Title: Multiple targets of balancing selection in *Leishmania donovani* complex parasites.

Authors

Cooper Alastair Grace[1]

Sarah Forrester[1]

Vladimir Costa Silva[2]

Aleksander Aare[1]

Hannah Kilford[1]

Yen Peng Chew[1,3]

Sally James[1]

Dorcas L. Costa[2]

Jeremy C. Mottram[1]

Carlos C. H. N. Costa[2]

Daniel C. Jeffares[1]

Affiliations

1. York Biomedical Research Institute and Department of Biology, University of York, York YO10 5DD, United Kingdom
2. Instituto de Doenças do Sertão, Instituto de Doenças Tropicais Natan Portella, Centro de Ciências da Saúde da Universidade Federal do Piauí, Rua Artur de Vasconcelos 151-Sul, 64001-450 Teresina-PI, Brazil
3. Institute of Molecular Plant Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

Abstract

The *Leishmania donovani* species complex are the causative agents of visceral leishmaniasis, which cause 20-40,000 fatalities a year. Here, we conduct a screen for balancing selection in this specie complex. We sequence 93 isolates of *L. infantum* from Brazil and used 387

publicly-available *L. donovani* and *L. infantum* genomes, to describe the global diversity of this species complex. We identify five genetically-distinct populations that are sufficiently represented by genomic data to search for signatures of selection. We show that multiple metrics identify genes with robust signatures of balancing selection. We produce a curated set of 19 genes with robust signatures, including zeta toxin, nodulin-like and flagellum attachment proteins. Candidate genes were generally not shared between populations, consistent with divergent rather than long-term balancing selection in these species. This study highlights the extent of genetic divergence between *L. donovani complex* parasites and provides candidate genes for further study.

Main text

Introduction

Intracellular *Leishmania* parasites cause the neglected infectious disease leishmaniasis in over 80 countries. Visceral leishmaniasis (VL) is the most severe form of the disease, caused by *Leishmania donovani* and *Leishmania infantum.* Yearly cases of VL are estimated at a minimum of 50,000, with a fatality of $\geq$ 95% if untreated, and occur primarily in the Indian subcontinent (ISC), Bangladesh, Sudan, South Sudan, Ethiopia and Brazil (World Health Organisation, 2020). After transmission by sand flies, *Leishmania* promastigotes are taken up by macrophages, and develop into amastigotes which proliferate. These processes require evasion and active modulation of mammalian host or sand fly vector immune responses (Atayde *et al.*, 2016; Dong, Filho and Olivier, 2019). *Leishmania* species contain genomes that are primarily diploid and sexually-recombining. Amongst their unusual features are the use of constitutively-transcribed polycistronic genes and supernumerary chromosomes with unstable polidy (Dumetz *et al.*, 2017).

In *Plasmodium* parasites that have been studied more intensively, some host-interacting genes maintain high genetic diversity due to balancing selection (BS) within populations (Ochola-Oyier *et al.*, 2019), in some cases with a bias towards polymorphic residues to be exposed on surface epitopes (Guy *et al.*, 2018). Given the competitive interaction between

Leishmania cells and host immune cells, balancing selection may also be present in *Leishmania*. Other drivers of balancing selection, such as heterozygote advantage (overdominance) or alleles that confer fitness differentially in the sand fly vector and the mammal host are also possible, that are expected to generate similar genetic signals (Charlesworth, 2006). In all these scenarios, genomic signatures of BS can highlight genes that are important for transmission, host immune evasion or ecological adaptation.

Results

L. donovani complex genome data and population structure

To date there have been no published studies of BS in *Leishmania* species. To identify the most likely targets of BS we examined genome-scale polymorphism data from *L. donovani* species complex, covering the main global foci of East Africa, the Indian subcontinent (ISC), Brazil and Europe. We utilised 229 *L. donovani* isolates from the ISC (Imamura *et al.*, 2016), 43 *L. donovani* isolates from Ethiopia (Zackay *et al.*, 2018), 25 *L. infantum* isolates from Brazil (Carnielli *et al.*, 2018), and 87 *L. donovani* isolates from a variety of locations including Sudan (14 strains), France (6) and Israel (10) (Franssen *et al.*, 2020). Additionally, we sequenced 93 *L. infantum* isolates from Piauí state, Brazil (**Supplementary Figure 1**, **Supplementary Table 9**). This produced a data set of 477 sequenced isolates from the *L. donovani* complex. To detect genetic variants in these genomes we mapped reads from all isolates to the *L. donovani* BPK282A1 reference genome, and applied variant calling methods and filtering to identify single-nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms (indels). We validated the variant calling and filtering using the *de novo* assembly of *L. infantum* isolate MHOM/BR/06/MA01A (see Methods). This resulted in 339,367 SNPs and 14,383 indels.

We used the ADMIXTURE clustering tool (Alexander and Lange, 2011) to assign isolates to populations. This analysis indicated that this collection can be clustered into between 8 and 11 populations (**Figure 1**, and **Supplementary Figure 2-3**). Principal component analysis and phylogenetic analysis were consistent with the majority of the sequenced isolates we examined being members of five relatively well-sampled populations (**Figure 1**) where

isolates could be assigned with ≥ 99% confidence by ADMIXTURE. This included two populations from the Indian Sub-Continent (ISC1, ISC2), two from East Africa (EA1, EA2) and a Brazil-Mediterranean population (Bra-Med). The remaining isolates were assigned to populations of <6 isolates (*n*=44), which we did not analyse further.

Although hybrids have been shown to occur between *L. donovani* complex populations in both East Africa and Turkey (Rogers *et al.*, 2014; Cotton *et al.*, 2020), large proportions of the genetic variation within these five populations is generally derived from independent ancestry since fixation index (F_{ST}) values range from 0.27 to 0.90 (**Supplementary Table 7**). We do not include *L. donovani* complex hybrids from Turkey (Rogers *et al.*, 2014) in our analysis, because hybrid populations may contain balanced alleles that give an appearance of BS. This, and the under-sampling of VL-endemic regions between Europe and India, render our data unsuited to studying the true extent of global gene flow in this species, so we do not analyse this further here.

Phylogenetic analysis provides some qualitative insight to the history of these species (**Figure 1**, **Panel C**). The basal position of EA2 and early-branching position of EA1 support the relative age of these populations in East Africa, as does the high genetic diversity in this region and genetic distance between these populations, consistent with previous studies (Gelanew *et al.*, 2010, 2014; Ferreira *et al.*, 2012; Teixeira *et al.*, 2017; Zackay *et al.*, 2018; Cotton *et al.*, 2020; Franssen *et al.*, 2020). The high nucleotide diversity of EA1 (**Table 1**) is reflected in the branch lengths in this clade of the phylogeny. In contrast, the smaller Indian subcontinent population we identify as ISC1 here (equivalent to the ISC5 group identified by (Imamura *et al.*, 2016)) produced short terminal branches in the phylogeny, and lower genetic diversity (**Table 1**, **Figure 2**), consistent with previous genomic analyses indicating that is an emergent population (Imamura *et al.*, 2016) and with epidemiological evidence indicating that it arose in the 1970s after the malaria elimination programme (Dye and Wolpert, 1988; Bhattacharya *et al.*, 2006; Thakur, 2007; Muniaraj, 2014; Dhiman and Yadav, 2016). The 93 Brazilian isolates we examined here cluster within isolates originating from Mediterranean countries (**Figure 1**, **Panel D**), consistent with a relatively recent European origin for *L. infantum* in Brazil (<400 years; (Kuhls *et al.*, 2011). Short branches in the Brazilian clade, low genetic diversity and excess of rare alleles (**Figure 2**) are all consistent with a relatively

recent population bottleneck and an expanding population. In contrast, both East African and populations and the main population from the Indian subcontinent (ISC2) show higher genetic diversity, and are likely to have been endemic in larger populations for a longer period.

As well as the very low diversity, the Brazil-Mediterranean and smaller East African (EA2) population contain more indel polymorphisms (**Table 1**), with a SNP:indel ratio of 6:1 and 5:1, respectively, compared to EA1 (11:1), ISC1 (10:1) and ISC2 (9:1), which are more typical of eukaryote polymorphisms where SNPs typically outnumber indels (Mullaney *et al.*, 2010; Jeffares *et al.*, 2015). Extensive variant-calling quality control showed that this excess of indels is unlikely to be artefactual (**Supplementary Figure 5**). It is possible that this is due to the accumulation of weakly-deleterious indel alleles when the Brazil population was established from European *L. infantum* populations (Boité *et al.*, 2019).

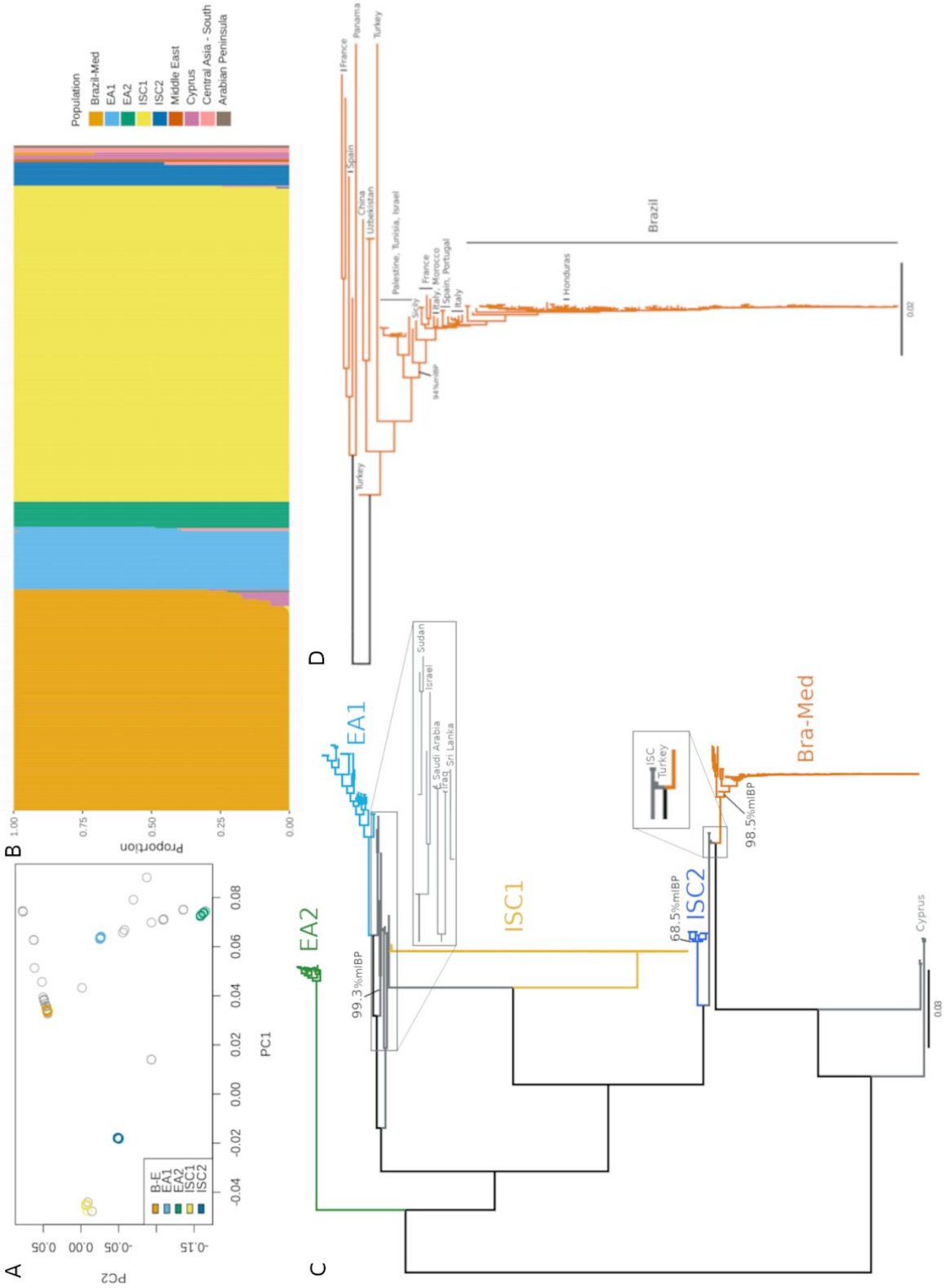


Figure 1: Population structure of the *L. donovani* complex. Panel A: Principal component analysis (PCA); **Panel B**: ADMIXTURE analysis indicated between 8-11 populations, here *K*=9. Cross validation error values are available as **Supplementary Figure 2**. ADMIXTURE

plots for *K*=8, 10 and 11 populations available as **Supplementary Figure 3**. Isolates in grey are those not confidently assigned to one of the five major populations (Brazil-Med, EA1, EA2, ISC1 and ISC2) by ADMIXTURE. **Panel C**: Maximum likelihood phylogeny, based upon a SNP alignment of 477 sequences with 283378 variable sites. All visible branches are maximally supported (100% mlBP) unless indicated. The tree is midpoint rooted, with the unrooted tree available as **Supplementary Figure 4**. The scale bar represents the number of nucleotide changes per site. Country names in grey indicate origins of the minor populations. **Panel D**: Brazil-Mediterranean tree, based upon a SNP alignment of 158 sequences with 81018 variable sites, midpoint rooted. Scale bar and support are as in Panel A. A single sample isolated in China (Franssen *et al.*, 2020), indicative of worldwide movement, is the only demographic exception in the Brazil-Mediterranean sample collection. Data and tree files are available in supplementary documents.

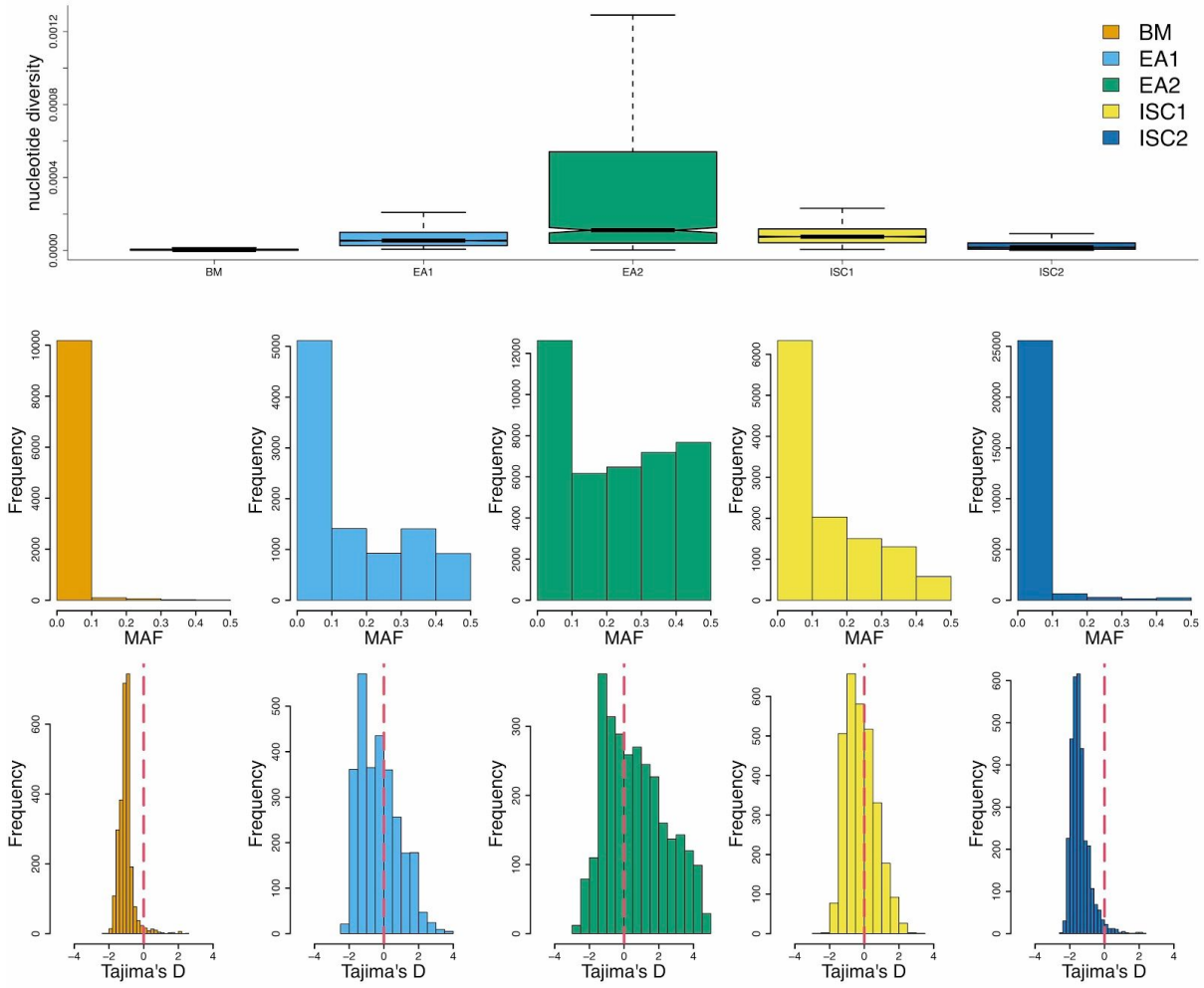


Figure 2. Population genetic statistics. Upper panel: nucleotide diversity (π); middle panel: Minor allele frequency (MAF); Lower panel: Tajima's *D*.

Table 1. Population statistics for *L. donovani* complex populations.

Population	Source	Isolates	No. of 'pure'	SNPs	Indels	π x10^{-6}	Tajima *D*	Mean MAF
EA1	East Africa	41	41	3033	705	424	0.70	0.23
EA2	East Africa	18	18	970	575	87	-0.20	0.14
ISC1	India	225	211	2689	551	6.3	-1.04	0.009
ISC2	India	15	15	3103	716	84	-0.23	0.17
BM	Brazil, Med.	133	127	8886	1578	28	-1.42	0.02

Screening for balancing selection in multiple populations

Genetic diversity summary statistics vary considerably between populations (**Figure 2**, **Table 1)**, indicating that these populations have been through varied demographic histories. To select groups of strains that will approximate panmictic populations as close as is practical, we used the ADMIXTURE analysis to identify isolates that were confidently assigned to one population (*n*=433), rather than being inter-population hybrids. This selection resulted in 59 isolates from East Africa (two populations of 41 and 18), 226 from the ISC (two populations of 211 and 15) and 127 from Brazil-Mediterranean (**Table 1**).

To identify the most likely candidates of BS in these populations we applied the *NCD2* test (Bitarello *et al.*, 2018) and *Betascan2* test (Siewert and Voight, 2020) in 10 kb windows to each of the five populations. We selected genomic windows that were in the 99th percentile of either test as likely target regions. We make the assumption that protein-coding genes will drive the strongest signals of BS. To identify the genes within these windows that are likely targets, we calculated nucleotide diversity (π) and Tajima's *D* (Tajima, 1989) for each gene, and selected genes in the 90th percentile of either statistic as the plausible targets. We combine the *NCD2/Betascan2* statistics with the π/Tajima's *D* gene-based statistics by selecting genes that were outliers in both categories (*NCD2 or Betascan2* and π or Tajima's *D*). This intersection identified 33 genes (**Supplementary Table 5**). Finally, we manually vetted these to remove 'hitchhiker', genes whose high diversity was likely due to their close proximity to a BS 'driver' gene. We also removed genes with suspicious read coverage, since gene duplications produce strong artifactual signals of BS (**Supplementary Figure 15**). These procedures identified 19 vetted candidate genes (**Table 2,** justification for vetting is genes in **Supplementary Table 5**). In the EA1 population, where the most candidate genes were discovered, nucleotide diversity is 39-fold higher than the genome-wide median. Diversity is also elevated in genome regions surruonding BS target genes, and remains significantly higher up to 130 kb from the BS targets in this population (**Figure O**).

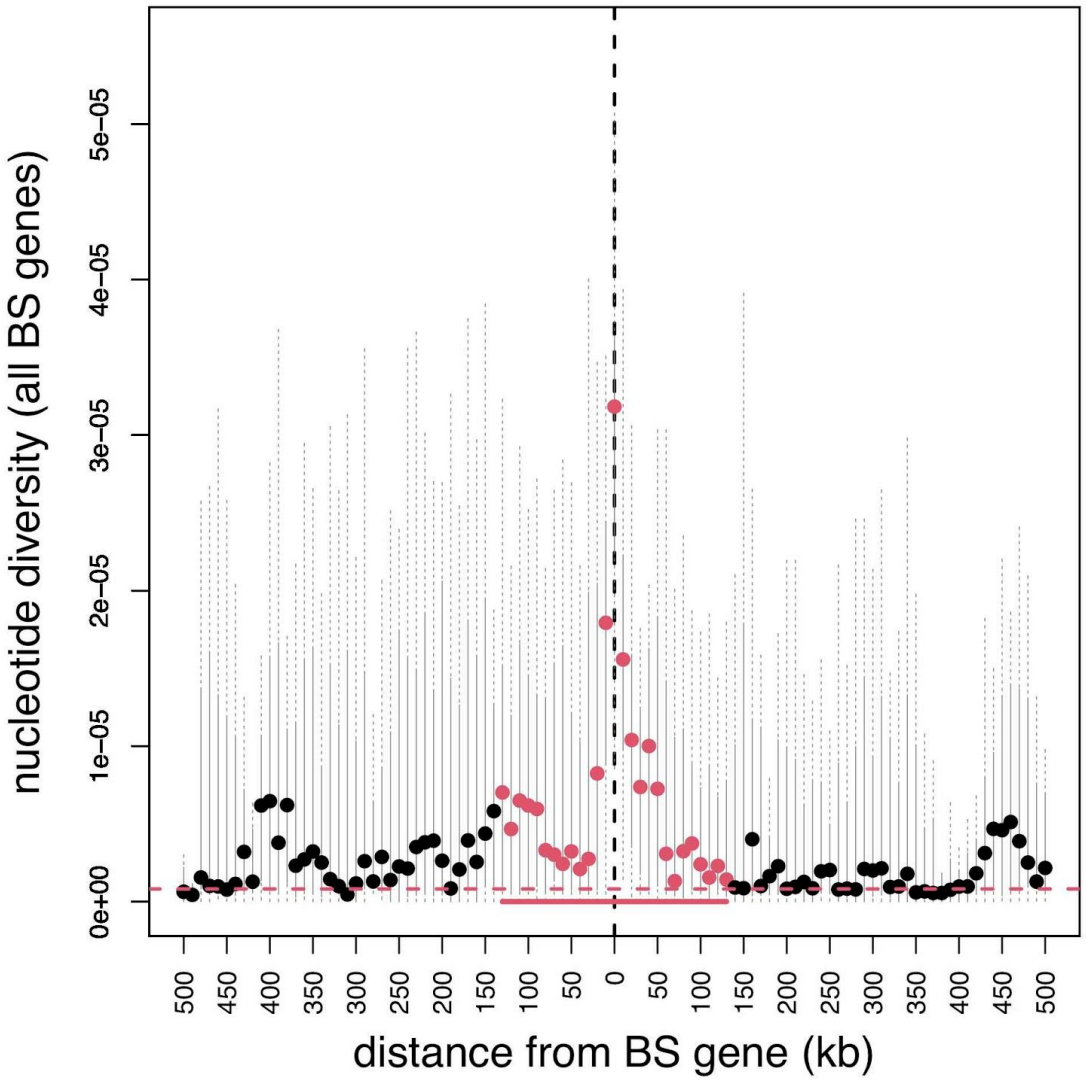


Figure O. Diversity is significantly elevated in BS target regions. For the 24 BS candidate genes discovered in the EA1 population (including those that did not pass manual vetting), we show the distribution of nucleotide diversity (grey vertical lines) in the 10 kb closest to the gene, and every 10 kb up and downstream to 500 kb. Median diversity is indicated with filled circles, red where the diversity is significantly higher than the genome-wide distribution and black otherwise. Diversity is higher in BS candidate genes at each distance up to 130 kb from the target gene (one-sided Mann-Whitney test $P = 5 \times 10^{-4}$), indicated by the horizontal red bar. The dashed horizontal line shows the median nucleotide diversity for the genome in the EA1 population.

In general, we do not find convincing evidence for long-term BS (between populations) in these candidate genes. While genomic windows in the 130 kb around these target regions are enriched for SNPs that co-segregate between populations (**Supplementary Figure 16**), an excess of co-segregating SNPs would be expected from any high allele frequency set due to occasional migration alone, without BS. Also, the 24 genes that are BS candidates in the EA1 population do not show significantly elevated Tajma's *D* values in any other population (**Supplementary Figure 19**), or higher *Betacan* scores (data not shown).

Candidate gene	**Description**	**Population**	**Tajima's *D***	**Variants (nonsyn/ synon)**	**Tests NCD2/Beta**
LdBPK_161760.1	FLAM3, flagellum attachment protein in *L. mexicana*, see (Sunter *et al.*, 2019)	ISC2	3.1	12/4	NCD2
LdBPK_341740.1	Zeta toxin protein 1, conserved in Trypanosomes, see (Srivastava *et al.*, 2019)	EA2	3.3	35/18	both
LdBPK_363870.1	Mitogen activated kinase-like protein, conserved in Trypanosomes	EA1	3.9	10/9	both
LdBPK_291600.1	Nodulin-like, conserved in Trypanosomes	ISC2	3.2	8/9	both
LdBPK_170210.1	Unknown function, conserved in *Leishmania*	EA1	3.0	6/9	Beta
LdBPK_280190.1	Unknown function, conserved in *Leishmania*, contains helix domains	EA1	2.9	12/3	Beta
LdBPK_282030.1	p21-C-terminal region-binding protein, conserved in Trypanosomes	EA1	1.9	9/6	both

LdBPK_301540.1	Rad17 cell cycle checkpoint clamp protein (hypothetical protein on TriTrypDB), conserved in Trypanosomes. involved in chromatin binding and DNA repair, see (Nunes *et al.*, 2011)	EA1	4.0	8/14	both
LdBPK_311120.1	emp24/gp25L/p24/GOLD family, conserved in Trypanosomes, involved in golgi vesicle transportation	EA1	2.9	4/2	both
LdBPK_311170.1	Unknown function, conserved in *Leishmania*, adenylate cyclase regulatory protein-like	ISC2	3.1	8/2	both
LdBPK_312260.1	Unknown function, conserved in *Leishmania*	EA1	4.3	20/6	both
LdBPK_312550.1	2Fe-2S iron-sulfur cluster binding domain, only conserved in *L. donovani* and *L. infantum*, cofactor, implicated in redox metabolism e.g. (Kumar, Chauhan and Singh, 2019)	EA1	4.3	15/4	Beta
LdBPK_330840.1	Nuclear LIM interactor-interacting (NLI) factor-like phosphatase, conserved in *Leishmania*. no literature	EA1	4.8	27/17	both
LdBPK_350960.1	Unknown function, conserved in Trypanosomes	EA1	2.6	3/4	both
LdBPK_361900.1	Ras-like small GTPase, conserved in *Leishmania*	EA1	3.7	6/3	NCD2
LdBPK_363830.1	Unknown function - shares >40% similarity with	EA1	3.8	6/8	both

	tectonic/cilia protein, conserved across Trypanosomes - see (Dean *et al.*, 2016)				
LdBPK_366210.1	Unknown function, conserved in *Leishmania*	EA1	4.1	6/8	both
LdBPK_300960.1	Hypothetical protein, conserved in *Leishmania*	EA1	4.2	10/10	Beta
LdBPK_312990.1	Clathrin and VPS/Zinc finger RING-type	EA1	4.3	14/11	both

Table 2: Highly enriched candidate list of genes under balancing selection. Details of the method are described in **Supplementary Figure 7** and **Supplementary Text 2**. All comparisons of tests for each population are available as **Supplementary Figures 8-12**, respectively. Where protein function is 'unknown' on TriTrypDB, we subjected each protein to BLASTp searches to obtain homology to other known proteins and ascertain conservation across trypanosomes.

Genes under balancing selection in the *L. donovani* complex

Our manual vetting of BS candidates retained 19 genes (**Table 2**), of which 15 were discovered in the EA1 population. We did not discover any reliable candidates in the two populations that appear to be expanding following a bottleneck (BM and ISC1), nor in the population consisting entirely of *L. infantum* isolates (BM). **Figure 3** illustrates the variety of robust genetic signatures that implicate four genes. All vetted genes contain similarly robust signatures (**Supplementary Figure 18**).

Several genes caught our attention as interesting targets of selection. LdBPK_291600.1 encodes a transmembrane protein containing a nodulin-like domain. Such proteins have been implicated in membrane transport and iron homeostasis (Laranjeira-Silva *et al.*, 2018) in *Leishmania*.

The zeta toxin domain protein (LdBPK_341740.1, **Figure 3**) is indicated as a BS target by both *Betascan2* and *NCD2* metrics in the East African population EA2. The gene also has high nucleotide diversity in EA1 (**Supplementary Table 2**). The zeta domain is positioned at 744-861aa, with two non-synonymous variants resulting in the changes Leu747Thr and Ser752Phe, respectively, from the reference genome. The zeta toxin is part of the Type-II toxin-antitoxin (TA) module identified in prokaryotes, with homologues only recently discovered in *Leishmania* (Srivastava *et al.*, 2019). The toxin component of the TA module acts against cellular processes such as translation, and is neutralised by the antitoxin component in favourable conditions. Sharing similar functional domains and activity with the *E. coli* homologue (Srivastava *et al.*, 2019), *L. donovani* zeta toxin may therefore also be implicated in stress response and/or virulence (Rocker and Meinhart, 2015).

The *FLAM3* gene (LdBPK_161760.1) encodes a flagellum attachment zone (FAZ) protein essential for host interaction (Sunter *et al.*, 2015, 2019; An and Li, 2018). Our strict criteria for BS was only indicated in the Indian subcontinent population ISC2, but this gene has high nucleotide diversity in four of the five populations (**Supplementary Table 2**). The *FLAM3* protein contains a clustered mitochondria (CLU) domain and a domain of repeats (Sunter *et al.*, 2015). The majority of variants in ISC2 occur between these domains, with none falling within the CLU domain (**Supplementary Table 8**).

LdBPK_363870.1 (**Figure 3**) encodes a protein with similarity to mitogen-activated protein (MAP) kinases, a large family of enzymes implicated in antimony drug resistance (Ashutosh *et al.*, 2012; Raj *et al.*, 2019). While candidacy was indicated by BS tests only in the East African population EA1, nucleotide diversity is also high in ISC1,2 and EA2 populations (**Supplementary Table 2**). The candidate is an orthologue of *L. mexicana* LmxM.36.3680, and therefore an endomembrane member of the STE family (Baker *et al.*, 2020).

Two candidate genes, LdBPK_311120.1 and LdBPK_361900.1, encoding a member of the emp24/gp25L/p24/GOLD family and a ras-like GTPase, respectively, are implicated in protein trafficking and export (Anantharaman and Aravind, 2002; Field, 2005). Other than this commonality, we do not observe any strong signals of functional enrichment in our 19 BS targets. There were no significant enrichments for gene ontology containing more than

one gene and no significant enrichment for proteins with transmembrane domains or predicted signal peptides, or any enrichment in particular regions of the genome.

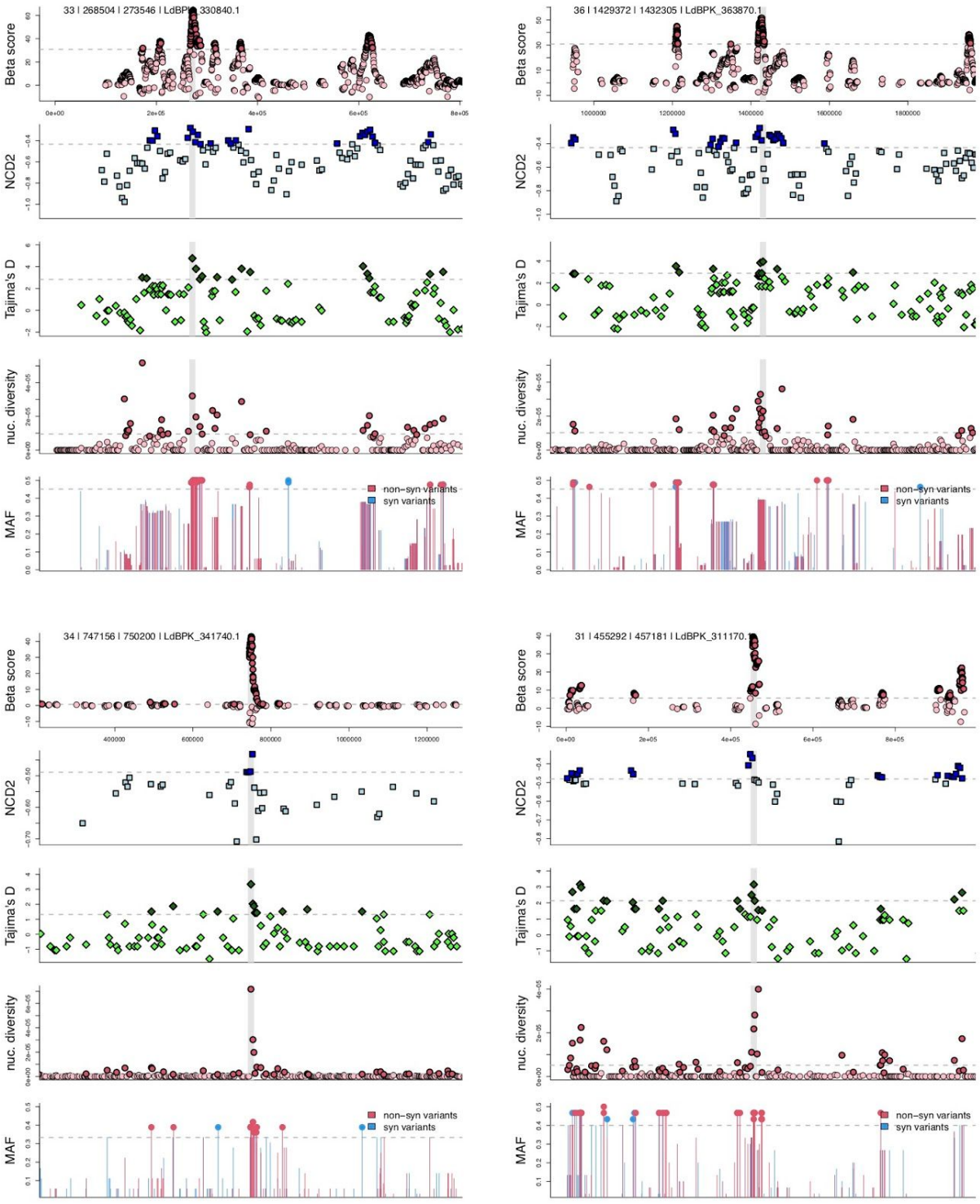


Figure 3. Candidate genes show multiple genetic signatures of balancing selection. Four each of four genes we show *Betascan2*, *NCD2*, Tajima's *D*, nucleotide diversity (π), and minor allele frequency (MAF) in a 500 kb window around the candidate gene. The location of the candidate gene is indicated by a vertical grey bar. The population-specific 90th percentile for each metric is shown as a horizontal dashed line, scores that are above this are drawn in darker shades, or plotted with a filled dot for MAF. Panel titles indicate the chromosome, gene start and end coordinates, and gene ID. Genes and populations where BS detected are; top left NLI interacting factor-like phosphatase-like phosphatase LdBPK_330840.1 (EA1), top right mitogen activated kinase-like protein LdBPK_363870.1 (EA1), bottom left putative Zeta toxin LdBPK_341740.1 (EA2), bottom right hypothetical protein LdBPK_311170.1 (ISC2). Similar plots are shown for all candidate genes in **Supplementary Figure 18**.

Discussion

Here we sequence 93 strains of *L. infantum* from Brazil, contributing to a worldwide collection of *L. donovani* complex isolates along with previous analyses (Imamura *et al.*, 2016; Carnielli *et al.*, 2018; Zackay *et al.*, 2018; Franssen *et al.*, 2020). Our analysis is consistent with these previous studies showing that the *L. infantum* population in Brazil contains very little genetic diversity (Carvalho *et al.*, 2020; Schwabl *et al.*, 2021).

This is the first study of balancing selection in *Leishmania* species. Relatively few studies have attempted balancing selection screenings in parasites (reviewed by Weedall and Conway, 2010). Our study uses the *Betacan2* (Siewert and Voight, 2020) and *NCD2* (Bitarello et al., 2018) metrics that have been developed recently and have been shown outperform classic metrics under models of BS, such as Tajima's *D* (Tajima, 1989). These analytic tools, and the use of multiple populations should produce an analysis that is at least as reliable as previous screens for balancing selection searches within parasites such as *Plasmodium* (Tetteh *et al.*, 2009, 6 genes; Amambua-Ngwa *et al.*, 2012, 25 genes; Nygaard *et al.*, 2010, 19 genes). Our candidate list of 19 genes is of a similar size, but differences in populations, ploidy, and methods preclude a direct comparison.

Our analysis suggests that the performance of *Betascan2* and *NCD2* BS tests are dependent on population demography and history, despite claims that the tests are robust to demography (Siewert and Voight, 2017, 2020; Bitarello *et al.*, 2018; Stern and Lee, 2020). This is likely due to our pragmatic screening criteria, that required >5 variants per genomic window (*Betascan2, NCD2*) and >5 variants per gene (Tajima's D and π metrics). The low-diversity Brazil and ISC1 population contain far fewer regions that satisfy these criteria. We expect this is because the population bottlenecks experienced by these populations has reduced genetic diversity, ablating strong BS signatures. To our knowledge, this has not been reported in other species. Only two populations (EA1 and ISC2) appear to be population-sTable 1nd panmictic, with sufficient nucleotide diversity to identify BS candidates effectively. Although model-based methods to detect BS in multiple populations are available (DeGiorgio, Lohmueller and Nielsen, 2014; Cheng and DeGiorgio, 2019, 2020), we chose not to employ these because we do not believe that *L. donovani* complex populations are sufficiently understood to model meaningfully at this stage. Our approach was to divide the samples into multiple populations, with the expectation that this would alleviate some of the issues resulting from population structure.

The candidate genes under balancing selection that we have identified possess varying functions within the parasite. The flagellum attachment gene *FLAM3* is a striking candidate in view of the importance of the encoded protein in parasite cellular structure, proliferation and differentiation (Sunter *et al.*, 2019). Furthermore, candidate genes LdBPK_311120.1 and LdBPK_361900.1, which encode a member of the emp24/gp25L/p24/GOLD family and a ras-like GTPase, respectively, may influence trafficking of virulence factors, and subsequently interaction with their host. Given the lack of detailed studies of these candidates in the *L. donovani* species complex, studies of the cell biology of these proteins will be useful next steps.

It is possible that balancing selection of single-copy genes is not the most important mechanism that maintains diversity within the *L. donovani* complex, or protozoan parasites generally. The effects of multi-copy gene families encoding RIFIN, STEVOR and PfEMP1 variant surface antigens in pathogenesis of *Plasmodium* parasites are well-described, for example (Wahlgren, Goel and Akhouri, 2017). These genes are typically removed from BS

screens, because multi-copy genes will produce artefactual signals of BS when analysed with current bioinformatics methods (**Supplementary Figure 17**). While variants in duplicated regions or multiple copies of genes may allow the parasite to maintain diversity, it is an open question whether this diversity is maintained by neutral processes or balancing selection.

In summary, our description of the major populations of the *L. donovani* species complex present provides insight into the global population dynamics of this parasite. The identification of genes with robust signatures of balancing selection shows that balancing selection operates in this parasite, and provides candidate genes for the study of host-parasite and host-vector interactions.

Materials and Methods

Ethics

Samples from Brazil were obtained as part of a broad study for genomic studies in the Laboratory of Leishmaniasis at the Institute of Tropical Medicine Natan Portella, approved by the Research Ethics Committee of the Federal University of Piauí (approval ID number 0116/2005). All methods were performed according to the approved guidelines and regulations. A written informed consent was obtained from all study participants or their legal guardians.

Strain culture and genome sequencing

Bone marrow aspirates were obtained from the routine diagnosis of patients admitted to the Natan Portella Tropical Diseases Institute in Teresina-PI, Brazil. Aspirates were inoculated into a mixed culture medium NNN (Neal, Novy, Nicolle) containing 2 ml of Schneider's medium supplemented with 10% fetal bovine serum, 2% urine and penicillin 10,000 U/ml, and streptomycin 10 mg/ml. The positive isolates in mixed media were expanded in Schneider's liquid medium under the same conditions mentioned above. Extraction of DNA from the parasites was performed after washing to remove culture medium, using Qiagen Blood and Tissue kit was used according to the manufacturer's recommendation.

Genome sequencing was performed on Illumina HiSeq 2500, or similar machines to produce paired end 150 nt reads. The majority (95%) of the samples were sequenced to provide mapped read coverage of ≥ 30x (mean 97x, minimum 19x). Raw sequencing reads were submitted to NCBI's sequencing read archive (SRA) under the BioProject accession PRJNA702997 (note to reviewers: data are scheduled to be released upon publication).

Sequence analysis/variant calling

Publicly available *L. donovani* complex data was downloaded in FASTQ format from the European Nucleotide Archive (ENA: https://www.ebi.ac.uk/ena). Full list of strain names/ENA numbers in **Supplementary Table 9**. The *L. donovani* reference genome (strain BPK282A1) was downloaded from TriTrypDB (version Nov 2019). Strain reads were mapped to the reference using bwa v.0.7.17 (Li and Durbin, 2009), converted to bam, sorted, indexed and duplicates removed with SAMtools v.1.9 (Li *et al.*, 2009).
For each strain, SNPs and indels were called using The Genome Analysis Toolkit (GATK) HaplotypeCaller v.4.1.0.0 (Depristo *et al.*, 2011) using the 'discovery' genotyping mode Freebayes v.1.3.2 (https://github.com/ekg/freebayes) accepting calls with a minimum alternative allele read count ≥ 5. We accepted calls discovered by both methods, merged all VCFs and regenotyped with Freebayes. The regenotyped VCF was sorted with Picard SortVcf (https://broadinstitute.github.io/picard/) and indexed with GATK IndexFeatureFile. SNP hard-filtering was performed with BCFtools (https://samtools.github.io/bcftools/) on biallelic variants only, to remove sites with any of the following: DPRA < 0.73 or > 1.48; QA or QR < 100; SRP or SAP > 2000; RPP or RPPR > 3484; PAIRED or PAIRPAIREDR < 0.8; MQM or MQMR < 40. As chromosome 31 is generally supernumerary, we specified DPB < 30401 or > 121603 to be removed, and for remaining chromosomes, DPB < 18299 or > 73197 (<0.5x or >2x median DPB). Biallelic indels were filtered to remove sites with any of the following: DPRA < 0.73 or > 1.48; QA or QR < 100; SRP or SAP > 2000.

With this variant filtering we observed a correlation between minor allele frequency (MAF) and read coverage at SNP sites (**Supplementary Figure 15**). Modelling showed that duplications resulted in a systematic bias against calling rare alleles. We therefore removed any SNP/indel sites where the mean variant coverage within the ADMIXTURE-defined population was ≥1.5x larger than the median coverage (corresponding to triploid sites in a

generally diploid chromosome), or ≥1.25x larger than the median coverage for chromosome 31 (corresponding to tetraploid sites in a generally triploid chromosome). We also removed sites where coverage was highly variable, by excluding sites in the upper 5th percentile of the standard deviation. In each population this filtered ~5-7% of sites. After this filtering, the correlation between MAF and read coverage was either far less significant or removed completely. This filtering retained 10377 out of a possible 10778 sites in population ISC1; 9781 out of a possible 10227 sites in population ISC2; 40127 out of a possible 41957 sites in EA1; 11757 out of a possible 12365 sites in EA2 and 26884 out of a possible 28281 sites in Brazil-Med.

VCF annotation was performed with the snpEff v.4.3 package (Cingolani *et al.*, 2012) using the default Leishmania_donovani_BPK282A1 database included with the software. SnpSift filter with the option "ANN[*].EFFECT has 'missense_variant'" was used to extract nonsynonymous sites.

Mapping coverage was ascertained by SAMtools bedcov for each gene in the multipopulation VCF.

To validate the variant filtering we produced a *de novo* assembly of the MHOM/BR/06/MA01A *L. infantum* isolate from Brazil (Carnielli *et al.*, 2018), mapped Illumina reads from the same isolate to the assembly, and called SNPs and indels as above. All calls should be heterozygous sites, or errors. Initial variant calling identified 4 SNPs and 23 indels, after filtering no SNPs or indels remained, consistent with a very low false positive call per strain. The MHOM/BR/06/MA01A *de novo* assembly will be described elsewhere. Briefly, the assembly was produced using Oxford Nanopore Technology (ONT) reads to 110x coverage, assembled with Canu v.1.9 (Koren *et al.*, 2017), polished once using ONT reads using Nanopolish (v.0.9.2) (Loman, Quick and Simpson, 2015) and thrice with Illumina reads using Pilon (v.1.22) (Walker *et al.*, 2014).

Repeat regions were determined as follows. Intergenic coordinates in *L. donovani* were extracted from the annotation .gff, downloaded from TriTrypDB (version Nov 2019) with BEDtools (v.2.27.1 (Quinlan and Hall, 2010)) complement with default parameters.

Intergenic regions were then extracted from the genome using BEDtools getfasta. Repeat regions in *L. donovani* were identified by nBLASTing (v.2.7.1; (Altschul *et al.*, 1990)) intergenic regions against the rest of the genome, removing redundant hits and those <200nt in length. Resulting coordinates were converted to bed format for filtering.

Phylogenetic analysis

VCF containing all variants from all 477 isolates was converted to PHYLIP format using vcf2phylip (available at https://github.com/edgardomortiz/vcf2phylip/tree/v2.0). This produced an alignment of 283378 sites. IQ-TREE v.1.5.5 (Nguyen *et al.*, 2015) was used to perform Maximum Likelihood (ML) phylogenetic analysis with the model GTR+ASC, which includes ascertainment bias correction, with 1000 bootstrap replicates and 1000 UFBOOT (Hoang *et al.*, 2018) approximations to produce ML support values. The resulting tree was visualised with Figtree v.1.4.4 (available at http://tree.bio.ed.ac.uk/). Treefiles are available in supplementary material.

Population and diversity analysis

For all population analyses we utilised only biallelic SNPs, pruning linked sites ($r^2 > 0.5$) in 2kb windows with a step size of 1 with PLINK v.1.9 (Purcell *et al.*, 2007) using the option --indep-pairwise 2kb 1 0.5. This produced 194351 SNPs from the initial 353301 (158950 variants removed). ADMIXTURE v.1.3 (Alexander and Lange, 2011) was run with K = 1-12. Principal component coordinates were produced with PLINK v.1.9.

VCFs were filtered with VCFtools v.0.1.15 (Danecek *et al.*, 2011) and option --mac 1. Repeat regions (see above) were also filtered out of VCFs at this stage.

Prior to balancing selection tests, mixed ancestry strains were removed from population VCFs. Tajima's D, π and MAF were calculated on unpruned variants using VCFtools. Tests for balancing selection used biallelic SNPs and indels from each population. Copy-number variant and duplicated genome regions were removed from this analysis, as these regions will produce biases in allele frequencies towards common alleles, producing artifactual signals of BS (**Supplementary Text 1, Supplementary Figure 6**). Variant calling for multi-copy regions was beyond the scope of this study.

The NCD2 test used software provided by Bitarello *et al.* (2018), using windows of 1, 5 and 10 kb with step sizes of 0.5, 2.5 and 5kb, respectively. 10 kb windows sizes were used in this study. A list of fixed differences between *L. donovani* populations (total 285 isolates) and the 'pure' Brazilian-Mediterranean *L. infantum* population (127 isolates) was used as the outgroup for analysis. Fixed differences were determined using bcftools isec called on VCFs of all *L. donovani* populations and *L. infantum* containing fixed variants, resulting in 76284 fixed difference sites. Resulting target frequencies (*tf*) and Equation 2 of (Bitarello *et al.*, 2018) were used to generate *Ztf-IS* scores, with the exception of using the standard deviation (SD) for each number of informative sites (IS) rather than simulated SD. *P* values for each window were calculated by assigning a rank based on *Ztf-IS* score and dividing by the total number of scanned windows.

The BetaScan2 test (Siewert and Voight, 2020), using default parameters, using the file format generated from the variants using glactools (Renaud 2017; available at: https://github.com/grenaud/glactools), using the Brazil-Mediterranean *L. infantum* population as the outgroup. We performed the test on each individual population.

Gene Ontology analysis

Gene Ontology (GO) descriptions and gene details for the *L. donovani* BPK282A1 reference genome were downloaded from TriTrypDB. GO enrichment analysis was performed using the PANTHER service on tritrypdb.org. Proteins that were classed as 'hypothetical' or of 'unknown function' were BLASTed against the non-redundant protein sequences (nr) database of NCBI to obtain possible identity by shared homology, and to determine conservation across trypanosomes.

Author contributions

CAG - performed data curation, sequence analysis and variant calling, phylogenetics, population and diversity analysis, balancing selection tests, gene ontology analysis, contributed to writing the manuscript
SJF - data curation, sequence analysis and variant calling pipelines, contributions to data analysis and writing the manuscript
VC - performed isolate culture and DNA extraction for Brazil samples
AA - contributed to population and diversity analysis
HK - contributed to population and diversity analysis
YPC - contributed to data analysis
SJ - performed library preparation and sequencing of isolates from Brazil
DLC - contributed to isolate collection in Brazil
JCM - obtained funding, contributed to data analysis and writing the manuscript
CHNC - obtained funding, collected and cultured isolates from Brazil, contributed to the manuscript
DCJ - obtained funding, initiated the project, supervised student work, assisted with data analysis, wrote the manuscript

Acknowledgements

This work has been produced as part of the UK:Brazil Joint Centre Partnership in Leishmaniasis. This project was undertaken on the Viking Cluster, which is a high performance computer facility provided by the University of York. We are grateful for computational support from the University of York High Performance Computing service, Viking and the Research Computing team. We acknowledge Shoumit Dey and Martina Stoycheva for contributions to the VCF pipeline. We acknowledge the support of John Davies for the Brazil *Leishmania infantum* assembly and The Genomics and Bioinformatics Laboratory at The University of York for their assistance.

Supplementary data

Supplementary Table 9 lists all isolates used in this study, including source, location, and ENA accession numbers, including genome data from Brazil (produced in this work). All

sequence data generated in this study have been deposited under the NCBI BioProject accession PRJNA702997 and will be released upon publication. All Betascan2, NCD2, Tajima's *D* and nucleotide diversity metrics and Supplementary Tables are available from the Figshare repository https://figshare.com/account/home#/projects/94292 (Note to reviewers: this place holder will be replaced when the article is accepted).

Funding

SJF was supported by a Wellcome Seed Award in Science to DCJ (208965/Z/17/Z). CAG was supported by MRC Newton as a component of the UK:Brazil Joint Centre Partnership in Leishmaniasis to JCM (MR/S019472/1).

References

Alexander, D. H. and Lange, K. (2011) 'Enhancements to the ADMIXTURE algorithm for individual ancestry estimation', *BMC bioinformatics*, 12, p. 246. doi: 10.1186/1471-2105-12-246.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of molecular biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Amambua-Ngwa, A. *et al.* (2012) 'Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites', *PLoS genetics*, 8(11), p. e1002992. doi: 10.1371/journal.pgen.1002992.

Anantharaman, V. and Aravind, L. (2002) 'The GOLD domain, a novel protein module involved in Golgi function and secretion', *Genome biology*, 3(5), p. research0023. doi: 10.1186/gb-2002-3-5-research0023.

An, T. and Li, Z. (2018) 'An orphan kinesin controls trypanosome morphology transitions by targeting FLAM3 to the flagellum', *PLoS pathogens*, 14(5), p. e1007101. doi: 10.1371/journal.ppat.1007101.

Ashutosh *et al.* (2012) 'Downregulation of mitogen-activated protein kinase 1 of *Leishmania donovani* field isolates is associated with antimony resistance', *Antimicrobial agents and chemotherapy*, 56(1), pp. 518–525. doi: 10.1128/AAC.00736-11.

Atayde, V. D. *et al.* (2016) '*Leishmania* exosomes and other virulence factors: Impact on innate immune response and macrophage functions', *Cellular immunology*, 309, pp. 7–18. doi: 10.1016/j.cellimm.2016.07.013.

Baker, N. *et al.* (2020) 'Systematic functional analysis of *Leishmania* protein kinases identifies regulators of differentiation or survival', *Cold Spring Harbor Laboratory*. doi: 10.1101/2020.09.06.279091.

Bhattacharya, S. K. *et al.* (2006) 'Elimination of leishmaniasis (kala-azar) from the Indian subcontinent is technically feasible & operationally achievable', *The Indian journal of medical research*, 123(3), pp. 195–196. Available at: https://www.ncbi.nlm.nih.gov/pubmed/16778303.

Bitarello, B. D. *et al.* (2018) 'Signatures of Long-Term Balancing Selection in Human Genomes', *Genome biology and evolution*, 10(3), pp. 939–955. doi: 10.1093/gbe/evy054.

Boité, M. C. *et al.* (2019) 'Trans-Atlantic Spill Over: Deconstructing the Ecological Adaptation of *Leishmania infantum* in the Americas', *Genes*, 11(1). doi: 10.3390/genes11010004.

Carnielli, J. B. T. *et al.* (2018) 'A *Leishmania infantum* genetic marker associated with miltefosine treatment failure for visceral leishmaniasis', *EBioMedicine*, 36, pp. 83–91. doi: 10.1016/j.ebiom.2018.09.029.

Carvalho, K. S. S. *et al.* (2020) 'Application of next generation sequencing (NGS) for descriptive analysis of 30 genomes of *Leishmania infantum* isolates in Middle-North Brazil', *Scientific reports*, 10(1), p. 12321. doi: 10.1038/s41598-020-68953-9.

Charlesworth, D. (2006) 'Balancing selection and its effects on sequences in nearby genome regions', *PLoS genetics*, 2(4), p. e64. doi: 10.1371/journal.pgen.0020064.

Cheng, X. and DeGiorgio, M. (2019) 'Detection of Shared Balancing Selection in the Absence of Trans-Species Polymorphism', *Molecular biology and evolution*, 36(1), pp. 177–199. doi: 10.1093/molbev/msy202.

Cheng, X. and DeGiorgio, M. (2020) 'Flexible Mixture Model Approaches That Accommodate Footprint Size Variability for Robust Detection of Balancing Selection', *Molecular biology and evolution*, 37(11), pp. 3267–3291. doi: 10.1093/molbev/msaa134.

Cingolani, P. *et al.* (2012) '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(2), pp. 80–92. doi: 10.4161/fly.19695.

Cotton, J. A. *et al.* (2020) 'Genomic analysis of natural intra-specific hybrids among Ethiopian isolates of *Leishmania donovani*', *PLoS neglected tropical diseases*, 14(4), p. e0007143. doi: 10.1371/journal.pntd.0007143.

Danecek, P. *et al.* (2011) 'The variant call format and VCFtools', *Bioinformatics* , 27(15), pp. 2156–2158.

Dean, S. *et al.* (2016) 'Cilium transition zone proteome reveals compartmentalization and differential dynamics of ciliopathy complexes', *Proceedings of the National Academy of Sciences of the United States of America*, 113(35), pp. E5135–43. doi: 10.1073/pnas.1604258113.

DeGiorgio, M., Lohmueller, K. E. and Nielsen, R. (2014) 'A model-based approach for identifying signatures of ancient balancing selection in genetic data', *PLoS genetics*, 10(8), p. e1004561. doi: 10.1371/journal.pgen.1004561.

Depristo, M. A. *et al.* (2011) 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature genetics*, 43(5), pp. 491–498. Available at: http://www.nature.com/ng/journal/v43/n5/full/ng.806.html.

Dhiman, R. C. and Yadav, R. S. (2016) 'Insecticide resistance in phlebotomine sandflies in Southeast Asia with emphasis on the Indian subcontinent', *Infectious diseases of poverty*, 5(1), p. 106. doi:

10.1186/s40249-016-0200-3.

Dong, G., Filho, A. L. and Olivier, M. (2019) 'Modulation of Host-Pathogen Communication by Extracellular Vesicles (EVs) of the Protozoan Parasite *Leishmania*', *Frontiers in cellular and infection microbiology*, 9, p. 100. doi: 10.3389/fcimb.2019.00100.

Dumetz, F. *et al.* (2017) 'Modulation of Aneuploidy in *Leishmania donovani* during Adaptation to Different In Vitro and In Vivo Environments and Its Impact on Gene Expression', *mBio*, 8(3). doi: 10.1128/mBio.00599-17.

Dye, C. and Wolpert, D. M. (1988) 'Earthquakes, influenza and cycles of Indian kala-azar', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 82, pp. 843–850.

Ferreira, G. E. M. *et al.* (2012) 'The genetic structure of *Leishmania infantum* populations in Brazil and its possible association with the transmission cycle of visceral leishmaniasis', *PloS one*, 7(5), p. e36242.

Field, M. C. (2005) 'Signalling the genome: the Ras-like small GTPase family of trypanosomatids', *Trends in parasitology*, 21(10), pp. 447–450. doi: 10.1016/j.pt.2005.08.008.

Franssen, S. U. *et al.* (2020) 'Global genome diversity of the *Leishmania donovani* complex', *eLife*, 9. doi: 10.7554/eLife.51243.

Gelanew, T. *et al.* (2010) 'Inference of Population Structure of *Leishmania donovani* Strains Isolated from Different Ethiopian Visceral Leishmaniasis Endemic Areas', *PLoS Neglected Tropical Diseases*, p. e889. doi: 10.1371/journal.pntd.0000889.

Gelanew, T. *et al.* (2014) 'Multilocus sequence and microsatellite identification of intra-specific hybrids and ancestor-like donors among natural Ethiopian isolates of *Leishmania donovani*', *International journal for parasitology*, 44(10), pp. 751–757. doi: 10.1016/j.ijpara.2014.05.008.

Guy, A. J. *et al.* (2018) 'Proteome-wide mapping of immune features onto *Plasmodium* protein three-dimensional structures', *Scientific Reports*. doi: 10.1038/s41598-018-22592-3.

Hoang, D. T. *et al.* (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular biology and evolution*, 35(2), pp. 518–522. doi: 10.1093/molbev/msx281.

Imamura, H. *et al.* (2016) 'Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent', *eLife*, 5(MARCH2016), p. 56. Available at: http://elifesciences.org/lookup/doi/10.7554/eLife.12613.

Jeffares, D. C. *et al.* (2015) 'The genomic and phenotypic diversity of Schizosaccharomyces pombe', *Nature genetics*, 47(3), pp. 235–241.

Koren, S. *et al.* (2017) 'Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation', *Genome research*, 27(5), pp. 722–736. doi: 10.1101/gr.215087.116.

Kuhls, K. *et al.* (2011) 'Comparative microsatellite typing of New World *Leishmania infantum* reveals low heterogeneity among populations and its recent old world origin', *PLoS neglected tropical diseases*, 5(6), p. e1155.

Kumar, A., Chauhan, N. and Singh, S. (2019) 'Understanding the Cross-Talk of Redox Metabolism and Fe-S Cluster Biogenesis in *Leishmania* Through Systems Biology Approach', *Frontiers in cellular and infection microbiology*, 9, p. 15. doi: 10.3389/fcimb.2019.00015.

Laranjeira-Silva, M. F. *et al.* (2018) 'A MFS-like plasma membrane transporter required for *Leishmania* virulence protects the parasites from iron toxicity', *PLoS pathogens*, 14(6), p. e1007140. doi: 10.1371/journal.ppat.1007140.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map (SAM) Format and SAMtools', *Bioinformatics* . Available at: http://bioinformatics.oxfordjournals.org/cgi/reprint/btp352v1.pdf.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics* , 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Loman, N. J., Quick, J. and Simpson, J. T. (2015) 'A complete bacterial genome assembled de novo using only nanopore sequencing data', *Nature methods*, 12(8), pp. 733–735. doi: 10.1038/nmeth.3444.

Mullaney, J. M. *et al.* (2010) 'Small insertions and deletions (INDELs) in human genomes', *Human molecular genetics*, 19(R2), pp. R131–6. doi: 10.1093/hmg/ddq400.

Muniaraj, M. (2014) 'The lost hope of elimination of Kala-azar (visceral leishmaniasis) by 2010 and cyclic occurrence of its outbreak in India, blame falls on vector control practices or co-infection with human immunodeficiency virus or therapeutic modalities?', *Tropical parasitology*, 4(1), pp. 10–19. doi: 10.4103/2229-5070.129143.

Nguyen, L.-T. *et al.* (2015) 'IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular biology and evolution*, 32(1), pp. 268–274. doi: 10.1093/molbev/msu300.

Nunes, V. S. *et al.* (2011) 'The Hus1 homologue of *Leishmania major* encodes a nuclear protein that participates in DNA damage response', *Molecular and biochemical parasitology*, 177(1), pp. 65–69. doi: 10.1016/j.molbiopara.2011.01.011.

Nygaard, S. *et al.* (2010) 'Long- and short-term selective forces on malaria parasite genomes', *PLoS genetics*, 6(9), p. e1001099. doi: 10.1371/journal.pgen.1001099.

Ochola-Oyier, L. I. *et al.* (2019) 'Few *Plasmodium falciparum* merozoite ligand and erythrocyte receptor pairs show evidence of balancing selection', *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 69, pp. 235–245. doi: 10.1016/j.meegid.2019.02.004.

Purcell, S. *et al.* (2007) 'PLINK: A tool set for whole-genome association and population-based linkage analyses', *American journal of human genetics*, 81(3), pp. 559–575. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17701901&retmode=ref&cmd=prlinks.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics* , 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Raj, S. *et al.* (2019) 'Biochemical characterization and chemical validation of *Leishmania* MAP Kinase-3 as a potential drug target', *Scientific reports*, 9(1), p. 16209. doi: 10.1038/s41598-019-52774-6.

Rocker, A. and Meinhart, A. (2015) 'A cis-acting antitoxin domain within the chromosomal toxin–antitoxin module EzeT of *Escherichia coli* quenches toxin activity', *Molecular microbiology*, 97(3), pp. 589–604. doi: 10.1111/mmi.13051.

Rogers, M. B. *et al.* (2014) 'Genomic Confirmation of Hybridisation and Recent Inbreeding in a

Vector-Isolated *Leishmania* Population', *PLoS genetics*, 10(1), p. e1004092.

Schwabl, P. *et al.* (2021) 'Colonization and genetic diversification processes of *Leishmania infantum* in the Americas', *Communications biology*, 4(1), p. 139. doi: 10.1038/s42003-021-01658-5.

Siewert, K. M. and Voight, B. F. (2017) 'Detecting Long-Term Balancing Selection Using Allele Frequency Correlation', *Molecular biology and evolution*, 34(11), pp. 2996–3005. doi: 10.1093/molbev/msx209.

Siewert, K. M. and Voight, B. F. (2020) 'BetaScan2: Standardized Statistics to Detect Balancing Selection Utilizing Substitution Data', *Genome biology and evolution*, 12(2), pp. 3873–3877. doi: 10.1093/gbe/evaa013.

Srivastava, A. *et al.* (2019) 'Identification and functional characterization of a bacterial homologue of Zeta toxin in *Leishmania donovani*', *FEBS letters*, 593(11), pp. 1223–1235. doi: 10.1002/1873-3468.13429.

Stern, D. B. and Lee, C. E. (2020) 'Evolutionary origins of genomic adaptations in an invasive copepod', *Nature ecology & evolution*, 4(8), pp. 1084–1094. doi: 10.1038/s41559-020-1201-y.

Sunter, J. D. *et al.* (2015) 'Modulation of flagellum attachment zone protein FLAM3 and regulation of the cell shape in *Trypanosoma brucei* life cycle transitions', *Journal of cell science*, 128(16), pp. 3117–3130. doi: 10.1242/jcs.171645.

Sunter, J. D. *et al.* (2019) '*Leishmania* flagellum attachment zone is critical for flagellar pocket shape, development in the sand fly, and pathogenicity in the host', *Proceedings of the National Academy of Sciences of the United States of America*, 116(13), pp. 6351–6360. doi: 10.1073/pnas.1812462116.

Tajima, F. (1989) 'Statistical method for testing the neutral mutation hypothesis by DNA polymorphism', *Genetics*, 123(3), pp. 585–595.

Teixeira, D. G. *et al.* (2017) 'Comparative analyses of whole genome sequences of *Leishmania infantum* isolates from humans and dogs in northeastern Brazil.', *International journal for parasitology*, 47(10-11), pp. 655–665.

Tetteh, K. K. A. *et al.* (2009) 'Prospective identification of malaria parasite genes under balancing selection', *PloS one*, 4(5), p. e5568. doi: 10.1371/journal.pone.0005568.

Thakur, C. P. (2007) 'A new strategy for elimination of kala-azar from rural Bihar', *The Indian journal of medical research*, 126(5), pp. 447–451. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18160749.

Wahlgren, M., Goel, S. and Akhouri, R. R. (2017) 'Variant surface antigens of *Plasmodium falciparum* and their roles in severe malaria', *Nature reviews. Microbiology*, 15(8), pp. 479–491. doi: 10.1038/nrmicro.2017.47.

Walker, B. J. *et al.* (2014) 'Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement', *PloS one*, 9(11), p. e112963. doi: 10.1371/journal.pone.0112963.

Weedall, G. D. and Conway, D. J. (2010) 'Detecting signatures of balancing selection to identify targets of anti-parasite immunity', *Trends in parasitology*, 26(7), pp. 363–369. doi: 10.1016/j.pt.2010.04.002.

World Health Organisation (2020) *Leishmaniasis*. Available at:

https://www.who.int/news-room/fact-sheets/detail/leishmaniasis (Accessed: 21 January 2021).

Zackay, A. *et al.* (2018) 'Genome wide comparison of Ethiopian *Leishmania donovani* strains reveals differences potentially related to parasite survival', *PLoS genetics*, 14(1), p. e1007133. Available at: http://dx.plos.org/10.1371/journal.pgen.1007133.

Supplementary Text

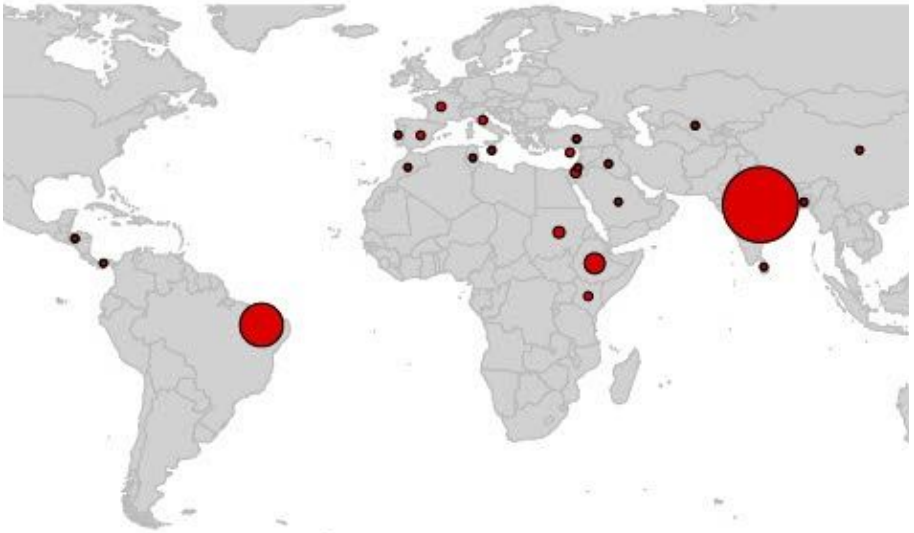
Supplementary Text 1. Duplications distort allele frequencies

Initial analysis of the allele frequencies of these populations showed a tail of high MAF alleles, that would be consistent with strong balancing selection in some sites. However, closer inspection of these data showed that all populations had a strong positive correlation between read coverage and MAF (**Supplementary Figure 6**). Very high copy regions, which included multiple duplicated genes, were significantly enriched for common alleles. This bias is most likely due to systematic under-calling of rare alleles in duplicated genes, combined with the appearance of 'balanced' allele frequencies caused by variants fixed in one duplication (**Supplementary Figure 17**). These processes will distort allele frequencies, and result in artefacts in many of the methods used to detect balancing selection, such as Tajima's *D* (Tajima, 1989) and *NCD2* (Bitarello *et al.*, 2018), and *BetaScan/Betascan2* which utilise correlated alleles (Siewert and Voight, 2017, 2020). As the missing rare alleles are generally newer, this process will also reduce population distances (F_{ST}), an expectation of long term balancing selection (Charlesworth, 2006).

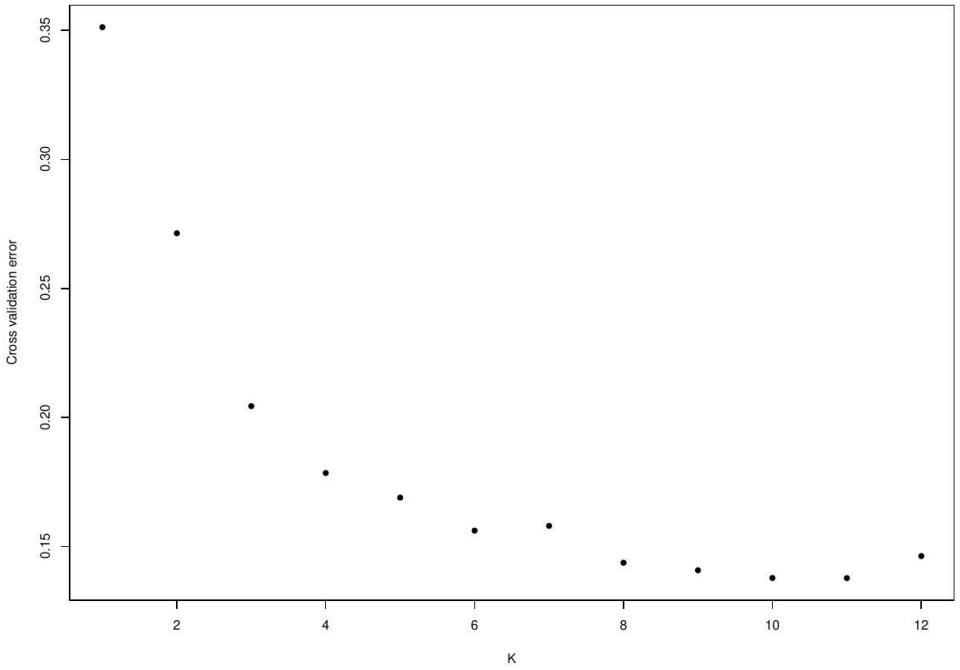
Supplementary Text 2. Full description of BS tests and justifications

To identify the most likely candidates of BS in these populations we applied a search strategy that used genomic window analyses, followed by gene-centric analysis (**Supplementary Figure 6**). To identify genomic windows that contain signals consistent with BS we applied two tests, *BetaScan2* (Siewert and Voight, 2020) and *NCD2 (Bitarello et al., 2018)* over 10 kb windows around each segregating site, for each population. *BetaScan2* detects regions with alleles 'balanced' at correlated frequencies and a deficit of substitutions compared to the outgroup, while the *NCD2* test detects regions with an excess of alleles near a target frequency (0.5 in our case). Selection of outlier regions with *Betascan2* and/or *NCD2* scores in any population identified 258 genes (13 were identified in more than one population).
For gene-centric analysis, we identified specific genes that may be BS targets within these windows, therefore we calculated nucleotide diversity (π) and Tajima's *D* ((Tajima, 1989; Bitarello *et al.*, 2018; Siewert and Voight, 2020) for all genes in all populations (**Supplementary Tables A, B**). We selected genes in the 90th percentile of either statistic for any population as the most likely targets of balancing selection. Neither of these gene sets were significantly enriched for any Gene Ontology categories after multiple-test correction (**Supplementary Tables C, D**). Selection of genes with both genomic window tests (*NCD2* or *Betascan*) and gene-centric metrics (Tajima's *D* or π) identified 33 genes (**Supplementary Table 5**).

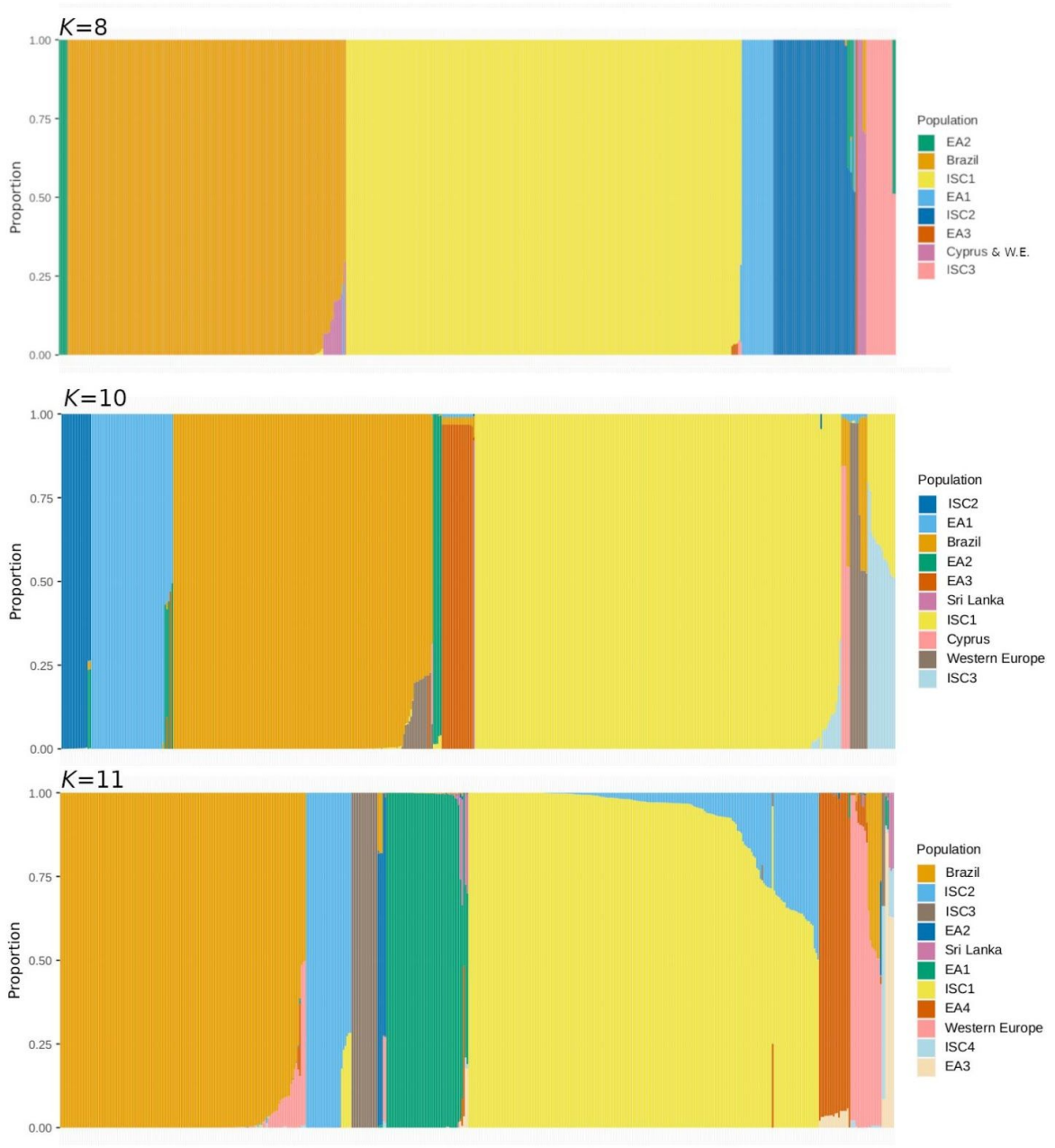
Supplementary Figures



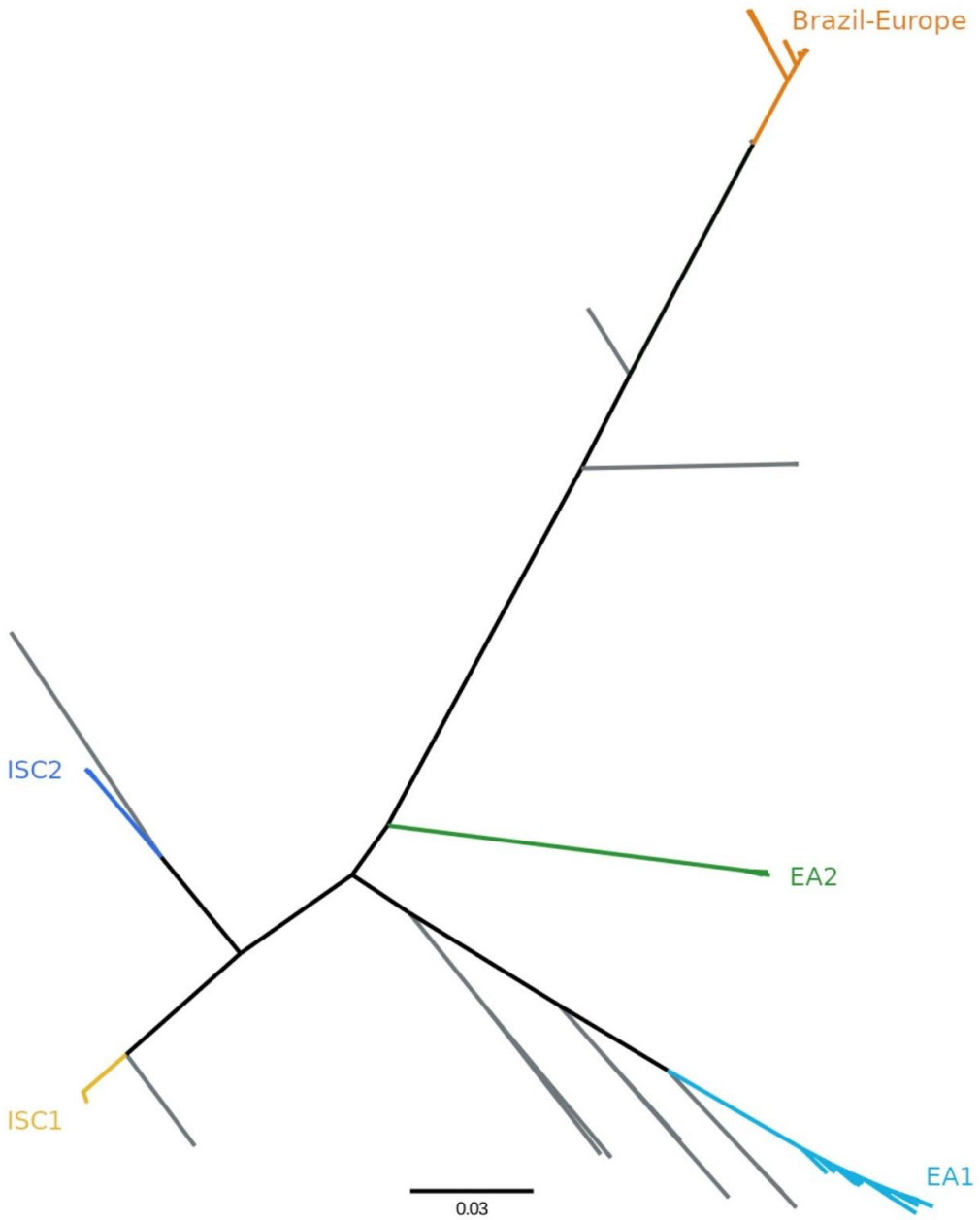
Supplementary Figure 1. Map of isolate origins. Size of the circle corresponds to the number of isolates used from this location.



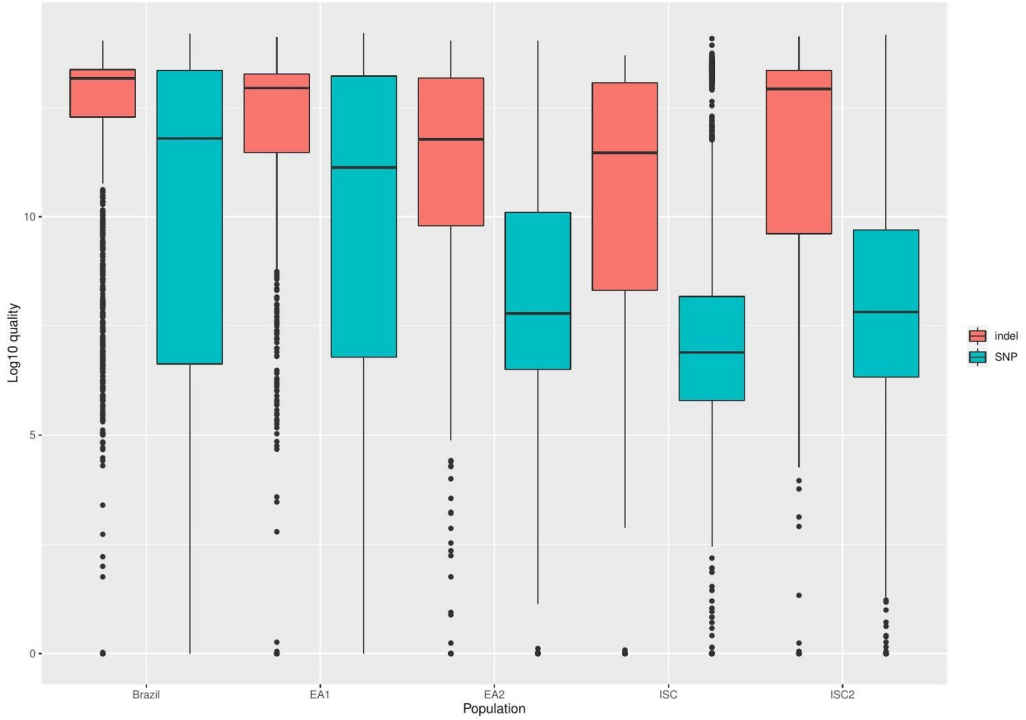
Supplementary Figure 2. ADMIXTURE cross-validation error for *K*=1-12. ADMIXTURE analysis of 477 isolates indicated a true population number between 8-11.



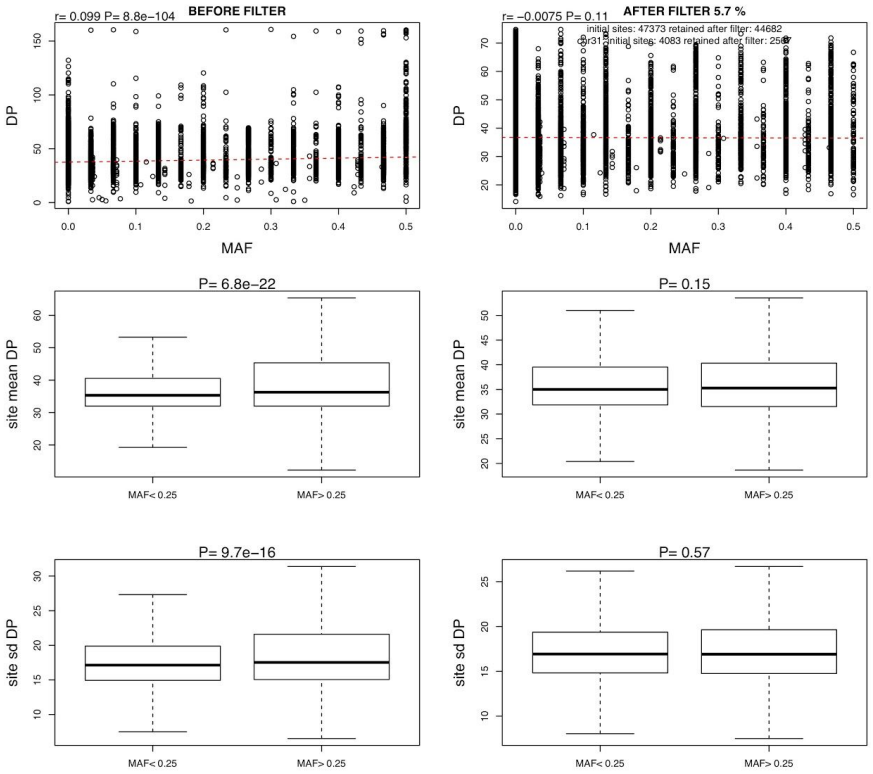
Supplementary Figure 3. ADMIXTURE results for *K*=8, 10 and 11. Isolates assigned to populations when *K*= 8, 10 and 11. Each vertical bar represents a single isolate. We proceeded with *K*=9 (**Figure 1**) in line with previous analysis by (Franssen *et al.*, 2020). Populations are named according to their predominant origin. WE = Western Europe.



Supplementary Figure 4. Unrooted phylogenetic tree of 477 isolates. Maximum likelihood phylogeny, based upon a SNP alignment of 477 sequences with 283378 variable sites. All visible branches are maximally supported (100% mlBP) unless indicated. The scale bar represents the number of nucleotide changes per site. Branches in grey are those isolates assigned to minor populations which we do not analyse further.

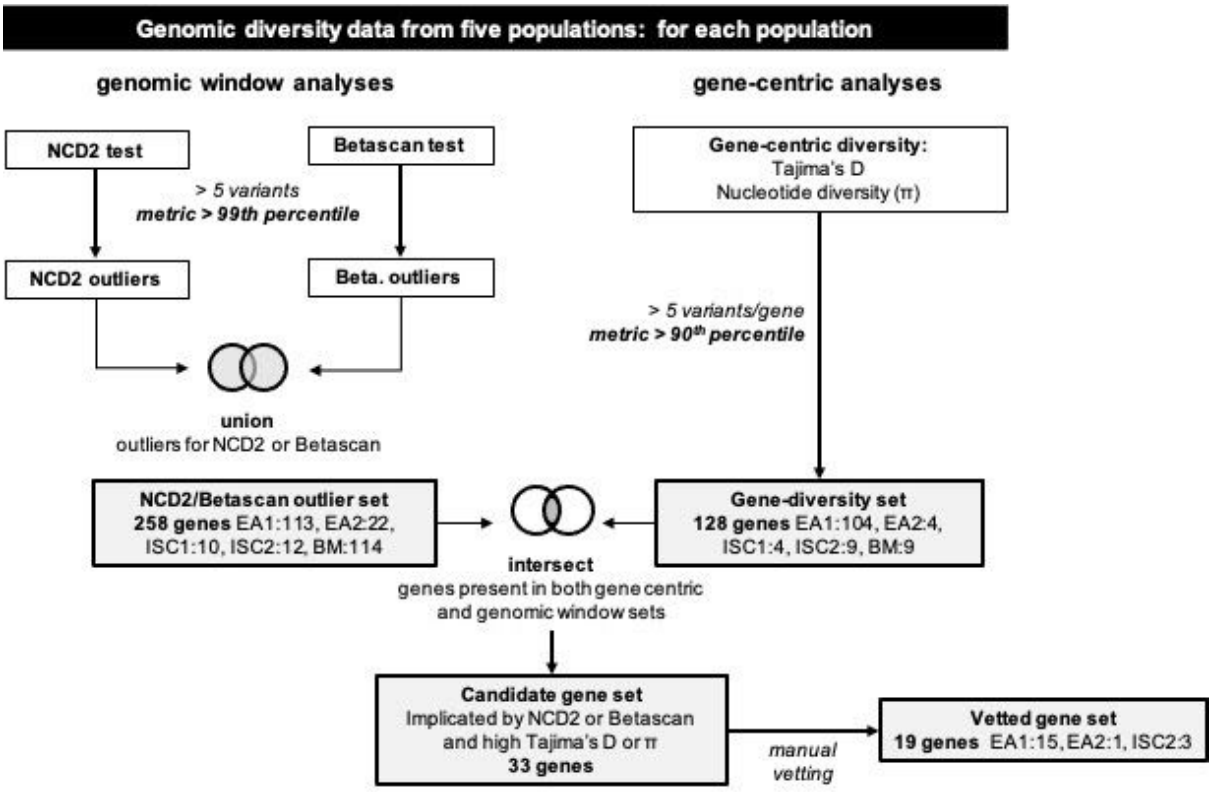


Supplementary Figure 5. The excess of indels is unlikely to be artefactual. Log10 QUAL scores for SNP and indel positions are displayed for each population.

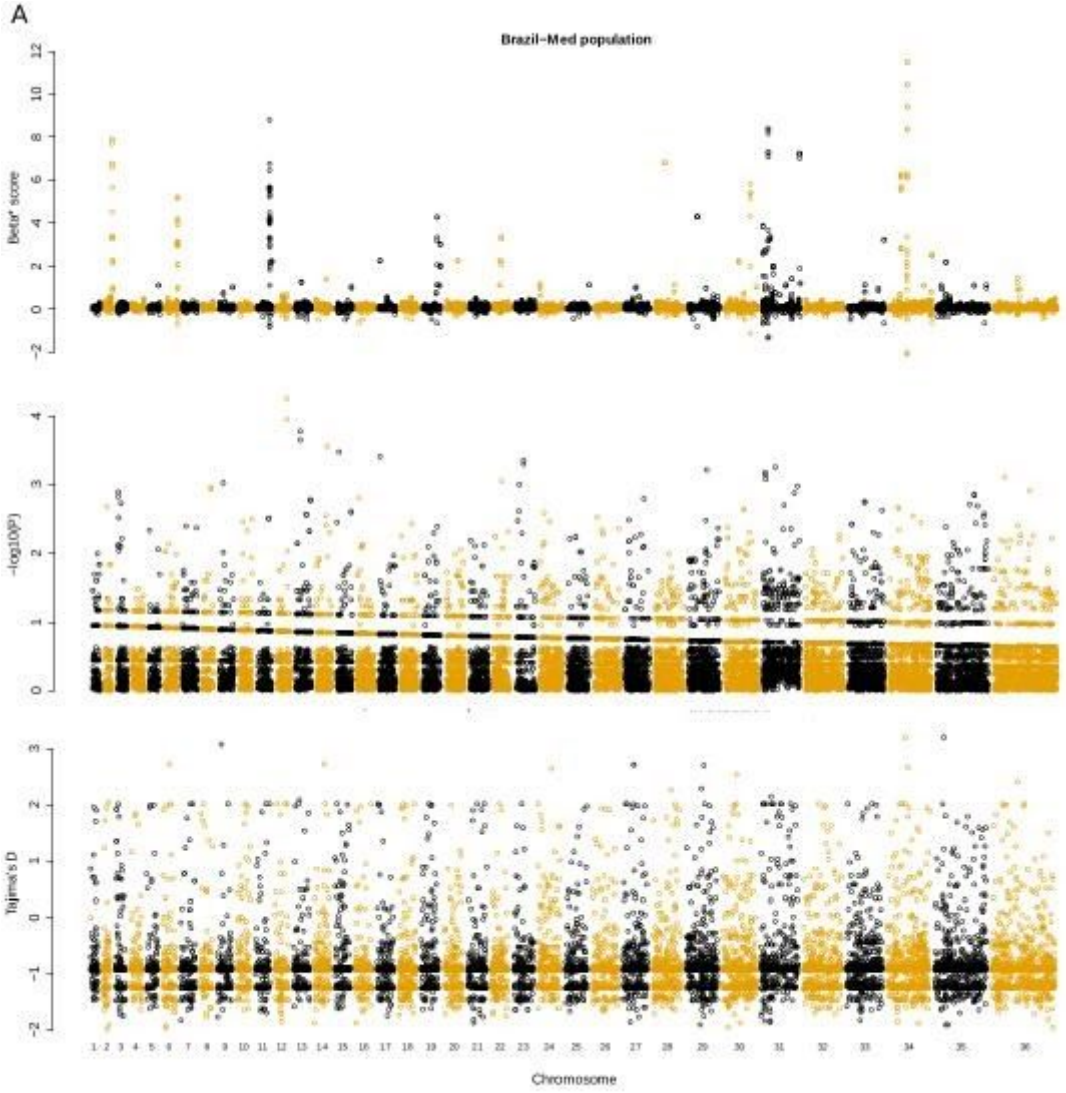


Supplementary Figure 6. Duplications systematically bias allele frequencies. Data shown are from the ISC2 population. Left three panels show data prior to depth filtering. Top panel, a positive correlation between total read depth (DP) and minor allele frequency (MAF). Middle panel, sites with MAF > 0.25 are significantly higher read depth. Lower panel, sites

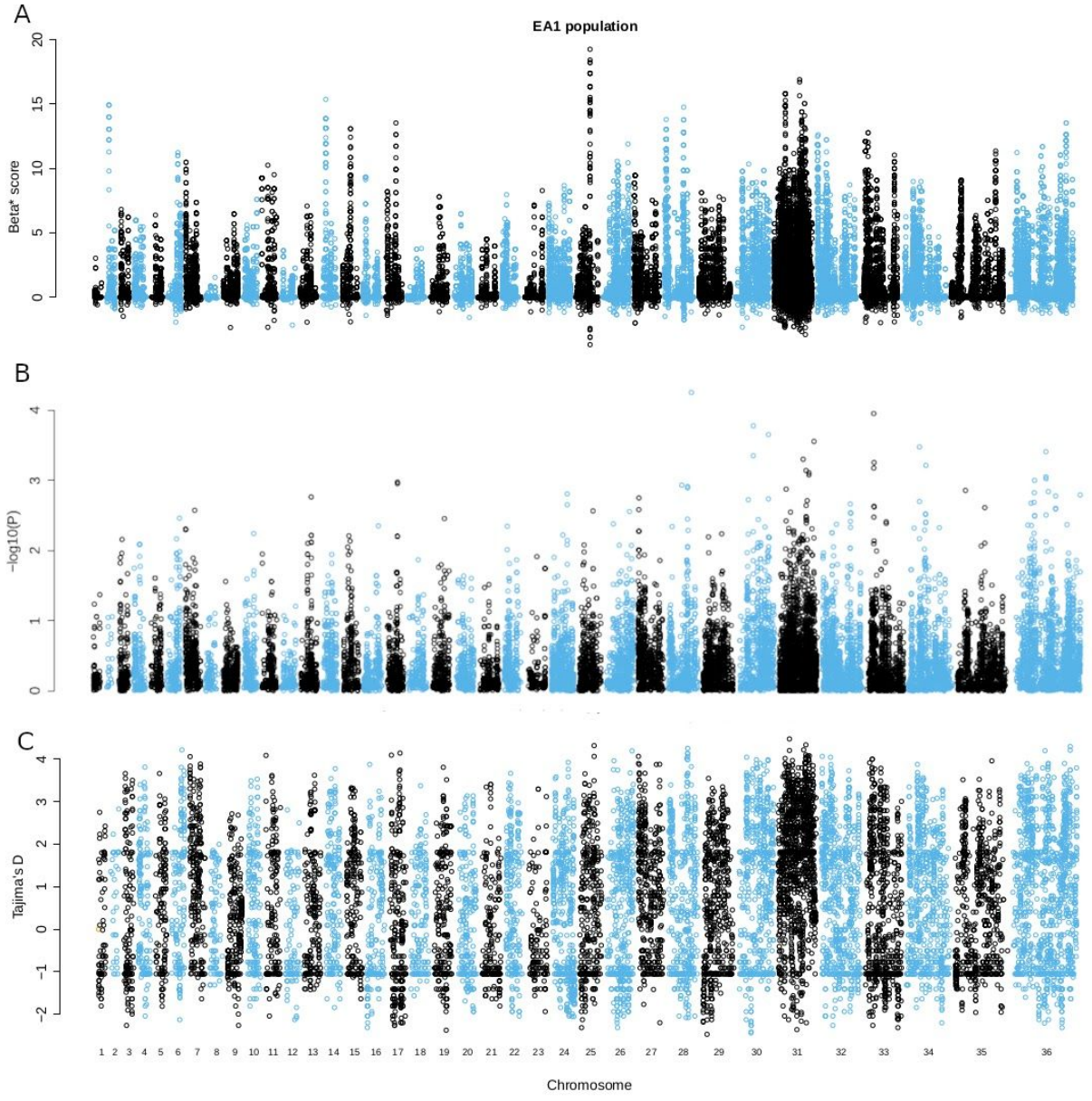
with MAF > 0.25 are significantly higher variation in (standard deviation, sd) in read depth. Right three panels show the same presentations of data after read depth filtering. None of the relationships remain statistically significant.



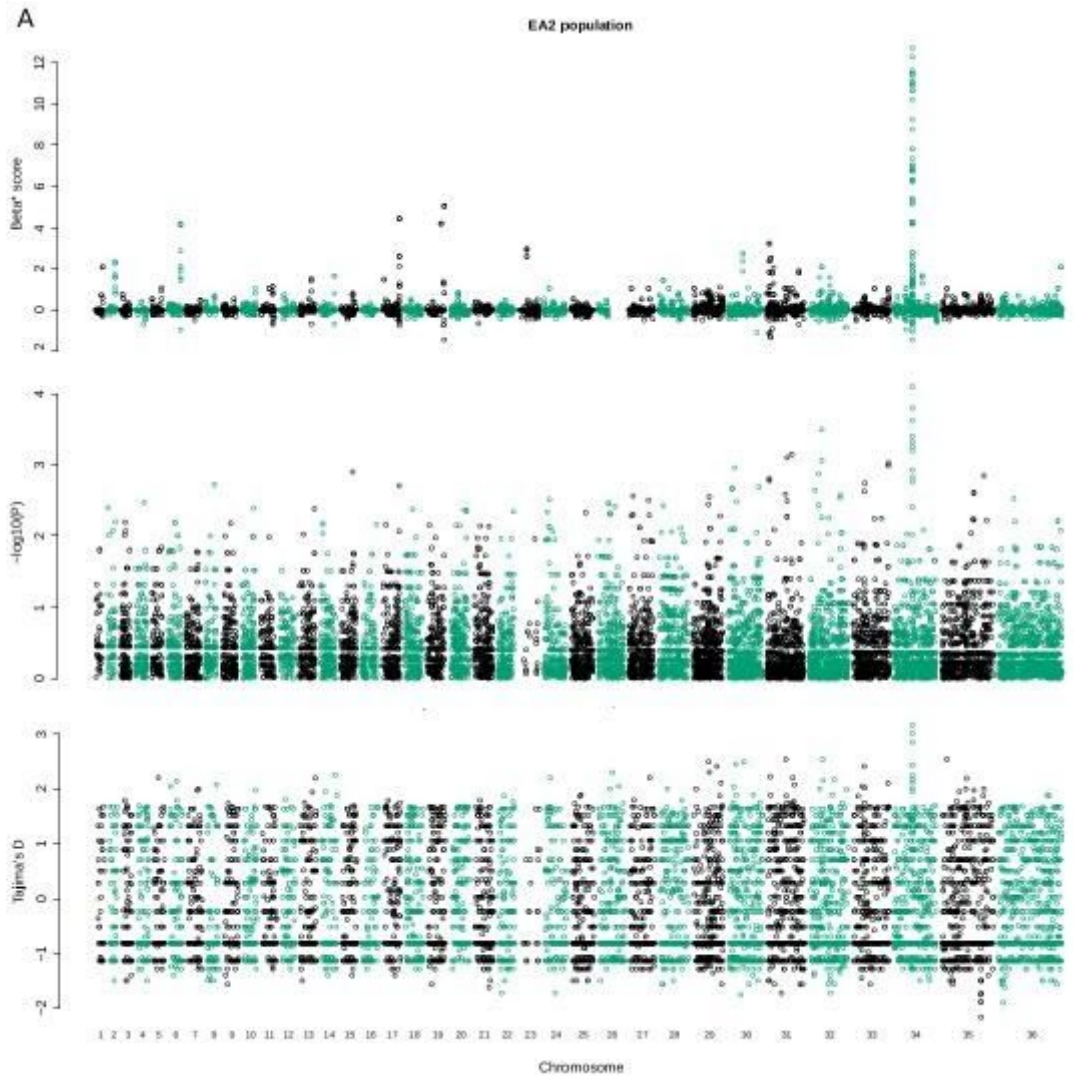
Supplementary Figure 7. Search strategy employed in balancing selection search. Full description in **Supplementary Text 1**.



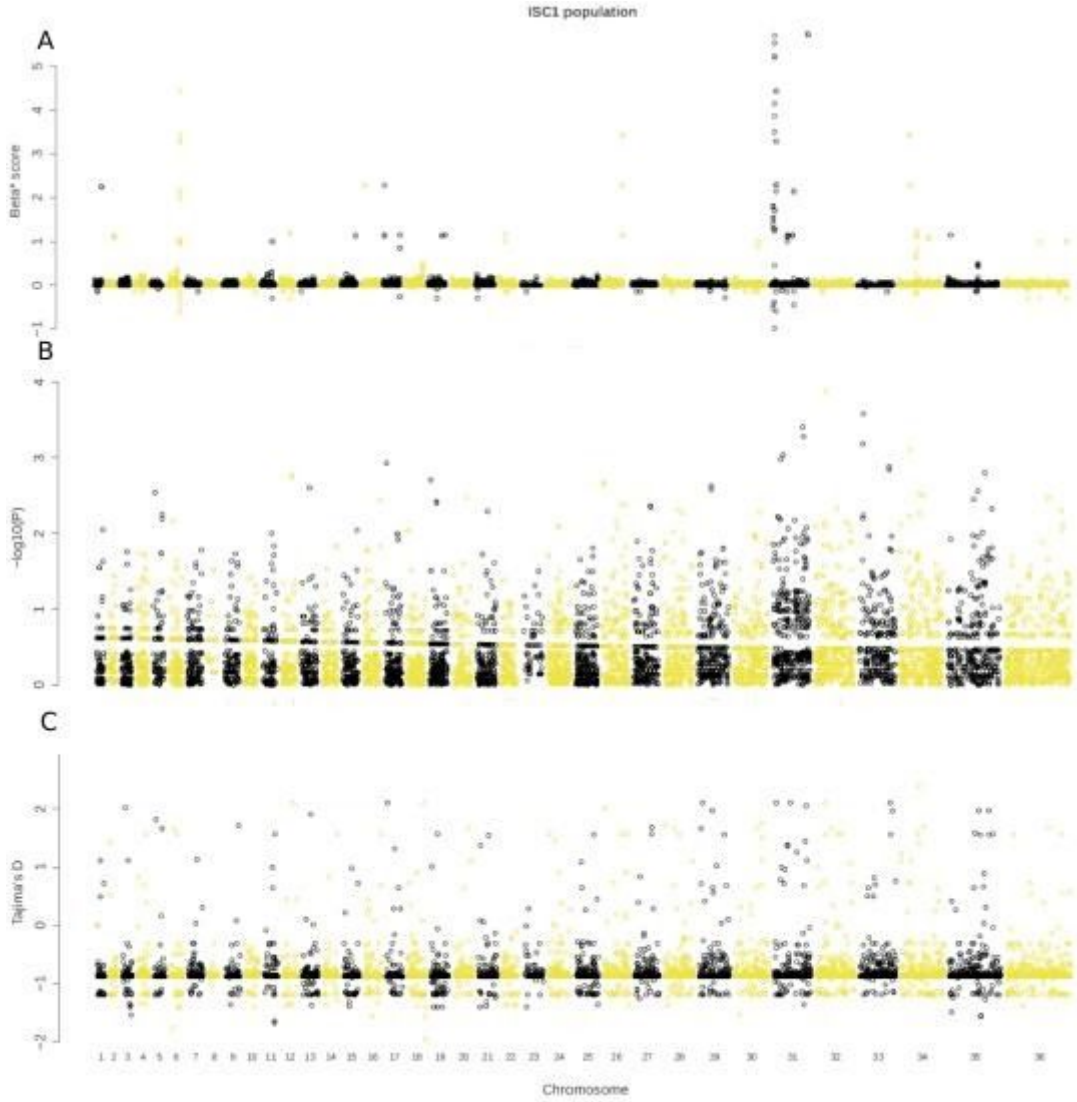
Supplementary Figure 8. Comparison of balancing selection tests for population BM. Panels from top: Betascan2, NCD2, Tajima's *D*.



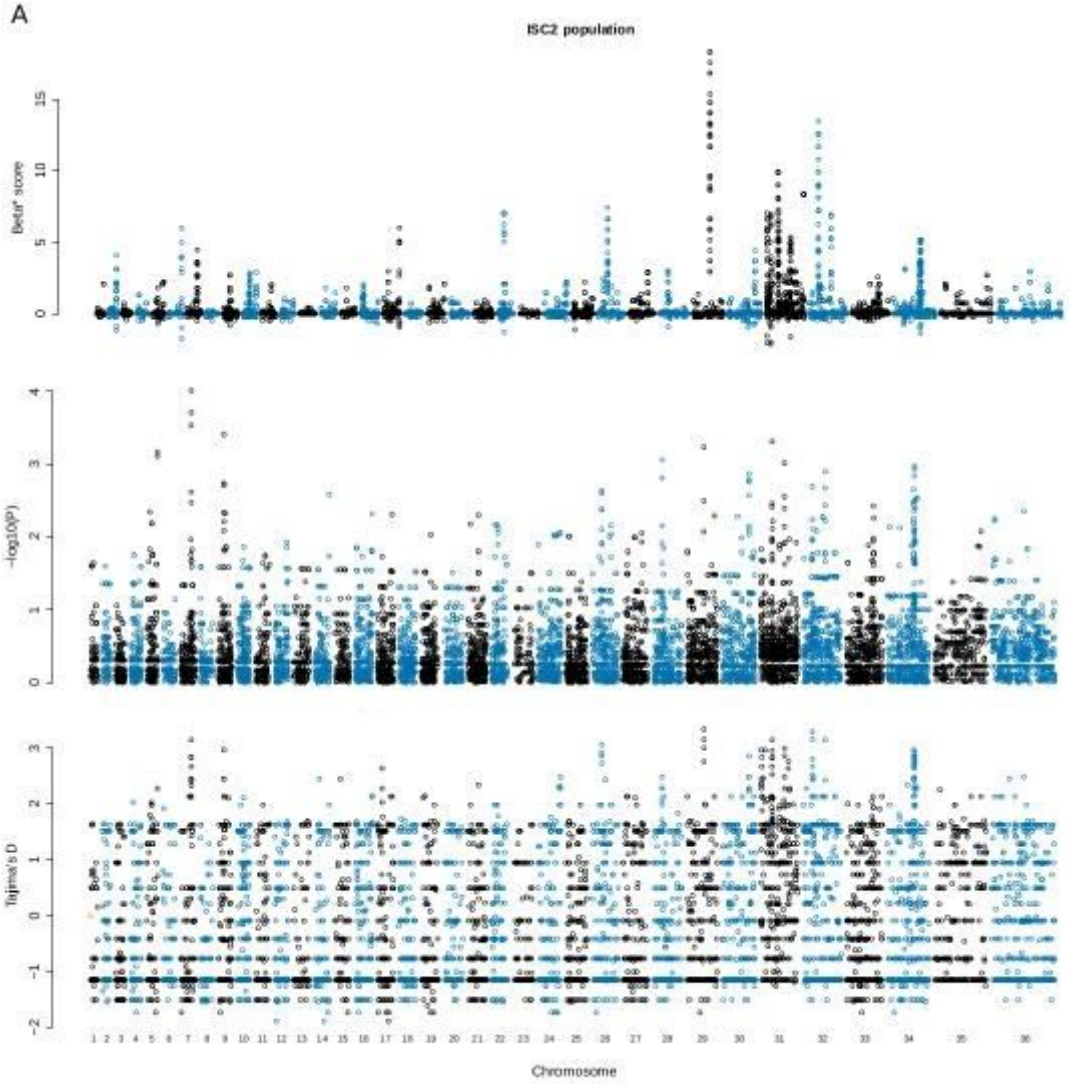
Supplementary Figure 9. Comparison of balancing selection tests for population EA1. Panels from top: Betascan2, NCD2, Tajima's *D*.



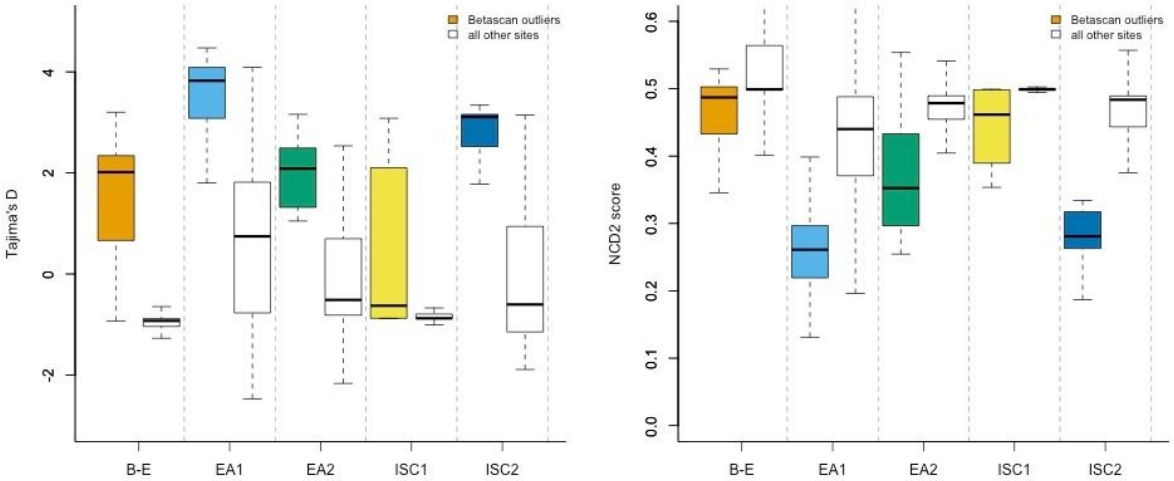
Supplementary Figure 10. Comparison of balancing selection tests for population EA2. Panels from top: Betascan2, NCD2, Tajima's *D*.



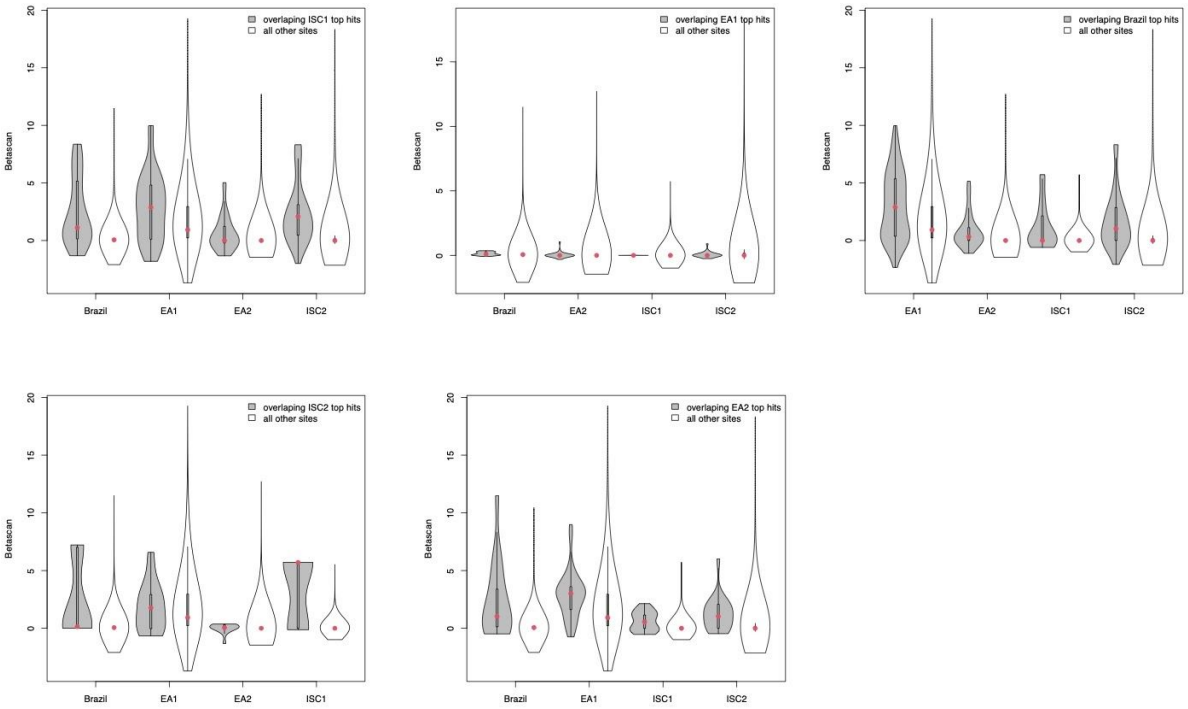
Supplementary Figure 11. Comparison of balancing selection tests for population ISC1. Panels from top: Betascan2, NCD2, Tajima's *D*.



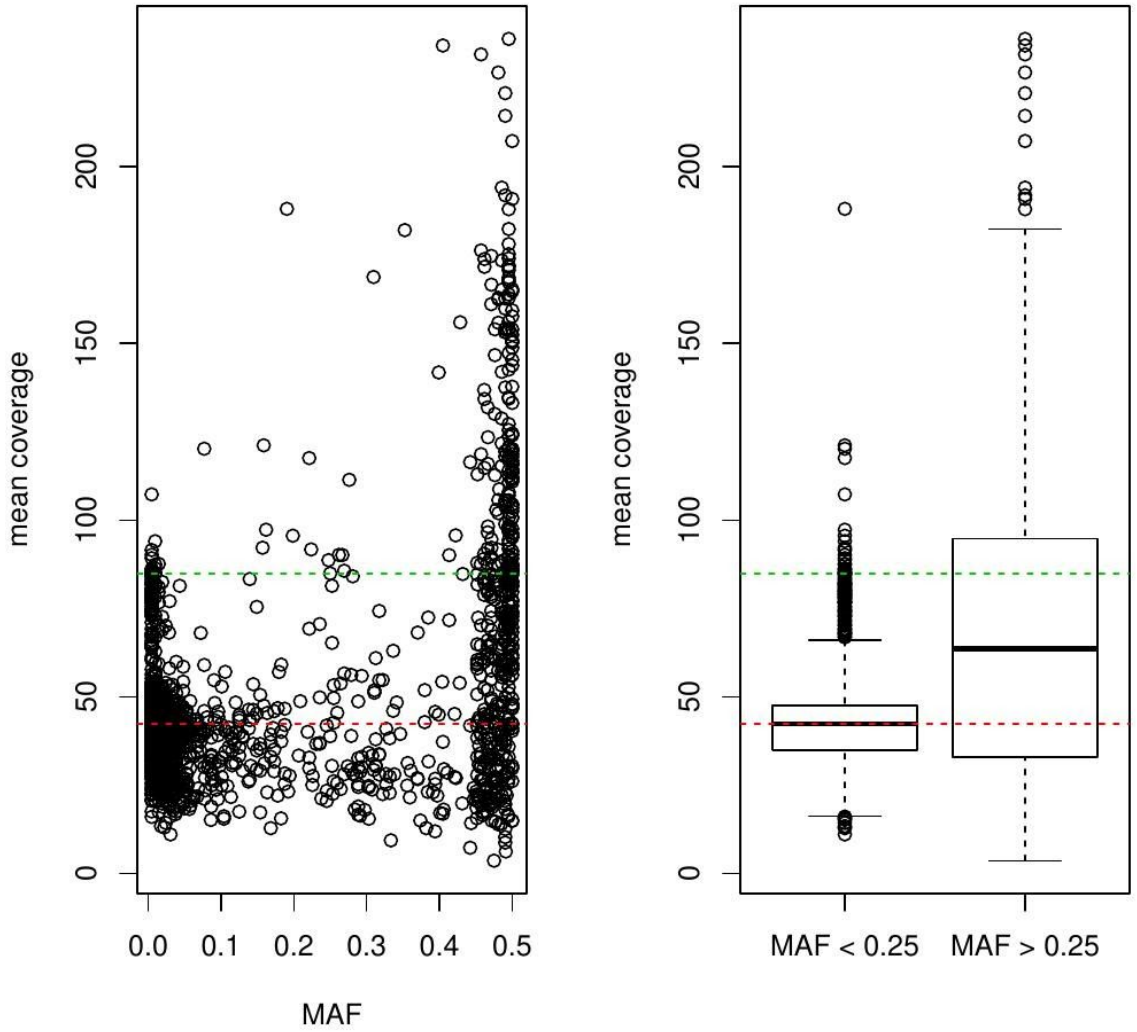
Supplementary Figure 12. Comparison of balancing selection tests for population ISC2. Panels from top: Betascan2, NCD2, Tajima's *D*.



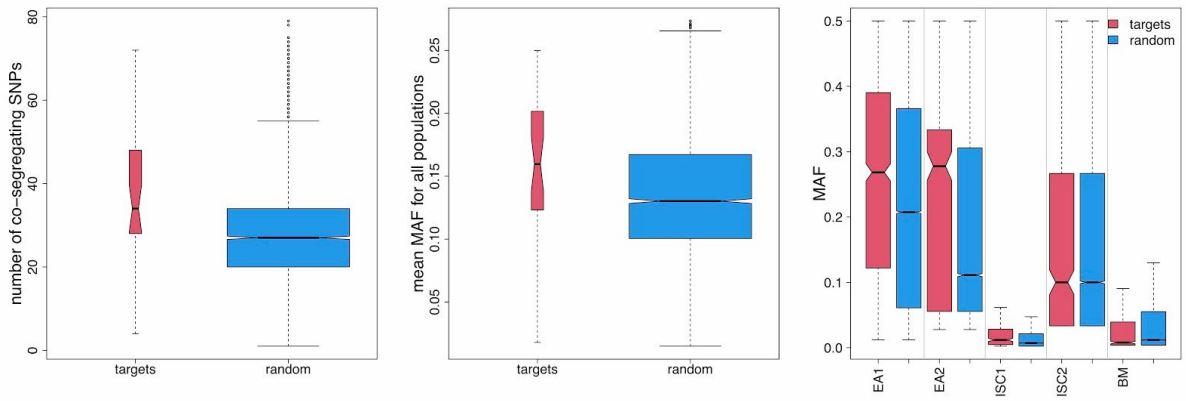
Supplementary Figure 13. Consistency between tests for balancing selection. In each population, *Betascan* outliers in the 99th percentile of the $\beta^{(2)}$ statistic (Siewert and Voight, 2020) were enriched for high Tajima's *D* (left panel) values, and low *NCD2* scores (right panel). Note that *NCD2* scores are expected to be *low* for sites subject to BS (Bitarello *et al.*, 2018), while Tajima's *D* will return *higher* values for BS sites.



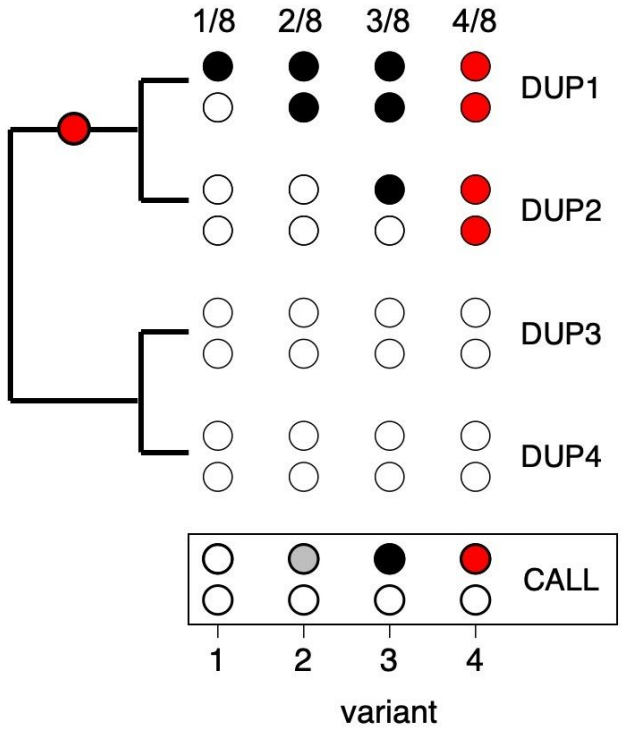
Supplementary Figure 14. Betascan hits were generally predictive of higher scores in other populations.



Supplementary Figure 15. Correlation between MAF and read coverage at SNP sites. Modelling showed that duplications resulted in a systematic bias against calling rare alleles. We removed any SNP/indel sites where the mean variant coverage within the ADMIXTURE-defined population was ≥1.5x larger than the median coverage (corresponding to triploid sites in a generally diploid chromosome), or ≥1.25x larger than the median coverage for chromosome 31 (corresponding to tetraploid sites in a generally triploid chromosome). We also removed sites where coverage was highly variable, by excluding sites in the upper 5th percentile of the standard deviation.



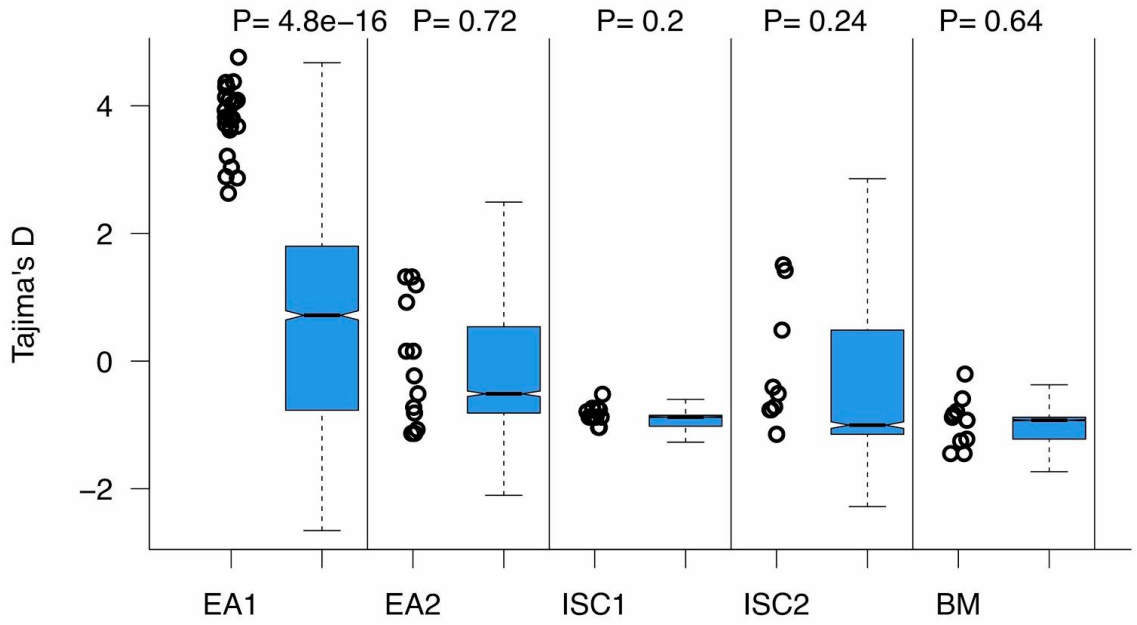
Supplementary Figure 16. Candidates for BS are enriched for high MAF co-segregating SNPs. Left panel; for the 33 candidate BS genes, we show the number of variants (SNPs/indels) in 130 kb proximal genomic windows that segregate between more than one population, compared to 330 length- and chromosome-matched genomic windows chosen at random. These distributions are significantly different (Mann-Whitney P = 0.0001118). Middle panel, these variants have higher mean minor allele frequencies (MAF, means calculated between all populations) compared to matched random controls (Mann-Whitney P = 0.001374). Right panel, comparison of the MAF of co-segregating SNPs between populations indicates that target MAF is significantly elevated in the EAst African populations EA1 (where most targets were discovered) and EA2.



Supplementary Figure 17. Systematic under-calling of rare alleles in duplicated genes. We consider variant calling in one strain, in a gene that contains four copies, and four variant positions. Duplications have arisen with some arbitrary phylogeny. In a diploid chromosome,

there are eight haplotypes. The likely consensus variant call by a variant-caller that assumes a single diploid site is shown in the box below. *Loss of rare alleles.* Rare alleles, such as singletons that have arisen in one copy (variant 1) will be represented by ⅛th of the reads, so are unlikely to be called (two open dots in consensus call). More common alleles, such as those homozygous in one duplication (variant 2), may or not be called, depending on details of variant calling, and stochastic read counts for the alleles (grey dot in consensus call). More common variants in duplications are increasingly likely to be called as heterozygous sites in the consensus call (variant 3). *Appearance of balanced alleles.* In the rare case that a variant becomes fixed in one clade of the duplication phylogeny (variant 4, and red dot in phylogeny), it will be called as a heterozygous site. Depending on the age of the duplication, such a site may be called as a heterozygous site in many/all strains. This will produce the appearance of a balanced polymorphism.

Supplementary Figure 18. Vignettes of all candidate genes. Panels from top: Betascan2, NCD2, Tajima's *D*, nucleotide diversity, MAF and coverage. Note to reviewers: this figure has been uploaded to figshare.



Supplementary Figure 19. For the 24 candidate BS genes that were discovered in the EA1 population, we show the Tajima *D* values in all populations (open circles), relative to the genome-wide distribution (blue boxes). All p-values are Mann-Whitney tests of candidate BS genes compared to all other genes.

Supplementary Tables

Supplementary Table 1. Tajima's *D* values for all genes in all populations.

Supplementary Table 2. Nucleotide diversity values for all genes in all populations.

Supplementary Table 3. Gene diversity set enrichments.

Supplementary Table 4. Betascan/NCD2 enrichments.
Supplementary Table 5. Candidate gene set and justifications.
Supplementary Table 6. Candidate gene set enrichments.
Supplementary Table 7. FST values for each population.
Supplementary Table 8. Annotated variants in candidate genes.
Supplementary Table 9. List of isolates (note to reviewers: isolates sequenced in this study have temporary sample accession numbers assigned by NCBI. We will update this before publication and when data have been released).