6 Doolittle, W.F. (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* 14, 307–311
7 Wolf, Y.I. *et al.* (2000) No footprints of primordial introns in a eukaryotic genome. *Trends Genet.* 16, 333–334
8 Ochman, H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304
9 Ponting, C.P. *et al.* (1999) Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.* 289, 729–745
10 de la Cruz, F. and Davies, J. (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* 8, 128–133
11 Lefebvre, L. *et al.* (1998) Abnormal maternal behaviour and growth retardation associated with loss of the imprinted gene *Mest*. *Nat. Genet.* 20, 163–169

C.P. Ponting
MRC Functional Genetics Unit,
Dept of Human Anatomy and Genetics,
University of Oxford, South Parks Road,
Oxford, UK OX1 3QX.
e-mail: Chris.Ponting@anat.ox.ac.uk

Birth and death of duplicated genes in completely sequenced eukaryotes

Andreas Wagner

Gene and genome duplications are commonly regarded as being of major evolutionary significance. But how often does gene duplication occur? And, once duplicated, what are the fates of duplicated genes? How do they contribute to evolution? In a recent article, Lynch and Conery analyze divergence between duplicate genes from six eukaryotic genomes. They estimate the rate of gene duplication, the rate of gene loss after duplication and the strength of selection experienced by duplicate genes. They conclude that although the rate of gene duplications is high, so is the rate of gene loss, and they argue that gene duplications could be a major factor in speciation.

The importance of evolution by gene duplication, first forcefully advocated in Ohno's visionary book[1], is now universally accepted. More than a third of a typical eukaryotic genome consists of duplicate genes and gene families. Gene duplications are thus a key force of genome evolution. After duplication, gene duplicates often experience relaxed evolutionary constraints. This promotes functional diversification of duplicates and biochemical innovation through mutations and recombination. And yet another role for closely related gene duplicates is suggested by recent theoretical population genetic work[2–4]: duplicate genes with redundant functions might insulate an organism against otherwise deleterious mutations.

Despite its obvious significance, many questions remain about the details of the gene duplication process:

- What is the rate at which gene duplications occur?
- Once a gene is duplicated, what are the chances that the duplication becomes fixed in a population?
- How long does it take until such fixation?
- Do many duplicates evolve new functions?
- How long does it take until one of the duplicates suffers degenerative mutations and becomes silenced?
- Do the vast majority of gene duplicates become silenced?

In a study unprecedented in both scale and resolution, Lynch and Conery[5] now answer some of these questions. They analyze duplicate genes in three completely sequenced eukaryotic genomes (*Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Caenorhabditis elegans*) and in three others for which hundreds or thousands of duplicate genes are available (*Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana*). However, before discussing their work in greater detail, I will embed it in the context of some previous work.

Previous contributions

Most quantitative questions about the process of gene duplications and subsequent gene loss have eluded conclusive answers, but not from lack of interest by the scientific community. Take the question about the rate of gene loss after duplication, for example. When addressing it, one must distinguish gene loss after whole-genome duplication (polyploidization) from gene loss after duplication of single genes. Some studies have addressed gene loss after polyploidization[6–11]. They ask how many duplicate genes a genome retains long (>50 Myr) after a genome duplication. These studies rely on a range of approaches, from studying electrophoretic separation of isozymes at duplicate loci, to analyzing complete genome sequence information. Their range of answers is similarly broad. Between 50% and 92% of all duplicate genes appear to get lost eventually, depending on the study consulted. Where complete sequence information is available[10,11], the loss estimates are substantially higher than 50%. For instance, in the case of yeast, 92% of all genes could have been lost since a genome duplication that occurred some 100 Myr ago[10].

The consensus of earlier studies on single gene duplications is that the vast majority of gene duplicates should get lost. Most of these studies rely on population genetic models[12–19]. Albeit shrouded in mathematical argument, one of their key messages is plain: because both duplicates have identical functions after duplication, one of them is free to degenerate through loss-of-function mutations. It is just a matter of when. From this perspective, one might ask why gene duplicates persist at all. And why are eukaryotic genomes full of them?

Existing models could underestimate the rate of gene loss because of the limited perspective on gene function that many of them take. According to this perspective, a gene either functions properly or it harbors a deleterious mutation. However, genes can have many functions, each of which can be independently affected by mutations that are not necessarily deleterious if two copies are available. Consider the multiple spatio-temporal expression domains of many developmental genes. Distinct modular enhancers often establish these domains. Mutations affect modular enhancers

 0168–9525/01/$ – see front matter PII: S0168-9525(01)02243-0

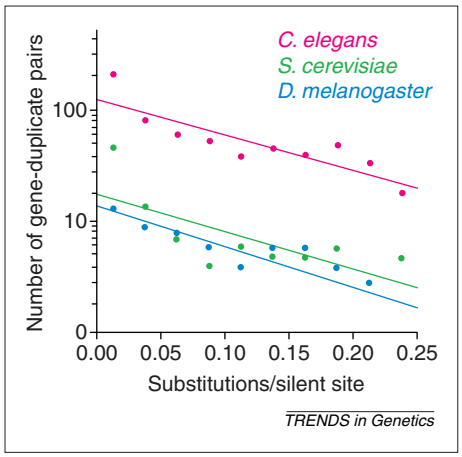

Fig. 1. Survival of gene duplicates based on the complete sequences of *Caenorhabditis elegans* (pink), *Saccharomyces cerevisiae* (green) and *Drosophila melanogaster* (blue). Redrawn, with permission, from Ref. 5.

independently and can thus eliminate part of an expression domain. The result of such partially degenerative mutations in one of two gene duplicates is that gene duplicates evolve overlapping functions from initially identical functions. And, as opposed to loss of a gene with a completely redundant function, loss of a gene with overlapping functions might not be easily tolerated. If partially redundant duplicate genes are abundant, as suggested by many recent genetic studies on developmental genes, then models using a simple all-or-none model of gene function could drastically underestimate the rate of gene loss.

A whole-genome study on single gene duplications

There is no end to the number of arguments one can conceive for why gene loss should be either frequent or rare. But the underlying question can only be answered by analyzing empirical data, data that has only become available recently. Lynch and Conery[5] present the first whole-genome based study – three whole genomes and three partial ones – to answer this and other quantitative questions about the process of gene duplication. They first identify duplicate gene pairs by (1) searching for similar sequence pairs using gapped BLAST for all available translated open reading frames, (2) eliminating ambiguously aligned amino acid sequences from each gene pair thus found, and (3) aligning nucleotide sequences guided by the remaining amino acid sequence alignment. From this nucleotide alignment, they estimate the fraction of synonymous (silent) nucleotide substitutions, S, and that of nonsynonymous (amino acid replacement) nucleotide substitutions, R, per nucleotide site for each identified gene pair[20].

These quantities are useful for two reasons. First, S provides a crude measure of the time since duplication for each gene pair. The reason is that synonymous nucleotide substitutions are not subject to the same strong selection pressures as nonsynonymous or replacement substitutions. Thus, they accumulate at a stochastic rate proportional to time[20]. Second, the ratio R/S provides a measure of the selection pressure a gene pair is subject to. If a duplicate gene pair shows $R/S \approx 1$ (i.e. amino acid replacement substitutions occur at the same rate as synonymous substitutions), then few or no amino acid replacement substitutions have been eliminated since the gene duplication. In other words, the duplicate genes are under few or no selective constraints. More frequent is the case of 'purifying selection' ($R/S < 1$). Here, some replacement substitutions have been purged by natural selection, presumably because of their deleterious effects. The smaller R/S is, the greater is the number of eliminated substitutions and the greater is the selective constraint under which two genes evolve. (The case of $R/S > 1$, indicating frequently occurring advantageous mutations, is rare.)

How do R and S help to answer how fast genes are lost after duplication? Lynch and Conery bin closely related gene duplicates ($0.01 < S < 0.25$, corresponding to less than 25% divergence at synonymous sites) into several categories according to S. If gene duplications occur at an approximately constant rate and if duplication products survive indefinitely, then each bin should contain the same number of gene pairs. But if genes get lost after duplication, then the number of duplicates per bin should decrease with increasing S. The faster this number decreases, the greater the rate of gene loss. Lynch and Conery find a rapid and nearly constant decrease in the number of duplicates per bin (Fig. 1), from which they can estimate the half-life of duplicate genes. It ranges from 3 million to 7 million years for their study taxa and suggests that more than 90% of duplicates disappear before 50 million years have elapsed. Thus, their analysis supports the idea of pervasive gene loss after duplication.

In the absence of genome-scale data, it has been just as difficult to determine the rate of gene duplications as to determine the rate of gene loss. The rate of duplications is thus the most noteworthy among other rates that Lynch and Conery estimate[5]. They obtain it from information on (1) the number of extremely closely related gene pairs ($S < 0.01$), (2) the total (estimated) number of genes per genome, and (3) the independently estimated time[20] until two duplicates attain a divergence of $S = 0.01$. They arrive at duplication rates ranging from 0.002 (fruit fly) to 0.02 (nematode) per gene, per million years. Gene conversion is not likely to bias their estimates, partly because of the generally weak linkage between closely related duplicates. This average duplication rate per gene is remarkably high. The authors point out that it is of the order of the mutation rate per nucleotide site. In other words, gene duplications could be just as important as point mutations as a source of variation. Based partly on this finding, they argue that gene duplications and subsequent gene loss might have a significant role in speciation, because the loss of different gene copies in two isolated populations could cause genetic differences to accumulate rapidly.

A further important result is that distantly related, and thus old, duplicated gene pairs show a tenfold decrease in their R/S ratio, reflecting increased selective constraints compared with recent duplicates. Why is this significant? Because it confirms – on a genome-wide scale – the existence of something essential for gene duplication to be a biochemical innovator: relaxed selective constraints after duplication.

Limitations

What price, if any, do the authors pay for their peerless rate estimates, obtained by analyzing thousands of gene pairs? It is a price shared by most genome-level studies, the sacrifice of biologically relevant detail. The author's data show that for most values of S, divergence at replacement sites R varies by as much as two orders of magnitude. Buried in all this scatter is much – some might say most – relevant biological information about gene functions and the evolutionary constraints they imply.

Similarly, the statistical approach of calculating the half-life of a gene provides only a caricature of genome evolution. We

know many gene families that have persisted over a billion years, an event of astronomical improbability if one takes the short average half-lives of duplicates at face value. And, one might say such conserved gene families are most interesting biologically because they provide information on gene functions required for all or most living things.

Thus, genome-wide analyses of evolutionary rates have to be taken with a grain of salt. However, they avoid the nagging doubts about sample bias which small-scale studies are vulnerable to. Their quantitative results provide a dynamic picture of genome evolution, and they can put old debates to rest – finally and conclusively.

References

1 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag
2 Wagner, A. (2000) The role of pleiotropy, population size fluctuations, and fitness effects of mutations in the evolution of redundant gene functions. *Genetics* 154, 1389–1401
3 Nowak, M.A. *et al.* (1997) Evolution of genetic redundancy. *Nature* 388, 167–171
4 Force, A. *et al.* (1999) Preservation of duplicate genes by subfunctionalization. *Am. Zoologist* 39, 460–460
5 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155
6 Ferris, S.D. and Whitt, G.S. (1976) Loss of duplicate gene expression after polyploidization. *Nature* 265, 258–260
7 Ferris, S.D. and Whitt, G.S. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.* 12, 267–317
8 Allendorf, F.W. *et al.* (1975) In *Isozymes Vol. IV: Genetics and Evolution* (Markert, C.L., ed.), pp. 415–432, Academic Press
9 Nadeau, J.H. and Sankoff, D. (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* 147, 1259–1266
10 Seoighe, C. and Wolfe, K.H. (1998) Extent of genomic rearrangement after genome duplication in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 95, 4447–4452
11 The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
12 Watterson, G.A. (1983) On the time for gene silencing at duplicate loci. *Genetics* 105, 745–766
13 Kimura, M. and King, J.L. (1979) Fixation of a deleterious allele at one of two duplicate loci by mutation pressure and random drift. *Proc. Natl. Acad. Sci. U. S. A.* 76, 2858–2861
14 Maruyama, T. and Takahata, N. (1981) Numerical studies of the frequency trajectories in the process of fixation of null genes at duplicated loci. *Heredity* 46, 49–57
15 Nei, M. and Roychoudhury, A.K. (1973) Probability of fixation of nonfunctional genes at duplicate loci. *Am. Nat.* 107, 362–372
16 Clark, A.G. (1994) Invasion and maintenance of a gene duplication. *Proc. Natl. Acad. Sci. U. S. A.* 91, 2950–2954
17 Ohta, T. (1987) Simulating evolution by gene duplication. *Genetics* 115, 207–213
18 Ohta, T. (1988) Time for acquiring a new gene by duplication. *Proc. Natl. Acad. Sci. U. S. A.* 85, 3509–3512
19 Walsh, J.B. (1995) How often do duplicated genes evolve new functions? *Genetics* 139, 421–428
20 Li, W-H. (1997) *Molecular Evolution*, Sinauer

A. Wagner
University of New Mexico, Dept of Biology, 167A Castetter Hall, Albuquerque, NM 817131-1091, USA.
e-mail: wagnera@unm.edu

Bakers yeast rises to the challenge: reconstitution of mammalian steroid receptor signalling in *S. cerevisiae*

Iain J. McEwan

Steroid hormones are an important class of signalling molecule, regulating a diverse range of processes in metazoan eukaryotes. The actions of these hormones are mediated by intracellular receptor proteins that act as ligand-activated transcription factors. The ability to reconstitute steroid receptor signalling in the budding yeast, *Saccharomyces cerevisiae*, provides a genetically tractable model system in which to investigate steroid receptor structure and function. Through targeted disruption and genetic screening, an increasing number of genes have been identified that are likely to have a role in steroid receptor action.

Bakers yeast, *Saccharomyces cerevisiae*, is often referred to as a 'simple' eukaryote. However, the ability to grow yeast on synthetically defined media and the availability of protocols for efficient transformation with plasmid DNA, combined with well-established genetic tools, makes this yeast an attractive model organism for studying processes from less genetically tractable multicellular organisms. Thus, yeast has been used to unravel the molecular details of eukaryotic cell-division[1], gene transcription[2], DNA repair[3] and, more recently, ageing[4]. Moreover, the ability to reconstitute novel pathways from metazoa in this single-celled eukaryote is helping to illuminate the mechanisms involved in apoptosis[5] and signalling by steroid receptors and related proteins[6,7]. A recent study, by Sitcheran *et al.*[8], highlighted how *S. cerevisiae* can be a powerful tool in elucidating the molecular details of signalling by glucocorticoid steroid hormones.

'In the field of observation, chance favours only the prepared mind.'
Louis Pasteur

Steroid hormones have diverse physiological actions in multicellular organisms, being involved in development, reproduction, salt balance and the immune system. The actions of these hormones are mediated by intracellular receptor proteins that function as hormone-activated transcription factors (Ref. 9 and references therein). The receptor proteins have a characteristic domain structure consisting of a DNA-binding domain (DBD) flanked by a ligand-binding domain (LBD) within the C terminus of the protein and a highly variable N-terminal domain (NTD) that is involved in transcriptional transactivation. The general steps of the receptor signalling pathway have been known for some years. In the absence of hormone, steroid receptors are found in a complex with molecular chaperones, such as hsp90, in the cytoplasm or nucleus. Upon binding of the hormone, this complex dissociates and the receptor binds to specific DNA response elements to regulate target gene expression.

Why use yeast to investigate steroid receptor action?

Yeast cells do not contain members of the steroid receptor superfamily, but, by expression of the receptor proteins, receptor signalling can be reconstituted

 0168–9525/01/$ – see front matter PII: S0168-9525(01)02273-9