Infection, Genetics and Evolution 20 (2013) 188–196

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

# MLSTest: Novel software for multi-locus sequence data analysis in eukaryotic organisms

Nicolás Tomasini [a,*], Juan J. Lauthier [a], Martin S. Llewellyn [b], Patricio Diosque [a]

[a] Unidad de Epidemiología Molecular (UEM), Instituto de Patología Experimental, Universidad Nacional de Salta-CONICET, Av. Bolivia 5150, CP4400 Salta, Argentina
[b] Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, University of Wales, Bangor, Deiniol Road, Bangor, Gwynedd LL572UW, UK

ARTICLE INFO

*Article history:*
Received 31 May 2013
Received in revised form 29 August 2013
Accepted 31 August 2013
Available online 8 September 2013

*Keywords:*
MLSTest
Software
Multi-locus sequence typing
MLST
Incongruence
Concatenation

ABSTRACT

Multi-locus sequence typing (MLST) is a frequently used genotyping method whose goal is the unambiguous assignment of microorganisms to genetic clusters. MLST typically involves analysis of DNA sequence results generated from several house-keeping gene loci. MLST remains the gold standard for molecular typing of many bacterial pathogens. Eukaryotic pathogens have also been the subject of MLST, however, few tools are available to deal with diploid sequence data. Here we present novel software for MLST data analysis tailored towards diploid Eukaryotes: MLSTest. This software meets various methods used in MLST and introduces some novel methodologies for the evaluation of the data set. In addition to construction of allelic profiles and basic clustering analysis, the MLSTest looks for network structures that suggest genetic exchange in BURST graphs. Additionally, it uses several simple methods for tree construction with the advantage of managing heterozygous or three-state sites. Additionally, the software analyses whether concatenation of fragments from different genes is suitable for the data set using different tests (bionj-incongruence length difference test, Templeton test). It evaluates how the incongruence is distributed across the tree using a variation of the localized incongruence length difference test based on a modified neighbour joining algorithm. We tested the last method in simulated datasets. We showed that is conservative (adequate type I error rate) and moderately to highly powerful as well as useful to localize incongruences in two bacterial and two eukaryotic MLST datasets. MLSTest was also designed for developing MLST schemes. It thus has tools to optimize locus combinations and to reduce the number of targets required for typing. MLSTest also analyses whether the discriminatory power of the typing scheme is increased by including more loci. We evaluated the software over simulated and real datasets from bacterial and eukaryotic microorganisms. The software is freely available at http://www.i-pe.unsa.edu.ar/software.



## 1. Introduction

Multi-locus Sequence Typing (MLST)(Maiden et al., 1998) is a method originally developed for the analysis of bacterial population genetic diversity (Dingle et al., 2001; Enright et al., 2000; Nallapareddy et al., 2002), and later used for typing diploid organisms such *as Candida* spp. (Bougnoux et al., 2002; Odds, 2010; Robles et al., 2004), *Aspergillus fumigatus* (Bain et al., 2007), the *Fusarium solani* species complex (Debourgogne et al., 2010), *Leishmania spp.* (Mauricio et al., 2006) and *Trypanosoma cruzi* (Yeo et al., 2011; Lauthier et al., 2012), among others. MLST evaluates genetic variation at internal fragments of housekeeping genes. Different sequences at each locus are considered as distinct alleles. The combination of alleles of several loci generates an allelic profile for each strain (also called Sequence type or ST). Analyses based on allelic profiles have the advantage of buffering the effect of recombination observed in many micro-organism species. This is because multiple differences between two strains at a single locus may be due to a single recombination event and not due to multiple mutations. Consequently, there is no weighting to reflect the number of differences between different alleles. Once defined, identity between allelic profiles can be represented in a distance matrix and analysed by classic agglomerative methods of clustering like UPGMA or neighbour Joining. Another method widely used for Bacterial MLST is BURST (**B**ased **U**pon **R**elated **S**equence **T**ypes) and its derivatives, eBURST (Feil et al., 2004) and goeBURST (Francisco et al., 2009) (which not only determine clonal complexes but they also infer the relationships between strains within these clusters). These methods are fast and suitable to analyse large population datasets relevant in a clinical/epidemiological context.

MLST does not necessarily require a phylogenetic component. Indeed, the purpose of MLST is to provide a system of strain

* Corresponding author. Address: IPE – Fac de Ciencias de la Salud., UNSa., Av Bolivia 5150, Argentina. Tel./fax: +54 387 4255333.
*E-mail address:* nicotomasini@yahoo.com.ar (N. Tomasini).

1567-1348/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.08.029

identification easily interoperable between centres within a simple, cluster-based, analytical framework. Deeper phylogenetic analyses: Bayesian methods; coalescent and maximum likelihood tests; can also be addressed with raw sequence data. However, the real usefulness of MLST – to provide rapid, robust and easily databased strain information to health professionals – requires more accessible tools.

Most classical tools used in MLST were developed for haploid organisms (e.g. bacteria). The application of these same tools to diploid or even aneuploid organisms is often problematic. A significant component of such error arises from the common practice of ignoring heterozygous or multi-state sites, which are considered as ambiguous information. However analytical difficulties associated with MLST are several-fold and not all associated with ploidy: first, allelic profiles summarize multiple changes in a sequence as a single difference, which results in loss of resolution. For example, in *Candida albicans* (Tavanti et al., 2005a) approximately 45% of the strains could not be assigned to any cluster by eBURST analysis. Second, congruence is not usually evaluated when MLST data are analysed. Congruence among loci means an agreement between phylogenetic information from different loci. This agreement is achieved when different loci have a shared history. Incongruence implies that loci have different evolutionary histories, which is mainly caused by genetic exchange. This is important because artefactual clusters may be obtained in MLST analyses if there is incongruence between loci. Third, during the development of an MLST scheme loci numbers are required to be optimally small to save on cost and labour (Bougnoux et al., 2003; Lauthier et al., 2012). Similarly, optimal consensus MLST schemes need to be objectively derived when $\geqslant 2$ are present for the same organism (Lauthier et al., 2012; Yeo et al., 2011; Bougnoux et al.; 2003; Debourgogne et al., 2012; Ahmed et al., 2011; Boonsilp et al., 2013).

In this manuscript we introduce MLSTest: novel software to assist with the development and analysis of MLST schemes. Our aim is to improve the efficiency and efficacy of these schemes, especially in the context of diploid or non-haploid organisms, while adhering to the straightforward clustering-based approach that underpins the success of MLST as a pathogen typing strategy.

## 2. Software description

### 2.1. A general overview of MLSTest

*Sequence manipulations*: Multiple FASTA sequences can be loaded simultaneously. The user can view the alignment/polymorphic sites/codons and/or amino acids interchangeably. Multiple options to modify, concatenate and export alignments into different file formats are available.

Allele calling: The software can assign alleles, determine allelic profiles and calculate associated measures like typing efficiency, discriminatory power with its confidence interval (Severiano et al., 2011). MLSTest also has two options to handle heterozygous sites (described in Section 2.2).

*Clustering*: MLSTest implements UPGMA, Neighbour-Joining (NJ), BIO-Neighbour Joining (BIONJ), with different node support measures. MLSTest calculates consensus trees summarizing the information of individual fragment trees (consensus trees are based on branch frequency into individual loci trees). Multidimensional scaling plots can also be created from pairwise distance matrices. MLSTest is able to make basic clustering using allelic profiles with the BURST algorithm.

*Congruence*: MLSTest implements tools to analyse congruence between genetic loci, including the Incongruence length difference test (implemented using BIONJ) (Zelwer and Daubin, 2004) and the CADM test (Congruence Among Distance Matrices test) (Campbell et al., 2011) and different measures of localized incongruence, which allows identifying nodes affected by incongruence into the concatenation based tree (described in Section 2.3). Additionally, the level of congruence among different fragments is useful to analyse the genetic structure degree of populations. In this regard, datasets with strong congruence among fragments implies a strong genetic structure. In addition, MLSTest can generate a BURST modified diagram that represents all recombination events within a dataset (described in Section 2.4).

*MLST scheme development*: The software includes several tools to screen different combinations of loci in order to: 1-determine the minimum number of fragments required for obtaining the maximum discriminatory power (described in Section 2.5.1) and 2-the best combination of fragments according different criteria as discriminatory power, identification of certain predefined groups and cluster supports (described in Section 2.5.1). Finally, MLSTest had a graph viewer that allows viewing trees and BURST diagrams, to edit and export them in different formats.

*Environment*: MLSTest is written in Visual Basic and runs in Microsoft Windows.

### 2.2. Managing diploid sequences

MLSTest has two options for managing heterozygous sites in the analyses:

*Average States*: The distance between two bases is calculated as the mean of distances among all possible resolutions of the heterozygosis (see example in Fig. 1). Three-state bases are also allowed.

*SNP duplication and heterozygosis resolution* (Tavanti et al, 2005b): Invariable sites are removed from the alignment. Then, polymorphic sites are duplicated and heterozygosities are resolved (see example in Fig. 1).

### 2.3. Testing for incongruence

#### 2.3.1. Overall incongruence and localized incongruence

Several tests to detect significant incongruence between genetic loci in a dataset have been proposed (Planet, 2006). The classical Incongruence Length Difference (ILD) test implementing the maximum parsimony method (Farris et al., 1994; 1995) is frequently used (Planet, 2006). MLSTest implements a variant of the ILD test called BIONJ-ILD (Zelwer and Daubin, 2004) which is a variant.

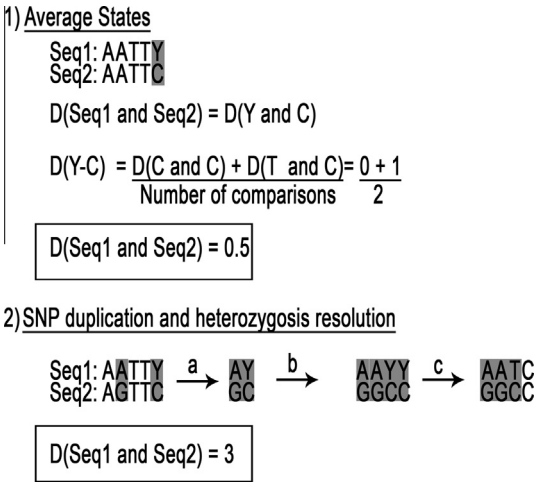


**Fig. 1.** Examples of different methods for calculating distance between sequences containing heterozygous sites. Polymorphic sites are highlighted. D, distance; a, invariable sites are deleted; b, polymorphic sites are duplicated; c, double heterozygous bases are resolved.

The BIONJ-ILD test uses BIO-neighbour Joining method instead of parsimony. The null hypothesis assumes congruent branching patterns between different genetic loci. BIONJ-ILD analyses whether contradictory phylogenetic signals are distributed at random among different gene fragments (random homoplasy) or whether they are concentrated in certain fragments (incongruent loci). ILD is calculated as the sum of the branch lengths in the tree based on concatenation minus the sum of the branch lengths of all trees of individual fragments. When there is full congruence, the ILD = 0. However, when there is incongruence, ILD > 0. Statistical significance of the ILD value is evaluated using a permutation test. $P < 0.05$ suggests that at least one locus is incongruent but does not indicate how this incongruence is distributed through the tree. Sometimes, incongruence is located in a few branches and it would be not necessary to doubt about the reliability of the entire tree. Furthermore, MLSTest is able to show the number of individual trees that are topologically incompatible with a certain node into the concatenated tree. In this way incongruence can be localized to clades in the tree.

#### 2.3.2. Localized incongruence length difference

In order to identify the branches in the concatenated sequence tree that have a locus or loci with statistically significant incongruence, we developed a novel algorithm based on the localized incongruence length difference (LILD) for parsimony (Thornton and DeSalle, 2000). We implement LILD using a neighbour joining method. First, we enforce a node *x* (a node in the tree based on concatenation) in the tree specific for each locus as described into the appendix, Section 5.1.1. Second, the sum of the length differences between constrained and optimal topology for each locus is calculated. This value represents the nj-Localized Incongruence Length Difference (nj-LILD). nj-LILD suggests the incongruence level for the analysed branch but does not indicate the statistical significance of the value. Statistical significance may be evaluated using two different methods, a permutation test or a modified templeton test (see appendix, Section 5.1.2). The null hypothesis for the former is the random distribution – across all loci – of phylogenetic signals that contradicts the branch under analysis. A significant *p* value means at least one fragment is incongruent with the clustering proposed by the analysed branch. By contrast, for a certain branch into the tree based on concatenation, the templeton test says whether phylogenetic signals of loci that are topologically incompatible with it are statistically well supported.

### 2.4. Testing for BURST suitability

BURST methods are based on a simple model of clonal bacterial evolution at which an ST increases in frequency and then diversify. Clusters or clonal complexes are identified by these algorithms based on an arbitrary group definition. Every ST within a group has an arbitrary minimum number of identical alleles in common with at least one other ST in the group. Usually, the group definition is at least *n*-1 shared alleles (where *n* is the number of loci). Other group definitions are possible (i.e., *n*-2 or *n*-3). MLSTest includes two useful functions to assess BURST analysis suitability. First, a BURST over all group definitions will allow the user to observe the number of clusters and singletons conformed under all possible group definitions. In addition, a dendrogram is constructed showing the relationships among groups over all group definitions. The algorithm for the dendrogram involves hierarchical clustering from stringent (*n*-1) to more relaxed (e.g. *n*-7) group definitions. The procedure is repeated until all STs are joined to the tree. The goal of the dendrogram is to show the distribution of groups in a range of group definitions from 0 (root) to *n* (leaves) shared alleles.

Additionally, the relationships among STs in a BURST group (which was defined by *n*-1 shared alleles) may be represented as a set of connections among STs (these connection are also termed Single Locus Variants or SLVs). The eBURST (Feil et al., 2004) and goe BURST graphs (Francisco et al., 2009) display only connections following certain criteria around an assumption of predominantly clonal evolution. However, eukaryotic organisms may not always fit a criterion of clonal evolution. In fact, genetic exchange may be observed by SLV connections that form networks (not represented in standard eBURST or goeBURST graphs) without internal SLV connections. MLSTest is able to identify network structures in BURST groups to assist with the identification of recombinant strains.

### 2.5. Selecting an optimal typing scheme

MLSTest includes several options to optimize MLST schemes. That is, to reduce the number of loci used for typing whilst maximising discriminatory power.

#### 2.5.1. Optimum number of loci

Genotypic diversity (GD) which is defined here in a simple way as the number of different genotypes is estimated for every possible combination of 2 to *n*-1 loci. The results are presented in a table showing the number of loci, the number of possible combinations and the minimum, mean and maximum Genotypic Diversity (GD). Maximum GD assists with the selection of the minimum locus number. Mean GD can inform the researcher whether adding a further locus to the full scheme will increase the observed GD. When mean GD reaches an asymptote, adding further loci is sub-optimal (Arnaud-Haond et al., 2005).

#### 2.5.2. Testing all possible combinations

For a determined number of loci (selected by the user), all possible combinations may be evaluated by concatenation and NJ method. The number of possible combinations is determined by the combinatory formula:

$$Ncomb(n,x) = n!/[x!(n-x)!] \quad (1)$$

where *n* is the number of loci into the dataset and *x* is the number of loci in the reduced scheme. Three criteria for scheme selection are considered under this option. First, the software determines observed genotypic diversity (GD) to select those schemes with best discriminatory power. Second, cluster/typing unit monophyly is evaluated based on the selected markers. Monophyly indicates whether the selected loci combination is able to identify predefined groups (i.e., known clades identified by genome sequencing/other markers). Third, bootstrap cluster support is evaluated in order to select combinations of loci that result in robust cluster assignment. The software gives flexibility in criteria implementation and several options are available for the user. Bootstrapping and alternative branch support methods (fast-bootstrapping and clade significance) are detailed in appendix (Section 5.1.3).

## 3. Results and discussion

We evaluated several features of MLSTest on real or simulated datasets. First, we analysed the results of different ways of managing diploid sequences in a real dataset of *Leishmania donovani complex*. Second, we evaluated the performance of the novel nj-LILD test to detect incongruence and recombination. Two bacterial and two eukaryotic MLST datasets are analysed as examples. In addition, we analysed a dataset of *Aspergillus fumigatus* as an example of network structures in a BURST diagram. Lastly, we evaluated

alternative methods of bootstrap in order to calculate branch support when multiple combinations of loci are analysed.

### 3.1. Managing diploid sequences in the Leishmania donovani complex

In the past single nucleotide polymorphism (SNP) duplication has been proposed and used in order to resolve ambiguous bases in alignments (Chen et al., 2006; Odds et al., 2007; Tavanti et al., 2005a; Tavanti et al., 2005b; Yeo et al., 2011), however this strategy has a key flaw: the resolution of two different heterozygoses in the same position for two different samples may still be ambiguous. For example, the distance between heterozygoses Y and K has two possible resolutions: Y/K = TC/TG, distance = 1 or Y/K = CT/TG, distance = 2, which are not currently accounted for. The same ambiguous resolution occurs with other pairs of heterozygous sites: W/M, W/R, W/Y, W/K, M/R, M/Y, M/S, R/K, R/S and Y/S (see Cornish-Bowden (1985) for further detail about heterozygosis coding letters). Additionally, bootstrap values are sometimes different to those obtained by using average states method (See Section 2.2). Differences arise because SNP duplication deletes constant sites (which are not informative for tree topology but they are for bootstrap significance), thus bootstrap value are unusually high in datasets with few polymorphic sites. To evaluate the different approaches to diploid data treatment we assessed support for 22 branches in a tree of *L. donovani* complex (see appendix Section 5.3 for further details of the dataset). We observed significantly lower bootstrap values using the average states method when compared to the bootstrap values obtained by SNP duplication ($p < 0.0001$ for a Wilcoxon sum rank test, data not shown). Differences between bootstrap values obtained by the two methods were inflated by $\geqslant 10$ percentage points in 40% of the branches for 1000 bootstrap replications, and we propose the average states method as the superior methodology.

### 3.2. nj-LILD test error rates in simulated datasets

Analysing congruence among loci is a first step to determine whether concatenation is a reliable option to define relationships among strains or isolates. Additionally this approach can define the level of recombination in a dataset. Table 1 shows the results of the nj-LILD test with a permutation test and nj-LILD with a Templeton test for simulated congruent data sets (based on identical tree topologies showed in Fig. 2A) in different evolutionary conditions (see appendix, Section 5.2). The probability of rejecting the true hypothesis ($H_0$) of congruence (type I error rate) was thus evaluated. To be acceptable, the type I error rate should be lower or around the level of significance used in the test (in this case alpha = 5%). The two tests reject the congruence hypothesis less or around 5% of cases. Consequently, the type I error rates were within the acceptable range irrespective of topological symmetry, rate variation between branches and substitution rates. Type I error rates were also acceptable for a complex tree based on simulations of *C. glabrata* FSK gene tree (data not shown).

**Table 1**
Type I error rate for nj-LILD tests in simulated datasets.

| Substitution rate | ASYM | | SYM | | |
|---|---|---|---|---|---|
| | CER | VER | CER | VER | 4xCER + 3xVER |
| 0.01 | 4.4/3.4 | 6.4/4.3 | 3.8/<0.1 | 5.1/1.3 | 6.5/0.5 |
| 0.002 | 2.1/<0.1 | <0.1/<0.1 | 2.3/<0.1 | 1.7/<0.1 | 2.4/<0.1 |

The values represent the percentage of the 1000 simulations for which the true hypothesis of congruence is rejected for nj-LILD using permutation test (before slash) and templeton test (after slash). Results are given in relation to the tree topologies (ASYM, asymmetric, SYM, symmetric), to the substitution rate, to the variability of evolutionary rates between branches (CER, constant evolutionary rate; VER, variable evolutionary rate).

We additionally evaluated the power of the test (probability of correctly rejecting null hypothesis). Table 2 shows power of the test for 1000 simulated datasets of seven loci where 1 or 2 loci had evolved with an event of horizontal gene transfer (which generates incongruence) as is showed in Fig. 2. We observed that the permutation test is powerful for substitution rates of 0.01 changes per site per branch, consistent with evolutionary rates at housekeeping loci in *Neisseria meningitidis* or *Haemophilus influenzae* datasets (data not shown). Power was moderate to low at lower substitution rates and correlated with the number of incongruent loci. Additionally, the power was variable for different branches. We observed that the test was more powerful in branches that contain the donor strain in the lateral gen transfer event than in the receptor branch (data not shown).

In our approach, multiple branches are tested per tree. Thus, a statistical correction for multiple comparisons (like the Bonferroni correction) is required in order to avoid an increase in type I error rate. However, by decreasing the alpha cut-off, Bonferroni correction reduces the power. To test Bonferroni's impact on power, we simulated a complex phylogeny based on the FSK locus for 20 STs of *C. glabrata* dataset. We randomly shuffled the sequences corresponding to a determined clade of 8 STs and then analysed whether nj-LILD tests were able to detect significant incongruence within the clade after a Bonferroni correction. Even with a Bonferroni correction, we were still able to detect within clade incongruence in 72% of simulations, using permutations to evaluate nj-LILD value. By comparison to permutation test, the Templeton test had lower power in almost all cases. We did not test the full range of parameters in simulations used to evaluate bionj-ILD test (like gamma distribution parameters) (Zelwer and Daubin, 2004) but we consider that the range of used parameters is according to common MLST datasets (datasets were simulated using likelihood parameters of real datasets).

### 3.3. Analysing genetic structure and recombination within natural datasets

We applied different tests to two prokaryotic (*N. meningitidis* and *H. influenzae*) and two eukaryotic datasets (*Candida glabrata* and *Trypanosoma cruzi*) (see appendix, Section 5.3) in order to evaluate population genetic structure. A summary of levels of incongruence for different datasets is detailed in Table 3.

*N. meningitidis*: *The N. meningitidis* tree showed significant overall incongruence based on bionj-ILD test ($p < 0.001$). We also observed high topological incongruence. Permutation and Templeton based nj-LILD tests were highly significant after Bonferroni correction for those branches with high topological incongruence (Table 3). As such, we detected high incongruence between trees based on concatenated and individual loci. Bootstrap values were >80% in 53% (9/17) of the branches despite the incongruence among loci, suggesting that this support measure alone is not sufficient to define clusters in a multilocus analysis. Additionally, an extended-majority rule consensus tree of single locus topologies was distinct to that derived from concatenated loci, as one would expect in a frequently recombining population. Overall our analysis is consistent previous work, i.e., that the population of *N. meningitidis* is a diverse and freely recombining group of different genotypes, from which emerge certain well-defined clusters (Feil et al., 2000; 2001; Perez-Losada et al., 2006).

*H. influenzae*: The *H. influenzae* dataset also had a significant bionj-ILD test ($p < 0.001$). However, the mean topological incongruence and the proportion of branches with high topological incongruence were lower than *N. meningitidis* (Table 3). All

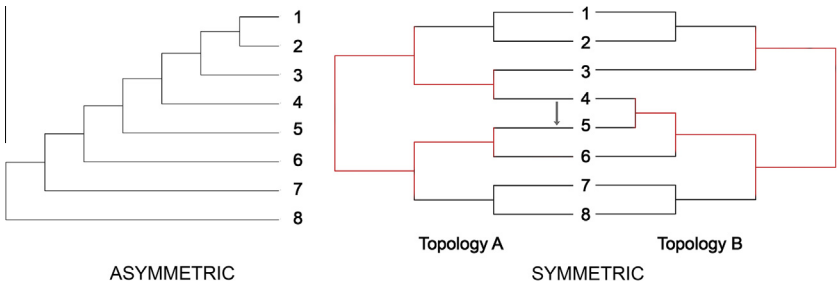


**Fig. 2.** Topologies along which the nucleotide sequences were evolved in order to evaluate nj-LILD tests. These trees follow a molecular clock (Constant evolutionary rate). Topologies with Variable Evolutionary Rate among branches were made based on these and alternating short and long branch lengths with a length ratio of three as is described in Zelwer and Daubin (2004). Topology B represents a recent lateral gene transfer involving taxa 4 and 5 (showed by an arrow in topology A). Highlighted in red are showed branches incongruent between the two topologies. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
Power of nj-LILD tests in simulated datasets.

| Substitution rate | 6xTopology A + 1xTopology B | | 5xTopology A + 2xTopology B | |
|---|---|---|---|---|
| | CER | VER | CER | VER |
| 0.01 | 96.4/71.8 | 92.1/72.1 | 99.5/99.7 | 98.5/97.4 |
| 0.002 | 42.7 (18–59)[1]/6.6 (0–19) | 42.6 (18–59)/6.6 (0-19) | 62.5 (33–82)/16.2 (1–51) | 57.9 (33–82)/15.5 (1–48) |

The values represent the percentage of the 1000 simulations for which the false hypothesis of congruence is rejected for nj-LILD using permutation test (before slash) and templeton test (after slash). Results are given in relation to the number of incongruent trees, to the substitution rate, to the variability of evolutionary rates between branches (CER, constant evolutionary rate; VER, variable evolutionary rate).
[1] Represent the range of power for different branches.

branches with high topological incongruence were significant after Bonferroni correction for nj-LILD using Templeton and permutation tests. Seven branches (41%) had lower topological incongruence (2 or less trees) but permutation nj-LILD test was still significant for 6 out of these branches. Here, it is important to note that statistically significant incongruence does not always mean incorrect clustering (Hipp et al., 2004). In this case it may mean that at least one locus probably had different evolutionary history from the rest. Then, it is important to analyse after nj-LILD test how many and which loci are implicated. We used here encapsulated strains of *H. influenzae* which are classified in different serotypes. Our results for this reduced dataset are in concordance with the serotype classification with the exception of one ST of the serotype *a* which clustered outside of the group as the observed by Meats (2003) who detected paraphyly for serotypes *a* and *b*. Although clonal structure of *H. influenzae* is still debated (Feil et al., 2001; Meats et al., 2003; Perez-Losada et al., 2006) and is suggested that recombination is higher in non-encapsulated strains (Meats et al., 2003), our results suggests significant lower levels of recombination in relation to *N. meningitidis* for encapsulated strains.

*C. glabrata*: The *C. glabrata* tree showed several branches (27 out of 56) with high topological incongruence (at least 4/6 loci with topological incongruence). Twenty-one of these branches were significant after Bonferroni correction for Templeton or nj-LILD permutation tests. Furthermore, we observed poor correspondence between the tree based on concatenation and the extended-majority rule tree with the exception of six clades (all of them had non-significant nj-LILD). Although no sexual or parasexual cycle is known for *C. glabrata,* recombination has been suggested (Dodgson et al., 2005). Thus our results are consistent with previous data and a level of recombination enough to generate substantial incongruence among loci.

*T. cruzi*: *T. cruzi* tree showed low to moderate topological incongruence (Table 3) with significant inconsistency in three branches; just one (5%) remained significant after Bonferroni correction for nj-LILD permutation test. The incongruence is produced by one strain (TEP6) that belong to cluster TcVI but is showed as an outlier in the tree due to an apparent Loss of Heterozygosis (LOH) in one locus as has been previously described (Lauthier et al., 2012).

**Table 3**
Different measures of localized incongruence for *Neisseria meningitidis*, *Haemophilus influenzae*, *Candida glabrata* and *Trypanosoma cruzi*.

| Dataset | *n*[a] | Branches[b] | SMTI[c] (95%CI)[d] | % high TI[e] | % nj-LILD[f] |
|---|---|---|---|---|---|
| *N. meningitidis* | 7 | 17 | 5.67 (4.69–6.37) | 88.2 | 94.1 |
| *H. influenzae* | 7 | 17 | 2.80 (2.03–3.64) | 11.8 | 82.3 |
| *C. glabrata* | 6 | 56 | 3.57 (2.87–4.20) | 48.2 | 34 |
| *T. cruzi* | 10 | 20 | 1.96 (1.26–2.87) | 10 | 5 |

[a] Number of loci in the MLST scheme.
[b] Number of analysed branches.
[c] SMTI: Standardised Mean Topological Incongruence. In order to make comparisons among datasets, the value represents the standardised number of loci that are topologically incongruent per branch in a 7 loci scheme.
[d] Confidence interval of SMTI using 500 bootstrap replications.
[e] %high TI: Percentage of branches with high topological incongruence. Topological incongruence for a branch in the concatenated loci tree was arbitrarily considered as high when the proportion of incongruent loci with it was $\geqslant 0.66$.
[f] Percentage of branches with significant p value after Bonferroni correction. The p value was calculated with 500 permutations.

In combination our data show that nj-LILD is useful to detect incongruence and to analyse uncertain clustering when is combined with other measures of incongruence; for example the number of individual gene fragments that are topologically incongruent. The combination of both measures gives outcomes for a branch:

1. Low topological incongruence with non-significant nj-LILD. If branch bootstrap support is low, it is likely that low level of polymorphic sites to define the branch.
2. Low topological incongruence with significant nj-LILD. This incongruence is produced by a single or few loci. Clustering is still evident, bust some genetic exchange has occurred.
3. High topological incongruence with not significant nj-LILD. Topological incongruence is not statistically significant, perhaps due to low number of polymorphic sites to define the cluster. Little confidence may be placed in clustering.

4. High topological incongruence with significant nj-LILD. It is probably due to incongruences among all loci or a wrong clustering due to just one or two well-supported loci which are strongly incongruent with the others. Clustering should be also carefully analysed here.

Finally, although nj-LILD implementing Templeton test had low power in simulated datasets with single point incongruences, it was useful for real datasets where incongruences are generally generated by several loci at time. Additionally, Templeton test is faster than permutation test and we propose implementation of Templeton test when larger datasets are analysed, in order to obtain an overall view and as a first step to detect branches with high levels of incongruences.

### 3.4. Detection of network structures in a BURST graph of Aspergillus fumigatus

With the previous methods we evaluated suitability of a bifurcating tree diagram to represent the relationships among strains based on alignments when there is genetic exchange. Although eBURST methods based on STs attempt to reduce erroneous clustering due to recombination events, it remains a bifurcating tree and could fail to represent relationships among STs if genetic exchange is relatively frequent. MLSTest is able to show ST relationships that do not fit the bifurcating tree model. As an example, we analysed the *A. fumigatus* dataset to look for network structures in BURST diagrams. Network structures represent direct events of recombination. Fig. 3 shows observed network structures and the corresponding eBURST graph. Fifteen out of twenty-three (65%) STs from the first eBURST group were conforming network structures. Genetic exchange may be relatively frequent for this dataset and a tree-like graph should not be useful to represent all the relationships among STs. Recombination mediated by a sexual cycle has been proposed for this organism (Paoletti et al., 2005) and it may have a certain importance in natural populations (reviewed in Varga and Toth, 2003) and could be the cause of the observed network. Finally, although analysing network structures in diploid organisms do not directly indicate genetic exchange, it suggests that the implicated strains cannot be well represented in tree-like graphs.

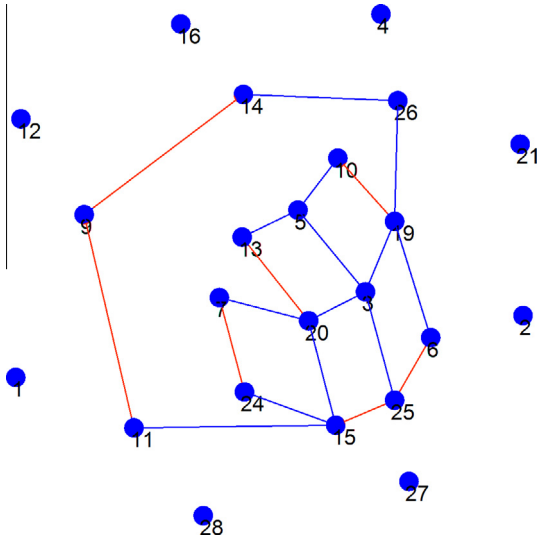


**Fig. 3.** Network structure observed in a BURST diagram of *Aspergillus fumigatus* dataset. Just connections that implicated in the network are showed. Red branches show SLVs not represented in classical eBURST graph. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### 3.5. Alternative supporting measures for combinatorial analysis

Once we have determined clusters in a dataset and these clusters are supported and they are not artefacts due to incongruence, we may want to reduce the typing scheme in order to use the minimum number of loci to determine clusters or STs. As a large

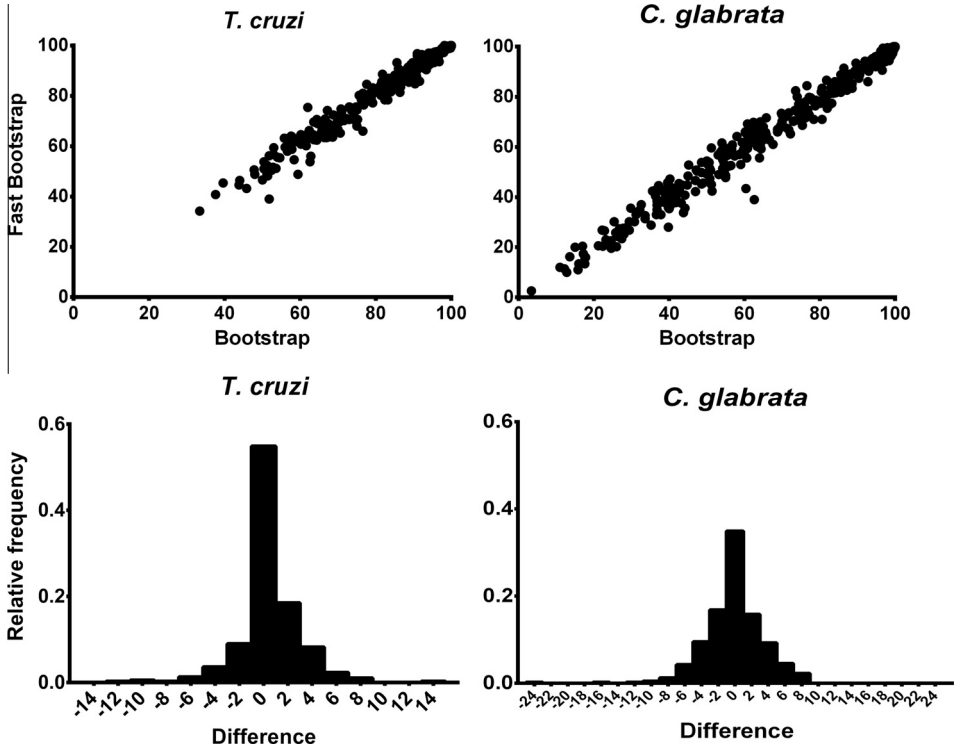


**Fig. 4.** Correlation between bootstrap and fast-bootstrap values (upper graphs); and distribution of the difference between both values (lower graphs) for branches in trees resulting of all combinations of three loci for *T. cruzi* and *C. glabrata* datasets.

**Table 4**
Correlation among bootstrap probability and 1-neighbour joining based clade significance in different real and simulated datasets.

| | *T. cruzi* | *A. fumigatus* | *C. glabrata* | *C. glabrata* simulated[a] |
|---|---|---|---|---|
| Spearman (*p* value) | 0.92 (0.000062) | 0.75 (0.00033) | 0.51 (0.00016) | 0.7 (0.0000002) |
| Branches | 20 | 24 | 53 | 56 |
| Mean bootstrap probability | 0.816 | 0.434 | 0.475 | 0.581 |
| Mean 1-njCS | 0.794 | 0.637 | 0.509 | 0.619 |

[a] The NJ tree of *C. glabrata* was used to simulate a dataset with 6 loci evolved under congruence hypothesis. Branches with length zero were discarded from the analysis.

number of combinations of loci are possible (see Eq. (1)), we propose two different alternative methods in order to obtain faster support measures for these combinations. First, a modified bootstrap algorithm (fast-bootstrap) was developed (see appendix, Section 5.1.3). We analysed the correlation between bootstrap value and the fast-bootstrap value on *T. cruzi* and *C. glabrata* datasets. We analysed all possible combinations of three loci for the seven loci dataset of *T. cruzi* and for the 6 loci dataset of *C. glabrata*. We observed a high correlation between the two measures in both datasets for 500 replications (Fig. 4, $r_{spearman} = 0.99$, $p < 0.01$ for both datasets). Ninety per cent of the branches had a difference lower than 5% and 7% between the two measures for *T. cruzi* and *C. glabrata*, respectively (Fig. 4).

In theory, since resampling probability of a site in fast-bootstrapping is different to the used by classical bootstrap, some differences could appear for branches that have no uniform support distribution among loci. For example, if a branch is supported just by one locus, but not for others, the bootstrap value could differ from the fast-bootstrap value. Additionally, some differences could be obtained if the sequence length strongly varies among loci. In this sense, and for multiple loci analyses, the fast-bootstrapping has more biological sense than classical bootstrap because sites are resampled within each locus.

Second, we implemented a Templeton derived method (neighbour joining Clade significance, njCS) to evaluate branch support. We compared the bootstrap support with the njCS in three publicly available datasets (Table 4). The Spearman coefficient was significant but variable among datasets. The lower value was obtained for *C. glabrata*. High levels of incongruence (bionj-ILD, topological incongruence, nj-LILD) and low bootstrap values for most branches were observed for this dataset. Additionally, njCS was generally more conservative (Branch support was lower) than bootstrap in this dataset. Simulating the tree of *C. glabrata* under congruence, a twofold increase into the Spearman coefficient was observed although bootstrap values were still low. This requires further evaluation and njCS should be considered carefully (or even avoided at all) in highly incongruent or poorly supported datasets. Despite this, njCS is really faster than bootstrap in several situations, particularly when the number of groups to test is lower than the number of replications of bootstrap. So, njCS would be useful as first approach in order to reduce the combinations to be analysed by bootstrapping in datasets that show low levels of character conflict.

## 4. Concluding remarks

MLSTest is new user friendly software which brings together a set of both new and pre-existing tools for multi-locus analyses (particularly for MLST approach) of haploid, and especially diploid microorganisms. This new software was designed to develop and to optimize MLST schemes and cluster assignment. It is also useful to evaluate the current methods used for clustering (eBURST and trees based on concatenation of alignments) and to analyse population structure. We also developed several novel tools to localize incongruences and to approximate branch support faster than bootstrapping. We think that MLSTest will fulfil many of the requirements of scientists who use multi-locus sequence data for molecular epidemiology of pathogenic microorganisms.

## 5. Appendix

### 5.1. Detailed methods

#### 5.1.1. A modified neighbour-joining algorithm for constrain trees

Constrained trees are useful for statistical comparison of tree topologies (the shape of the tree) at branch level. For example, constraining a branch in a tree is useful to evaluate whether topologies with a certain branch are or not significantly better than others without the branch. In this sense, the method is useful to evaluate node support or significance of topological incongruence for a certain branch when multiple loci are considered.

A constrained tree is defined here for two cases: first, as a tree where a suboptimal node for the dataset is imposed to appear in the tree; or second, as a tree where an optimal node is constrained not to appear. In the context of maximum parsimony or minimum evolution criteria, the constrained tree is defined as the shortest tree with or without the node under evaluation. As the neighbour-joining method is a greedy algorithm that approximates to minimum evolution tree, we implemented here a modified algorithm in order to constrain trees in a fast way. The NJ method is an agglomerative algorithm. At each cycle, the OTU (operational taxonomic unit) pair to be agglomerated is selected based on a least-square approach. First, the NJ algorithm is modified in order to impose a suboptimal node *x* (a node that not exist in the NJ-tree) to appear. Instead of selecting the pair that minimize the least square equation among all possible pairs, it is selected among the pairs that are topologically congruent with node *x*. This procedure creates a suboptimal tree that is topologically congruent with the node *x*. Second, a similar procedure is used in order to constrain a node, i.e., avoiding the occurrence of this node in the NJ-tree. In each agglomerative cycle the pair that minimizes the least square equation is selected among all pairs except for the pair that forms the node under test.

Algorithm to impose a branch in a tree:

Define *NodeToTest* as the node to be imposed to the tree
*Step 1*: Load $n$ sequences
*Step 2*: Calculate *Dm* as a Distance matrix of size $n \times n$
Initialize the number of OTUs as $n \leftarrow r$
*Step 3*: Calculates *TDm* as a Transformed Distance matrix*
*Step 4*: Select the pair $TDm_{xy}$ with minimum value in TDm
*Step 5*: check if the pair is topologically compatible with *NodeToTest*
(If True then go to step 6
Else delete $TDm_{xy}$ and go to step 4)
*Step 6*: Join the selected OTUs as a node
*Step 7*: Reduce the distance matrix*
*Step 8*: decrease $r$ ($r \leftarrow r-1$)
*Step 9*: Return to step 3 while $r$ is greater than 3
*(for details of step 3 or step 7 see Saitou and Nei (1987))

To constrain a branch avoiding their occurrence into the tree, the step 3a is replaced by:

Step 5b: Check if the pair forms the same bipartition than *NodeToTest*
(If True then delete $TDm_{xy}$ and go to step 3
Else go to step 4)

*5.1.2. Testing significance of localized incongruence*

In order to test significance for nj-LILD value we implemented two different methods.

*Permutation test*: The method uses a permutation of characters among different loci as was previously proposed for the traditional LILD test by Thornton and DeSalle (2000). At each replication, the alignments for each locus are rebuilt based on a random selection (without reposition) of sites of the concatenated alignment. Then, the nj-LILD is calculated for this replica (nj-LILDp). Finally, the proportion of times that nj-LILDp is higher or equal to the nj-LILD is the probability that a so high nj-LILD may be produced just by random.

*Templeton test*: Once we have optimal and constrained topologies for each locus, we can use the Templeton test in order to determine as overall whether optimal topologies are significantly better supported than constrained topologies enforcing the node *x*. This test identifies the sites in the alignment that support one topology over the other, arranges these sites in rank order of their degree of support for one topology over the other, and assesses whether this rank order is significantly different from random under the Wilcoxon rank-sum test. In order to determine whether a site into the alignment support one topology over the alternative, we implemented the Pauplin Formula for tree length (Pauplin, 2000). With this formula we are able to calculate tree length for a single site and determine whether one topology is or not shorter than the other one for such site. Finally, the length differences over each site are tested with Wilcoxon rank-sum test. A significant *p* value indicates that those loci which trees are topologically incongruent with the node *x* are significantly supported. In other words, a *p* value < alpha means that topological incongruence is significant.

*5.1.3. Branch support for multiple combination analyses*

*Bootstrap*: Proposed by Felsenstein (1985) to evaluate branch support. This method remakes the alignment selecting sites at random from the original alignment. Then, the tree is made based on this resampled alignment. The procedure is repeated a user-defined number of times. Then, the proportion of times that a branch appears in the replications is the bootstrap value for that branch. This method is really time-consuming for analysing all possible combinations of certain number of loci. Because of this, we included into the software several options to reduce the number of combinations that needs to be evaluated by bootstrapping.

*Fast-bootstrap*: In the bootstrapping procedure, most of the computational time is invested in tree reconstruction in every replication; however a considerable time is required to make the resampling of sites. Fast-bootstrap is proposed here to resample when multiple combinations are analysed. It is based on the fact that when all combinations of loci are bootstrapped, every locus is resampled several times more than the number of replications. For example, the number of possible combinations of 5 loci in a data set of 10 loci is 252. From these, 126 have the locus *x*, so this locus is bootstrapped 126×R times in the analysis of all combinations, where *R* is the number of replications defined by the user. The fast bootstrapping method reduces to 1×R the times that a locus is resampled. The algorithm makes this in two steps (Supplementary pseudocode). First, characters for each locus are resampled for R replications, and distance matrices are made for each replication and stored in memory. In the second step, the tree for each combination is made and the branch support for *R* replications is calculated. In order to do the last, at each replication, one stored distance matrix of the first step is selected by random for each locus into the combination. Then, the selected matrices are summed. Finally, the bootstrapped tree is made based on the new matrix and evaluated. In classical bootstrapping the total number of resampled sites for all combinations of loci is proportional to the number of combinations multiplied by the number of loci to select. Instead, the total number of resampling sites in fast-bootstrapping is proportional just to the number of loci in the dataset. This method requires more memory to store distance matrices but is twice to ten times faster than bootstrapping when number of combination and size of combinations is high.

*Neighbour Joining based Clade significance (njCS)*: This method is similar to clade significance proposed for parsimony (Lee, 2000) which is based on the Bremer support (Bremer and K.r., 1994). MLSTest calculates clade significance using the modified templeton test (described in Section 5.1.2) to evaluate whether the optimal tree with a clade *x* is significantly better than the tree with the clade *x* constrained not to appear. Clade significance is proposed here as a fast method to discard combinations with low supported clusters to reduce the number of combinations to be analysed by bootstrapping in order to find a good combination.

*5.2. Simulations*

In order to assess the rate of type I error in nj-LILD tests, 1000 simulated datasets of 7 loci were evolved along the topologies showed in Fig. 2 (Asymmetric and Symmetric topology A) under the congruence hypothesis using Seq-Gen v1.3.2 (Rambaut and Grassly, 1997). In order to evaluate variable evolutionary rate among branches, the topologies in Fig. 2 were modified alternating short and long branches with a length ratio of three as is described in Zelwer and Daubin (2004). Type I error rate was measured as the percentage of times that congruence is rejected by the test in these datasets. Additionally, in order to evaluate type II error rate, 1000 datasets of six or five loci evolved along the topology showed in Fig. 2 (topology A) plus one or two locus evolved along the topology showed in Fig. 2 (topology B) were simulated representing a recent lateral gene transfer among strains 4 and 5. Type II error rate was determined as the percentage of times that the hypothesis of congruence is accepted in these datasets. The results were showed as power of the test (100 – type II error rate). The branch lengths for simulations were established in an average of 0.01 or 0.002 changes per site, the length of the simulated alignments was set to 500 nucleotides. The evolutionary model was set to Kimura two-parameters with a proportion of invariable sites of 0.66. The model was selected based on the best model that fit the *T. cruzi* dataset published by Lauthier et al. (2012) using jMODELTEST (Posada, 2009). A complex topology was also evaluated for type I error rate, 1000 simulated datasets of seven loci were generated under the congruence hypothesis based on a tree of 20 random STs of the *FSK* locus for *C. glabrata* dataset. In order to analyse the power of the tests, shuffling of sequences in a *FSK* tree clade of 8 STs was made in order to simulate incongruence. The percentage of trees that had branches with significant *p* value within the shuffled clade was calculated.

*5.3. Datasets and data analysis*

*C. glabrata* dataset using the MLST scheme proposed by Dodgson et al. (2003) and *A. fumigatus* using the MLST scheme proposed by Bain et al. (2007) were downloaded from MLST.net (http://cglabrata.mlst.net/) and pubmlst (http://pubmlst.org/afumigatus/) databases, respectively. Twenty random STs for *Neisseria meningitides* were downloaded from pubmlst (http://pubmlst.org/neisseria/) and twenty representative STs of different serotypes (encapsulated strains) of *H. influenzae* were downloaded from mlst.net (http://haemophilus.mlst.net/). *T. cruzi* and *L. donovani complex* datasets was based on sequences published by Lauthier et al. (2012) and Mauricio et al. (2006), respectively. Correlation between bootstrap and fast-bootstrap *p* values and between bootstrap and 1-njCS was estimated with spearman correlation coefficient using INFOSTAT (InfoStat, 2009).

## Acknowledgements

This work was supported by the European Union Seventh Framework Programme, contract number 223034 (ChagasEpiNet). We thank to Mathew Yeo, Paula Ruybal, Cesar Gómez Hernández, Anahí Alberti D'Amato, Paula Ragone, Mercedes Monje Rumi and Cecilia Pérez Brandán for commentaries and suggestions about the application.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2013.08.029.

## References

Ahmed, A., Thaipadungpanit, J., Boonsilp, S., Wuthiekanun, V., Nalam, K., Spratt, B.G., Aanensen, D.M., Smythe, L.D., Ahmed, N., Feil, E.J., Hartskeerl, R.A., Peacock, S.J., 2011. Comparison of two multi-locus sequence based genotyping schemes for *Leptospira* species. PLoS Negl. Trop. Dis. 5, e1374.

Arnaud-Haond, S., Alberto, F., Teixeira, S., Procaccini, G., Serrao, E.A., Duarte, C.M., 2005. Assessing genetic diversity in clonal organisms: low diversity or low resolution? Combining power and cost efficiency in selecting markers. J. Hered. 96, 434–440.

Bain, J.M., Tavanti, A., Davidson, A.D., Jacobsen, M.D., Shaw, D., Gow, N.A., Odds, F.C., 2007. Multilocus sequence typing of the pathogenic fungus *Aspergillus fumigatus*. J. Clin. Microbiol. 45, 1469–1477.

Boonsilp, S., Thaipadungpanit, J., Amornchai, P., Wuthiekanun, V., Bailey, M.S., Holden, M.T., Zhang, C., Jiang, X., Koizumi, N., Taylor, K., Galloway, R., Hoffmaster, A.R., Craig, S., Smythe, L.D., Hartskeerl, R.A., Day, N.P., Chantratita, N., Feil, E.J., Aanensen, D.M., Spratt, B.G., Peacock, S.J., 2013. A single multilocus sequence typing (MLST) scheme for seven pathogenic *Leptospira* species. PLoS Negl. Trop. Dis. 7, e1954.

Bougnoux, M.E., Morand, S., d'Enfert, C., 2002. Usefulness of multilocus sequence typing for characterization of clinical isolates of *Candida albicans*. J. Clin. Microbiol. 40, 1290–1297.

Bougnoux, M.E., Tavanti, A., Bouchier, C., Gow, N.A., Magnier, A., Davidson, A.D., Maiden, M.C., D'Enfert, C., Odds, F.C., 2003. Collaborative consensus for optimized multilocus sequence typing of *Candida albicans*. J. Clin. Microbiol. 41, 5265–5266.

Bremer, K.r., 1994. Branch support and tree stability. Cladistics 10, 295–304.

Campbell, V., Legendre, P., Lapointe, F.J., 2011. The performance of the congruence among distance matrices (CADM) test in phylogenetic analysis. BMC Evol. Biol. 11, 64.

Chen, K.W., Chen, Y.C., Lo, H.J., Odds, F.C., Wang, T.H., Lin, C.Y., Li, S.Y., 2006. Multilocus sequence typing for analyses of clonality of *Candida albicans* strains in Taiwan. J. Clin. Microbiol. 44, 2172–2178.

Cornish-Bowden, A., 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. 13, 3021–3030.

Debourgogne, A., Gueidan, C., Hennequin, C., Contet-Audonneau, N., de Hoog, S., Machouart, M., 2010. Development of a new MLST scheme for differentiation of *Fusarium solani species complex* (FSSC) isolates. J. Microbiol. Methods 82, 319–323.

Debourgogne, A., Gueidan, C., de Hoog, S., Lozniewski, A., Machouart, M., 2012. Comparison of two DNA sequence-based typing schemes for the *Fusarium solani species complex* and proposal of a new consensus method. J. Microbiol. Methods 91, 65–72.

Dingle, K.E., Colles, F.M., Wareing, D.R., Ure, R., Fox, A.J., Bolton, F.E., Bootsma, H.J., Willems, R.J., Urwin, R., Maiden, M.C., 2001. Multilocus sequence typing system for *Campylobacter jejuni*. J. Clin. Microbiol. 39, 14–23.

Dodgson, A.R., Pujol, C., Denning, D.W., Soll, D.R., Fox, A.J., 2003. Multilocus sequence typing of *Candida glabrata* reveals geographically enriched clades. J. Clin. Microbiol. 41, 5709–5717.

Dodgson, A.R., Pujol, C., Pfaller, M.A., Denning, D.W., Soll, D.R., 2005. Evidence for recombination in *Candida glabrata*. Fungal Genet. Biol. 42, 233–243.

Enright, M.C., Day, N.P., Davies, C.E., Peacock, S.J., Spratt, B.G., 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. J. Clin. Microbiol. 38, 1008–1015.

Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1994. Testing significance of incongruence. Cladistics 10, 315–319.

Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1995. Constructing a Significance Test for Incongruence. Systematic Biology 44, 570–572.

Feil, E.J., Enright, M.C., Spratt, B.G., 2000. Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. Res. Microbiol. 151, 465–469.

Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E., Zhou, J., Spratt, B.G., 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA 98, 182–187.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P., Spratt, B.G., 2004. EBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. 186, 1518–1530.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Francisco, A.P., Bugalho, M., Ramirez, M., Carrico, J.A., 2009. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinform. 10, 152.

Hipp, A.L., Hall, J.C., Sytsma, K.J., 2004. Congruence versus phylogenetic accuracy: revisiting the incongruence length difference test. Syst. Biol. 53, 81–89.

InfoStat, Grupo InfoStat, 2009, FCA, Universidad Nacional de Córdoba, Argentina.

Lauthier, J.J., Tomasini, N., Barnabe, C., Rumi, M.M., D'Amato, A.M., Ragone, P.G., Yeo, M., Lewis, M.D., Llewellyn, M.S., Basombrio, M.A., Miles, M.A., Tibayrenc, M., Diosque, P., 2012. 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.

Lee, M.S.Y., 2000. Tree robustness and clade significance. Sys. Biol. 49, 829–836.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA 95, 3140–3145.

Mauricio, I.L., Yeo, M., Baghaei, M., Doto, D., Pratlong, F., Zemanova, E., Dedet, J.P., Lukes, J., Miles, M.A., 2006. Towards multilocus sequence typing of the *Leishmania donovani complex*: resolving genotypes and haplotypes for five polymorphic metabolic enzymes (ASAT, GPI, NH1, NH2, PGD). Int. J. Parasitol. 36, 757–769.

Meats, E., Feil, E.J., Stringer, S., Cody, A.J., Goldstein, R., Kroll, J.S., Popovic, T., Spratt, B.G., 2003. Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J. Clin. Microbiol. 41, 1623–1636.

Nallapareddy, S.R., Duh, R.W., Singh, K.V., Murray, B.E., 2002. Molecular typing of selected *Enterococcus faecalis* isolates: pilot study using multilocus sequence typing and pulsed-field gel electrophoresis. J. Clin. Microbiol. 40, 868–876.

Odds, F.C., 2010. Molecular phylogenetics and epidemiology of *Candida albicans*. Future Microbiol. 5, 67–79.

Odds, F.C., Bougnoux, M.E., Shaw, D.J., Bain, J.M., Davidson, A.D., Diogo, D., Jacobsen, M.D., Lecomte, M., Li, S.Y., Tavanti, A., Maiden, M.C., Gow, N.A., d'Enfert, C., 2007. Molecular phylogenetics of *Candida albicans*. Eukaryot. Cell 6, 1041–1052.

Paoletti, M., Rydholm, C., Schwier, E.U., Anderson, M.J., Szakacs, G., Lutzoni, F., Debeaupuis, J.P., Latge, J.P., Denning, D.W., Dyer, P.S., 2005. Evidence for sexuality in the opportunistic fungal pathogen *Aspergillus fumigatus*. Curr. Biol. 15, 1242–1248.

Pauplin, Y., 2000. Direct calculation of a tree length using a distance matrix. J. Mol. Evol. 51, 41–47.

Perez-Losada, M., Browne, E.B., Madsen, A., Wirth, T., Viscidi, R.P., Crandall, K.A., 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect. Genet. Evol. 6, 97–112.

Planet, P.J., 2006. Tree disagreement: measuring and testing incongruence in phylogenies. J. Biomed. Inform. 39, 86–102.

Posada, D., 2009. Selection of models of DNA evolution with jModelTest. Methods Mol. Biol. 537, 93–112.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

Robles, J.C., Koreen, L., Park, S., Perlin, D.S., 2004. Multilocus sequence typing is a reliable alternative method to DNA fingerprinting for discriminating among strains of *Candida albicans*. J. Clin. Microbiol. 42, 2480–2488.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Severiano, A., Carrico, J.A., Robinson, D.A., Ramirez, M., Pinto, F.R., 2011. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. PLoS One 6, e19539.

Tavanti, A., Davidson, A.D., Fordyce, M.J., Gow, N.A., Maiden, M.C., Odds, F.C., 2005a. Population structure and properties of *Candida albicans*, as determined by multilocus sequence typing. J. Clin. Microbiol. 43, 5601–5613.

Tavanti, A., Davidson, A.D., Johnson, E.M., Maiden, M.C., Shaw, D.J., Gow, N.A., Odds, F.C., 2005b. Multilocus sequence typing for differentiation of strains of *Candida tropicalis*. J. Clin. Microbiol. 43, 5593–5600.

Thornton, J.W., DeSalle, R., 2000. A new method to localize and test the significance of incongruence: detecting domain shuffling in the nuclear receptor superfamily. Syst. Biol. 49, 183–201.

Varga, J., Toth, B., 2003. Genetic variability and reproductive mode of *Aspergillus fumigatus*. Infect. Genet. Evol. 3, 3–17.

Yeo, M., Mauricio, I.L., Messenger, L.A., Lewis, M.D., Llewellyn, M.S., Acosta, N., Bhattacharyya, T., Diosque, P., Carrasco, H.J., Miles, M.A., 2011. Multilocus sequence typing (MLST) for lineage assignment and high resolution diversity studies in *Trypanosoma cruzi*. PLoS Negl. Trop. Dis. 5, e1049.

Zelwer, M., Daubin, V., 2004. Detecting phylogenetic incongruence using BIONJ: an improvement of the ILD test. Mol. Phylogenet. Evol. 33, 687–693.