Chem. Senses 26: 151–159, 2001

Updating the *str* and *srj* (*stl*) Families of Chemoreceptors in *Caenorhabditis* Nematodes Reveals Frequent Gene Movement Within and Between Chromosomes

Hugh M. Robertson

Department of Entomology, University of Illinois at Urbana—Champaign, 505 S. Goodwin, Urbana, IL 61801, USA

Correspondence to be sent to: H. M. Robertson, Department of Entomology, University of Illinois at Urbana—Champaign, 505 S. Goodwin, Urbana, IL 61801, USA. e-mail hughrobe@uiuc.edu

Abstract

The seven transmembrane receptor (*str*) and *srj* (renamed from *stl*) families of chemoreceptors have been updated and the genes formally named following completion of the *Caenorhabditis elegans* genome sequencing project. Analysis of gene locations revealed that 84% of the 320 genes and pseudogenes in these two families reside on the large chromosome V. Movements to other chromosomes, especially chromosome IV, have nevertheless been relatively common, but only one has led to further gene family diversification. Comparisons with homologs in *C. briggsae* indicated that 22.5% of these genes have been newly formed by gene duplication since the species split, while also showing that four have been lost by large deletions. These patterns of gene evolution are similar to those revealed by analysis of the equally large *srh* family of chemoreceptors, and are likely to reflect general features of nematode genome dynamics. Thus large random deletions presumably balance the rapid proliferation of genes and their degeneration into pseudogenes, while gene movement within and between chromosomes keeps these nematode genomes in flux.

Introduction

The nematode *Caenorhabditis elegans* has a large repertoire of chemoreceptor genes (Troemel *et al.*, 1995; Bargmann, 1998; Robertson, 1998, 2000). Approximately 800 genes and pseudogenes belong to one major superfamily consisting of the large seven transmembrane receptor (*str*), *srh*, *srd* and related families (Robertson, 1998, 2000), while another approximately 300 belong to the *sra*, *srb*, *sre*, *srg* and other families (Bargmann, 1998) (H.M. Robertson, unpublished data). In addition to revealing interesting patterns of molecular evolution within these families, with implications for the chemosensory capabilities of this and related nematodes, I have previously employed these large gene families to illuminate several aspects of the genome dynamics of these nematodes. Analysis of 179 genes and 71 pseudogenes in the large *str* family and a smaller related family revealed abundant patterns of recent and more ancient gene duplications, commonly in large tandem arrays (Robertson, 1998). Genes were frequently reduced to pseudogene status; these pseudogenes were all formed relatively recently. Mapping of intron evolution on a phylogenetic tree revealed frequent losses and just one gain within the *str* family. Although indications of abundant gene movement were observed, these were not analysed further.

Analysis of the large *srh* gene family, with 214 genes and 90 pseudogenes, confirmed that these patterns of gene evolution were common (Robertson, 2000), although in this case seven intron gains were inferred within the family. Analysis of the patterns of DNA deletions within the *srh* family showed that removal of pseudogenes probably results from the common occurrence of large deletions. Completion of sequencing of the *C. elegans* genome, and comparison with *srh* chemoreceptor orthologs in the partially completed sequence of the *C. briggsae* genome, also revealed that some genes have been completely lost from the *C. elegans* genome, while perhaps 28% of the *srh* family chemoreceptors in *C. elegans* have been newly formed since the split with *C. briggsae*. Finally, 82% of the *srh* family genes and pseudogenes occur on the large chromosome V; mapping of gene location on a phylogenetic tree revealed that movements to other chromosomes have been common (27 altogether), but only twice have led to amplification of new gene lineages on other chromosomes.

With completion of the sequencing of the *C. elegans* genome (*C. elegans* Sequencing Consortium, 1998; *C. elegans* Genome Consortium, 1999) it is now possible to provide a complete description and formal naming of the *str* family, as well as the related family previously called *stl* but here renamed *srj*. Phylogenetic analysis of these two gene

families confirms several of the genome dynamics inferred from the *srh* family, including loss of *C. briggsae* orthologs, recent formation of many genes within *C. elegans* and the frequent occurrence of movements of genes between chromosomes. In addition, preliminary analysis of gene location within chromosome V revealed frequent gene movement within it.

Materials and methods

The public DNA databases were searched using TBLASTN for relatives of all major gene lineages of the *str* and *srj* (*stl*) families (Robertson, 1998) and a PSI-BLASTP search was used to identify any additional annotated proteins that might belong within these two families (Altschul *et al.*, 1997). Genes and their conceptual translations were aligned by eye with the original *str*/*stl* family datasets, but because alignment of the transmembrane (TM) regions 4 and 5 was not simple, a protein alignment obtained using Clustal X (Jeanmougin *et al.*, 1998) at default settings was employed for the phylogenetic analyses. This alignment agreed, in large part (particularly the blocks of TM1, 2, 3, 6 and 7 alignment), with the manual alignment [see Figure 1 of Robertson (Robertson, 1998)]. Clustal X performs better at aligning the hydrophobic membrane-spanning TM4 and 5 regions. Manual adjustment of the Clustal X alignment was nevertheless necessary to correct some gaps, particularly those resulting from deletions within pseudogenes. All amino acid positions were employed for the phylogenetic analyses to provide the maximum possible information within subfamilies. Phylogenetic analysis was performed using neighbor-joining (NJ) followed by tree-bisection-and-reconnection branch swapping as implemented by PAUP* v4.0b2a (PPC) for the Macintosh (Swofford, 1998). Bootstrap analysis employed 1000 NJ replications. Additional bootstrap analyses were performed on four subsets of the dataset represented by the four segments of Figure 1. In addition a maximum parsimony (MP) analysis was undertaken employing the heuristic algorithm of PAUP*, with tree-bisection-and-reconnection branch swapping and 100 interations of random sequence addition. The frequencies of synonymous (K_s) and non-synonymous (K_a) changes were computed following Nei and Gojobori (1986) using the Macintosh program KsKaCalc (H. Akashi, personal communication).

Results

The updated *str* and *srj* families

The updated *str* (seven transmembrane receptor) family consists of 189 genes and 74 pseudogenes. The related *srj* family consists of 39 genes and 18 pseudogenes. This family was previously called the *stl* family; however, this gene name has been reserved, so this family is being renamed in the sr (serpentine receptor) gene name series initiated by Troemel *et al.* (1995). In addition, 23 homologs in these two families are available from the partial genome sequence of the congener *C. briggsae* (see below). Conceptual translations were aligned with the previous dataset (Robertson, 1998), but for the phylogenetic analysis an alignment generated with Clustal X was employed. Phylogenetic analysis of this large dataset of 343 protein sequences is difficult. MP analysis, which was employed for these two families previously, using the heuristic algorithm of PAUP* yielded six equally parsimonious trees 37 647 steps long with a consistency index of 0.15 and required 7 days on a 300 MHz G3 Power Macintosh computer to examine >2 trillion trees. However, this island of very similar trees was found only once, leaving the possibility that shorter trees exist. Therefore NJ was employed, followed by the heuristic minimum evolution (ME) algorithm of PAUP*, which examined >5 million rearrangements using tree-bisection-and-reconnection branch swapping resulting in a tree 0.2% shorter. This tree is shown in four sections in Figure 1, with the *srj* family designated as the outgroup based on its location in preliminary phylogenetic analyses of the entire superfamily.

The genes and pseudogenes are given formal names in the *str* and *srj* series according to their location in this phylogenetic tree (gene fragments encoding less than half of the typical amino acid length of these receptors were excluded from the pseudogene set). Genes C42D4.5, C50C10.7 and M7.13 in the large (DN)P subfamily have already been named *str-1*, *-2* and *-3* (Troemel *et al.*, 1997; Dwyer *et al.*, 1998; Peckol *et al.*, 1999) and the *odr-10* name is also retained for gene C53B7.5 in the small *odr-10* subfamily (Sengupta *et al.*, 1996).

Two new subfamilies are recognized in the *str* family. The DA subfamily consists of two divergent annotated genes on cosmid B0213 identified by the PSI-BLASTP search. The D(SP) subfamily, consisting of many newly identified genes in two overlapping yeast artificial chromosomes (YACs; Y9C9 and Y17G9) was previously a small basal lineage of the large (DN)P subfamily. In addition, two highly divergent proteins (STR-4/W06D12.4 and STR-5/Y40H7A.1) were identified in the PSI-BLASTP search that are distantly related to the DP subfamily in both the NJ and MP trees, but there was no bootstrap support for this relationship so they have not been assigned to a subfamily. Similarly, STR-94/F07C3.8 is highly divergent, does not cluster confidently with the EP subfamily and has a different placement in MP trees [at the base of the (DE)P subfamily] both in Robertson (1998) and the present analysis, so it has not been assigned to a subfamily. Otherwise all the subfamilies are as recognized in Robertson (1998); all were supported by bootstrapping at the 70% level and most at the 95% level, as they were with MP on the original dataset. Also as before, there was little bootstrap support for the relationships of the subfamilies within the *str* family. Within the larger *str* subfamilies, the basal architecture was also seldom well supported by

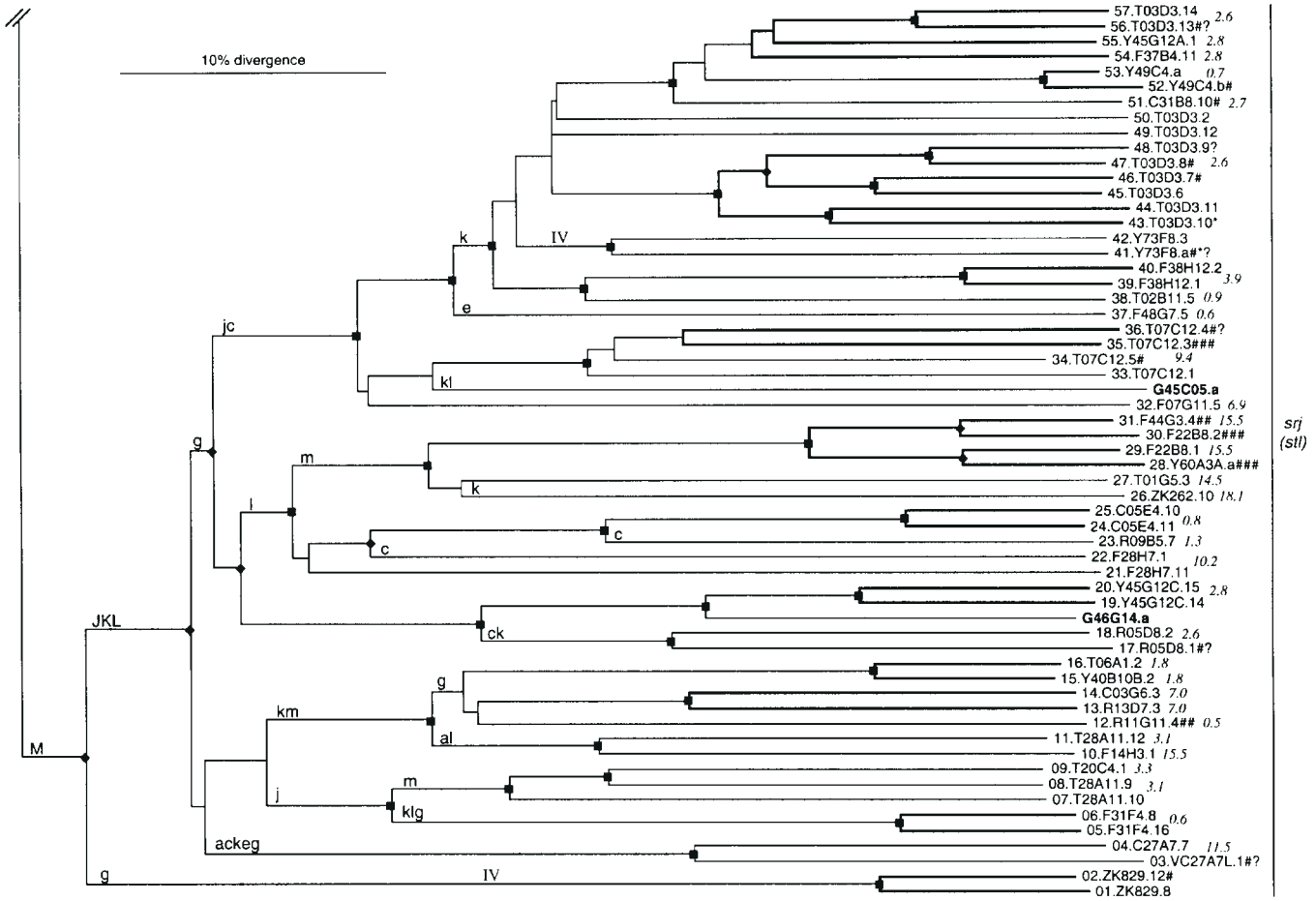

Figure 1 Phylogenetic tree relating members of the *str* and *srj* families of chemoreceptors. Subfamilies are indicated on the right. Bootstrap support of >95% is indicated by a square on the relevant node, with a diamond indicating bootstrap support >70%. Inferred jumps of genes from chromosome V to another chromosome are indicated by roman numerals above the middle of the relevant branch. The approximate location of chromosome V genes within that chromosome is indicated in Mbp from the left end after the gene name. Lower case letters above the base of the relevant branch indicate inferred intron loss while upper case letters indicate inferred intron gain. Double thickness lines connect genes that were inferred to have arisen by gene duplication since the *C. elegans*/*C. briggsae* species split. *C. briggsae* genes are indicated in bold type, start with the letter G and are not numbered. *C. elegans* genes are assigned gene numbers in *str* and *srj* series. Pseudogene status is indicated by symbols after each gene name: #, frameshift or large insertion or deletion; *, in-frame stop codon; ?, loss of start codon or questionable intron boundary.

bootstrapping, and commonly was somewhat different in the most parsimonious trees identified.

Chromosomal location

As in the nuclear receptor superfamily (Sluder *et al.*, 1999) and the *srh* family (Robertson, 2000), the vast majority, 267 (84%) of these 320 genes and pseudogenes are located on the large chromosome V, with just 40 on IV, seven on X, three on II, two on III and one on I. Mapping of these gene locations on the phylogenetic tree allowed inference of inter-chromosomal gene movements; these are indicated by roman numerals above the middle of the appropriate tree branch in Figure 1. Even the canonical *odr-10*/C53B7.3 gene is one of these, a recent gene duplication from *str-112*/F10D2.4 that moved to the X chromosome.

Three additional aspects of this analysis are remarkable. First, 20 of the chromosome IV genes are clustered on overlapping YACs and cosmids and comprise most of the newly recognized D(SP) subfamily (*str-152* to *-175*). These all appear to have resulted from duplication of an ancestral gene that moved from chromosome V a long time ago and formed this subfamily. Remarkably, four members of the subfamily now reside on chromosome V, but each involved a separate movement back to chromosome V, both on phylogenetic grounds [their independent clustering within the D(SP) subfamily is strongly supported by bootstrapping] and because they are widely disparately located on chromosome V.

Second, one other movement to chromosome IV in subfamily D(SA) has led to the formation of one gene and three pseudogenes on cosmid C34D4 (*str-48* to *-51*); however, just five other movements have led to gene duplications at the

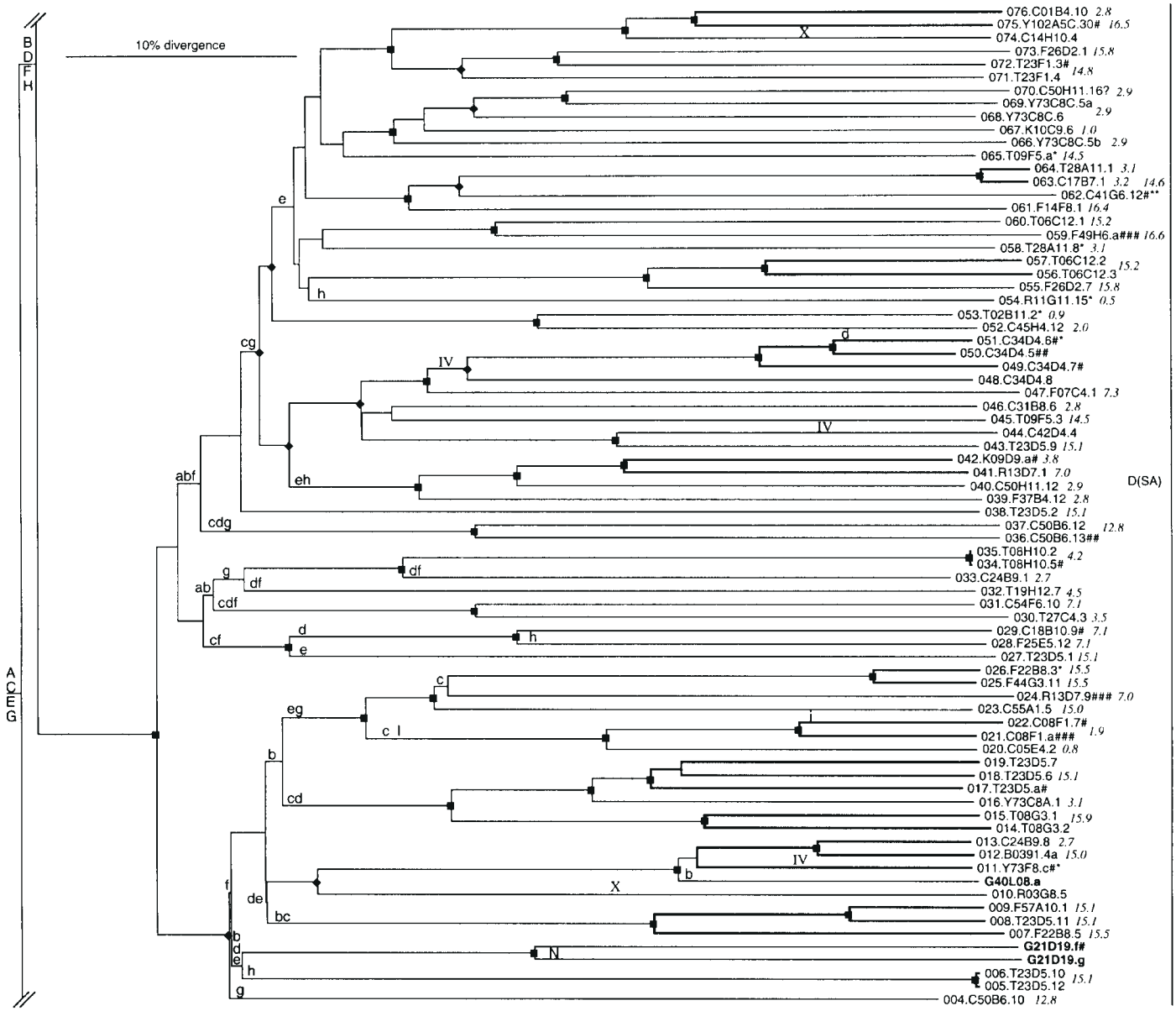

Figure 1(B)

new chromosomal location, all pairs, and four of them include a pseudogene. The remaining non-chromosome V genes and pseudogenes are all singletons. Altogether 14 movements to chromosome IV, six to X, three to II, two to III and one to I were inferred, making a total of 30 movements between chromosomes in the two families [including the return of the four D(SP) subfamily genes to chromosome V]. The independence of all of these gene movements was strongly supported by bootstrapping in the NJ/ME analysis and they were also clearly separate in the MP analysis, commonly occurring in different subfamilies or divergent gene lineages within the larger subfamilies.

Third, a further enigmatic feature of these gene movements is that three independent movements to chromosome IV have resulted in genes on cosmid C42D4 [*str-44* in the D(SA) subfamily and *str-1* and *str-249* in the (DN)P subfamily], while another led to formation of the tandem cluster on the overlapping cosmid C34D4 mentioned above in the D(SA) subfamily. The independence of these events in the phylogenetic tree is convincing; they are not adjacent genes on these two cosmids, but it seems remarkable that they are so closely linked within 64 kbp on a chromosome of 10.7 Mbp. This is the only obvious instance of a possible 'hotspot' for gene insertion on to a new chromosome: the other chromosome IV genes are fairly evenly distributed across the chromosome, as are those on chromosomes X, III and II.

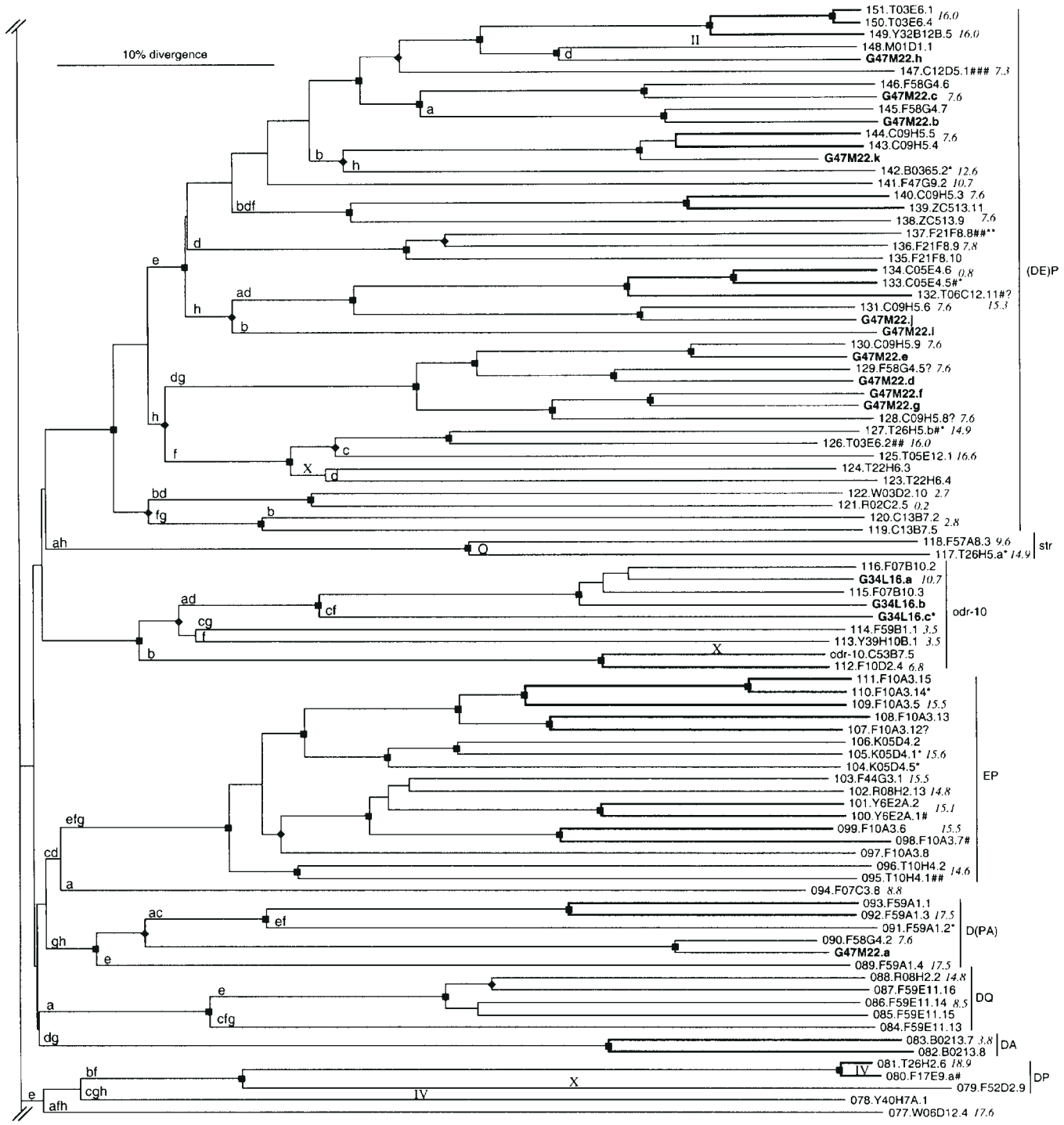

Figure 1(C)

Gene movement within chromosome V

Gene movement within a chromosome was obvious in the original dataset (Robertson, 1998), but there is no simple way to quantify gene movement within a chromosome, in part because complete contigs are not yet available. Nevertheless, it is possible to undertake a preliminary analysis using the provisional chromosomal locations provided by the Entrez genome server at NCBI (http://www.ncbi.nlm.nih.gov/pmgifs/genomes/6239.html). The approximate location of all chromosome V genes in Mbp from the left end of this 20.6 Mbp chromosome is given after the gene name in Figure 1 (some clones/genes are

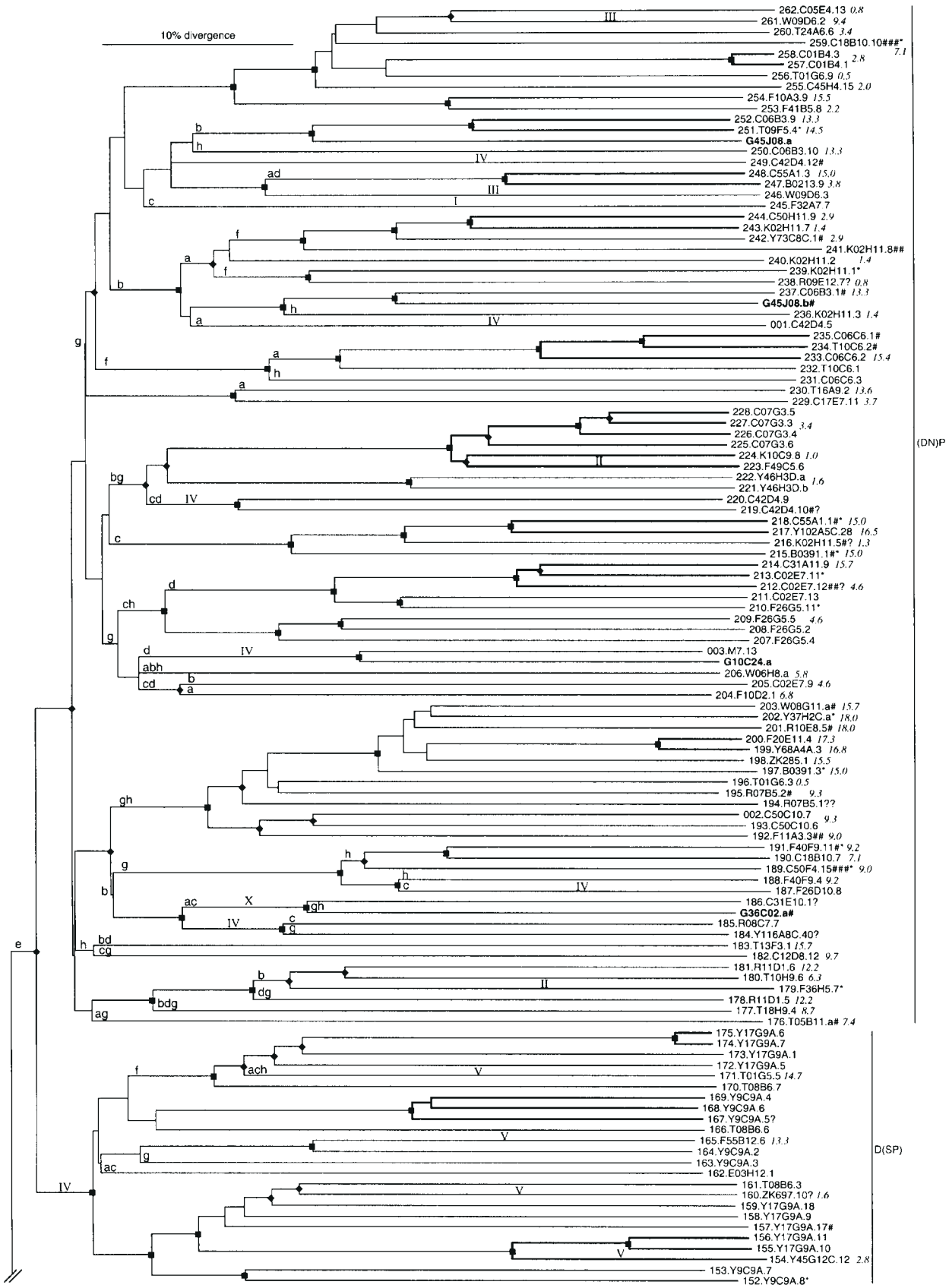
10% divergence
262.C05E4.13 0.8
261.W09D6.2 9.4
260.T24A6.6 3.4
259.C18B10.10###* 7.1
258.C01B4.3 2.8
257.C01B4.1
256.T01G6.9 0.5
255.C45H4.15 2.0
254.F10A3.9 15.5
253.F41B5.8 2.2
252.C06B3.9 13.3
251.T09F5.4* 14.5
G45J08.a
250.C06B3.10 13.3
249.C42D4.12#
248.C55A1.3 15.0
247.B0213.9 3.8
246.W09D6.3
245.F32A7.7
244.C50H11.9 2.9
243.K02H11.7 1.4
242.Y73C8C.1# 2.9
241.K02H11.8##
240.K02H11.2 1.4
239.K02H11.1*
238.R09E12.7? 0.8
237.C06B3.1# 13.3
G45J08.b#
236.K02H11.3 1.4
001.C42D4.5
235.C06C6.1#
234.T10C6.2#
233.C06C6.2 15.4
232.T10C6.1
231.C06C6.3
230.T16A9.2 13.6
229.C17E7.11 3.7
228.C07G3.5
227.C07G3.3 3.4
226.C07G3.4
225.C07G3.6
224.K10C9.8 1.0
223.F49C5.6
222.Y46H3D.a 1.6
221.Y46H3D.b
220.C42D4.9
219.C42D4.10#?
218.C55A1.1#* 15.0
217.Y102A5C.28 16.5
216.K02H11.5#? 1.3
215.B0391.1#* 15.0
214.C31A11.9 15.7
213.C02E7.11*
212.C02E7.12##? 4.6
211.C02E7.13
210.F26G5.11*
209.F26G5.5 4.6
208.F26G5.2
207.F26G5.4
003.M7.13
G10C24.a
206.W06H8.a 5.8
205.C02E7.9 4.6
204.F10D2.1 6.8
203.W08G11.a# 15.7
202.Y37H2C.a* 18.0
201.R10E8.5# 18.0
200.F20E11.4 17.3
199.Y68A4A.3 16.8
198.ZK285.1 15.5
197.B0391.3* 15.0
196.T01G6.3 0.5
195.R07B5.2# 9.3
194.R07B5.1??
002.C50C10.7 9.3
193.C50C10.6
192.F11A3.3## 9.0
191.F40F9.11#* 9.2
190.C18B10.7 7.1
189.C50F4.15###* 9.0
188.F40F9.4 9.2
187.F26D10.8
186.C31E10.1?
G36C02.a#
185.R08C7.7
184.Y116A8C.40?
183.T13F3.1 15.7
182.C12D8.12 9.7
181.R11D1.6 12.2
180.T10H9.6 6.3
179.F36H5.7*
178.R11D1.5 12.2
177.T18H9.4 8.7
176.T05B11.a# 7.4
175.Y17G9A.6
174.Y17G9A.7
173.Y17G9A.1
172.Y17G9A.5
171.T01G5.5 14.7
170.T08B6.7
169.Y9C9A.4
168.Y9C9A.6
167.Y9C9A.5?
166.T08B6.6
165.F55B12.6 13.3
164.Y9C9A.2
163.Y9C9A.3
162.E03H12.1
161.T08B6.3
160.ZK697.10? 1.6
159.Y17G9A.18
158.Y17G9A.9
157.Y17G9A.17#
156.Y17G9A.11
155.Y17G9A.10
154.Y45G12C.12 2.8
153.Y9C9A.7
152.Y9C9A.8*
(DN)P
D(SP)

Figure 1(D)

not yet entered into this database, but their position was ascertained through overlap with those that are). The genes are fairly evenly distributed along the length of this chromosome and simple inspection of these locations shows that genes must have moved around on this chromosome frequently. For example, within the well-resolved *srj* family, at least 22 movements around chromosome V can be inferred from the tree. Even the terminal lineage of several genes on cosmid T03D3 and neighboring cosmids (C31B8, F37B4 and Y45G12A) involves several non-contiguous genes. Similarly high rates of movement within chromosome V are revealed by well-supported regions of the major *str* subfamilies. On the other hand, some lineages such as the entire EP subfamily have remained in the same region of chromosome V, even if not all the genes remain contiguous. As noted above, the same is true for most of the D(SP) subfamily on chromosome IV.

Intron evolution

Mapping of intron losses from the ancestral condition of eight introns within the *str* family on the phylogenetic tree was not as simple as before (Robertson, 1998), because the independence of many losses inferred on basal branches within the large subfamilies is seldom supported by bootstrapping (Figure 1). Details of relationships within the subfamilies were somewhat different in this NJ/ME tree from the MP trees obtained earlier and now, and as before there are regions of the large D(SA) and (DN)P subfamilies where some relatively minor rearrangements would reduce inferred numbers of intron losses. On the other hand, there is sometimes underestimation of likely intron losses, for example in that gene *str-94*/F07C3.8 probably lost introns c and d independently of the EP subfamily ancestor. In the *str* family, 177 losses were inferred in Figure 1, with ≥150 of these being convincingly independent; the independence of 122 was supported by ≥70% bootstrapping (no matter how the trees are rearranged, more losses than those whose independence was supported by bootstrapping must be inferred to explain the current distribution of introns). In the *srj* family, 28 of the 30 inferred losses in Figure 1 were convincingly independent and supported by bootstrapping, yielding a total of ≥178 intron losses for the two families.

Only one intron gain was previously noted within the *str* family (Robertson, 1998), but two more were recognized here. First, a new homolog from *C. briggsae*, G21D19.g near the base of the D(SA) subfamily, has a new intron called intron n between the positions of introns e and f. The *C. elegans* ortholog has been lost, so it is unclear if this is a unique addition within *C. briggsae*. Second, pseudogene *str-117*/T26H5.a* in the *str* subfamily appears to have a novel intron near the C-terminus, beyond the position of intron h, that was not recognized in the original reconstruction (it is named o). In addition, a minor adjustment to the timing of acquisition of introns j, k and l near the base of the *srj* family has been made; the most parsimonious mapping suggests that they were acquired after the first lineage of the family diverged (Figure 1).

C. briggsae homologs

Twenty-three homologs of the *str* and *srj* family members have been identified among the 8 Mbp or 8% of the *C. briggsae* genome available; their relationships are shown in Figure 1. In my previous analysis, Table 1 in Robertson (Robertson, 1998), the levels of similarity of orthologs between the two species within the *str* family varied rather widely, from 68–87% amino acid identity for the 11 genes on *C. briggsae* cosmid G47M22 to 57–61% for three other orthologous pairs where at least one was a pseudogene. In contrast, orthologous pairs in the *srh* family revealed less variability in levels of amino acid identity, averaging 68% (range 56–77%) (Robertson, 2000). Five newly recognized orthologs in the *str* and *srj* families (Table 1) have divergence levels more in line with these (64–78%), and an additional eight orthologous gene pairs in the *srd* and two smaller families have identities ranging from 53 to 74% (H. M. Robertson, unpublished). I have therefore chosen 70% amino acid identity as an average value for the divergence of *C. elegans*/*C. briggsae* chemoreceptor orthologous pairs. Inspection of Figure 1 shows that many closely related pairs, some triplets and even two quadruples and sextuples of chemoreceptors in the *str* and *srj* families within *C. elegans* are more closely related to each other than this (branches highlighted in bold). Altogether 72 gene duplications were inferred to have occurred within *C. elegans* since the species split, forming 22.5% of the two families.

These orthologous comparisons also show that these genes are under considerable selective pressure, because the frequency of synonymous or silent changes (K_s) (base changes that do not change the encoded amino acids) was always far greater than that for non-synonymous or replacement changes (K_a). Nevertheless, four of the *C. briggsae* genes have no orthologs in *C. elegans*, so these have apparently been lost by large deletions. The orthologs of G21D19.f# and g were apparently lost as part of a large deletion that also removed the orthologs of G21D19b, c, d and e in the *srh* family (Robertson, 2000). As was true for the *srh* family (Robertson, 2000), large deletions were common in the many pseudogenes in these two families (nine of them longer than 100 bp) and 13 long terminal truncations of at least 100 bp reducing genes to fragments were observed.

Discussion

Complete description of these two large families of chemoreceptors allows consideration of several new aspects of their molecular evolution not systematically addressed previously (Robertson, 1998). Like the large *srh* family (Robertson, 2000), there is abundant evidence for movement of genes across chromosomes. Unlike the *srh* family, where most of these movements were from chromosome V to II,

Table 1 Comparison of *C. briggsae* chemoreceptor genes with their *C. elegans* orthologs[a]

C. briggsae gene	*C. elegans* gene	Encoded aa identity (%)	$K_s \pm$ SE	$K_a \pm$ SE	K_s/K_a ratio	Introns	Synteny[b]
G46G14.a	*srj-19*/Y45G12C.14	70	3.03 ± 1.64	0.23 ± 0.02	13.2	seven shared	L and R
	srj-20/Y45G12C.15	73	2.78 ± 1.16	0.22 ± 0.02	12.7	seven shared	L and R
G45C05.a	no ortholog						
G21D19.f# and g	no orthologs						
G40L08.a[c]	*str-11*/Y73F8.c#*	73	1.02 ± 0.15	0.32 ± 0.03	3.2	four shared[d]	none
	str-12/B0391.4a	76	3.32 ± 2.38	0.20 ± 0.02	16.6	four shared[d]	none
	str-13/C24B9.8	75	2.02 ± 0.44	0.20 ± 0.02	10.1	four shared[d]	none
G34L16.a	*str-116*/F07B10.2	78	3.40 ± 2.65	0.18 ± 0.02	18.9	six shared	L
G34L16.b	*str-115*/F07B10.3	72	NC[e]	0.22 ± 0.02	–	six shared	L
G34L16.c*	no ortholog						
G10C24.a	*str-3*/M7.13	64	1.50 ± 0.23	0.30 ± 0.02	5.0	five shared	L and R

[a]*C. briggsae str* family genes on cosmids G47M22, G45J08 and G36C02 are not included [see Table 1 in Robertson (Robertson, 1998)]; however, one correction is required, in that intron h in *C. briggsae* gene G47M22.b has now been recognized.

[b]Synteny of flanking genes or regions to the left (L) or right (R) of the inferred orthologs is indicated.

[c]G40L08.a does not exhibit synteny with flanking genes of any of the three closely related *C. elegans* genes (all three have moved from the region of chromosome V that is syntenic with G40L08); nevertheless, these three genes appear to have originated within *C. elegans* from a G40L08.a ortholog (Figure 1).

[d]G40L08.a lost intron b (see Figure 1).

[e]KsKaCalc could not estimate this value because synonymous changes are saturated.

almost 50% of these inter-chromosome movements in the *str* and *srj* families have been from chromosome V to IV. As was true for the *srh* family, most of these movements involve a single gene, implying that they involve relatively short stretches of DNA and that most such newly located genes do not lead to formation of new gene lineages but, rather, are probably eventually removed by deletion. Just one movement has led to a new gene lineage, the formation of the D(SP) subfamily on chromosome IV. That chromosome V somehow provides a more hospitable environment for these chemoreceptors is also suggested by four independent movements of D(SP) subfamily genes back to chromosome V.

Preliminary analysis of chromosomal location within chromosome V shows that these genes have moved around frequently within the chromosome, leading to an even distribution of gene numbers along its length. Presumably most of these transpositions involve a similar mechanism to movement across chromosomes. Inversions might also be involved in movements within chromosome V, but their frequency and importance is unknown in nematodes. Extensive simulation studies to develop null models would be required to determine whether movement within chromosomes is more common than between them. It does appear, however, that movements to other chromosomes usually lead to loss of the gene, rather than the formation of new gene lineages so frequently seen within chromosome V.

As before (Robertson, 1998, 2000), comparisons with *C. briggsae* provide information about the patterns of gene evolution in these chemoreceptor families and genome dynamics in general. Almost a quarter (22.5%) of *str* and *srj* family genes and pseudogenes in *C. elegans* appear to have been newly formed by gene duplications since the species split. This process is clearly not limited to these and other chemoreceptors, because 60% of the *C. elegans* genome consists of gene families that have been found in nematodes but not in yeast, *Drosophila* or mammals (*C. elegans* Sequencing Consortium, 1998; *C. elegans* Genome Consortium, 1999; Rubin *et al.*, 2000). As expected with the finishing of the *C. elegans* genome sequencing project, orthologs were found for several of the *C. briggsae* genes identified before (Robertson, 1998), as well as some new ones. Nevertheless, four *C. briggsae* members of these two families do not have orthologs in *C. elegans*; they have apparently been lost by three large deletions. As was true for the *srh* family (Robertson, 2000), examination of pseudogenes and gene fragments also revealed the common occurrence of large deletions in the *C. elegans* genome. Presumably it is these kinds of event that remove the many newly forming pseudogenes and maintain the small size of this nematode genome in the face of rampant gene duplication.

Acknowledgements

I thank the Genome Sequencing Centers at Washington University, St Louis, USA and the Sanger Centre, Cambridge, UK, for communication of DNA sequence data prior to publication, John Spieth and Richard Durbin for their assistance in annotating these nematode genes, and John Spieth and Chris Michelsen for assistance in analyzing chromosomal locations. This work was

supported by NSF grant IBN 96-04095. The amino acid alignment file has been submitted to the EMBL alignment database (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/) with accession number ds42124. Gene and protein alignments are also available from the author at hughrobe@uiuc.edu.

References

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

Bargmann, C.I. (1998) *Neurobiology of the* Caenorhabditis elegans *genome.* Science, 282, 2028–2033.

***C. elegans* Genome Consortium** (1999) *How the worm was won.* Trends Genet., 15, 51–58.

***C. elegans* Sequencing Consortium** (1998) *Genome sequence of the nematode* C. elegans*: a platform for investigating biology.* Science, 282, 2012–2018.

Dwyer, N.D, Troemel, E.R., Sengupta, P. and **Bargmann, C.I.** (1998) *Odorant receptor localization to olfactory cilia is mediated by ODR-4, a novel membrane-associated protein*. Cell, 93, 455–466.

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G. and **Gibson, T.J.** (1998) *Multiple sequence alignment with Clustal X.* Trends Biochem. Sci., 23, 403–405.

Nei, M. and **Gojobori, T.** (1986) *Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions*. Mol. Biol. Evol., 3, 418–426.

Peckol, E.L, Zallen, J.A., Yarrow, J.C. and **Bargmann, C.I.** (1999) *Sensory activity affects sensory axon development in* C. elegans. Development, 126, 1891–1902.

Robertson, H.M. (1998) *Two large families of chemoreceptor genes in the nematodes* Caenorhabditis elegans *and* Caenorhabditis briggsae *reveal extensive gene duplication, diversification, movement, and intron loss.* Genome Res., 8, 449–463.

Robertson, H.M. (2000) *The large srh family of chemoreceptor genes in* Caenorhabditis *nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses*. Genome Res., 10, 192–203.

Rubin, G.M. and **54 coauthors** (2000) *Comparative genomics of the eukaryotes*. Science, 287, 2204–2215.

Sengupta, P., Chou, J.H. and **Bargmann, C.I.** (1996) odr-10 *encodes a seven transmembrane domain olfactory receptor required for responses to the odorant diacetyl*. Cell, 84, 899–909.

Sluder, A.E., Mathews, S.W., Hough, D., Yin, V.P. and **Maina, C.V.** (1999) *The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes.* Genome Res., 9, 103–120.

Swofford, D.L. (1998) PAUP*: *Phylogenetic Analysis Using Parsimony and Other Methods, Version 4*. Sinauer Press, New York.

Troemel, E.R., Chou, J.H., Dwyer, N.D., Colbert, H.A. and **Bargmann, C.I.** (1995) *Divergent seven transmembrane receptors are candidate chemosensory receptors in* C. elegans. Cell, 83, 207–218.

Troemel, E.R., Kimmel, B.E. and **Bargmann, C.I.** (1997) *Reprogramming chemotaxis responses: sensory neurons define olfactory preferences in* C. elegans. Cell, 91, 161–169.

Accepted September 21, 2000