Duplication and concerted evolution in a master sex determiner under balancing selection

Eyal Privman1,2, Yannick Wurm1,2,†,‡ and Laurent Keller1,†

1Department of Ecology and Evolution, University of Lausanne, Lausanne 1015, Switzerland
2Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland

The transformer (tra) gene is a key regulator in the signalling hierarchy controlling all aspects of somatic sexual differentiation in Drosophila and other insects. Here, we show that six of the seven sequenced ants have two copies of tra. Surprisingly, the two paralogues are always more similar within species than among species. Comparative sequence analyses indicate that this pattern is owing to the ongoing concerted evolution after an ancestral duplication rather than independent duplications in each of the six species. In particular, there was strong support for inter-locus recombination between the paralogues of the ant Atta cephalotes. In the five species where the location of paralogues is known, they are adjacent to each other in four cases and separated by only few genes in the fifth case. Because there have been extensive genomic rearrangements in these lineages, this suggests selection acting to conserve their synteny. In three species, we also find a signature of positive selection in one of the paralogues. In three bee species where information is available, the tra gene is also duplicated, the copies are adjacent and in at least one species there was recombination between paralogues. These results suggest that concerted evolution plays an adaptive role in the evolution of this gene family.

1. Introduction

In insects, tra has been found to be a conserved upstream component inducing female development by regulating sex-specific alternative splicing of downstream genes such as doublesex [1,2]. However, the primary signal activating tra varies among species. In some species, tra activation is controlled by a sex chromosome system, whereas in others (especially the 200,000 species of the order Hymenoptera that include the ants, bees, wasps and sawflies) there is a haplodiploid sex determination system where diploids develop into females and haploids into males [3]. Many hymenopterans use the complementary sex determination (CSD) mechanism, of which the honeybee Apis mellifera is the prototypical example. Heterozygosity at a single locus (the CSD locus) triggers female development in diploid individuals, while haploid individuals have a hemizygous CSD genotype and thus develop into males [4]. Likewise, homozygous diploid individuals have one allele and thus develop into diploid males, but these are generally unviable or sterile [5]. Selection against diploid males should thus favour rare alleles at the CSD locus by balancing selection [6].

The CSD locus was molecularly identified in the honeybee Ap. mellifera and found to be a homologue of tra [7]. In this species, the tra locus is duplicated [8]. One copy, named csd, is the primary signal of the sex determination pathway. It activates the second copy, named feminizer (fem), which is more conserved and retains the ancestral function of regulating doublesex and the downstream factors in the signal transduction cascade. Following the duplication csd was subject to positive selection, consistent with neofunctionalization of this paralogue. Two paralogues of tra were also found in the two other species of the genus Apis (Apis cerana and Apis dorsata). A gene tree, including these six sequences and a single homologue in the sister genus Bombus supported the view of a duplication in the Apis lineage, after the split from Bombus [8]. The finding of many highly diverged csd alleles in
the three honeybee species is consistent with balancing selection acting on this locus to reduce production of diploid males [8].

The recent sequencing of seven ant genomes [9–14] provides an opportunity to investigate the gene duplication, sequence evolution, and selective forces acting on these genes in Hymenoptera. We will use the names csd and fem for the two paralogues in honeybees, but we will name them traA and traB in ants because their functional roles are still unknown. We chose these names rather than tra1 and tra2 because the name tra2 is already in use for a more distant parologue that exists in many insects, including flies [15] and ants (data not shown).

2. Material and methods

(a) Identification of tra homologues

Homologues were identified in the seven ant genome assemblies [9–14] using translated BLAST [16] with the protein sequence of Acromyrmex echinatior tra [10] as the query. For Solenopsis invicta, we used an improved assembly of the genome (O. Riba-Grognuz 2012, unpublished data). Gene models were constructed using MAKER [17] given as evidence homologous tra protein sequences and assembled transcriptome sequences for each species. MAKER gene models were manually corrected using APOLLO [18]. Coding sequences and predicted protein sequences are included in the electronic supplementary material.

(b) Previously published tra homologues

Genbank accession numbers for sequences used in phylogenetic and reconstruction analyses (amino acid, nucleotide sequences): Ap. mellifera fem (AA86667, NM_001134828) and csd (AA86665, NM_001101569); Ap. dorsata fem (ABV56232, EU100939) and csd (ABV56216, EU100933); Ap. cerana fem (ABV56230, EU100936); and csd (ABV56877, EU100908); Bonnus terrestris fem (XP_003402358, XM_003402310); Nasua vespivrae fem (NP_011128299).

Apis csd alleles for recombination tests: Ap. mellifera: NM_001101569, AY350615, AY350617, AY350618, AY350616, EU100898, EU100896, EU100894, EU100892, EU100890, EU100888, EU100886, EU100899, EU100897, EU100895, EU100893; AY350697, EU100891, EU100889, EU100887, EU100885, EU100883, EU100881, EU100879, EU100877, EU100875, EU100873, EU100871, EU100869, EU100867, EU100865, EU100863, EU100861, EU100859, EU100857, EU100855, EU100853, EU100851, EU100849, EU100847, EU100845, EU100843, EU100841, EU100839, EU100837, EU100835, EU100833, EU100831, EU100829, EU100827, EU100825, EU100823, EU100821, EU100819, EU100817, EU100815, EU100813, EU100811, EU100809, EU100807, EU100805, EU100803, EU100801; Ap. cerana: EU100916, EU100914, EU100912, EU100910, EU100908, EU100906, EU100904, EU100902, EU100900, EU100915, EU100913, EU100911, EU100909, EU100907, EU100905, EU100903, EU100901; Ap. dorsata: EU100935, EU100933, EU100931, EU100929, EU100927, EU100925, EU100923, EU100921, EU100919, EU100917, EU100915, EU100913, EU100911, EU100909, EU100907, EU100905, EU100903, EU100901; Ap. cerana: EU100926, EU100924, EU100922, EU100920, EU100918.

(c) Alignment and phylogeny reconstruction

Protein sequences were aligned and the gene tree reconstructed using the simultaneous alignment and phylogeny Bayesian reconstruction algorithm implemented in BALI-PHY [19], with the LG + gamma substitution model, indel model RS07. Eight BALI-PHY chains were run in parallel until convergence. The posterior decoding alignment was used for all subsequent analyses. Unreliably aligned residues (approximately unbiased score less than 80%) were masked by replacing them with missing data characters (‘X’). The alignment was mapped back to the coding sequences to produce the corresponding codon sequence alignment (corresponding codons masked as ‘NNN’). The majority consensus tree was constructed (including splits with posterior probability greater than 50%). The maximum a posteriori tree was used for downstream analyses that require a fully bifurcating tree (recombination and positive selection tests).

(d) Positive selection and rate shift tests

The branch-site test for positive selection [20] implemented in PAML [21] was run on the masked BALI-PHY codon alignment to test each branch of the BALI-PHY maximum a posteriori tree for a dN/dS ratio (the omega parameter) greater than one. The null model (fixed omega of one) and the alternative model (free omega greater than one) were optimized for each branch and used to calculate the likelihood ratio test (LRT). The LRT confidence scores were adjusted to control the false discovery rate (FDR) using the Benjamini–Hochberg method [22]. A posterior probability of greater than 95% for a dN/dS ratio (the omega parameter) greater than 1 was considered significant evidence for positive selection in specific codon sites.

To test for acceleration of the evolutionary rate following duplication, we used PAML’s clade model C (six free parameters), which we ran once for each branch as the foreground branch. The C model allows the foreground branch to have a different rate than the rest of the tree for the third rate category [23]. We used M3 as the null model (five free parameters), because it allows three discrete rate categories similar to the C model, but no rate variation across the tree [24]. For each terminal branch, we calculated the LRT comparing the two models. We used the χ²-squared distribution with 1 d.f. to check for statistical significance.

(e) Recombination tests

Phylogenetic ‘splits networks’ were reconstructed based on evolutionary distances using the NEIGHBOURNet algorithm [25] implemented in the SPLITSTree package [26]. Networks were visualized using the ‘equal angle’ algorithm [27]. The following tests were performed on the masked DNA alignments using the RDP4 package [28]: RDP [29] with internal and external reference, MAXCHI [30], GENECONV [31], BOOTSCAN [32] with neighbour joining trees, CHIMAERA [33], SISCAN [34], 3SEQ [35], LARD [36] and PHYLPRO [37]. We used RDP4 to look for statistical support for specific recombination events. We applied different methods to datasets ranging in their sequence divergence: protein sequences from the full dataset, including the wasp and bee outgroups (18 sequences), coding DNA sequences from all seven ant species (13 sequences), from the five more closely related ant species (nine sequences), the four more closely related species (seven sequences) and the two most closely related species (Ac. echinatior and Atta cephalotes), together with S. invicta tra as an outgroup (four sequences).

We similarly analysed the honeybee tra homologues (fem and csd) from Ap.mellifera, Ap.cerana and Ap. dorsata. There are 43, 17 and 19 published csd allele sequences in Ap. mellifera, Ap. cerana and Ap. dorsata, respectively. To test for recombination between the two paralogues, we analysed sequence alignments that include the fem sequence and one csd allele sequence from each Apis species, and the B. terrestris fem as the outgroup (seven sequences). The sequences were codon-aligned and unreliably aligned codons were masked as ‘NNN’ using PRANK [38] and GUIDANCE [39,40]. To test different csd alleles, we repeated these analyses 100 times, every time with a different random choice of one csd allele per species. We applied the aforementioned recombination tests to each of the 100 sequence alignments.

We used the ‘alignment uncertainty (AU) test of phylogenetic tree selection’ [41] to further confirm inter-locus recombination events that were inferred by the above tests in the ant and bee sequence alignments. Alignments were split into putative recombinant and non-recombinant regions according to the recombination points inferred by RDP4. Maximum-likelihood (ML) trees were constructed for each sub-alignment using PrinML v. 3.0.1beta.
Table 1. Duplicated tra homologues found in the sequenced ant genomes.

<table>
<thead>
<tr>
<th>species</th>
<th>tra homologues</th>
<th>genomic organization</th>
<th>protein product</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. saltator</td>
<td>traA traB</td>
<td>at ends of disjoint scaffolds in the genome</td>
<td></td>
</tr>
<tr>
<td>L. humile</td>
<td>traA traB</td>
<td>seven genes between traA and traB</td>
<td>C-terminal region of traA and traB is truncated</td>
</tr>
<tr>
<td>C. floridanus</td>
<td>traA traB</td>
<td>adjacent</td>
<td></td>
</tr>
<tr>
<td>P. barbatus</td>
<td>traA traB</td>
<td>adjacent</td>
<td>traB appears to be missing some internal fragments</td>
</tr>
<tr>
<td>S. invicta</td>
<td>traA traB</td>
<td>adjacent</td>
<td>N-terminal region of traB is truncated</td>
</tr>
<tr>
<td>At. cephalotes</td>
<td>traA traB</td>
<td>adjacent</td>
<td></td>
</tr>
<tr>
<td>Ac. echinatior</td>
<td>traA traB</td>
<td>adjacent</td>
<td></td>
</tr>
</tbody>
</table>

Comparative sequence analyses provided evidence for recombination between tra paralogues, in support of the hypothesis of concerted evolution. Phylogenetic network reconstruction revealed many alternative tree topologies, consistent with the presence of recombinant fragments in the paralogues (figure 2a). A more focused analysis of the leafcutter ant clade that contains the most closely related ant sequences of At. cephalotes and Ac. echinatior (and therefore those with the clearest phylogenetic signal) showed high bootstrap support for two different gene trees (figure 2b). This indicates a recombination event between the paralogues of At. cephalotes (there is only one homologue in Ac. echinatior). The use of the nine statistical methods for the detection of specific recombination events (implemented in the RDP4 package [28]) also provided strong support for a recombination event (by seven of nine methods) in At. cephalotes in a 497 bp segment of the 3’ region of the gene (table 2 and figure 3a). Importantly, the occurrence of interlocus recombination is strongly supported by three methods having higher detection power compared with the other methods—RDP, MAXCHI and CHIMAERA [33,45]. We further validated this inference using a likelihood-based phylogenetic test (the AU test). The gene tree topology inferred from the putative recombinant region was incompatible with the non-recombinant region of the alignment (p = 0.035) while the topology from the non-recombinant region was not significantly incompatible with the recombinant region (p = 0.17). A statistically significant signature of specific recombination events could not be detected in the other species either because of a genuine lack of recombination or because the divergence time and evolutionary distances between these pairs of paralogues and the nearest orthologues in other species was too large to allow detection of recombination.

We applied the same tests to the parologue pairs (csd and fem) in the threeApis honeybee species. In contrast to ants, a phylogenetic network reconstructed for the bees paralogues gave low bootstrap support for alternative topology (see the electronic supplementary material, figure S1), implying an ancestral duplication before speciation of the three Apis species. Nevertheless, evidence of recombination and gene conversion was found by several test statistics. We used the suite of nine tests to search for gene conversion events.
in 79 alleles of the csd loci in the three *Apis* species. We applied the tests to 100 sequence sets that each included a different random choice of one csd allele for each species. In 87 of these samples, at least one of the test statistics was significant (figure 4a). In particular, RDP, *March, Chimæra* and *Bootscan* found recombination in 39, 80, 67 and 48 samples, respectively. In total, 121 gene conversion events were inferred (77 in *Ap. dorsata*, 25 in *Ap. mellifera* and 19 in *Ap. cerana*; figure 4b, electronic supplementary material, S2 and S3), providing strong evidence that gene conversion occurred between the fem and csd genes in the three honeybee lineages.

We validated these inferred recombinant csd alleles with the phylogenetic AU test. We applied the AU test to the 48 samples that received a significant score from the *Bootscan* method, because this indicates a phylogenetic difference between the putative recombinant and the non-recombinant regions of the alignment. These included 63 *Bootscan*-supported putative gene conversion events. The gene tree topology inferred from the recombinant region was significantly incompatible with the non-recombinant region of the alignment in 17 of the samples while the topology from the non-recombinant region was significantly incompatible with the recombinant region in two of the samples (FDR adjusted \( p < 0.05 \); see the electronic supplementary material, table S2). Thus, a robust phylogenetic signal in the recombinant regions of the high-scoring alleles supports the results of the suite of test statistics.

Further support for the inference of gene conversion is found in the form of fem-specific substitutions that were copied to the csd allele. We searched for such substitutions in the top scoring alignment, which contains the putatively recombinant *Ap. mellifera* csd allele AY352276. A recombination event between *Ap. mellifera* fem and this csd allele was inferred (by seven of nine methods) in a 207 bp segment at the 3' end of the gene (table 2). Figure 3b shows the RDP results for this alignment, plotting the pairwise similarity between the *Ap. mellifera* csd allele, the *Ap. mellifera* fem and an *Ap. cerana* csd allele. In the inferred recombinant region the *Ap. mellifera* csd allele shows a higher similarity to the paralogous *Ap. mellifera* fem than to the orthologous *Ap. cerana* csd. We found three substitutions in this region where this *Ap. mellifera* csd allele shares the same derived nucleotide occurring in the *Ap. mellifera* fem sequence but not in any of the other csd alleles. No such substitutions are found in the non-recombinant region of the alignment (Fisher's exact test \( p = 0.01 \)). This suggests a gene conversion event copying this region from fem to csd in the *Ap. mellifera* lineage.

In spite of this evidence for gene conversion, the *Apis* gene tree suggests duplication prior to speciation of the three *Apis* species (figure 1). This pattern contrasts with the ant gene tree that implies multiple duplications after the ants' speciation. This difference can be explained by the much greater divergence time between the ants than the bees used in this study. The ant lineages began to speciate more than 115 million years ago [46] while the honeybees speciated less than 40 million years ago [47]. As a result, gene conversion events in honeybees may be limited to a smaller subset of the gene sequence, and thus not yet result in sufficient homogenization among paralogues to make them more similar than orthologues. The same explanation may account

---

**Figure 1.** Gene tree of *tra* homologues in Hymenoptera. Stars indicate putative independent gene duplication events. Phylogeny was reconstructed using BAli-Phy from predicted protein sequences. Posterior probability support is indicated for each branch. Forward green arrows indicate significant acceleration of the evolutionary rate, and plus signs indicate significant evidence for positive diversifying selection. Backward blue arrow indicates significantly reduced rate (detailed results presented in tables 3 and 4). Species name abbreviations: *nvit*, *Nasonia vitripennis*; *amel*, *Apis mellifera*; *acer*, *Apis cerana*; *ador*, *Apis dorsata*; *hsal*, *Harpegnathos saltator*; *lhum*, *Linepithema humile*; *cflo*, *Camponotus floridanus*; *pbar*, *Pogonomyrmex barbatus*; *sinv*, *Solenopsis invicta*; *aceh*, *Acromyrmex echinatior*; *aced*, *Atta cephalotes*. Gene name abbreviations: *tra*, transformer; *fem*, feminizer; *csd*, complementary sex determiner.
for the abovementioned difference in the phylogenetic network results for ants and bees.

(b) Adaptive evolution of \textit{tra} paralogues

Two lines of evidence indicate that natural selection acts in a similar manner on \textit{tra} homologues in ants as in bees. First, \textit{tra} paralogues are adjacent loci in the genomes of the three \textit{Apis} species [7], as well as in four out of five ant species that have two paralogues with known relative position (in the fifth species the paralogues are separated by seven genes; in the sixth they were found on different scaffolds in the genome assembly so their relative position is unknown; table 1). In ants, the close location is unlikely to be owing to chance alone because extensive genomic rearrangements occurred in these lineages: a comparative analysis of the genomes of the four species where the two \textit{tra} paralogues are adjacent reveals that only 1.07 per cent of neighbouring pairs of genes in any of these four species remain adjacent in all four genomes (see the electronic supplementary material, table S1). This indicates selection for maintaining synteny of the paralogues. Such physical proximity between \textit{tra} paralogues probably facilitates concerted evolution [48]. Second, there is evidence for adaptive sequence evolution in \textit{tra} sequences in ants as was previously described in bees [8]. Branch-site tests provided evidence for positive selection (dN/dS > 1) in at least three codon sites in \textit{Linepithema humile \textit{traB}} (table 3) and accelerated evolutionary rate in at least 24, 13 and 15 codon sites in the \textit{traB} sequences of \textit{L. humile}, \textit{C. floridanus} and \textit{At. cephalotes}, respectively (table 4). Thus, selection appears to be a significant driver of

Figure 2. Evidence for recombination between \textit{tra} paralogues in ants from phylogenetic splits networks constructed using \textit{SplitsTree} for coding sequences from (a) the seven ant species and for (b) a narrowed dataset of the attine ants (\textit{Ac. echinatior} and \textit{At. cephalotes}) with \textit{S. invicta \textit{traA}} as an outgroup. Parallelogram branches represent alternative tree topologies corresponding to putative recombination events. Bootstrap support for two alternative topologies is indicated on the parallelogram. Species name abbreviations as in figure 1.
sequence divergence between the two paralogues in both ants and bees.

4. Discussion

The comparison of ant genomes revealed the presence of two *tra* copies in six of the seven species, with the two paralogues being invariably more similar within- than between-species. This striking pattern is inferred to result from the action of concerted evolution over at least 115 million years of evolution. While the alternative hypothesis of multiple independent duplications cannot be completely ruled out, three lines of evidence support the view that gene conversion between the two paralogues has occurred in both ants and bees. First, gene conversion events were found in the ant and bee sequences with support from multiple different recombination test statistics. Second, likelihood-based phylogenetic tests confirmed that recombinant sequences support a different tree topology than the non-recombinant sequences. Third, we identified *fem*-specific substitutions that are also found in a *csd* allele of *Ap. mellifera*. In the light of these results, we interpret the ant gene tree as an ancestral duplication at least 115 million years ago followed by concerted evolution between the two paralogues. The detection of relatively recent gene conversion events suggests that concerted evolution is acting over long evolutionary time in this gene family.

Two lines of evidence suggest that selection pressures are responsible for the evolutionary patterns that we observe. First, the longstanding action of concerted evolution and the conservation of the paralogues in near proximity to one another suggest selection to maintain the ability of paralogues to recombine. Second, we show that one of two paralogues experienced accelerated evolution and/or positive selection in several ant and bee lineages. Hence, we hypothesize that gene conversion between paralogues serves as a mechanism to generate novel recombinant alleles of the fast-evolving paraglogue. Novel alleles would be selected for under balancing selection, which is known to act on the *csd* loci of honeybees.

Selection for concerted evolution at a locus under balancing selection may seem counterintuitive because concerted evolution is better known for homogenizing gene sequences, as in ribosomal RNA genes [44]. However, a diversifying effect of recombination has been previously described in antigens of pathogenic bacteria and protozoa (e.g. [49,50], reviewed in [51]), in vertebrate major histocompatibility complex immune genes [52–54], and in metazoan *prdm9* genes [55], all of which are under balancing selection [55–58]. Inter-allelic recombination and gene conversion was first proposed to contribute to the extremely high polymorphism of vertebrate MHC genes (reviewed in [53,54]) and was subsequently shown to generate novel alleles more rapidly than point mutations in birds [52].

Schmieder et al. [59] recently conducted similar analyses to those reported here. An important difference in our studies is that we used multiple methodologies, which are more conservative and robust to artefacts. The use of more conservative methodologies is important because the highly variable and complex patterns of sequence evolution in this gene family could conceivably lead to false inference of recombination. *tra* paralogues in different lineages were found to evolve with considerably different evolutionary rates, including episodes of positive selection. At least in the three *Apis* species where multiple allele sequences are available, one of the paralogues (*csd*) was shown to evolve under strong balancing selection, which generates a pool of highly diverged alleles. Moreover, considerable portions of the protein are repetitive sequences, which are notoriously difficult to align, which may lead to artefacts in molecular evolutionary analyses such as positive selection inference [40,60].

Because of the problems mentioned above, we took special care in our choice of methodologies, aiming for robustness to potential artefacts. There were three important differences between our analyses and those done by Schmieder et al. [59]. First, we used a superior Bayesian method for joint alignment and phylogeny reconstruction (BALI-PHY) to minimize artefacts owing to alignment errors. We also masked unreliable parts of the alignments to prevent alignment errors from affecting the inference of recombination and positive selection [40]. This difference may be responsible to the lesser extent of positive selection that we find. Only one ant gene (*L. humile* *traB*) had a significant signal in our analyses, compared with five in the results of Schmieder and colleagues. In two other lineages (*traB* of *C. floridanus* and *At. cephalotes*), we found acceleration of the evolutionary rate, but not necessarily positive selection.

Second, while Schmieder and colleagues used *GeneConv* to test for recombination, we use eight additional test statistics, including RDP, *MaxChi* and *Chimaera* that have been shown to have superior detection power compared with other methods [33,45]. Furthermore, our analyses include phylogenetic methods (RDP and *BootsScan*), which detect different phylogenetic relationships among the sequences in the inferred recombinant and non-recombinant regions of the alignment. Such methods provide important robustness to the variation in selection pressure that we observe among the paralogues. RDP and other phylogenetic methods

<table>
<thead>
<tr>
<th>recombinant paralogues</th>
<th>RDP</th>
<th><em>GeneConv</em></th>
<th><em>BootsScan</em></th>
<th><em>MaxChi</em></th>
<th><em>Chimaera</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ap. mellifera</em> csd</td>
<td>$1.98 \times 10^{-3}$</td>
<td>$1.59 \times 10^{-2}$</td>
<td>$2.35 \times 10^{-1}$</td>
<td>$4.13 \times 10^{-4}$</td>
<td>$4.92 \times 10^{-2}$</td>
</tr>
<tr>
<td><em>At. cephalotes</em> traA</td>
<td>$1.86 \times 10^{-9}$</td>
<td>n.s.</td>
<td>n.s.</td>
<td>$2.46 \times 10^{-2}$</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Table 2. Evidence for recombination between *tra* paralogues in honeybees and in ants. (Analyses were based on coding sequence alignments. The *Ap. mellifera* recombination event involves the *fem* sequence AAS86667 and the *csd* sequence AY352276.)
**Figure 3.** Pairwise identity plots calculated by RDP for the *tra* coding sequences from (a) the two leaf cutter ant species *At. cephalotes* and *Ac. echinatior* and from (b) the two honeybee species *Ap. mellifera* (csd allele AY352276, fem AAS86667) and *Ap. dorsata* (csd allele EU100926). Shaded area indicates the putative recombinant region.

**Figure 4.** Recombination test statistics applied to *fem* sequences and samples of *csd* alleles of *Ap. mellifera*, *dorsata* and *cerana*. (a) The number of samples that received a positive result for each test statistic. (b) The number of instances of inferred recombination events in these samples classified by test and species.
were shown to be more robust than non-phylogenetic methods in simulation studies involving variation of selection, including positive selection, as well as asymmetric tree topology and sequence divergence [61]. We did not find the gene conversion events reported by Schmieder and colleagues in the longer branches of the ant phylogeny (Harpegnathos saltator, C. floridanus and P. barbatus), perhaps as a result of the use of these more conservative methodologies. However, our analyses allow one to confidently infer gene conversion in the shorter branches (At. cephalotes).

Third, we validated the inferred recombination events by a superior, likelihood-based phylogenetic test for the significance of the topological difference between the recombinant and the non-recombinant regions. These methodologies are more robust to potential artefacts, thus providing greater confidence in the inference of concerted evolution.

Other differences relative to the study by Schmieder and colleagues include our testing of the entire pool of Apis csd alleles, which allowed robust detection of multiple gene conversion events in all three Apis species, and the identification of an Ap. mellifera csd allele with a clear pattern of gene conversion supported by unique derived substitutions. The inclusion of the second S. invicta paralogue and the two L. humile paralogues provided further evidence for positive selection and better resolution of the phylogeny.

Gene conversion between the paralogues could have interesting implications in terms of their molecular function. Molecular studies in honeybees demonstrated that zygosity of csd regulates alternative splicing of fem mRNA [8], and it has been proposed that the mechanism by which heterozygosity of the csd gene affects fem involves the formation of a heterodimer of csd protein products [62]. Gene conversion events such as the one reported

| Table 3. Tests for positive selection on tra homologues in ants using the PAML branch-site model A. (Species abbreviations are identical to those used in figure 1. Statistically significant results are indicated in bold.) |
|-------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| branch | In likelihood for H0 | In likelihood for H1 | p-value | significant after FDR correction | number of sites with P(positive selection) >95% | rate in branch (ω) |
|-------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| hsal traA | -7782.67 | -7787.09 | 0.00 | yes | 0 | 139.80 |
| hsal traB | -7785.79 | -7785.84 | 0.74 | no | 0 | 1.53 |
| 1hum traA | -7784.31 | -7786.87 | 0.02 | yes | 0 | 57.31 |
| 1hum traB | -7775.84 | -7780.05 | 0.00 | yes | 3 | 7.36 |
| cflo traA | -7787.10 | -7787.10 | 1.00 | no | 0 | 1.00 |
| cflo traB | -7775.85 | -7777.25 | 0.09 | no | 2 | 2.54 |
| pbar traA | -7786.37 | -7786.37 | 1.00 | no | 0 | 1.00 |
| pbar traB | -7785.87 | -7785.87 | 1.00 | no | 0 | 1.00 |
| sinv traA | -7787.10 | -7787.10 | 1.00 | no | 0 | 1.00 |
| sinv traB | -7783.78 | -7785.18 | 0.09 | no | 0 | 260.08 |
| acep traA | -7787.08 | -7787.08 | 1.00 | no | 0 | 1.00 |
| acep traB | -7780.86 | -7781.33 | 0.33 | no | 3 | 2.64 |
| aech tra | -7786.12 | -7786.77 | 0.26 | no | 0 | 23.13 |

| Table 4. Tests for shifts in the evolutionary rate of tra homologues in ants using the PAML clade model C. (Species abbreviations are identical to those used in figure 1.) |
|-------------------------|-----------------|-----------------|-----------------|-----------------|
| branch | In likelihood for H0 | In likelihood for H1 | p-value | significant after FDR correction | number of sites with P(rate shift) >95% | rate in branch (ω) |
|-------------------------|-----------------|-----------------|-----------------|-----------------|
| hsal traA | -7774.10 | -7772.45 | 0.07 | no | 5 | 0 |
| hsal traB | -7774.10 | -7773.08 | 0.15 | no | 5 | 0.77 |
| 1hum traA | -7774.10 | -7774.00 | 0.66 | no | 4 | 0.69 |
| 1hum traB | -7774.10 | -7763.06 | 0.00 | yes | 24 | 3.27 |
| cflo traA | -7774.10 | -7773.53 | 0.29 | no | 3 | 999.00 |
| cflo traB | -7774.10 | -7766.44 | 0.00 | yes | 13 | 1.43 |
| pbar traA | -7774.10 | -7774.08 | 0.85 | no | 3 | 0.47 |
| pbar traB | -7774.10 | -7773.86 | 0.48 | no | 3 | 0.32 |
| sinv traA | -7774.10 | -7769.29 | 0.00 | yes | 3 | 0.03 |
| sinv traB | -7774.10 | -7774.08 | 0.84 | no | 3 | 0.44 |
| acep traA | -7774.10 | -7774.09 | 0.90 | no | 3 | 0.47 |
| acep traB | -7774.10 | -7768.58 | 0.00 | yes | 15 | 1.44 |
| aech tra | -7774.10 | -7772.51 | 0.07 | no | 0 | 1.44 |
here in the *Ap. mellifera* csd allele may alter the specificity of the variable C-terminal domains of the protein (RS repeat and proline-rich domains), which were suggested to participate in dimer formation [62]. Thereby, partial gene conversion events could generate novel alleles with altered binding specificities.

Other partial gene conversion events, as exemplified by the *At. cephalotes* paralogues, may affect the N-terminal domain of the protein. In contrast to the variability of the C-terminal domains within- and between-species, the N-terminal domain is highly conserved among all ant and bee homologues. Such a contrasting pattern is comparable to the contrast between purifying selection at the *fen* locus and positive selection at the neighbouring csd locus. Owing to their tight linkage, positive selection on csd could lead to a selective sweep (‘hitchhiking effect’) of linked deleterious alleles that may arise in *fen*. The hitchhiking effect associated with positive selection on csd was even observed in other loci farther away from *fen* and csd [63]. Therefore, selection is expected to favour recombinant haplotypes, where the linkage between the positively and negatively selected alleles is broken. Similarly, positive selection on the C-terminal domains of csd and *traB* could lead to a sweep of linked deleterious mutations in the N-terminal domain. Therefore, gene conversion events that copy the conserved sequence of the N-terminal domain from the slow-evolving parologue (*fen/trnA*) onto the fast-evolving parologue (*csd/trnB*) may act as a mechanism rescuing the deteriorating molecular function of this domain.

In conclusion, this study provides strong support for concerted evolution based on comparative sequence analyses of the two *tra* paralogues in ants and bees. We also find evidence for natural diversifying selection acting on one gene in each pair of paralogues, possibly resulting from balancing selection. These observations suggest that concerted evolution may be an important mechanism for diversifying the allele pool of genes under balancing selection.

This work was supported by several grants from the Swiss NSF and an ERC advanced grant. We thank Darren Martin for advice on the use of RDP4 and on the robustness of recombination inference to variation of selection. We thank Martin Beye, Leo Beukeboom and two reviewers for their useful comments on the manuscript. We thank Sanne Nygaard for input at early stages of this work.

References