Bulletin of Entomological Research (2011) **101**, 383–391
doi:10.1017/S0007485310000556

First published online 21 December 2010

Analysis of the workers head transcriptome of the Asian subterranean termite, *Coptotermes gestroi*

F.C. Leonardo[1], A.F. da Cunha[2], M.J. da Silva[3], M.F. Carazzolle[1,4], A.M. Costa-Leonardo[5], F.F. Costa[6] and G.A. Pereira[1]*

[1]Laboratório de Genômica e Expressão, Departamento de Genética Evolução e Bioagentes, Instituto de Biologia, Universidade Estadual de Campinas, CEP: 13083-970, Campinas, São Paulo, Brazil: [2]Laboratório de Bioquímica e Biotecnologia molecular, Departamento de Genética e Evolução, Universidade Federal de São Carlos, Via Washington Luis Km 235, CEP:13565-905, São Carlos, São Paulo, Brazil: [3]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Avenida Cândido Rondon, 400, CEP: 13083-875, Campinas, São Paulo, Brazil: [4]Centro Nacional de Processamento de Alto Desempenho em São Paulo, Rua Saturnino Brito, 45, Cidade Universitária, CEP: 13083-889, Universidade Estadual de Campinas, São Paulo, Brazil: [5]Departamento de Biologia, Unesp, Univ. Estadual Paulista, Avenida 24 A Universidade Estadual Paulista, Avenida 24[a], Bela Vista, CEP: 13506-900, Rio Claro, São Paulo, Brazil: [6]Centro de Hematologia e Hemoterapia, Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Rua Carlos Chagas, 480, CEP: 13083-878, Campinas, São Paulo, Brazil

Abstract

The lower termite, *Coptotermes gestroi* (Isoptera: Rhinotermitidae), is originally from Southeast Asia and has become a pest in Brazil. The main goal of this study was to survey *C. gestroi* transcriptome composition. To accomplish this, we sequenced and analyzed 3003 expressed sequence tags (ESTs) isolated from libraries of worker heads. After assembly, 695 uniESTs were obtained from which 349 have similarity with known sequences. Comparison with insect genomes demonstrated similarity, primarily with genes from *Apis mellifera* (28%), *Tribolium castaneum* (28%) and *Aedes aegypti* (10%). Notably, we identified two endogenous cellulases in the sequences, which may be of interest for biotechnological applications. The results presented in this work represent the first genomic study of the Asian subterranean termite, *Coptotermes gestroi*.

Keywords: *Coptotermes gestroi*, ESTs, cDNA library, Isoptera, cellulose

(Accepted 20 August 2010)

Introduction

In recent years, the genomes of key insects such as *Drosophila melanogaster*, *Anopheles gambiae*, *Apis mellifera*, *Bombyx mori*, *Tribolium castaneum* and *Acyrthosiphon pisum* have been sequenced. Although novel sequencing methods,

*Author for correspondence
Fax: 0055-19-35216235
E-mail: goncalo@unicamp.br

such as pyrosequencing (Ronaghi, 2001), promise a new era in genomics, sequencing of cDNA libraries produced from whole body or selected tissues is still an efficient strategy to gain insights into genetics and physiology of non-model insects.

Termites are social insects that can be roughly separated into two major groups, higher and lower termites, according to the symbionts found in their gut (Grassé, 1949). Higher termites harbor only bacteria in the digestive tract and include only species of the family Termitidae. In addition to bacterial flora, the lower termites also present symbiotic protozoa in their gut, which contribute to their digestive processes. A variety of molecular studies have been conducted to understand caste polymorphism in termites. Wu-Scharf *et al.* (2003) described 88 expressed sequence tags (ESTs), generated from a polyphenic cDNA library from *Reticulitermes flavipes* (Wu-Scharf *et al.*, 2003). Among these sequences, they found genes that were caste-specifically expressed (Scharf *et al.*, 2003, 2005), including two hexamerin encoding genes. With the use of RNAi, it was discovered that dual silencing of both hexamerins leads to significant increases in JH-dependent presoldier differentiation (Zhou *et al.*, 2006).

Since most of the termites are xylophagous, they have the ability to digest lignocellulose due to lignocellulose enzymes present in their digestive tract and in their gut symbionts (Nakashima *et al.*, 2002; Scharf *et al.*, 2003). These enzymes have demonstrated a great potential for bioenergy applications and led to a new era in research (Scharf & Tartar, 2008).

Digestive genomic studies include symbiont metatranscriptomic sequencing of cDNA clones, representing expressed genes of eukaryotic (protistan) symbionts from the lower the termite, *Reticulitermes speratus* (Todaka *et al.*, 2007). This study revealed diverse cellulase and hemicellulase sequences in this termite, while another investigation was conducted to elucidate the symbiont metagenome sequencing of a higher *Nasutitermes* species from Costa Rica (Warnecke *et al.*, 2007). This study produced over 100 million bases of DNA sequence and revealed genes that relate to numerous aspects of microbial life in the gut microenvironment. Most recently, dual-host symbiont transcriptome sequencing was accomplished to determine genes expressed in the gut of the termite, *Reticulitermes flavipes*. The results increased the number of genes of the host and symbiotic glycosyl hydrolases families, supporting previous models of lignin degradation and host-symbiotic collaboration in hemicellulose digestion (Tartar *et al.*, 2009).

Among the Rhinotermitidae, *Coptotermes gestroi* is marked as the most destructive pest termite, damaging structural wood in the urban areas in Southeast Asia (Kirton, 2005). This endemic Southeast Asian species has spread in various regions around the world: Marquesas Islands (Pacific Ocean), Mauritius and Reunion (Indian Ocean), Brazil, Barbados, West Indian Islands, southern Mexico and the United States (Jenkins *et al.*, 2007).

Although *C. gestroi* has had an important economic impact, to date, only a few sequences have been produced from this species, and most of these correspond to ribosomal or mitochondrial fragments (Jenkins *et al.*, 2007). In this study, we report the identification of 695 *C. gestroi* ESTs, interpreting their possible role in the biology of this species. Among these ESTs, we identified two cellulases: endo-β-1,4-glucanase and β-glucosidase. Comparative analysis showed that these enzymes are very similar to counterparts found in other termite species and, therefore, are likely to represent genuine termite enzyme coding genes and not enzymes from the symbiotic fauna of the workers' digestive tract. Quantitative PCR of different developmental stages was performed to demonstrate the comparative expression of these cellulases.

Experimental procedures

Biological samples and RNA extraction

Workers of *C. gestroi* were obtained from a field colony located in the city of Rio Claro, São Paulo State, Brazil (22°23′S, 47°32′W). In order to avoid contamination by gut symbiotic protists, only the workers' heads were used in the experiment. Prior to RNA isolation, the workers had their heads cut off, and these were frozen at –80°C. Total RNA was isolated from heads using TRIzol reagent (Invitrogen).

Generation of ESTs (expressed sequence tags)

Aliquots of 1.85 μg of total RNA were reverse transcribed utilizing SuperScript II Reverse Transcriptase (Invitrogen), primer Oligo dT [5′-GGCGGCCGCACAACTTTGTACAAG-AAAGTTGGGT$(T)_{19}$-3′] and primer Poly G (5′-TCGTCGGG-GACAACTTTGTACAAAAAAGTTGG-3′). First-strand cDNA synthesis occurred at 42°C for 60 min in a total volume of 10 μl. The cDNAs contained in 2 μl of each RT-product were then amplified by long-distance polymerase chain reaction (LD-PCR), which is a technique that enables the amplification of large cDNAs from an uncloned pool of mRNA extension products using the same primers in a PCR mix (5.0 μl 10 × Advantage 2 PCR buffer, 1.0 μl primer Olido dT, 1.0 μl primer poly G, 1.0 μl enzyme 50 × Advantage 2 polymerase mix and 40 μl distilled water). The amplification protocol consisted of an initial step at 95°C for 20 s, followed by 20 cycles of 95°C for 5 s and 68°C for 6 min.

Aliquots of the PCR products (5 μl) were run on 1% agarose gels and stained with ethidium bromide. The PCR product will produce a *smear* that ranges from 0.1 to 4.0 Kb, representing the most abundant transcripts. The amplification products of over 600 bp were then extracted from the agarose gels (Wizard–SV gel and PCR Clean-up System, Promega) and ligated into a pDONR-222 vector (Invitrogen). Ligation was transformed by eletroporation into competent *E. coli* DH10B-cells and selected for kanamycin resistance. Plasmid DNA was purified and sequenced using M13 primer (5′–GTAAAACGACGGCCAG-3′). High throughput EST sequencing was performed using dye-terminator chemistry on an ABI 3700 sequencer.

Sequence analysis

The EST sequences were base-called and screened for vector sequences using PHRED and cross-match software (Ewing & Green, 1998). Poly-A repeats and low quality sequences were trimmed (Telles & da Silva, 2001), and the high quality ones (at least 100 bp with Phred > 20) were clustered and assembled by CAP3 (Huang & Madan, 1999). The clusters were submitted to automatic and manual annotation, where the sequences were used as queries for similarity searches by BLASTn, tBLASTx and BLASTx (Altschul *et al.*, 1997) against GenBank (nucleotide and amino acid databases). Database searches were performed under default settings, with a threshold e-value of lower than 10^{-5} regarded as significant. In addition, we compared the

Table 1. Primers used in the relative gene expression experiment.

Primers	Sequence 5–3′	Concentration
β-glucosidase F	GATAATAGAGTTGTACCGTGG	150 nM
β-glucosidase R	ACTCCACCGTAGTCCGAGAAAC	
Endo-β-1,4-glucanase F	GTGCCTTCAACTGGGATAACAA	150 nM
Endo-β-1,4-glucanase R	TGTATGCCTGCTTGCTTGTGA	
β-actin F	GTTGTCAAATGTAATCCATCC	300 nM
β-actin R	CTGACGTACTGTCGCCTGT	
α-tubulin F	CCCTCTCCTTCGCCCTTCAA	150 nM
α-tubulin R	GAGGATCTCGCTGCTCTGGA	

C. gestroi sequences with those of the 18,122 Isoptera EST sequences available in a public database (www.ncbi.nlm.nih.gov) using tBLASTx with a threshold e-value of lower than 10^{-10}.

The clusters that presented 'no database matches' (no hits) were analyzed by the ESTScan program (Iseli *et al.*, 1999), employing the Drosophila model for CDS prediction. GC content of the coding and non-coding sequences and GC percentage at different positions (GC1, GC2 and GC3) were obtained using the predicted CDS and comparing them with CDS of the hits clusters. Functional classification was performed using the Gene Ontology database (www.geneontology.org).

Relative gene expression

Total RNA was extracted from the whole body of different castes and developmental stages (larvae, soldiers, alates, queens and workers) using QIAGEN RNAeasy Mini Kit. RNA samples (5 µg) were incubated with 1 U DNaseI (Invitrogen Corp., Rockville, MD, USA) for 15 min at room temperature, and EDTA was added to a final concentration of 2 mM to stop the reaction. The DNaseI enzyme was subsequently inactivated by incubation at 65°C for 5 min. DNaseI-treated RNA samples were then reverse-transcribed with Superscript III and RNaseOut (Invitrogen Corp., Carlsbad, CA, USA) for 50 min at 50°C, 15 min at 70°C. cDNA samples were quantified using a Nanodrop spectrophotometer (ND-1000; Nanodrop Technologies Inc., Wilmington, DE, USA). Synthetic oligonucleotide primers were designed to amplify cDNA for the genes encoding the two cellulases (Primer ExpressTM, Applied Biosystems, Foster City, CA, USA). For primer sequences and concentrations, see table 1. Primers were synthesized by Invitrogen Corp. (Carlsbad, CA, USA) and IDT (Coralville, IA, USA). All samples were assayed in a 12-µl volume, containing 5 ng cDNA (3 µl), 6 µl SYBR Green Master Mix PCR (Applied Biosystems), and 3 µl-specific primers in a MicroAmp Optical 96-well reaction plate (Applied Biosystems), using the Step one Plus (Applied Biosystems), as described previously. Gene expression analyses were performed with the geNorm program (Vandesompele *et al.*, 2002). Two replicas were run on the plate for each sample, and the data are mean values of three independent preparations. Results are expressed as mRNA levels of each gene studied, normalized according to β-actin and α-tubulin expressions.

Statistical analyses

To evaluate expression patterns (real-time PCR), one-way analyses of variance (ANOVA) were performed separately for each gene using caste (larvae, nymph, workers and soldiers) as fixed factors. Tukey's multiple comparison test was applied to reveal pair-wise differences. Statistical analyses were performed using Graph Pad Prism (version 5.0).

Results

cDNA library construction and analysis

We generated and analyzed 4697 sequences of a cDNA library from worker heads of the subterranean termite *C. gestroi* (general scheme in fig. 1). From the remaining sequences, 3003 were considered of appropriate quality – presenting at least 100 bases with Phred ≥ 20. These sequences were clustered by CAP3 into 245 contigs and 450 singlets, and the EST size of the contigs is shown in fig. 1. Sequences were submitted to similarity searches by BLAST against GenBank and, after manual annotation, resulted in 321 (46.19%) hits and 374 (53.81%) no hits. All these sequences were deposited in GenBank´s dbEST database and are available using the accession numbers GT566617–GT568042.

From the 321 ESTs, 42 unigenes presented similarity with non-insect genes such as bacteria, protozoa, fungi and plants. The other 271 ESTs were assigned as the termite's sequences and were compared to other insect genomes.

ESTScan analysis

The ESTScan analyses applied to no-hits sequences identified 129 predicted CDS. In order to validate these sequences, a GC content analysis was performed, and the results were compared with GC content from *C. gestroi* hits and sequences of other Isoptera (table 2).

Gene ontology classification

For the putative genes with known function, we assigned biological processes by using level three of the Gene Ontology (GO) classification (Ashburner *et al.*, 2000). There is a clear prevalence for ESTs representing cellular physiological process and metabolism (fig. 2). This result was also found in a previous study of *Apis mellifera*'s ESTs (Nunes *et al.*, 2004).

Genome comparison

We compared *C. gestroi* sequences with those from insects whose genomes have been completely sequenced. The distribution of the respective best hits were with *A. mellifera* (28%) and *Tribolium castaneum* (28%), followed by *Aedes aegypti* (10%), *Anopheles gambiae* (8%) and *Drosphila sp.* (7%).

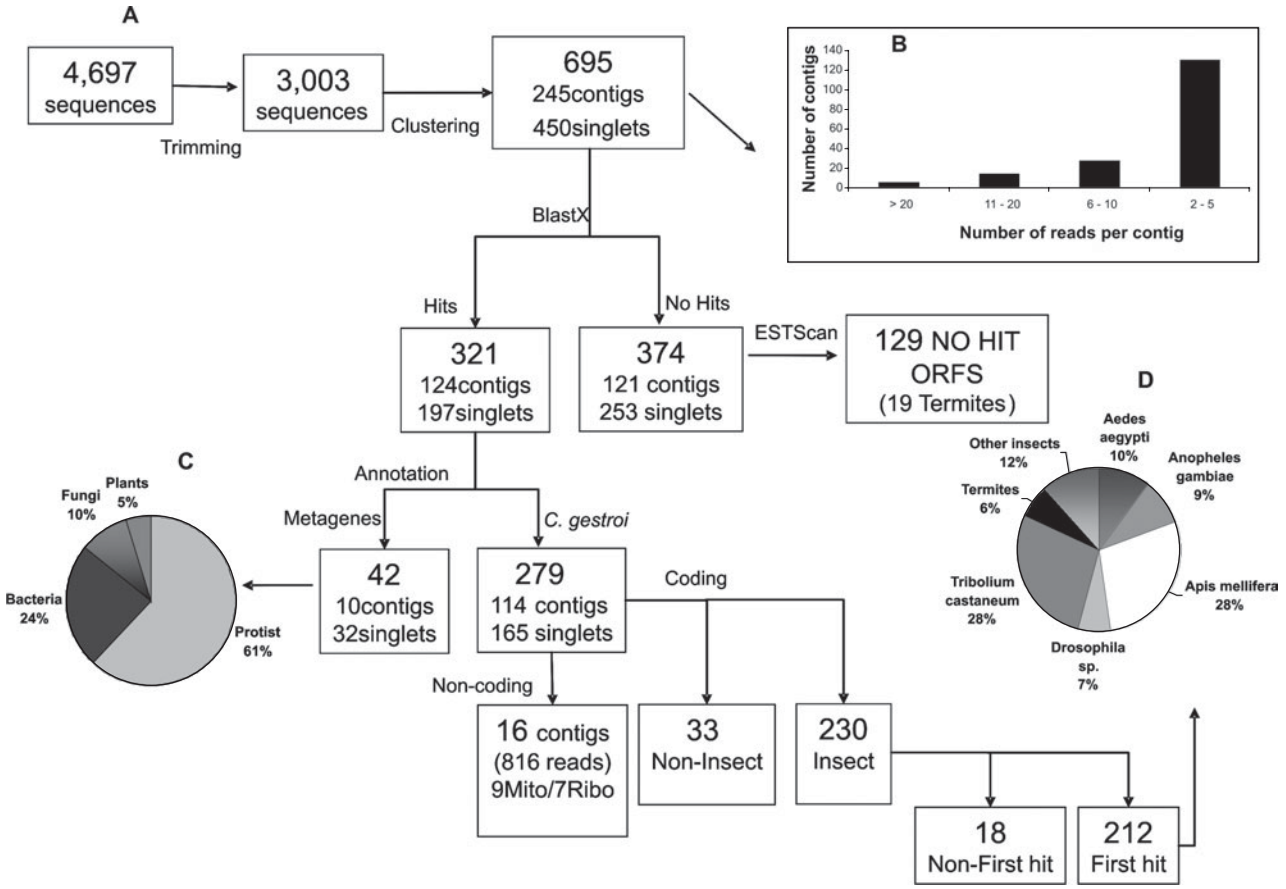

Fig. 1. General scheme representing how the sequences were grouped and their manual annotation. (a) Fluxogram generated by the treatment of the sequences, (b) distribution of reads in clusters, (c) metagenes, and (d) first hit comparison with insect genomes.

Table 2. Base composition (%GC) at different positions for coding sequences of *C. gestroi*'s library and Isoptera sequences available in a public database (GenBank).

GC position	Isoptera	*C. gestroi* hits	*C. gestroi* no hits
GC1	52 ± 5	53 ± 6	54 ± 9
GC2	40 ± 7	40 ± 5	43 ± 10
GC3	52 ± 10	49 ± 10	55 ± 8

Even though the number of sequences we obtained for *C. gestroi* is relatively low, an interesting finding was that we obtained ESTs representing two cellulases: endo-β-1,4-glucanase (EC 3.2.1.4) (fig. 3) and β-glucosidase (EC 3.2.1.2.1) (fig. 4), which have been previously identified in salivary glands of other termites species.

Gene expression analyzes

To validate and analyze the expression of the putative cellulases found in the cDNA library, relative gene expression experiment was performed using RNA samples derived from different castes and developmental stages (larvae, alates, soldiers, queens and workers). Figure 5 shows relative expression levels of both cellulases in all castes. The statistical analysis showed that the expression of both genes is higher in workers compared to the other developmental stages.

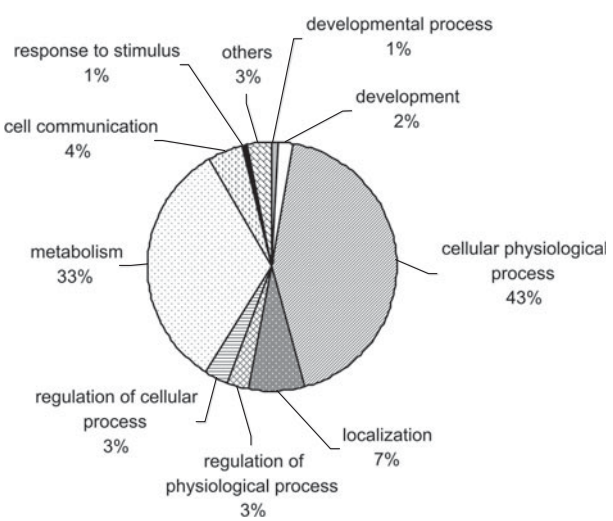

Fig. 2. Gene ontology classification of *C. gestroi* cDNA library sequences.

```
R.flavipes          EDLTGGYYDAGDFVKFGFPMAYTVTVLAWGVIDYESAYSAAGALDSGRKALKYGTDYFLK 120
R.speratus          EDLTGGYYDAGDFVKFGFPMAYTVTVLAWGVIDYESAYSAAGALDSGRKALKYGTDYFLK 120
C.formosanus        EDLTGGYYDAGDFVKFGFPMAYTVTVLAWGLVDYESAYSTAGALDDGRKALKWGTDYFLK 120
C.acinaciformis     EDLTGGYYDAGDFVKFGFPLAYTATVLAWGLVDYEAGYSSAGATDDGRKAVKWATDYLLK 120
C.gestroi           --------------------------------------------------WGTDYFLK 8
                                                                       :.***:**

R.flavipes          AHTAANEFYGQVGQGDVDHAYWGRPEDMTMSRPAYKIDTSKPGSDLAAETAAALAATAIA 180
R.speratus          AHTAANEFYGQVGQGDVDHAYWGRPEDMTMSRPAYKIDTSKPGSDLAAETAAALAATAIA 180
C.formosanus        AHTAANEFYGQVGQGDVDHAYWGRPEDMTMSRPAYKIDTSKPGSDLAAETAAALAATAIA 180
C.acinaciformis     AHTAATELYGQVGDGDADHAYWGRPEDMTMARPAYKIDASRPGSDLAGETAAALAAASIV 180
C.gestroi           AHTAASEFYGQVGQGDVDHAYWGRPEDMTMSRPAYKIDTSKPGSDLAAETAAALAATAIV 68
                    *****.*:*****:**.*************:*******:*:******.********::*.

R.flavipes          YKSADATYSNNLITHAKQLFDFANNYRGKYSDSITDAKNFYASGDYKDELVWAAAWLYRA 240
R.speratus          YKSADATYSNNLITHAKQLFDFANNYRGKYSDSITDAQNFYASGDYKDELVWAAAWLYRA 240
C.formosanus        YKSADSTYSNNLITHAKQLFDFANNYRGKYSDSITDAKNFYASGDYKDELVWAAAWLYRA 240
C.acinaciformis     FKGVDSSYSDNLLAHAKQLFDFADNYRGKYSDSITQASNFYASGDYKDELVWAATWLYRA 240
C.gestroi           YKSVDSTYSNNLITHAKQLFDFANNYRGKYSDSITDAKNFYASGDYKDELVWAAAWLYRA 128
                    :*..*::**:**::*********:***********:*.****************:*****

R.flavipes          TNDNTYLTKAESLYNEFGLGNWNGAFNWDNKISGVQVLLAKLTSKQAYKDKVQGYVDYLI 300
R.speratus          TNDNTYLTKAESLYNEFGLGNWNGAFNWDNKISGVQVLLAKLTSKQAYKDKVQGYVDYLI 300
C.formosanus        TNDNTYLTKAESLYNEFGLGSWNGAFNWDNKISGVQVLLAKLTSKQAYKDKVQGYVDYLV 300
C.acinaciformis     TNDNTYLTKAESLYNEFGLGNWNGAFNWDNKVSGVQVLLAKLTSKQAYKDTVQGYVDYLI 300
C.gestroi           TNDNAYLTKAESLYNEFGLGSWNGAFNWDNKISGVQVLLVKLTSKQAYKDKVQGYV-YLI 187
                    ****:***************.**********:*******.**********.***** **:
```

Fig. 3. Multiple alignments of endo-β-1,4-glucanase from termites: *Reticulitermes flavipes* (AY572862.2), *Reticulitermes speratus* (AB008778.2) *Coptotermes formosanus* (AB058671.1), *Coptotermes gestroi* (GT567748.1) and *Coptotermes acinaciformis* (AF336120). GenBank accession numbers are in parentheses. Asterisks represent the same amino acid residues.

```
C.formosanus        TSQDPEWPESASSWLRVVPWGFRKELNWIANEYGNPPIFITENGFSDYGGVNDTNRVLYY 417
N.takasagoensis     LSKDPNWPESASSWLRVVPWGFRKELNWIANAYGNPPIYVTENGFSDYGGLNDTNRVLYY 412
N.koshunensis       LTQDAAWPISASSWLKVVPWGFRKELNWIKNEYNNPPVFITENGFSDYGGLNDTGRVHYY 420
C.gestroi           -------PRES-----IKQWYQR-------------------------------RVLYY 16
                           * .:     :  *  *                                ** **

C.formosanus        TEHLKEMLKAIHIDGVNVIGYTAWSLIDNFEWLRGYTERFGIHAVNFIDPSRPRTPKESA 477
N.takasagoensis     TEHLKEMLKAIHIDGVNVLGYTAWSLLDNFEWLRGYTERFGIHEVNFNDPSRPRTPKESA 472
N.koshunensis       TEHLKEMLKAIHEDGVNVIGYTAWSLMDNFEWLRGYSEKFGIYAVDFEDPARPRIPKESA 480
C.gestroi           TEHLKEMLKAIHIDGVNVIGYTAWSLIDNFEWLRGYSERFGIHEVNFIDPSRPRTPKESA 76
                    ************ *****:*******:*********:*:***: *:* **:*** *****

C.formosanus        RVLTEIFKTRQIPERFRD---------------------------------------- 495
N.takasagoensis     KVLTEIFNTRKIPDRFLD---------------------------------------- 490
N.koshunensis       KVLAEIMNTRKIPERFRD---------------------------------------- 498
C.gestroi           RILTEIFSTRQIPERFRDLHIQGATTYTKKKINSNTELLLVPLKTTVRSWKTVVLNKHLQ 136
                    ::*:**:.**:**:** *
```

Fig. 4. Multiple alignments of β-glucosidase from termites: *Coptotermes gestroi* (GT566634.1), *Coptotermes formosanus* (GQ911585.1), *Nasutitermes takasagoensis* (AB508959.1) and *Neotermes koshunensis* (AB073638.2). GenBank accession numbers are in parentheses. Asterisks represent the same amino acid residues.

Discussion

In this study, we generated a set of *Coptotermes gestroi* ESTs, permitting a first glance at its transcriptome. We did not use whole bodies of these insects because the digestive tract of lower termites has a dilated portion in the hindgut region that contains several microbial symbionts, such as flagellated protists and prokaryotes (Ohkuma, 2008). Therefore, in order to minimize contamination by sequences from such endosymbionts, we prepared cDNA from RNA extracted only from heads of termite workers, collected in the field.

From the assembled ESTs, 374 did not match any known sequence (no hits). At first glance, this result suggests that these sequences could represent specific termite genes that have not been identified by comparison with GenBank since only 115 *C. gestroi* sequences are available in this collection. In order to confirm their coding potential, an open reading frame (ORF) analysis was carried out. ESTScan analyses showed that 129 (34.49%) sequences yielded a positive result and indeed represent potential new termite-or species-specific genes. Two analyses support this hypothesis. Firstly, 19 of these 129 sequences showed counterparts in a databank of termite EST sequences. Secondly, the specific GC content (table 1) of the individual bases within the codons, i.e. GC1, GC2 and GC3, of 'no hit' ORFs is very similar to that of ORFs representing genes of other termites (GenBank) or from

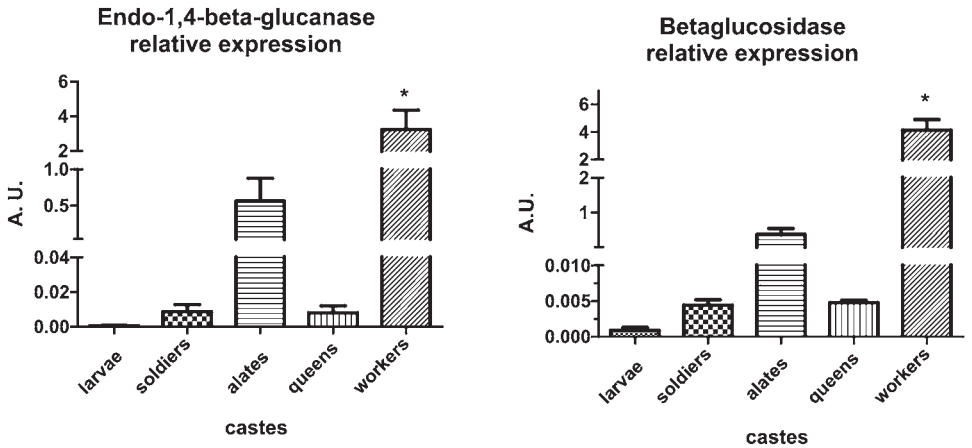

Fig. 5. Relative expression levels of (a) endo-β-1,4-glucanase and (b) β-glucosidase in different castes measured by real-time PCR. The Y-axis indicates gene expression in arbitrary units and data are mean values of three independent preparations. Significant differences are indicated (workers have a significant difference compared to the other castes; Tuckey: *, $P<0.05$).

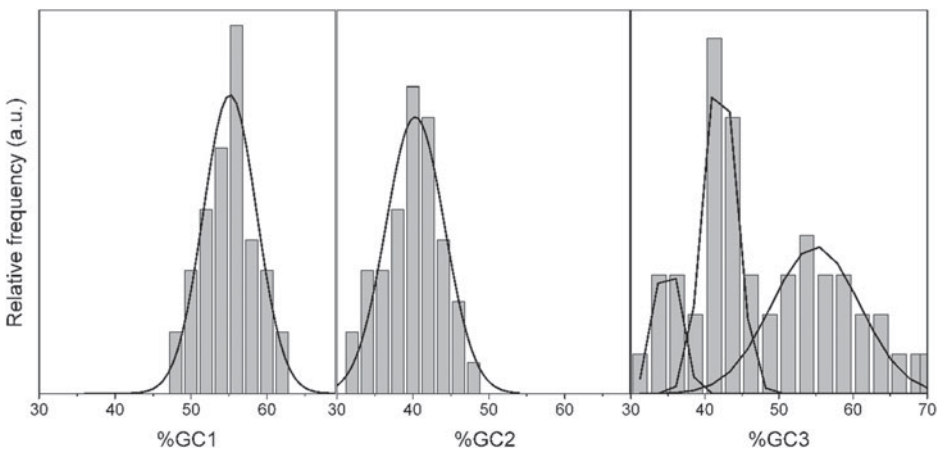

Fig. 6. Distribution of GC content at different positions in the *C. gestroi* hits sequences.

the 266 ORFs of *C. gestroi* reported in this work. This pattern suggests that these sequences are indeed suitable for translation and may be treated as true genes. Another explanation for the apparently large quantity of no hits is that they may represent 'taxonomically restricted genes' (TRGs). Every eukaryotic genome contains 10–20% of genes without any significant sequence similarity to genes of other species that are classified as 'orphans' or TRGs (Wilson *et al.*, 2005). Although such genes have arisen in the genomes of every group of organisms studied so far, they have received comparatively little attention, and their functions remain largely unknown (Khalturin *et al.*, 2009).

The 245 remaining no hits were also 'no ORFs' and might have been considered an artifact. However, non-hypothesis-driven gene expression studies on honeybee (Colonello-Frattini & Hartfelder, 2009) and fruit fly (Tupy *et al.*, 2005) transcripts also identified 'no ORF' ESTs that were clearly expressed in different organs. These sequences are thought to have a potential function as nuclear RNAs (Sawata *et al.*, 2004), micro RNAs (Bartel, 2004) or long non-coding RNAs (Mercer *et al.*, 2009). Therefore, the 'no ORF' *C. gestroi* sequences should be analyzed further to identify possible functions.

When comparing *C. gestroi* sequences with those of other insects, the distribution of the first (best) hit demonstrated higher similarity to the honey bee than to dipteran species (fig. 1d). At first glance, this could suggest some similarity based on common molecular mechanisms to suit social life. However, the similarity between *C. gestroi* and *A. mellifera* is equivalent to the similarity to the flour beetle, *Tribolium castaneum*, a non-social insect species. The lack of a clear pattern may reflect the relative paucity of sequences from hemimetabolous species in the GenBank database. This kind of developmental profile may have an association with defined genomic structures that are not yet available for comparison.

The sequences were categorized by gene ontology in biological processes (fig. 2). A large portion of the genes (43%) was classified as having a role in cellular physiological process and metabolism, which represent basic functions. This result was also found in a previous study of *Apis mellifera* ESTs (Nunes *et al.*, 2004).

We analyzed the dispersion of GC content at the individual positions from putative hit-ORFs. As expected, at positions GC1 and GC2, the dispersion is lower (Sabater-Muñoz *et al.*, 2006) and presents a homogeneous Gaussian distribution (fig. 6). At GC3, dispersion is wider; but, curiously, we identified three distribution patterns. We hypothesized that theses distributions could represent different classes of genes, but the analysis with the available set of genes failed to indicate a clear pattern.

Even though we were careful not to contaminate the samples with gut material, 42 uniESTs clearly derived from non-insect genes were also identified: ten sequences from various types of bacteria, 26 from protist, four from fungi and two from plants. Some bacterial sequences produced positive hits against a databank from marine metagenomics (Kannan *et al.*, 2007), soil fungi and soil bacteria, probably representing DNA from soil particles. Most of the protist sequences presented *Trichomonas vaginalis* as a first hit. This does not mean, however, that this species is present in *C. gestroi*. This result probably indicates the presence of parabasalian protist, which has been reported in the termite digestive tract (Ohkuma, 2008).

The most representative group found in the library was composed of genes with enzymatic activity. Two cellulase-encoding sequences were of particular interest to us because of their potential biotechnology applications. Only around ten years ago the first termite endogenous cellulase was reported (Tokuda *et al.*, 1997), changing the general belief that termite cellulase is produced only by the symbiotic microbes presented in their digestive tract.

The role of these celullases in *C. gestroi* has not yet been elucidated, but the possibility of mimetizing termites for use of cellulose to convert biomass into energy resources has attracted much interest (Scharf & Tartar, 2008). Termites initiate the digestion of wood by mechanical degradation of the substrate with their powerful mandibles. Unlike some biotechnology processes, these mild conditions allow digestive enzymes to act quickly, with the first cellulolytic attack probably occurring in the mouth. Since we prepared cDNA of workers' heads, the cellulases we identified probably are present in the salivary glands. This would be consistent with previous reports that found a gene similar to the *C. gestroi* endo-β-1,4-glucanase (GT567748.1) gene expressed in the salivary glands of *C. formosanus* (Nakashima *et al.*, 2002) and *R. flavipes* (Zhou *et al.*, 2007).

After this initial digestion step, the substrate passes to the hindgut where a more complex enzymatic combination – possibly a dual cellulolytic system between termites and symbionts – works together to accomplish cellulose degradation and lignin removal (Todaka *et al.*, 2007). The other enzyme, β-glucosidase (GT566634.1), should possibly act in this step. Enzymes responsible for lignin degradation were not found in our library. However, recent studies reported that analyses of cDNA libraries from host termites and symbiotic protists presented genes related to lignin degradation (Kudo, 2009; Tartar *et al.*, 2009). A research conducted with the dampwood termite, *Zootermopsis angusticollis*, demonstrated that there is lignin degradation in the gut of this wood-feeding insect (Geib *et al.*, 2008).

To verify the expression of the cellulases in *C. gestroi* species, a quantitative expression analysis was conducted using RNA samples from different castes (larvae, workers, nymphs, alates, soldiers and queens). Both genes presented a higher expression in workers, compared to the other castes ($P < 0.05$). This result was expected considering that this caste executes the feeding of the colony, and it was consistent with a previous study conducted with *Reticulitermes flavipes* that found high levels of expression of this cellulase in workers, nymphs, alates and supplementary reproductives (Scharf *et al.*, 2005). The β-glucosidase gene is similar to another homolog, the Neofem 2, which is believed to play a role in reproductive suppression in the termite, *Cryptotermes secundus* (Weil *et al.*, 2007). A knockdown of Neofem 2 was conducted in queens, making the workers believe that the colony was queenless (Korb *et al.*, 2009). Nevertheless, our experiments show that, in *C. gestroi*, the expression of this gene is extremely low in queens, supporting the idea that this enzyme works together with others to efficiently degrade cellulose. Earlier RT-PCR experiments demonstrate that workers of the lower termite, *Neotermes koshunensis*, are known to express β-glucosidase in the salivary glands (Tokuda *et al.*, 2002), supporting our findings since we generated a cDNA library from workers' heads. A differential study of the digestive cellulase expression among castes in two termite species showed that minor workers of *Nasutitermes takasagoensis* presented higher rates of expression of β-glucosidase, when compared to soldiers, medium and major workers, revealing a division of labor controlled by nutrition (Fujita *et al.*, 2008). Even though the rates were differentially expressed, the site of expression was higher in the midgut for all castes. In the termite *Hodotermopsis sjostesti*, however, the expression was higher in the salivary glands for the worker caste and in the hindgut for the soldiers (Fujita *et al.*, 2008). Further investigation are been carried out to locate the sites of expression of these cellulases in *C. gestroi*.

In summary, this is the first survey of the *C. gestroi* transcriptome, with a focus on the heads of workers, the caste involved in cellulose degradation. As previously discussed, investigations of the termite's lifestyle, combined with molecular analyses of its physiology and symbionts, indicate that integration of mechanical processes and enzymatic treatment of wood may be the key for efficient cellulose utilization. Clearly, these insects deserve further in-depth molecular studies to elucidate the metabolic pathways employed in lignocellulose degradation.

We are currently investigating the properties of these putative cellulase enzymes and their possible utilization in biotechnological processes.

Acknowledgements

We thank Klaus Hartfelder for helpful comments on previous versions of the manuscript and Nicolan Conran, HEMOCENTRO-UNICAMP, for help with English revision. This study was supported by FAPESP, Brazil (No. 06/59086-8 and No. 08/50114-4).

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17), 3389–3402.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinsk, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarski, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000) Gene

ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**(1), 25–29.

Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**(2), 281–297.

Colonello-Frattini, N. & Hartfelder, K. (2009) Differential gene expression profiling in mucus glands of honey bee (Apis mellifera) drones during sexual maturation. *Apidologie* **40**, 481–495.

Ewing, B. & Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**(3), 186–194.

Fujita, A., Miura, T. & Matsumoto, T. (2008) Differences in cellulose digestive systems among castes in two termite lineages. *Phisiological Entomology* **33**, 73–82.

Geib, S.M., Filley, T.R., Hatcher, P.G., Hoover, K., Carlson, J.E., del Mar Jimenez-Gasco, M., Nakagawa-Izumi, A., Sleighter, R.L. & Tien, M. (2008) Lignin degradation in wood-feeding insects. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 12932–12937.

Grassé, P.P. (1949) Ordre des Isoptères ou termites. pp. 408–544 *in* Grassé, P.P. (*Ed.*) *Traité de Zoologie*. v. 9, Masson, Paris. *Physiological Entomology* **33**, 73–82.

Huang, X. & Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9**(9), 868–877.

Iseli, C., Jongeneel, C.V. & Bucher, P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. pp. 138–148 in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, AAAi Press, Menlo Park, CA, USA.

Jenkins, T.M., Jones, S.C., Lee, C.Y., Forschler, B.T., Chen, Z., Lopez-Martinez, G., Gallagher, N.T., Brown, G., Neal, M., Thistleton, B. & Kleinschmidt, S. (2007) Phylogeography illuminates maternal origins of exotic *Coptotermes gestroi* (Isoptera: Rhinotermitidae). *Molecular Phylogenetics and Evolution* **42**(3), 612–621.

Kannan, N., Taylor, S.S., Zhai, Y., Venter, J.C. & Manning, G. (2007) Structural and functional diversity of the microbial kinome. *Public Library of Science Biology* **5**(3), e17.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T.C. (2009) More than just orphans: are taxonomically restricted genes important in evolution? *Trends in Genetics* **25**(9), 404–413.

Kirton, L.G. (2005) The importance of accurate termite taxonomy in the broader perspective of termite management. pp. 1–7 *in* Lee, C.-Y. & Robinson, W.H. (*Eds*) *Proceedings of the Fifth International Conference on Urban Pests*. P & Y Design Network, Penang, Malaysia.

Korb, J., Weil, T., Hoffmann, K., Foster, K.R. & Rehli, M. (2009) A gene necessary for reproductive suppression in termites. *Science* **324**, 758.

Kudo, T. (2009) Termite-Microbe Symbiotic System and Its Efficient Degradation of Lignocellulose. *Bioscience Biotechnology and Biochemistry* **73**(12), 2561–2567.

Mercer, T.R., Dinger, M.E. & Mattick, J.S. (2009) Long non-coding RNAs: insights into functions. *Nature Reviews Genetics* **10**(3), 155–159.

Nakashima, K., Watanabe, H., Saitoh, H., Tokuda, G. & Azuma, J.I. (2002) Dual cellulose digesting system of the wood-feeding termite, *Coptotermes formosanus* Shiraki. *Insect Biochemistry and Molecular Biology* **32**(7), 777–784.

Nunes, F.M., Valente, V., Sousa, J.F., Cunha, M.A., Pinheiro, D. G., Maia, R.M., Silva, W.A., Araujo, D.D., Costa, M.C.R., Martins, W.K., Carvalho, A.F., Monesi, N., Nascimento, A.M., Peixoto, P.M.V., Silva, M.F.R., Ramos, R.G.P., Reis, L.F.L., Dias-Neto, E., Souza, S.J., Simpson, A.J.G., Zago, M.A., Soares, A.E.E., Bitondi, M.M. G., Espreafico, E.M., Espindola, F.S., Paco-Larson, L., Simoes, Z.L.P., Hartfelder, K. & Silva, W.A. (2004) The use of Open Reading frame ESTs (ORESTES) for analysis of the honey bee transcriptome. *BioMed Central Genomics* **5**(1), 84–1.

Ohkuma, M. (2008) Symbioses of flagellates and prokaryotes in the gut of lower termites. *Trends in Microbiology* **16**(7), 345–352.

Ronaghi, M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Research* **11**(1), 3–11.

Sabater-Muñoz, B., Legeai, F., Rispe, C., Bonhomme, J., Dearden, P., Dossat, C., Duclert, A., Gauthier, J.P., Ducray, D.G., Hunter, W., Dang, P., Kambhampati, S., Martinez-Torres, D., Cortes, T., Moya, A., Nakabachi, A., Philippe, C., Prunier-Leterme, N., Rahbé, Y., Simon, J.C., Stern, D.L., Wincker, P. & Tagu, D. (2006) Large-scale gene discovery in the pea aphid *Acyrthosiphon pisum* (Hemiptera). *Genome Biology* **7**(3), R21.

Sawata, M., Takeuchi, H. & Kubo, T. (2004) Identification and analysis of the minimal promoter activity of a novel non coding nuclear RNA gene, AncR-1, from the honeybee (*Apis mellifera L.*). *Rna* **10**(7), 1047–1058.

Scharf, M.E. & Tartar, A. (2008) Termites digestome as sources for novel lignocellulases. *Biofuels, Bioproducts and Biorefining* **6**(2), 540–552.

Scharf, M.E., Wu-Scharf, D., Pittendrigh, B.R. & Bennett, G.W. (2003) Caste- and development-associated gene expression in a lower termite. *Genome Biology* **4**(10), R62.

Scharf, M.E., Wu-Scharf, D., Zhou, X., Pittendrigh, B.R. & Bennett, G.W. (2005) Gene expression profiles among immature and adult reproductive castes of the termite *Reticulitermes flavipes*. *Insect Molecular Biology* **14**(1), 31–44.

Tartar, A., Wheeler, M.M., Zhou, X., Coy, M.R., Boucias, D.G. & Scharf, M. (2009) Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite *Reticulitermes flavipes*. *Biotechnology for Biofuels* **2**, 25.

Todaka, N., Moriya, S., Saita, K., Hondo, T., Kiuchi, I., Hirotoshi, T., Ohkuma, M., Piero, C., Hayashisaky, Y. & Kudo, T. (2007) Enviromental cDNA analysis og the genes involved in lignocellulose digestion in the symbiotic protist community of Reticulitermes speratus. *Federation of European Microbiological Societies Microbiology Ecology* **59**, 592–599.

Tokuda, G., Watanabe, H., Matsumoto, T. & Noda, H. (1997) Cellulose digestion in the wood-eating higher termite, *Nasutitermes takasagoensis* (Shiraki): distribution of cellulases and properties of endo-beta-1,4-glucanase. *Zoological Science* **14**(1), 83–93.

Tokuda, G., Saito, H. & Watanabe, H. (2002) A digestive β-glucosidase from the salivary glands of the termite, *Neotermes koshunensis* (Shiraki): distribution, characterization and isolation of its precursor cDNA by 5′- and 3′-RACE amplifications with degenerate primers. *Insect Biochemistry and Molecular Biology* **32**(12), 1681–1689.

Telles, G. & da Silva, F. (2001) Trimming and clustering sugarcane ESTs. *Genetics and Molecular Biology* **24**, 17–23.

Tupy, J.L., Bailey, A.M., Dailey, G., Evans-Holm, M., Siebel, C. W., Misra, S., Celniker, S.E. & Rubin, G.M. (2005) Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America* **102**(15), 5495–5500.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. & Speleman, F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology* **3**(7), RESEARCH0034.

Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T.H., Stege, J.T., Cayouette, M., McHardy, A. C., Djordjevic, G., Aboushadi, N., Sorek, R., Tringe, S.G., Podar, M., Martin, H.G., Kunin, V., Dalevi, D., Madejska, J., Kirton, E., Platt, D., Szeto, E., Salamov, A., Barry, K., Mikhailova, N., Kyrpides, N.C., Matson, E.G., Ottesen, E.A., Zhang, X., Hernández, M., Murillo, C., Acosta, L.G., Rigoutsos, I., Tamayo, G., Green, B.D., Chang, C., Rubin, E.M., Mathur, E.J., Robertson, D.E., Hugenholtz, P. & Leadbetter, P. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**(7169), 560–565.

Watanabe, H. & Tokuda, G. (2001) Animal cellulases. *Cellular and Molecular Life Sciences* **58**(9), 1167–1178.

Weil, T., Rehli, M. & Korb, J. (2007) Molecular basis for the reproductive division of labour in a lower termite. *BMC Genomics* **8**, 198.

Wilson, G.A., Bertrand, N., Patel, Y., Hughes, J.B., Feil, E.J. & Field, D. (2005) Orphans as taxonomically restricted and ecologically important genes. *Microbiology* **151**, 2499–2501.

Wu-Scharf, D., Scharf, M.E., Pittendrigh, B.R. & Bennett, G.W. (2003) Expressed sequence tags from a polyphenic *Reticulitermes flavipes* (Isoptera: Rhinotermitidae) cDNA library. *Sociobiology* **41**(2), 479–490.

Zhou, X., Oi, F.M. & Scharf, M.E. (2006) Social exploitation of hexamerin: RNAi reveals a major caste-regulatory factor in termites. *Proceedings of the National Academy of Sciences of the United States of America* **103**(12), 4499–4504.

Zhou, X., Smith, J.A.M., Oi, F.M., Koehler, P.G., Bennett, G.W. & Scharf, M.E. (2007) Correlation of cellulose gene expression and cellulolytic activity throughout the gut of the termite Reticulitermes flavipes. *Gene* **395**, 29–39.