OPEN ACCESS Freely available online

PLoS one

Odorant Binding Proteins of the Red Imported Fire Ant, *Solenopsis invicta*: An Example of the Problems Facing the Analysis of Widely Divergent Proteins

Dietrich Gotzek[1]*[9¤], Hugh M. Robertson[2,9], Yannick Wurm[1], DeWayne Shoemaker[3]

1 Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland, **2** Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **3** Center for Medical, Agricultural, and Veterinary Entomology, United States Department of Agriculture – Agricultural Research Service, Gainesville, Florida, United States of America

Abstract

We describe the odorant binding proteins (OBPs) of the red imported fire ant, *Solenopsis invicta*, obtained from analyses of an EST library and separate 454 sequencing runs of two normalized cDNA libraries. We identified a total of 18 putative functional OBPs in this ant. A third of the fire ant OBPs are orthologs to honey bee OBPs. Another third of the OBPs belong to a lineage-specific expansion, which is a common feature of insect OBP evolution. Like other OBPs, the different fire ant OBPs share little sequence similarity (~20%), rendering evolutionary analyses difficult. We discuss the resulting problems with sequence alignment, phylogenetic analysis, and tests of selection. As previously suggested, our results underscore the importance for careful exploration of the sensitivity to the effects of alignment methods for data comprising widely divergent sequences.

Citation: Gotzek D, Robertson HM, Wurm Y, Shoemaker D (2011) Odorant Binding Proteins of the Red Imported Fire Ant, *Solenopsis invicta*: An Example of the Problems Facing the Analysis of Widely Divergent Proteins. PLoS ONE 6(1): e16289. doi:10.1371/journal.pone.0016289

Editor: Corrie Moreau, Field Museum of Natural History, United States of America

Received September 16, 2010; **Accepted** December 10, 2010; **Published** January 31, 2011

Funding: Part of this study was funded by USDA-AFRI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gotzekd@si.edu

9 These authors contributed equally to this work.

¤ Current address: Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, D.C., United States of America

Introduction

Chemosensory systems play a central role in the way insects perceive their surroundings and are critical to finding mates, food, and oviposition sites. These olfactory and gustatory systems rely on at least two distinct protein families to translate environmental chemical signals into action potential. Since these proteins are thought to be the first interactants with the odorant semiochemicals they pose an important discriminatory filter during perception of chemosensory stimuli. Odorant binding proteins (OBPs) and chemosensory proteins (CSPs) are small, water-soluble, extracellular proteins, which bind hydrophobic semiochemicals in the lymphatic cavities of the sensory organs and transport them to the second class of proteins, the chemoreceptors [1]. Odorant binding proteins were first thought to have highly specific binding affinities to certain semiochemicals and to be exclusively expressed in the antennae of insects. However, both hypotheses have proven not to be correct. Although some OBPs appear to be exclusively involved in odor detection, others are expressed in various tissues and during different life stages (see [2] for a review), which suggests that the protein family can serve multiple functions. Whole genome surveys have shown that OBPs and CSPs are highly divergent protein families and are characterized by lineage-specific expansions, presumably driven largely by adaptation. To date, most insect genomes have been shown to contain around 40–55 OBPs and 4–8 CSPs [3]. The honey bee, *Apis mellifera*, is unusual in that it contains a low number of OBPs, only 21, and no significant expansion of CSPs [4]. Until recently [5,6], no OBPs and only CSPs had been found in the antennae of ants, causing Calvello et al. [7] to speculate that functionally, the OBPs have been replaced by CSPs in these Hymenopterans. This hypothesis is consistent with the large number of CSPs in the red imported fire ant, *Solenopsis invicta* Buren, 1972, which possesses at least 14 CSPs [6]. However, the number of OBPs in this ant has not been determined.

For the present study, we attempted to identify and enumerate the full repertoire of OBPs in this ant. While such an endeavor previously was not feasible, the recent development of genomic resources for this ant now affords us with such an opportunity. The first such resource was an expressed sequence tag (EST) project in which >22,000 cDNAs were sequenced from both ends using Sanger termination methods, resulting in 21,715 ESTs representing 11,864 putatively different transcripts [7]. González et al. [6] recently described the chemosensory proteins (CSPs) revealed by the Sanger-based EST project, and here we describe the OBPs. The EST library [9] was augmented with data from two sequencing runs of massively parallel pyrosequencing using the Roche 454 FLX machine generating a total of 533,091 reads averaging 236 bp long and mined for sequences encoding OBPs.

To date, only one OBP has been described in detail from any ant, the locus *general protein-9* (*Gp-9*), which is implicated in

regulating colony queen number in *S. invicta* and closely related fire ants [10,11,12,13]. The *Gp-9* locus is unusual for an OBP in several ways – it displays high levels of variation in the coding region, is highly expressed, and found in the hemolymph of all castes [12,13]. To provide the foundation for future studies of other fire ant OBPs, identification of all members of the OBP gene family in *S. invicta* is needed, and is the goal of the present study. We also use our data to further emphasize a general problem facing studies of widely divergent molecular sequences (which is one characteristic of insect OBPs) since the results obtained heavily depended on the underlying multiple sequence alignment method used, stemming no doubt from the large sequence divergence of these proteins. Our study highlight the necessity to carefully consider whether current analytical methods are adequate to analyze increasingly divergent molecular sequences (e.g., [14]) as well as the importance of investigating the influence of alignment methods on results.

Results

Identification of OBPs

The final assembly contains 18 contigs encoding *S. invicta* OBPs (SiOBPs), summarized in Table 1. One additional sequence similar to an OBP was also found (SiJWD04CAE), but it was so highly degenerate that it was not named and was dropped from all further analyses (this sequence shares the closest sequence identity with SiOBP3 [*Gp-9*]). Only a few of these contigs were full-length in the automated assembly, but manual re-assembly of the reads that belong to each contig allowed extension of 5′ and/or 3′ ends, generally yielding at least the entire coding sequence, and generally reaching a polyA tail. It is not possible to be confident that the 5′ ends of these contigs are the true transcription start site, so the cDNA lengths given in Table 1 are not necessarily definitive. It appears that the automated assembly was conservative in trimming reads for low quality ends, and in not extending contigs beyond apparent length differences in the constituent reads. Although most sequences derive from contigs comprising large numbers of 454 reads of ~250 bases, eleven also have longer Sanger reads from the earlier published EST project [7], indeed three sequences are entirely from Sanger reads, with SiOBP18 being derived from a single Sanger read. Together with SiOBP17, these are also the two most problematic sequences. SiOBP17 appears to be partially unspliced with apparent intronic sequence interrupting the coding region, which is otherwise full-length, while the SiOBP18 read encodes only an internal part of this OBP, despite being quite long. The numbers of 454 reads contributing to each contig gives a rough estimate of their expression levels, with several clearly being well-expressed; SiOBP3, which has already been extensively studied as *Gp-9*, has an extremely large number of reads. The manual assembly of the 454 reads for several OBPs revealed that commonly more than one polyadenylation site was employed (listed in Table 1), and for those we employed the longest 3′ UTR available. Contig sequences encoding SiOBPs 1–16, excluding SiOBP3 which is already highly represented in GenBank as *Gp-9*, have been submitted to GenBank (HQ853350–HQ853364).

Table 1. Details of the *Solenopsis invicta* odorant binding proteins.

Gene	cDNA	TotAA	MatAA	454	Sanger	C	PolyA
SiOBP1	857	139	120	99	4	6	multi
SiOBP2	804	152	135	41	0	4	single
SiOBP3	631	153	134	>1200**	8	6	single
SiOBP4	638	153	134	14	0	6	none
SiOBP5	730	144	122	34	0	6	multi
SiOBP6	591	146	128	4	0	6	none
SiOBP7	623	133	116	335	2	6	single
SiOBP8	634	153	126	0	3	4	single
SiOBP9	859	129	109	21	0	6	multi
SiOBP10	747	147	131	43	1	6	single
SiOBP11	662	149	125	15	0	6	single
SiOBP12	936	174	154	35	4	6	multi
SiOBP13	740	160	144	14	9	6	single
SiOBP14	781	162	146	80	0	6	single
SiOBP15	894	162	140	112	9	6	multi
SiOBP16	660	171	155	40	4	6	multi
SiOBP17N*	834	168	148	0	1	6	single
SiOBP18N	654	>77	>77	0	1	6	none

*This single Sanger read appears to be partially unspliced and frameshifted.
**The total number of 454 reads contributing to this SiOBP3/*Gp-9* contig is unclear, because it strangely assembled in several different non-overlapping contigs.
The columns are: Gene – number we are assigning; cDNA – length of cDNA in base pairs, excluding polyA tail; TotAA – conceptual precursor protein length including signal sequence; MatAA – mature secreted protein length excluding signal sequence according to PSORTII; 454 – number of 454 reads contributing to contig; Sanger – number of Sanger reads contributing to contig; C – number of conserved cysteines; PolyA – presence of single or multiple poly-adenylation sites.
doi:10.1371/journal.pone.0016289.t001

Multiple sequence alignment

Due to the significant sequence divergence of the OBPs used in this study (overall ~20% protein sequence identity), we were skeptical of the accuracy of any single multiple sequence alignment (MSA) to infer homologous amino acid residues of these divergent proteins. Hence, we compared six MSA methods, which employ widely different alignment methodologies and have been shown to perform well and/or are commonly used (Table 2). Additionally, we conducted simultaneous alignment and topology inference in a Bayesian framework using BAli-Phy for both the *Apis* and *Solenopsis* OBPs (AmOBPs; [4] and SiOBPs, respectively). Since this approach is generally considered to be conceptually superior to the generally used two phase methods, which separate alignment estimation and tree topology inference [15,16], we considered the alignments and topologies derived from these searches to be the "true" tree.

It is common practice to account for the wide divergence between OBPs by removing signal peptides and less often the C-terminal residues prior to multiple sequence alignment and, hence, to restrict the following analyses to the presumed more conserved "core" of the proteins [e.g., 17,18,19,20]. However, Wong *et al.* [14] advise against eliminating difficult blocks from alignments, since some of these may still contain informative sites and their removal does not necessarily result in more concordant inferences. Additionally, they show that it is possible to make inferences despite considerable alignment uncertainty. Hence we did not remove areas of uncertain alignment, especially since the AU plots of both the *Solenopsis* and *Apis* BAli-Phy alignments suggest that there are still high quality alignment blocks within these "problematic" areas to warrant their inclusion in the overall alignment procedure (Figure 1b). This is especially true for the signal peptides, which are most often removed before analyses

Table 2. Details of the multiple sequence alignment (MSA) methods used and maximum likelihood phylogenies estimated from them.

rank	alignment	version	length	core length	LnL	parsimony	tree size	average aLRT	RF distance ant/bee	% seq.identity	reference
	BAli-Phy	2.0.2	253	130	−5736.1899	1229	14.24144		na	0.222	[51]
1	PRANK	1.0	332	152	***−10671.224***	***2284***	***28.84963***	0.86275	***4*/2**	***0.223***	[67]
2	MUSCLE	3.6	206	117	−11160.1619	2520	34.76337	***0.877306***	***4***/4	0.203	[68]
3	MAFFT	6	209	115	−10900.89887	2448	34.152	0.866611	8/8	0.203	[69]
4	CLUSTALW	2.0.12	197	111	−10966.46551	2474	36.24757	0.843056	8/10	0.191	[70]
5	OPAL	1.0.3	219	127	−11178.67281	2511	37.78651	0.83475	8/10	0.203	[71]
6	SATCHMO	2.06	232	121	−11159.04352	2521	42.12012	0.792278	12/10	0.193	[72]

We define the core length as the number of character positions from the first to the last of the characteristic cysteine residues (C1–C6) of the OBPs. The log-likelihoods (LnL), parsimony informative characters, tree size, and average approximate likelihood-ratio tests (aLRT) are derived from the ML analyses. Robinson-Foulds tree distances (RF distance) are calculated by comparing the ant and bee MAP trees to the ML trees derived under the other MSA methods. Best scores of the MSAs compared to the BAli-Phy MAP are in highlighted in bold italics.

doi:10.1371/journal.pone.0016289.t002

[17,18,19,20]. Moreover, we do not consider the "core" sequences to be inherently more informative than the outside areas, since the lengths of the core (which we define as ranging from C1 to C6) differed greatly between MSA methods (Table 2). Preliminary analyses also suggest that removing the outer areas do not change significantly the topology derived from them (data not shown). Additionally, both the Steel [21] and Xia [22] tests indicated high levels of sequence saturation for our dataset for all MSAs (not shown), suggesting that the dataset contains little useful evolutionary signal.

Phylogenetic analyses

Despite the great difference in alignment lengths and the pronounced sequence saturation as shown by the Steel and Xia tests, most MSAs still yielded highly similar tree topologies. Several clades were consistently recovered and the midpoint root was generally placed in the same position across all MSAs (Figure 1a, Figure 2). So despite the obvious problems to align the widely divergent OBP dataset, we conclude there is enough phylogenetic information in the alignments to at least draw tentative conclusions regarding the evolution of fire ant OBPs. The maximum likelihood and two Bayesian searches recovered highly similar tree topologies, with the Bayesian trees generally being less resolved, especially at the deeper nodes.

Selection analyses

Forêt and Maleszka [4] described evidence of positive selection in the AmOBP expansion, so we used estimates of dN/dS (ω) to examine whether the same was true of the SiOBP expansion. Given the uncertainties of alignment and topology, we conducted site-specific tests of selection [23,24,25,26] on the two best (PRANK, MUSCLE), the shortest (CLUSTAL), and the *Solenopsis* MAP alignments (Table 3). Site specific analyses of all OBPs combined showed no evidence of positive selection for either the PRANK or MUSCLE alignments. The ant MAP alignment, however, showed a signature of positive selection using the M1a (neutral)–M2a (selection) comparison, but not the M7–M8 comparison, which has been shown to be less robust (but more powerful) than the M1a–M2a comparison [23]. For the M1a–M2a comparison, the Bayes Empirical Bayes (BEB) method identified two amino acid positions in the core (aa81 with PP = 0.991 and aa128 with PP = 0.979) as being under positive selection ($\omega = 2.9485$). Even though the M7–M8 comparison was not significant, the BEB indicated the same sites (aa81 and aa128) have elevated ω estimates ($\omega = 1.8266$). The CLUSTAL alignment contained evidence of positive selection for both tests (M1a–M2a: $\omega = 2.0973$, aa20 PP = 0.972, aa25 PP = 0.981, aa49 PP = 0.981, aa70 PP = 0.955, aa133 PP0.995, aa177 PP 0.976, aa178 PP = 0.999; M7–M8: $\omega = 3.5508$, aa178 PP = 0.964). The two amino acid positions in the core of the CLUSTAL alignment identified to be under positive selection (aa70 and aa133) are not identical to those of the MAP alignment, suggesting that the tests of positive selection using different alignments are not picking up the same evolutionary signals.

We tested whether these signatures of positive selection were associated with the ant-specific expansion, which we tested using branch-specific tests of selection [27,28]. Oddly enough, the LRT comparing the null and alternative hypotheses showed significant differences in the PRANK and MUSCLE MSAs, suggesting episodes of positive selection on this branch. However, in both cases the estimates of ω for this branch were <1 and even lower than the estimate of ω across all other branches. This pattern is consistent with relaxed selection, especially since it is coupled with a rapid gene expansion in this clade. The explanation of increased purifying selection to explain this pattern seems less likely to us. However, the branch-specific test for selection averages the estimates of ω across the whole sequence length and as a result may lack power [29] and obscure episodes of positive selection restricted to one or very few sites. Hence, we also applied branch-site analyses of selection [26,30] on the branch leading to the ant-specific expansion. These tests were not significant for any of the datasets, supporting our interpretation of lack of positive selection.

Discussion

We identified a total of 19 OBPs in *S. invicta*, of which 18 appear to be putatively functional. The red imported fire ant thus appears to possess a small set of OBPs similar to that of the honey bee *Apis mellifera* (21 OBPs [4]). Although this estimate may slightly change with the assembly and annotation of the complete fire ant genome [31], the fire ant OBP repertoire is one of the smallest reported among insects, with only the pea aphid *Acyrthosiphon pisum* and the body louse, *Pediculus humanus*, appearing to have fewer OBPs (15 and 5, respectively; [20,32]). Preliminary scans of the coding regions (CDS) and peptide libraries of the jumping ant, *Harpegnathos saltator*, and the carpenter ant, *Camponotus floridanus*, genomes (both version 3.3 [33]) found twelve and seven OBPs

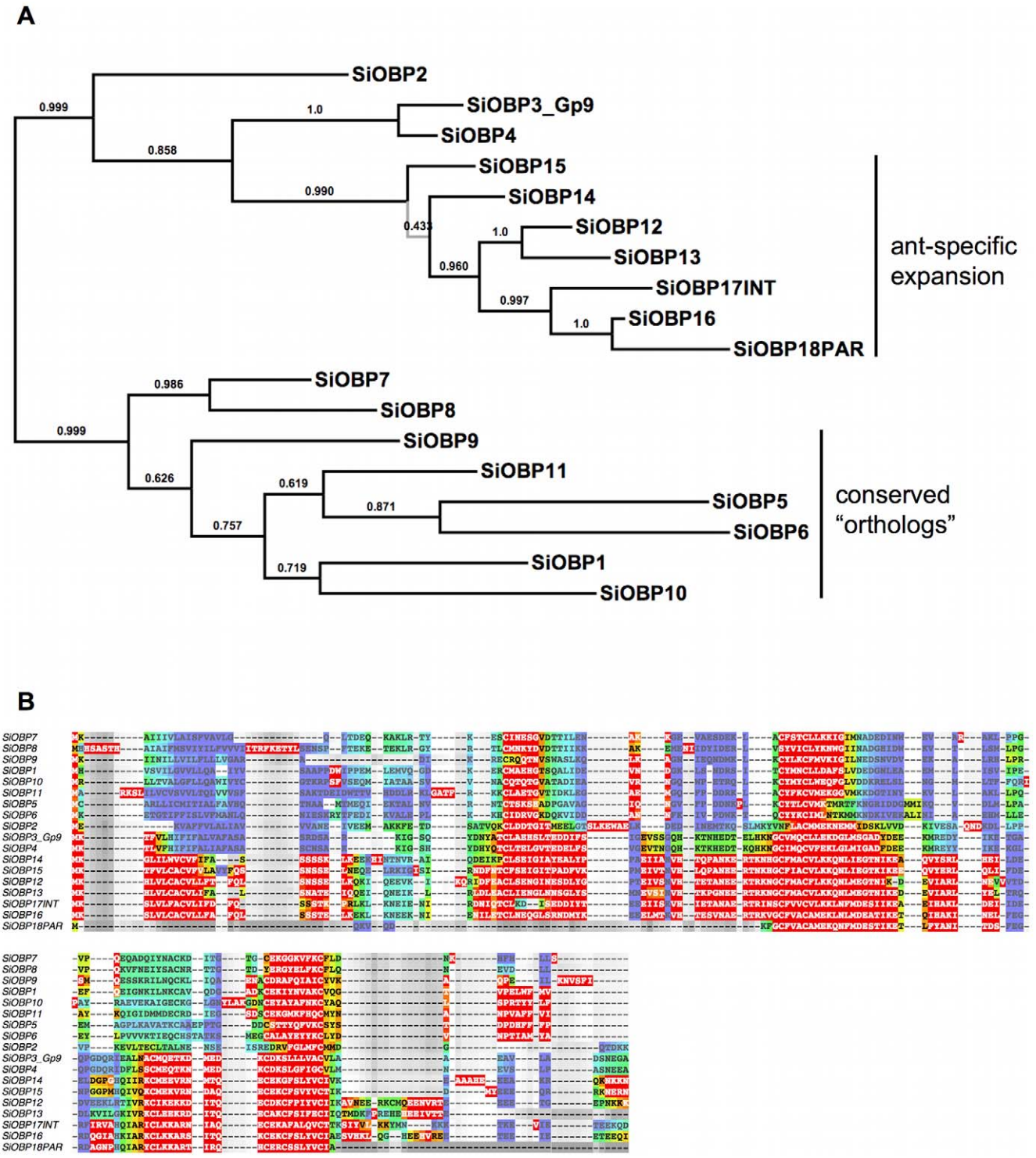

Figure 1. Maximum a-posteriori (MAP) phylogeny and alignment of the *Solenopsis invicta* (SiOBP) odorant binding proteins. A. The *S. invicta* MAP phylogeny. The branch in grey is collapsed in the 50% consensus tree. Branch support is posterior probabilities derived from 3241 samples taken after the burn-in was discarded. Even though the node support in the conserved ortholog clade is relatively poor, the exact same topology of the orthologs was recovered in the honey bee MAP tree (not shown), suggesting that the branching pattern is accurate. B. The *S. invicta* MAP-AU plot. The quality of the alignment is indicated through a heat map. Red (warm colors) indicates areas of high quality alignment, blue (cold colors) signifies areas of low certainty. Note that there are considerable high quality alignment blocks in the N-terminal signal peptide and the C-terminal protein tail.
doi:10.1371/journal.pone.0016289.g001

respectively. While additional annotation efforts on these genomes likely will increase the number of OBPs to comparable levels of *Solenopsis* and *Apis*, it does appear that the social Hymenoptera in general possess relatively few OBPs. Ongoing and future genome projects in other bees and ants will prove important to address this issue.

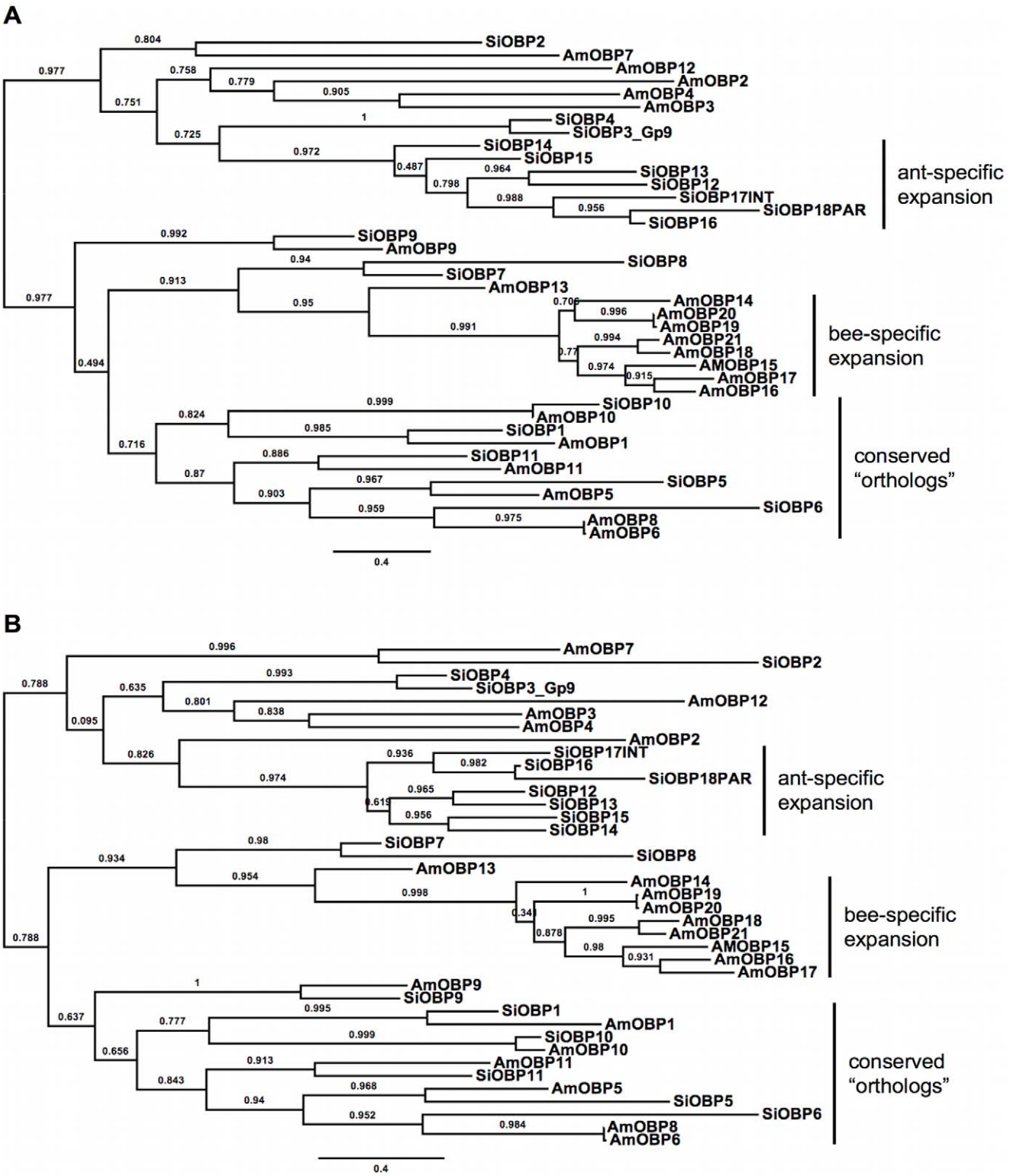

Figure 2. Maximum likelihood phylogenies of the fire ant OBPs (SiOBPs) and honey bee OBPs (AmOBPs). The phylogenies are based on the two best alignments (top: MUSCLE, bottom: PRANK). All trees are midpoint rooted in the absence of a suitable outgroup. Branch support is SH-like aLRT derived from PhyML analyses.
doi:10.1371/journal.pone.0016289.g002

Multiple sequence alignment

The MSAs resulted in alignments of widely different lengths and quality (Table 2). The best alignment method, as measured by the Robinson-Foulds distance to the BAli-Phy topologies, was PRANK followed by MUSCLE. These two methods also produced the "best" fitting trees to the data by any measurement

Table 3. Results of the selection analyses for the best alignment method (PRANK), and two others (CLUSTAL and MUSCLE) and the MAP dataset of *Solenopsis*.

Site model

	M1a	M2a	LRT	M7	M8	LRT
CLUSTAL	−16240.8767	−16219.8	*42.1534****	−16048.4291	−16046.0607	*4.7369****
MUSCLE	−16330.4985	−16330.3	0.4105	−16270.4	−16269.6	1.6646
PRANK	−15838.686	−15838.7	0	−15772.4	−15772.4	0
MAP	−8454.7161	−8446.69	*16.0498****	−8399.03	−8398.04	1.9778

Branch model

	H_o	H_a	LRT
CLUSTAL	−16184.9809	−16183.7734	2.4152
MUSCLE	−16478.8518	−16474.628	*8.4477****
PRANK	−15891.3104	−15886.5809	*9.4588****
MAP	−8453.96731	−8452.55841	2.8178

Branch-site model

	H_o	H_a	LRT
CLUSTAL	−16112.0983	−16111.4054	1.3857
MUSCLE	−16334.5808	−16333.4641	2.2334
PRANK	−15842.8377	−15841.8544	1.9667
MAP	−8424.3691	−8424.3426	0.0529

Given are the log-likelihoods of the null hypotheses (H_0), which assume no selection, and alternative hypotheses (H_a), which allow for positive selection. Likelihood-ratio tests (LRT) of positive selection are conducted to compare the two hypotheses. Levels of significance are 3.84 at 5% and 6.63 at 1% for the site model and 3.84 at 5% and 5.99 at 1% for the branch and branch-site models, following the χ_1^2 distribution to guide against violations of model assumptions. Statistically significant LRTs for positive selection are indicated by italics and *** for $p \ll 0.01$. Note that inference of positive selection greatly depends on the alignment method used.
doi:10.1371/journal.pone.0016289.t003

(LnL, tree length, branch support). The other MSAs (MAFFT, CLUSTAL, OPAL, SATCHMO) fared worse and were never "best" by any measure. MAFFT, however, came in second to PRANK in the estimates of LnL, parsimony, tree length, branch support, and percent sequence identity. The quality of the alignments do not seem to be contingent upon the total lengths or the core lengths, since PRANK is by far the longest alignment and MUSCLE is the second shortest. Additionally, both the Steel [21] and Xia [22] tests indicated high levels of sequence saturation for our dataset for all MSAs (not shown), suggesting that the OBP alignments contained little evolutionary signal.

Also, the AU plots (Figure 1b) suggest that the common removal of signal peptides [17,18,19,20] may not be necessary, since these areas still possess high quality alignment blocks.

Phylogenetic analyses

The phylogenetic relationships of the 18 functional fire ant OBPs (SiOBPs) to the 21 OBPs described from the honey bee, *Apis mellifera* (AmOBPs [4]) are shown in Figure 2. We named the SiOBPs in a numerical series attempting as best possible to use the same numbers for those showing high conservation and presumed orthology with a subset of the honey bee OBPs (Figure 2). These are OBPs 1, 5, 6, 9, 10, and 11 (AmOBP6 and 8 are almost identical in encoded amino acid sequence, but derived from adjacent slightly different genes). Our phylogenetic assessment of orthology in these OBPs is robust across all alignment methods (despite moderate branch support in some cases), suggesting that the assignment is accurate. The phylogenetic analysis indicates that these conserved orthologs may constitute a monophyletic lineage, albeit without high branch support. Even though not all MSAs recovered the same relationship among these orthologs, nor their monophyly, the *Solenopsis* and *Apis* BAli-Phy trees share identical branching patterns for these orthologs, suggesting that the phylogenetic information within these sequences was conserved during cladogenesis. This branching pattern was also recovered by the PRANK alignment method, even though the branch support for the deeper nodes is relatively poor.

Phylogenetic analyses also suggest a close relationship between these same orthologs and a bee-specific clade (AmOBPs 14–21), which is comprised of OBPs encoded by a tandem array that are distinct in having lost a pair of the six usually conserved cysteines (so-called "C-minus" OBPs) and also exhibiting signals of positive selection [4]. Although AmOBP13 is also in this tandem array, this OBP has six cysteines and is not expressed in adult antennae but rather in late larval and early pupal stages [4]. SiOBP7 and SiOBP8 are sister to the C-minus expansion and AmOBP13, but with weak support. SiOBP8 has lost the same pair of cysteines (Table 1), apparently independently of the losses in the honey bee, which in turn are independent of other losses of this pair of cysteines in other C-minus OBPs in other insects [4].

The other half of the tree contains another mixture of AmOBP and SiOBP lineages. SiOBP2 has lost the same pair of cysteines as SiOBP8, and this loss also seems to be independently derived, since it always clusters with AmOBP7 (except in the SATCHMO alignment, not shown) with modest branch support. AmOBPs 2–4, and 12 cluster together with weak support. SiOBP3 is *GP-9*, the OBP implicated in control of social behavior in these ants [10,12], and SiOBP4 apparently is a paralog: These proteins share only 68% amino acid identity, but are co-linear. SiOBP4 is 87% identical to a supposed divergent ortholog of *GP-9* from an unidentified "thief ant" species (GenBank AAW80681 [34]). This suggests that the supposed thief ant *GP-9* is more likely an ortholog

of SiOBP4 and that *GP-9* may be restricted to the fire ants (*geminata* species group [35,36]). While these proteins have no consistent relationship to any of the honey bee OBPs, SiOBP3 and SiOBP4 are the sister group to a seven-gene ant-specific OBP expansion (SiOBP12–18), which itself is close in size and in rate of radiation to the C-minus AmOBP14–21 gene expansion.

It is tempting to speculate that like the OBPs of the bee-specific expansion, these relatively young ant-specific OBPs might well constitute a major fraction of those expressed in the antennae and thus may serve as part of the primary olfactory OBPs in *S. invicta*. However, whether any of these proteins are directly involved in olfaction remains to be demonstrated. Circumstantial evidence suggests that this is unlikely. In fact, the use of OBPs in ant chemosensation has been questioned. Previous studies were unable to identify any members of this protein family in ant antennae [37,38,39], which led Calvello *et al.* [7] to speculate that ants may prefer to use CSPs instead of OBPs for olfaction, which could explain the expansion of CSPs in *S. invicta*. More recently however, three OBPs have been documented [5,6, R. Renthal personal communication] in the antennae of red imported fire ant workers (SiOBP15 [OBP1 of Wang *et al.* [7]], SiOBP3 [*GP-9*], and SiOBP2). None of these proteins appear to be orthologous to any AmOBPs, which have been shown to be expressed in the bee antennae. While the bee OBP data suggest that expression in antennae (and the concomitant presumed use in chemosensation) is phylogenetically preserved, this view may well be biased because half of the AmOBPs tested belong to the rapid bee specific expansion [4].

Selection analyses

The varied and mixed results of the selection analyses suggest that any selection analyses of OBPs be viewed with healthy skepticism. As Wong *et al.* [14] demonstrated, alignment variability is positively and significantly correlated with the number of non-synonymous substitutions, which could explain our positive results for the site- and branch-specific tests of selection and those of Forêt and Maleszka [4]. More recently, Fletcher and Yang [40] showed that alignment errors can lead to a high number of false positives for the branch-site test of positive selection. Even the best performing MSA method (PRANK) did not have the false-positives under control, but nonetheless did fare better than the other alignment methods (MAFFT, MUSCLE, and CLUSTAL) [40]. However, our branch-site tests of selection did not reveal any evidence of positive selection on the branch leading to the ant-specific expansion for any of the alignments used, suggesting that alignment error may not have been an important issue for these analyses. Thus, we are left in the unfortunate position of not being able to conclude confidently the nature of selective forces, if any, shaping the evolution of OBPs in *S. invicta* (and the honey bee), except to say that, like in other insects, lineage-specific expansions are a common feature of Hymenopteran OBP evolution and that their OBPs are widely divergent.

Perhaps more importantly, our data suggest that inferences drawn from analyses of widely divergent molecular sequences are to be regarded with skepticism, since the outcome heavily depends on the resulting alignment chosen. While these issues have been raised previously [e.g., 14,40,41,42,43,44,45], such analyses are becoming increasingly commonplace, especially with the advent of next-generation DNA sequencing platforms and the rapid increase in genomic data, yet, many researchers appear not to consider the estimation of molecular sequence alignment as an exploratory phase of data analysis [46]. Rather, the inference of tree topology is explored much more often, where the judicious choice and use of underlying models, optimality criteria, branch support measures, etc. are a mandatory consideration in virtually all publications and the potentially different outcomes are discussed critically. This apparent lack of attention to MSA methods perhaps stems from an era when the study of molecular sequences was limited to what could be successfully amplified, which likely led to biased analyses of closely related sequences. In any case, we concur with earlier studies that there is an increasing need for awareness for the necessity of careful and critical data exploration during all stages of molecular evolutionary analyses [14,44,46].

Materials and Methods

Identification of loci

Odorant binding proteins and chemoreceptors were identified using BLAST searches [47] of the combined EST and preliminary 454 sequencing data using the fruit fly [48] and honey bee OBPs [4]) as query. The fire ant genes thus identified were then iteratively used as BLAST queries against the same fire ant sequence database until no further new *Solenopsis* loci were found. After we had concluded all our analyses, we also used BLAST searches against the predicted proteins and CDS of the recently released *Camponotus floridanus* and *Harpegnathos saltator* genomes v. 3.3 [33] using the *Apis* and *Solenopsis* OBP amino acid sequences as queries. Given the incomplete annotation of the genomes and the low number of OBPs recovered, we chose not to perform analyses including the other ant OBPs, but instead defer to future researchers that can make use of the several other ant genomes currently being sequenced to address this issue more fully [31].

Multiple sequence alignment

Expecting the generally divergent nature of OBPs sequences (~20% amino acid identity over all sequences) to make the sequence alignment problematic [49], we used several multiple sequence alignment (MSA) methods to evaluate potential different outcomes of using six alignment approaches (Table 2), which differ greatly in popularity and general approach to the MSA problem [50]. We used default parameters for all alignment estimates. Nucleotide (codon) alignments were based on the amino acid alignments.

In addition, we used BAli-Phy 2.0.2 [51] to simultaneously estimate the alignment and phylogeny of the each species' OBPs in a Bayesian framework [52]. Since BAli-Phy is computationally intensive and generally considered to be too slow to be efficiently used with more than a dozen sequences, we conducted these analyses for both the ant and bee datasets independently. Additionally, we removed six bee OBPs from the well-supported C-minus expansion [4] to reduce computational burden. We used default parameters for each run of 100,000 generations. Stationarity of the searches was verified using Tracer 1.5 [53]. 9999 samples were removed in the burn-in. The lowest effective sample size (ESS) for any parameter estimate was 802.3378, suggesting that we had run the analyses sufficiently long to enable meaningful estimates from the posterior sampling.

The alignments were compared using a range of ad hoc heuristic criteria. First, we visually compared alignments for congruence in their ability to align sections of the alignments (especially the inner core) using AltAVisT [54] and the overall sequence identity calculated from each alignment. We then tested for sequence saturation using both the Steel (for amino acids; [21]) and the Xia (for nucleotides [22]) methods [55] using DAMBE [56]. Finally we compared their ability to capture phylogenetic signal relative to the other alignment methods (using ML trees; see below). To this end, we compared log-likelihoods, tree length (measured by parsimony steps of the phylogeny and ML tree size),

and the average of aLRT branch support [57] as well as the Robinson-Foulds tree distance [58] to the ant and bee MAP trees using the TreeDist program in the PHYLIP 3.69 package [59].

Phylogenetic analyses

We used the ProtTest server [60] to estimate the best-fitting model of amino acid substitution for each alignment using the Bayesian information criterion (BIC [61]). Tree topologies were optimized starting from an initial BioNJ tree. Phylogenetic hypotheses under the maximum likelihood criterion were derived from the amino acid alignments using PhyML3 [62]. We implemented the model consistently chosen by the BIC (LG [63]) while estimating the proportion of invariable sites (+I) and gamma shape parameter (+Γ) with 4 rate categories. Tree searches started from five random starting trees and used SPR and NNI to optimize topologies. Branch lengths were optimized and branch support was estimated using the SH-like aLRT [57]. We also employed MrBayes 3.1.2 [64] to compare phylogenetic hypotheses derived from the amino acid and nucleotide datasets. Due to computational burden of the Bayesian analyses, we only performed these on the two best alignments (MUSCLE and PRANK). For each alignment, we performed two searches using different models of sequence evolution. For the amino acid dataset we employed model averaging [65] to incorporate model selection in the Markov Chain Monte Carlo (MCMC) search. For the nucleotide codon alignment we implemented the GTR+I+Γ model. Four chains were run for 5 million generations (one cold and three heated; temperature = 0.02–0.03). Samples from the MCMC were taken every 1000^{th} generation. All other parameters were left at program defaults. Convergence was assessed by measuring average standard deviations of split frequencies, potential scale reduction factor (PSRF) values, plateauing of log-likelihoods values, and ESS values >100.

Selection analyses

We conducted analyses of positive selection using the codeml program in the PAML 4.3 package [66]. Since codeml requires a fully resolved tree, we used the ML trees of the PRANK, MUSCLE, CLUSTAL, and BAli-Phy alignments as input. These represent the two "best", the longest and shortest alignments. We estimated branch lengths under the F3×4 codon model on the respective topologies. We conducted site-specific tests of selection [23,24,25,26]. We were also specifically interested in whether positive selection had influenced the divergence of the ant-specific expansion. Hence, we performed branch-specific tests of selection [27,28] on the branch leading to this clade. However, under certain circumstances the branch-specific test of selection can lack power and so we also used the branch-site test of selection [26,30] implementing the Bayes empirical Bayes (BEB [26]) method to identify sites under selection. To ensure that the analyses had converged properly, we repeated each analysis three times from different starting parameter options and under different codon models.

Acknowledgments

Kim Walden provided assistance with the analyses. Robert Renthal provided helpful comments and unpublished data. Three anonymous reviewers provided helpful comments which greatly improved the manuscript.

Author Contributions

Conceived and designed the experiments: DG HMR DS. Performed the experiments: DG HMR YW DS. Analyzed the data: DG HMR. Contributed reagents/materials/analysis tools: YW DS. Wrote the paper: DG HMR.

References

1. Vogt RG (2003) Biochemical diversity of odor detection. In: Blomquist G, Vogt R, eds. Insect Pheromone Biochemistry and Molecular Biology. London: Elsevier. pp 391–445.
2. Pelosi P, Zhou JJ, Ban LP, Calvello M (2006) Soluble proteins in insect chemical communication. Cell Mol Life Sci 63: 1658–1676.
3. Sánchez-Gracia A, Vieira FG, Rozas J (2009) Molecular evolution of the major chemosensory gene families in insects. Heredity 103: 208–216.
4. Forêt S, Maleszka R (2006) Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (*Apis mellifera*). Genome Res 16: 1404–1413.
5. Leal WS, Ishida Y (2008) GP-9s are ubiquitous proteins unlikely involved in olfactory mediation of social organization in the red imported fire ant, *Solenopsis invicta*. PLoS One 3: e3762.
6. González D, Zhao Q, McMahan C, Velasquez D, Haskins WE, et al. (2009) The major antennal chemosensory protein of red imported fire ant workers. Insect Mol Biol 18: 395–404.
7. Calvello M, Brandazza A, Navarrini A, Dani F, Turillazzi S, et al. (2005) Expression of odorant-binding proteins and chemosensory proteins in some Hymenoptera. Insect Biochem Mol Biol 35: 297–307.
8. Wang J, Jemielity S, Uva P, Wurm Y, Gräff J, et al. (2007) An annotated cDNA library and microarray for large-scale gene-expression studies in the ant *Solenopsis invicta*. Genome Biol 8: R9.
9. Wurm Y, Uva P, Ricci F, Wang J, Jemielity S, et al. (2009) Fourmidable: a database for ant genomics. BMC Genomics 10: 5.
10. Ross KG (1997) Multilocus evolution in fire ants: effects of selection, gene flow and recombination. Genetics 145: 961–974.
11. Krieger MJ, Ross KG (2002) Identification of a major gene regulating complex social behavior. Science 295: 328–32.
12. Gotzek D, Ross KG (2007) Genetic regulation of colony social organization in fire ants: an integrative overview. Q Rev Biol 82: 201–226.
13. Gotzek D, Ross KG (2009) Current status of a model system: the gene *Gp-9* and its association with social organization in fire ants. PLoS One 4: e7713.
14. Wong KM, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. Science 319: 473–476.
15. Felsenstein J (2004) Inferring Phylogenies. Sunderland: Sinauer Associates. 664 p.
16. Boussau B, Daubin V (2009) Genomes as documents of evolutionary history. Trends Ecol Evol 25: 224–232.
17. Zhou JJ, He X-L, Pickett JA, Field LM (2008) Identification of odorant-binding proteins of the yellow mosquito *Aedes aegypti*: genome annotation and comparative analyses. Insect Mol Biol 17: 147–163.
18. Gong D-P, Zhang H-J, Zhao P, Xia Q-Y, Xiang Z-H (2009) The odorant binding protein gene family from the genome of silkworm, *Bombyx mori*. BMC Genomics 10: 332.
19. Vieira FG, Sánchez-Gracia A, Rozas J (2009) Comparative genomic analyses of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. Genome Biol 8: R235.
20. Zhou JJ, Vieira FG, He X-L, Smadja C, Liu R, et al. (2010) Genome annotation and comparative analyses of the odorant-binding proteins and chemosensory proteins in the pea aphid, *Acyrthosiphon pisum*. Insect Mol Biol 19(Suppl. 2): 113–122.
21. Steel M, Lockhart PJ, Penny D (1993) Confidence in evolutionary trees from biological sequence data. Nature 364: 440–442.
22. Xia XH, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Mol Phyl Evol 26: 1–7.
23. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–936.
24. Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000) Codon-substitution models for heterogenous selection pressure at amino acid sites. Genetics 155: 431–449.
25. Wong WSW, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168: 1041–1051.
26. Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107–1118.
27. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573.
28. Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46: 409–418.
29. Yang Z (2006) Computational Molecular Evolution. Oxford: Oxford University Press. 376 p.
30. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472–2479.

31. Smith CD, Smith CR, Mueller U, Gadau J (2010) Ant genomics: strength in numbers. Mol Ecol 19: 31–35.
32. Kirkness EF (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Nat'l Acad Sci USA 107: 12168–12173.
33. Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, et al. (2010) Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Science 329: 1068–1071.
34. Krieger MJ, Ross KG (2005) Molecular evolutionary analyses of the odorant-binding protein gene *Gp-9* in fire ants and other *Solenopsis* species. Mol Biol Evol 22: 2090–2103.
35. Pitts JP, McHugh JV, Ross KG (2005) A cladistic analysis of the fire ants of the *Solenopsis saevissima* species-group (Hymenoptera: Formicidae). Zool Scripta 34: 493–505.
36. Trager J (1991) The fire ants of the *Solenopsis geminata* group. J New York Entomol Soc 99: 141–198.
37. Ishida Y, Chiang V, Leal WS (2002) Protein that makes sense in the Argentine ant. Naturwissenschaften 89: 505–7.
38. Guntur KV, Velasquez D, Chadwell L, Carroll C, Weintraub S, et al. (2004) Apolipophorin-III-like protein expressed in the antenna of the red imported fire ant, *Solenopsis invicta* Buren (Hymenoptera: Formicidae). Arch Insect Biochem Physiol 57: 101–110.
39. Ozaki M, Wada-Katsumata A, Fujikawa K, Iwasaki M, Yokohari F, et al. (2005) Ant nestmate and non-nestmate discrimination by a chemosensory sensillum. Science 309: 311–314.
40. Fletcher W, Yang Z (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol 27: 2257–2267.
41. Morrison DA, Ellis JT (1997) Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa. Mol Biol Evol 14: 428–441.
42. Rost B (1999) Twilight zone of protein sequence alignment. Prot Engineering 12: 85–94.
43. Ogden TH, Rosenberg MS (2006) Multiple sequence alignment accuracy and phylogenetic inference. Syst Biol 55: 314–328.
44. Martin W, Roettger M, Lockhart PJ (2007) A reality check for alignments and trees. Trends Genet (23: 478–480.
45. Opperdoes FR (2009) Phylogenetic analysis using protein sequences. In: Lemey P, Salemi M, Vandamme A-M, eds. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2nd edition. Cambridge: Cambridge University Press. pp 313–331.
46. Morrison DA (2009) Why would phylogeneticists ignore computerized sequence alignment? Syst Biol 58: 150–158.
47. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
48. Hekmat-Scafe DS, Scafe CR, McKinney AJ, Tanouye MA (2002) Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. Genome Res 12: 1357–1369.
49. Pelosi P, Calvello M, Ban L (2005) Diversity of odorant-binding proteins and chemosensory proteins in insects. Chem Senses 30: i291–i292.
50. Higgins D, Lemey P (2009) Multiple sequence alignment. In: Lemey P, Salemi M, Vandamme A-M, eds. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2nd edition. Cambridge: Cambridge University Press. pp 68–108.
51. Suchard MA, Redelings BD (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22: 2047–2048.
52. Redelings BD, Suchard MA (2006) Joint Bayesian estimation of alignment and phylogeny. Syst Biol 54: 401–418.
53. Rambaut A, Drummond AJ (2007) Tracer v1.5. Available: http://beast.bio.ed.ac.uk/Tracer.
54. Morgenstern B, Goel S, Sczyrba A, Dress A (2003) AltAVisT: A WWW tool for comparison of alternative multiple alignments. Bioinformatics 19: 425–426.
55. Xia X (2009) Assessing substitution saturation with DAMBE. In: Lemey P, Salemi M, Vandamme A-M, eds. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2nd edition. Cambridge: Cambridge University Press. pp 615–630.
56. Xia XH, Xie Z (2001) DAMBE: Software package for data analysis in molecular biology and evolution. J Hered 92: 371–373.
57. Anisimova M, Gascuel O (2006) Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. Syst Biol 55: 539–552.
58. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53: 131–147.
59. Felsenstein J (2009) PHYLIP (Phylogeny Inference Package) version 3.69. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
60. Abascal F, Zardoya R, Posada D (2005) ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21: 2104–2105.
61. Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 53: 793–808.
62. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
63. Le SQ, Gascuel O (2008) An Improved General Amino-Acid Replacement Matrix. Mol Biol Evol 25: 1307–20.
64. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
65. Huelsenbeck JP, Joyce P, Lakner C, Ronquist F (2008) Bayesian analysis of amino acid substitution models. Phil Trans R Soc B 363: 3941–3953.
66. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.
67. Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320: 1632–1635.
68. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–97.
69. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics 9: 286–298.
70. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) ClustalW and ClustalX version 2.0. Bioinformatics 23: 2947–2948.
71. Wheeler TJ, Kececioglu JD (2007) Multiple alignments by aligning alignments. Bioinformatics 23: i559–i568.
72. Edgar RC, Sjolander K (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics 19: 1404–1411.