneous landscape resistance. Parasites are sampled (i.e., host/vectors captured, parasites isolated and DNA extracted) and DNA is sequenced.

D. Cost-distance analysis. Metrics of dissimilarity are calculated among individual genotypes (e.g., based on genome-wide single nucleotide polymorphisms) and correlated to cost-distances (computed in, e.g., PATHMATRIX[599]) separating these individuals (see main text) for preliminary validation of the resistance surface constructed in step A. GEA interactions are also explored based on various landscape features (see Box 5.1).

E. Data simulation and iterative resistance surface modification. Using tools like CDPOP[325], spatially explicit changes in population genetic structure are simulated as functions of individual-based movement, reproduction, mortality and dispersal[581]. These models predict patterns of gene flow (i.e., connectivity) between individuals based on the resistance surface constructed in step A and GEAs detected in step D. Simulated connectivity measures are then compared to empirical estimates from step D to further validate the resistance surface. Surface components (e.g., conductivity values (see Fig. 5.2, step A3)) are iteratively re-weighted until connectivity matches the observed (i.e., **pattern-process modelling**).

F. Landscape model validation. The refined landscape resistance surface underlying parasite dispersal in the phase one sampling area can now be extrapolated regionally. At a second, independent site, parasites are sampled, sequenced and genotyped. Cost-distance analysis and the goodness-of-fit between simulated and empirical connectivity at the second site determine the power of the resistance map.

5.4 What makes landscape genomics such a powerful tool to study parasitic disease?

5.4.1 Accuracy in detection, precision in prediction

Spatially explicit models of parasite dispersal have traditionally been fitted and validated against occupancy and abundance data[601,602]. Genetic structure of the disease agents still rarely replaces these response variables despite several clear advantages for host-vector-parasite systems. Interpretation of occupancy and abundance data is complicated by imperfect detection, and many zoonoses (e.g., Chagas disease and leishmaniasis, for which surveillance is under-resourced, diagnostics are substandard, and symptoms are inconsistent) are highly prone to this bias. Pairwise genetic data are robust to detection bias and offer far greater resolution to inference on parasite dispersal, chiefly because genotypes not only identify individuals but also dynamic associations of alleles, their origin and putative location of intermediate genotypes. Predicting when and where genotypes and alleles end up in the landscape based on their current spatial distribution is critical for disease control. For example, increased resistance, virulence and transmission potential often arise only when certain sets of genes come to combine[154,563,603], each uniquely routed by ecological features of variable resistance-to-movement and selective force. As spatially explicit, individual-based modelling turns ‘genotype-based’, intricate demographic and evolutionary interactions (e.g., heterosis or selection for/against specific alleles and reproductive modes) become decipherable from neutral and adaptive genetic structure in space and time.

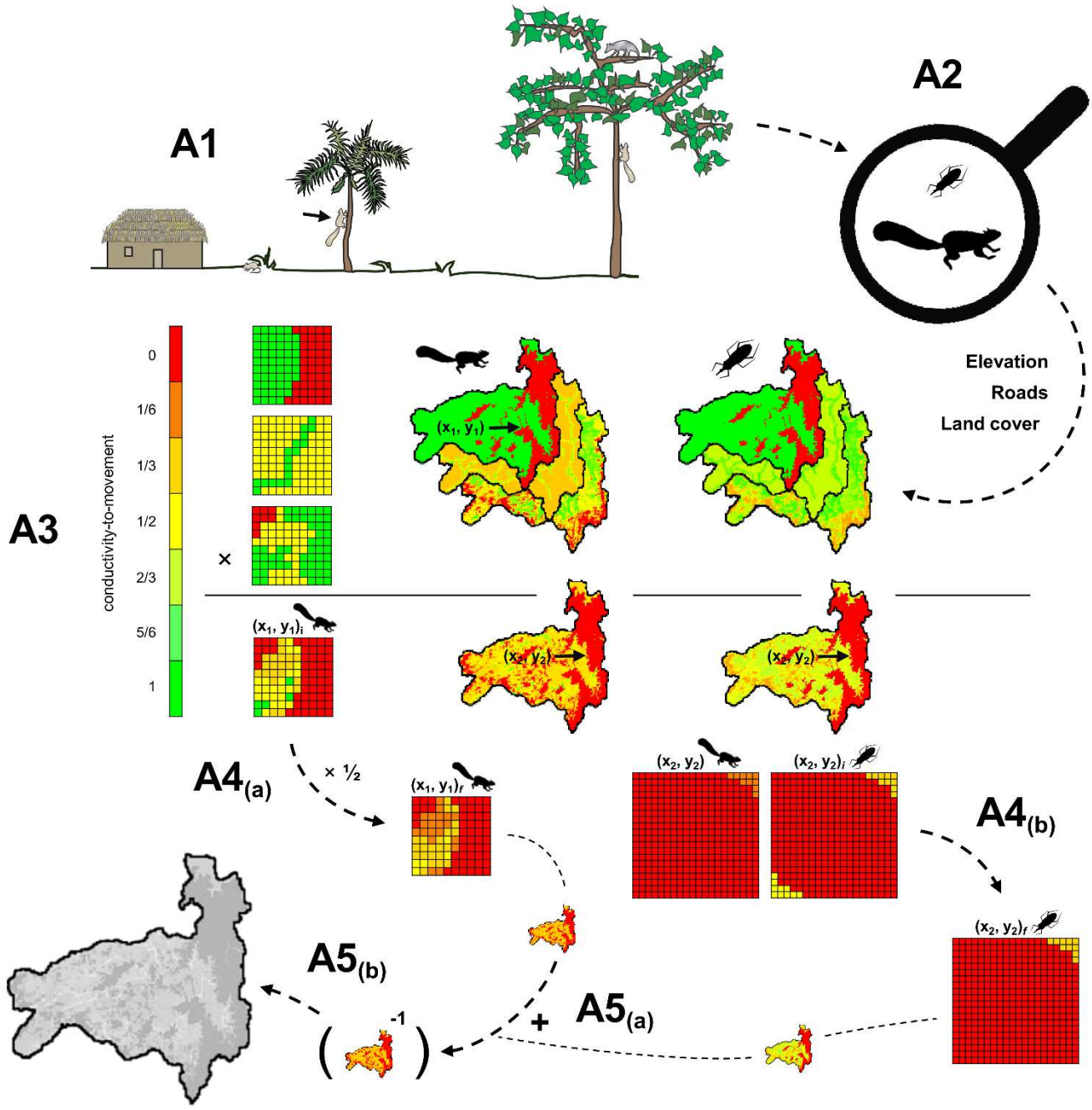


Figure 5.2 Composing a resistance map for the regional transmission of Chagas disease. Resistance surface construction, the first step in cost-distance analysis (Fig. 5.1, step A), allows multi-species parasitic disease systems to fold neatly into the landscape genomic approach. We work though this key translational step by example of *T. cruzi* transmission in southern Ecuador.

A1. Specification of principal host/vector species. As host/vector specification founds all further analyses, factors relating to transmission competence must be thoroughly examined, e.g., abundance, vagility, physiological and life-history traits determining susceptibility, tolerance and transmission intensity. Studies on eclectic (also 'host-fitting'[149] parasites such as *T. cruzi* may require that spatial study extent be reduced to scales at which limiting agents emerge. In Loja Province (ca. 100 km x 100 km), *Sciurus stramineus* is specified as principal *T. cruzi* host based on the rodent's year-round arboreal nesting, i.e., triatomine habitat that holds against limiting vegetation phenology at the domestic-sylvatic interface[412]. This triatomine association is supported by other randomized sampling[509] and blood meal analyses that link high infection tolerance to short-lived species with high reproductive rates[604]. *Rhodnius ecuadoriensis* is specified as primary *T. cruzi* vector based on its ecology, defecation and feeding patterns[605] and wide distribution of sylvatic and synanthropic populations in southern Ecuador[509].

Figure 5.2 (legend continued)

A2. Specification of landscape features underlying host/vector movement. Data modelling[606] and algorithmic approaches[607] specify land-cover type and elevation as two principal determinants of triatomine movement at the scale applied in Loja. Analyses of triatomine genetic structure also suggest a strong influence of human transport (i.e., roads) on dispersal at this scale[608]. These three features also regulate host movement. *S. stramineus* is native to the Andes and, despite declines from land-use change, populations are now common from 2,000 m to sea-level (similar to *R. ecuadoriensis*[609]) in various forest and man-made environments[610].

A3. Composition of conductivity surfaces. Remote-sensing data on elevation, land-cover and roads are rasterized and re-coded to conductivity-to-movement scores. In this case, re-coding is coarse (e.g., for both host and vector, conductivity = 1 if elevation ≤ 2,000 m), given that ecological traits of *S. stramineus* (e.g., habitat/trophic flexibility[610]) and *R. ecuadoriensis* (e.g., microhabitat selection[609]) likely buffer continuous landscape effects on movement. The product of the three scores is then taken for each cell to generate host and vector conductivity surfaces.

A4. Abridgement and weighting of conductivity surfaces. The distribution of raster cells that allow for host movement is now abridged based on vector distribution relative to host movement rate and infection time. Cells conducive to vector movement are corrected based on host distribution in the same manner: if the distance to the nearest cell where host-vector interaction is possible (i.e., where *S. stramineus* conductivity is non-zero) exceeds maximum parasite carriage distance by the vector (equable to *R. ecuadoriensis* dispersal range (ca. 2,000 m, based on Schweigmann et al. (1988)[611]) when infection does not compromise lifespan and movement (e.g., Schaub (1988)[612] and Castro et al. (2014)[613]), the vector conductivity score is re-coded to zero. Once abridged, host conductivity values are scaled by a coefficient that quantifies host relative to vector competence in dispersing *T. cruzi*, weighing in factors such as vagility and transmission intensity (e.g., see specifications by Devillers et al. (2008)[601]).

A5. Conversion to a composite resistance surface. The refined conductivity surfaces for *S. stramineus* and *R. ecuadoriensis* are merged by addition, then inverted to generate the resistance surface.

Early attempts to apply landscape genetic methodologies to infectious agents have yielded unprecedented precision in disease prediction and surveillance. For example, by coupling spatial analysis with phylogenetic methods, Biek et al. (2007) demonstrated segregated dispersal trajectories and intermittent expansions among the viral lineages of an explosive rabies outbreak in the mid-Atlantic United States[614]. Unrelated to selection on novel variants (given few, irregular changes at adaptive loci), dispersal patterns were explained by viral spread into low-elevation raccoon habitats and restrained dispersal behind the wave front. This elevation-based patterning was recently affirmed using cost-distance approaches akin to those outlined in Fig. 5.1[572] (see also applications on principal rabies hosts[615]).

As cost-distance methods begin to spread through virus research, cases from vector-borne disease systems remain few and far between, but all the more compelling. In West Africa, for example, Bouyer et al. (2015) built 'friction' maps to model least-cost paths between *Glossina palpalis* populations and the 'main tsetse belt' of the region[578]. Paths were then ranked by cost to identify isolated eradication targets in the fight against African trypanosomiasis. Medley et al. (2015) also used landscape genetics to study disease vectors, decomposing the invasion process of *Aedes albopictus* through the United States[577]. Here, deft study design (e.g., flower vases recognized as preexistent larvae repositories in 26 cities

connected by various traffic intensities) and high-resolution (30 m) land-cover sensing provided for MRDM on range-wide data at multiple spatial scales. Results depict occasions of long-distance, human-aided *Ae. albopictus* expansion followed by stepping-stone dispersal as a function local landscape. Unfortunately, however, both of these studies did not advance their powerful resistance models to simulation for additional validation, refinement and extrapolation, i.e., steps E and F in Fig. 5.1 (yet, see an intriguing follow-up study[616] on the scope of landscape genetic simulation modelling to evaluate pattern-process relationships such as those inferred by Medley et al. (2015)).

5.4.2 Power to explore the unorthodox and unknown

As isolation-by-resistance featured prominently in the studies above, landscape effects on non-neutral genetic structure have been largely discounted so far. Yet, dispersal outcomes are without question also shaped by context-dependent adaptive change (and vice versa – see Box 5.1), sometimes to profound effect (e.g., hybridization under insecticidal pressure[617]). To this end, landscape genomics' pioneering approach to simultaneously detect divergent alleles and their ecological drivers, then to visualize and simulate both neutral and selection-driven structure in heterogeneous space, has received special attention (Box 5.1). There is considerable scope for the use of landscape genomic tools to study adaptive genetic change in parasitic disease and clear advantages over classic population genetic approaches.

Among several parasite species, reproduction is not uniform, with clonal propagation interspersed by unorthodox modes of genetic exchange. Especially for parasitic protozoa, these episodes of recombination remain incompletely defined both in mechanism and extent (see Chapter 2). Traditional approaches to detect targets of selection scan for excess genetic differentiation between discrete populations (e.g., outlier analyses such as BAYESCAN[460]). However, methods to define such populations *a posteriori* (e.g., Pritchard et al. (2000)[335]) rely on assumptions of Mendelian sexuality and are thus liable to distort results at the earliest stage of analysis when applied to parasitic species. In contrast, landscape genomics' correlative GEA methods (see Box 5.1 and Forester et al. (2015)[593]) are individual-based and make few assumptions about the underlying reproductive mode.

Host-vector-parasite systems are also inclined to subtle, step-wise adaptive change, i.e., weak selection on individual alleles[618]. In parasites, this tendency relates to high mutation rates and population sizes[563,618,619], as well as elevated gene redundancy and ploidy[6]. In hosts and vectors, the effect likely arises from prevalent polygenic, epistatic and pleiotropic control of interaction traits[620,621]. Simulations show how quickly differentiation-based methods lose power to detect adaptive change as selection intensity weakens, reaching complete impotence at levels still easily managed by correlative alternatives[357]. The latter

take further leverage from study designs that prioritize environmental representation over genetic sampling intensity per site, a strategy counter to classic methods based on clustered sampling. These arguments were recently taken from simulation to reality in coastal Kenya, where Mackinnon et al. (2016) applied environmental association analysis to genotypes obtained from a hospital serving ethnic groups long segregated among ecotypes of contrasting malaria prevalence[621]. After rejecting dozens of disappointing candidates proposed by methods of the past, this search for resistance loci exposed several divergent genes that mitigate brain inflammation, a symptom of severe malaria. Moreover, the study detected subtle clines in the sickle-cell mutation β^S, signs of balancing selection seldom distinguished at such fine spatial scales.

Naturally, landscape genomics' potential to enhance resolution and power in the study of parasitic disease also has its caveats. A few areas of concern are introduced in Box 5.2.

Box 5.2 Limitations of landscape genomics to study parasitic disease

As landscape genetics is just entering its teenage years, uncertainties come and go. Lasting concerns relate primarily to statistical power (e.g., high type I error due to non-independence, multicollinearity and multiple testing) and empirical sampling design (e.g., how to select spatio-temporal scales). These issues affect the entire body of landscape genetics/genomics and are under extensive treatment[622,623], increasingly aided by simulation software[624]. We therefore turn to caveats of particular relevance to applications on parasitic disease.

We share concerns that high-resolution model output from simulations of gene flow is easily generated, taken for precision and misapplied[625]. Ethical arguments for immediate translation and high visibility of research on human disease (e.g., Quick et al. (2016)[626] intensify this risk. Also, our framework will sometimes rely on limited 'expert knowledge' to elaborate core model input (i.e., the multi-species resistance surface). Moreover, resistance-to-movement may involve variables (e.g., soil conditions for helminths[627]) and scales (e.g., micro-geographic differentiation in *Plasmodium*[628]) for which empirical data are unavailable.

We also emphasize that landscape genomics may miss principal causes and consequences of disease spread for phenotypes of non-heritable or complex genetic basis. Pathogenicity, for example, can regulate disease spread[629] and founds on complex epistatic host-parasite interactions. Not only is genetic structural variation known to underlie pathogenic differences ((e.g., Behnke et al. (2011)[630]), host tolerance (likely of low heritability itself[631]) further modifies infection outcomes. Classic models of dispersal skirt this complexity by directly implementing phenotypic data (e.g., infection intensity, clinical forms), and classic approaches to detect adaptive variation have adjusted to search beyond the single locus. Meanwhile, landscape genomics continues to define and apply genotypes as proxies for phenotypes with limited discretion. For example, environmental resistance may differ among genetic structural variants[21], but standard metrics of dissimilarity do not measure such differentiation. Indeed, defining and interpreting genetic structure is often troublesome and tempts to simplifying but spurious assumptions. Such shortcuts through our framework require caution. For example, in step A1 (Fig. 5.2), resorting to analysis of host/parasite genetic co-structure to distinguish principal host species (see Mazé-Guilmo et al. (2016)[592]) is rather hazardous, as is linking GEAs to local adaptation while slighting other forms of selection (see Bierne et al. (2011)[632]). Clearly, landscape genomic tools require discreet handling and refinement based on underlying hypotheses, and interdisciplinary complementation remains indispensable to the study of parasitic disease.

5.5 Prospects

5.5.1 Conservation genomics in reverse

In conservation biology, landscape genomics strives to identify 'conservation units', i.e., genetically unique subpopulations to be preserved and/or managed distinctly to sustain biodiversity of the whole[633]. In epidemiology, spatial genomics are crucial to identifying operational units that maximize the reach of surveillance and control. Apprised of such epidemiological units and their distributions, insecticidal campaigns (often too indiscriminate to be sustainable in the past[634]), for example, might aim precisely to rule out pivotal hybridization outcomes observed *in vitro* (see below) or capitalize on high landscape resistance to gene flow (see Bouyer et al. (2015)[578]), while diagnostic approaches might be differentiated based on particular genotypes expected to arrive in a region. A look at leishmaniasis further elaborates these points. Hundreds of thousands, primarily the poor, fall victim to this neglected zoonosis every year, with cases ranging from self-healing cutaneous infection to severe disfigurement and fatal visceral disease. The distinct pathologies ascribe to certain subsets of *Leishmania* species[635], yet these may also proliferate as natural hybrids of enhanced virulence, resistance and plasticity[170]. For good reason, therefore, underdeveloped molecular surveillance strategies are now remonstrated in such places as Colombia, where massive efforts to innovate this area are currently underway[636]. Elsewhere, especially in Brazil, much effort has been devoted to ecological niche modelling (ENM) to inform *Leishmania* control. While such occupancy-based correlative and algorithmic methods provide essential guidance, direction is generally less immediate. For example, ENM rates nearly all of Amazonia at current risk to leishmaniasis and projects southward vector expansion under climate change[637], but what next? Where are limited intervention resources to be allocated, and when? Might temperatures be approaching tipping points to rapid proliferation of disease? In a landscape genomic cost-distance framework that models connectivity and genotype movement in the very process of identifying resistance variables, simulation-based analysis may promptly transition to such questions. For example, after pattern-process modelling American marten (*Martes americana*) dispersal in the Rocky Mountains, Wasserman et al. (2012) proceeded right to forward-simulation of population structure in a warming climate[638]. Results not only detail gradual habitat and population fragmentation through space and time, but specify imminent warming thresholds beyond which genetic connectivity plummets to levels that threaten extinction. Translating such innovations from landscape genetic/genomic conservation studies offers to accelerate progress towards high-impact solutions against pervasive disease under global change.

5.5.2 Groundwork for genetic modification in disease control

In sub-Saharan Africa, the burden of neglected diseases such as leishmaniasis is far outweighed by that of malaria. As existing control strategies cannot keep pace (e.g., ca. 400,000 malaria deaths in 2015[639]), the swift replacement of natural vector populations through transgenic, *Plasmodium*-refractory types offers much appeal. However, this approach depends on mating among transgenic and natural mosquitos in populations unlikely to be panmictic (in fact, cryptic speciation is rather notorious to *Anopheles gambiae*, principal malaria vector of the sub-Sahara[640]). Therefore, patterns and processes of genetic connectivity and reproductive isolation in the target environment must be well understood to legitimize transgenic release and predict its manifold effects[641]. Landscape genomic tools are designed precisely to forward such understanding. For example, after identifying key drivers of dispersal from cost-distance analyses applied to native vector populations (e.g., as in Medley et al. (2015)[577]), transgenic genotypes could be placed into landscape genomic simulation modelling of mating, selection and dispersal in the landscape. Should the transgene confer environment-dependent fitness costs (see Marrelli et al. (2006)[642]), various simulators could also integrate this information to forecast gene flow and consequent distribution of refractory types through the environment[624]. Simulations might also explore to what extent transgene fitness costs must be reduced or inheritance must be biased (transgenesis methods often exploit 'selfish genetic elements'[643]) for effective replacement of native vector populations. Finally, based on resultant equilibrium conditions, *Plasmodium* dispersal could be modelled among remnant (e.g., reproductively isolated) vector and human populations in the framework outlined above. Here, resistance surface construction offers to incorporate temperature-dependent vectorial capacity (e.g., changes in *Anopheles* immunity and *Plasmodium* fitness[644]) and other theoretical updates on disease spread in heterogeneous space. In times to come, these explorations will help disambiguate and enhance the potential of transgenic release strategies as well as consider how standard methods best round off novel efforts to defeat malaria and other major parasitic disease.

5.6 Concluding remarks

Here, we claim a strategic place for host-vector-parasite interactions to join spatially explicit analyses of genetic connectivity. This integration not only allies molecular epidemiology with landscape ecology, but advances both into the realm of 'landscape community genomics'[645], only just envisioned to explore previously impenetrable eco-evolutionary causes and consequences of genomic structure. First inroads would be well-timed to seek out the potential of landscape genomics in forecasting land use, climate change and intervention impacts on parasite dispersal. Parallel efforts underway across various

disciplines offer ample opportunity to validate and synthesize results to 'best practices' for sustainable disease control. Novel genome-typing strategies (e.g., pathogen GLST combined with restriction site-associated vector DNA analysis[646]) that now place individual-based, multi-species genomic analyses within possibility of a single study also impel research on interactions between genotype-genotype factors (e.g., hybridization and co-evolution) and disease heterogeneity in the environment. However, no single study can or should take on too many questions at once. Only following clear hypotheses on a few factors of interest can landscape genomic methods such as those presented above be adequately tuned and, when necessary, replaced. Indeed, the framework presented here is just that – a framework, and discretion is advised. We hope to have placed helpful rails, not unchangeable rules, into challenging new terrain for the study and prevention of parasitic disease.

Box 5.3 Glossary.

Cost-distance: The cumulative resistance of intervening landscapes to the movement of individuals (or populations, etc.) between a pair of sites. These 'distances' are typically calculated by scoring landscape variables (e.g., elevation) based on (putative) resistance-to-movement, plotting resistance scores into a raster grid (see 'resistance surface' below) and adding up grid values along the path(s) of interest.

Genotype-by-environment association (GEA): A correlation between genetic and environmental variation and possible effect of natural selection. In landscape genomics, specialized regression models are applied to genome-wide data collected in heterogeneous landscapes to detect these GEAs as environment-related clines in allele frequencies.

Isolation-by-distance (IBD): In the IBD model, the probability that an individual disperses to any site in the landscape depends only on its distance to that location. Here, no matter the heterogeneity in the landscape, 'cost-distances' (see above) between sites relate directly to straight-line Euclidean distances, given that landscape features are not considered to resist movement and/or modify paths of dispersal.

Isolation-by-resistance (IBR): Unlike for 'IBD' (see above), Euclidean distances do not suffice to predict the level of dispersal between a pair of sites in the presence of IBR. Rather, the probability that an individual disperses from one site to another depends also on the resistance of the intervening landscape to the movement of that individual (see 'cost-distance', above).

Landscape genetic simulation modelling: A spatially explicit modelling framework to simulate the actions and reactions of organisms and attendant genetic structure in heterogeneous space. Simulations are generally individual-based, such that these actions and reactions (e.g., dispersal, mating, survival) depend not only on user-defined landscape heterogeneity but also on inter-individual differences in age, sex, fitness, etc.

Pattern-process modelling: A modelling scheme that evaluates whether an underlying process inferred through empirical induction can produce the patterns (e.g., population genetic structure) observed in the data, and how well (i.e., at what precision, accuracy and repeatability) it can do so.

Resistance surface: A representation of the landscape, often in raster form, in which each location (e.g., raster cell) is assigned a cost or resistance value which affects movement and gene flow through the landscape.

Chapter 6

General discussion

6.1 Final synopsis

This dissertation infiltrated the secret lives of *T. cruzi* and *L. infantum* using a combination of classic population genetic theory and modern genomic tools.

Our WGS analyses on *T. cruzi* aimed to understand population structure and underlying reproductive mechanisms at an endemic transmission focus of Chagas disease in Loja Province. By specializing polymorphism analysis for short-read data on repeat-rich genomic DNA and verifying new hypotheses with targeted subcloning and re-sequencing experiments, we apprehended microgeographic reproductive polymorphism and, importantly, a focus of sexual recombination within the study landscape. As elaborated in Section 6.2 below, these findings contradict long-standing dogma about the role of genetic exchange in shaping contemporary *T. cruzi* population structure and various experimental and long-read sequencing follow-up studies are planned.

Our WGS analyses on *L. infantum* aimed to understand the extent and mechanisms by which this species has re-diversified after bottlenecking into the New World during the colonial era. Disentangling the influences of selection and demographic changes on the subtle genetic heterogeneity we exposed across Brazil required a variety of computational methods (coalescence modelling, copy number analyses, phylogenetics on phased and simulated genotypes, etc.) as well as phenotypic assays and gene dose measurement in monoclonal subcultures using qPCR. Section 6.3 elaborates on our discoveries as well as on shortcomings and follow-up needed to substantiate the theories we have proposed.

Complementary to the above WGS studies, we also developed GLST, a culture-free genotyping system that rapidly constructs genome-wide amplicon libraries for a pathogen of interest by co-amplifying hundreds of target SNPs from DNA extracted directly from the vector or host. GLST is advantageous when the pathogen is difficult or expensive to enrich representatively *ex vivo* or when its genome is known to contain large amounts of invariant or analytically intractable DNA. We experienced these limitations with *T. cruzi* in Chapter 2 and therefore provided first proof-of-principle for GLST using metagenomic extracts from *T. cruzi*-infected triatomines. As Section 6.4 explains, we must now extend proof-of-principle to other important pathogen systems and sample types. This work is already

underway with *Plasmodium vivax*, a very poorly culturable parasite[347] and the primary cause of malaria outside of sub-Saharan Africa[647].

Finally, we also explored the possibility to enhance predictions on parasite dispersal and selection by explicitly coupling high-resolution genetic and spatial analyses in a landscape genomic resistance modelling approach. Our theoretical framework is novel in the epidemiological context but has not yet been verified with empirical data. Section 6.5 discusses future implementation in light of the complex sources of parasite population genetic structure we encountered in Chapters 2 and 3.

6.2 Key findings, limitations and prospects of research from Chapter 2

Chapter 2's population genomic evidence for meiotic sex in clones of *T. cruzi* field isolates arguably represents the most important finding of this dissertation. *T. cruzi* has for decades been considered a paradigm of predominant clonal evolution for which nuclear genetic exchange is too rare to modulate population structure[16] and might only occur via parasexual mechanisms observed *in vitro* by Gaunt et al. (2003)[171]. The occurrence of sex in contemporary *T. cruzi* populations is so important because it accelerates genetic and phenotypic diversification. Biomarkers and genetic bases of important biomedical properties such as drug susceptibility, pathogenicity and tissue tropism become less stable and less reliably attributable to particular lineages or groups. This complicates taxonomy, diagnostics and drug design[648].

Exposing meiotic signatures in Chapter 2's WGS data was challenging due to the extremely repetitive nature of the *T. cruzi* genome. Less than 50% of the genome proves to be reliably mappable based on rates at which virtual reads (i.e., sequences created by cutting the reference assembly into segments of lengths equal to those of Illumina reads) do not correctly map back to the positions they are cut from[390]. We managed this dilemma by rigorous masking and, more importantly, by placing special focus on qualitative analyses robust to artefactual diversity that can distort inference from poorly mapping genomes. In Fig. 2.3, for example, we demonstrate linkage decay curves instead of simply quantifying mean linkage between variant sites. In Figs. 2.4 and 2.5, we pinpointed specific tree topologies that discontinuously represent the data rather than simply concluding that phylogenetic instability occurs within and among chromosomes. We also managed poor mappability[390] by exploiting three types of comparative analyses as controls. The first and most frequent type of comparison involved the parallel analysis of parasite groups representing different geographic and phylogenetic partitions of the sample set. Linkage (Fig. 2.3), recombination (Tbl. 2.2), F_{IS} (Supplementary Fig. 2.3), tree topology weighting[400] (Fig. 2.4) and somy calculations (Fig. 2.6), for example, were divided among (or organized

to visually separate) samples from El Huayco, Ardanza and Bella Maria communities in Loja Province, Ecuador. Differences observed between sample groups subjected to identical methodological procedures could thereby begin to be considered to represent true biological phenomena rather than systematic error. The second type of comparative approach involved the use of simulated control data. For example, we simulated recombinant and non-recombinant chromosome evolution to determine statistical power in Hardy-Weinberg null-hypothesis testing (Supplementary Fig. 2.3) and to validate recombination analyses with LDhat[398] (Tbl. 2.2). We also tested the effects of sample size reduction to include only one *T. cruzi* clone per infection source (e.g., Supplementary Tbl. 2.3 and Supplementary Fig. 2.7a). In our third comparative approach, we obtained Illumina reads for TcI-Sylvio (the same reads Talavera-Lopez et al. (2018) used in combination with PacBio data to construct the current reference assembly[265,306]) to optimize many methods, e.g., to specify mapping and variant filtration criteria or to calibrate the windowed somy estimation approach (Supplementary Fig. 2.14).

We achieved robust inference from Chapter 2's Illumina dataset with the above control strategies in place, but the heavy masking integral to our short-read analyses also substantially restricted precision and scope. For example, we could not comprehensively define recombination breakpoint positions across the genome in order to derive more precise estimates on the frequency of meiosis in the Bella Maria group. We also could not leverage full genomic resolution toward reliable divergence estimation (e.g., using BEAST) for Clusters 1 and 2, and analyses largely excluded information from large repetitive gene families such as DGF, mucin, MASP and GP63[265]. Many of these gene families encode surface proteins of central importance to the parasite's interactions with the vector/host immune system and its response to drugs[77,306]. They therefore represent precisely those parts of the *T. cruzi* genome in which the modification and transfer of sequence diversity through sexual reproduction is of most applied interest.

There are clearly two ways forward. First, the TcI-Sylvio reference assembly must be enhanced and further field samples sequenced using long-read systems that have rapidly advanced in the last decade. The current reference assembly was built using ca. 2 kb reads produced by the first-generation PacBio RS platform. The new RS II platform with C4 chemistry now produces average read-lengths of 10 – 15 kb and Oxford Nanopore Technologies have achieved read-lengths beyond 2 Mb, i.e., easily spanning across multi-copy gene families and even entirely encompassing most chromosomal contigs established for TcI-Sylvio thus far. Future sequencing projects should also be applied to single *T. cruzi* cells, e.g., library preparation involving fluorescence activated cell sorting or microfluidic partitioning strategies such as those introduced by 10x Genomics (www.10xgenomics.com).

Single-cell sequencing can expose haploid gametes (e.g., on the basis of genome-wide homozygosity) within diploid cell populations, pinpoint progenitors of recombinant genotypes or distinguish individual chromosomal somy differences (e.g., trisomies inferred in Fig. 2.6) when mosaic aneuploidy occurs[649]. Somy differences are important because they can expose potential non-meiotic or failed meiotic recombination events[173,650]. They may also play a role in reproductive isolation or relate to mating switches or mating types[651]. Monosomy on chromosome 13, for example, was frequently observed in putatively recombinant (and otherwise diploid) *T. cruzi* genomes from Bella Maria. This somy pattern was not observed in genomes from El Huayco and Ardanza, where trisomies were more common and clonality appeared to prevail.

The second way to advance understanding of genetic exchange in *T. cruzi* leads into the lab. It may be possible to specify the frequency, anatomical site and life cycle stage at which genetic exchange occurs using *in vitro* and *in vivo* models by creating mixed infections with potentially hybridizing parasite cells (e.g., clones from Bella Maria) modified to express distinct bioluminescent (i.e., luciferases) and/or fluorescent proteins. Cells exhibiting co-fluorescence can indicate hybrid progeny formation and can be further analyzed[652]. Genetic engineering tools have, like sequencing technologies, rapidly advanced in recent years and a CRISPR/Cas9 gene editing system[653] required for fluorescent hybrid detection has already been established for *T. cruzi* by collaborators at LSHTM[654]. The system integrates T7 RNA polymerase and Cas9 genes into ribosomal gene arrays and uses PCR products as guide RNA. The homology-directed repair templates encode luciferase/mNeonGreen or luciferase/mScarlet fusion proteins and are likewise transfected as PCR products, i.e., no cloning step is required. This enables high-throughput genome-editing and fluorescence tracking throughout the parasite life cycle, also in fixed cells (mNeonGreen and mScarlet fluorescence does not require ATP)[654].

The Machado group at Universidade Federal de Minas Gerais in Brazil also describes an alternative to genome-editing to create distinctly fluorescing *T. cruzi* lines[655]. The approach cultures parasites in media containing the nucleoside analogues 5′-chloro-2′-deoxyuridine or 5′-iodo-2′-deoxyuridine. These molecules are incorporated into the parasite DNA (replacing thymidine) and give distinct signals (red and green, respectively) after immunostaining cells. This method represents a valuable complement to fluorescent tagging via genome-editing because it is less likely to involve side effects on cellular physiology or alter survival and fitness. Genome-editing, however, also enables gene expression detection, overexpression or knockout studies, not just the color-tagging of parasite cells. For example, the CRISPR/Cas9 system could be used to complement transcriptomic studies in profiling the activity and timing of molecular machinery suspected to underlie genetic exchange processes

inferred from Chapter 2. These studies could begin to focus on parts of the 'meiosis detection tool kit', a small set of genes with homologs in animals, plants, fungi and protists that are expressed only during meiosis and for which null mutations do not affect other cellular processes or traits[656]. These include SPO11, a topoisomerase required for DNA double strand breaks that initiate meiotic recombination; HOP1, a synaptonemal complex protein that promotes chromosome pairing by oligomerizing at double strand break regions; and DMC1, also a vital synaptonemal complex protein that facilitates homologous recombination by forming specialized filaments with single-stranded DNA[656]. SPO11 is interesting because it occurs in two homologs within *T. cruzi* and related trypanosomatid genomes, but the function of these homologs have not yet been analyzed[657]. HOP1 and DMC1 are interesting because their expression is known to coincide with *T. b. brucei* gamete production in the salivary glands of the tsetse fly[658]. Both genes are also expressed in insect-stage metacyclic *L. major* promastigotes, although an association to meiotic division has not been confirmed in the latter genus[659]. Next to the meiosis detection toolkit, it would also be interesting to further profile the activity of RAD51, a recombinase protein homolog of DMC1 that was recently shown to promote the occurrence of fused-cell hybrids in *T. cruzi* epimastigote culture[655]. RAD51 is known to interact with DMC1 prior to meiotic synapsis but also represents the strand exchange protein vital to mitotic recombinational repair of eukaryotic DNA[660].

In establishing the mechanism and frequency of genetic recombination in *T. cruzi*, above types of genetic engineering studies will also help determine the potential for experimental quantitative genetic approaches to help identify the genetic bases of epidemiologically relevant traits. If meioses can be frequently induced in the lab, then crossing systems could conceivably create a spectrum of genetic and phenotypic diversity among hybrid progeny using parental lines that differ in biomedical properties such as drug susceptibility or virulence. Phenotyping and long-read WGS applied to these progeny would then enable regression analyses that predict causal variants or gauge the extent of polygenicity underlying the phenotype. The sequencing of hybrids and progenitors could also help establish how often recombination occurs in the surface-gene families we were unable to interrogate in short-read analysis. These are very ambitious objectives, but they are enheartened by previous successes using forward genetics with related trypanosomatid species. For example, quantitative trait locus (QTL) mapping by Morrison et al. (2009)[661] associated levels of spleen and liver enlargement in *T. brucei*-infected mice to sequence variation within a 100 kb region of interest in the parasite's ca. 26 Mb genome[662]. Additional markers have narrowed down the splenomegaly QTL to a set of just 52 genes and reverse genetic tools are being designed to further specificize causal variants[663].

6.3 Key findings, limitations and prospects of research from Chapter 3

Chapter 3 used a combination of phylogenomic and phenotyping approaches in its attempt to reconstruct *L. infantum* divergence histories and tease apart the roles of demographic changes and selection processes during range expansion into the New World. Like in Chapter 2, a key finding from this research was the presence of genetic exchange. It was not simply the detection of genetic exchange, however, that was most interesting in our study of *L. infantum* given that a considerable body of experimental[154,244,245,293,294] and field evidence[246,295,297,298,664] for periodic meiosis-like sex within and between *Leishmania* species has accumulated over recent decades. Several studies have also demonstrated novel phenotypic variability among *Leishmania* hybrids and their parental lines[154,295,294]. What was most important from our observations of genetic exchange in *L. infantum* was that we could reconcile the frequent and unmistakable presence of intra-specific hybridizations with recent (post-Columbian) parasite demographic restructuring linked to range expansion (divergent contact between bottlenecked subpopulations, perhaps subpopulations that separately entered the New World) and that these hybridizations restore gene function at a genetic locus that controls sensitivity to miltefosine, a front-line anti-leishmanial drug[258,259] (see Fig. 3.6a). The abundant signs of both outcrossing and endogamic genetic exchange (recall ubiquitous excess homozygosity in Fig. 3.3) may also explain the success we had with fastsimcoal2[315] models in which genetic (re-) connectivity between demes was simulated to involve classic Mendelian mating events.

Among the weaker elements of Chapter 3 was our suggestion that convergent selection processes (not simply founder effects) have contributed to the widespread proliferation of *L. infantum* isolates with genomes in which the recently identified miltefosine sensitivity locus has been fully deleted from chr31. This theory hinges on only two pieces of evidence. First, we did not observe perfect monophyly for deletion-carrying (Del) isolates because the phylogenetic positions of a small group on non-deletion type (NonDel) isolates were found to nest within the former clade (Fig. 3.4). Absence of monophyly is consistent with multiple deletion origins as opposed to widespread deletion inheritance from a common ancestral mutant. However, it could be that cryptic introgression has broken up Del monophyly and misled us to the conclusion that these isolates represent a true paraphyletic group. However, Del paraphyly was supported by the second piece of evidence that deletion locus boundaries covary with phylogenetic variation in the dataset. Distinct, phylogenetically correlated deletion architecture suggests the occurrence of independent (convergent) deletion origins in different clades. Our method of deletion locus boundary detection, however, was based solely on read-depth analysis and certainly requires additional validation, e.g., using Sanger sequencing of PCR amplicons that span the junctions formed between deletion breakpoints.

Furthermore, it should be verified whether repeat motifs around the deletion locus support the generation of variable deletion sizes via homologous recombination or whether such variation more likely stems from smaller INDEL mutations subsequent to a shared ancestral deletion event.

Pursuing the possibility that selection is contributing to *L. infantum* genetic differentiation in the New World should also involve better measurement of parasite fitness proxies in host and vector stages, ideally using larger sample sizes than those used in our phenotypic (ecto-3'-nucleotidase and ecto-ATPase activity) assays. First steps in this direction might, for example, compare Del vs. NonDel susceptibility to neutrophil extracellular traps. Parasite capture by these web-like chromatin structures represents an essential component of the host immune response during early stages of infection and is known to vary in efficacy depending on the level of ecto-3'-nucleotidase activity in *L. infantum* promastigotes[448]. Our collaborators at FIOCRUZ have already begun this research. Considering parasite performance in the vector, it would be interesting to assess infectivity (e.g., quantify post-bloodmeal parasitaemia at different timepoints) and transmissivity (e.g., quantify infective dose after parasite maturation/migration to the salivary glands) of Del and NonDel isolates. It would be especially informative to perform these assays on both *Lu. longipalpis* and *Lu. cruzi* vectors given a loose association observed between the geographic distributions of these sand fly species and *L. infantum* population genetic subdivision in western Brazil. The use of induced or natural (second-generation) parasite hybrids could also significantly improve the ability of host and vector infection assays to advance understanding on the evolutionary significance of the chr31 deletion and other sequence variants. Inference in Chapter 3 was often challenged by the fact that phenotypes and sequence variants of interest occurred on very few genetic backgrounds (one of two phylogenetically divergent groups contained most non-deletion type isolates and the other, very homogeneous group contained all deletion-carrying isolates), making the dataset's few hybrids very valuable in helping expose confounding kinship effects (e.g., in Supplementary Fig. 3.3, see how samples such as the putative F_2 recombinant NonDel_MT_3210 help clarify that genome-wide gene copy number variation is predicted by geographic (state) origin and not – as it might first appear – by presence/absence of the chr31 deletion locus).

Assessing more hybrids and detecting more divergent subpopulations will not only improve genotype-genotype and genotype-phenotype association studies but also help to resolve whether *L. infantum* was just once or many times introduced to the New World. This question was another that could not be resolved definitively in Chapter 3. We recommend attention to the strongest possible sources of vicariance (e.g., the Andes) and regions representing different (e.g., Spanish and French) colonization histories in future sampling

designs. Additional, possibly GLST-based *L. infantum* typing in Brazil (e.g., filling sampling gaps in Paraná, Goiás and Tocantins) and other Latin American countries (e.g., where lower-elevation corridors cross the Andes in northern Colombia or southern Ecuador) might also facilitate landscape genetic approaches towards a more conclusive reconstruction of New World colonization events (see further discussion in Section 6.5).

All these future efforts, however, are likely to experience low American *L. infantum* genetic diversity as a limiting factor in some stage of analysis as we did time and again in Chapter 3. Diversity was too low, for example, to make use of intra-chromosomal linkage patterns for robust verification of chr31 deletion convergence or to clarify whether backcrossing events are responsible for the nested positions of some NonDel isolates within the Del clade (see further above). Establishing whether the low levels of polymorphism expected to occur in most American *L. infantum* datasets (e.g., see the short branch lengths in Fig. 3.4's phylogenetic tree, including those of geographically disparate, Honduran isolates) are sufficient to answer the question of interest is therefore paramount to study proposal and sampling design. Simulated genotypes, perhaps involving 'spiked mutations'[399] as did our controls from Chapter 2, may help achieve the power analyses required. It is also recommended to complement analysis of the nuclear genome with that of the kinetoplast DNA, which we did not yet complete in Chapter 3. A number of studies on *T. cruzi* have reported mitochondrial recombination without detectable nuclear genetic exchange[170]. If this phenomenon is occurring in *L. infantum*, kinetoplast sequence variation may expose past demographic processes that are not chronicled in the nuclear genome. One may also find cases where genetic signals of interest have become homogenized in both nuclear and kinetoplast minicircle sequences but remain pure in the maxicircle DNA (unlike minicircles, maxicircles do not appear to mosaicize (or maintain heteroplasmies) in *Leishmania* spp.)[3].

6.4 Key findings, limitations and prospects of research from Chapter 4

Chapter 4 developed a multiplexed amplicon sequencing strategy we refer to as GLST to measure genome-wide pathogen sequence variation using uncultured sample types. The simple PCR-based 'genome-typing' strategy is valuable because culture-based methods often introduce selection bias and require resources inaccessible to many labs, especially those operating in less developed countries and/or where endemic pathogen transmission is most relevant to public health. We provided proof-of-principle by applying GLST to metagenomic DNA extracts from the intestinal tracts of naturally infected triatomines. GLST detected 368 SNP variants in 203 *T. cruzi* amplicons co-amplified from these vector samples and hundreds more in amplicon libraries created for TcIII, TcIV and TcVI reference clones. GLST thereby achieved important resolution benchmarks, including the detection of

isolation-by-distance relationships within TcI and the potential for multiple-lineage analysis. However, the study only used one uncultured sample type and the sample set represented a medley of donations from collaborators without a specific epidemiological question in mind. It is therefore now clearly the next step to demonstrate the transferability of GLST to different sample types and pathogen systems whilst simultaneously pursuing a specific research goal beyond that of method development. We have therefore already designed a second GLST panel for *P. vivax* with the intention of tracking a major malaria epidemic emanating from the Venezuelan Amazon. Desperate socioeconomic circumstances have led a growing number of people to work in illegal gold mining areas, especially at a mine known as Las Cristinas in Bolivar state. Frequent migrations to/from Las Cristinas (often by immunologically naïve people), rapid deforestation and the general collapse of health infrastructure (no drugs, diagnostics, vector control, etc.) are fueling a malaria outbreak of unprecedented proportions in the region and reshaping malaria epidemiology at the national scale[665].

Our new *P. vivax* GLST panel co-amplifies 107 SNP loci identified in WGS data by Oliveira et al. (2017)[666]. This publicly available sequencing project contains the short-read data of 84 *P. vivax* clones from Mexico, Peru, Colombia and Brazil. We singled out these 107 PCR-multiplexable SNP loci because each locus shows polymorphism in clones from all four countries of the study. Each locus, however, is polymorphic in at most 50% of each country's clones. Finally, each locus represents noncoding DNA. We expect these criteria will maximize our chances to detect neutral sequence variation suitable for epidemiological tracking within Venezuela. The new study will apply GLST to DNA extracted from FTA cards containing the blood of *P. vivax*-infected patients visiting the Instituto de Medicina Tropical of the Universidad Central de Venezuela in Caracas (many thanks to Oscar Noya) as well as from desiccated mosquitos captured in Bolivar state (many thanks to Jorge Moreno). We have metainformation revealing that the majority of malaria patients at the clinic in Caracas contracted their *P. vivax* infections during travels to Bolivar state. Fellow PhD student Antonella Bacigalupo has already successfully amplified the blood spot samples in first-round GLST reaction and hopes to achieve the same for the mosquito sample set (huge thanks to Marnie Davidson for preparing metagenomic extracts). The idea is to determine whether our GLST measurement in blood and mosquitos can predict the metainformation we have about the malaria patients' prior whereabouts in Bolivar state. The mosquito sample set also covers intra- and peri-domestic collections between 2014 and 2017 such that we can assess spatio-temporal changes to parasite genetic diversity during an exceptionally steep rise in malaria prevalence (316,401 *P. vivax* infections were recorded in Venezuela in 2017 vs. 62,850 in 2014[665]). Furthermore, DNA was extracted separately from

head and abdominal sections (another huge thanks to Marnie Davidson) of *Anopheles darlingi* and *An. albitarsis* such that it may be possible to compare parasite genetic diversity and genotype-specific transmissivity (i.e., colonization of the salivary gland) between primary (*An. darlingi*) and secondary (*An. albitarsis*) vector species.

More generally, it would also be interesting to compare and complement epidemiological inferences made from neutral GLST marker sets with those from vir gene or – in the case of *P. falciparum* – var gene analysis. These are hypervariable, *Plasmodium*-specific multi-copy gene families key to antigenic diversity and cytoadherence[667]. Some subsections are also amenable to conventional amplicon sequencing, i.e., using a single primer pair in PCR[668]. Clustering analysis of var gene DBLα amplicon reads has become a powerful approach in studying immune selection but diversity is often considered too complex to be tractable for dispersal studies beyond the most microgeographic of scales[668,669].

Comparison of our *P. vivax* GLST panel to the 71-SNP barcode recently introduced by Benavente et al. (2020) is also of high interest. The authors used linkage block tagging[670] and machine-learning methods to find SNPs with maximal predictive power for *P. vivax* source tracking (to the country level) but did not yet design a delivery system for these SNPs. Both barcode design and implementation occurred *in silico* using WGS data and design did not test amenability to (multiplexable) PCR or other non-WGS genotyping techniques. Combining our focus on panel adjustability and multiplexable deliverability with the elaborate power optimization strategies demonstrated by Benavente et al. is an exciting prospect for future research.

6.5 Key concepts, limitations and prospects of research from Chapter 5

Chapter 5 proposed a new landscape genetic framework to better understand the spread of vector-borne disease through heterogeneous environment. We defined a pattern-process modelling workflow that compares observed parasite genetic structure with that simulated over a digital resistance surface summarizing hypothesized effects of (remotely sensed) landscape features on parasite transmission among vectors and hosts. However, the complexity of our step-by-step illustrations in Figs. 5.1 and 5.2 reflects the fact that implementation may not be so straight-forward in many contexts.

New insights into *T. cruzi* and *L. infantum* demography from Chapters 2 to 4 suggest that parasite genetic datasets can harbor complex genetic variation controlled by unmeasured or non-environmental processes and traits. Such features may confound landscape genetic analysis. For example, we suggested that reproductive polymorphism genetically segregates sympatric TcI populations in Loja Province, southern Ecuador. It is unclear whether this

polymorphism has any (measurable) association to the intervening or local environment. In the New World *L. infantum* system, we emphasized historic demographic changes that may confound or obscure the detection of contemporary landscape genetic effects (the 'ghosts of landscape past'[361]). Genetic disorganization from bottlenecks and secondary contact resulting from range expansion, for example, are likely to compound possible influences of ecological variation (e.g., changes in Lu. *longipalpis* vs. *Lu. cruzi* abundance, transitions between savanna (cerrado) and semideciduous forest or to urban zones) on *L. infantum* diversity in southwestern Brazil. Extensive follow-up is required to understand how best to incorporate (or whether one must avoid study foci containing) these complexities in a landscape genetic simulation modelling approach.

We hope to contribute to this follow-up in our upcoming attempts to use landscape genetic simulation to resolve whether *L. infantum* expanded into the New World from a single or multiple introduction events. These attempts will not yet involve Chapter 5's framework in its fully-fledged form but rather exploit selected simulation features at larger spatial scales. Specifically, we plan to simulate *L. infantum* gene flow on a rudimentary resistance raster (incorporating only road networks because these represent a strong proxy for urbanization and the dispersal of both dogs and sandflies[671]) and focus on the effect of one vs. two input (founder) groups. The landscape genetic simulator CDMetaPOP[326] can handle multiple input locations, e.g., one in in the Northeast of Brazil (e.g., Fortaleza or Recife) and another at a Spanish colonial port such as Buenos Aires, Argentina, which is not separated by the Andes from Brazil.

This question on single vs. multiple introductions aside, it will also be interesting to test our landscape genetic framework in southeastern Brazil, e.g., in landscapes within the states of Espírito Santo and/or Rio de Janeiro, where *Lutzomyia* distribution probabilities are heterogeneous[672] and parasite sampling is more likely to involve genotypes belonging to a single invasion process from the Atlantic Coast. This system might even prove more tractable to landscape genetic simulation than that of *T. cruzi* in Loja Province because *L. infantum* host/vector spectrum is much less complex. *L. infantum* essentially uses just three host/vector species in the New World (dogs, humans and *Lutzomyia longipalpis*) and cryptic niche differentiation (e.g., haplotype-specific vectorial capacity[673] or segregated arboreal and terrestrial transmission cycles[149]) is unlikely to complicate analyses given the very little time the parasite has been evolving on the American continent.

We must also verify our landscape genetic predictive framework on *T. cruzi* as initially proposed in the rural landscapes of Loja Province. Implementation in Loja will be guided by recent landscape genetic models of vector dispersal by fellow PhD student Luis Hernández

(dissertation currently in review). Using a genetic algorithm (GA) resistance surface optimization approach[674], Luis exposed road configuration as a primary determinant of gene flow in *Rhodnius ecuadoriensis*, the primary vector of Chagas disease in the study region. It will be interesting to quantify the extent to which *T. cruzi* gene flow mirrors this relationship (perhaps with higher sensitivity due to higher mutation rates[675]) and more generally to examine hypotheses of parasite-vector genetic co-structure in the landscape. Co-structure analysis is especially intriguing in this system because 'paired genome-typing' (i.e., acquiring genome-wide SNP data from both the parasite infrapopulation and the vector individual representing each infection) has become increasingly viable with the arrival of GLST for *T. cruzi* and Luis' 2b-RAD system for *R. ecuadoriensis*[646]. Identifying local landscape conditions where parasite-vector genotype pairs deviate from global patterns of covariation may help refine landscape genetic models for each protagonist.

Finally, we also aim to implement landscape genetic approaches in more densely populated regions, specifically in the Metropolitan District of Caracas (MDC) in Venezuela. Chapter 4 illuminated extraordinary levels of TcI diversity within the MDC and its patchwork of urban, semi-urban and sylvatic environments spread across complex altitudinal relief. It is critical to understand how regional parasite diversity is filtering into the city and threatening human lives. Additional leverage using GLST is also especially promising here because we can integrate the prolific citizen science triatomine collection program managed by our collaborators at the Universidad Central de Venezuela. Not only will this help satisfy data-hungry landscape genetic simulators but it is very important to help build public awareness on the risks of vector-borne disease. This social component is especially relevant when the public health benefit of the project may not be immediate or when uncertainties like ours on reproductive polymorphism in *T. cruzi* may complicate initial aims of research.

6.6 Final reflections

This PhD plunged into a great complexity of research topics and bioinformatic techniques. Thousands of hours were spent in a grueling virtual underworld where sanity can easily be lost. Several big pictures could nevertheless be apprehended, and sanity has remained relatively intact. We advanced fundamental theory on two dangerous parasite genera by exposing meiotic population genetic signatures in *T. cruzi* and reconciling hidden diversification and convergence processes in *L. infantum* with the evolutionarily recent spillover of visceral leishmaniasis to the Americas from the Old World. Both of these research outputs have important applied consequences. Sex creates new, potentially harmful phenotypic diversity and thereby complicates surveillance and treatment. On the bright side, however, it also brings new opportunities to quantitative genetic research. Hidden

diversification in severely bottlenecked *Leishmania* populations dismisses the common assumption that spillover events reduce parasite diversity to the extent that only host and/or environmental variation is likely to explain variability in disease phenotypes. This realization is crucial to a better understanding of unexpected clinical outcomes of American visceral leishmaniasis observed in recent years. This dissertation also introduced new technical and conceptual frameworks for epidemiological research. Chapter 4's pathogen barcoding technology substantially reduces the costs of genome-wide polymorphism analysis and can therefore help studies achieve spatial sampling designs required for meaningful inference. We observed time and again in this dissertation that the ability to maximize sample sizes and to optimize spatial sampling configurations is paramount to study success. Chapter 5's landscape genomic framework is notable in that it repurposes a traditionally conservation genetic study apparatus for the opposite objective of eradicating parasitic disease. The framework requires high sampling effort, but ideally this downside will encourage project designs that generate additional value in the process of data collection. It is only ethical that field expeditions simultaneously serve to screen at-risk human populations and bring medical attention when infections are found. General awareness-building is also essential, especially considering diseases such as Chagas for which infection is largely preventable so long as one knows that triatomines transmit the parasite and that simple lifestyle changes can minimize triatomine colonization of the domestic environment.

Many such lessons spring from this PhD's quest to advance epidemiological theory and pathogen surveillance tools. While demonstrating to what great extent whole-genome or genome-wide polymorphism analysis can help clarify fundamental biological questions on important vector-borne parasites, the dissertation also demonstrates that this power remains contingent on many elements of study design. While genomic analysis is increasingly advertised as 'push-button' exercise[676,677], various examples described herein emphasize that computational pipelines can require very careful honing and that pre-sequencing study decisions (e.g., spatial sampling design and strategies used to characterize multiclonal infections) are as important as ever to robust inference. Continuing to advance cross-disciplinary research platforms is also key because complex disease systems can only be understood so far when analyses on parasite, vector, host and environmental variables remain discrete. This dissertation should provide an important reference for the great amount of work that lies ahead.

References

1. Pérez-Molina, J. A. & Molina, I. Chagas disease. *Lancet* **391**, 82–94 (2018).
2. Montgomery, S. P., Starr, M. C., Cantey, P. T., Edwards, M. S. & Meymandi, S. K. Neglected parasitic infections in the United States: Chagas disease. *Am. J. Trop. Med. Hyg.* **90**, 814–818 (2014).
3. Conteh, L., Engels, T. & Molyneux, D. H. Socioeconomic aspects of neglected tropical diseases. *Lancet* **375**, 239–247 (2010).
4. Burza, S., Croft, S. L. & Boelaert, M. Leishmaniasis. *Lancet* **392**, 951–970 (2018).
5. Alvar, J. et al. Leishmaniasis worldwide and global estimates of its incidence. *PLoS One* **7**, e35671 (2012).
6. Ramírez, J. D. & Llewellyn, M. S. Reproductive clonality in protozoan pathogens--truth or artefact? *Mol. Ecol.* **23**, 4195–4202 (2014).
7. de Meeûs, T. et al. Population genetics and molecular epidemiology or how to "débusquer la bête". *Infect. Genet. Evol.* **7**, 308–332 (2007).
8. Steinauer, M. L., Blouin, M. S. & Criscione, C. D. Applying evolutionary genetics to schistosome epidemiology. *Infect. Genet. Evol.* **10**, 433–443 (2010).
9. Gilabert, A. & Wasmuth, J. D. Unravelling parasitic nematode natural history using population genetics. *Trends Parasitol.* **29**, 438–448 (2013).
10. Tibayrenc, M. & Ayala, F. J. The clonal theory of parasitic protozoa: 12 years on. *Trends Parasitol.* **18**, 405–410 (2002).
11. Andras, J. P. & Ebert, D. A novel approach to parasite population genetics: experimental infection reveals geographic differentiation, recombination and host-mediated population structure in *Pasteuria ramosa* , a bacterial parasite of *Daphnia*. *Mol. Ecol.* **22**, 972–986 (2013).
12. Telleria, J. & Tibayrenc, M. American trypanosomiasis, Chagas disease: one hundred years of research. *Elsevier* (2010).
13. Zingales, B. et al. The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. *Infect. Genet. Evol.* **12**, 240–253 (2012).
14. Ramírez, J. D. et al. Contemporary cryptic sexuality in *Trypanosoma cruzi*. *Mol. Ecol.* **21**, 4216–4226 (2012).
15. Messenger, L. A., Miles, M. A. & Bern, C. Between a bug and a hard place: *Trypanosoma cruzi* genetic diversity and the clinical outcomes of Chagas disease. *Expert Rev. Anti-infect. Ther.* **13**, 995–1029 (2015).

16. Tibayrenc, M. & Ayala, F. J. *Trypanosoma cruzi* and the model of predominant clonal evolution; in American trypanosomiasis, Chagas disease: one hundred years of research (2nd edition) 475–495. *Elsevier* (2017).
17. Maurício, I. L., Stothard, J. R. & Miles, M. A. The strange case of *Leishmania chagasi*. *Parasitol. Today* **16**, 188–189 (2000).
18. Kuhls, K. et al. Comparative microsatellite typing of New World *Leishmania infantum* reveals low heterogeneity among populations and its recent Old World origin. *PLoS Negl. Trop. Dis.* **5**, e1155 (2011).
19. Leblois, R., Kuhls, K., François, O., Schönian, G. & Wirth, T. Guns, germs and dogs: on the origin of *Leishmania chagasi*. *Infect. Genet. Evol.* **11**, 1091–1095 (2011).
20. Laffitte, M.-C. N., Leprohon, P., Papadopoulou, B. & Ouellette, M. Plasticity of the *Leishmania* genome leading to gene copy number variations and drug resistance. *F1000Res.* **5**, 2350 (2016).
21. Rogers, M. B. et al. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania. Genome Res.* **21**, 2129–2142 (2011).
22. Rougeron, V., de Meeûs, T., Kako-Ouraga, S., Hide, M. & Bañuls, A. L. 'Everything you always wanted to know about sex (but were afraid to ask)' in *Leishmania* after two decades of laboratory and field analyses. *PLoS Pathog.* 6, e 1001004 (2010).
23. Balkenhol, N., Cushman, S., Storfer, A. & Waits, L. Landscape genetics: concepts, methods, applications. *John Wiley & Sons* (2015).
24. Reisen, W. K. Landscape epidemiology of vector-borne diseases. *Annu. Rev. Entomol.* **55**, 461–483 (2010).
25. LaDeau, S. L., Allan, B. F., Leisnham, P. T. & Levy, M. Z. The ecological foundations of transmission potential and vector-borne disease in urban landscapes. *Funct. Ecol.* **29**, 889–901 (2015).
26. Chagas disease in Latin America: an epidemiological update based on 2010 estimates. *Wkly. Epidemiol. Rec.* **90**, 33–43 (2015).
27. Vos, T. et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *Lancet* **386**, 743–800 (2015).
28. Wygodzinsky, P. W. & Lent, H. Revision of the Triatominae (Hemiptera, Reduviidae), and their significance as vectors of Chagas' disease. *Bulletin of the AMNH* **163**, 123–520 (1979).

29. WHO expert committee. Control of Chagas disease. *World Health Organ. Tech. Rep. Ser.* **905**, 1–109 (2002).
30. Stevens, L. et al. Kissing bugs, the vectors of Chagas. *Adv. Parasitol.* **75**, 169–192 (2011).
31. Nagajyothi, F. et al. Mechanisms of *Trypanosoma cruzi* persistence in Chagas disease. *Cell. Microbiol.* **14**, 634–643 (2012).
32. Jurberg, J. & Galvão, C. Biology, ecology, and systematics of Triatominae (Heteroptera, Reduviidae), vectors of Chagas disease, and implications for human health. *Denisia* **19**, 1095–1116 (2006).
33. Jansen, A. M., Xavier, S. C. das C. & Roque, A. L. R. *Trypanosoma cruzi* transmission in the wild and its most important reservoir hosts in Brazil. *Parasit. Vectors* **11**, 502 (2018).
34. Carreira, J. C., Jansen, A. M., de Nazareth Meirelles, M., Costa e Silva, F. & Lenzi, H. L. *Trypanosoma cruzi* in the scent glands of *Didelphis marsupialis*: the kinetics of colonization. *Exp. Parasitol.* **97**, 129–140 (2001).
35. Pérez-Molina, J. A., Perez, A. M., Norman, F. F., Monge-Maillo, B. & López-Vélez, R. Old and new challenges in Chagas disease. *Lancet Infect. Dis.* **15**, 1347–1356 (2015).
36. Lee, B. Y., Bacon, K. M., Bottazzi, M. E. & Hotez, P. J. Global economic burden of Chagas disease: a computational simulation model. *Lancet Infect. Dis.* **13**, 342–348 (2013).
37. Sales Junior, P. A. et al. Experimental and clinical treatment of Chagas disease: a review. *Am. J. Trop. Med. Hyg.* **97**, 1289–1303 (2017).
38. Jackson, Y. et al. Tolerance and safety of nifurtimox in patients with chronic Chagas disease. *Clin. Infect. Dis.* **51**, e69-75 (2010).
39. Pérez-Molina, J. A. et al. Nifurtimox therapy for Chagas disease does not cause hypersensitivity reactions in patients with such previous adverse reactions during benznidazole treatment. *Acta Trop.* **127**, 101–104 (2013).
40. Bonney, K. M. Chagas disease in the 21st century: a public health success or an emerging threat? *Parasite* **21**, 11 (2014).
41. Tarleton, R. L., Gürtler, R. E., Urbina, J. A., Ramsey, J. & Viotti, R. Chagas disease and the London Declaration on Neglected Tropical Diseases. *PLoS Negl. Trop. Dis.* **8**, e3219 (2014).
42. Abad-Franch, F. et al. Ecology, evolution, and the long-term surveillance of vector-borne Chagas disease: a multi-scale appraisal of the tribe Rhodniini (Triatominae). *Acta Trop.* **110**, 159–177 (2009).

43. Westenberger, S. J., Barnabé, C., Campbell, D. A. & Sturm, N. R. Two hybridization events define the population structure of *Trypanosoma cruzi*. *Genetics* **171**, 527–543 (2005).
44. de Freitas, J. M. et al. Ancestral genomes, sex, and the population structure of *Trypanosoma cruzi*. *PLoS Pathog.* **2**, e24 (2006).
45. Tibayrenc, M., Ward, P., Moya, A. & Ayala, F. J. Natural populations of *Trypanosoma cruzi*, the agent of Chagas disease, have a complex multiclonal structure. *Proc. Natl. Acad. Sci. USA* **83**, 115–119 (1986).
46. Tibayrenc, M. & Ayala, F. J. Reproductive clonality in protozoan pathogens--truth or artifact? A comment on Ramírez and Llewellyn. *Mol. Ecol.* **24**, 5778–5781 (2015).
47. Ramírez, J. D. & Llewellyn, M. S. Reproductive clonality in protozoan pathogens – truth or artefact? Response to Tibayrenc and Ayala. *Mol. Ecol.* **24**, 5782–5784 (2015).
48. Tibayrenc, M. Genetic epidemiology of parasitic protozoa and other infectious agents: the need for an integrated approach. *Int. J. Parasitol.* **28**, 85–104 (1998).
49. Fernandes, O. et al. Brazilian isolates of *Trypanosoma cruzi* from humans and triatomines classified into two lineages using mini-exon and ribosomal RNA sequences. *Am. J. Trop. Med. Hyg.* **58**, 807–811 (1998).
50. Brisse, S., Verhoef, J. & Tibayrenc, M. Characterisation of large and small subunit rRNA and mini-exon genes further supports the distinction of six *Trypanosoma cruzi* lineages. *Int. J. Parasitol.* **31**, 1218–1226 (2001).
51. Yeo, M. et al. Origins of Chagas disease: *Didelphis* species are natural hosts of *Trypanosoma cruzi* I and armadillos hosts of *Trypanosoma cruzi* II, including hybrids. *Int. J. Parasitol.* **35**, 225–233 (2005).
52. Souto, R. P., Fernandes, O, Macedo, A. M., Campbell, D. A. & Zingales, B. DNA markers define two major phylogenetic lineages of *Trypanosoma cruzi*. *Mol. Chem. Parasitol.* 78, 127–137 (1996).
53. Kawashita, S. Y., Sanson, G. F., Fernandes, O., Zingales, B. & Briones, M. R. Maximum-likelihood divergence date estimates based on rRNA gene sequences suggest two scenarios of *Trypanosoma cruzi* intraspecific evolution. *Mol. Biol. Evol.* **18**, 2250–2259 (2001).
54. Marcili, A. et al. A new genotype of *Trypanosoma cruzi* associated with bats evidenced by phylogenetic analyses using SSU rDNA, cytochrome b and histone H2B genes and genotyping based on ITS1 rDNA. *Parasitology* **136**, 641–655 (2009).
55. Brenière, S. F., Waleckx, E. & Barnabé, C. Over six thousand *Trypanosoma cruzi* strains classified into discrete typing units (DTUs): attempt at an inventory. *PLoS Negl. Trop. Dis.* **10**, e0004792 (2016).

56. Llewellyn, M. S. et al. Genome-scale multilocus microsatellite typing of *Trypanosoma cruzi* discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection. *PLoS Pathog.* **5**, e1000410 (2009).
57. Cosentino, R. O. & Agüero, F. A simple strain typing assay for *Trypanosoma cruzi*: discrimination of major evolutionary lineages from a single amplification product. *PLoS Negl. Trop. Dis.* **6**, e1777 (2012).
58. Messenger, L. A., Yeo, M., Lewis, M. D., Llewellyn, M. S. & Miles, M. A. Molecular genotyping of *Trypanosoma cruzi* for lineage assignment and population genetics. *Methods Mol. Biol.* **1201**, 297–337 (2015).
59. Fernandes, O. et al. The complexity of the sylvatic cycle of *Trypanosoma cruzi* in Rio de Janeiro state (Brazil) revealed by the non-transcribed spacer of the mini-exon gene. *Parasitology* **118**, 161–166 (1999).
60. Lisboa, C. V., Pinho, A. P., Monteiro, R. V. & Jansen, A. M. *Trypanosoma cruzi* (Kinetoplastida, Trypanosomatidae): biological heterogeneity in the isolates derived from wild hosts. *Exp. Parasitol.* **116**, 150–155 (2007).
61. Kerr, C. L. et al. Lineage-specific serology confirms Brazilian Atlantic forest lion tamarins, *Leontopithecus chrysomelas* and *Leontopithecus rosalia*, as reservoir hosts of *Trypanosoma cruzi* II (TcII). *Parasit. Vectors* **9**, 584 (2016).
62. Miles, M. A. et al. The molecular epidemiology and phylogeography of *Trypanosoma cruzi* and parallel research on *Leishmania*: looking back and to the future. *Parasitology* **136**, 1509–1528 (2009).
63. Messenger, L. A., Ramirez, J. D., Llewellyn, M. S., Guhl, F. & Miles, M. A. Importation of hybrid human-associated *Trypanosoma cruzi* strains of southern South American origin, Colombia. *Emerg. Infect. Dis.* **22**, 1452–1455 (2016).
64. Tibayrenc, M., Neubauer, K., Barnabé C., Guerrini, F., Skarecky, D. & Ayala, F. J. Genetic characterization of six parasitic protozoa: parity between random-primer DNA typing and multilocus enzyme electrophoresis. *Proc. Natl. Acad. Sci. USA* **90**, 1335–1339 (1993).
65. Souto, R. P. & Zingales, B. Sensitive detection and strain classification of *Trypanosoma cruzi* by amplification of a ribosomal RNA sequence. *Mol. Biochem. Parasitol.* **62**, 45–52 (1993).
66. Tibayrenc, M. Population genetics of parasitic protozoa and other microorganisms. *Adv. Parasitol.* **36**, 47–115 (1995).
67. Brisse, S., Barnabé, C. & Tibayrenc, M. Identification of six *Trypanosoma cruzi* phylogenetic lineages by random amplified polymorphic DNA and multilocus enzyme electrophoresis. *Int. J. Parasitol.* **30**, 35–44 (2000).

68. Baptista, R. P. et al. Assembly of highly repetitive genomes using short reads: the genome of discrete typing unit III *Trypanosoma cruzi* strain 231. *Microb. Genomics* **4**, e000156 (2018).
69. Lewis, M. D. et al. Recent, independent and anthropogenic origins of *Trypanosoma cruzi* hybrids. *PLoS Negl. Trop. Dis.* **5**, e1363 (2011).
70. Machado, C. A. & Ayala, F. J. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of *Trypanosoma cruzi*. *Proc. Natl. Acad. Sci. USA* **98**, 7396–7401 (2001).
71. Flores-López, C. A. & Machado, C. A. Analyses of 32 loci clarify phylogenetic relationships among *Trypanosoma cruzi* lineages and support a single hybridization prior to human contact. *PLoS Negl. Trop. Dis.* **5**, e1272 (2011).
72. Brisse, S. et al. A phylogenetic analysis of the *Trypanosoma cruzi* genome project CL Brener reference strain by multilocus enzyme electrophoresis and multiprimer random amplified polymorphic DNA fingerprinting. *Mol. Biochem. Parasitol.* **92**, 253–263 (1998).
73. Brisse, S. et al. Evidence for genetic exchange and hybridization in *Trypanosoma cruzi* based on nucleotide sequences and molecular karyotype. *Infect. Genet. Evol.* **2**, 173–183 (2003).
74. Sturm, N. R., Vargas, N. S., Westenberger, S. J., Zingales, B. & Campbell, D. A. Evidence for multiple hybrid groups in *Trypanosoma cruzi*. *Int. J. Parasitol.* **33**, 269–279 (2003).
75. Sturm, N. R. & Campbell, D. A. Alternative lifestyles: the population structure of *Trypanosoma cruzi*. *Acta Trop.* **115**, 35–43 (2010).
76. Tomasini, N. & Diosque, P. Evolution of *Trypanosoma cruzi*: clarifying hybridisations, mitochondrial introgressions and phylogenetic relationships between major lineages. *Mem. Inst. Oswaldo Cruz* **110**, 403–413 (2015).
77. El-Sayed, N. M. et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* **309**, 409–415 (2005).
78. O'Dea, A. et al. Formation of the Isthmus of Panama. *Sci. Adv.* **2**, e1600883 (2016).
79. Tomazi, L., Kawashita, S. Y., Pereira, P. M., Zingales, B. & Briones, M. R. S. Haplotype distribution of five nuclear genes based on network genealogies and Bayesian inference indicates that *Trypanosoma cruzi* hybrid strains are polyphyletic. *Genet. Mol. Res.* **8**, 458–476 (2009).
80. Miles, M. A. et al. Isozymic heterogeneity of *Trypanosoma cruzi* in the first autochthonous patients with Chagas' disease in Amazonian Brazil. *Nature* **272**, 819–821 (1978).

81. Mendonça, M. B. A. et al. Two main clusters within *Trypanosoma cruzi* zymodeme 3 are defined by distinct regions of the ribosomal RNA cistron. *Parasitology* **124**, 177–184 (2002).
82. Coura, J. R., Junqueira, A. C. V., Fernandes, O., Valente, S. A. S. & Miles, M. A. Emerging Chagas disease in Amazonian Brazil. *Trends Parasitol.* **18**, 171–176 (2002).
83. Tomasini, N. & Diosque, P. Phylogenomics of *Trypanosoma cruzi*: few evidence of TcI/TcII mosaicism in TcIII challenges the hypothesis of an ancient TcI/TcII hybridization. *Infect. Genet. Evol.* **50**, 25–27 (2017).
84. Tomasini, N. Introgression of the kinetoplast DNA: an unusual evolutionary journey in *Trypanosoma cruzi*. *Curr. Genomics* **19**, 133–139 (2018).
85. Ramírez, J. C., Torres, C., Curto, M. de los A. & Schijman, A. G. New insights into *Trypanosoma cruzi* evolution, genotyping and molecular diagnostics from satellite DNA sequence analysis. *PLoS Negl. Trop. Dis.* **11**, e0006139 (2017).
86. Murphy, W. J. et al. Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614–618 (2001).
87. Townsend, J. P. & Lopez-Giraldez, F. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. *Syst. Biol.* **59**, 446–457 (2010).
88. Nabhan, A. R. & Sarkar, I. N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief. Bioinform.* **13**, 122–134 (2012).
89. Barnabé, C., Mobarec, H. I., Jurado, M. R., Cortez, J. A. & Brenière, S. F. Reconsideration of the seven discrete typing units within the species *Trypanosoma cruzi*, a new proposal of three reliable mitochondrial clades. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **39**, 176–186 (2016).
90. Legendre, P. & Makarenkov, V. Reconstruction of biogeographic and evolutionary networks using reticulograms. *Syst. Biol.* **51**, 199–216 (2002).
91. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
92. Pollard, D. A., Iyer V. N., Moses, A. M. & Eisen, M. B. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* **2**, e173 (2006).
93. Wu, G. A. et al. Genomics of the origin and evolution of *Citrus*. *Nature* **554**, 311–316 (2018).
94. Vago, A. R. et al. Genetic characterization of *Trypanosoma cruzi* directly from tissues of patients with chronic Chagas disease: differential distribution of genetic types into diverse organs. *Am. J. Pathol.* **156**, 1805–1809 (2000).

95. Burgos, J. M. et al. Molecular identification of *Trypanosoma cruzi* I tropism for central nervous system in Chagas reactivation due to AIDS. *Am. J. Trop. Med. Hyg.* **78**, 294–297 (2008).
96. Burgos, J. M. et al. Molecular identification of *Trypanosoma cruzi* discrete typing units in end-stage chronic Chagas heart disease and reactivation after heart transplantation. *Clin. Infect. Dis.* **51**, 485–495 (2010).
97. Mantilla, J. C., Zafra, G. A., Macedo, A. M. & González, C. I. Mixed infection of *Trypanosoma cruzi* I and II in a Colombian cardiomyopathic patient. *Hum. Pathol.* **41**, 610–613 (2010).
98. Sánchez, L. V. et al. Temporal variation of *Trypanosoma cruzi* discrete typing units in asymptomatic Chagas disease patients. *Microbes Infect.* **15**, 745–748 (2013).
99. Miles, M. A., Apt B, W., Widmer, G., Povoa, M. M. & Schofield, C. J. Isozyme heterogeneity and numerical taxonomy of *Trypanosoma cruzi* stocks from Chile. *Trans. R. Soc. Trop. Med. Hyg.* **78**, 526–535 (1984).
100. Luquetti, A. O. et al. *Trypanosoma cruzi*: zymodemes associated with acute and chronic Chagas' disease in central Brazil. *Trans. R. Soc. Trop. Med. Hyg.* **80**, 462–470 (1986).
101. Ortiz, S., Zulantay, I., Apt, W., Saavedra, M. & Solari, A. Transferability of *Trypanosoma cruzi* from mixed human host infection to *Triatoma infestans* and from insects to axenic culture. *Parasitol. Int.* **64**, 33–36 (2015).
102. 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).
103. Dean, M. P., Jansen, A. M., Mangia, R. H. R., Gonçalves, A. M. & Morel C. M. Are our laboratory 'strains' representative samples of *Trypanosoma cruzi* populations that circulate in nature? *Mem. Inst. Oswaldo Cruz* **79**, 19–24 (1984).
104. Macedo, A. M. & Segatto, M. Implications of *Trypanosoma cruzi* intraspecific diversity in the pathogenesis of Chagas disease; in American trypanosomiasis, Chagas disease: one hundred years of research (eds. Telleria, J. & Tibayrenc, M.) 489–522. *Elsevier* (2010).
105. Freitas, J. M., Lages-Silva, E., Crema, E., Pena, S. D. J. & Macedo, A. M. Real time PCR strategy for the identification of major lineages of *Trypanosoma cruzi* directly in chronically infected human tissues. *Int. J. Parasitol.* **35**, 411–417 (2005).
106. Lages-Silva, E. et al. Variability of kinetoplast DNA gene signatures of *Trypanosoma cruzi* II strains from patients with different clinical forms of Chagas' disease in Brazil. *J. Clin. Microbiol.* **44**, 2167–2171 (2006).

107. Diosque, P. et al. Multilocus enzyme electrophoresis analysis of *Trypanosoma cruzi* isolates from a geographically restricted endemic area for Chagas' disease in Argentina. *Int. J. Parasitol.* **33**, 997–1003 (2003).
108. Virreira, M., Serrano, G., Maldonado, L. & Svoboda, M. *Trypanosoma cruzi*: typing of genotype (sub)lineages in megacolon samples from Bolivian patients. *Acta Trop.* **100**, 252–255 (2006).
109. del Puerto, R. et al. Lineage analysis of circulating *Trypanosoma cruzi* parasites and their association with clinical forms of Chagas disease in Bolivia. *PLoS Negl. Trop. Dis.* **4**, e867 (2010).
110. Cura, C. I. et al. *Trypanosoma cruzi* discrete typing units in Chagas disease patients from endemic and non-endemic regions of Argentina. *Parasitology* **139**, 516–521 (2012).
111. Virreira, M. et al. Congenital Chagas disease in Bolivia is not associated with DNA polymorphism of *Trypanosoma cruzi*. *Am J. Trop. Med. Hyg.* **75**, 871–879 (2006).
112. Virreira, M. et al. Comparison of *Trypanosoma cruzi* lineages and levels of parasitic DNA in infected mothers and their newborns. *Am. J. Trop. Med. Hyg.* **77**, 102–106 (2007).
113. Corrales, R. M. et al. Congenital Chagas disease involves *Trypanosoma cruzi* sub-lineage IId in the northwestern province of Salta, Argentina. *Infect. Genet. Evol.* **9**, 278–282 (2009).
114. Miles, M. A. et al. Do radically dissimilar *Trypanosoma cruzi* strains (zymodemes) cause Venezuelan and Brazilian forms of Chagas' disease? *Lancet* **1**, 1338–1340 (1981).
115. Añez, N. et al. Predominance of lineage I among *Trypanosoma cruzi* isolates from Venezuelan patients with different clinical profiles of acute Chagas' disease. *Trop. Med. Int. Health* **9**, 1319–1326 (2004).
116. Ramírez, J. D. et al. Chagas cardiomyopathy manifestations and *Trypanosoma cruzi* genotypes circulating in chronic chagasic patients. *PLoS Negl. Trop. Dis.* **4**, e899 (2010).
117. Guhl, F. Epidemiología molecular de *Trypanosoma cruzi*. *Rev. Esp. Salud Pública* **87**, 1–8 (2013).
118. Guhl, F. & Ramírez, J. D. Retrospective molecular integrated epidemiology of Chagas disease in Colombia. *Infect. Genet. Evol.* **10**, 148–154 (2013).
119. Ramírez, J. D. et al. Molecular epidemiology of human oral Chagas disease outbreaks in Colombia. *PLoS Negl. Trop. Dis.* **7**, e2041 (2013).

120. Dario, M. A. et al. Ecological scenario and *Trypanosoma cruzi* DTU characterization of a fatal acute Chagas disease case transmitted orally (Espírito Santo state, Brazil). *Parasit. Vectors* **9**, 477 (2016).
121. Marcili, A. et al. *Trypanosoma cruzi* in Brazilian Amazonia: lineages TCI and TCIIa in wild primates, *Rhodnius* spp. and in humans with Chagas disease associated with oral transmission. *Int. J. Parasitol.* **39**, 615–623 (2009).
122. Alarcón de Noya, B. et al. Orally-transmitted Chagas disease: epidemiological, clinical, serological and molecular outcomes of a school microepidemic in Chichiriviche de la Costa, Venezuela. *Parasite Epidemiol. Control* **1**, 188–198 (2016).
123. Jiménez, P., Jaimes, J., Poveda, C. & Ramírez, J. D. A systematic review of the *Trypanosoma cruzi* genetic heterogeneity, host immune response and genetic factors as plausible drivers of chronic chagasic cardiomyopathy. *Parasitology* **146**, 269–283 (2019).
124. Burgos, J. M. et al. Direct molecular profiling of minicircle signatures and lineages of *Trypanosoma cruzi* bloodstream populations causing congenital Chagas disease. *Int. J. Parasitol.* **37**, 1319–1327 (2007).
125. Guhl, F. Geographical distribution of Chagas disease; in American trypanosomiasis, Chagas disease: one hundred years of research (2nd edition) (eds. Telleria, J. & Tibayrenc, M.) 89–112. *Elsevier* (2017).
126. Roellig, D. M., Ellis, A. E. & Yabsley, M. J. Genetically different isolates of *Trypanosoma cruzi* elicit different infection dynamics in raccoons (*Procyon lotor*) and Virginia opossums (*Didelphis virginiana*). *Int. J. Parasitol.* **39**, 1603–1610 (2009).
127. Morocoima, A. et al. *Trypanosoma cruzi* III from armadillos (*Dasypus novemcinctus novemcinctus*) from northeastern Venezuela and its biological behavior in murine model. Risk of emergency of Chagas' disease. *Exp. Parasitol.* **132**, 341–347 (2012).
128. Orozco, M. M. et al. New sylvatic hosts of *Trypanosoma cruzi* and their reservoir competence in the humid Chaco of Argentina: a longitudinal study. *Am. J. Trop. Med. Hyg.* **88**, 872–882 (2013).
129. Llewellyn, M. S. et al. *Trypanosoma cruzi* IIc: phylogenetic and phylogeographic insights from sequence and microsatellite analysis and potential impact on emergent Chagas disease. *PLoS Negl. Trop. Dis.* **3**, e510 (2009).
130. Jansen, A. M., Xavier, S. C. C. & Roque, A. L. R. The multiple and complex and changeable scenarios of the *Trypanosoma cruzi* transmission cycle in the sylvatic environment. *Acta Trop.* **151**, 1–15 (2015).

131. Curtis-Robles, R. et al. Epidemiology and molecular typing of *Trypanosoma cruzi* in naturally-infected hound dogs and associated triatomine vectors in Texas, USA. *PLoS Negl. Trop. Dis.* **11**, e0005298 (2017).
132. Vandermark, C. et al. *Trypanosoma cruzi* strain TcIV infects raccoons from Illinois. *Mem. Inst. Oswaldo Cruz* **113**, 30–37 (2018).
133. Widmer, G., Marinkelle, C. J., Guhl, F. & Miles, M. A. Isozyme profiles of *Trypanosoma cruzi* stocks from Colombia and Ecuador. *Ann. Trop. Med. Parasitol.* **79**, 253–257 (1985).
134. Saravia, N. G., Holguín, A. F., Cibulskis, R. E. & D'Alessandro, A. Divergent isoenzyme profiles of sylvatic and domiciliary *Trypanosoma cruzi* in the eastern plains, piedmont, and highlands of Colombia. *Am. J. Trop. Med. Hyg.* **36**, 59–69 (1987).
135. Herrera, C. et al. Identifying four *Trypanosoma cruzi* I isolate haplotypes from different geographic regions in Colombia. *Infect. Genet. Evol.* **7**, 535–539 (2006).
136. Falla, A. et al. Haplotype identification within *Trypanosoma cruzi* I in Colombian isolates from several reservoirs, vectors and humans. *Acta Trop.* **110**, 15–21 (2009).
137. Cura, C. I. et al. *Trypanosoma cruzi* I genotypes in different geographic regions and transmission cycles based on a microsatellite motif of the intergenic spacer of spliced leader genes. *Int. J. Parasitol.* **40**, 1599–1607 (2010).
138. Herrera, C. et al. Genetic variability and phylogenetic relationships within *Trypanosoma cruzi* I isolated in Colombia based on miniexon gene sequences. *J. Parasitol. Res.* **2009**, e897364 (2010).
139. Zumaya-Estrada, F. A. et al. North American import? Charting the origins of an enigmatic *Trypanosoma cruzi* domestic genotype. *Parasit. Vectors* **5**, 226 (2012).
140. Ocaña-Mayorga, S., Llewellyn, M. S., Costales, J. A., Miles, M. A. & Grijalva, M. J. Sex, subdivision, and domestic dispersal of *Trypanosoma cruzi* lineage I in southern Ecuador. *PLoS Negl. Trop. Dis.* **4**, e915 (2010).
141. Costales, J. A. et al. *Trypanosoma cruzi* population dynamics in the central Ecuadorian coast. *Acta Trop.* **151**, 88–93 (2015).
142. Messenger, L. A. et al. Multiple mitochondrial introgression events and heteroplasmy in *Trypanosoma cruzi* revealed by maxicircle MLST and next generation sequencing. *PLoS Negl. Trop. Dis.* **6**, e1584 (2012).
143. Nouvellet, P., Dumonteil, E. & Gourbière, S. The improbable transmission of *Trypanosoma cruzi* to human: the missing link in the dynamics and control of Chagas disease. *PLoS Negl. Trop. Dis.* **7**, e2505 (2013).

144. Alkmim-Oliveira, S. M. et al. *Trypanosoma cruzi* experimental congenital transmission associated with TcV and TcI subpatent maternal parasitemia. *Parasitol. Res.* **112**, 671–678 (2013).
145. Cruz, L. et al. Comparative study of the biological properties of *Trypanosoma cruzi* I genotypes in a murine experimental model. *Infect. Genet. Evol.* **29**, 110–117 (2015).
146. Hernández, C. et al. Molecular diagnosis of Chagas disease in Colombia: parasitic loads and discrete typing units in patients from acute and chronic phases. *PLoS Negl. Trop. Dis.* **10**, e0005112 (2016).
147. Barnabé, C. et al. Putative panmixia in restricted populations of *Trypanosoma cruzi* isolated from wild *Triatoma infestans* in Bolivia. *PloS One* **8**, e82269 (2013).
148. Lima, V. S., Jansen, A. M., Messenger, L. A., Miles, M. A. & Llewellyn, M. S. Wild *Trypanosoma cruzi* I genetic diversity in Brazil suggests admixture and disturbance in parasite populations from the Atlantic Forest region. *Parasit. Vectors* **7**, 263 (2014).
149. Messenger, L. A. et al. Ecological host fitting of *Trypanosoma cruzi* TcI in Bolivia: mosaic population structure, hybridization and a role for humans in Andean parasite dispersal. *Mol. Ecol.* **24**, 2406–2422 (2015).
150. Roman Maldonado, F., Iniguez, A., Yeo, M. & Jansen, A. Multilocus sequence typing: genetic diversity in *Trypanosoma cruzi* I (TcI) isolates from Brazilian didelphids. *Parasit. Vectors* **11**, 107 (2018).
151. Roman, F. et al. Dissecting the phyloepidemiology of *Trypanosoma cruzi* I (TcI) in Brazil by the use of high resolution genetic markers. *PLoS Negl. Trop. Dis.* **12**, e0006466 (2018).
152. Hamilton, P. B., Gibson, W. C. & Stevens, J. R. Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. *Mol. Phylogenet. Evol.* **44**, 15–25 (2007).
153. Lukeš, J., Skalický, T., Týč, J., Votýpka, J. & Yurchenko, V. Evolution of parasitism in kinetoplastid flagellates. *Mol. Biochem. Parasitol.* **195**, 115–122 (2014).
154. Volf, P. et al. Increased transmission potential of *Leishmania major*/*Leishmania infantum* hybrids. *Int. J. Parasitol.* **37**, 589–593 (2007).
155. Cortes, S. et al. *In vitro* and *in vivo* behaviour of sympatric *Leishmania* (*V.*) *braziliensis*, *L.* (*V.*) *peruviana* and their hybrids. *Parasitology* **139**, 191–199 (2012).
156. Balmer, O., Beadell, J. S., Gibson, W. & Caccone, A. Phylogeography and taxonomy of *Trypanosoma brucei*. *PLoS Negl. Trop. Dis.* **5**, e961 (2011).
157. Ferreira, M. U., Nunes, M. da S. & Wunderlich, G. Antigenic diversity and immune evasion by malaria parasites. *Clin. Diagn. Lab. Immunol.* **11**, 987–995 (2004).

158. Heitman, J. Sexual reproduction and the evolution of microbial pathogens. *Curr. Biol.* **16**, R711-725 (2006).
159. Arenas, M. et al. Mutation and recombination in pathogen evolution: relevance, methods and controversies. *Infect. Genet. Evol.* **63**, 295–306 (2018).
160. Tibayrenc, M. & Ayala, F. J. Towards a population genetics of microorganisms: the clonal theory of parasitic protozoa. *Parasitol. Today* **7**, 228–232 (1991).
161. Tibayrenc, M., Cariou, M. L., Solignac, M. & Carlier, Y. Arguments génétiques contre l'existence d'une sexualité actuelle chez *Trypanosoma cruzi*: implications taxinomiques. C. R. Acad. Sci. 293 207–209 (1981).
162. Tibayrenc, M., Kjellberg, F. & Ayala, F. J. A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. *Proc. Natl. Acad. Sci. USA* **87**, 2414–2418 (1990).
163. Oliveira, R. P. et al. Probing the genetic population structure of *Trypanosoma cruzi* with polymorphic microsatellites. *Proc. Natl. Acad. Sci. USA* **95**, 3776–3780 (1998).
164. Lauthier, J. J. et al. Candidate targets for multilocus sequence typing of *Trypanosoma cruzi*: validation using parasite stocks from the Chaco Region and a set of reference strains. *Infect. Genet. Evol.* **12**, 350–358 (2012).
165. Tibayrenc, M. & Ayala, F. J. The population genetics of *Trypanosoma cruzi* revisited in the light of the predominant clonal evolution model. *Acta Trop.* **151**, 156–165 (2015).
166. Tibayrenc, M. & Ayala, F. J. How clonal are *Trypanosoma* and *Leishmania*? *Trends Parasitol.* **29**, 264–269 (2013).
167. Ramírez, J. D., Duque, M. C., Montilla, M., Cucunubá, Z. & Guhl, F. Natural and emergent *Trypanosoma cruzi* I genotypes revealed by mitochondrial (Cytb) and nuclear (SSU rDNA) genetic markers. *Exp. Parasitol.* **132**, 487–494 (2012).
168. Baptista, R. de P. et al. Evidence of substantial recombination among *Trypanosoma cruzi* II strains from Minas Gerais. *Infect. Genet. Evol.* **22**, 183–191 (2014).
169. Wahlund, S. Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas* **11**, 65–106 (2014).
170. Messenger, L. A. & Miles, M. A. Evidence and importance of genetic exchange among field populations of *Trypanosoma cruzi*. *Acta Trop.* **151**, 150–155 (2015).
171. Gaunt, M. W. et al. Mechanism of genetic exchange in American trypanosomes. *Nature* **421**, 936–939 (2003).

172. Carrasco, H. J., Frame, I. A., Valente, S. A. & Miles, M. A. Genetic exchange as a possible source of genomic diversity in sylvatic populations of *Trypanosoma cruzi*. *Am. J. Trop. Med. Hyg.* **54**, 418–424 (1996).
173. Bennett, R. J. The parasexual lifestyle of *Candida albicans*. *Curr. Opin. Microbiol.* **28**, 10–17 (2015).
174. Lewis, M. D. et al. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in *Trypanosoma cruzi* populations and expose contrasts between natural and experimental hybrids. *Int. J. Parasitol.* **39**, 1305–1317 (2009).
175. Ramírez, J. D., Tapia-Calle, G. & Guhl, F. Genetic structure of *Trypanosoma cruzi* in Colombia revealed by a high-throughput nuclear multilocus sequence typing (nMLST) approach. *BMC Genet.* **14**, 96 (2013).
176. Berry, A. S. F. et al. Sexual reproduction in a natural *Trypanosoma cruzi* population. *PLoS Negl. Trop. Dis.* **13**, e0007392 (2019).
177. Lima, V. D. S. et al. Expanding the knowledge of the geographic distribution of *Trypanosoma cruzi* TcII and TcV/TcVI genotypes in the Brazilian Amazon. *PloS One* **9**, e116137 (2014).
178. Reis-Cunha, J. L. et al. Whole genome sequencing of *Trypanosoma cruzi* field isolates reveals extensive genomic variability and complex aneuploidy patterns within TcII DTU. *BMC Genomics* **19**, 816 (2018).
179. Chen, G., Bradford, W. D., Seidel, C. W. & Li, R. Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. *Nature* **482**, 246–250 (2012).
180. Mannaert, A., Downing, T., Imamura, H. & Dujardin, J.-C. Adaptive mechanisms in pathogens: universal aneuploidy in *Leishmania*. *Trends Parasitol.* **28**, 370–376 (2012).
181. Tihon, E., Imamura, H., Dujardin, J.-C., Abbeele, J. V. D. & Van den Broeck, F. Discovery and genomic analyses of hybridization between divergent lineages of *Trypanosoma congolense*, causative agent of animal African trypanosomiasis. *Mol. Ecol.* **26**, 6524–6538 (2017).
182. Tibayrenc, M. & Ayala, F. Hybridization in *Trypanosoma congolense* does not challenge the predominant clonal evolution model. A comment on Tihon et al., 2017, Mol. Ecol. *Mol. Ecol.* **27**, 3421–3424 (2018).
183. Van den Broeck, F., Tavernier, L. J. M., Vermeiren, L., Dujardin, J.-C. & Van den Abbeele, J. Mitonuclear genomics challenges the theory of clonality in *Trypanosoma congolense*. Reply to Tibayrenc and Ayala. *Mol. Ecol.* **27**, 3425–3431 (2018).
184. Tibayrenc, M. & Ayala, F. J. How clonal is *Trypanosoma congolense*? A necessary clarification of the predominant clonal evolution model. *Acta Trop.* **190**, 28–29 (2019).

185. Chappuis, F. et al. Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? *Nat. Rev. Microbiol.* **5**, S7–S16 (2007).
186. Lenk, E. J. et al. Socioeconomic benefit to individuals of achieving 2020 targets for four neglected tropical diseases controlled/eliminated by innovative and intensified disease management: human African trypanosomiasis, leprosy, visceral leishmaniasis, Chagas disease. *PLoS Negl. Trop. Dis.* **12**, e0006250 (2018).
187. Azevedo, T. S. de et al. Risk mapping of visceral leishmaniasis in Brazil. *Rev. Soc. Bras. Med. Trop.* **52**, e20190240 (2019).
188. Okwor, I. & Uzonna, J. Social and economic burden of human leishmaniasis. *Am. J. Trop. Med. Hyg.* **94**, 489–493 (2016).
189. Rijal, S., Koirala, S., Van der Stuyft, P. & Boelaert, M. The economic burden of visceral leishmaniasis for households in Nepal. *Trans. R. Soc. Trop. Med. Hyg.* **100**, 838–841 (2006).
190. Ponte-Sucre, A. et al. Drug resistance and treatment failure in leishmaniasis: a 21st century challenge. *PLoS Negl. Trop. Dis.* **11**, e0006052 (2017).
191. McGwire, B. S. & Satoskar, A. R. Leishmaniasis: clinical syndromes and treatment. *QJM* **107**, 7–14 (2014).
192. Travi, B. L., Cordeiro-da-Silva, A., Dantas-Torres, F. & Miró G. Canine visceral leishmaniasis: diagnosis and management of the reservoir living among us. *PLoS Negl. Trop. Dis.* **12**, e0006082 (2018)
193. Sevá, A. P. et al. Canine-based strategies for prevention and control of visceral leishmaniasis in Brazil. *PLoS One* **11**, e0160058 (2016).
194. Moreno, J. Assessment of vaccine-induced immunity against canine visceral leishmaniasis. *Front. Vet. Sci.* **6**, 168 (2019).
195. Sadlova, J. et al. *Leishmania donovani* development in *Phlebotomus argentipes*: comparison of promastigote- and amastigote-initiated infections. *Parasitology* **144**, 403–410 (2017).
196. Dawes, B. Advances in Parasitology. *Academic Press* (1977).
197. Quinnell, R. J. & Courtenay, O. Transmission, reservoir hosts and control of zoonotic visceral leishmaniasis. *Parasitology* **136**, 1915–1934 (2009).
198. Takala, S. L. & Plowe, C. V. Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. *Parasite Immunol.* **31**, 560–573 (2009).
199. Blake, D. P. et al. Population, genetic, and antigenic diversity of the apicomplexan *Eimeria tenella* and their relevance to vaccine development. *Proc. Natl. Acad. Sci. USA* **112**, E5343–E5350 (2015).

200. Imamura, H. et al. Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent. *eLife* **5**, e12613 (2016).
201. Costanzo, R. O. Leishmaniosis visceral: a 101 años del primer caso diagnosticado en las Américas. *Mem. Inst. Investig. Cienc. Salud* **10**, 100–104 (2012).
202. Penna, H. A. Leishmaniose visceral no Brasil. *Bras. Med.* **18**, 940–950 (1934).
203. Conti, R., Lane, V., Montebello, L. & Pinto Junior, V. Visceral leishmaniasis epidemiologic evolution in timeframes, based on demographic changes and scientific achievements in Brazil. *J. Vectorborne Dis.* **53**, 99–104 (2016).
204. Deane, L. M. & Deane, M. P. Observações sobre a transmissão da leishmaniose visceral no Ceará. *O. Hospital.* **48**, 347–364 (1955).
205. Costa, C. H., Pereira, H. F. & Araújo, M. V. Visceral leishmaniasis epidemic in the state of Piauí, Brazil, 1980-1986. *Rev. Saude Publica* **24**, 361–372 (1990).
206. da Silva, E. R. Leishmaniose visceral (calazar) na Ilha de São Luís, Maranhão, Brasil: evolução e perspectivas. *Rev. Soc. Bras. Med. Trop.* **30**, 359–368 (1997).
207. Jeronimo, S. et al. An urban outbreak of visceral leishmaniasis in Natal, Brazil. *Trans. R. Soc. Trop. Med. Hyg.* **88**, 386–388 (1994).
208. Marzochi, M. C. A., Marzochi, K. B. F. & Carvalho, R. W. Visceral leishmaniasis in Rio de Janeiro. *Parasitol. Today* **10**, 37–40 (1994).
209. Bevilacqua, P. D., Paixão, H. H., Modena, C. M. & Castro, M. C. P. S. Urbanização da leishmaniose visceral em Belo Horizonte. *Arq. Bras. Med. Veterinária E Zootec.* **53**, 1–8 (2001).
210. Furlan, M. B. G. Epidemia de leishmaniose visceral no município de Campo Grande-MS, 2002 a 2006. *Epidemiol. Serviços Saúde* **19**, 16–25 (2010).
211. Gontijo, C. M. F. & Melo, M. N. Leishmaniose visceral no Brasil: quadro atual, desafios e perspectivas. *Rev. Bras. Epidemiol.* **7**, 338–349 (2004).
212. Romero, G. A. S. & Boelaert, M. Control of visceral leishmaniasis in Latin America—a systematic review. *PLoS Negl. Trop. Dis.* **4**, e584 (2010).
213. Monteiro, D. C. S. et al. *Leishmania infantum* infection in blood donors, northeastern Brazil. *Emerg. Infect. Dis.* **22**, 739–740 (2016).
214. Organização Pan-Americana da Saúde. Leishmanioses. informe epidemiológico das Américas, dezembro 2019. *Inf. Leishmanioses 7, 8* (2019).
215. Sellon, R. K. et al. Endemic visceral leishmaniasis in a dog from Texas. *J. Vet. Intern. Med.* **7**, 16–19 (1993).
216. Pigott, D. M. et al. Global distribution maps of the leishmaniases. *eLife* **3**, e02851 (2014).

217. Satragno, D. et al. Autochthonous outbreak and expansion of canine visceral leishmaniasis, Uruguay. *Emerg. Infect. Dis.* **23**, 536–538 (2017).
218. Lainson, R. & Rangel, E. F. *Lutzomyia longipalpis* and the eco-epidemiology of American visceral leishmaniasis, with particular reference to Brazil: a review. *Mem. Inst. Oswaldo Cruz* **100**, 811–827 (2005).
219. Travi, B. L. et al. *Lutzomyia evansi*, an alternate vector of *Leishmania chagasi* in a Colombian focus of visceral leishmaniasis. *Trans. R. Soc. Trop. Med. Hyg.* **84**, 676–677 (1990).
220. Travi, B. L. et al. Bionomics of *Lutzomyia evansi* (Diptera: Psychodidae) vector of visceral leishmaniasis in northern Columbia. *J. Med. Entomol.* **33**, 278–285 (1996).
221. Aguilar, C. M. et al. Urban visceral leishmaniasis in Venezuela. *Mem. Inst. Oswaldo Cruz* **93**, 15–16 (1998).
222. Feliciangeli, M. D., Rodriguez, N., De Guglielmo, Z. & Rodriguez, A. The re-emergence of American visceral leishmaniasis in an old focus in Venezuela. *Parasite* **6**, 113–120 (1999).
223. Brito, V. N. de et al. Phlebotomine fauna, natural infection rate and feeding habits of *Lutzomyia cruzi* in Jaciara, state of Mato Grosso, Brazil. *Mem. Inst. Oswaldo Cruz* **109**, 899–904 (2014).
224. de Oliveira, E. F. et al. Vector competence of *Lutzomyia cruzi* naturally demonstrated for *Leishmania infantum* and suspected for *Leishmania amazonensis*. *Am. J. Trop. Med. Hyg.* **96**, 178–181 (2017).
225. Coelho, M. V., Cunha, A. S. & Falcão, A. R. Notas sobre um foco de calazar no sudoeste do estado de Goiás. *Rev. Bras. Malariol. Doenças. Trop.* **17**, 143–148 (1965).
226. Oliveira, A. C., Batista, S. M. & Falcão, A. L. Calazar em Minas Gerais. Revisão dos dados epidemiológicos obtidos até 1958. *Hospital RJ* **56**, 625–643 (1959).
227. Momen, H., Grimaldi Júnior, G. & Deane, L. M. *Leishmania infantum*, the aetiological agent of American visceral leishmaniasis (AVL)? *Mem. Inst. Oswaldo Cruz* **82**, 447–448 (1987).
228. Maurício, I. L., Howard, M. K., Stothard, J. R. & Miles, M. A. Genomic diversity in the *Leishmania donovani* complex. Parasitology **119**, 237–246 (1999).
229. Rioux, J. A. et al. Taxonomy of *Leishmania*. Use of isoenzymes. Suggestions for a new classification. *Ann. Parasitol. Hum. Comp.* **65**, 111–125 (1990).
230. Cupolillo, E., Grimaldi, G. J. & Momen, H. A general classification of New World *Leishmania* using numerical zymotaxonomy. *Am. J. Trop. Med. Hyg.* **50**, 296–311 (1994).

231. Lainson, R., Shaw, J. J., Silveira, F. T. & Braga, R. R. American visceral leishmaniasis: on the origin of *Leishmania* (*Leishmania*) *chagasi*. *Trans. R. Soc. Trop. Med. Hyg.* **81**, 517 (1987).
232. Shaw, J. J. Taxonomy of the genus *Leishmania*: present and future trends and their implications. *Mem. Inst. Oswaldo Cruz* **89**, 471–478 (1994).
233. Sherlock, Í. A. et al. Natural infection of the opossum *Didelphis albiventris* (Marsupialia, Didelphidae) with *Leishmania donovani*, in Brazil. *Mem. Inst. Oswaldo Cruz* **79**, 511–511 (1984).
234. Travi, B. L., Osorio, Y., Becerra, M. T. & Adler, G. H. Dynamics of *Leishmania chagasi* infection in small mammals of the undisturbed and degraded tropical dry forests of northern Colombia. *Trans. R. Soc. Trop. Med. Hyg.* **92**, 275–278 (1998).
235. Lainson, L. & Shaw, J. J. New World leishmaniasis: the neotropical *Leishmania* species; in Topley and Wilson's microbiology and microbial infections (9th edition) (eds. Cox, F. E. G., Kreier, J. P. & Wakelin, D) 242–266. *Hodder Arnold* (1998).
236. Combes, C. Fitness of parasites: pathology and selection. *Int. J. Parasitol.* **27**, 1–10 (1997).
237. Lipsitch, M. & Moxon, E. R. Virulence and transmissibility of pathogens: what is the relationship? *Trends Microbiol.* **5**, 31–37 (1997).
238. Courtenay, O., Macdonald, D. W., Lainson, R., Shaw, J. J. & Dye, C. Epidemiology of canine leishmaniasis: a comparative serological study of dogs and foxes in Amazon Brazil. *Parasitology* **109 (Pt 3)**, 273–279 (1994).
239. Courtenay, O., Quinnell, R. J., Garcez, L. M., Shaw, J. J. & Dye, C. Infectiousness in a cohort of Brazilian dogs: why culling fails to control visceral leishmaniasis in areas of high transmission. *J. Infect. Dis.* **186**, 1314–1320 (2002).
240. Luppi, M. M. et al. Visceral leishmaniasis in captive wild canids in Brazil. *Vet. Parasitol.* **155**, 146–151 (2008).
241. Courtenay, O., Quinnell, R. J., Garcez, L. M. & Dye, C. Low infectiousness of a wildlife host of *Leishmania infantum*: the crab-eating fox is not important for transmission. *Parasitology* **125**, 407–414 (2002).
242. Lainson, R. Evolution, classification and geographical distribution. *Academic Press* (1987).
243. Excoffier, L., Foll, M. & Petit, R. J. Genetic consequences of range expansions. *Annu. Rev. Ecol. Evol. Syst.* **40**, 481–501 (2009).
244. Akopyants, N. S. et al. Demonstration of genetic exchange during cyclical development of *Leishmania* in the sand fly vector. *Science* **324**, 265–268 (2009).

245. Inbar, E. et al. The mating competence of geographically diverse *Leishmania major* strains in their natural and unnatural sand fly vectors. *PLoS Genet.* **9**, e1003672 (2013).
246. Cotton, J. A. et al. Genomic analysis of natural intra-specific hybrids among Ethiopian isolates of *Leishmania donovani*. *bioRxiv* 516211 (2019), doi:10.1101/516211.
247. Canestrelli, D. et al. The rangled evolutionary legacies of range expansion and hybridization. *Trends Ecol. Evol.* **31**, 677–688 (2016).
248. Ferreira, G. E. M. et al. The genetic structure of *Leishmania infantum* populations in Brazil and its possible association with the transmission cycle of visceral leishmaniasis. *PLoS One* **7**, e36242 (2012).
249. Campos-Ponce, M., Ponce, C., Ponce, E. & Maingon, R. D. C. *Leishmania chagasi*/*infantum*: further investigations on *Leishmania* tropisms in atypical cutaneous and visceral leishmaniasis foci in Central America. *Exp. Parasitol.* **109**, 209–219 (2005).
250. Ponce, C. et al. *Leishmania donovani chagasi*: new clinical variant of cutaneous leishmaniasis in Honduras. *Lancet* **337**, 67–70 (1991).
251. Noyes, H., Chance, M., Ponce, C., Ponce, E. & Maingon, R. *Leishmania chagasi*: genotypically similar parasites from Honduras cause both visceral and cutaneous leishmaniasis in humans. *Exp. Parasitol.* **85**, 264–273 (1997).
252. Warburg, A., Saraiva, E., Lanzaro, G. C., Titus, R. G. & Neva, F. Saliva of *Lutzomyia longipalpis* sibling species differs in its composition and capacity to enhance leishmaniasis. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **345**, 223–230 (1994).
253. Davies, C. R. & Mazloumi Gavgani, A. S. Age, acquired immunity and the risk of visceral leishmaniasis: a prospective study in Iran. *Parasitology* **119**, 247–257 (1999).
254. Maciel, B. L. L. et al. Association of nutritional status with the response to infection with *Leishmania chagasi*. *Am. J. Trop. Med. Hyg.* **79**, 591–598 (2008).
255. Lindoso, J. A. et al. Visceral leishmaniasis and HIV coinfection in Latin America. *PLoS Negl. Trop. Dis.* **8**, e3136 (2014).
256. Alam, M. Z. et al. Multilocus microsatellite typing (MLMT) reveals genetic homogeneity of *Leishmania donovani* strains in the Indian subcontinent. *Infect. Genet. Evol.* **9**, 24–31 (2009).
257. 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).
258. Carnielli, J. B. T. et al. A *Leishmania infantum* genetic marker associated with miltefosine treatment failure for visceral leishmaniasis. *EBioMedicine* **36**, 83–91 (2018).

259. Carnielli, J. B. T. et al. The *Leishmania infantum* miltefosine sensitivity locus (poster). *BSP trypanosomiasis and leishmaniasis seminar in Granada* (2020).
260. Dumetz, F. et al. Modulation of aneuploidy in *Leishmania donovani* during adaptation to different *in vitro* and *in vivo* environments and its impact on gene expression. *mBio* **8**, e00599 (2017).
261. Bhattacharya, A. & Ouellette, M. New insights with miltefosine unresponsiveness in Brazilian *Leishmania infantum* isolates. *EBioMedicine* **37**, 13–14 (2018).
262. Callejas-Hernández, F., Rastrojo, A., Poveda, C., Gironès, N. & Fresno, M. Genomic assemblies of newly sequenced *Trypanosoma cruzi* strains reveal new genomic expansion and greater complexity. *Sci. Rep.* **8**, 1–13 (2018).
263. Berná, L. et al. Expanding an expanded genome: long-read sequencing of *Trypanosoma cruzi*. *Microb. Genomics* **4**, e000177 (2018).
264. Peacock, C. S. et al. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat. Genet.* **39**, 839–847 (2007).
265. Franzén, O. et al. Comparative genomic analysis of human infective *Trypanosoma cruzi* lineages with the bat-restricted subspecies *T. cruzi marinkellei*. *BMC Genomics* **13**, 531 (2012).
266. Clayton, C. & Shapira, M. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Mol. Biochem. Parasitol.* **156**, 93–101 (2007).
267. Clayton, C. Regulation of gene expression in trypanosomatids: living with polycistronic transcription. *Open Biol.* **9**, 190072.
268. Li, C.-H. et al. Roles of a *Trypanosoma brucei* 5′→3′ exoribonuclease homolog in mRNA degradation. *RNA* **12**, 2171–2186 (2006).
269. Schwede, A. et al. The role of deadenylation in the degradation of unstable mRNAs in trypanosomes. *Nucleic Acids Res.* **37**, 5511–5528 (2009).
270. Milone, J., Wilusz, J. & Bellofatto, V. Identification of mRNA decapping activities and an ARE-regulated 3' to 5' exonuclease activity in trypanosome extracts. *Nucleic Acids Res.* **30**, 4040–4050 (2002).
271. Ubeda, J.-M. et al. Modulation of gene expression in drug resistant *Leishmania* is associated with gene amplification, gene deletion and chromosome aneuploidy. *Genome Biol.* **9**, R115 (2008).
272. Leprohon, P. et al. Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. *Nucleic Acids Res.* **37**, 1387–1399 (2009).
273. Ubeda, J.-M. et al. Genome-wide stochastic adaptive DNA amplification at direct and inverted DNA repeats in the parasite *Leishmania*. *PLoS Biol.* **12**, e1001868 (2014).

274. Bussotti, G. et al. *Leishmania* genome dynamics during environmental adaptation reveal strain-specific differences in gene copy number variation, karyotype instability, and telomeric amplification. *mBio* **9**, e01399-18 (2018).
275. Drini, S. et al. Species- and strain-specific adaptation of the HSP70 super family in pathogenic trypanosomatids. *Genome Biol. Evol.* **8**, 1980–1995 (2016).
276. Zhang, W. W. et al. Genetic analysis of *Leishmania donovani* tropism using a naturally attenuated cutaneous strain. *PLoS Pathog.* **10**, e1004244 (2014).
277. Garvey, E. P. & Santi, D. V. Stable amplified DNA in drug-resistant *Leishmania* exists as extrachromosomal circles. *Science* **233**, 535–540 (1986).
278. Grondin, K., Papadopoulou, B. & Ouellette, M. Homologous recombination between direct repeat sequences yields P-glycoprotein containing amplicons in arsenite resistant *Leishmania. Nucleic Acids Res.* **21**, 1895–1901 (1993).
279. Sterkers, Y., Lachaud, L., Crobu, L., Bastien, P. & Pagès, M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*. *Cell. Microbiol.* **13**, 274–283 (2011).
280. McConville, M. J. & Naderer, T. Metabolic pathways required for the intracellular survival of *Leishmania. Annu. Rev. Microbiol.* **65**, 543–561 (2011).
281. Mukherjee, A. et al. Telomeric gene deletion and intrachromosomal amplification in antimony-resistant *Leishmania. Mol Microbiol.* **88**, 189–202 (2013).
282. Barja, P. P. et al. Haplotype selection as an adaptive mechanism in the protozoan pathogen *Leishmania donovani*. *Nat. Ecol. Evol.* **1**, 1961–1969 (2017).
283. Downing, T. et al. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res.* **21**, 2143–2156 (2011).
284. Vanaerschot, M. et al. Genetic markers for SSG resistance in *Leishmania donovani* and SSG treatment failure in visceral leishmaniasis patients of the Indian subcontinent. *J. Infect. Dis.* **206**, 752–755 (2012).
285. Monte-Neto, R. et al. Locus deletion and point mutation in the aquaglyceroporin AQP1 gene in antimony resistant *Leishmania* (*Viannia*) *guyanensis. PLoS Negl. Trop. Dis.* **9**, e0003476 (2015).
286. Coelho, A. C. et al. Multiple mutations in heterogeneous miltefosine-resistant *Leishmania major* population as determined by whole genome sequencing. *PLoS Negl. Trop. Dis.* **6**, e1512 (2012).
287. Laffitte, M.-C. N., Leprohon, P., Légaré, D. & Ouellette, M. Deep-sequencing revealing mutation dynamics in the miltefosine transporter gene in *Leishmania infantum* selected for miltefosine resistance. *Parasitol. Res.* **115**, 3699–3703 (2016).

288. Mondelaers, A. et al. Genomic and molecular characterization of miltefosine resistance in *Leishmania infantum* strains with either natural or acquired resistance through experimental selection of intracellular amastigotes. *PloS One* **11**, e0154101 (2016).
289. Pérez-Victoria, F. J., Gamarro, F., Ouellette, M. & Castanys, S. Functional cloning of the miltefosine transporter. A novel P-type phospholipid translocase from *Leishmania* involved in drug resistance. *J. Biol. Chem.* **278**, 49965–49971 (2003).
290. Pérez-Victoria, F. J., Sánchez-Cañete, M. P., Castanys, S. & Gamarro, F. Phospholipid translocation and miltefosine potency require both *L. donovani* miltefosine transporter and the new protein LdRos3 in *Leishmania* parasites. *J. Biol. Chem.* **281**, 23766–23775 (2006).
291. Bhandari, V. et al. Drug susceptibility in *Leishmania* isolates following miltefosine treatment in cases of visceral leishmaniasis and post kala-azar dermal leishmaniasis. *PLoS Negl. Trop. Dis.* **6**, e1657 (2012).
292. Berg, M., Mannaert, A., Vanaerschot, M., Auwera, G. V. D. & Dujardin, J.-C. (Post-) genomic approaches to tackle drug resistance in *Leishmania. Parasitology* **140**, 1492–1505 (2013).
293. Sadlova, J. et al. Visualisation of *Leishmania donovani* fluorescent hybrids during early stage development in the sand fly vector. *PloS One* **6**, e19851 (2011).
294. Romano, A. et al. Cross-species genetic exchange between visceral and cutaneous strains of *Leishmania* in the sand fly vector. *Proc. Natl. Acad. Sci. USA* **111**, 16808–16813 (2014).
295. Nolder, D., Roncal, N., Davies, C. R., Llanos-Cuentas, A. & Miles, M. A. Multiple hybrid genotypes of *Leishmania* (*Viannia*) in a focus of mucocutaneous leishmaniasis. *Am. J. Trop. Med. Hyg.* **76**, 573–578 (2007).
296. Pimenta, P. F. et al. Evidence that the vectorial competence of phlebotomine sand flies for different species of *Leishmania* is controlled by structural polymorphisms in the surface lipophosphoglycan. *Proc. Natl. Acad. Sci. USA* **91**, 9155–9159 (1994).
297. Schwenkenbecher, J. M. et al. Microsatellite analysis reveals genetic structure of *Leishmania tropica*. *Int. J. Parasitol.* **36**, 237–246 (2006).
298. Rogers, M. B. et al. Genomic confirmation of hybridisation and recent inbreeding in a vector-isolated *Leishmania* population. *PLoS Genet.* **10**, e1004092 (2014).
299. Gouzelou, E. et al. Multilocus microsatellite typing (MLMT) of strains from Turkey and Cyprus reveals a novel monophyletic *L. donovani* sensu lato group. *PLoS Negl. Trop. Dis.* **6**, e1507 (2012).

300. Gelanew, T. et al. Multilocus sequence and microsatellite identification of intra-specific hybrids and ancestor-like donors among natural Ethiopian isolates of *Leishmania donovani*. *Int. J. Parasitol.* **44**, 751–757 (2014).
301. Pfennig, K. S., Kelly, A. L. & Pierce, A. A. Hybridization as a facilitator of species range expansion. *Proc. R. Soc. B Biol. Sci.* **283**, 20161329 (2016).
302. Budd, A. F. & Pandolfi, J. M. Evolutionary novelty is concentrated at the edge of coral species distributions. *Science* **328**, 1558–1561 (2010).
303. Chunco, A. J. Hybridization in a warmer world. *Ecol. Evol.* **4**, 2019–2031 (2014).
304. Pool, J. E., Hellmann, I., Jensen, J. D. & Nielsen, R. Population genetic inference from genomic sequence variation. *Genome Res.* **20**, 291–300 (2010).
305. Weir, W. et al. Population genomics reveals the origin and asexual evolution of human infective trypanosomes. *eLife* **5**, e11473 (2016).
306. Talavera-Lopez, C. et al. Repeat-driven generation of antigenic diversity in a major human pathogen, *Trypanosoma cruzi*. *bioRxiv* 283531 (2018), doi:10.1101/283531.
307. Bragg, J. G., Supple, M. A., Andrew, R. L. & Borevitz, J. O. Genomic variation across landscapes: insights and applications. *New Phytol.* **207**, 953–967 (2015).
308. McManus, K. F. et al. Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. *PLoS Genet.* **13**, e1006560 (2017).
309. Goodhead, I. et al. Whole-genome sequencing of *Trypanosoma brucei* reveals introgression between subspecies that is associated with virulence. *mBio* **4**, e00197 (2013).
310. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
311. 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).
312. 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).
313. 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).
314. 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).

315. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
316. Pudlo, P. et al. Reliable ABC model choice via random forests. *Bioinformatics* **32**, 859–866 (2016).
317. Bradburd, G. S., Ralph, P. L. & Coop, G. M. Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution* **67**, 3258–3273 (2013).
318. Manzoni, C. et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.* **19**, 286–302 (2016).
319. Holderegger, R. & Wagner, H. H. A brief guide to landscape genetics. *Landsc. Ecol.* **21**, 793–796 (2006).
320. Storfer, A. et al. Putting the 'landscape' in landscape genetics. *Heredity* **98**, 128–142 (2006).
321. Wagner, H. H. & Fortin, M.-J. A conceptual framework for the spatial analysis of landscape genetic data. *Conserv. Genet.* **14**, 253–261 (2013).
322. Schoville, S. D. et al. Adaptive genetic variation on the landscape: methods and cases. *Ann. Rev. Ecol. Evol. Sys.* **43**, 23–43 (2012).
323. Legendre, P. & Legendre, L. F. J. Numerical ecology. *Elsevier* (2012).
324. Mims, M. C. et al. Simulating demography, genetics, and spatially explicit processes to inform reintroduction of a threatened char. *Ecosphere* **10**, e02589 (2019).
325. Landguth, E. L. & Cushman, S. A. CDPOP: a spatially explicit cost distance population genetics program. *Mol. Ecol. Resour.* **10**, 156–161 (2010).
326. Landguth, E. L., Bearlin, A., Day, C. C. & Dunham, J. CDMetaPOP: an individual-based, eco-evolutionary model for spatially explicit simulation of landscape demogenetics. *Methods Ecol. Evol.* **8**, 4–11 (2017).
327. Ramsey, J. M. et al. Distribution of domestic Triatominae and stratification of Chagas disease transmission in Oaxaca, Mexico. *Med. Vet. Entomol.* **14**, 19–30 (2000).
328. Costa, J., Dornak, L. L., Almeida, C. E. & Peterson, A. T. Distributional potential of the *Triatoma brasiliensis* species complex at present and under scenarios of future climate conditions. *Parasit. Vectors* **7**, 238 (2014).
329. Garrido, R. et al. Potential impact of climate change on the geographical distribution of two wild vectors of Chagas disease in Chile: *Mepraia spinolai* and *Mepraia gajardoi*. *Parasit. Vectors* **12**, 478 (2019).
330. Grijalva, M. J., Villacís, A. G., Ocaña-Mayorga, S. et al. Limitations of selective deltamethrin application for triatomine control in central coastal Ecuador. *Parasit. Vectors* **4**, 20 (2011).

331. de Noya, B. A. & González, O. N. An ecological overview on the factors that drives to *Trypanosoma cruzi* oral transmission. *Acta Trop.* **151**, 94–102 (2015).
332. Forattini, O. P. Biogeography, origin and distribution of Triatominae domicile dispersal in Brazil. *Rev. Saude Publica* **14**, 265–299 (1980).
333. Berry, A. S. F. et al. Dispersal patterns of *Trypanosoma cruzi* in Arequipa, Peru. *bioRxiv* 838235 (2019), doi:10.1101/838235.
334. Gillespie, J. H. Population genetics: a concise guide. *JHU Press* (2010).
335. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
336. Excoffier, L. & Heckel, G. Computer programs for population genetics data analysis: a survival guide. *Nat. Rev. Genet.* **7**, 745–758 (2006).
337. Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* **4**, 981–994 (2003).
338. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
339. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
340. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
341. Putman, A. I. & Carbone, I. Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecol. Evol.* **4**, 4399–4428 (2014).
342. Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M. & Holderegger, R. A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* **24**, 4348–437 (2015).
343. Prugnolle, F. & de Meeus, T. Apparent high recombination rates in clonal parasitic organisms due to inappropriate sampling design. *Heredity* **104**, 135–140 (2009).
344. Garin, Y. J.-F. et al. Virulence of *Leishmania infantum* is expressed as a clonal and dominant phenotype in experimental infections. *Infect. Immun.* **69**, 7365–7373 (2001).
345. Choisy, M. & de Roode, J. C. Mixed infections and the evolution of virulence: effects of resource competition, parasite plasticity, and impaired host immunity. *Am. Nat.* **175**, E105-118 (2010).
346. Taylor, A. R. et al. Estimation of malaria haplotype and genotype frequencies: a statistical approach to overcome the challenge associated with multiclonal infections. *Malar. J.* **13**, 102 (2014).

347. Noulin, F., Borlon, C., Van Den Abbeele, J., D'Alessandro, U. & Erhart, A. 1912-2012: a century of research on *Plasmodium vivax in vitro* culture. *Trends Parasitol.* **29**, 286–294 (2013).
348. Cuypers, B. et al. Multiplexed spliced-leader sequencing: a high-throughput, selective method for RNA-seq in trypanosomatids. *Sci. Rep.* **7**, 1–11 (2017).
349. Domagalska, M. A. et al. Genomes of intracellular *Leishmania* parasites directly sequenced from patients. *bioRxiv* 676163 (2019), doi:10.1101/676163.
350. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939 (2011).
351. Ponstingl, H. & Ning, Z. SMALT - a new mapper for DNA sequencing reads (poster). *F1000Posters* **1**, 313 (2010).
352. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
353. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
354. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
355. Dormann, C. F. et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27–46 (2013).
356. Strasburg, J. L. et al. What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 364–373 (2012).
357. de Mita, S. et al. Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.* **22**, 1383–1399 (2013).
358. Frichot, E., Schoville, S. D., de Villemereuil, P., Gaggiotti, O. E. & François, O. Detecting adaptive evolution based on association with ecological gradients: orientation matters! *Heredity* **115**, 22–28 (2015).
359. Lotterhos, K. E. & Whitlock, M. C. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol. Ecol.* **24**, 1031–1046 (2015).
360. de Villemereuil, P., Frichot, É., Bazin, É., François, O. & Gaggiotti, O. E. Genome scan methods against more complex models: when and how much should we trust them? *Mol. Ecol.* **23**, 2006–2019 (2014).
361. Landguth, E. L. et al. Quantifying the lag time to detect barriers in landscape genetics. *Mol. Ecol.* **19**, 4179–4191 (2010).

362. Landguth, E. L. & Schwartz, M. K. Evaluating sample allocation and effort in detecting population differentiation for discrete and continuously distributed individuals. *Conserv. Genet.* **15**, 981–992 (2014).
363. Willing, E.-M., Dreyer, C. & van Oosterhout, C. Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. *PLoS One* **7**, e42649 (2012).
364. Landguth, E. L. et al. Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. *Mol. Ecol. Resour.* **12**, 276–284 (2012).
365. Tucker, J. M., Schwartz, M. K., Truex, R. L., Pilgrim, K. L. & Allendorf, F. W. Historical and contemporary DNA indicate fisher decline and isolation occurred prior to the European settlement of California. *PLoS One* **7**, e52803 (2012).
366. Tucker, J. M., Schwartz, M. K., Truex, R. L., Wisely, S. M. & Allendorf, F. W. Sampling affects the detection of genetic subdivision and conservation implications for fisher in the Sierra Nevada. *Conserv. Genet.* **15**, 123–136 (2013).
367. Lloyd-Smith, J. O. et al. Epidemic dynamics at the human-animal interface. *Science* **326**, 1362–1367 (2009).
368. White, L. A., Forester, J. D. & Craft, M. E. Dynamic, spatial models of parasite transmission in wildlife: their structure, applications and remaining challenges. *J. Anim. Ecol.* **87**, 559–580 (2018).
369. Mejía-Jaramillo, A. M. et al. Genotyping of *Trypanosoma cruzi* in a hyper-endemic area of Colombia reveals an overlap among domestic and sylvatic cycles of Chagas disease. *Parasit. Vectors* **7**, 108 (2014).
370. Pennington, P. M. et al. The Chagas disease domestic transmission cycle in Guatemala: Parasite-vector switches and lack of mitochondrial co-diversification between *Triatoma dimidiata* and *Trypanosoma cruzi* subpopulations suggest non-vectorial parasite dispersal across the Motagua valley. *Acta Trop.* **151**, 80–87 (2015).
371. Correa Antonialli, S. A., Torres, T. G., Paranhos Filho, A. C. & Tolezano, J. E. Spatial analysis of American visceral leishmaniasis in Mato Grosso do Sul state, central Brazil. *J. Infect.* **54**, 509–514 (2007).
372. Bern, C., Martin, D. L. & Gilman, R. H. Acute and congenital Chagas disease. *Adv. Parasitol.* **75**, 19–47 (2011).
373. Coura, J. R. & Viñas, P. A. Chagas disease: a new worldwide challenge. *Nature* **465**, S6–7 (2010).
374. Simpson, A. G. B., Stevens, J. R. & Lukeš, J. The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol.* **22**, 168–174 (2006).

375. Benne, R. et al. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**, 819–826 (1986).
376. Agabian, N. Trans splicing of nuclear pre-mRNAs. *Cell* **61**, 1157–1160 (1990).
377. Ramesh, M. A., Malik, S.-B. & Logsdon, J. M. A Phylogenomic inventory of meiotic genes: evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. *Curr. Biol.* **15**, 185–191 (2005).
378. Tait, A. Evidence for diploidy and mating in trypanosomes. *Nature* **287**, 536–538 (1980).
379. MacLeod, A. et al. Allelic segregation and independent assortment in *T. brucei* crosses: proof that the genetic system is Mendelian and involves meiosis. *Mol. Biochem. Parasitol.* **143**, 12–19 (2005).
380. Peacock, L., Bailey, M., Carrington, M. & Gibson, W. Meiosis and haploid gametes in the pathogen *Trypanosoma brucei*. *Curr. Biol.* **24**, 181–186 (2014).
381. Ravel, C. et al. First report of genetic hybrids between two very divergent *Leishmania* species: *Leishmania infantum* and *Leishmania major*. *Int. J. Parasitol.* **36**, 1383–1388 (2006).
382. Inbar, E. et al. Whole genome sequencing of experimental hybrids supports meiosis-like sexual recombination in *Leishmania*. *PLoS Genet.* **15**, e1008042 (2019).
383. Rougeron, V. et al. Extreme inbreeding in *Leishmania braziliensis*. *Proc. Natl. Acad. Sci. USA* **106**, 10224–10229 (2009).
384. Minning, T. A., Weatherly, D. B., Flibotte, S. & Tarleton, R. L. Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in *Trypanosoma cruzi* strains revealed by array comparative genomic hybridization. *BMC Genomics* **12**, 139 (2011).
385. Grijalva, M. J. et al. Comprehensive survey of domiciliary triatomine species capable of transmitting Chagas disease in southern Ecuador. *PLoS Negl. Trop. Dis.* **9**, e0004142 (2015).
386. 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).
387. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).
388. Picard Toolkit. *Broad Institute*, http://broadinstitute.github.io/picard.
389. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

390. Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS One* **7**, e30377 (2012).
391. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
392. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
393. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
394. R: the R project for statistical computing. *CRAN*, https://www.r-project.org/.
395. Goudet, J. hierfstat, a package for r to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186 (2005).
396. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
397. Tsai, I. J., Bensasson, D., Burt, A. & Koufopanou, V. Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. *Proc. Natl. Acad. Sci. USA* **105**, 4957–4962 (2008).
398. McVean, G. A. T. et al. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
399. Ewing, A. D. et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).
400. Martin, S. H. & van Belleghem, S. M. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* **206**, 429–438 (2017).
401. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
402. Rutherford, K. et al. Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945 (2000).
403. Excoffier, L. & Smouse, P. E. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics* **136**, 343–359 (1994).
404. Mark Welch, D. & Meselson, M. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* **288**, 1211–1215 (2000).
405. Bennett, R. J. & Johnson, A. D. Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains. *EMBO J.* **22**, 2505–2515 (2003).

406. Bennett, R. J., Forche, A. & Berman, J. Rapid mechanisms for generating genome diversity: whole ploidy shifts, aneuploidy, and loss of heterozygosity. *Cold Spring Harb. Perspect. Med.* **4**, a019604 (2014).
407. Smith, K. N. & Nicolas, A. Recombination at work for meiosis. *Curr. Opin. Genet. Dev.* **8**, 200–211 (1998).
408. Tihon, E., Imamura, H., Dujardin, J.-C. & Van Den Abbeele, J. Evidence for viable and stable triploid *Trypanosoma congolense* parasites. *Parasit. Vectors* **10**, 468 (2017).
409. Almeida, L. V. et al. Chromosomal copy number variation analysis by next generation sequencing confirms ploidy stability in *Trypanosoma brucei* subspecies. *Microb. Genomics* **4**, e000223 (2018).
410. Zeledón, R., Alvarado, R. & Jirón, L. F. Observations on the feeding and defecation patterns of three triatomine species (Hemiptera: Reduviidae). *Acta Trop.* **34**, 65–77 (1977).
411. Gürtler, R. E. & Cardinal, M. V. Reservoir host competence and the role of domestic and commensal hosts in the transmission of *Trypanosoma cruzi*. *Acta Trop.* **151**, 32–50 (2015).
412. Grijalva, M. J., Terán, D. & Dangles, O. Dynamics of sylvatic Chagas disease vectors in coastal Ecuador is driven by changes in land cover. *PLoS Negl. Trop. Dis.* **8**, e2960 (2014).
413. Schenck, R. A. & Vrijenhoek, R. C. Spatial and temporal factors affecting coexistence among sexual and clonal forms of *Poeciliopsis*. *Evol. Int. J. Org. Evol.* **40**, 1060–1070 (1986).
414. Verduijn, M. H., Dijk, P. J. V. & Damme, J. M. M. V. Distribution, phenology and demography of sympatric sexual and asexual dandelions (*Taraxacum officinale* s.l.): geographic parthenogenesis on a small scale. *Biol. J. Linn. Soc.* **82**, 205–218.
415. Vorburger, C., Lancaster, M. & Sunnucks, P. Environmentally related patterns of reproductive modes in the aphid *Myzus persicae* and the predominance of two 'superclones' in Victoria, Australia. *Mol. Ecol.* **12**, 3493–3504 (2003).
416. Lehto, M. P. & Haag, C. R. Ecological differentiation between coexisting sexual and asexual strains of *Daphnia pulex*. *J. Anim. Ecol.* **79**, 1241–1250 (2010).
417. Haag, C. R. & Ebert, D. A new hypothesis to explain geographic parthenogenesis. *Ann. Zool. Fenn.* **41**, 539–544 (2004).
418. Gibson, W. & Garside, L. Genetic exchange in *Trypanosoma brucei brucei*: variable chromosomal location of housekeeping genes in different trypanosome stocks. *Mol. Biochem. Parasitol.* **45**, 77–89 (1991).

419. Lin, X. et al. αADα hybrids of *Cryptococcus neoformans*: evidence of same-sex mating in nature and hybrid fitness. *PLoS Genet.* **3**, e186 (2007).
420. Wakeley, J. & Aliacar, N. Gene genealogies in a metapopulation. *Genetics* **159**, 893–905 (2001).
421. González-Martínez, S. C., Ridout, K. & Pannell, J. R. Range expansion compromises adaptive evolution in an outcrossing plant. *Curr. Biol.* **27**, 2544-2551.e4 (2017).
422. Hartfield, M., Wright, S. I. & Agrawal, A. F. Coalescent times and patterns of genetic diversity in species with facultative sex: effects of gene conversion, population structure, and heterogeneity. *Genetics* **202**, 297–312 (2016).
423. Gibson, W., Peacock, L., Ferris, V., Williams, K. & Bailey, M. The use of yellow fluorescent hybrids to indicate mating in *Trypanosoma brucei*. *Parasit. Vectors* **1**, 4 (2008).
424. Peacock, L., Ferris, V., Bailey, M. & Gibson, W. Fly transmission and mating of *Trypanosoma brucei brucei* strain 427. *Mol. Biochem. Parasitol.* **160**, 100–106 (2008).
425. Peacock, L., Ferris, V., Bailey, M. & Gibson, W. Mating compatibility in the parasitic protist *Trypanosoma brucei*. *Parasit. Vectors* **7**, 78 (2014).
426. 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).
427. Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236 (1963).
428. Clement, M., Posada, D. & Crandall, K. A. TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* **9**, 1657–1659 (2000).
429. Petit, R. J. Early insights into the genetic consequences of range expansions. *Heredity* **106**, 203–204 (2011).
430. Kolbe, J. J. et al. Genetic variation increases during biological invasion by a Cuban lizard. *Nature* **9,** 177–181 (2004).
431. Gradoni, L. & Gramiccia, M. *Leishmania infantum* tropism: strain genotype or host immune status? *Parasitol. Today* **10**, 264–267 (1994).
432. Sulahian, A., Garin, Y. J., Pratlong, F., Dedet, J. P. & Derouin, F. Experimental pathogenicity of viscerotropic and dermotropic isolates of *Leishmania infantum* from immunocompromised and immunocompetent patients in a murine model. *FEMS Immunol. Med. Microbiol.* **17**, 131–138 (1997).
433. Guerbouj, S., Guizani, I., Speybroeck, N., Le Ray, D. & Dujardin, J. C. Genomic polymorphism of *Leishmania infantum*: a relationship with clinical pleomorphism? *Infect. Genet. Evol.* **1**, 49–59 (2001).

434. Dumetz, F. et al. Molecular preadaptation to antimony resistance in *Leishmania donovani* on the Indian subcontinent. *mSphere* **3**, e00548 (2018).
435. Ishikawa, E. a. Y. et al. Genetic variation in populations of *Leishmania* species in Brazil. *Trans. R. Soc. Trop. Med. Hyg.* **96 Suppl 1**, S111-121 (2002).
436. Cupolillo, E. et al. Genetic polymorphism and molecular epidemiology of *Leishmania* (*Viannia*) *braziliensis* from different hosts and geographic areas in Brazil. *J. Clin. Microbiol.* **41**, 3126–3132 (2003).
437. Victoir, K. & Dujardin, J.-C. How to succeed in parasitic life without sex? Asking *Leishmania*. *Trends Parasitol.* **18**, 81–85 (2002).
438. Cortes, S. et al. Risk factors for canine leishmaniasis in an endemic Mediterranean region. *Vet. Parasitol.* **189**, 189–196 (2012).
439. Calvo-Álvarez, E. et al. First evidence of intraclonal genetic exchange in trypanosomatids using two *Leishmania infantum* fluorescent transgenic clones. *PLoS Negl. Trop. Dis.* **8**, e3075 (2014).
440. Akhoundi, M. et al. A historical overview of the classification, evolution, and dispersion of *Leishmania* parasites and sandflies. *PLoS Negl. Trop. Dis.* **10**, e0004349 (2016).
441. Lainson, R. & Shaw, J. J. New World leishmaniasis; in Topley and Wilson's microbiology and microbial infections (10th edition) (eds. Cox, F. E. G., Wakelin, D., Gillespie, S. H. & Despommier, D. D.) 313–349. *Hodder Arnold* (2005).
442. Satragno, D. et al. Autochthonous outbreak and expansion of canine visceral leishmaniasis, Uruguay. *Emerg. Infect. Dis.* **23**, 536–538.
443. Departamento de Vigilância Epidemiológica. Manual de vigilância e controle da leishmaniose visceral. *Ministério da Saúde* (2003).
444. Bern, C., Maguire, J. H. & Alvar, J. Complexities of assessing the disease burden attributable to leishmaniasis. *PLoS Negl. Trop. Dis.* **2**, e313 (2008).
445. Belli, A. et al. Widespread atypical cutaneous leishmaniasis caused by *Leishmania* (*L.*) *chagasi* in Nicaragua. *Am. J. Trop. Med. Hyg.* **61**, 380–385 (1999).
446. De Lima, H. et al. Cutaneous leishmaniasis due to *Leishmania chagasi*/*Le. infantum* in an endemic area of Guarico State, Venezuela. *Trans. R. Soc. Trop. Med. Hyg.* **103**, 721–726 (2009).
447. Freitas-Mesquita, A. L. et al. Inhibitory effects promoted by 5′-nucleotides on the ecto-3′-nucleotidase activity of *Leishmania amazonensis*. *Exp. Parasitol.* **169**, 111–118 (2016).

448. Guimarães-Costa, A. B. et al. 3'-nucleotidase/nuclease activity allows *Leishmania* parasites to escape killing by neutrophil extracellular traps. *Infect. Immun.* **82**, 1732–1740 (2014).
449. Guimarães-Costa, A. B. et al. *Leishmania amazonensis* promastigotes induce and are killed by neutrophil extracellular traps. *Proc. Natl. Acad. Sci. USA* **106**, 6748–6753 (2009).
450. Rosenzweig, D. et al. Retooling *Leishmania* metabolism: from sand fly gut to human macrophage. *FASEB J.* **22**, 590–602 (2008).
451. Carnielli, J. B. T. et al. Natural resistance of *Leishmania infantum* to miltefosine contributes to the low efficacy in the treatment of visceral leishmaniasis in Brazil. *Am. J. Trop. Med. Hyg.* **101**, 789 (2019).
452. Romero, G. A. S. et al. Efficacy and safety of available treatments for visceral leishmaniasis in Brazil: a multicenter, randomized, open label trial. *PLoS Negl. Trop. Dis.* **11**, e0005706 (2017).
453. Sunyoto, T., Potet, J. & Boelaert, M. Why miltefosine—a life-saving drug for leishmaniasis—is unavailable to people who need it the most. *BMJ Glob. Health* **3**, e000709 (2018).
454. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
455. Gower, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338 (1966).
456. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
457. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
458. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
459. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
460. Foll, M. & Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977–993 (2008).

461. Marco-Sola, S. & Ribeca, P. Efficient alignment of Illumina-like high-throughput sequencing reads with the genomic multi-tool (GEM) mapper. *Curr. Protoc. Bioinforma.* **50**, 11.13.1–11.13.20 (2015).
462. Warnes, G. R. et al. gplots: various R programming tools for plotting data (2020).
463. Oksanen, J. et al. vegan: community ecology package (2019).
464. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods.* **17**, 261–272 (2020).
465. Fox, J. et al. car: companion to applied regression (2019).
466. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2 ΔΔCt method. *Methods.* **25**, 402–408 (2001).
467. Young, D. G. Guide to the identification and geographic distribution of *Lutzomyia* sand flies in Mexico, the West Indies, Central and South America (Diptera: Psychodidae). *PN* (1994).
468. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
469. Blomberg, S. P., Garland, T. J. & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**, 717–745 (2003).
470. Paletta-Silva, R. et al. *Leishmania amazonensis*: characterization of an ecto-3′-nucleotidase activity and its possible role in virulence. *Exp. Parasitol.* **129**, 277–283 (2011).
471. Vieira, D. P., Paletta-Silva, R., Saraiva, E. M., Lopes, A. H. & Meyer-Fernandes, J. R. *Leishmania chagasi*: an ecto-3′-nucleotidase activity modulated by inorganic phosphate and its possible involvement in parasite–macrophage interaction. *Exp. Parasitol.* **127**, 702-707 (2011).
472. Hallatschek, O., Hersen, P., Ramanathan, S. & Nelson, D. R. Genetic drift at expanding frontiers promotes gene segregation. *Proc. Natl. Acad. Sci. USA* **104**, 19926–19930 (2007).
473. Peischl, S. et al. Relaxed selection during a recent human expansion. *Genetics* **208**, 763–777 (2018).
474. Graciá, E. et al. Surfing in tortoises? Empirical signs of genetic structuring owing to range expansion. *Biol. Lett.* **9**, 20121091 (2013).
475. Melo-Ferreira, J., Alves, P. C., Rocha, J., Ferrand, N. & Boursot, P. Interspecific x-chromosome and mitochondrial DNA introgression in the Iberian hare: selection or allele surfing? *Evolution* **65**, 1956–1968 (2011).
476. Peischl, S., Dupanloup, I., Kirkpatrick, M. & Excoffier, L. On the accumulation of deleterious mutations during range expansions. *Mol. Ecol.* **22**, 5972–5982 (2013).

477. Gottlieb, M. Enzyme regulation in a trypanosomatid: effect of purine starvation on levels of 3'-nucleotidase activity. *Science* **227**, 72–74 (1985).
478. Cohn, C. S. & Gottlieb, M. The acquisition of purines by trypanosomatids. *Parasitol. Today* **13**, 231–235 (1997).
479. Leite, P. M. et al. Ecto-nucleotidase activities of promastigotes from *Leishmania* (*Viannia*) *braziliensis* relate to parasite infectivity and disease clinical outcome. *PLoS Negl. Trop. Dis.* **6**, e1850 (2012).
480. Alizon, S. Transmission-recovery trade-offs to study parasite evolution. *Am. Nat.* **172**, E113-21 (2008).
481. Buckingham-Jeffery, E. et al. Spatio-temporal modelling of *Leishmania infantum* infection among domestic dogs: a simulation study and sensitivity analysis applied to rural Brazil. *Parasit. Vectors* **12**, 215 (2019).
482. Vasudevan, G., Ullman, B. & Landfear, S. M. Point mutations in a nucleoside transporter gene from *Leishmania donovani* confer drug resistance and alter substrate selectivity. *Proc. Natl. Acad. Sci. USA* **98**, 6092–6097 (2001).
483. Rastrojo, A. et al. Genomic and transcriptomic alterations in *Leishmania donovani* lines experimentally resistant to antileishmanial drugs. *Int. J. Parasitol.* **8**, 246–264 (2018).
484. Patino, L. H., Muskus, C. & Ramírez, J. D. Transcriptional responses of *Leishmania* (*Leishmania*) *amazonensis* in the presence of trivalent sodium stibogluconate. *Parasit. Vectors* **12**, 348 (2019).
485. Kuhls, K. et al. Multilocus microsatellite typing (MLMT) reveals genetically isolated populations between and within the main endemic regions of visceral leishmaniasis. *Microbes Infect.* **9**, 334–343 (2007).
486. 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**, e05166 (2015).
487. Hall, M. D. et al. Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. *eLife* **8**, e46402 (2019).
488. 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).
489. Miotto, O. et al. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nat. Genet.* **47**, 226–234 (2015).
490. Auburn, S. et al. Genomic analysis of a pre-elimination Malaysian *Plasmodium vivax* population reveals selective pressures and changing transmission dynamics. *Nat. Commun.* **9**, 2585 (2018).

491. 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).
492. Dvorak, J. A., Hartman, D. L. & Miles, M. A. *Trypanosoma cruzi*: correlation of growth kinetics to zymodeme type in clones derived from various sources. *J. Protozool.* **27**, 472–474 (1980).
493. Lima, F. M. et al. Interclonal variations in the molecular karyotype of *Trypanosoma cruzi*: chromosome rearrangements in a single cell-derived clone of the G strain. *PLoS One* **8**, e63738 (2013).
494. Kumar, N. et al. Efficient subtraction of insect rRNA prior to transcriptome analysis of *Wolbachia-Drosophila* lateral gene transfer. *BMC Res. Notes* **5**, 230 (2012).
495. Oyola, S. O. et al. Efficient depletion of host DNA contamination in malaria clinical sequencing. *J. Clin. Microbiol.* **51**, 745–751 (2013).
496. Feehery, G. R. et al. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PLoS One* **8**, e76096 (2013).
497. Melnikov, A. et al. Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol.* **12**, R73 (2011).
498. Schuenemann, V. J. et al. Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science* **341**, 179–183 (2013).
499. Metsky, H. C. et al. Zika virus evolution and spread in the Americas. *Nature* **546**, 411–415 (2017).
500. Cowell, A. N. et al. Selective whole-genome amplification is a robust method that enables scalable whole-genome sequencing of *Plasmodium vivax* from unprocessed clinical samples. *mBio* **8**, e02257 (2017).
501. Hintzsche, J. D., Robinson, W. A. & Tan, A. C. A survey of computational tools to analyze and interpret whole exome sequencing data. *Int. J. Genomics* **2016**, e7983236 (2016).
502. Gampawar, P. et al. Evaluation of the performance of AmpliSeq and SureSelect exome sequencing libraries for Ion Proton. *Front. Genet.* **10**, 856 (2019).
503. Nag, S. et al. High throughput resistance profiling of *Plasmodium falciparum* infections based on custom dual indexing and Illumina next generation sequencing-technology. *Sci. Rep.* **7**, 2398 (2017).
504. Momčilović, S., Cantacessi, C., Arsić-Arsenijević, V., Otranto, D. & Tasić-Otašević, S. Rapid diagnosis of parasitic diseases: current scenario and future needs. *Clin. Microbiol. Infect.* **25**, 290–309 (2019).

505. Arias, A. et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* **2**, vew016 (2016).
506. Park, J. et al. Determining genotypic drug resistance by ion semiconductor sequencing with the Ion AmpliSeq^{TM} TB panel in multidrug-resistant *Mycobacterium tuberculosis* isolates. *Ann. Lab. Med.* **38**, 316–323 (2018).
507. Ferrario, C. et al. A genome-based identification approach for members of the genus *Bifidobacterium*. *FEMS Microbiol. Ecol.* **91**, fiv009 (2015).
508. Makowsky, R. et al. Genomic diversity and phylogenetic relationships of human papillomavirus 16 (HPV16) in Nepal. *Infect. Genet. Evol.* **46**, 7–11 (2016).
509. Grijalva, M. J., Suarez-Davalos, V., Villacis, A. G., Ocaña-Mayorga, S. & Dangles, O. Ecological factors related to the widespread distribution of sylvatic *Rhodnius ecuadoriensis* populations in southern Ecuador. *Parasit. Vectors* **5**, 17 (2012).
510. Nascimento, J. D. et al. Taxonomical over splitting in the *Rhodnius prolixus* (Insecta: Hemiptera: Reduviidae) clade: Are *R. taquarussuensis* (da Rosa et al., 2017) and *R. neglectus* (Lent, 1954) the same species? *PLoS One* **14**, e0211285 (2019).
511. Velásquez-Ortiz, N. et al. *Trypanosoma cruzi* infection, discrete typing units and feeding sources among *Psammolestes arthuri* (Reduviidae: Triatominae) collected in eastern Colombia. *Parasit. Vectors* **12**, 157 (2019).
512. Caicedo-Garzón, V. et al. Genetic diversification of *Panstrongylus geniculatus* (Reduviidae: Triatominae) in northern South America. *PLoS One* **14**, e0223963 (2019).
513. Carrasco, H. J., Torrellas, A., García, C., Segovia, M. & Feliciangeli, M. D. Risk of *Trypanosoma cruzi* I (Kinetoplastida: Trypanosomatidae) transmission by *Panstrongylus geniculatus* (Hemiptera: Reduviidae) in Caracas (Metropolitan District) and neighboring states, Venezuela. *Int. J. Parasitol.* **35**, 1379–1384 (2005).
514. Carrasco, H. J. et al. Geographical distribution of *Trypanosoma cruzi* genotypes in Venezuela. *PLoS Negl. Trop. Dis.* **6**, e1707 (2012).
515. Nakad, B. C. C. et al. Genetic variability of *Panstrongylus geniculatus* (Reduviidae: Triatominae) in the Metropolitan District of Caracas, Venezuela. *Infect. Genet. Evol.* **66**, 236–244 (2018).
516. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
517. You, F. M. et al. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9**, 253 (2008).
518. Sonnhammer, E. L. & Hollich, V. Scoredist : a simple and robust protein sequence distance estimator. *BMC Bioinformatics* **6**, 108 (2005).

519. Cummings, K. L. & Tarleton, R. L. Rapid quantitation of *Trypanosoma cruzi* in host tissue by real-time PCR. *Mol. Biochem. Parasitol.* **129**, 53–59 (2003).
520. PhiX Sequencing Control V3. https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html.
521. Access Array System for Illumina Sequencing Systems (user guide). https://docplayer.net/78505463-Access-array-system-for-illumina-sequencing-systems.html.
522. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS One* **6**, e17288 (2011).
523. Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).
524. Purcell, S. et al. PLINK: a tool let for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
525. Ritland, K. Inferences about inbreeding depression based on changes of the inbreeding coefficient. *Evolution* **44**, 1230–1241 (1990).
526. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887–893 (2005).
527. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462 (1995).
528. Šavrič, B., Jenny, B. & Jenny, H. Projection wizard – an online map projection selection tool. *Cartogr. J.* **53**, 177–185 (2016).
529. Wiens, J. J. & Morrill, M. C. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst. Biol.* **60**, 719–731 (2011).
530. Slatkin, M. Isolation by distance in equilibrium and non-equilibrium populations. *Evol. Int. J. Org. Evol.* **47**, 264–279 (1993).
531. Llewellyn, M. S. The molecular epidemiology of *Trypanosoma cruzi* infection in wild and domestic transmission cycles with special emphasis on multilocus microsatellite analysis (PhD dissertation). *London School of Hygiene & Tropical Medicine* (2008).
532. Shibata, H. et al. The use of PCR in detecting toxoplasma parasites in the blood and brains of mice experimentally infected with *Toxoplasma gondii*. *Kansenshogaku Zasshi* **69**, 158–163 (1995).
533. Yang, H., Golenberg, E. M. & Shoshani, J. Proboscidean DNA from museum and fossil specimens: an assessment of ancient DNA extraction and amplification techniques. *Biochem. Genet.* **35**, 165–179 (1997).

534. Ramos, R. A. N. et al. Quantification of *Leishmania infantum* DNA in the bone marrow, lymph node and spleen of dogs. *Rev. Bras. Parasitol. Vet.* **22**, 346–350 (2013).
535. Schubert, G. et al. Targeted detection of mammalian species using carrion fly – derived DNA. *Mol. Ecol. Resour.* **15**, 285–294 (2014).
536. Côté, N. M. L. et al. A new high-throughput approach to genotype ancient human gastrointestinal parasites. *PLoS One* **11**, e0146230 (2016).
537. Cencig, S., Coltel, N., Truyens, C. & Carlier, Y. Parasitic loads in tissues of mice infected with *Trypanosoma cruzi* and treated with AmBisome. *PLoS Negl. Trop. Dis.* **5**, e1216 (2011).
538. Souza, R. T. et al. Genome size, karyotype polymorphism and chromosomal evolution in *Trypanosoma cruzi*. *PLoS One* **6**, e23042 (2011).
539. Reithinger, R., Lambson, B. E., Barker, D. C. & Davies, C. R. Use of PCR to detect *Leishmania* (*Viannia*) spp. in dog blood and bone marrow. *J. Clin. Microbiol.* **38**, 5 (2000).
540. Wen, C. et al. Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLoS One* **12**, e0176716 (2017).
541. Storfer, A., Patton, A. & Fraik, A. K. Navigating the interface between landscape genetics and landscape genomics. *Front. Genet.* **9**, 68 (2018).
542. Erben, E. D. High-throughput methods for dissection of trypanosome gene regulatory networks. *Curr. Genomics* **19**, 78–86 (2018).
543. Quinlan, A. R & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**, 841– 842 (2010).
544. Aurrecoechea, C. et al. EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic Acids Res.* **45**, D581–D591 (2017).
545. Linck, E. & Battey, C. J. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol. Ecol. Resour.* **19**, 639–647 (2019).
546. Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. Inferring species tees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29**, 1917–1932 (2012).
547. Piry, S. et al. GENECLASS2: a software for genetic assignment and first-generation migrant detection. *J. Hered.* **95**, 536–539 (2004).
548. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224–1228 (2013).

549. Anderson, E. C. & Thompson, E. A. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217–1229 (2002).
550. Graffelman, J., Jain, D. & Weir, B. A genome-wide study of Hardy–Weinberg equilibrium with next generation sequence data. *Hum. Genet.* **136**, 727–741 (2017).
551. Sefid Dashti, M. J. & Gamieldien, J. A practical guide to filtering and prioritizing genetic variants. *Biotechniques* **62**, 18–30 (2017).
552. Kaplinski, L., Andreson, R., Puurand, T. & Remm, M. MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics* **21**, 1701–1702 (2005).
553. Etherington, T. R. Python based GIS tools for landscape genetics: visualising genetic relatedness and measuring landscape connectivity. *Methods Ecol. Evol.* **2**, 52–55 (2011).
554. 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).
555. 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).
556. 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**, e1005348 (2017).
557. 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).
558. 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).
559. Nunes, M. C. P. 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).
560. Vazquez-Prokopec, G. M. et al. Coupled heterogeneities and their impact on parasite transmission and control. *Trends Parasitol.* **32**, 356–367 (2016).
561. Vazquez-Prokopec, G. M., Spillmann, C., Zaidenberg, M., Gürtler, R. E. & Kitron, U. Spatial heterogeneity and risk maps of community infestation by *Triatoma infestans* in rural northwestern Argentina. *PLoS Negl. Trop. Dis.* **6**, e1788 (2012).

562. Blanquart, F., Kaltz, O., Nuismer, S. L. & Gandon, S. A practical guide to measuring local adaptation. *Ecol. Lett.* **16**, 1195–1205 (2013).
563. Roper, C., Pearce, R., Nair, S., Sharp, B., Nosten, F. & Anderson, T. Intercontinental spread of pyrimethamine-resistant malaria. *Science* **305**, 1124 (2004).
564. Fitzpatrick, S., Feliciangeli, M. D., Sanchez-Martin, M. J., Monteiro, F. A. & Miles, M. A. Molecular genetics reveal that silvatic *Rhodnius prolixus* do colonise rural houses. *PLoS Negl. Trop. Dis.* **2**, e210 (2008).
565. Manel, S., Schwartz, M. K., Luikart, G. & Taberlet, P. Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.* **18**, 189–197 (2003).
566. Cushman, S. A., Shirk, A. & Landguth, E. L. Separating the effects of habitat area, fragmentation and matrix resistance on genetic differentiation in complex landscapes. *Landsc. Ecol.* **27**, 369–380 (2012).
567. Wasserman, T. N., Cushman, S. A., Littell, J. S., Shirk, A. J. & Landguth, E. L. Population connectivity and genetic diversity of American marten (*Martes americana*) in the United States northern Rocky Mountains in a climate change context. *Conserv. Genet. 142 529-541* **14**, 529–541 (2013).
568. Sommer, S., McDevitt, A. D. & Balkenhol, N. Landscape genetic approaches in conservation biology and management. *Conserv. Genet.* **14**, 249–251 (2013).
569. Real, L. A. et al. Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. *Proc. Natl. Acad. Sci. USA.* **102**, 12107–12111 (2005).
570. Blanchong, J. A. et al. Landscape genetics and the spatial distribution of chronic wasting disease. *Biol. Lett.* **4**, 130–133 (2008).
571. Biek, R. & Real, L. A. The landscape genetics of infectious disease emergence and spread. *Mol. Ecol.* **19**, 3515–3531 (2010).
572. Dellicour, S., Rose, R. & Pybus, O. G. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics* **17**, 82 (2016).
573. Savage, A. E., Becker, C. G. & Zamudio, K. R. Linking genetic and environmental factors in amphibian disease risk. *Evol. Appl.* **8**, 560–562 (2015).
574. Criscione, C. D. et al. Landscape genetics reveals focal transmission of a human macroparasite. *PLoS Negl. Trop. Dis.* **4**, e665 (2010).
575. Morgan, K., McGaughran, A., Ganeshan, S., Herrmann, M. & Sommer, R. J. Landscape and oceanic barriers shape dispersal and population structure in the island nematode *Pristionchus pacificus*. *Biol. J. Linn. Soc.* **12**, 1–15 (2014).
576. Liang, L., Liu, Y., Liao, J. & Gong, P. Wetlands explain most in the genetic divergence pattern of *Oncomelania hupensis*. *Infect. Genet. Evol.* **27**, 436–444 (2014).

577. Medley, K. A., Jenkins, D. G. & Hoffman, E. A. Human-aided and natural dispersal drive gene flow across the range of an invasive mosquito. *Mol. Ecol.* **24**, 284–295 (2015).
578. Bouyer, J. et al. Mapping landscape friction to locate isolated tsetse populations that are candidates for elimination. *Proc. Natl. Acad. Sci. USA.* **112**, 14575–14580 (2015).
579. Leo, S. S., Gonzalez, A. & Millien, V. Multi-taxa integrated landscape genetics for zoonotic infectious diseases: deciphering variables influencing disease emergence. *Genome* **59**, 349–361 (2016).
580. Sprehn, C. G., Blum, M. J., Quinn, T. P. & Heins, D. C. Landscape genetics of *Schistocephalus solidus* parasites in threespine stickleback (*Gasterosteus aculeatus*) from Alaska. *PLoS One* **10**, e0122307 (2015).
581. Landguth, E., Cushman, S. A. & Balkenhol, N. Simulation modeling in landscape genetics; in Landscape Genetics (eds. Balkenhol, N., Cushman, S. A., Storfer, A. T. & Waits, L. P.) 99–113. *John Wiley & Sons* (2015).
582. Bern, C. Chagas' Disease. *N. Engl. J. Med.* **373**, 456–466 (2015).
583. Manel, S. & Holderegger, R. Ten years of landscape genetics. *Trends Ecol. Evol.* **28**, 614–621 (2013).
584. Bowcock, A. M. et al. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457 (1994).
585. Rousset, F. & Leblois, R. Likelihood-based inferences under isolation by distance: two-dimensional habitats and confidence intervals. *Mol. Biol. Evol.* **29**, 957–973 (2012).
586. Adriaensen, F. et al. The application of 'least-cost' modelling as a functional landscape model. *Landsc. Urban Plan.* **64**, 233–247 (2003).
587. McRae, B. H. Isolation by resistance. *Evolution* **60**, 1551–1561 (2006).
588. Spear, S. F., Balkenhol, N., Fortin, M.-J., McRae, B. H. & Scribner, K. Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis. *Mol. Ecol.* **19**, 3576–3591 (2010).
589. Cushman, S. A., McKelvey, K. S., Hayden, J. & Schwartz, M. K. Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. *Am. Nat.* **168**, 486–499 (2006).
590. Legendre, P. & Anderson, M. J. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* **69**, 1–24 (1999).
591. Geffen, E., Anderson, M. J. & Wayne, R. K. Climate and habitat barriers to dispersal in the highly mobile grey wolf. *Mol. Ecol.* **13**, 2481–2490 (2004).

592. Mazé-Guilmo, E., Blanchet, S., McCoy, K. D. & Loot, G. Host dispersal as the driver of parasite genetic structure: a paradigm lost? *Ecol. Lett.* **19**, 336–347 (2016).
593. Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L. & Lasky, J. R. Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Mol. Ecol.* **25**, 104–120 (2016).
594. Frichot, E., Schoville, S. D., Bouchard, G. & François, O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* **30**, 1687–1699 (2013).
595. Leggett, H. C., Buckling, A., Long, G. H. & Boots, M. Generalism and the evolution of parasite virulence. *Trends Ecol. Evol.* **28**, 592–596 (2013).
596. Galpern, P., Peres-Neto, P. R., Polfus, J. & Manseau, M. MEMGENE: spatial pattern detection in genetic distance data. *Methods Ecol. Evol.* **5**, 1116–1120 (2014).
597. Kawecki, T. J. & Ebert, D. Conceptual issues in local adaptation. *Ecol. Lett.* **7**, 1225–1241 (2004).
598. Gandon, S. & Nuismer, S. L. Interactions between genetic drift, gene flow, and selection mosaics drive parasite local adaptation. *Am. Nat.* **173**, 212–224 (2009).
599. Ray, N. PATHMATRIX: a geographical information system tool to compute effective distances among samples. *Mol. Ecol. Notes* **5**, 177–180 (2005).
600. Shah, V. B. & McRae, B. H. CIRCUITSCAPE: a tool for landscape ecology. *Proceedings of the 7th Python in Science Conference*, 62–66 (2008).
601. Devillers, H., Lobry, J. R. & Menu, F. An agent-based model for predicting the prevalence of *Trypanosoma cruzi* I and II in their host and vector populations. *J. Theor. Biol.* **255**, 307–315 (2008).
602. Slimi, R., El Yacoubi, S., Dumonteil, E. & Gourbière, S. A cellular automata model for Chagas disease. *Appl. Math. Model.* **33**, 1072–1085 (2009).
603. Snyder, L. A. S. & Saunders, N. J. The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as 'virulence genes'. *BMC Genomics* **7**, 128 (2006).
604. Gottdenker, N. L., Chaves, L. F., Calzada, J. E., Saldaña, A. & Carroll, C. R. Host life history strategy, species diversity, and habitat influence *Trypanosoma cruzi* vector infection in changing landscapes. *PLoS Negl. Trop. Dis.* 6, e1884 (2012).
605. Villacís, A. G., Arcos-Terán, L. & Grijalva, M. J. Life cycle, feeding and defecation patterns of *Rhodnius ecuadoriensis* (Lent & León 1958) (Hemiptera: Reduviidae: Triatominae) under laboratory conditions. *Mem. Inst. Oswaldo Cruz* **103**, 690–695 (2008).

606. Bustamante, D. M., Monroy, M. C., Rodas, A. G., Juarez, J. A. & Malone, J. B. Environmental determinants of the distribution of Chagas disease vectors in southeastern Guatemala. *Geospat. Health* **1**, 199–211 (2007).
607. Hernández, J., Núñez, I., Bacigalupo, A. & Cattan, P. E. Modeling the spatial distribution of Chagas disease vectors using environmental variables and people´s knowledge. *Int. J. Health Geogr.* **12**, 29 (2013).
608. Foley, E. A. Population structure of the Chagas disease vector, *Triatoma infestans*, at the urban-rural interface. *Mol. Ecol.* **22**, 5162–5171 (2013).
609. Abad-Franch, F., Palomeque, F. S., Aguilar, H. M. & Miles, M. A. Field ecology of sylvatic *Rhodnius* populations (Heteroptera, Triatominae): risk factors for palm tree infestation in western Ecuador. *Trop. Med. Int. Health* **10**, 1258–1266 (2005).
610. Merrick, M. J., Koprowski, J. L. & Gwinn, R. N. *Sciurus stramineus* (Rodentia: Sciuridae). *Mamm. Species* **44**, 44–50 (2012).
611. Schweigmann, N. et al. Dispersal flight by *Triatoma infestans* in an arid area of Argentina. *Med. Vet. Entomol.* **2**, 401–404 (1988).
612. Schaub, G. A. Developmental time and mortality of larvae of *Triatoma infestans* infected with *Trypanosoma cruzi*. *Trans. R. Soc. Trop. Med. Hyg.* **82**, 94–96 (1988).
613. Castro, L. A. et al. Flight behavior and performance of *Rhodnius pallescens* (Hemiptera: Reduviidae) on a tethered flight mill. *J. Med. Entomol.* **51**, 1010–1018 (2014).
614. Biek, R., Henderson, J. C., Waller, L. A., Rupprecht, C. E. & Real, L. A. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proc. Natl. Acad. Sci. USA* **104**, 7993–7998 (2007).
615. Rioux Paquette, S., Talbot, B., Garant, D., Mainguy, J. & Pelletier, F. Modelling the dispersal of the two main hosts of the raccoon rabies variant in heterogeneous environments with landscape genetics. *Evol. Appl.* **7**, 734–749 (2014).
616. Cushman, S. A. Pushing the envelope in genetic analysis of species invasion. *Mol. Ecol.* **24**, 259–262 (2015).
617. Clarkson, C. S. et al. Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation. Nat. Commun. **5**, 42–48 (2014).
618. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28**, 659–669 (2013).
619. Nair, S. et al. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Mol. Biol. Evol.* **24**, 562–573 (2007).

620. Hall, M. D. & Ebert, D. The genetics of infectious disease susceptibility: has the evidence for epistasis been overestimated? *BMC Biol.* **11**, 79 (2013).
621. Mackinnon, M. J. et al. Environmental correlation analysis for genes associated with protection against malaria. *Mol. Biol. Evol.* **33**, 1188–1204 (2016).
622. Balkenhol, N., Waits, L. P. & Dezzani, R. J. Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography* **3**, 818–830 (2009).
623. Anderson, C. D. et al. Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Mol. Ecol.* **19**, 3565–3575 (2010).
624. Hoban, S., Bertorelle, G. & Gaggiotti, O. E. Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.* **13**, 110–122 (2012).
625. Richardson, J. L., Brady, S. P., Wang, I. J. & Spear, S. F. Navigating the pitfalls and promise of landscape genetics. *Mol. Ecol.* **25**, 849–863 (2016).
626. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
627. Bethony, J. et al. Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. *Lancet* **367**, 1521–1532 (2006).
628. Anthony, T. J. et al. Fragmented population structure of *Plasmodium falciparum* in a region of declining endemicity. *J. Infect. Dis.* **191**, 1558–1564 (2005).
629. Tesson, S. V. M. et al. Integrating microorganism and macroorganism dispersal: modes, techniques and challenges with particular focus on co-dispersal. *Écoscience* **22**, 109–124 (2016).
630. Behnke, M. S. et al. Virulence differences in *Toxoplasma* mediated by amplification of a family of polymorphic pseudokinases. *Proc. Natl. Acad. Sci. USA* **108**, 9631–9636 (2011).
631. Miller, M. R., White, A. & Boots, M. The evolution of host resistance: tolerance and control as distinct strategies. *J. Theor. Biol.* **236**, 198–207 (2005).
632. Bierne, N., Welch, J., Loire, E., Bonhomme, F. & David, P. The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol. Ecol.* **20**, 2044–2072 (2011).
633. Allendorf, F. W., Hohenlohe, P. A. & Luikart, G. Genomics and the future of conservation genetics. *Nat. Rev. Genet.* **11**, 697–709 (2010).
634. Gürtler, R. E. Sustainability of vector control strategies in the Gran Chaco region: current challenges and possible approaches. *Mem. Inst. Oswaldo Cruz.* **104**, 52–59 (2009).

635. Kaye, P. & Scott, P. Leishmaniasis: complexity at the host–pathogen interface. *Nat. Rev. Microbiol.* **9**, 604–615 (2011).
636. Ramírez, J. D. et al. Taxonomy, diversity, temporal and geographical distribution of cutaneous leishmaniasis in Colombia: a retrospective study. *Sci. Rep.* **6**, 1–10 (2016).
637. Carvalho, B. M., Rangel, E. F., Ready, P. D. & Vale, M. M. Ecological niche modelling predicts southward expansion of *Lutzomyia* (*Nyssomyia*) *flaviscutellata* (Diptera: Psychodidae: Phlebotominae), vector of *Leishmania* (*Leishmania*) *amazonensis* in South America, under climate change. *PLoS One* **10**, e0143282 (2015).
638. Wasserman, T. N., Cushman, S. A., Shirk, A. S., Landguth, E. L. & Littell, J. S. Simulating the effects of climate change on population connectivity of American marten (*Martes americana*) in the northern Rocky Mountains, USA. *Landscape Ecol.* **27**, 211–225 (2012).
639. World malaria report. *WHO* (2015).
640. Crawford, J. E. et al. Reticulate speciation and barriers to introgression in the *Anopheles gambiae* species complex. *Genome Biol. Evol.* **7**, 3116–3131 (2015).
641. Lanzaro, G. & Tripet, F. Gene flow among populations of *Anopheles gambiae*: a critical review; in Ecological aspects for application of genetically modified mosquitoes (eds. Takken, W. & Scott, T. W.) 109–132. *Kluwer Acad. Publ.* (2004).
642. Marrelli, M. T., Moreira, C. K., Kelly, D., Alphey, L. & Jacobs-Lorena, M. Mosquito transgenesis: what is the fitness cost? *Trends Parasitol.* **22**, 197–202 (2006).
643. Gabrieli, P., Smidler, A. & Catteruccia, F. Engineering the control of mosquito-borne infectious diseases. *Genome Biol.* **15**, 535 (2014).
644. Murdock, C. C., Paaijmans, K. P., Cox-Foster, D., Read, A. F. & Thomas, M. B. Rethinking vector immunology: the role of environmental temperature in shaping resistance. *Nat. Rev. Microbiol.* **10**, 869–876 (2012).
645. Hand, B. K., Lowe, W. H., Kovach, R. P., Muhlfeld, C. C. & Luikart, G. Landscape community genomics: understanding eco-evolutionary processes in complex environments. *Trends Ecol. Evol.* **30**, 161–168 (2015).
646. Hernández-Castro, L. E. et al. 2b-RAD genotyping for population genomic studies of Chagas disease vectors: *Rhodnius ecuadoriensis* in Ecuador. *PLoS Negl. Trop. Dis.* **11**, e0005710 (2017).
647. Howes, R. E. et al. Global epidemiology of *Plasmodium vivax*. *Am. J. Trop. Med. Hyg.* **95**, 15–34 (2016).
648. Gibson, W. Liaisons dangereuses: sexual recombination among pathogenic trypanosomes. *Res. Microbiol.* **166**, 459–466 (2015).

649. Macaulay, I. C. & Voet, T. Single cell genomics: advances and future perspectives. *PLoS Genet.* **10**, e1004126 (2014).
650. Gibson, W. & Stevens, J. Genetic exchange in the Trypanosomatidae. *Adv. Parasitol.* **43**, 1–46 (1999).
651. Johnson, A. The biology of mating in *Candida albicans*. *Nat. Rev. Microbiol.* **1**, 106–116 (2003).
652. Gibson, W. & Peacock, L. Fluorescent proteins reveal what trypanosomes get up to inside the tsetse fly. *Parasit. Vectors* **12**, 6 (2019).
653. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
654. Costa, F. C. et al. Expanding the toolbox for *Trypanosoma cruzi*: a parasite line incorporating a bioluminescence-fluorescence dual reporter and streamlined CRISPR/Cas9 functionality for rapid *in vivo* localisation and phenotyping. *PLoS Negl. Trop. Dis.* **12**, e0006388 (2018).
655. Alves, C. L. et al. The recombinase Rad51 plays a key role in events of genetic exchange in *Trypanosoma cruzi*. *Sci. Rep.* **8**, 13335 (2018).
656. Schurko, A. M. & Logsdon, J. M. Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex. *BioEssays* **30**, 579–589 (2008).
657. Malik, S.-B., Ramesh, M. A., Hulstrand, A. M. & Logsdon, J. M. Protist homologs of the meiotic Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and lineage-specific loss. *Mol. Biol. Evol.* **24**, 2827–2841 (2007).
658. Peacock, L. et al. Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly. *Proc. Natl. Acad. Sci. USA* **108**, 3671–3676 (2011).
659. Inbar, E. et al. The transcriptome of *Leishmania major* developmental stages in their natural sand fly vector. *mBio* **8**, e00029 (2017).
660. Bishop, D. K. Rad51, the lead in mitotic recombinational DNA repair, plays a supporting role in budding yeast meiosis. *Cell Cycle* **11**, 4105–4106 (2012).
661. Morrison, L. J. et al. A major genetic locus in *Trypanosoma brucei* is a determinant of host pathology. *PLoS Negl. Trop. Dis.* **3**, e557 (2009).
662. Berriman, M. et al. The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**, 416–422 (2005).
663. Vaikkinen, H. J. Finding a gene for virulence in *Trypanosoma brucei* (PhD dissertation). *University of Glasgow* (2016).
664. Van den Broeck, F. et al. Ecological divergence and hybridization of neotropical *Leishmania* parasites. *bioRxiv* 824912 (2019), doi:10.1101/824912.

665. Grillet, M. E. et al. Venezuela's humanitarian crisis, resurgence of vector-borne diseases, and implications for spillover in the region. *Lancet Infect. Dis.* **19**, e149–e161 (2019).
666. Oliveira, T. C. de et al. Genome-wide diversity and differentiation in New World populations of the human malaria parasite *Plasmodium vivax*. *PLoS Negl. Trop. Dis.* **11**, e0005824 (2017).
667. Singh, V., Gupta, P. & Pande, V. Revisiting the multigene families: *Plasmodium* var and vir genes. *J Vector Borne Dis* 7 (2014).
668. Day, K. P. et al. Evidence of strain structure in *Plasmodium falciparum* var gene repertoires in children from Gabon, West Africa. *Proc. Natl. Acad. Sci. USA* **114**, E4103–E4111 (2017).
669. Tessema, S. K. et al. Phylogeography of var gene repertoires reveals fine-scale geospatial clustering of *Plasmodium falciparum* populations in a highly endemic area. *Mol. Ecol.* **24**, 484–497 (2015).
670. Xu, Z., Kaplan, N. L. & Taylor, J. A. TAGster: efficient selection of LD tag SNPs in single or multiple populations. *Bioinformatics* **23**, 3254–3255 (2007).
671. Oliveira, A. M. et al. Occurrence of *Lutzomyia longipalpis* and human and canine cases of visceral leishmaniasis and evaluation of their expansion in the northwest region of the state of São Paulo, Brazil. *Rev. Soc. Bras. Med. Trop.* **49**, 41–50 (2016).
672. Andrade-Filho, J. D. et al. Occurrence and probability maps of *Lutzomyia longipalpis* and *Lutzomyia cruzi* (Diptera: Psychodidae: Phlebotominae) in Brazil. *J. Med. Entomol.* **54**, 1430–1434 (2017).
673. Pech-May, A. et al. Genetic variation and phylogeography of the *Triatoma dimidiata* complex evidence a potential center of origin and recent divergence of haplogroups having differential *Trypanosoma cruzi* and DTU infections. *PLoS Negl. Trop. Dis.* **13**, e0007044 (2019).
674. Peterman, W. E. ResistanceGA: an R package for the optimization of resistance surfaces using genetic algorithms. *Methods Ecol. Evol.* **9**, 1638–1647 (2018).
675. Nieberding, C. M. & Olivieri, I. Parasites: proxies for host genealogy and ecology? *Trends Ecol. Evol.* **22**, 156–165 (2007).
676. Petkau, A. et al. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microb. Genomics* **3**, e000116 (2017).
677. Bogaerts, B. et al. Validation of a bioinformatics workflow for routine analysis of whole-genome sequencing data and related challenges for pathogen typing in a European national reference center: *Neisseria meningitidis* as a proof-of-concept. *Front. Microbiol.* **10**, 362 (2019).