JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1999

Gene Duplication and Gene Conversion in the *Caenorhabditis elegans* Genome

Colin Semple, Kenneth H. Wolfe

Genetics Department, University of Dublin, Trinity College, Dublin 2, Ireland

Received: 30 July 1998 / Accepted: 12 October 1998

Abstract. A comprehensive analysis of duplication and gene conversion for 7394 *Caenorhabditis elegans* genes (about half the expected total for the genome) is presented. Of the genes examined, 40% are involved in duplicated gene pairs. Intrachromosomal or *cis* gene duplications occur approximately two times more often than expected. In general the closer the members of duplicated gene pairs are, the more likely it is that gene orientation is conserved. Gene conversion events are detectable between only 2% of the duplicated pairs. Even given the excesses of *cis* duplications, there is an excess of gene conversion events between *cis* duplicated pairs on every chromosome except the X chromosome. The relative rates of *cis* and *trans* gene conversion and the negative correlation between conversion frequency and DNA sequence divergence for unconverted regions of converted pairs are consistent with previous experimental studies in yeast. Three recent, regional duplications, each spanning three genes are described. All three have already undergone substantial deletions spanning hundreds of base pairs. The relative rates of duplication and deletion may contribute to the compactness of the *C. elegans* genome.

Key words: *Caenorhabditis elegans* — Duplication — Gene conversion

Introduction

As more genomic sequence has become available gene duplication has become a common observation across a diverse range of organisms spanning bacteria (Labedan and Riley 1995; Brenner et al. 1995; Coissac et al. 1997), yeast (Wolfe and Shields 1997), plants (Ahn and Tanksley 1993; Frugoli et al. 1998), and mammals (Holland et al. 1994). The *Caenorhabditis elegans* genome sequencing project, although unfinished, has already revealed that many genes are found as members of families, sharing similarity with other members (Waterston and Sulston 1995; Sonnhammer and Durbin 1997; Robertson 1998). Closely related *C. elegans* genes, with up to 98% DNA sequence identity, have been described (Wilson et al. 1994; Waterston et al. 1997). Some duplicates occur close together in tandem arrays (Waterston and Sulston 1995).

Once a gene duplication has generated two ‘‘daughter’’ sequences, they can either diverge in sequence, sometimes aquiring different functions, or undergo concerted evolution. Gene conversion and unequal crossing over are the two most important mechanisms generating concerted evolution (reviewed by Li 1997). Gene conversion, which can be defined as the nonreciprocal transfer of information between two sequences, is involved in the homogenization of small tracts of DNA, usually between several and several hundred base pairs (Petes et al. 1991). Gene conversion leaves nonidentical flanking DNA on either side of the event. The homogenization of larger arrays of repetitive DNA is generally believed to involve unequal crossing-over (Szostak and Wu 1980), although more recent work in yeast (Gangloff et al.

Correspondence to: C. Semple, MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

Table 1. Observed duplications and numbers of genes involved

Chromosome	Approximate physical size (Mb)	Number of mapped genes	Proportion of genes that are duplicated	Proportion of all duplications involving this chromosome (number)[a]
I	15.40	236	0.31	0.02 (376)
II	18.27	1,488	0.39	0.22 (3,404)
III	13.46	1,385	0.32	0.13 (2,032)
IV	18.08	1,161	0.45	0.19 (3,028)
V	24.23	1,243	0.51	0.23 (3,573)
X	20.96	1,881	0.36	0.21 (3,245)
Total	110.40	7,394	0.40	1 (15,658)

[a] *Cis* (intrachromosomal) duplications are counted as two events involving the chromosome on which they occur.

1996) and lizards (Hillis et al. 1991) has suggested that gene conversion is the dominant mechanism here too. The extent of gene conversion is therefore important in understanding the evolution of multigene families.

There are many well-documented examples of gene conversion in multigene families including the yeast tandem rDNA array (Gangloff et al. 1994, 1996), the primate visual pigment genes (Shyue et al. 1994; Zhou and Li 1996) and the silkmoth chorion genes (Hibner et al. 1991). In *C. elegans* gene conversion has been shown to be involved in the evolution of heat shock protein genes (Russnak and Candido 1985) and collagen genes (Park and Kramer 1990).

Gene conversion events are often inferred simply by visual inspection, though statistical tests for their detection are available (Stephens 1985; Sawyer 1989; Hein 1993; Jakobsen et al. 1997). Sawyer's is the more general statistical test and, for a sample of three or more coding sequences, considers only the polymorphic synonomous sites which vary between them. Thus structure arising from recombination events can be identified, as opposed to that which arises from regions with different mutation rates (Sawyer 1989; Maynard Smith 1992).

The aim of this study was to broadly survey the extent of gene duplication across the sequenced parts of the *C. elegans* genome and examine the role of gene conversion in the maintenance of sequence similarity among duplicated members of gene families.

Materials and Methods

Data

The source of data was version 4.3 of ACeDB running data release WS2.4-17 [a *C. elegans* database compiled by Durbin and Thierry-Mieg (1996)]. The data analyzed comprised 1283 sequenced cosmids (totaling about 70 Mb of the estimated 100 Mb genome), version 12 of Wormpep [the database of 12178 predicted proteins, which is approximately 87% of the expected total (Waterston et al. 1997), and physical mapping data (Coulson et al. 1986).

Determination of Genomic Map Coordinates

The 1283 sequenced cosmids were linked together where possible according to their GenBank annotation, which details their 5′ and 3′ overlaps with other sequenced cosmids. Where gaps between cosmids existed their size was estimated from the ACeDB physical map. The outcome was a set of sequences corresponding to the central, gene-rich, sequenced sections of the six *C. elegans* chromosomes (Waterston et al. 1997). Each sequenced cosmid was assigned approximate (due to the presence of unsequenced gaps) beginning and end coordinates based on its position on the chromosome, from which chromosomal positions of the genes were calculated. Of the 12,178 proteins in Wormpep12, the DNA sequences coding for 7394 were assigned map coordinates, based upon their presence within mapped and sequenced cosmids. Some proteins were omitted, as only one protein product (the first variant listed in Wormpep12) was taken to represent each alternatively spliced gene. Certain cosmids could not be assigned positions because of the lack of sequenced neighbouring cosmids or physical map estimates. Most of the analysis in this study is based on the subset of 7394 proteins with known gene locations. The densities (number of genes per kilobase) of genes assigned positions for each autosome were similar (I, 0.238; II, 0.209; III, 0.205; IV, 0.212; V, 0.23) and suggest that the low number of mapped genes on chromosome I (Table 1) is due simply to the presence of larger unsequenced regions than on other chromosomes. In common with other studies (Waterston et al. 1997) the gene density on the X chromosome (0.153 gene per kb) was found to be lower than that on the autosomes.

Identification of Duplicated Genes

Each of the 12,178 proteins was searched against all others using BLASTP (Altschul et al. 1990) with the BLOSUM62 substitution matrix and the SEG filter, which masks regions of low compositional complexity (Wootton and Federhen 1993). All protein pairs with BLASTP scores greater than 150 (corresponding to a P value of $<10^{-13}$) were defined as the products of putatively duplicated genes. Using an absolute threshold for BLASTP similarity instead of a p value means that longer genes are more likely to exceed the threshold and be classed as duplicated. However, examination of the lengths of duplicated versus nonduplicated genes showed no significant difference.

Detection of Gene Conversion Events

The coding sequences of each duplicate gene pair were aligned using CLUSTALW version 1.4 (Thompson et al. 1994) with default settings. The alignments were then analyzed for evidence of gene conversion using Sawyer's (1989) method, performed by his program VTDIST3 (http://lado.wustl.edu/~sawyer/mbprogs/). This method involves the comparison of two sequences to an outgroup. Because many of the duplicated genes in *C. elegans* lacked an outgroup, it was necessary to use two modifications of Sawyer's method. The basis of the method is the identification of silent sites (synonymous codon positions) at which two DNA sequences agree (but differ from the outgroup) and the segmentation of the sequences according to contiguous stretches of these sites. Gene conversion increases the lengths of the stretches. The significance of these lengths is then estimated by comparison with values obtained from 10,000 artificial data sets constructed by randomly permuting the silent polymorphic sites. However, with a data set of only two sequences, each polymorphism distinguishes the two sequences but provides no other information. As a result all permuted scores are the same as the observed scores and no measure of significance can be made.

The first modification to Sawyer's method (proposed by S. Sawyer, personal communication) compensates for the lack of an outgroup by introducing an artificial sequence into the data set. This artificial sequence is distinct from the other two at all the silent sites at which they are monomorphic. As a result of the modification all silent sites in the two real sequences are treated as potential polymorphisms and the program discovers pairwise fragments between them. This modification means the method (in common with any comparison between only two sequences) cannot control for regions of similarity produced by selection or mutational ''cold spots.''

The second modification to Sawyer's (1989) method (proposed here) was the use of chi-square testing to verify the gene conversion breakpoints identified. This is related to Maynard Smith's (1992) method. Chi-square tests were performed for each event on the observed and expected (on the basis of the proportions outside the putatively gene converted region) numbers of identical and differing bases. All events were verified by visual inspection of the alignments from which they were generated, and the threshold of $\chi^2 \geqq 55$ was chosen arbitrarily on the basis that putative events with lower scores were unconvincing. The chi-square test requires substantial lengths of both converted and unconverted sequence within a gene so that this method is biased against both very long and very short gene conversions and the resulting data set is a conservative one.

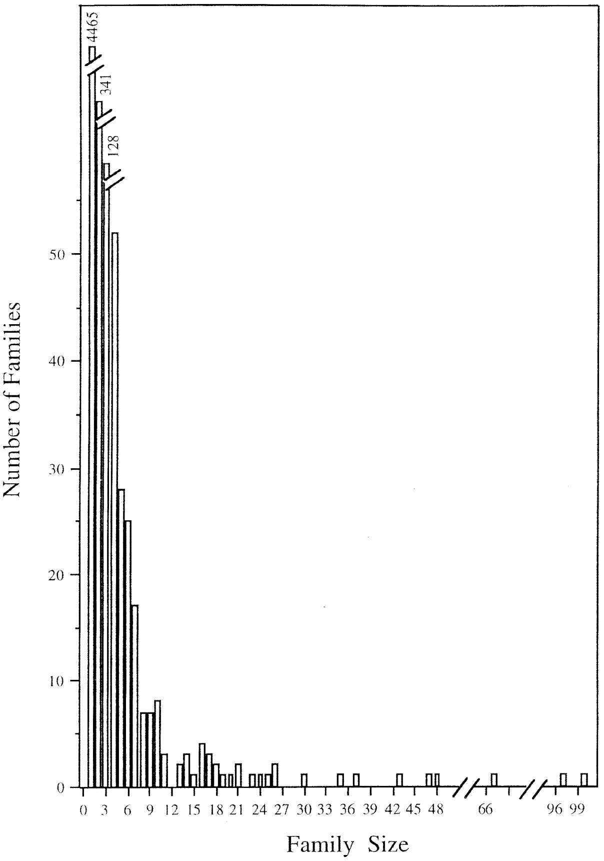

Fig. 1. Sizes of multigene families (mean number of members = 7; SD = 10). Multigene families were defined on a ''single-link'' basis. For example, if protein A hit proteins B and C with BLASTP scores of >150, then A, B, and C were included in the same family regardless of the score for the comparison of B and C.

Results

Duplications

Using a SEG-filtered BLASTP score of 150 as a threshold for sequence similarity, 7829 putatively duplicated gene pairs (i.e., significant hits between pairs of genes) were found. These pairs had a mean DNA sequence identity of 54% (range, 35%–99%) and a mean protein sequence identity of 62% (range, 32–100%). The 7829 hits involved only 2929 of the mapped genes because groups of duplications with shared member genes were common: 314 multigene families were identified (with between 3 and 100 members), as well as 341 gene pairs and 4465 (60%) singleton genes (Fig. 1). This is a conservative estimate of the real sizes of *C. elegans* multigene families and an overly generous estimate of the number of small families since the definition of a duplication used here was reasonably stringent. In particular, our criteria are more stringent than those of Sonnhammer and Durbin (1997), who reported 84 families with between 1 and 203 members in Wormpep11 (7299 proteins) using a more sensitive method which assigned proteins to families according to the Pfam motifs they contain (Sonnhammer et al. 1998). Our method of defining the members of multigene families was designed only to allow the associations in the present data to be described, but some overlaps exist with Sonnhammer and Durbins' (1997) results. For example, several of their descriptions of apparently nematode-specific families match families identified here.

Location of Duplicated Genes

The numbers of duplications varied across chromosomes, as did the proportion of genes involved (see Table 1). If duplicates were located at random in the genome, that is, if their location were dependent only on the number of mapped genes on each chromosome, then 19% of duplicate pairs would be expected to be on the same chromosome and 81% on different chromosomes (we refer to these as *cis* and *trans* duplications, respectively). The observed proportion of *cis* duplications (43%) is much higher than expected. Dot-matrix plots show that many of these *cis* duplications are closely spaced repeats

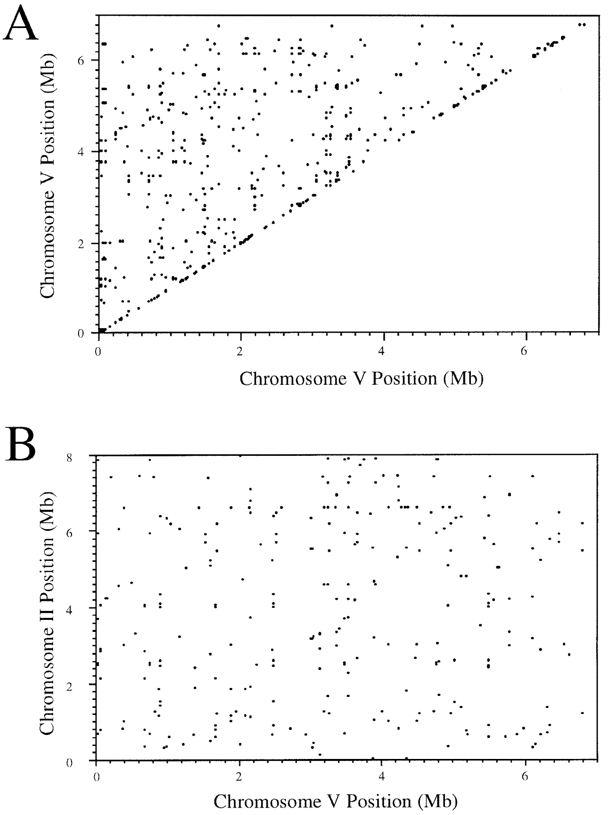

Fig. 2. Dot-matrix plots of the positions of duplicate pairs, where both members are on chromosome V **(A)** and where one member is on V and the other is on II **(B).** Self-hits have been removed in A so that the points near the main diagonal represent tandem gene duplications.

(Fig. 2). Of all the duplications, 976 (13%) were designated ''neighboring,'' where the pair is separated by five or fewer intervening genes, and of these, a subset of 438 (6%) was designated ''tandem,'' with no intervening genes between the members of the pair. Tandem duplications involved 721 genes (10% of all mapped genes). Generally the distance (kb) between members of gene pairs showed a strong bias to short distances, such that 29% of all *cis* duplications were categorized as neighboring or nearer (Fig. 3A). Correction of these data for the presence of multigene families made no difference to the shape of the distribution. The distances between *cis* pairs expressed in terms of the number of intervening genes shows the same bias, with 32% of pairs separated by <10 genes (Fig. 3B). The high proportion of closely spaced pairs means that the number of duplications along the chromosomes is variable. This may reflect the presence of ''hot spots'' for duplicative activity.

Expected frequencies of *cis* and *trans* duplications for each chromosome were calculated assuming that the probability of a duplication being *cis* or *trans* was dependent only on the number of mapped genes available on each chromosome. There was found to be a significant excess of *cis* duplications on every chromosome except I and III (Table 2). The excess for chromosome IV was large enough to obscure the fact that more than expected *trans* duplications occurred between chromosome IV and chromosome V and between chromosome IV and chromosome II (data not shown). In order to discover the reason for the departures from expectations, the data were reanalyzed with certain categories of duplication omitted.

Exclusion of all pairs less than 100 genes apart from the analysis was necessary to remove the excess of *cis* duplications from chromosomes II and IV but it was necessary to remove all those less than 500 genes apart from the data to obtain the same effect for V and X. Thus the excess of *cis* duplications was not simply a result of tandemly duplicated arrays of genes; more dispersed duplications over substantial fractions of the chromosome length are also involved. The observed values for *cis* duplications remained greater than expected even when all duplications involved in multigene families were removed from the analysis.

Orientation of Duplicated Genes

Approximately 50% of the genes on each chromosome occur on each strand. It follows that if the mechanism responsible for gene duplication were unbiased with respect to the orientation of the resulting duplicate, one would expect transcriptional orientation to be conserved 50% of the time. Across all *cis* duplications the figure was 61%. This compared with 80% of the neighboring duplications and 84% of the tandems. In general the closer the members of pairs were, the more likely it was that orientation was conserved (Fig. 3C).

Regional Duplications

Analysis of the yeast genome revealed large regional duplications of groups of neighboring genes (Wolfe and Shields, 1997). Dot-matrix analysis indicated that such duplications are not present in *C. elegans,* although there was evidence for three small, regional duplications. The regional duplications found occur within chromosomes V and X and between chromosome II and chromosome V, each involving three duplicated gene pairs with no intervening genes (Fig. 4). The regional duplication within chromosome X is inverted and the two regions are separated by 33,540 bp which spans five genes. The regional duplication within chromosome V is also inverted but the distance between the two regions is unknown, as sequencing in this area of the chromosome is unfinished, however, it exceeds 140 kb. In all three cases the DNA sequence identity between the duplicated regions is close to 100%, indicating that these duplications are recent events and at least one end point is located inside a gene. All three duplicated regions also contain deletions of between 100 and 1500 bp involving intronic, exonic, and intergenic DNA. There have been at least seven deletions involving genes and at least three indel

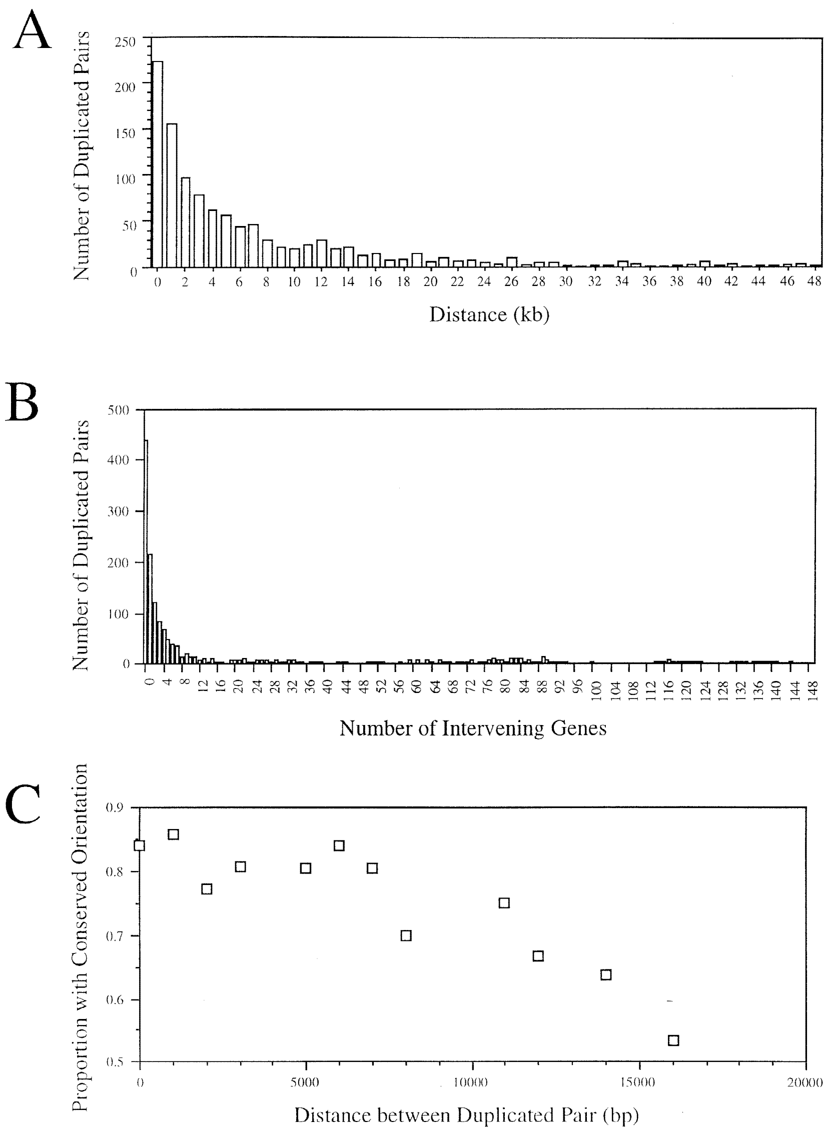

Fig. 3. **A** Distances between *cis* duplicated pairs (mean, 7783 bp; SD, 9665 bp). Distances were measured as the shortest possible distances between members of pairs. For example, the distance between the stop codon of the 5′ gene and the start codon of the 3′ gene is measured for genes lying in a head to tail configuration. **B** Number of intervening genes between *cis* duplicated pairs. **C** Distance between duplicated pairs versus proportion of pairs in which transcriptional orientation is conserved ($r = 0.88$, $p < 0.001$). Duplicate pairs were pooled into 1000-bp intervals and the proportion with conserved orientation was plotted for those containing more than five gene pairs.

Table 2. Observed numbers of *cis* and *trans* duplications, with expected numbers in parentheses (all χ^2 tests with 1 df)

	I		II		III		IV		V		X	
cis	7	(12)	846	(515)	314	(322)	562	(387)	895	(450)	723	(474)
trans	362	(357)	1712	(2043)	1404	(1396)	1904	(2079)	1783	(2228)	1799	(2048)
Total	369		2558		1718		2466		2678		2522	
χ^2	1.99		267.02		0.23		93.66		528.37		160.47	

events involving intergenic DNA. It is likely that the duplicated genes which have decreased in length (because of being incompletely duplicated or partly deleted) are pseudogenes. This has been the fate of one-third of the genes involved in the duplicated regions: W07G4.3, ZC317.6, C33D12.3, and M02F4.6 have lost exonic and intronic sequences from their 5′ ends; C33D12.2 has lost exonic and intronic sequences from its 3′ end; and ZC317.1 has undergone various small deletions shortening both exons and introns. This agrees with the observation that 29% of genes in a large *C. elegans* multigene family are pseudogenes (Robertson, 1998).

It is possible to test whether the number of putatively duplicated regions in the genome is significantly in excess of that expected by chance, that is, if duplicated genes were distributed randomly (Wolfe and Shields, 1997). The results of such a test suggested that the observed distribution of duplicated pairs was significantly different from that expected by chance. In 773 simulations the locations of genes were shuffled and then searched for regional duplications. The following criteria were used to define duplicated regions: at least three pairs of duplicated genes had to be involved, genes within each region had to be separated by no more than

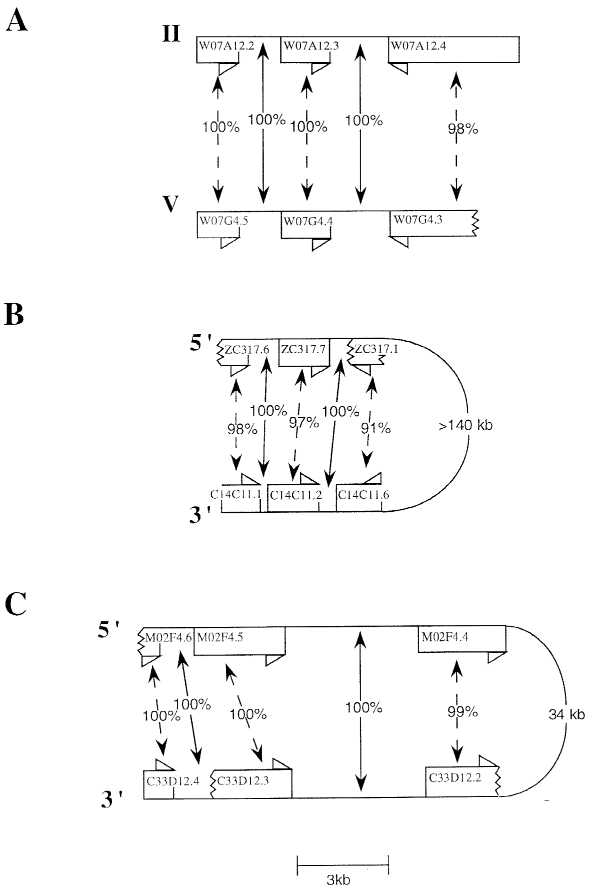

Fig. 4. The three regional duplications shown to scale: **A** the duplication between chromosome II and chromosome V; **B** the duplication within chromosome V; **C** the duplication within the X chromosome. All genes in the regions are shown as *boxes,* with *triangles* indicating their orientation. *Dashed arrows* link putative homologues and show percentage DNA identity. *Solid arrows* link putatively homologous intergenic spaces and indicate percentage DNA identity. All percentage identities refer to the sequence available for alignment allowing for gaps. *Jagged lines* indicate genes that have been truncated. The approximate distance between duplicated regions and their relative order along the chromosome are also shown for the duplications occurring within chromosomes V and X. The putative functions listed in ACeDB for these genes are as follows (other genes have no predicted functions): (A) W07A12.3 and W07G4.4, cytosolic aminopeptidases; W07G4.3, protein kinase; W07G4.5, collagen; (B) ZC317.1 and C14C11.6, helicases; (C) M02F4.5 and C33D12.3, potassium channels; M02F4.4 and C33D12.2, weak similarity to hemolysins; C33D12.3, ras-related protein.

10 intervening genes, and gene order and orientation had to be conserved between each of the two regions. Ten simulations produced one regional duplication each, and all these involved unique genes interspersed among three duplicated pairs. No simulations produced more than one regional duplication. The observed regional duplications are therefore highly unlikely to be artifacts, particularly given the conservation of intergenic DNA in addition to coding sequence.

Gene Conversion

A total of 526 gene conversion events were detected involving 143 (2%) of the 7829 duplicated pairs, using the modification of Sawyer's (1989) method described in Materials and Methods. Due to multiple events involving some genes, these conversion events involved only 212 genes. Most (78%) of the duplicated pairs showed evidence of more than one conversion event so that the mean number of events per gene pair was 3.75. This may be an artifact of Sawyer's method which identifies regions of complete sequence identity and does not allow for the possibility that nucleotide substitutions could occur subsequent to gene conversion. The method is also incapable of distinguishing recently produced chimeric genes (resulting from exon shuffling, for example) from gene conversion events. As expected, given the high proportion of genes in multigene families, the majority of duplicated pairs that underwent gene conversion (85%) were members of families. There was a significant negative correlation between multigene family size and gene conversion frequency (data not shown). This reflects increasing sequence divergence between members of larger families due to the ''single-link'' method of constructing multigene families. Melamed and Kupiec (1992) have demonstrated that for a given yeast gene, the frequency of gene conversion is proportional to the number of homologous sequences available for conversion. This effect might have been evident as a positive correlation between gene conversion frequency and the number of BLASTP hits for the genes in the present data. In fact there was a significant negative correlation (Fig. 5A) which was also seen when more stringent BLASTP score thresholds were applied (data not shown).

Gene conversion events were assumed to exceed the boundaries of a gene if they continued to the last base pair of that gene. Of all the conversion events, 97% did not exceed the boundaries of the two genes involved. The mean genomic conversion tract length for these genes was 117 bp but it varied widely, from 12 bp to 2958 bp (Fig. 6). The fact that so few events exceeded the boundaries of the genes involved may also be an artifact of the method of identification, since nucleotide substitutions occurring within the converted sequences are interpreted as the ends of converted tracts. The distribution of tract lengths is biased toward small sizes, with 65% of tracts being <80 bp long in spite of the fact that the number of events <20 bp was reduced by the failure of such short events to reach statistical significance during chi-square testing. It should be noted, however, that since gene conversion events were detected by analysis of coding regions, our method is biased toward detecting events spanning exons rather than introns. Most gene conversion events detected (85%) did not span introns, although the number of introns involved in events ranged from 0 to 9.

Since gene conversion is a form of homologous recombination, one might expect that the probability of gene conversion would be increased between pairs of genes that are more similar. To correct for the presence

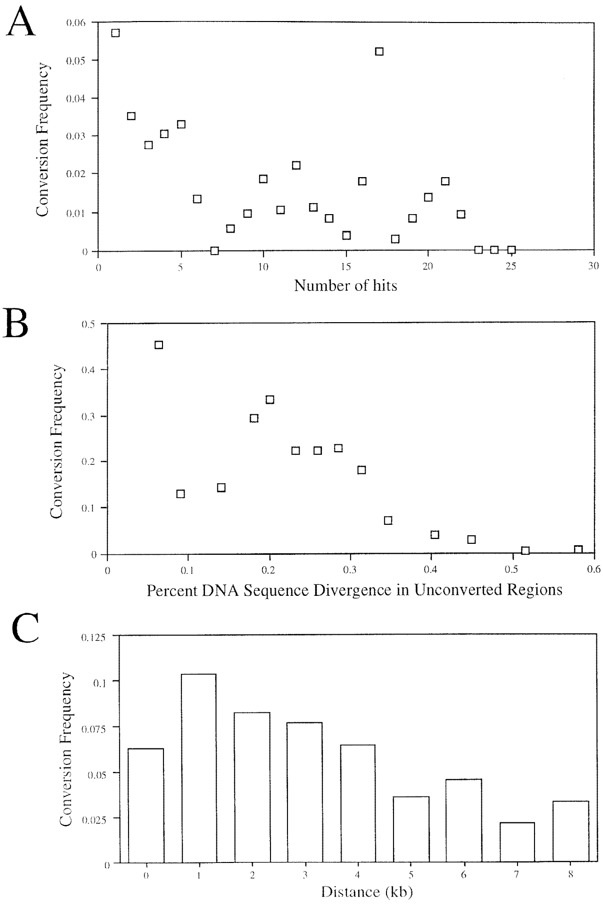

Fig. 5. Gene conversion frequency and **A** number of BLASTP hits (with genes binned on the basis of their number of hits) plotted for each bin with >100 duplicated gene pairs ($r = -0.54$, $p < 0.01$), **B** percentage DNA sequence divergence of unconverted regions pooled into bins of 10 gene pairs ($r = -0.78$, $p < 0.001$), and **C** distances between gene converted pairs in 1 kb intervals plotted for all intervals in which there were >20 duplicated gene pairs ($r = 0.82$, $p < 0.01$). In each case the gene conversion frequency was measured as the number of gene converted pairs divided by the number of duplicated gene pairs in each interval.

of multigene families, the number of conversion events with a given percentage sequence identity was divided by the total number of BLASTP hits showing this level of identity. This correction was carried out for successive bins of 10 genes with increasing sequence divergence. The percentage DNA sequence divergence between converted pairs of genes in unconverted regions was found to correlate negatively with the corrected frequency of conversion events (Fig. 5B).

Location of Gene Conversion Events

For *cis* gene conversion events it was possible to calculate the genomic distances between the pairs of genes involved. To correct for the presence of multigene families, the number of conversion events at a given distance was divided by the total number of BLASTP hits for that distance. At distances greater than 9 kb the numbers of duplications and conversion events were too small to estimate ratios reliably; for this reason, the data shown are for those events less than 9 kb. Frequency of conversion is negatively correlated with distance between gene pairs (Fig. 5C). However it should be noted that none of these conversion frequencies are normalized for the degree of divergence between converted pairs of genes.

Similar numbers of duplicated pairs involved in gene conversion events were detected on each chromosome, apart from chromosome I (Table 3). Expected probabilities of *cis* and *trans* gene conversions were calculated for each chromosome assuming that the events were dependent only on the number of *cis* and *trans* duplications involving each chromosome. It was found that even given the excesses of *cis* duplications shown already on every chromosome, there was an excess of *cis* conversion events (Table 3). This excess was statistically significant for chromosomes II, III, IV, and V. As with the data for the location of duplications, the conversion event data were reanalyzed with certain categories of duplication omitted. The *cis* conversion excess on chromosomes II, III, IV, and V was found to be attributable to gene conversion events between closely spaced duplicated gene pairs (<5 genes apart on II, <10 genes apart on III, <15 genes apart on IV, and tandem duplications on V). The higher frequency of *cis* gene conversion may be a result of lower sequence divergence between the duplicated genes involved.

Discussion

Our analysis indicates that the *C. elegans* genome contains an unexpectedly high number of intrachromosomal or *cis* gene duplications: approximately twice the number expected if duplicates were located at random. The excess of *cis* duplications involved tandemly duplicated arrays of genes as well as widely dispersed duplicated pairs separated by as much as half the total mapped chromosome length. A statistically significant excess was seen within each chromosome except the two smallest ones: chromosomes I and III. It is not possible to determine the cause of these differences between chromosomes from the present data. The small physical sizes of chromosomes I and III could be a consequence of the deficits of *cis* duplications seen on these chromosomes relative to the other four. This conclusion is consistent with the finding that genome size is a function of the extent of gene duplication (Coissac et al. 1997).

Tandem duplications involved 10% of all mapped proteins in this study and 29% of all *cis* duplications were categorized as neighboring or nearer. The excess of closely spaced pairs, even when a correction for multigene families is made, suggests that the origin of these genes was through a slippage mechanism rather than

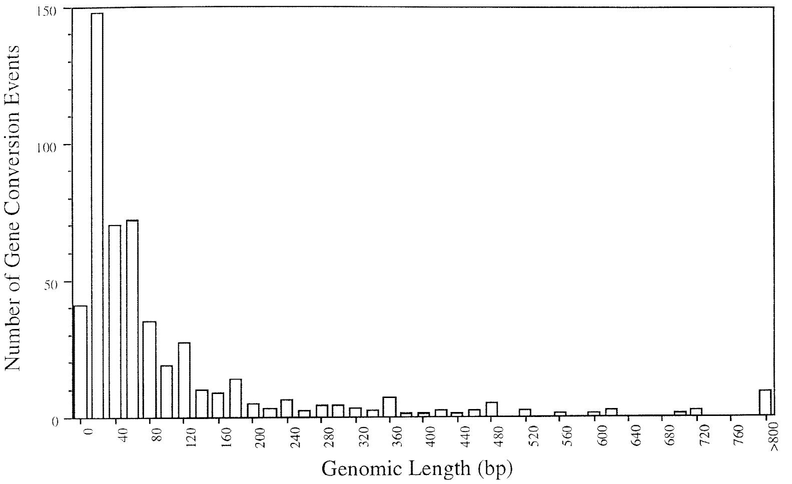

Fig. 6. Genomic length (i.e., including introns) of gene conversion tracts pooled into 20 bp intervals. Mean length = 117 bp; median = 58 bp; SD = 205.

Table 3. Observed numbers of gene conversions, with expected numbers in parentheses (all χ^2 tests with 1 df)

	I	II	III	IV	V	X
Cis	0 (0)	25 (11.58)	22 (6.03)	21 (9.57)	20 (10.36)	16 (11.76)
Trans	0 (0)	10 (23.45)	11 (26.97)	21 (32.43)	11 (20.64)	25 (29.25)
Total	0	35	33	42	31	41
χ^2	0	23.27	51.72	17.67	13.47	2.15

transposition. However, it should be noted that chimeric genes may also have been categorized as duplicated in this study if they show sufficiently strong similarity to a related gene. Wolfe and Shields (1997) have argued that gene duplications in the yeast genome were formed simultaneously through tetraploidy and then redistributed throughout the genome by translocation and deletion. It is impossible that the same phenomenon underlies the duplications described here, given the high frequency of closely spaced *cis* duplications. In agreement with this, analysis of dot-matrix plots failed to reveal widespread regional duplications of groups of adjacent genes.

There was evidence for three small but statistically robust regional duplications within chromosomes V and X and between chromosome II and chromosome V. All three showed high levels of protein sequence similarity and contained no intervening, unduplicated genes; it is therefore likely that they are recent duplications. Two of the three regional duplications involved genes that are thought to be members of operons, on the basis of their proximity to *trans*-splice sites. However, this represents only 4 genes of 18 involved in the duplicated regions, which is perhaps unsurprising since about 25% of *C. elegans* genes are expressed in operons (Blumenthal and Steward 1997). The fact that no older duplicated regions were found suggests two possibilities. It may be that regional duplications have been rare in the evolutionary history of *C. elegans;* a second possibility is that genomic rearrangements or deletions are fixed at a sufficiently high rate to remove the evidence for regional duplications. The presence of many closely spaced *cis* duplications in the genome suggests that translocations are not frequent. However, the presence of deletions spanning hundreds of base pairs in each (apparently recent) regional duplication suggests that deletions may become fixed at rates exceeding those of such duplications. This discrepancy may contribute to the compactness of the *C. elegans* genome (Waterston et al. 1997). The absence of older regional duplications, despite the evidence that regional duplications are formed (Fig. 4), suggests that they tend to decay into pseudogenes more rapidly than they can be recruited to new functions.

In contrast to the widespread evidence for gene duplication, gene conversion events were detected between only 2% of duplicated pairs and involving only 212 genes. Although gene conversion has been demonstrated in various organisms, it has only been investigated experimentally in detail only in yeast. These studies have provided information on spontaneous rates, tract lengths, and other characteristics (Petes et al. 1991). Our data represent the gene conversion events that remain detectable after the effects of genomic rearrangements and selection, so inferences cannot be made about spontaneously occurring characteristics of gene conversion. However, there are interesting parallels between our results and those of previous studies. For example, the rate of intrachromosomal or *cis* gene conversion has been estimated to be about three times that of *trans* or interchromosomal gene conversion in yeast (Petes and Hill 1988), and in our data the ratio is 3.7. As with yeast (Harris et al. 1993) we found that the frequency of gene conversion correlates negatively with DNA sequence divergence of the template sequences. Little is known about how the physical separation or the relative orientations of the sequences involved effect gene conversion frequency (Petes and Hill 1988). Our results suggest that the frequency of conversion is negatively correlated with the distance between gene pairs and that gene conversion events are more likely between genes with the same orientation.

It has been shown that the rate of gene conversion for a given gene increases in proportion to the number of available, identical donor sequences (Melamed and Kupiec 1992). By extension, one might assume that the larger the multigene family to which a gene belongs, the more likely it would be to undergo gene conversion. Our results show the opposite (Fig. 5A). Family size is negatively correlated with conversion frequency because larger families contain more divergent members. It would seem that the degree of sequence identity between divergent members is inadequate as a substrate for gene conversion. Alternatively, it may be that as the number of donor sequences increases so does the frequency of gene conversion but it becomes more difficult to detect substantial tracts.

In conclusion, there are substantial areas of agreement between previous laboratory work on gene conversion and the analysis presented here. The relative rates of *cis* and *trans* gene conversion and the negative correlation between conversion frequency and DNA sequence divergence for unconverted regions of converted pairs are consistent with previous studies. Other results, on the physical separation and relative orientations of gene converted sequences could be tested using existing methods in yeast (Petes and Hill 1988). This demonstrates that sequence analysis can complement the laboratory findings in this field.

Acknowledgments. We are grateful to S. Sawyer for advice on the use of his VTDIST3 program and to an anonymous reviewer for helpful comments. This work was carried out with funding from the EU 4th Framework Biotechnology Programme (Contract BIO4-CT95-0130).

References

Ahn S, Tanksley SD (1993) Comparative linkage maps of the rice and maize genomes. Proc Natl Acad Sci USA 90:7980–7984

Altschul SF, Gish W, Miller W, Myers EW, Lipman, DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Blumenthal T, Steward K (1997) RNA processing and gene structure. In: Riddle DL, Blumenthal T, Meyer BJ, Priess JR (eds) *C. elegans* II. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, p. 23

Brenner S, Hubbard ET, Murzin A, Chothia C (1995) Gene duplications in *Haemophilus influenzae.* Nature 378:140

Coissac E, Maillier E, Netter P (1997) A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol Biol Evol 14:1062–1074

Coulson AR, Sulston J, Brenner S, Karn J (1986) Toward a physical map of the genome of the nematode *C. elegans.* Proc Natl Acad Sci USA 83:7821–7825

Durbin R, Thierry-Mieg J (1996) In: The ACEDB Genomic Database, World Wide Web. [URL: ftp://ftp.sanger.ac.uk/pub/acedb]

Frugoli JA, McPeek MA, Thomas TL, McClung CR (1998) Intron loss and gain during evolution of the catalase gene family in angiosperms. Genetics 149:355–365

Gangloff S, Lieber MR, Rothstein R (1994) Transcription, topoisomerases and recombination. Experientia 50:261–269

Gangloff S, Zou H, Rothstein R (1996) Gene conversion plays the major role in controlling the stability of large tandem repeats in yeast. EMBO J 15:1715–1725

Harris S, Rudnicki KS, Haber JE (1993) Gene conversions and crossing over during homologous and homeologous ectopic recombination in *Saccharomyces cerevisiae.* Genetics 135:5–16

Hein, J (1993) A heuristic method to reconstruct the history of sequences subject to recombination. J Mol Evol 36:396–405

Hibner BL, Burke WD, Eickbush TH (1991) Sequence identity in an early chorion multigene family is the result of localised gene conversion. Genetics 128:595–606

Hillis DM, Moritz C, Porter CA, Baker RJ (1991) Evidence for biased gene conversion in concerted evolution of ribosomal DNA. Science 251:308–310

Holland PW, Garcia-Fernandez HJ, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. Development (Suppl):125–133

Jakobsen IB, Wilson SR, Easteal S (1997) The partition matrix: exploring variable phylogenetic signals along nucleotide sequence alignments. Mol Biol Evol 14:474–484

Labedan B, Riley M (1995) Gene products of *Escherichia coli:* sequence comparisons and common ancestries. Mol Biol Evol 12: 980–987

Li W-H (1997) Concerted evolution of multigene families. In: Li W-H (ed) Molecular evolution. Sinauer Associates, Sunderland, MA, p. 309

Maynard Smith J (1992) Analyzing the mosaic structure of genes. J Mol Evol 34:126–129

Melamed C, Kupiec M (1992) Effect of donor copy number on the rate of gene conversion in the yeast *Saccharomyces cerevisiae.* Mol Gen Genet 235:97–103

Park Y-S, Kramer JM (1990) Tandemly duplicated *C. elegans* collagen genes differ in their modes of splicing. J Mol Biol 211:395–406

Petes TD, Hill CW (1988) Recombination between repeated genes in microorganisms. Annu Rev Genet 22:147–168

Petes TD, Malone RE, Symington LS (1991) Recombination in yeast. In: Broach J, Jones E, Pringle J (eds) The molecular and cellular biology of the yeast *Saccharomyces:* genome dynamics, protein synthesis and energetics, Vol I. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, p. 407

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

Russnak RH, Candido EPM (1985) A locus encoding a family of small heat-shock genes in *C. elegans.* Two genes duplicated to form a 3.8 kilobase inverted repeat. Mol Cell Biol 5:1268–1278

Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Shyue S-K, Hewett-Emmett D, Sperling HG, Hunt DM, Bowmaker JM, Mollon JD, Li W-H (1994) Intronic gene conversion in the evolution of human X-linked color vision genes. Mol Biol Evol 11:548–551

Sonnhammer ELL, Durbin R (1997) Analysis of protein domain families in *C. elegans.* Genomics 46:200–216

Sonnhammer ELL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam—Multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26:320–322

Stephens JC (1985) Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol Biol Evol 2:539–556

Szostak JW, Wu R (1980) Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae.* Nature 284:426–430

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Waterston R, Sulston J (1995) The genome of *C. elegans.* Proc Natl Acad Sci USA 92:10836–10840

Waterston RH, Sulston JE, Coulson AR (1997) The Genome. In: Riddle DL, Blumenthal T, Meyer BJ, Priess JR (eds) *C. elegans* II. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY p. 23

Wilson R, et al. (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans.* Nature 368:32–38

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713

Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Comp Chem 17:149–163

Zhou Y-H, Li W-H (1996) Gene conversion and natural selection in the evolution of X-linked color vision genes in higher primates. Mol Biol Evol 13:780–783