A nuclear single-nucleotide polymorphism (SNP) potentially useful for the separation of Rhodnius prolixus from members of the Rhodnius robustus cryptic species complex (Hemiptera: Reduviidae)

Márcio G. Pavan a, Rafael D. Mesquita b, Gena G. Lawrence c, Cristiano Lazoski d, Ellen M. Dotson c, Sahar Abubucker e, Makedonka Mitreva e, Jennifer Randall-Maher e, Fernando A. Monteiro a,⇑

aLaboratório de Sistemática e Bioquímica, Instituto Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, Brazil
bCentro de Bionformática, Departamento de Bioquímica, Instituto de Química, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
cCenters for Disease Control and Prevention, Division of Parasitic Diseases and Malaria, Entomology Branch, 1600 Clifton Road NE, Atlanta, GA 30329, United States
dLaboratório de Bioinformática, Departamento de Bioquímica, Instituto de Química, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
eThe Genome Institute, Washington University School of Medicine, Campus Box 8501, St. Louis, MO 63108, United States

A R T I C L E   I N F O

Article history:
Received 19 June 2012
Received in revised form 16 October 2012
Accepted 23 October 2012
Available online 6 December 2012

Keywords:
Rhodnius
Cryptic species
SNP
Nuclear DNA

A B S T R A C T

The design and application of rational strategies that rely on accurate species identification are pivotal for effective vector control. When morphological identification of the target vector species is impractical, here we describe a non-coding, single-copy nuclear DNA fragment that contains a single-nucleotide polymorphism (SNP) with the potential to distinguish the important domestic Chagas disease vector, Rhodnius prolixus, from members of the four sylvatic Rhodnius robustus cryptic species complex. A total of 96 primer pairs obtained from whole genome shotgun sequencing of the R. prolixus genome (12,626 random reads) were tested on 43 R. prolixus and R. robustus s.l. samples. One of the seven amplicons selected (AmpG) presented a SNP, potentially diagnostic for R. prolixus, on the 280th site. The diagnostic nature of this SNP was then confirmed based on the analysis of 154 R. prolixus and R. robustus s.l. samples representing the widest possible geographic coverage. The results of a 60% majority-rule Bayesian consensus tree and a median-joining network constructed based on the genetic variability observed reveal the paraphyletic nature of the R. robustus species complex, with respect to R. prolixus. The AmpG region is located in the fourth intron of the Transmembrane protein 165 gene, which seems to be in the R. prolixus X chromosome. Other possible chromosomal locations of the AmpG region in the R. prolixus genome are also presented and discussed.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Chagas disease is transmitted through Trypanosoma cruzi infected feces of triatomine bugs (Hemiptera: Reduviidae) and is endemic in Latin America and in the Caribbean (Coura, 2007), ranking fourth in epidemiologic importance among all neglected diseases in those regions (Hotez et al., 2008). Although case incidence estimates show a 90% decline (from 500,000 to 50,000 new cases per year), 8 million people are still infected and 14,000 die every year (Senior, 2007). Domestic insect vectors are responsible for most of the disease transmission to humans. Because no vaccine is available, the elimination of household-infesting triatomines is the major goal of many Chagas disease control programs (Abad-Franch et al., 2010).

The design and implementation of any vector control intervention must rely on accurate species identification. However, the identification of Chagas disease vectors based on morphological characters alone can be technically challenging, since some groups of species are cryptic (i.e., isomorphic), and therefore not amenable to be distinguished in such a manner. This taxonomical issue is even more relevant when applied to the transmission of T. cruzi by Rhodnius prolixus and Rhodnius robustus s.l. to humans. Although morphologically indistinguishable, they play very different epidemiological roles: the former are efficient domestic vectors, and the latter are not, as all species of the complex are entirely sylvatic (Monteiro et al., 2003).

Rhodnius prolixus is one of the most important Chagas disease vectors in Latin America, particularly in Venezuela, Colombia, and Central America, due to its ability to efficiently colonize human dwellings. Although certain countries such as Guatemala have successfully controlled the transmission mediated by this species...
(Guhl et al., 2009), Venezuela and Colombia still face serious problems of re-infestation of insecticide-treated houses by sylvatic populations (Fitzpatrick et al., 2008; Lopez et al., 2007). These events are commonplace in endemic regions where vectors are autochthonous, and will demand the development of innovative surveillance/strategy that contemplate the complexities of sylvatic populations dynamics (Abad-Franch and Monteiro, 2005).

To allow for the accurate taxonomic identification of R. prolixus and the members of the R. robustus cryptic species complex, Pavan and Monteiro (2007) have developed a molecular and cost-effective mitochondrial DNA (mtDNA) multiplex-PCR assay. However, due to its mitochondrial-based nature, this method could generate mitochondrial DNA introgression (Fitzpatrick et al., 2008).

Thus, in order to overcome such limitation, we describe a new non-coding, single-copy nuclear DNA fragment containing a single-nucleotide polymorphism (SNP) with the potential to separate R. prolixus from members of the R. robustus cryptic species complex.

2. Materials and methods

2.1. Molecular identification of specimens and sampling

Since accurate taxonomic identification of R. prolixus and the four cryptic species of R. robustus [sensu Monteiro et al. (2003)] based on morphology alone is problematic, all specimens used in this study were identified through molecular taxonomy based on a fragment of the mitochondrial cytochrome b gene (Pavan and Monteiro, 2007). Moreover, we sequenced the ribosomal second internal transcribed spacer (ITS-2) of all specimens analyzed (data not shown), according to Marcilla et al. (2001), to identify possible introgression of mitochondrial DNA specimens.

For an exploratory screening aimed at finding unique, non-coding regions that separate these species, we tested 25 R. prolixus from two well established laboratory colonies kept at the Centers for Disease Control and Prevention (CDC) in Atlanta (19 specimens) and at the Medical Entomology Research and Training Unit in Guatemala (six specimens), plus two field-collected R. prolixus from Venezuela, two R. robustus I, four of each R. robustus II and III, and six R. robustus IV [sensu Monteiro et al. (2003)]. This approach was done during the initial phase of Rhodnius prolixus genome project, aimed to select a non-introgressed R. prolixus colony with low genetic variability. Fifty-two colony specimens, 43 R. prolixus and nine R. robustus II, and 102 field-collected insects, three R. prolixus, two R. robustus I, 74 R. robustus II, 19 R. robustus III, and four R. robustus IV (Fig. 1 and Table 1), were assayed to confirm the diagnostic applicability of possible species-specific SNPs.

2.2. DNA purification and library construction

DNA from the CDC and Guatemala Colony samples were extracted from ovaries or testis using the Promega genomic extraction kit. For all other Rhodnius specimens, one or two legs of a single Rhodnius specimen was placed in a 1.5 mL microtube, dipped into liquid nitrogen until frozen, and ground to a powder with Kimble/Kontes Pellet Pestles® prior to standard Genomic DNA Extraction with a Real Genomics™ kit. For all other specimens, one or two legs of a single Rhodnius specimen was placed in a 1.5 mL microtube, dipped into liquid nitrogen until frozen, and ground to a powder with Kimble/Kontes Pellet Pestles® prior to standard Genomic DNA Extraction with a Real Genomics™ kit.

From a small WGS plasmid library, 12,626 random Genome Survey Sequences (GSS) were generated as described by Abubucker et al. (2008), using genomic DNA extracted from a population of R. prolixus from the Atlanta colony, and running on an ABI 3730 automated sequencer. Ten thousand GSS reads were screened for non-coding regions, searched against a transposon database, then against RFAM (Gardner et al., 2011), and finally against a non-redundant protein database (with an e-value cutoff of 10−5). Regions with hits were masked. Vector sequences were removed and sequences with unmasked regions of 400–500 base pairs (bp) were extracted and sent through an in-house primer-calling pipeline. Primers with the following criteria were picked: (1) all bases of each primer in the pair must have a quality score of at least Q25; (2) minimum and maximum acceptable amplicon size was 400 and 500, respectively; and (3) excluded regions were 150-bp long (i.e., region with repetitive elements in the middle of the amplicon that would compromise primer annealing).

2.3. PCR amplification and DNA sequencing

From the putative unique, non-coding regions, 96 primer pairs were designed for the PCR amplification of regions of approximately 400–500-bp in length (Table A.1) during the survey sequencing approach.

PCR amplifications of the nuclear gene fragments were performed in a Veriti® 96 Well Thermal Cycler (Applied Biosystems), programmed for one denaturation step at 96 °C for 5 min, followed by 35 cycles at 94 °C for 30 s, 60 °C for 30 s, and 72 °C for 45 s, and a final 10-min extension step at 72 °C. Each 25 μL final volume reaction, contained 5–10 ng of DNA template, 2.5 μL buffer 10×, 2.5 mM of MgCl2, 1.25 units of Taq polymerase (Biotools), 0.5 mM of dNTPs, and 10 pmol of each primer. Purification of PCR products was performed with the Hi Yield™ Gel/PCR DNA Extraction Kit (Real Genomics™), and both fragment strands were subjected to fluorescent dye-terminator cycle sequencing reactions (ABI Prism® BigDye® Terminator v3.1 Cycle Sequencing Kit, Applied Biosystems), using the same primers above, and run on an ABI 3730 automated sequencer.

2.4. Analyses of generated DNA sequence data

The removal of primer sequences, editing of both forward and reverse strands, and the generation of a consensus sequence for each sample were done using the SEQMAN LASERGENE 7.0 program (DNASTar, Inc.). Sequences were aligned using the default parameters of CLUSTALW2 program (Larkin et al., 2007). Polymorphic sites were identified using MEGA 5 (Tamura et al., 2011). RECON (Bao and Eddy, 2002), and BLASTX (Altschul et al., 1997) were used to identify repeated regions and coding sequences, respectively.

A Bayesian phylogenetic tree constructed based on the diagnostic amplicons for 154 R. prolixus and R. robustus s.l. sequences was inferred in BEAST 1.7 (Drummond et al., 2012). Bayesian Information Criterion (BIC) in jMODELTEST (Posada, 2008) was used to elect Jukes and Cantor (1969) as the best-fit evolutionary model (delta

![Fig. 1. Geographical distribution of sampled specimens of R. prolixus and R. robustus s.l. Each location is numbered according to Table 1. The colors indicate the range of R. prolixus and the four members of the R. robustus cryptic species complex (R. robustus I-IV).](image-url)
Table 1

Information on the samples used in this study. Field-collected samples are numbered according to Fig. 1. \( N_A \) – number of individuals tested for seven amplicons (\( \text{AmpG} \) to \( \text{AmpG} \)). \( N_B \) – number of individuals sequenced for \( \text{AmpG} \).

<table>
<thead>
<tr>
<th>Species</th>
<th>Collection site/colony location</th>
<th>Geographical coordinates</th>
<th>( N_A )</th>
<th>( N_B )</th>
<th>Field/colony</th>
<th>Date of collection</th>
<th>GenBank accession numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R. ) prolixus</td>
<td>1. Colombia/CDC, Atlanta, USA</td>
<td>–</td>
<td>19</td>
<td>10</td>
<td>Colony</td>
<td>–</td>
<td>JQ432871 to 432880</td>
</tr>
<tr>
<td></td>
<td>2. Guatemala/NIERTU, Guatemala</td>
<td>–</td>
<td>6</td>
<td>10</td>
<td>Colony</td>
<td>–</td>
<td>JQ432881 to 432890</td>
</tr>
<tr>
<td></td>
<td>“Unknown”/JOC, Brazil</td>
<td>–</td>
<td>–</td>
<td>23</td>
<td>Colony</td>
<td>–</td>
<td>JQ432844 to 432866</td>
</tr>
<tr>
<td></td>
<td>3. Barinas, Venezuela</td>
<td>08°37'N 70°12'W</td>
<td>2</td>
<td>3</td>
<td>Field</td>
<td>2004</td>
<td>JQ432891 to 432893</td>
</tr>
<tr>
<td>( R. ) robustus I</td>
<td>4. Trujillo, Venezuela</td>
<td>09°22'N 70°25'W</td>
<td>2</td>
<td>2</td>
<td>Field</td>
<td>2004</td>
<td>JQ432894, JQ432895</td>
</tr>
<tr>
<td>( R. ) robustus II</td>
<td>5. Carauari, Amazonas, Brazil/INPA, Brazil</td>
<td>–</td>
<td>–</td>
<td>4</td>
<td>Colony</td>
<td>–</td>
<td>JQ432867 to 432870</td>
</tr>
<tr>
<td></td>
<td>6. Sucumbios, Ecuador</td>
<td>–</td>
<td>–</td>
<td>5</td>
<td>Colony</td>
<td>–</td>
<td>JQ432896 to 432900</td>
</tr>
<tr>
<td></td>
<td>7. Monte Negro, Rondônia, Brazil</td>
<td>10°15'N 63°17'W</td>
<td>4</td>
<td>2</td>
<td>Field</td>
<td>2004</td>
<td>JQ432961, JQ432962</td>
</tr>
<tr>
<td></td>
<td>8. Porto Velho, Rondônia, Brazil</td>
<td>08°44'5 63°25'W</td>
<td>45</td>
<td>Field</td>
<td>Apr/2009</td>
<td>JQ432916 to 432960</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9. Cacoal, Rondônia, Brazil</td>
<td>11°26'S 61°27'W</td>
<td>2</td>
<td>2</td>
<td>Field</td>
<td>Apr/2009</td>
<td>JQ432973, JQ432974</td>
</tr>
<tr>
<td></td>
<td>10. Santo Antônio, Rondônia, Brazil</td>
<td>12°33'S 63°13'W</td>
<td>–</td>
<td>1</td>
<td>Field</td>
<td>Feb/2010</td>
<td>JQ432971</td>
</tr>
<tr>
<td></td>
<td>11. Guajará Mirim, Rondônia, Brazil</td>
<td>10°47'5 65°20'W</td>
<td>–</td>
<td>2</td>
<td>Field</td>
<td>Jul/2010</td>
<td>JQ432969, JQ432970</td>
</tr>
<tr>
<td></td>
<td>12. Ouro Preto do Oeste, Rondônia, Brazil</td>
<td>10°43'5 62°15'W</td>
<td>–</td>
<td>7</td>
<td>Field</td>
<td>Oct/2009</td>
<td>JQ432963 to 432968, JQ432972</td>
</tr>
<tr>
<td></td>
<td>13. San Gabriel, Pando, Bolivia</td>
<td>10°16'S 67°03'W</td>
<td>–</td>
<td>2</td>
<td>Field</td>
<td>Sep/2007</td>
<td>JQ432912, JQ432913</td>
</tr>
<tr>
<td></td>
<td>14. Guayas, Santa Cruz, Bolivia</td>
<td>15°52'S 63°20'W</td>
<td>–</td>
<td>7</td>
<td>Field</td>
<td>Mar/2008</td>
<td>JQ432901 to 432907</td>
</tr>
<tr>
<td></td>
<td>15. San Ramón, Santa Cruz, Bolivia</td>
<td>16°29'S 62°29'W</td>
<td>–</td>
<td>1</td>
<td>Field</td>
<td>Mar/2008</td>
<td>JQ432908</td>
</tr>
<tr>
<td></td>
<td>16. Montero, Santa Cruz, Bolivia</td>
<td>17°17'S 63°08'W</td>
<td>–</td>
<td>3</td>
<td>Field</td>
<td>Mar/2008</td>
<td>JQ432909 to 432911</td>
</tr>
<tr>
<td></td>
<td>17. El Torno, Santa Cruz, Bolivia</td>
<td>18°04'S 63°18'W</td>
<td>–</td>
<td>2</td>
<td>Field</td>
<td>Sep/2007</td>
<td>JQ432914, JQ432915</td>
</tr>
<tr>
<td>( R. ) robustus III</td>
<td>18. Araguaína, Pará, Brazil</td>
<td>07°11'S 48°12'W</td>
<td>–</td>
<td>2</td>
<td>Field</td>
<td>Nov/2009</td>
<td>JQ432975, JQ432976</td>
</tr>
<tr>
<td></td>
<td>19. Ulianopolis, Pará, Brazil</td>
<td>03°43'S 47°29'W</td>
<td>–</td>
<td>7</td>
<td>Field</td>
<td>Nov/2009</td>
<td>JQ432977 to 432983</td>
</tr>
<tr>
<td></td>
<td>20. São Domingos do Araguaia, Pará, Brazil</td>
<td>05°30'S 48°44'W</td>
<td>–</td>
<td>3</td>
<td>Field</td>
<td>Nov/2009</td>
<td>JQ432984 to 432986</td>
</tr>
<tr>
<td></td>
<td>21. Rondon do Pará, Pará, Brazil</td>
<td>04°54'S 48°20'W</td>
<td>–</td>
<td>2</td>
<td>Field</td>
<td>Nov/2009</td>
<td>JQ432987, JQ432988</td>
</tr>
<tr>
<td></td>
<td>22. São Bento do Tocantins, Tocantins, Brazil</td>
<td>06°02'S 47°55'W</td>
<td>–</td>
<td>3</td>
<td>Field</td>
<td>Nov/2009</td>
<td>JQ432989 to 432991</td>
</tr>
<tr>
<td></td>
<td>23. Novo Repartimento, Pará, Brazil</td>
<td>04°20'S 49°50'W</td>
<td>4</td>
<td>2</td>
<td>Field</td>
<td>Aug/1998</td>
<td>JQ432992, JQ432993</td>
</tr>
<tr>
<td>( R. ) robustus IV</td>
<td>24. Santa Maria, Pará, Brazil</td>
<td>01°20'S 47°32'W</td>
<td>–</td>
<td>1</td>
<td>Field</td>
<td>Oct/2009</td>
<td>JQ432994</td>
</tr>
<tr>
<td></td>
<td>25. Cayenne, French Guiana</td>
<td>04°56'N 52°20'W</td>
<td>5</td>
<td>2</td>
<td>Field</td>
<td>Jan/2003</td>
<td>JQ432995, JQ432996</td>
</tr>
<tr>
<td></td>
<td>26. Rémire, French Guiana</td>
<td>04°53'N 52°16'W</td>
<td>1</td>
<td>1</td>
<td>Field</td>
<td>Aug/2001</td>
<td>JQ432997</td>
</tr>
</tbody>
</table>

| Total | | | | | | | 43 | 154 |

AIC = 0) for the data set. Tree prior was randomly generated and Yule process of speciation was imposed for all tree reconstructions. Two independent runs were performed for 10^7 generations, with a burn-in of 10^6 generations. Convergence of parameters and a proper mixing were confirmed by calculating the Effective Sample Size (ESS) in TRACER 1.5 (Drummond and Rambaut, 2007), excluding the initial 10% (burn-in) of each run. All considered parameters showed ESS values above 10^4. Runs were combined using LOGCOMBINE, and a maximum credibility tree based on the 20,000 trees generated (burn-in = 2000) and a posterior probability limit of 0.6 was produced using TREE ANNOTATOR (both programs part of the BEAST package). Statistical support for clades was assessed by the posterior probability method, and trees were visualized in FIGTREE v 1.0 (http://tree.bio.ed.ac.uk/software/figtree/).

Genealogy of the DNA sequences generated for the selected diagnostic amplicon was also inferred with NETWORK 4.6 (Fluxus-Engineering), using the Median-Joining network method (Bandelt et al., 1999), with a Maximum-Parimony (MP) calculation in the post-processing. Character weights and epsilon values were given a value of 10 (transition/transversion weighting was 1:1, since we analyzed a non-coding region). Also the Maximum-Composite Likelihood (M-CL) algorithm (Tamura et al., 2004) with 1000 bootstrap replicates of variance, available in MEGA 5 (Tamura et al., 2011), was used to construct intra and interspecific distance matrices.

2.5. Genomic search of selected amplicon


2.6. Selected amplicon genomic contig localization and chromosome synteny

2.6.1. Databases

The \( R. \) prolixus genome assembly v.3.0 and the hard-masked contig GL563069 ("N" masked transposons, repetitive and low complexity regions) were downloaded from VectorBase (http://www.vectorbase.org). Transposable elements and Rhodinus EST databases were kindly provided by Dr. J.M. Ribeiro. A. mellifera, B. terrestris and T. castaneum protein sequences were extracted from the NCBI non-redundant (nr) database using a Perl script that selected and extracted all sequences related to any NCBI taxonomic
index point (taxids used were respectively 7460, 30195 and 7070), using taxdump and gi-taxid-prot NCBI relational datasets (ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz, and ftp://ftp.ncbi.nih.gov/pub/taxonomy/gi_taxid_prot.dmp.gz). The T. castaneum genome assembly version 3.0 was downloaded from BeetleBase (ftp://bioinformatics.ksu.edu/pub/BeetleBase/3.0/Tribolium_genome_sequence.fasta). R. prolixus genes, transcripts and peptides belonging to prediction version C1 (VectorBase gene prediction pipeline using assembly v3.0) were also downloaded from VectorBase (named VectorBase prediction). Alternative gene prediction version Lagerblad 3.0 was done using GENEID v1.3 (Blanco et al., 2007) trained with a large R. prolixus EST database in genome assembly v3.0 (named Lagerblad and available in R. prolixus genomic browser at VectorBase). Transposable elements were identified in contig GL563069 using local RPSBLAST, BLASTN and TBLASTX (Tatusova, 2010), and transposable element databases.

2.6.2. Selected amplicon localization in genomic contigs

The selected amplicon region was searched on R. prolixus genomic contigs using BLASTN (Tatusova, 2010).

2.6.3. Genes and transposons curation in amplicon flanking regions

Genes from both predictions flanking the diagnostic amplicon region were searched for chimeras using BLASTP, NCBI nr protein, UNIPROT databases, and TBLASTN of Rhodnius EST database. Chimera genes were split manually to allow gene synteny analysis. Gene prediction revision was done manually on ARTEMIS (Rutherford et al., 2000) based on EST database hits and GENEWISE ( Birney et al., 2004) analysis using contig GL563069 and coded proteins best-hit in nr and UNIPROT databases. Transposable elements revision was also carried out in GENEWISE with BLASTN and TBLASTX hits, using TE database.

2.6.4. Gene synteny

Reciprocal BLASTP best-hit strategy (Tatusov et al., 1997) was used to identify homologs at the protein level for 14 genes in the diagnostic amplicon flanking region. We used in the comparison species with chromosome-mapped genomes (A. mellifera, B. terrestris and T. castaneum). The R. prolixus database included predicted proteins from both VectorBase and Lagerblad predictions. Bee and beetle homolog proteins GIs from the NCBI protein database (http://www.ncbi.nlm.nih.gov/protein) were used to identify coding genes and their chromosome location. Transposable elements were not considered in this analysis.

2.6.5. Whole contig synteny

Rhodnius prolixus hard-masked contig GL563069 was compared with all T. castaneum chromosomes separately using TBLASTX with a 10−30 e-value cutoff. The results were loaded in ACT (Rutherford et al., 2000) and limited by 50 or 70% identity cutoff when confirming selected gene homolog location, and 50% identity in GL563069 open reading frames (ORFs) larger than 100 bp when analyzing entire contig synteny.

3. Results

3.1. Amplicon selection

Seven amplicons were selected based on reproducibility, variability, and presence of species-specific sites across all taxa tested. In this first screening, 27 R. prolixus, two R. robustus I, four of each R. robustus II and III, and six R. robustus IV samples PCR-amplified and sequenced for these seven nuclear loci yielded 13 polymorphic sites, nine of which are SNPs (Table A2). One such SNP located on the 280th site of the 364-bp region amplified with primers AmpGF, 5′—GAG AGC TGA AGA TAG CAC AGC G, and AmpGR, 5′—TGA TAA CTG GAT TAG GCG CAC G was diagnostic for R. prolixus (i.e., all R. prolixus have an adenine, instead of a guanine, on that particular site). To further evaluate the reliability and usefulness of this particular SNP to serve as an important diagnostic marker for the identification of R. prolixus, an extensive effort was made to obtain the greatest amount of R. prolixus and R. robustus s.l. samples aimed at covering the widest geographical distribution for the species. A total of 154 specimens, 46 R. prolixus, two R. robustus I, 83 R. robustus II, 19 R. robustus III, and four R. robustus IV, were sequenced for the 364-bp AmpG region. The 280th site still proved to be diagnostic. This autapomorphic character of R. prolixus is preceded by 5′—GTCAATAAGA and followed by 5′—TCGTTGGTAG conserved sequences for all species analyzed.

Ten polymorphic sites were observed, eight of which are shared between R. prolixus and R. robustus I, and three of those are parsimony-informative sites (positions 174, 270, and 360; Table 2). The paraphyletic assemblage of R. robustus s.l., previously observed with mitochondrial cytochrome b and ribosomal DNA sequencing analyses (Monteiro et al., 2003), is corroborated here, based on a Bayesian phylogenetic tree reconstruction and a Median-Joining network with Maximum- Parsimony post-processing (Fig. 2). The R. robustus I clade is more closely related to the R. prolixus clade than to the other R. robustus clades. The R. robustus I and R. prolixus sister-species status is also evidenced by M-CL distance comparisons, as sequence divergence between these species is lower than between R. robustus I and R. robustus II-IV (0.6% in comparison with 1.1–1.7%; Table 3).

3.2. Amplicon G genomic contig and flanking region

The AmpG region is located in the R. prolixus genomic contig GL563069 from positions 1116977 to 1117340. It was not found in any other genome searched, not even when high e-values and smaller word sizes were used (allowing for more dissimilar matches to occur), which suggests that this region is unique.

Genes in the AmpG flanking region (contig GL563069 from 925995 to 1319022 positions) were curated (Fig. 3B) to remove chimeras and wrong predictions. Lagerblad genes 17948_72 and 17948_925995 to 1319022 positions) were curated (Fig. 3B) to remove chimeras and wrong predictions. Lagerblad genes 17948_72 and 17948_79 and 17948_84 were wrong predictions and then were re- moved (not shown). No chimera or false positive gene was found in the VectorBase prediction, but four genes were absent (17948_70 – 17948_79 and 17948_88 – 17948_89).

Table 2

<table>
<thead>
<tr>
<th>Species</th>
<th>74</th>
<th>80</th>
<th>167</th>
<th>174</th>
<th>270</th>
<th>280</th>
<th>321</th>
<th>337</th>
<th>341</th>
<th>360</th>
</tr>
</thead>
<tbody>
<tr>
<td>R. prolixus</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>R. robustus I</td>
<td>S</td>
<td>R</td>
<td>R</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>K</td>
<td>Y</td>
<td>S</td>
<td>T</td>
</tr>
<tr>
<td>R. robustus II</td>
<td>–</td>
<td>–</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>R. robustus III</td>
<td>–</td>
<td>–</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>R. robustus IV</td>
<td>–</td>
<td>–</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
**Fig. 2.** Phylogenetic relationship between *Rhodnius prolixus* and *R. robustus* I–IV haplotypes. Species are indicated by different colors. (A) Non-rooted 60% majority-rule Bayesian consensus tree obtained from all 154 *AmpG* sequences of *R. prolixus* and *R. robustus*. Posterior probability values higher than 0.9 are indicated near the nodes. Shared genotypes between different *R. robustus* species are represented by more than one circle in a single terminal branch. Note that *R. robustus* I specimens (red circles) are more closely related to *R. prolixus* (blue circles) than to the other *R. robustus* species, evidencing a paraphyletic assemblage. (B) Median-Joining network with Maximum-Parsimony post-processing. Each node (or circle) represents a unique haplotype. The size of each node is proportional to the number of specimens that shared the same haplotype, and the branch size is proportional to the number of mutational steps. Note that the distance (measured by branch sizes) between *R. prolixus* and *R. robustus* I is smaller than between *R. robustus* I and *R. robustus* II–IV, confirming their sister-species status. The *R. robustus* paraphyly is shown with the two circles, which highlight the *R. prolixus* + *R. robustus* I group, and the Amazonian *R. robustus* (II–IV), respectively.

**Table 3**

Maximum-Composite Likelihood (M-CL) estimates of intraspecific (in italics) and interspecific evolutionary divergences distance matrix. The analysis involved all 154 *R. prolixus* and *R. robustus* s.l. *AmpG* sequences. In parentheses, standard deviations calculated after 1000 bootstrap replications. Note that all comparisons between *R. robustus* I and *R. robustus* II–IV resulted in divergence levels higher than between *R. robustus* I and *R. prolixus*.

<table>
<thead>
<tr>
<th>Species</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <em>R. prolixus</em></td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. <em>R. robustus</em> I</td>
<td>0.005 (±0.003)</td>
<td>0.005 (±0.003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. <em>R. robustus</em> II</td>
<td>0.012 (±0.007)</td>
<td>0.015 (±0.006)</td>
<td>0.003 (±0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. <em>R. robustus</em> III</td>
<td>0.020 (±0.007)</td>
<td>0.017 (±0.007)</td>
<td>0.006 (±0.003)</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>5. <em>R. robustus</em> IV</td>
<td>0.014 (±0.006)</td>
<td>0.011 (±0.004)</td>
<td>0.004 (±0.003)</td>
<td>0.006 (±0.004)</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Fig. 3.** *AmpG* neighborhood. *Rhodnius prolixus* contig GL563069 (region from 1110523–1123517) that contains Transmembrane protein 165 (TP165). Exons are shown as gray rectangles with black outlines linked together. The *AmpG* region (1116877–1117340) is inside the fourth gene intron. Outcast_TE and RT_TE are in the fifth intron and represent an outcast class I transposable element and its reverse transcriptase (A). *Rhodnius prolixus* contig GL563069 (region from 925000 to 1319500) contains 14 genes and four transposons. *AmpG* region is shown as a black vertical dashed line. Genes were represented as gray arrows, and EST-supported genes as gray arrows with black outlines. Gene 8 had fully manual curation and EST support, and was represented as a gray outlined arrow (1 – Ephrin; 2 – Ribonuclease; 3 – Src tyrosine kinase; 4 – Odorant receptor; 5 – Sarcoplasmic Ca-binding protein; 6 – Small Gpa-rich tetratrico peptide repeat-containing protein alpha; 7 – X11Lbeta; 8 – transmembrane protein 165; 9 – interleukin 16; 10 – WNT4 protein; 11 – UNC-112 related protein; 12 – NADH dehydrogenase; 13 – KN motif and askirin repeat domain-containing protein; 14 – band 4.1-like protein 4). Transposable elements were represented as black arrows (a) hAT – class II; (b) outcast – class I; (c) mariner – class II; (d) L1 – class I) (B). Homologous genes were identified at the protein level using reciprocal best-hit strategy (Tatusov et al., 1997). Numbers mean chromosome number. X* = chromosome; X: nf = homologs not found; nm = homologs not mapped; * = not homologous, only a similar gene (C).
The only exception was in the found in the three selected genomes, when we used reciprocal *vitripennis* genome. This gene's original predictions thesis) were found: genomes (and their respective number of chromosomes in parenthesis) were (16), *D. miranda* Subclass (taxid 33340) was the closest taxonomic point that satisfies criteria: (1) to be the phylogenetically closest organisms with chromosome-mapped genomes available; and (2) to have the clos- est number of chromosomes to *. R. prolixus*, which was known to be 20 autosomes, plus XY (Panzer et al., 2007). *R. prolixus* (taxid 13249) taxonomic classification in the NCBI taxonomic web database (http://www.ncbi.nlm.nih.gov/taxonomy) was followed back by tracing the branches of the Insecta phylogenetic tree toward its trunk to search for organisms with chromosome-mapped genomes. Endopterygota Infraclass (taxid 33392), inside Neoptera Subclass (taxid 33340) was the closest taxonomic point that satisfied those two mentioned criteria, where the following 13 species genomes (and their respective number of chromosomes in parenthesis) were found: A. gambiae (4), D. pseudoobscura (7), A. mellifera (16), Drosophila miranda (6), B. terrestris (18), M. destructor (4), N. vitripennis (5), D. virilis (6), T. castaneum (10), D. simulans (6), D. yakuba (7), D. melanogaster (7) and A. aegypti (3). Based on the second criterion, we chose for the analysis two bees, A. mellifera and B. terrestris, and the beetle T. castaneum.

Almost all homolog genes of the AmpG flanking region were found in the three selected genomes, when we used reciprocal best-hit strategy in gene synteny analysis (Tatusov et al., 1997). The only exception was in the T. castaneum analyzed region, where we did not confirm a homolog for the WNT4 protein (GI 91086553; Fig. 3C), as the best-hit was not reciprocal to *R. prolixus*. However, this protein has the same chromosomal localization as UNC-112 related protein and NADH dehydrogenase (Fig. 3C), and thus was not discarded from analysis. X111Beta homolog searches in bees did not retrieve any similar gene, probably due to a lack of prediction, as the TBLASTN web search showed similar genomic regions in unmapped genomic scaffolds (data not shown). *Transmembrane protein 165* gene contained the AmpG region inside its fourth intron, but it was the only protein that exhibited homologs located in different chromosomes in all three compared insects (Fig. 3C). Many homologs *R. prolixus* genes were not found, or their chromosome localization did not suggest a pattern in bee genomes, especially B. terrestris. Considering A. mellifera, four of the first six homolog genes were localized in chromosome 1 (CHR1), but this pattern did not seem to overlap the AmpG region. Moreover, no other pattern was observed until the end of the gene block. T. castaneum homologs showed a mixed localization pattern in CHRX and CHR7 (six and five occurrences, respectively), but the two homologs flanking *Transmembrane protein 165* gene are located at the X chromosome (Fig. 3C).

Whole contig synteny of contig GL563069 was done with all T. castaneum chromosomes separately. Synteny results from the whole contig have confirmed the localization previously identified in reciprocal best-hit approach for the following genes: Ephrin, Snc tyrosin kinase, Small glutamine-rich tetraecopeptide repeat-containing protein alpha, X111Beta, *Transmembrane protein 165*, Interlukin 16, WNT4 protein, UNC-112 related protein, NADH dehydrogenase, KN motif and ankyrin repeat domain-containing protein and Band 4.1-like protein 4 (Fig. A.1). Although the Ribonuclease gene did not show similarity regions neither in CHR7 or CHRX (using an e-value < 10\(^{-10}\)), its homolog was found in reciprocal best-hit approach (Fig. A.1B). Genes Odorant receptor and Sarcoplasmic calcium-binding protein were confirmed to be located in CHR6 (data not shown). The overall synteny of this contig in T. castaneum CHRX, CHR7 and CHR4 showed 161, 158 and 52 hits, respectively (Fig. A.2) and rates of 14.6, 7.7, and 3.7 hits/Mbase, respectively. The synteny of all other chromosomes presented between 36 and 6 hits and 3.1–0.3 hits/Mbase.

3.3. Amplicon G chromosome localization

Arthropod choice for synteny analysis was based on the following criteria: (1) to be the phylogenetically closest organisms with chromosome-mapped genomes available; and (2) to have the closest number of chromosomes to *R. prolixus*, which was known to be 20 autosomes, plus XY (Panzer et al., 2007). *R. prolixus* (taxid 13249) taxonomic classification in the NCBI taxonomic web database (http://www.ncbi.nlm.nih.gov/taxonomy) was followed back by tracing the branches of the Insecta phylogenetic tree toward its trunk to search for organisms with chromosome-mapped genomes. Endopterygota Infraclass (taxid 33392), inside Neoptera Subclass (taxid 33340) was the closest taxonomic point that satisfied those two mentioned criteria, where the following 13 species genomes (and their respective number of chromosomes in parenthesis) were found: A. gambiae (4), D. pseudoobscura (7), A. mellifera (16), Drosophila miranda (6), B. terrestris (18), M. destructor (4), N. vitripennis (5), D. virilis (6), T. castaneum (10), D. simulans (6), D. yakuba (7), D. melanogaster (7) and A. aegypti (3). Based on the second criterion, we chose for the analysis two bees, A. mellifera and B. terrestris, and the beetle T. castaneum.

Almost all homolog genes of the AmpG flanking region were found in the three selected genomes, when we used reciprocal best-hit strategy in gene synteny analysis (Tatusov et al., 1997). The only exception was in the T. castaneum analyzed region, where we did not confirm a homolog for the WNT4 protein (GI 91086553; Fig. 3C), as the best-hit was not reciprocal to *R. prolixus*. However, this protein has the same chromosomal localization as UNC-112 related protein and NADH dehydrogenase (Fig. 3C), and thus was not discarded from analysis. X111Beta homolog searches in bees did not retrieve any similar gene, probably due to a lack of prediction, as the TBLASTN web search showed similar genomic regions in unmapped genomic scaffolds (data not shown). *Transmembrane protein 165* gene contained the AmpG region inside its fourth intron, but it was the only protein that exhibited homologs located in different chromosomes in all three compared insects (Fig. 3C). Many homologs *R. prolixus* genes were not found, or their chromosome localization did not suggest a pattern in bee genomes, especially B. terrestris. Considering A. mellifera, four of the first six homolog genes were localized in chromosome 1 (CHR1), but this pattern did not seem to overlap the AmpG region. Moreover, no other pattern was observed until the end of the gene block. T. castaneum homologs showed a mixed localization pattern in CHRX and CHR7 (six and five occurrences, respectively), but the two homologs flanking *Transmembrane protein 165* gene are located at the X chromosome (Fig. 3C).

Whole contig synteny of contig GL563069 was done with all T. castaneum chromosomes separately. Synteny results from the whole contig have confirmed the localization previously identified in reciprocal best-hit approach for the following genes: Ephrin, Snc tyrosin kinase, Small glutamine-rich tetraecopeptide repeat-containing protein alpha, X111Beta, *Transmembrane protein 165*, Interlukin 16, WNT4 protein, UNC-112 related protein, NADH dehydrogenase, KN motif and ankyrin repeat domain-containing protein and Band 4.1-like protein 4 (Fig. A.1). Although the Ribonuclease gene did not
prolixus itself, in contig GL563069, inside the fourth intron of Trans-
membrane protein 165 gene (TP165). Fourteen genes flanking the 
AmpG region were annotated and their homologs searched at the 
protein level, using reciprocal best-hit strategy (Tatusov et al., 
1997) in A. mellifera, B. terrestris and T. castaneum. These were the 
three arthropods that have shown the closest chromosome number 
and phylogenetic relationship to R. prolixus. Despite the detection 
of the fourth TP165 intron as the location of the AmpG region, 
this protein was the only one of the R. prolixus contig whose homologs 
were located on different chromosomes in all three species 
searched, and thus could not be used to infer AmpG localization. 
Considering all 14 gene homologs’ localization in all three species 
analyzed, T. castaneum showed the most homogenous pattern 
concerning genes location, with six and five occurrences in X and 7 
chromosomes (CHR X and CHR 7), respectively; notwithstanding, 
both AmpG flanking genes homologs were located in CHR X.

The whole contig synteny approach represents a more reliable 
method to look at genomic similarities, as DNA sequences are com-
pared along the complete extension of each contig. R. prolixus con-
tig GL563069 synteny in comparison with T. castaneum contig also 
pointed CHR X and CHR 7 with almost equal number of hits (161 
and 158) as the possible chromosome location, even though their 
lengths are very different. The frequency of hits in CHR X was 
two times higher than in CHR 7 (146,7 to 7,7 hits/Mbase, respec-
tively) and about ten times higher than all other chromosomes, 
showing that contig GL563069 shares more potential gene content 
with T. castaneum CHR X. Thus, our results suggest that this contig 
could be part of R. prolixus X chromosome.

DNA sequencing analysis of the AmpG region of R. prolixus and R. 
robustus s.l. samples shows that R. prolixus and R. robustus I share 
eight polymorphic sites, three of these are derived nucleotides 
(A, C, and A, at positions 174, 270, and 360, respectively; Table 2), 
and hence Maximum-Composite Likelihood divergence between 
these species reached 0.6% in contrast to 1.1–1.7% between R. robustus 
I and R. robustus II–IV. A Bayesian maximum credibility tree and a 
Median-Joining network with Maximum-Parsimony post-
processing further support the paraphyletic nature of R. robustus 
(Fig. 2), previously observed with cyt b and rDNA sequences 
(Monteiro et al., 2003).

In addition to its utility in taxonomy, this marker could be used 
in combination with the cyt b multiplex-PCR (Pavan and Monteiro, 
2007) for the detection of natural crossbreeding and introgression 
of mitochondrial DNA along putative hybrid zones, as reported by 
Abad-Franch et al. (1993) and Wigglesworth (1957), several laboratories world-
wide currently maintain colonies of this species. There is evidence, 
though, that some of these colonies have been contaminated with, 
or misidentified as, R. robustus s.l. (MCP, CL, EMD, FAM, unpub-
lished). The SNP here described could also be used to evaluate the 
identity and purity of R. prolixus colonies. However, although the 
possibility of using this SNP as a diagnostic marker for this import-
tant vector species seems very promising, its “universal” applicabil-
ity remains to be evaluated by the research community.

Disclosure statement

The authors declare no conflict of interest.

Role of the funding source

This work was supported by the Brazilian Research Council 
(CNPq), the Brazilian Ministry of Science, Technology and 
Innovation (MCTI), and NIH Grant No. U54 HG003079. The spon-
sors did not take part nor influenced the any aspect of this study.

Acknowledgements

To John Spieth at the Genome Institute at Washington University 
School of Medicine in St. Louis for the assistance with library 
construction and identification of nuclear loci. To Sebastião Aldo 
Valente, Luis Herman S. Gil, Toby Barrett, José R. Coura, Christine 
Aznar, and Celia Cordon-Rosales, for kindly providing specimens. 
We thank José Marcos Ribeiro for providing the Transposon data-
bases. The authors are also grateful to the comments of two anony-
mous referees that helped improve the original manuscript. We 
also thank the PTDIS-FIOCRUZ DNA sequencing core for running 
sequencing reactions.

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. 
2012.10.018.

References

Abad-Franch, F., Ferraz, G., Campos, P., Palomeque, P.S., Grijalva, M.J., Aguilar, H.M., 
Miles, M.A., 2010. Modeling disease vector occurrence when detection is 
imperfect: infestation of Amazonian palm trees by triatomine bugs at three 
Abubucker, S., Martin, J., Yin, Y., Fulkon, L., Yang, S.P., Hallsoworth-Pepin, K., Johnston, 
hookworm genome: analysis and classification of Ancylostoma canum survey 
Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 
in sequenced genomes. Genome Res. 12, 1269–1276.

Bargues, M.D., Marcilla, A., Dujardin, J.F., Mas-Coma, S., 2002. Triatomin vectors of 
Trypanosoma cruzi: a molecular perspective based on nuclear ribosomal DNA 
Bargues, M.D., Klioswicz, D.R., Panerza, F., Noireau, F., Marcilla, A., Perez, R., Rojas, 
M.G., O’Connor, J.E., Gonzalez-Candelas, F., Galvao, C., Jurberg, J., Carcavallo, 
Chagas disease main vector Triatoma infestans based on nuclear rDNA 

988–995.
Bioinform. 4, Unit 4.3.

Coura, J.R., 2007. Chagas disease: what is known and what is needed—a background 

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by 


Fitzpatrick, S., Feliciangeli, M.D., Sanchez-Martin, M.J., Monteiro, F.A., Miles, M.A., 
2008. Molecular genetics reveal that silvatic Rhodius prolixus do colonise rural 

Gardner, P.P., Daub, J., Tate, J., Moore, B.L., Osuch, I.H., Griffiths-Jones, S., Finn, R.D., 


Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, 
H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., 
2947–2948.


