Identification of a nematode chemosensory gene family

Nansheng Chen*†, Shraddha Pai*, Zhongying Zhao‡, Allan Mah‡, Rebecca Newbury§, Robert C. Johnsen‡, Zeynep Altun¶, Donald G. Moerman§, David L. Baillie‡, and Lincoln D. Stein*

*Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; ‡Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; ¶Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY 10461; and §Department of Zoology, University of British Columbia, Vancouver, BC, Canada V6T 1Z3

Communicated by Robert H. Waterston, University of Washington, Seattle, WA, November 15, 2004 (received for review June 15, 2004)

Taking advantage of the recent availability of the whole genome sequence of *Caenorhabditis briggsae*, a closely related nematode to *Caenorhabditis elegans*, we have examined the chemosensory gene superfamily by using comparative genomic methods. We have identified a chemosensory gene family, serpentine receptor class ab (*srab*), which exists in both species with 25 members in *C. elegans* and 14 members in *C. briggsae*. More than 20% of these gene models are reannotated. The *srab* family is similar to, but distinct from, the previously described serpentine receptor class a (*sra*) family and shows a differential expansion in *C. elegans* similar to that previously described for *sra*. The cellular expression patterns for multiple members of the *srab* family in both phasmid neurons in the tail and amphid neurons in the head supports the conclusion that they are chemosensory genes and suggests that they may play a role in integrating chemosensory inputs from both ends of the organism. The expansion of both the *srab* and *sra* gene families in *C. elegans* relative to *C. briggsae* is due to multiple rounds of tandem duplication and translocation of individual genes.

srab | duplication | cluster

Chemoreception, a term that encompasses olfaction, pheromonal sensation, hormonal signaling, sperm chemotaxis (1), and processes required for maintenance of the internal chemical milieu, is essential for animals in general (2, 3). The role of chemoreception in the nematode is even more prominent because worms such as *Caenorhabditis elegans* and *Caenorhabditis briggsae* are soil-dwelling and have no proper visual or auditory systems [but see Burr *et al.* (4), who showed crude light responsiveness in *C. elegans*].

A large portion of the genes of all sequenced animal genomes to date (≈1–5% for most known genomes) have been found to consist of confirmed and potential chemosensory genes (2, 5). Chemosensory genes in *C. elegans* were first identified by Troemel *et al.* (6) by searching the then-incomplete *C. elegans* genomic sequence, followed by GFP fusion expression studies of the promoters upstream of these putative chemosensory genes. A larger number of additional chemosensory genes have been reported at different stages of genome sequencing of *C. elegans* (5, 7, 8) by using similarity searches with known chemosensory genes as queries against the *C. elegans* sequences. More recently, Stein *et al.* (9) reported a large number of chemosensory genes for *C. briggsae*, the second nematode to be subjected to whole-genome sequencing. In that paper, we noted that *C. elegans* possesses almost 70% more chemosensory genes (718) than does *C. briggsae* (429) as defined by the distinct PFAM (Version 9.0) chemosensory gene families (10). Although each gene family is larger for *C. elegans* than that for *C. briggsae*, two families [7TM_5 and serpentine receptor class a (*sra*)] doubled in size.

To understand the basis of this differential gene family expansion, we have examined the genes of the *sra* gene family in *C. elegans* and *C. briggsae* in more detail. The *sra* gene family was chosen for the current study because it is relatively small (39 members in *C. elegans*, and 18 members in *C. briggsae*) and has been well studied, being the first identified chemosensory gene family in *C. elegans* (6). An understanding of *sra* gene family expansion will provide insights into other families in general. The questions we wished to answer were these: (*i*) Are the apparent differences in the size of the families in the two species real, or are they the result of an artifact such as missed gene predictions in *C. briggsae*? (*ii*) If real, what is the basis for gene expansion? In the course of this work, we discovered a putative chemosensory receptor family or subfamily, which added a third question to our list: (*iii*) Are the genes identified by sequence similarity searching putative chemosensory genes as determined by cellular expression pattern?

Methods

Data Mining. We started by searching WORMBASE (www.wormbase.org, Release WS110) for genes with PFAM motif PF02117, which is annotated as *C. elegans sra* chemosensory motif (PFAM 9.0) (11). To search for possibly misannotated genes, these genes (37 for *C. elegans*, and 18 for *C. briggsae*) were used as queries to BLAST (TBLASTN) (12) against whole-genome sequences of *C. elegans* and *C. briggsae*. An *e*-value cutoff value of 1×10^{-10} was used in this project because we did not want to have the BLAST hits contaminated by genes from other known families. Results obtained at different *e*-value cutoffs, ranging from 1×10^{-3} to 1×10^{-10}, demonstrated that we obtained a similar number of novel gene hits at these different values; however, we achieved many fewer contaminants from other known gene families with an *e* value of 1×10^{-10}. The BLAST hits were compared against the WORMBASE annotated gene models and were then classified into the following categories: (A) exact match to a query, (B) exact match with an existing gene model (most likely not annotated), (C) overlap with a query, (D) overlap with a known gene model, and (E) unknown fragments, which are hits that overlap with none of the annotated gene features. Hits in category A were ignored, and hits in category B were selected as new members. Hits in categories C and D were recruited as new members if the overlaps were longer than 100 bp. Hits in category E were subdivided into two types: those adjacent to a gene model of interest, and those adjacent to other E type hits. The former are putative missing exons of adjacent gene models, whereas the latter hits can potentially be combined to form *de novo* gene models.

The newly recruited genes and the WORMBASE *sra* genes were then used as queries to BLAST (BLASTP, $e = 1 \times 10^{-10}$) against whole proteomes of *C. elegans* and *C. briggsae* for potential hits.

To take the advantage of the existence of two genomes, the above procedures (TBLASTN and BLASTP) were also carried out by using genes from the opposite species as query, i.e., using *C. elegans* genes as queries to BLAST against the *C. briggsae* database, and *vice versa*.

Freely available online through the PNAS open access option.

Abbreviations: sra, serpentine receptor class a; srab, serpentine receptor class ab; TM, transmembrane domain.

†To whom correspondence should be addressed at: Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724. E-mail: chenn@cshl.org.

Ortholog Assignment. Because of gene duplication before and after speciation, orthologous relationship can have any of the following types: one-to-one, one-to-many, many-to-one, and many-to-many (Figs. 3 and 4) (13). To a first approximation, orthologs can be determined by identifying sets of related genes at the outer branches of the tree (Tables 3 and 4, which are published as supporting information on the PNAS web site). Also, we have assumed approximately equal rates of protein sequence evolution. A one-to-one relationship will appear as a pair of genes, one each from *C. elegans* and *C. briggsae*. This is the signature of a gene present in the common ancestor of the two species that has undergone neither gene amplification nor gene loss. A one-to-many relationship will appear as a cluster of related genes in which one gene is from either *C. elegans* or *C. briggsae* and the others are from the opposite species. This is the signature of a gene derived from a common ancestor that has undergone expansion in one species but not the other. In contrast, a gene that has lost its orthologue because of deletion or pseudogene conversion will appear as an "orphan" that has a long branch length to its closest neighbors in the opposite species.

Gene-Model Improvement. Chemosensory genes are examples of G protein-coupled receptors (3), which in turn are seven transmembrane domain (TM)-containing genes. We took the advantage of this fact to validate the gene models of the predicated genes by using the hidden Markov model-based program TMHMM (Fig. 5, which is published as supporting information on the PNAS web site) (13). For genes that did not contain seven TMs, we used the program GENEWISE (14, 15) to "repair" them. Genes that contained in-frame stop codons after the GENEWISE procedure were declared as hypothetical pseudogenes.

Phylogenetic Analysis. The procedure was described in Stein *et al.* (9). Briefly, multiple sequences were aligned by using CLUSTALW (16). The aligned result in PHYLIP (http://evolution.genetics.washington.edu/phylip.html) format was then fed into the programs SEQBOOT, PROTDIST, NEIGHBOR, and CONSENSE in the PHYLIP package (17) to construct a neighbor-joining tree. For bootstrap analyses, 1,000 data sets were created by the SEQBOOT program in the PHYLIP package.

Expression Assay Using Promoter::GFP Fusion. Promoter::GFP fusion constructs were prepared as described in ref. 18. Pictures were taken with a QImaging digital camera mounted on a Zeiss Axioskop and a Zeiss LSM5 Pascal confocal system mounted on an inverted Zeiss Axioskop.

Results

Data Mining. One possible trivial explanation for the apparent 2-fold expansion of *sra* genes in *C. elegans* relative to *C. briggsae* is that a systematic failure in gene prediction has caused an undercount of *sra* genes in the newly annotated *C. briggsae* genome. To address this issue, we initiated an exhaustive search for missed members of the family in both the *C. briggsae* and *C. elegans* genomes. We retrieved all of the PFAM-annotated *sra* chemosensory genes from WORMBASE (19) by using the WS110 reference release. There were 39 *C. elegans sra* genes, including 4 annotated pseudogenes and 18 *C. briggsae sra* genes. Two *C. elegans sra* genes are alternatively spliced (F44F4.5 and F49E12.5).

We then used the protein sequences of these *sra* genes as TBLASTN ($e = 1 \times 10^{-10}$) queries against the *C. elegans* and *C. briggsae* genome sequences. This procedure identified 16 new matches in *C. elegans* and 17 new matches in *C. briggsae*. All of the matches identified by this procedure had been previously identified by gene predictions programs but had not been annotated as belonging to the chemosensory gene superfamily. Of note is that all of the 17 genes identified in *C. briggsae* were discovered by using *C. elegans* genes as queries, demonstrating the usefulness of comparative genomics for gene discovery and annotation.

Table 1. List of *sra* and *sra*-like genes

Stage	Family	Method	*C. elegans*	*C. briggsae*
1	*sra*	PFAM	39 (4*)	18
	sra-like	TBLASTN	16	17
		BLASTP	10	2
		TBLASTN + BLASTP	28	17
2	*sra*	CLUSTALW + PHYLIP	41	23
	sra-like	CLUSTALW + PHYLIP	25	14
3	*sra*	GENEWISE	41 (9*)	23 (2*)
	sra-like	GENEWISE	25 (0*)	14 (3*)

Stages: 1, initial similarity screening; 2, after phylogenetic analysis; 3, after gene model repairing.
*Annotated as pseudogene in WORMBASE.

To further extend the search, we used the protein sequences from the newly identified chemosensory gene candidates as queries to BLASTP against the *C. elegans* and *C. briggsae* protein sets again using an *e* cutoff of 1×10^{-10}. Ten more chemosensory gene candidates were found in *C. elegans* and two more in *C. briggsae*. Subsequent analysis described below demonstrated that four of the *C. elegans sra* candidates and five of the *C. briggsae sra* candidates likely were pseudogenes. Taken together, we identified 26 *sra*-like genes for *C. elegans* and 19 *sra*-like genes for *C. briggsae*. Some genes are pseudogenes, which are discussed in the following sections (Table 1).

Surprisingly, although the genes identified by this data mining procedure had nucleotide- and protein-level similarity to known members of the *sra* family, all but two of them (see below) lacked the sra domain that defines the *sra* family in the PFAM database. To explore the relationship between the newly found genes and PFAM-defined *sra* genes, we constructed a merged data set of the *sra* and *sra*-like genes and pseudogenes for both *C. elegans* and *C. briggsae*. We then constructed a phylogenetic tree from this data set by using the CLUSTALW (16) and PHYLIP (17) packages (Fig. 1) as described in ref. 9. With four exceptions discussed below, the *sra* and *sra*-like genes segregate to two distinct sections of the tree. The exceptions are two *C. elegans* genes, C47A10.6 and T21H8.3, which contain PFAM-defined sra motifs but cluster with the *sra*-like genes. Similarly, two *C. briggsae sra*-like genes, CBG13454 and CBG13479,

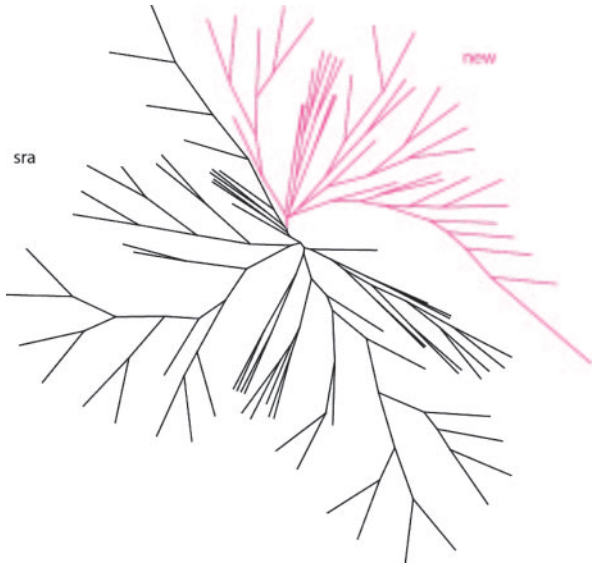

Fig. 1. Phylogenetic analysis of *sra* and *sra*-like genes. Spliced nucleotide sequences for *sra* and *sra*-like genes for *C. elegans* and *C. briggsae* were clearly segregated in the phylogenetic tree. Branches for *sra* genes are coded in black, and branches for *sra*-like genes are coded in red.

are placed in the *sra* portion of the tree, despite their not having a PFAM-defined sra domain. On the basis of the phylogenetic tree, we reassigned C47A10.6 to the *sra*-like gene set and CBG13454 and CBG13479 to the *sra* set. However, we kept the *sra* motif-containing gene T21H8.3, together with its two neighboring genes in the *C. elegans* genome, T21H8.2 and T21H8.4, in the *sra* gene family to avoid confusion, because they have been assigned *sra* family names. Accordingly, we assigned their orthologous genes, CBG07352, CBG07353, and CBG07355, as *C. briggsae sra* members. The final data sets (Table 1) comprised 41 *C. elegans sra* genes (including 9 pseudogenes), 23 *C. briggsae sra* genes (2 pseudogenes), 25 *C. elegans sra*-like genes (no pseudogenes), and 14 *C. briggsae sra*-like genes (3 pseudogenes).

Comparative Analysis of *sra* and *sra*-Like Genes. The phylogenetic tree suggests that the *sra* and *sra*-like gene sets diverged before the speciation of *C. elegans* and *C. briggsae*. To further explore this possibility, we examined the physical position of the gene families on the *C. elegans* and *C. briggsae* genomes. We retrieved the genomic coordinates of the *sra* and *sra*-like genes and pseudogenes from WORMBASE (Release WS110) (19) and compared their physical clustering patterns. Of 41 *sra* genes in *C. elegans*, 26 reside in 6 clusters on chromosome II, 8 reside on chromosome I as a single cluster, and the remaining 6 genes are on chromosomes IV, V, and X, respectively. In contrast, 23 of the 25 *sra*-like genes in *C. elegans* reside in seven clusters on chromosomes V. Two are located on chromosome II and IV. None of the *sra*-like genes are found on chromosomes I, III, or X. The distinct physical distribution of the two gene sets is most consistent with a model in which the common ancestor of the *sra* and the *sra*-like genes was first duplicated and the two copies then physically segregated onto chromosomes II and V, respectively, before further evolutionary divergence. This event presumably occurred in the common ancestor of the *C. elegans* and *C. briggsae* species.

***sra*-Like Genes Are Chemosensory.** Although the *sra*-like genes are clearly related to the *sra* genes, and may either represent a closely related family or a distinct subfamily of these genes, this does not necessarily imply that the identified genes play a chemosensory role in the lifecycle of the organism. To address this question, we used two approaches to find evidence that the *sra*-like genes are chemosensory. First, we used the microarray-based gene expression clusters described by Kim *et al.* (20) to examine whether the *sra*-like genes were temporally coexpressed with known chemosensory genes. Using the "expression topology map" reported by these authors, we determined that the *sra* genes cluster in mountains #0 (18 members), #9 (6 members), #13 (4 members), #3 (3 members), and #10 (3 members). The *sra*-like genes have a similar distribution over the expression map and are clustered at expression mountains # 0 (13 members), #9 (3 members), #3 (4 members), #10 (4 members), and #13 (2 members). This finding suggests that in *C. elegans*, the expression patterns of the *sra* and *sra*-like genes are similar at different life stages of the organism and under different pharmacological, genetic, and environmental conditions. This observation, in turn, implies that the two sets of genes play similar physiological roles. Corresponding expression data in *C. briggsae* are not yet available.

Our second approach was to directly assay the anatomic expression pattern of the *sra*-like genes in *C. elegans*. To do this, we generated promoter::GFP fusion transgenic *C. elegans* lines as described in *Methods*. Of the seven genes attempted, we were able to successfully express six fused gene constructs. In each of the transgenic lines, we observed GFP fluorescence limited mostly to the head (amphid and labial) and the tail (phasmid) chemosensory neurons (Fig. 2 and Table 2). One *sra*-like promoter::GFP fusion construct (T20D4.1) was exclusively expressed in a pair of tail phasmid neurons PHAL/R (Fig. 2), two (C04F5.4 and C36C5.6) were exclusively expressed in a pair of the head amphid neurons

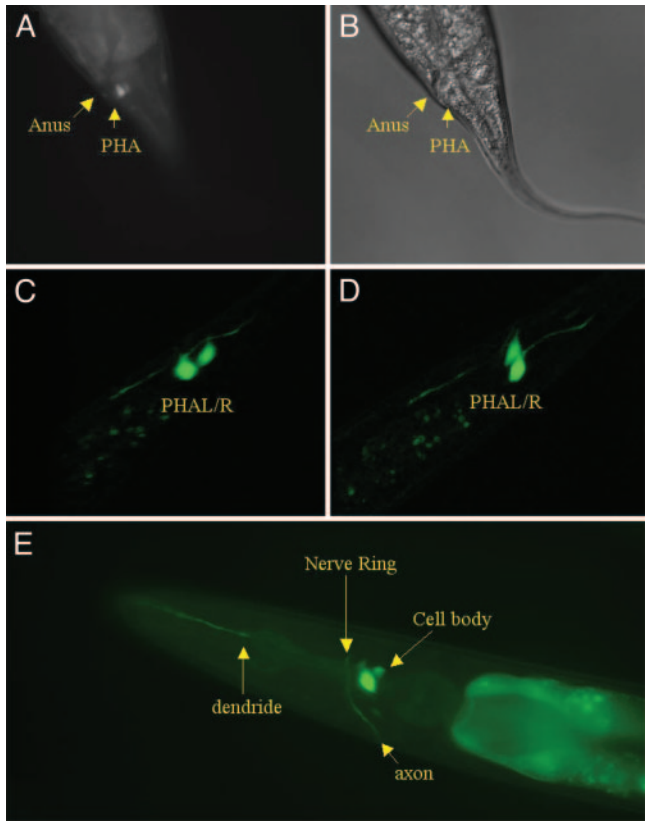

Fig. 2. Cellular expression patterns of *sra*-like genes. Shown is the tail of a *C. elegans* with expression of T20D4.1 promoter fused with GFP coding region. (*A*) A pair of PHA GFP-positive neurons. (*B*) Differential interference contrast view of the tail shown in *A*. (*C* and *D*) Three-dimensional reconstructed views of a pair of PHA neurons in the *C. elegans* tail with expression of GFP fused with T20D4.1 promoter.

(Fig. 2*E*), and two (T21H8.4 and C47A10.6) were expressed in both head (phasmid for T21H8.4, and labial for C47A10.6) and tail neurons (phasmid PHAL/R and PHBL/R for T21H8.4, and phasmid PHCL/R for C47A10.6). C47A10.6 and C33G8.5 also showed medium to strong expression in scattered nonchemosensory neurons. Together, these data support a chemosensory role for the *sra*-like genes.

Selective Expansion of *sra* and *sra*-Like Genes in *C. elegans*. After a thorough reannotation effort, we could confirm our earlier observations that the sra gene family is selectively expanded in *C. elegans* relative to *C. briggsae*. Somewhat surprisingly, however, we found that the *sra*-like genes were also more abundant in *C. elegans* by about the same 2-fold ratio, despite the fact that phylogenetic evidence suggested that the *sra* and *sra*-like gene sets diverged

Table 2. GFP promoter fusion expression patterns in *C. elegans*

Gene	Head neurons	Tail neurons
T20D4.1	—*	PHB
T21H8.4	ASJ and ADL	PHA and PHB
C33G8.5	AWB, I2, I4, and others (NSM, pm6, RID, VA, VB, DVC, AFD, RME, AIA, and PVT)	—
C36C5.6	AWA	—
C47A10.6	IL1s, IL2s, and others (PVP, BDU, CEPD, CEPV, PVD, SDQL, ADA, or RMG)	PHC and others (LUA)
C04F5.4	AWB	—
T20D4.18	—	—

*No GFP expression detected.

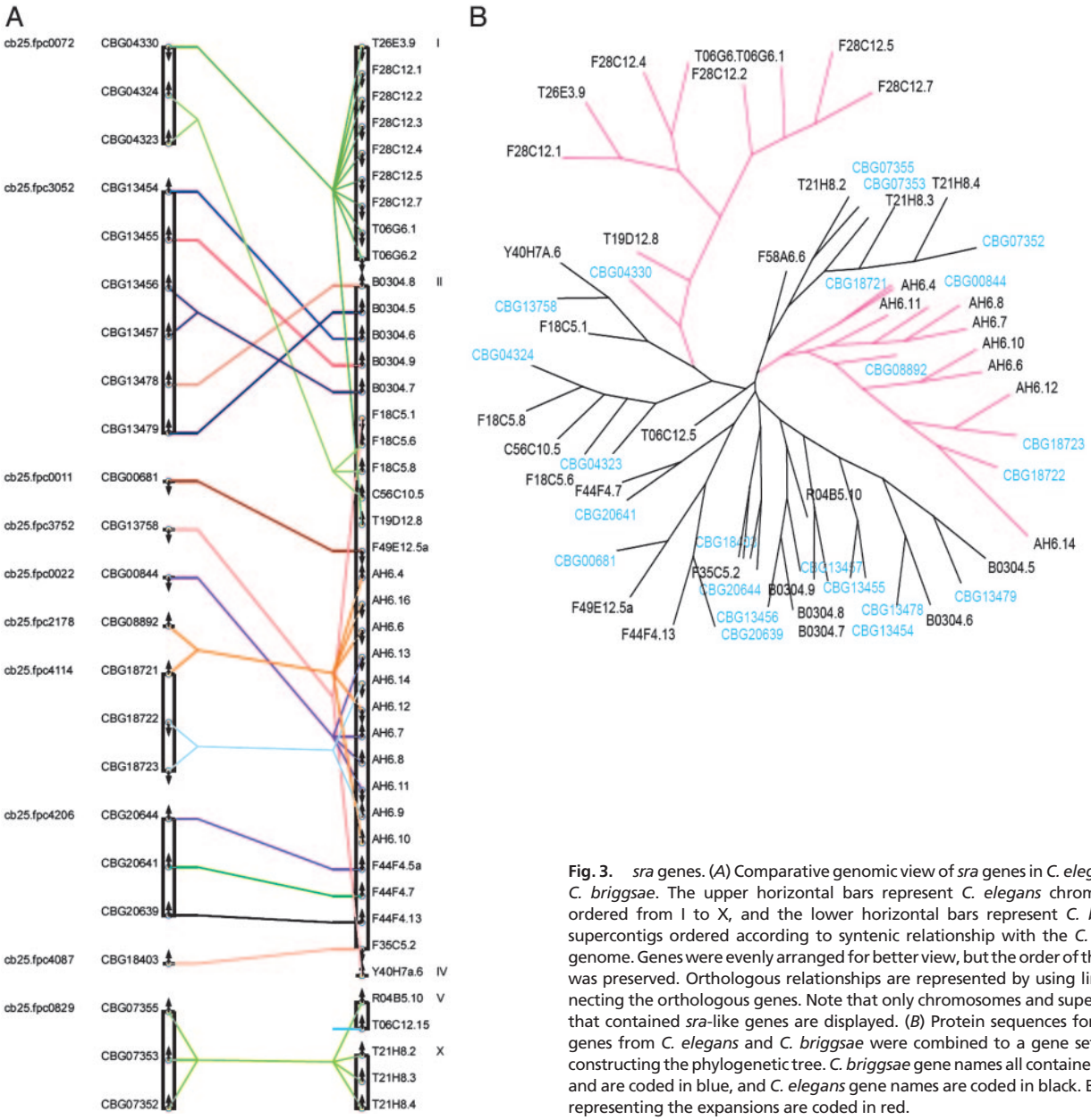

Fig. 3. *sra* genes. (*A*) Comparative genomic view of *sra* genes in *C. elegans* and *C. briggsae*. The upper horizontal bars represent *C. elegans* chromosomes ordered from I to X, and the lower horizontal bars represent *C. briggsae* supercontigs ordered according to syntenic relationship with the *C. elegans* genome. Genes were evenly arranged for better view, but the order of the genes was preserved. Orthologous relationships are represented by using lines connecting the orthologous genes. Note that only chromosomes and supercontigs that contained *sra*-like genes are displayed. (*B*) Protein sequences for the *sra* genes from *C. elegans* and *C. briggsae* were combined to a gene set before constructing the phylogenetic tree. *C. briggsae* gene names all contained "CBG" and are coded in blue, and *C. elegans* gene names are coded in black. Branches representing the expansions are coded in red.

before the segregation of the two species. This finding suggests that some evolutionary process has been driving the differential composition of the *sra* and *sra*-like gene sets in the two species.

There are two general models for this differential composition. One is that the *sra* and *sra*-like genes were expanded in *C. elegans* versus its sister species by tandem duplication and translocation, and the other is that there is an accelerated loss of *C. briggsae* genes, either by conversion into pseudogenes or by gene deletion. A third explanation, that a subset of *C. briggsae sra* and *sra*-like genes were simply missed because of poor annotation, was effectively eliminated by the data-mining results reported herein. Two additional explanations, (*i*) large deletions in *C. briggsae* of the expanded clusters and (*ii*) incomplete *C. briggsae* sequence, are unlikely because of the family-selective fashion of the expansion of gene families and the fact that more genes have been annotated in the whole *C. briggsae* genome (9). However, these two explanations cannot be completely ruled out.

There are 18 more *sra* genes (including pseudogenes) in *C. elegans* than in *C. briggsae*. According to the definition of orthologous relationship detailed the *Methods*, there are altogether 10 one-to-one orthologous pairs (Fig. 3 *A* and *B* and Table 3). The majority of these one-to-one orthologous pairs fall within larger conserved synteny blocks (data not shown). The differential composition of *C. elegans* vs. *C. briggsae* genes is in great part due to orthologous relationships of just 3 *C. briggsae* genes that are orthologous to multiple *C. elegans* genes. The remaining 5 additional *C. elegans* genes are involved in many-to-many relationships or, in the case of a single *C. elegans* pseudogene (Table 3), are orphans.

A single *C. briggsae* gene, CBG04330, corresponds to 10 *C. elegans* orthologs. Looking at this event more closely, we find that the *C. elegans* gene that is most similar to CBG04330 in the orthologs set, T19D12.8, resides on chromosome I, whereas the 9 less similar *C. elegans* genes reside in a tandemly duplicated cluster on chromosome II (Fig. 3 *A* and *B*). This finding is consistent with a series of events in which the common ancestor of CBG04330 and T19D12.8 first underwent a duplication event in the *C. elegans* lineage. The duplicate was then translocated to chromosome II, and

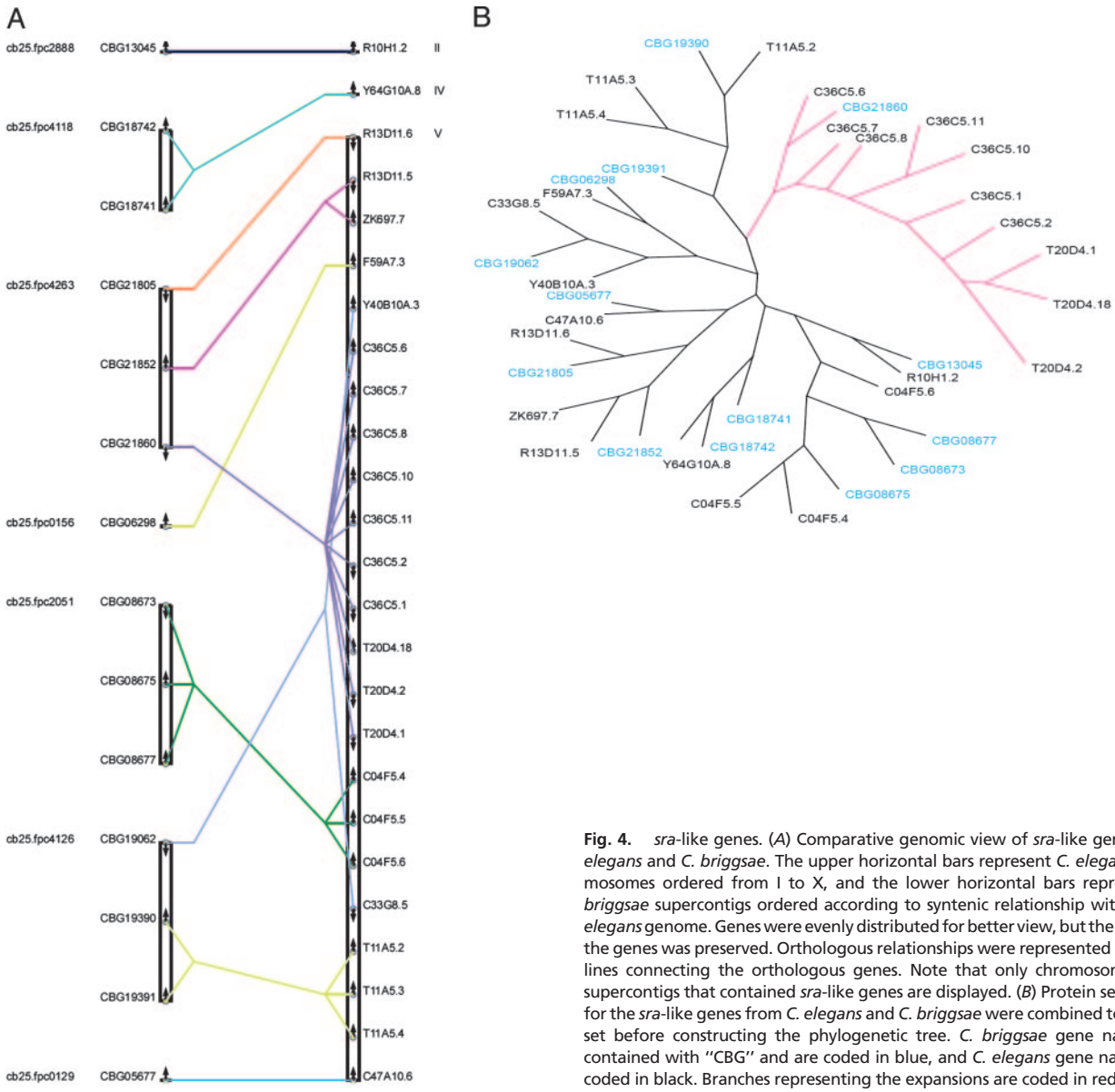

Fig. 4. *sra*-like genes. (*A*) Comparative genomic view of *sra*-like genes in *C. elegans* and *C. briggsae*. The upper horizontal bars represent *C. elegans* chromosomes ordered from I to X, and the lower horizontal bars represent *C. briggsae* supercontigs ordered according to syntenic relationship with the *C. elegans* genome. Genes were evenly distributed for better view, but the order of the genes was preserved. Orthologous relationships were represented by using lines connecting the orthologous genes. Note that only chromosomes and supercontigs that contained *sra*-like genes are displayed. (*B*) Protein sequences for the *sra*-like genes from *C. elegans* and *C. briggsae* were combined to a gene set before constructing the phylogenetic tree. *C. briggsae* gene names all contained with "CBG" and are coded in blue, and *C. elegans* gene names are coded in black. Branches representing the expansions are coded in red.

this was followed by a series of tandem duplications to give rise to the current chromosome II cluster.

The other many-to-one and many-to-many orthologous sets show a similar pattern in which the *C. elegans* orthologs reside in clusters of tandemly duplicated genes. Together, these observations suggest that a small number of differential expansion events drove the expansion of the *sra* gene family in *C. elegans*.

A similar pattern of selective expansion in *C. elegans* was seen for the PFAM-identified *sra*-like genes. Every *sra*-like gene in *C. elegans* has an orthologous partner in *C. briggsae* (25 for *C. elegans* and 14 for *C. briggsae*, pseudogenes included; Fig. 4 *A* and *B* and Table 4). No orphans were identified for *sra*-like genes, arguing against the model that genes in *C. briggsae* were lost. The great majority of the orthologous relationships are of the one-to-many kind in which a single *C. briggsae* gene corresponds to multiple *C. elegans* genes. The most extreme case is the orthologous relationship that includes *C. briggsae* gene CB21860, which corresponds to 10 genes in *C. elegans*. Further supporting the expansion model, we found that in the one-to-many orthologous groups, the *C. elegans* genes were usually physically close to each other (Fig. 6, which is published as supporting information on the PNAS web site).

Gene-Model Improvement. Having identified the orthologs between the two species, we were able to use this information to refine the gene-model predictions. First, we took advantage of the fact that chemosensory genes are G protein-coupled receptors containing seven TMs (3, 21) by using the hidden Markov model-based program TMHMM (22) to identify those members of the *sra* and *sra*-like families that were missing one or more TM. In the *sra*-like gene set, 16 of 25 genes had seven predicted TMs, and 7 of 28 had six predicted TMs. For all of the genes with fewer or more than seven predicted TMs, we attempted to repair the gene models by using intra- and interspecies homology data as described in *Methods*. In this way, the gene models for seven *C. elegans* and four *C. briggsae sra*-like genes were successfully repaired, restoring seven TMs to five *C. elegans* genes (C36C5.6, T20D4.1, T20D4.2, T20D4.18, and T11A5.4) and two *C. briggsae* genes (CBG21860 and CBG18742). The gene-repairing procedure identified two *C. briggsae* genes (CBG08675 and CBG18741) in the *sra*-like gene set as hypothetical pseudogenes by virtue of having premature stop codons. In summary, after gene-model repairing, there are 21 *C. elegans sra*-like genes with seven TMs, and three *C. elegans sra*-like genes with six TMs. In *C. briggsae*, there are eight *sra*-like genes with

seven TMs, seven *sra*-like genes with six TMs, and two hypothetical pseudogenes.

A similar procedure was applied to the *sra* gene set. Results indicated that five *C. elegans sra* genes (AH6.12, B0304.9, F44F4.13, Y40H7A.6, and F49E12.5) could be repaired. Five *C. elegans sra* genes (R04B5.10, F28C12.1, F18C5.1, R04B5.10, and B0304.7) are hypothetical pseudogenes, in addition to the four hypothetical pseudogenes annotated by WORMBASE (Release WS110). Three *C. briggsae sra* genes (CBG19390, CBG13454, and CBG13479) are hypothetical pseudogenes (Table 3).

The repaired gene models have been submitted to the curators of WORMBASE.

Discussion

Because a significant portion of predicted gene models, especially the G protein-coupled receptors in the case of *C. elegans* (23), are likely imperfect with inappropriate intron–exon splicing sites, missing introns and exons, and other defects, careful examination and improvement of predicted gene models is a necessary prerequisite to comparative protein-family analysis.

We have identified a set of 25 genes in *C. elegans* and 14 genes in *C. briggsae* (Table 3) that are related to the *sra* family of nematode chemosensory genes but do not contain the sra protein domain signature that is the defining characteristic of the *sra* family. These *sra*-like genes could either be considered a distinct subfamily of *sra* or a separate family in its own right. Although the distinction between these two possibilities is largely a semantic one, a number lines of evidence argue that it would be better to consider these as distinct families rather than subfamilies.

First, the *sra* and *sra*-like genes are easily distinguished on the basis of their phylogenetic relatedness (Fig. 1), and the distinction between the two sets clearly precedes the speciation of *C. elegans* and *C. briggsae*. The two sets of genes reside in different regions of the genome, with the *sra*-like genes present on chromosomes V and the *sra* genes typically found on chromosomes I and II.

A stronger argument for declaring the *sra*-like genes to be a distinct family comes from the cellular expression pattern. The *sra* genes are reported to have a cellular expression pattern (6) that is distinct from the pattern we observed for the *sra*-like genes. The genes *sra-1* and *sra-6* are expressed in male spicules and in the neurons SPD and SPV. The genes *sra-7* and *sra-9* are expressed in the amphid ASK neuron. The gene *sra-10* is expressed in URX sensory neuron, the AVB interneuron, and a pharyngeal neuron. The gene *sra-11* is expressed in AIY interneuron. Strikingly, none of these *sra* genes with known expression pattern is expressed in the PHA, PHB, or PHC neurons of the nematode chemosensory system. In contrast, half (three of six) of the *sra*-like genes that we assayed with promoter::GFP constructs were expressed in the phasmid PHA/PHB neurons (Table 2 and Fig. 2). Also, none of these six genes was found to be expressed in male-specific neurons.

On the basis of these arguments, we propose to declare the *sra*-like genes a separate family of chemosensory genes, and propose the name serpentine receptor class ab (*srab*) for this family.

The expression pattern of the *srab* genes is biologically intriguing. Of the six promoters successfully expressed in transgenic organisms, one was exclusively expressed in the tail phasmid neurons, two were exclusively expressed in a head amphid neuron, and two were expressed both in the head and tail neurons as well as a limited number of other cells (Table 2 and Fig. 2). A recent report has provided evidence that *C. elegans* can integrate chemosensory input from both the head and the tail to coordinate behavior (24). The expression of several of these genes (e.g., C47A10.6) in both the head and tail neurons suggests that they may play a role at the molecular level in integrating the chemical messages received at these two sites.

By examining the orthologous regions of the two species, we have demonstrated that the difference in size of the *sra* and *srab* families between *C. elegans* and *C. briggsae* is most likely due to just a few tandem duplication events in the *C. elegans* lineage, followed in some cases by a translocation of a portion of the region to another region of the genome. It is intriguing that this mechanism of expansion affects both the *sra* and *srab* families at roughly the same rate, even though the two families were separated before the divergence of *C. elegans* from *C. briggsae*. Furthermore, the increased rate of tandem duplication in *C. elegans* does not seem to be a general feature of multigene families, because most other large nematode gene families, including other chemosensory receptor types, do not show a differential increase in size. This observation suggests that the difference in family size may be adaptive, although the nature of the adaptation is obscure.

The identification of the *srab* gene family, the insights gained into the mechanisms of gene family evolution, and the practical importance of the gene-model improvements all demonstrate the importance of comparative genomics in the study of nematode chemoreception. We are eager to extend these methodologies to other putative chemosensory receptor gene families, to develop a comprehensive catalog of this large and biologically important superfamily. The endeavor will be assisted in coming months by the planned sequencing of the genomes for three more related nematode species (www.genome.gov/10002154). Ultimately, the full identification of chemosensory genes in *C. elegans* and other nematodes will help our understanding of the evolution of olfaction in general and will assist in studying the physiology of chemoreception in *C. elegans*.

We thank Dr. David Hall for assistance in identifying neurons in *C. elegans*; Drs. Nancy Hawkins and Hugh Robertson for fruitful discussion; Dr. Jonathan Hodgkin for communication regarding the CGC gene names; Drs. Zachary Mainen, Josh Dubnau, Andrew Neuwald, Tristan Fiedler, and Sheldon McKay for critical reading of the manuscript; and the reviewers for their critical suggestions. Sheldon McKay designed primers for PCR reactions. D.L.B. and D.G.M. are supported by grants from Genome British Columbia, Genome Canada, and the Natural Sciences and Engineering Research Council (Canada). N.C. and L.D.S. are supported by a National Human Genome Research Institute grant. S.P. was supported by the Olney Fund.

1. Spehr, M., Gisselmann, G., Poplawski, A., Riffell, J. A., Wetzel, C. H., Zimmer, R. K. & Hatt, H. (2003) *Science* **299,** 2054–2058.
2. Buck, L. B. (2000) *Cell* **100,** 611–618.
3. Mombaerts, P. (2004) *Nat. Rev. Neurosci.* **5,** 263–278.
4. Burr, A. H. (1985) *Photochem. Photobiol.* **41,** 577–582.
5. Robertson, H. M. (1998) *Genome Res.* **8,** 449–463.
6. Troemel, E. R., Chou, J. H., Dwyer, N. D., Colbert, H. A. & Bargmann, C. I. (1995) *Cell* **83,** 207–218.
7. Robertson, H. M. (2000) *Genome Res.* **10,** 192–203.
8. Robertson, H. M. (2001) *Chem. Senses* **26,** 151–159.
9. Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., *et al.* (2003) *PLoS Biol.* **1,** E45.
10. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. (2002) *Nucleic Acids Res.* **30,** 276–280.
11. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999) *Nucleic Acids Res.* **27,** 260–262.
12. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
13. Remm, M. & Sonnhammer, E. (2000) *Genome Res.* **10,** 1679–1689.
14. Birney, E. & Durbin, R. (2000) *Genome Res.* **10,** 547–548.
15. Birney, E., Clamp, M. & Durbin, R. (2004) *Genome Res.* **14,** 988–995.
16. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
17. Felsenstein, J. (1988) *Annu. Rev. Genet.* **22,** 521–565.
18. Hobert, O. (2002) *BioTechniques* **32,** 728–730.
19. Harris, T. W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Chan, J., *et al.* (2004) *Nucleic Acids Res.* **32,** D411–D417.
20. Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N. & Davidson, G. S. (2001) *Science* **293,** 2087–2092.
21. Mombaerts, P. (1999) *Science* **286,** 707–711.
22. Sonnhammer, E. L., von Heijne, G. & Krogh, A. (1998) *Proc. Int. Conf. Intel. Syst. Mol. Biol.* **6,** 175–182.
23. Reboul, J., Vaglio, P., Rual, J. F., Lamesch, P., Martinez, M., Armstrong, C. M., Li, S., Jacotot, L., Bertin, N., Janky, R., *et al.* (2003) *Nat. Genet.* **34,** 35–41.
24. Hilliard, M. A., Bargmann, C. I. & Bazzicalupo, P. (2002) *Curr. Biol.* **12,** 730–734.