Gene duplications and genetic redundancy in *C. elegans*[*]

Alison Woollard[§], Genetics Unit, Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK

Table of Contents

Abstract

Evolutionary innovation requires genetic raw materials upon which selection can act. The duplication of genes is of fundamental importance in providing such raw materials. Gene duplications are very widespread in *C. elegans* and appear to arise more frequently than in either *Drosophila* or yeast. It has been proposed that the rate of duplication of a gene is of the same order of magnitude as the rate of mutation per nucleotide site, emphasising the enormous potential that gene duplication has for generating substrates for evolutionary change.

The fate of duplicated genes is discussed. Complete functional redundancy seems unstable in the long term. Most models require that equality amongst duplicated genes must be disrupted if they are to be preserved. There are various ways of achieving inequality, involving either the nonfunctionalization of one copy, or one copy acquiring some novel, beneficial function, or both copies becoming partially compromised so that both copies are required to provide the overall function that was previously provided by the single ancestral gene. Examples of *C. elegans* gene duplications that appear to have followed each of these pathways are considered.

1. Gene duplications and genetic redundancy

The duplication of genes is of paramount importance in providing raw materials for the evolution of genetic diversity. It is therefore of interest to consider *C. elegans* gene duplications in the context of other sequenced genomes. Two recent estimates of the overall extent of gene duplications in *C. elegans* have come up with very similar figures. There are thought to be around 1,200 gene families containing two or more paralogues in *C. elegans* (Cavalcanti et al., 2003; Gu et al., 2002). The total number of genes in the paranome is around 6,000, or ~ 32% of

[*]Edited by Philip Anderson and Jonathan Hodgkin. Last revised June 8, 2005. Published June 25, 2005. This chapter should be cited as: Woollard, A. Gene duplications and genetic redundancy in *C. elegans* (June 25, 2005), *WormBook*, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.2.1, http://www.wormbook.org.

[§]To whom correspondence should be addressed. E-mail: woollard@bioch.ox.ac.uk

the genome, where the paranome is defined as the set of proteins that have one or more paralogues, that is, those that are not singletons. In addition, 7.1% of the duplicated genes in the worm are thought to have resulted from block duplications (duplication events involving more than one gene; Cavalcanti et al., 2003). Simple sequence duplications range in size from hundreds of bases to tens of kilobases and copies may be dispersed or clustered (The *C. elegans* Sequencing Consortium, 1998). Local clusters of duplicated genes are easily identified and there are 402 such clusters (where a cluster is defined as a group of N genes that are similar within a window of 2N genes along the chromosome and where N = 3 or more) distributed throughout the genome (The *C. elegans* Sequencing Consortium, 1998). The worm genome contains twice the number of local gene duplications as *Drosophila* (Rubin et al., 2000), and also differs from the fly genome in terms of the distribution of gene duplicates. Gene duplications are found randomly throughout the fly genome, whereas in the worm duplicated genes are mostly clustered in the recombinogenic segments of the autosomal arms (Rubin et al., 2000).

The rate of origin of gene duplicates in *C. elegans* over the past few hundred thousand years appears to be substantially higher than that for *Drosophila* or yeast (Lynch and Conery, 2000). Lynch and Conery (2000) propose a per-gene rate of duplication of 0.02 per million years for *C. elegans*, compared with a rate of 0.002 duplications per gene per million years in *Drosophila* and 0.008 in yeast. Gu et al. (2002) have reported exactly the same figure for *C. elegans*, compared with a rate of only 0.0014 for *Drosophila* (Gu, 2003).

Given the range of values predicted for different species, it has been proposed that 50% of genes in a genome would be expected to duplicate, on average, at least once on time scales of 35 to 350 million years (Lynch and Conery, 2000). Thus, even in the absence of direct amplification of the entire genome (polyploidization), gene duplication has substantial potential for generating substrates for evolutionary innovation. Indeed, it has been proposed that the rate of duplication of a gene is of the same order of magnitude as the rate of mutation per nucleotide site (Lynch and Conery, 2000).

What is the mechanism of gene duplication? Replication slippage and unequal exchange are often invoked as an explanation for closely spaced gene duplicates, but these mechanisms would be expected to give rise to tandem duplicates pointing in the same direction. A recent study found, however, that up to 69% of *C. elegans* duplicate genes reside in the inverse orientation, especially young duplicates (Katju and Lynch, 2003). Inversions are considered by these authors to be part and parcel of the original duplication event, rather than secondary rearrangements (Katju and Lynch, 2003). It has been hypothesized that inverted duplications could be generated by an illegitimate recombination event during DNA replication, involving strand switching by the DNA polymerase or strand misalignment-realignment. RNA-mediated transposition is thought unlikely to have played a significant role in gene duplication within the *C. elegans* genome because of the very small proportion of duplicate genes that lack introns relative to the original copy (Katju and Lynch, 2003). The mechanisms responsible for gene duplication are therefore, in general, unlikely to respect gene boundaries. This is borne out by the Katju and Lynch study, which analysed 290 gene pairs with ≤ 10% divergence at synonymous sites (Ks) within the *C. elegans* genome. They found that the average duplication span of 1.4 kb is less than the average gene length in *C. elegans* (2.5 kb), suggesting that partial gene duplications are frequent (Katju and Lynch, 2003).

Around half of the *C. elegans* gene duplicates with very low levels of synonymous site substitution were found to also contain unique coding sequence not present in the other copy, in addition to the region of close homology (Katju and Lynch, 2003). Thus, structural heterogeneity between duplicate genes is common. Chimeric duplicates may well have creative potential, especially when they act in conjunction with shuffling events.

What fate awaits duplicate genes? Three alternative theories have been proposed (Lynch, 2002; Lynch and Conery, 2000). The first is the nonfunctionalization of one copy by the accumulation of degenerative mutations. The second is neofunctionalization. In this scenario, one copy acquires a novel, beneficial function and becomes preserved by natural selection, with the other copy retaining the original function. A variation of this scenario is that the two genes could acquire divergent functions but maintain a functional overlap. Selection would then act on these divergent properties and indirectly maintain the functional overlap. Selection could also act on a newly emergent property unique to the combined action of two closely related genes, or on some enhanced efficiency or fidelity achieved by the combined action of two such genes (Thomas, 1993). In the third scenario, subfunctionalization, both copies become partially compromised by the accumulation of mutations, to a point where their total capacity is reduced to the level of the single-copy ancestral gene.

Whatever the mechanism, most models for conserving redundancy over the course of evolution require some degree of symmetry-breaking: equality amongst duplicated genes must be disrupted if they are to be preserved. The fate awaiting most gene duplicates in *C. elegans*, as well as in most other species studied, is likely to be silencing

and ultimate loss, rather than preservation (Lynch and Conery, 2000). It has been estimated that the average half-life of a gene duplicate is around 4 million years (Lynch and Conery, 2000). The propensity of recently duplicated genes to become nonfunctional pseudogenes is borne out in a recent study by Mounsey et al. (2002), who found a lower than expected success rate in generating expression patterns for recently duplicated genes, suggesting that recently duplicated genes are less likely to be expressed. Given our current state of knowledge of gene function in *C. elegans*, it is impossible to estimate the actual proportion of duplicated genes in the genome that are redundant. This will become clearer as sensitized screens and combinatorial RNAi approaches seek to address the general problem of assigning gene function.

One recent study examined the relationship between the prevalence of gene duplications and ontogeny in *Caenorhabditis* species (Castillo-Davis and Hartl, 2002). It was found that genes expressed after embryogenesis had a significantly greater number of duplicates than those expressed early in development. This was found to be true in both the *C. elegans* and *C. briggsae* genomes. For example, 18.36% of early-expressed genes (n=1,280) were found to have detectable paralogs in the *C. elegans* genome versus 35.31% of late-expressed genes (n=1,014). For *C. briggsae*, the figures are even more striking, 6.7% (n=165) and 38.8% (n=237), respectively (Castillo-Davis and Hartl, 2002). Therefore, duplicated copies of early-expressed genes appear to be selectively lost. Based on earlier calculations of the rate of origin of gene duplicates in *C. elegans* (Lynch and Conery, 2000) and the divergence between *C. elegans* and *C. briggsae*, Castillo-Davis and Hartl estimated that more than 40% of all genes are expected to have duplicated at least once in both the *C. elegans* and *C. briggsae* lineages since their divergence (Castillo-Davis and Hartl, 2002). The proportion of duplicated genes in the late-expressed class thus falls close to this estimate, whereas the proportion of duplicated genes in the early-expressed class is much lower than expected. It is suggested that the selective loss of duplicates of early-expressed genes reflects developmental constraint (Castillo-Davis and Hartl, 2002). A related study found that gene duplications were more prevalent amongst the set of conserved, slowly-evolving genes versus the set of non-conserved genes in the *C. elegans* and *S. cerevisiae* genomes, suggesting that slowly evolving genes may be the main source of new genes in eukaryotic genomes (Davis and Petrov, 2004).

11 of 33 of the largest clusters of duplicated genes in *C. elegans* consist of olfactory receptors (Rubin et al., 2000). Olfactory receptors are seven-transmembrane G-protein coupled receptors (GPCRs), each of which specifically recognises a set of odorant and tastant chemicals, allowing the worm to sense and respond to its chemical environment. GPCRs comprise by far the largest multigene family in *C. elegans*, with over 1000 members, making up ~ 5% of the genome (Bargmann, 1998), although around a third of this group of genes are thought to be pseudogenes (Bargmann, 1998). Thus many duplications in this large multigene family are fated to become nonfunctional. Several recent studies have charted the molecular evolution of some of these genes and found that processes of duplication, diversification and movement that have led to these large gene families are very much ongoing. For example, the 3126 bp inverse orientation duplication that duplicates the 5' half of gene T08H10.2 to give rise to pseudogene T08H10.a is thought to have occurred very recently as the duplicated sequences are identical (Robertson, 1998). It has been suggested that strong selective pressures are at work for the continued functionality of these genes because the occurrence of synonymous changes in duplicated GPCR genes is 11-fold higher than the occurrence of amino acid substitutions or nonsynonymous changes (Robertson, 1998; Robertson, 2000). Indeed, it has been proposed that natural selection might somehow favour duplications of genes that are generally involved in responses to environmental stress and pathogens, in organisms facing a challenging and dynamic molecular environment (Lespinet et al., 2002). The massive lineage specific expansion of worm odorant/chemosensory receptors could be just one example of this.

Other large gene families in *C. elegans* include C-type lectins, hormone receptors, collagens and serine/threonine/tyrosine protein kinases. A full description of major protein-coding gene families, including a discussion of family size distribution, can be found in Genomic classification of protein-coding gene families. An example of a *C. elegans* transcription factor gene family in which duplications abound is the T-box family. There are 21 T-box genes in the *C. elegans* genome (WormBase release WS132, 2004), most of which lack clear orthologs in other species. There are 4 pairs of genes that are likely to have arisen from relatively recent duplication events and a functional analysis of 2 of these pairs has been reported. *tbx-37* and *tbx-38* (83% amino acid identity) have redundant functions in mesoderm induction in *C. elegans* embryos (Good et al., 2004), whereas *tbx-8* and *tbx-9* (59% amino acid identity), have overlapping, but probably not completely redundant, functions in embryonic morphogenesis (Pocock et al., 2004). The most recent duplication in the T-box gene family would appear to be the one that gave rise to the genes Y59E9AR.3 (*tbx-30*) and Y59E9AR.5 (*tbx-30.1*). These two sequences are 100% identical and situated in inverse orientations, ~ 2 kb apart. Biological function has been ascribed to *tbx-30* (Pocock et al., 2004), but it is unclear at present whether or not *tbx-30.1* is expressed.

There are several other good examples of duplicated gene pairs that have undergone neofunctionalization. A recent one is the *fbf-1/fbf-2* gene pair involved in germ line development in *C. elegans*. *fbf-1* and *fbf-2* encode closely related RNA binding proteins that reciprocally regulate the expression of one another in order to modulate the size of the germ line mitotic region (Lamont et al., 2004). Thus, there appears to be a benefit unique to the combined action of these two closely related genes; the duplication event in this case has provided an opportunity for fine-tuning of a developmental process.

A good potential example of the subfunctionalization of a duplicated gene pair in *C. elegans* is the case of the Notch-like receptors *lin-12* and *glp-1*. In *C. elegans*, *lin-12* and *glp-1* posess both overlapping and separate biological functions. When both gene functions are removed, larval lethality results (the Lag – Lin And Glp phenotype; Lambie and Kimble, 1991). However, in *C. briggsae* and *C. remanei*, the Lag phenotype is seen when only *lin-12* expression is silenced (Rudel and Kimble, 2002). It is suggested that ancestral functions may have been divided between *lin-12* and *glp-1* in *C. elegans* after duplication in such a way that their total capacity became reduced to the level of the single-copy ancestral gene. In *C. briggsae* and *C. remanei*, on the other hand, *lin-12* appears to have retained this ancestral function following duplication (Rudel and Kimble, 2002).

An interesting example of the apparent gradual evolutionary demise of a duplicated gene is presented in the case of *elt-4*. *elt-4* encodes a truncated GATA type zinc finger transcription factor of just 8.1kDa (72 amino acids), and is situated ~ 5 kb upstream of the highly conserved *elt-2*, proposed to be the original copy of the duplicate pair (Fukushige et al., 2003). The bigger *elt-2* gene is expressed in intestinal cells and is required for intestinal development (Fukushige et al., 1998). While it is clear that *elt-4* is expressed in the intestine, no effect was found of deleting *elt-4*, in the gut or elsewhere, either alone or in combination with *elt-2*. Furthermore, experiments in yeast and *in vitro* could not demonstrate any role for ELT-4 in activating or repressing transcription, or indeed any specific DNA binding activity (Fukushige et al., 2003). The *elt-2*/*elt-4* duplication event is thought to have occurred only in the *C. elegans* lineage, after the point at which *C. elegans* and *C. briggsae* diverged. Thus, the proposal is that although *elt-4* may have conferred some selective advantage to *C. elegans* in the past (hence the high level of conservation of the zinc finger domain), its ultimate evolutionary fate will be disappearance from the *C. elegans* genome (Fukushige et al., 2003).

Genome sequence analysis is thus providing the community with far more than simply the informational content of genomes. Gene duplications are widespread in the *C. elegans* genome and provide raw material for evolutionary novelty. The data are a snapshot of evolutionary time, from which we can glimpse into the past as well as, perhaps, seek to prophesy the future.

2. Acknowledgements

I would like to thank members of my laboratory and Andreas Russ for useful discussions and the reviewers for helpful comments.

3. References

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

Castillo-Davis, C.I., and Hartl, D.L. (2002). Genome evolution and developmental constraint in *Caenorhabditis elegans*. Mol. Biol. Evol. *19*, 728–735. Abstract

Cavalcanti, A.R.O., Ferreira, R., Gu, Z., and Li, W.-H. (2003). Patterns of gene duplication in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. J. Mol. Evol. *56*, 28–37. Abstract Article

Davis, J.C., and Petrov, D.A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. PLoS. Biol. *2*, E55. Epub 2004 Mar 16. Abstract Article

Fukushige, T., Goszczynski, B., Tian, H., and McGhee, J.D. (2003). The evolutionary duplication and probable demise of an endodermal GATA factor in *Caenorhabditis elegans*. Genetics *165*, 575–588. Abstract

Fukushige, T., Hawkins, M.G., and McGhee, J.D. (1998). The GATA-factor *elt-2* is essential for formation of the *Caenorhabditis elegans* intestine, Dev. Biol. *198*, 286–302. Abstract

Good, K., Ciosk, R., Nance, J., Neves, A., Hill, R.J., and Priess, J.R. (2004). The T-box transcription factors TBX-37 and TBX-38 link GLP-1/Notch signaling to mesoderm induction in *C. elegans* embryos, Development *131*, 1967–1978. Epub 2004 Mar 31. Abstract Article

Gu, X. (2003). Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet. *19*, 354–356. Abstract Article

Gu, Z., Cavalcanti, A., Chen, F.-C., Bouman, P., and Li, W.-H. (2002). Extent of gene duplication in the genomes of *Drosophila*, nematode and yeast. Mol. Biol. Evol. *19*, 256–262. Abstract

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature *421*, 231–237. Article

Katju, V., and Lynch, M. (2003). The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics *165*, 1793–1803. Abstract

Lambie, E.J., and Kimble, J. (1991). Two homologous regulatory genes, *lin-12* and *glp-1*, have overlapping functions. Development *112*, 231–240. Abstract

Lamont, L.B., Crittenden, S.L., Bernstein, D., Wickens, M., Kimble, J. (2004). FBF-1 and FBF-2 regulate the size of the mitotic region in the *C. elegans* germline. Dev. Cell *7*, 697–707. Abstract Article

Lespinet, O., Wolf, Y.I., Koonin, E.V., and Aravind, L. (2002). The role of lineage-specific expansion in the evolution of eukaryotes. Genome Res. *12*, 1048–1059. Abstract Article

Lynch, M. (2002). Genomics. Gene duplication and evolution. Science *297*, 945–947. Abstract Article

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151–1155. Abstract Article

Mounsey, A., Bauer, P., and Hope, I.A. (2002). Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. Genome Res. *12*, 770–775. Abstract Article

Pocock, R., Ahringer, J., Mitsch, M., Maxwell, S., and Woollard, A. (2004). A regulatory network of T-box genes and the *even-skipped* homologue *vab-7* controls patterning and morphogenesis in *C. elegans*. Development *131*, 2373–2385. Epub 2004 Apr 21. Abstract Article

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. Abstract

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. Abstract Article

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C. R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. (2000). Comparative genomics of the eukaryotes. Science *287*, 2204–2215. Abstract Article

Rudel, D., and Kimble, J. (2002). Evolution of discrete Notch-like receptors from a distant gene duplication in *Caenorhabditis*. Evol. Dev. *4*, 319–333. Abstract Article

The *C. elegans* Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science *282*, 2013–2018. Abstract

Thomas, J.H. (1993). Thinking about genetic redundancy. Trends Genet. *9*, 395–399. Abstract Article