



ELSEVIER

# Turning the clock back on ancient genome duplication

Cathal Seoighe

Complete genome sequence data led rapidly to the conclusion that ancient genome duplications had shaped the genomes of the model organisms *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. Recent contributions have gone on to refine date estimates for these duplications and, in the case of *Arabidopsis*, to infer additional, more ancient, rounds of duplication by reconstructing gene order before the most recent duplication event. It is becoming widely accepted that an ancient duplication occurred before the radiation of the ray-finned fish. However, despite methodological advances and the availability of complete genome sequence data the debate over whether very ancient genome duplications have occurred early in the vertebrate lineage has not yet been fully resolved.

## Addresses

South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville 7535, South Africa  
e-mail: cathal@sanbi.ac.za

## Current Opinion in Genetics & Development 2003, 13:636–643

This review comes from a themed issue on  
Genomes and evolution  
Edited by Evan Eichler and Nipam Patel

0959-437X/\$ – see front matter  
© 2003 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2003.10.005

## Abbreviation

My million years

## Introduction

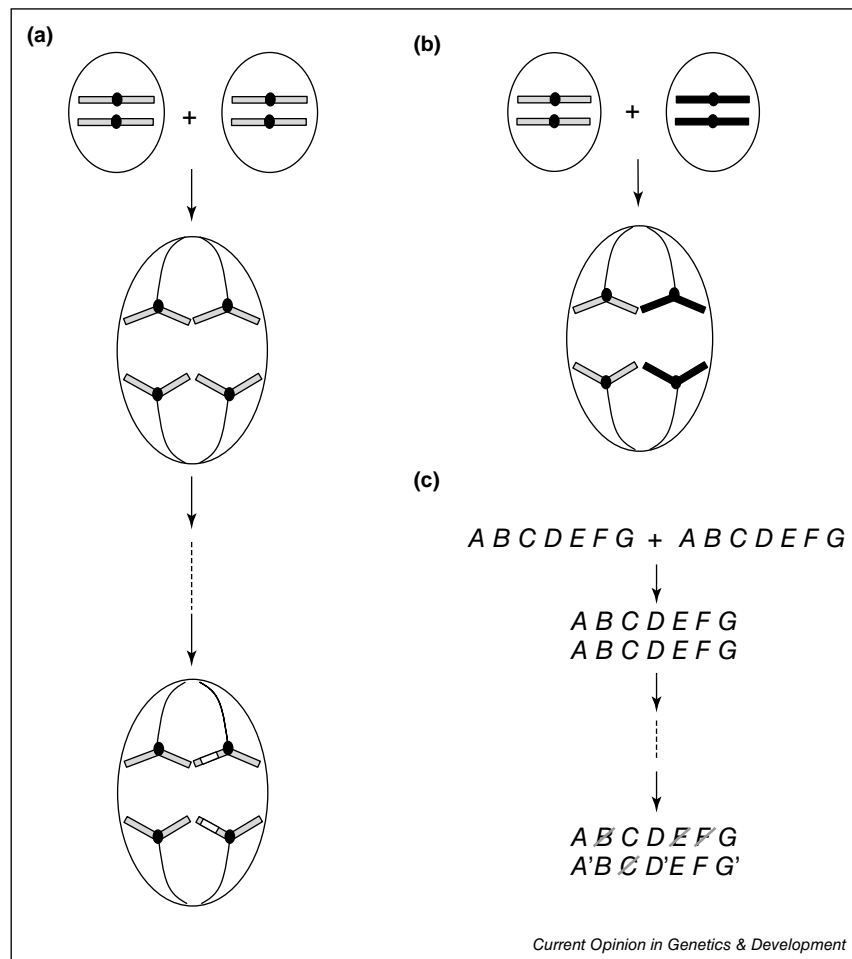
The availability of complete genome sequence data for several eukaryotes has increased the power of researchers to detect the traces of very ancient genome duplication events that have been obscured by chromosomal rearrangement and the deletion of duplicated genes [1]. The increasingly apparent ubiquity of genome duplication supports the long-held view that duplication of entire genomes has played a pivotal role in evolution [1,2]. Genome duplication or, more generally, polyploidy may result from the merger of more than one genome from the same species or from different species, termed auto-polyploidy and allopolyploidy, respectively (Figure 1). Following genome duplication in a diploid organism, chromosomal rearrangements and loss of duplicated genes can restrict pairing between homologous chromosomes and lead to the restoration by degrees of diploid chromosomal behaviour, through a process known as diploidization [1]. Convincing evidence of genome duplication has been uncovered in the genomes of *Saccharomyces cerevisiae*

[3,4,5\*\*] and *Arabidopsis thaliana* [6–8,9\*,10,11\*\*,12]. In a range of other eukaryotes that have been sequenced completely, varying degrees of evidence for ancient genome duplication have been uncovered (Table 1). As the debates over the occurrence of genome duplications continue, insights gained from investigation of recently polyploid plants, including many of the world's most economically important crops [13], may contribute to our ability to model evolution following genome duplication and thereby improve our understanding of its contribution to the evolutionary process.

## Inferring ancient polyploidy from complete genome sequence data

The majority of the yeast genome can be assigned to non-overlapping, partially duplicated segments that are the remnants of a single complete genome duplication [3]. These blocks have been thinned out by the loss of redundant gene copies and are bounded by chromosomal rearrangements [3,14]. Whereas successive duplications of individual segments of the genome would be likely to result in overlapping and multiple copy duplicated regions, the absence of these implies that the duplicated blocks have been produced by a single duplication event [3]. A flexible algorithm to delineate partially duplicated regions of this kind in genetic maps or genome sequence has recently been described [15]. Using the same approach, as well as other lines of argument, the debate over ancient polyploidy in *Arabidopsis* has become a matter of how many and when rather than whether ancient genome duplications have occurred [6–8,9\*,10,11\*\*,12]. By contrast, the completion of the first draft of the human genome sequence has failed to resolve the controversy over the contribution of genome duplication to the evolution of the lineage leading to the vertebrates [1,16,17]. Susumu Ohno [2] originally proposed that two or three rounds of genome duplication have occurred in the vertebrate lineage and the proposal that there have been exactly two rounds has since gained considerable ground (the 2R hypothesis; see [1] and [18] for reviews). The occurrence of more than one genome duplication, separated by an unknown time interval, complicates the task of finding the duplicated segments that have been used as evidence of ancient polyploidy [1]. In addition, the duplication events proposed under the 2R hypothesis are much more ancient than genome duplications that have been inferred in yeast and *Arabidopsis* and the duplicated regions that remain are likely to be sparse and substantially reshuffled by chromosomal rearrangement. No genome duplication as ancient as the duplications proposed under the 2R hypothesis has yet been proven.

Figure 1



Genome duplication of a diploid organism with  $2n = 2$ . **(a)** Genome duplication through the merger of two diploid nuclei from the same species (autotetraploidy). Initially duplicated chromosomes pair up and segregate randomly during cell division (tetrasomic inheritance). After a period of time, duplicated chromosomes differentiate, possibly as a result of structural changes or deletion of genetic material (represented here by white bands on chromosome arms). When duplicated chromosomes are sufficiently differentiated chromosomal pairing is again specific, marking a return to the normal diploid mode of inheritance (disomic inheritance). **(b)** Genome duplication through the merger of two diploid nuclei from different species (allotetraploidy). In the example illustrated, specific pairing and segregation of chromosomes occurs from the outset, although an intermediate state is possible, in which some chromosomes pair specifically and others randomly (segmental allotetraploidy). **(c)** Gene order on a segment of a chromosome is shown with genes represented by letters A–G. Immediately following whole genome duplication each gene is represented in duplicate and the genome is completely redundant. After a period of time, redundancy is reduced through gene loss and divergence of duplicate genes.

### Is the 2R hypothesis untestable?

Several kinds of evidence have been advanced in support of the 2R hypothesis. McLysaght, Hokamp and Wolfe [19\*\*] have reported 244 large duplicated blocks of  $\geq 4$  paralogues in the human genome and suggest that at least one round of polyploidy has occurred in the vertebrate lineage. Panopoulou *et al.* [20] have argued in favour of the 2R hypothesis on the basis of the numbers of genes in gene families of the vertebrates and related invertebrates. Other researchers have inferred ancient polyploidy from the observation of large numbers of duplicated genes that share roughly the same dates of

origin [21\*\*]. However, Lynch and Conery [7] suggest a typical rate of origin of new gene duplicates in the order of 0.01 per gene per My so that in 100 My the number of single-gene duplications should be approximately equal to the total number of genes in the organism. Given the generally high rate of gene duplication, caution should be exercised when using bursts of duplicate genes with a common time of origin spanning  $>100$  My to infer genome duplication. It is difficult to disprove that an apparent burst of duplication activity was not the result of a reduction in the probability of gene loss rather than a whole-genome duplication event.

Table 1

## Evidence for ancient genome duplications in completely sequenced eukaryotes.

Organism	Evidence for genome duplication	Dates (My)
<i>Saccharomyces cerevisiae</i>	Non-overlapping duplicated genomic segments with segment orientation conserved relative to centromere [3] Gene order in related organisms [5**]	100–150 [3,5**,44*]
<i>Arabidopsis thaliana</i>	Non-overlapping segments from successive rounds of genome duplication [11**,9*] Large number of duplicated genes with common date of origin [7] Gene order in related organisms [12]	Three rounds of duplication inferred $\alpha$ : 14.5–86 $\beta$ : 170–235 $\gamma$ : 300 [11**]
Vertebrate lineage	Large number of duplicated genes with common date of origin [21**] Number and phylogenetic analysis of specific paralogous regions (e.g. [31]) Whole genome analysis of duplicated segments [19**]	430–750 [21**]
Rice	Presence of many paralogous genomic regions (Genome duplication has been suggested [40*] but a more detailed analysis is required.)	40–50 [40*] (Date estimates were based on a single rate of amino acid divergence for the proteins studied and may not be accurate)
Ray-finned fishes ( <i>Fugu</i> and zebrafish)	Number and phylogenetic analysis of Hox regions [28] Large number of duplicated genes with common date of origin [30]	150–450 [28,29*] (This duplication is in addition to the proposed duplication shared by all vertebrates)

In spite of recent optimism [22] inspired by the age distribution of duplicated human genes [21\*\*] and the presence of large numbers of paralogous regions in the human genome [19\*\*], conclusive proof of even a single genome duplication in the vertebrate lineage remains elusive. The task of providing this proof is by no means trivial and any proof will have to contend with competing hypotheses (e.g. repeated segmental duplications that are known to occur in the human genome [23]). It is worth noting that most of the tests of genome duplication described above could provide a proof but could not disprove a genome duplication and there is generally a lack of clearly defined falsifiable predictions for the genome duplication model. The impact of whole genome duplication on comparative gene order (reviewed below) may provide some of the most testable predictions of a genome duplication hypothesis yet. If so, the power to finally resolve the 2R debate may lie within the gene orders of invertebrates such as amphioxus and the sea squirt, *Ciona intestinalis* [18], the genome of which has recently been published in draft form [24].

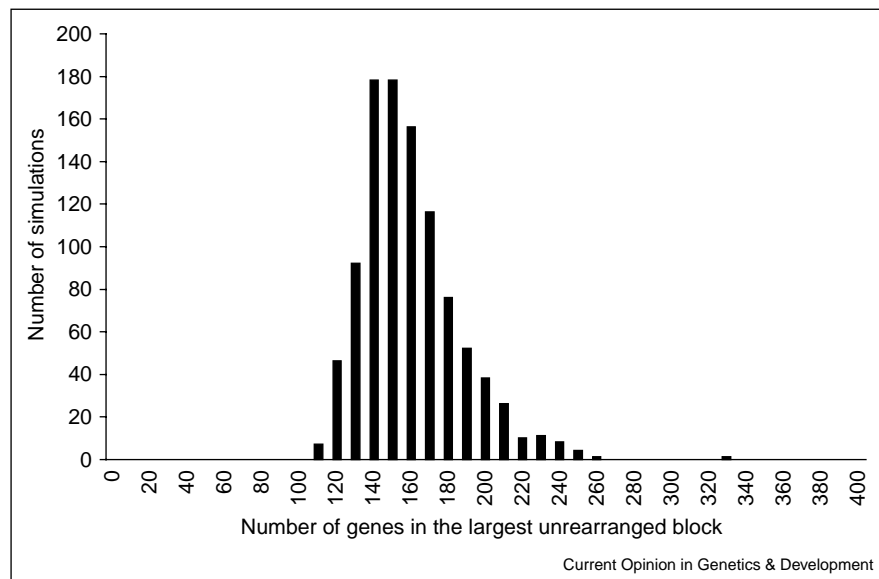
### Quadruplicated paralogous regions – the remnants of 2R?

The discovery of quadruplicated paralogous regions in the human genome provided some of the earliest evidence for the 2R hypothesis [1,18], although the origin of these regions continues to be debated [18,25–27]. Hughes, da Silva and Friedman [25] devised a parsimony test to determine whether large-scale duplication provides a good explanation of the quadruplicated regions containing the mammalian Hox clusters. The number of events (gene deletions and duplications) that would be

required to explain the duplicated genes in the Hox regions was compared under the competing hypotheses of large-scale duplications and tandem duplications of individual genes followed by translocations. Hughes, da Silva and Friedman [25] favoured the latter model, but proponents of the 2R hypothesis responded rapidly with a demonstration that the parsimony approach used failed to favour genome duplication even for the *Arabidopsis* duplicated segments that are widely accepted to have originated from genome duplication [26] and that the parsimony test was, therefore, not appropriate. A detailed analysis of the Hox cluster regions also supported an origin through segmental or whole-genome duplications rather than single-gene duplications [27]. The discovery and phylogenetic analysis of seven Hox clusters in zebrafish [28] and the fact that coelacanth appears to have just four [29\*] has led to the belief that a whole-genome duplication occurred in the ray-finned fishes after the divergence from the lobe-finned fishes [29\*,30]. This is further supported by the observation that many zebrafish paralogues appear to have originated in the ancestor of the ray-finned fishes [30].

Analysis of the human MHC paralogous regions has indicated that they were formed by successive *en-bloc* duplications that occurred between 766 and 528 My ago [31]. However, the high degree of macrorearrangement apparent from comparison of the human and mouse genomes — 242 macrorearrangements and a far larger number of microrearrangements [32\*] — suggests a rate of chromosomal rearrangement that should be inconsistent with conserved blocks of the size reported (the quadruplicated MHC regions contain ~1240 genes in total

Figure 2



The maximum expected number of genes in duplicated blocks on the human genome assuming chromosomal rearrangements occur at random intergenic positions. The graph was produced from 1000 simulations, assuming 40,000 human genes. The number of rearrangements after a putative genome duplication was assumed to be 500 and each rearrangement resulted in four breakpoints in the map of duplicated regions. The average number of genes between breakpoints was just 20.

[31]). If we assume, conservatively, that one macrorearrangement event occurred per My in the human lineage [32,33] and that rearrangements occur at random intergenic locations, then the average and the maximum number of genes between macrorearrangement breakpoints can be estimated from simulation (Figure 2). By the same reasoning, it seems unlikely that the largest paralogous region identified in the human genome by McLysaght, Hokamp and Wolfe [19] (a 41 Mb region on Chromosome I) was maintained after the proposed ancestral vertebrate genome duplication (again assuming random chromosomal rearrangements). If these large paralogous regions are the remnants of genome duplication early in the vertebrate lineage, it may be interesting to explore the reasons that they have been maintained intact for such a long period of time.

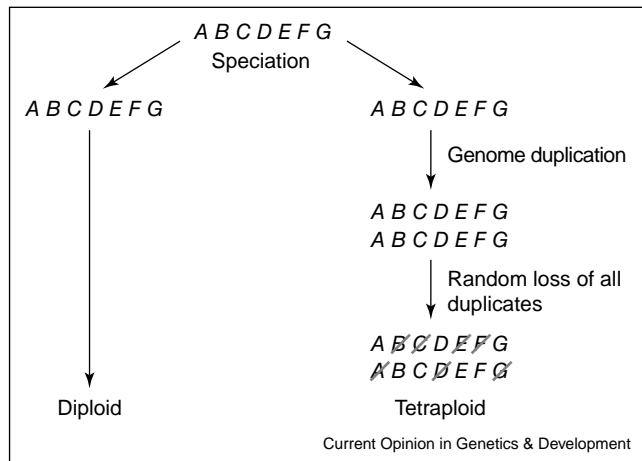
### Untangling successive rounds of duplication in *Arabidopsis*

Genome sequencing of *Arabidopsis* revealed large, partially duplicated chromosomal segments that are the likely remnants of ancient polyploidy events [6–8,9,11,12,34,35,36]. In apparently independent developments, two separate groups [9,11] interpolated gene order before the most recent genome duplication to search for more ancient duplicated blocks left over from earlier rounds of duplication. Bowers *et al.* [11] inferred three duplications at ~14.5–86, 170–235 and 300 My and labeled them  $\alpha$ ,  $\beta$  and  $\gamma$ , respectively. Blanc, Hokamp and Wolfe [9] were less specific about the more ancient

duplications and were content to infer that earlier duplications had occurred. Three ancient duplications in the *Arabidopsis* lineage have also been inferred on the basis of the number of copies of paralogous regions and detailed analysis of specific paralogous regions [35,36]. Subdivision of duplicated genes into age-classes had previously been used to infer that at least four large-scale duplication events have taken place in *Arabidopsis* in the past 200 My [6] but the methodology used — based on a single molecular clock applied to amino acid sequence divergence — has since been criticized [1,37].

The identification of the duplicated chromosomal segments that are indicative of genome duplication becomes increasingly difficult if the proportion of genes that are retained in duplicate is small [1]. However, it remains possible to infer a complete genome duplication because of its impact on gene order even if not one single gene has been maintained in duplicate (Figure 3; [38]). Duplicated segments in *Arabidopsis* that were previously undetectable because of extensive deletion of duplicated genes have been recovered by comparison to the completely sequenced rice genome [38]. Using an elegant method to display and analyze gene proximity in related yeast species, Wong, Butler and Wolfe *et al.* [5] were able to refine the map of *S. cerevisiae* duplicated regions significantly. Pebusque *et al.* [39] had previously used gene linkage in unduplicated genomes to infer large-scale duplications in human. In the same way, the 2R debate may eventually be resolved by comparing vertebrate gene

Figure 3



Gene order of seven genes (A–G) in a tetraploid and a diploid that diverged from a common ancestor before genome duplication in the tetraploid lineage. Despite the fact that no genes are maintained in duplicate in this example it is clear that the two chromosome segments in the tetraploid are related because they both contain genes that are on the same chromosomal segment in the diploid with conservation of gene order.

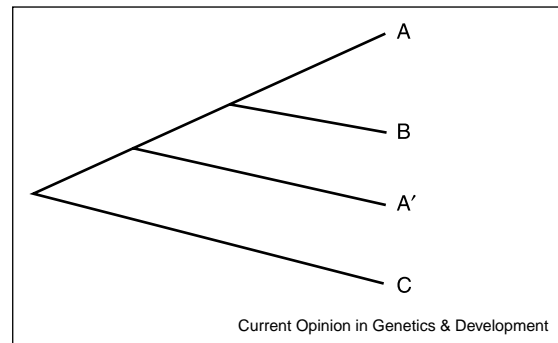
orders with the gene orders of invertebrates, even if only a small proportion of vertebrate gene pairs have been retained in duplicate.

### Dating ancient duplications

The dates at which genome duplications have occurred have been estimated from the divergence times of duplicate gene-pairs from chromosomal segments believed to originate from the duplication event. These dates, in turn, can be estimated under the assumption of a molecular clock. Date estimates that have applied a single molecular clock to amino acid sequences [6,40<sup>\*</sup>] are likely to be vulnerable to specifics of the protein sets that were used [1,37]. Date estimates using synonymous substitution rates are likely to be more accurate, although synonymous sites tend to be saturated in highly diverged sequences and synonymous substitution rates are more variable than previously believed [41]. Even when divergence times of duplicated genes can be estimated accurately, they provide only a lower limit on the date of genome duplication because they indicate only the date at which the progenitor genomes diverged. In the case of allopolyploids, this may differ significantly from the date of polyploidy.

Upper and lower bounds on the dates of ancient duplications have frequently been inferred from the phylogenetic tree topologies of duplicated gene-pairs and orthologues from related species [42] (Figure 4). These methods are not directly subject to violations of the molecular clock assumption, but they depend on the availability of orthologous sequences from related organisms that have not undergone duplication and the result-

Figure 4



Estimation of dates of gene duplication from phylogenies. Homologous genes from organism A, B and C are shown. Two copies of the gene (denoted A and A') are found in organism A. If the phylogenetic tree is accurate the duplication event occurred before organisms A and B diverged but after the divergence of A and C.

ing dates often correspond to very broad ranges and are only as good as the date estimates of the speciations that bound the duplication event. Recently, accurate estimates of the age of gene duplications have been derived by combining the strengths of the molecular clock and phylogenetic approaches [19<sup>\*\*</sup>,21<sup>\*\*</sup>].

### Implications for comparative gene order

If duplicated genes are lost randomly from the genome then genes that were adjacent before the duplication event may be on different chromosomes after duplication through differential gene loss [43]. This should be taken into account in estimates of chromosomal rearrangement distances between organisms if large-scale duplication events have occurred in either organism after their divergence. Wong, Butler and Wolfe [5<sup>\*\*</sup>] used the expected degree of disruption of adjacent genes between a paleopolyploid and an unduplicated related genome to infer that genome duplication in the *Saccharomyces cerevisiae* lineage post-dates the divergence of *S. cerevisiae* from *Kluyveromyces lactis*. Langkjaer *et al.* [44<sup>\*</sup>], however, have provided phylogenetic evidence that the duplication occurred before this speciation. The discrepancy could be explained by the presence of older duplicates within the yeast duplicated blocks as suggested by Friedman and Hughes [45], although it seems unlikely that a large proportion of between-block paralogues correspond to more ancient duplications in yeast. Another possible explanation is that, in addition to a genome duplication that preceded speciation, subsequent large-scale duplications, or possibly a whole genome duplication occurred in the *K. lactis* lineage. This would explain the relatively low levels of conserved adjacency [5<sup>\*\*</sup>] between *S. cerevisiae* and *K. lactis*.

### Modeling evolution after polyploidy

Because of the prevalence of recent polyploids among the plants and the ease with which synthetic polyploids can

be engineered, plants provide an ideal means of studying the immediate consequences of polyploidy (see [13] for review). Complex interactions between regulatory networks of the genomes in the case of allopolyploids, and to a lesser extent, also in the case of autopolyploids may lead to significantly altered patterns of gene expression in a newly formed polyploid [46]. Ten out of 40 genes from a study of allotetraploid cotton (*Gossypium hirsutum*) showed divergent expression patterns between copies derived from the two diploid progenitors [47\*]. This is in spite of the relatively recent date of tetraploidy in cotton (~1.5 My ago [48\*]). Gene silencing through methylation and rapid gene loss, possibly resulting from intragenomic recombination has been observed in synthetic allotetraploids [49,50]. These results are important for models of the evolution of genetic redundancy following genome duplication. The observation of differential expression between homeologues in synthetic allotetraploids [49,51] points to non-mutational mechanisms that can lead to complimentary degenerative subfunctionalization as proposed by Force *et al.* [52] and is relevant to estimations of the proportion of duplicated genes that are likely to be retained following genome duplication [53].

### Function, expression and sequence divergence of duplicate genes

Using microarray data and duplicated genes from *S. cerevisiae* Gu *et al.* [54] have found significant correlation between  $K_s$  (the number of synonymous substitutions per synonymous site) and expression divergence, although an earlier study [55] had not detected a correlation. The implication of this finding is that expression of duplicated genes continues to diverge in proportion to the time since duplication. Duplicated genes have also been shown to evolve more rapidly than unduplicated genes [56–58], probably because of functional redundancy, and the role of functional redundancy as well as positive selection in the evolution of novel protein functionality through gene duplication has recently been emphasized [59\*].

### Conclusions

Given the prevalence of polyploidy in a range of extant species it is not surprising that genome duplication has occurred and that traces of ancient duplications persist in organisms that are now diploid. Perhaps more surprising is the fact that genome duplications have occurred sufficiently rarely for their effects to be observed in relatively clearly delineated and non-overlapping duplicated regions in an organism such as yeast. Exploration of factors such as expression level and gene function that could determine the fate of duplicated genes [60] may benefit from comparison of the gene sets that have been retained in the successive rounds of genome duplication in *Arabidopsis*. The question of whether there have been genome duplications early in the vertebrate lineage remains difficult to resolve definitively, partly because these duplication events, if they did occur, are much more

ancient than the genome duplications that have been detected in other organisms. A putative genome duplication in the ancestor of the vertebrates could be over five times older than genome duplication in yeast and more than ten times older than the most recent genome duplication in *Arabidopsis*. Proving that a genome duplication as old as this has occurred is difficult and disproving it is even more so, but improved analysis and increasing amounts of data are bringing more ancient duplication events within reach.

### Update

Recently Papp, Pal and Hurst [61] have shown that changes in dosage of protein complex subunits tend to be deleterious. They suggest that genes coding for subunits of protein complexes are more likely to be retained in duplicate following a complete genome duplication event and less likely to be duplicated singly because of their sensitivity to dosage effects. As an example they cite the over-representation of ribosomal genes among genes that have been retained in duplicate following complete genome duplication in *Saccharomyces cerevisiae*. An excess of duplicated protein complex subunits among genes duplicated during a particular period could in the future be used as evidence of a genome-scale duplication event. In an unrelated development Vandepoele, Simillion and Van De Peer [62] question the suggestion by Goff *et al.* [40\*] that a genome-duplication event shaped the rice genome and instead suggest that duplicated segments in rice are more likely to have resulted from aneuploidy.

### Acknowledgements

I am grateful to Victoria Nembaware and Chris Gehring for critical comments and suggestions.

### References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
  - of outstanding interest
1. Wolfe KH: **Yesterday's polyploids and the mystery of diploidization.** *Nat Rev Genet* 2001, **2**:333-341.
  2. Ohno S: *Evolution by Gene Duplication*. New York: Springer Verlag; 1970.
  3. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**:708-713.
  4. Seoighe C, Wolfe KH: **Updated map of duplicated regions in the yeast genome.** *Gene* 1999, **238**:253-261.
  5. Wong S, Butler G, Wolfe KH: **Gene order evolution and paleopolyploidy in hemiascomycete yeasts.** *Proc Natl Acad Sci USA* 2002, **99**:9272-9277.
- A graphical representation of neighbouring genes in closely related organisms. These proximity plots were interpreted to indicate that genome duplication occurred in *S. cerevisiae* after divergence from *K. lactis* which is at odds with the findings of Langkjaer *et al.* [44\*].
6. Vision TJ, Brown DG, Tanksley SD: **The origins of genomic duplications in *Arabidopsis*.** *Science* 2000, **290**:2114-2117.
  7. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290**:1151-1155.

8. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**:796-815.
9. Blanc G, Hokamp K, Wolfe KH: **A recent polyploidy • superimposed on older large-scale duplications in the *Arabidopsis* genome**. *Genome Res* 2003, **13**:137-144.
- One of the first papers to use interpolation of pre-duplication gene order to discover duplicated segments corresponding to earlier rounds of duplication. The website (<http://wolfe.gen.tcd.ie/athal/dup>) that complements this paper provides a useful tool for exploring the duplicated blocks.
10. Ermolaeva MD, Wu M, Eisen JA, Salzberg SL: **The age of the *Arabidopsis thaliana* genome duplication**. *Plant Mol Biol* 2003, **51**:859-866.
11. Bowers JE, Chapman BA, Rong J, Paterson AH: **Unravelling • angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events**. *Nature* 2003, **422**:433-438.
- Interpolation of gene order before the most recent *Arabidopsis* genome duplication was used to infer two additional rounds of duplication. Duplications were dated using a phylogenetic methodology. Non-overlapping segments from independent rounds of genome duplication provide strong evidence of whole genome duplications in *Arabidopsis*.
12. Ku HM, Vision T, Liu J, Tanksley SD: **Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny**. *Proc Natl Acad Sci USA* 2000, **97**:9121-9126.
13. Wendel JF: **Genome evolution in polyploids**. *Plant Mol Biol* 2000, **42**:225-249.
14. Seoighe C, Wolfe KH: **Extent of genomic rearrangement after genome duplication in yeast**. *Proc Natl Acad Sci USA* 1998, **95**:4447-4452.
15. Hampson S, McLysaght A, Gaut B, Baldi P: **LineUp: statistical detection of chromosomal homology with application to plant comparative genomics**. *Genome Res* 2003, **13**:999-1010.
16. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.
17. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al.*: **The sequence of the human genome**. *Science* 2001, **291**:1304-1351.
18. Makalowski W: **Are we polyploids? A brief history of one hypothesis**. *Genome Res* 2001, **11**:667-670.
19. McLysaght A, Hokamp K, Wolfe KH: **Extensive genomic • duplication during early chordate evolution**. *Nat Genet* 2002, **31**:200-204.
- Numerous duplicated blocks in the human genome, containing as many as 29 duplicated genes, are reported. Through randomization it was shown that blocks containing more than five duplicated genes were very unlikely to have been created by individual mutations. Duplications were dated using the molecular clock and a phylogenetic approach. It would be interesting to see if the larger blocks, some of which may be the result of more recent segmental duplications, have earlier dates of divergence.
20. Panopoulou G, Hennig S, Groth D, Krause A, Poustka AJ, Herwig R, Vingron M, Lehrach H: **New evidence for genome-wide duplications at the origin of vertebrates using an *Amphioxus* gene set and completed animal genomes**. *Genome Res* 2003, **13**:1056-1066.
21. Gu X, Wang Y, Gu J: **Age distribution of human gene families • shows significant roles of both large- and small-scale duplications in vertebrate evolution**. *Nat Genet* 2002, **31**:205-209.
- An accurate method of dating duplication events appears to indicate that there has been a burst of duplication activity coinciding approximately with the time that the genome duplications should have occurred under the 2R model.
22. Spring J: **Genome duplication strikes back**. *Nat Genet* 2002, **31**:128-129.
23. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome**. *Science* 2002, **297**:1003-1007.
24. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM *et al.*: **The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins**. *Science* 2002, **298**:2157-2167.
25. Hughes AL, da Silva J, Friedman R: **Ancient genome duplications did not structure the human Hox-bearing chromosomes**. *Genome Res* 2001, **11**:771-780.
26. Gu X, Huang W: **Testing the parsimony test of genome duplications: a counterexample**. *Genome Res* 2002, **12**:1-2.
27. Larhammar D, Lundin LG, Hallbook F: **The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications**. *Genome Res* 2002, **12**:1910-1920.
28. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL *et al.*: **Zebrafish hox clusters and vertebrate genome evolution**. *Science* 1998, **282**:1711-1714.
29. Koh EG, Lam K, Christoffels A, Erdmann MV, Brenner S, Venkatesh B: **Hox gene clusters in the Indonesian coelacanth, *Latimeria menadoensis***. *Proc Natl Acad Sci USA* 2003, **100**:1084-1088.
- The authors report that the coelacanth genome has four Hox clusters that are analogous to the mammalian Hox clusters. This suggests that the proposed genome duplication in the ray-finned fishes took place after the divergence of the ray-finned and lobe-finned fishes.
30. Taylor JS, Braasch I, Frickey T, Meyer A, Van De Peer Y: **Genome duplication, a trait shared by 22000 species of ray-finned fish**. *Genome Res* 2003, **13**:382-390.
31. Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H: **Evidence of en bloc duplication in vertebrate genomes**. *Nat Genet* 2002, **31**:100-105.
32. Pevzner P, Tesler G: **Genome rearrangements in mammalian • evolution: lessons from human and mouse genomes**. *Genome Res* 2003, **13**:37-45.
- The number of rearrangements that have occurred since human and mouse diverged is higher than previously thought. This has implications for the expected lengths of paralogous regions that have been maintained after ancient polyploidy.
33. Nadeau JH, Taylor BA: **Lengths of chromosomal segments conserved since divergence of man and mouse**. *Proc Natl Acad Sci USA* 1984, **81**:814-818.
34. Blanc G, Barakat A, Guyot R, Cooke R, Delseny M: **Extensive duplication and reshuffling in the *Arabidopsis* genome**. *Plant Cell* 2000, **12**:1093-1101.
35. Ziolkowski PA, Blanc G, Sadowski J: **Structural divergence of chromosomal segments that arose from successive duplication events in the *Arabidopsis* genome**. *Nucleic Acids Res* 2003, **31**:1339-1350.
36. Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van De Peer Y: **The hidden duplication past of *Arabidopsis thaliana***. *Proc Natl Acad Sci USA* 2002, **99**:13627-13632.
- The authors show how additional duplicated blocks can be inferred by careful analysis of a genome that has undergone successive rounds of ancient genome duplication. They infer that three, and probably no more than three, genome duplications have occurred in *Arabidopsis*.
37. Raes J, Vandepoele K, Simillion C, Saeys Y, Van De Peer Y: **Investigating ancient duplication events in the *Arabidopsis* genome**. *J Struct Funct Genomics* 2003, **3**:117-129.
38. Vandepoele K, Simillion C, Van De Peer Y: **Detecting the • undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice**. *Trends Genet* 2002, **18**:606-608.
- A demonstration of the use of comparative genomics to uncover additional duplicated regions in a paleopolyploid genome, in this case the genome of *Arabidopsis* by comparison with the rice genome, but the approach is generally applicable.
39. Pebusque MJ, Coulier F, Birnbaum D, Pontarotti P: **Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution**. *Mol Biol Evol* 1998, **15**:1145-1159.
40. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*)**. *Science* 2002, **296**:92-100.

Evidence for a whole-genome duplication in rice is presented in the supplementary data, but no proof of genome duplication is provided. Date estimates for a duplication event do not take account of the fact that paralogues may evolve more rapidly than orthologues and a single value for the rate of amino acid sequence divergence is used.

41. Zhang L, Vision TJ, Gaut BS: **Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana***. *Mol Biol Evol* 2002, **19**:1464-1473.
42. Kellogg EA: **Genome evolution: it's all relative**. *Nature* 2003, **422**:383-384.
43. Lundin LG: **Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse**. *Genomics* 1993, **16**:1-19.
44. Langkjaer RB, Cliften PF, Johnston M, Piskur J: **Yeast genome duplication was followed by asynchronous differentiation of duplicated genes**. *Nature* 2003, **421**:848-852.  
The date of genome duplication was estimated using a phylogenetic approach and is at odds with what has been proposed by Wong, Butler and Wolfe [5\*\*] using gene order data. The resolution of this disparity will be a test of the different methods that were used.
45. Friedman R, Hughes AL: **Gene duplication and the structure of eukaryotic genomes**. *Genome Res* 2001, **11**:373-381.
46. Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V, Martienssen RA: **Understanding mechanisms of novel gene expression in polyploids**. *Trends Genet* 2003, **19**:141-147.
47. Adams KL, Cronn R, Percifield R, Wendel JF: **Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing**. *Proc Natl Acad Sci USA* 2003, **100**:4649-4654.  
Organ-specific changes in gene expression were found to be consistent between different related species that shared a common allotetraploidy ~1.5 My ago [48\*]. This was interpreted as evidence that the major part of the differentiation in gene expression between 'homeologous' gene-pairs that the authors have observed either occur at the onset of polyploidization or very shortly afterwards.
48. Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF: **Rate variation among nuclear genes and the age of polyploidy in gossypium**. *Mol Biol Evol* 2003, **20**:633-643.  
The estimated date of divergence of one of the constituent genomes of tetraploid cotton from its closest living diploid relative is given (1.5 My). This provides an upper-limit on the date at which allopolyploid cotton was formed.
49. Kashkush K, Feldman M, Levy AA: **Gene loss, silencing and activation in a newly synthesized wheat allotetraploid**. *Genetics* 2002, **160**:1651-1659.
50. Song K, Lu P, Tang K, Osborn TC: **Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution**. *Proc Natl Acad Sci USA* 1995, **92**:7719-7723.
51. Chen ZJ, Pikaard CS: **Transcriptional analysis of nucleolar dominance in polyploid plants: biased expression/silencing of progenitor rRNA genes is developmentally regulated in Brassica**. *Proc Natl Acad Sci USA* 1997, **94**:3442-3447.
52. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 1999, **151**:1531-1545.
53. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization**. *Genetics* 2000, **154**:459-473.
54. Gu Z, Nicolae D, Lu HH, Li WH: **Rapid divergence in expression between duplicate genes inferred from microarray data**. *Trends Genet* 2002, **18**:609-613.
55. Wagner A: **Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate**. *Proc Natl Acad Sci USA* 2000, **97**:6579-6584.
56. Nembaware V, Crum K, Kelso J, Seoighe C: **Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs**. *Genome Res* 2002, **12**:1370-1376.
57. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications**. *Genome Biol* 2002, **3**:RESEARCH0008.
58. Seoighe C, Johnston CR, Shields DC: **Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation**. *Mol Biol Evol* 2003, **20**:484-490.
59. Zhang J, Zhang YP, Rosenberg HF: **Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey**. *Nat Genet* 2002, **30**:411-415.  
The authors provide an outstanding example of the evolution of a protein with novel properties through gene duplication. One copy of the duplicated digestive enzyme studied underwent rapid positive selection in response to a change in diet while the other copy remained unchanged to perform the remaining functions of the protein. This work has implications for the dual roles of relaxed constraint and positive selection in the evolution of novel functionality through gene duplication.
60. Seoighe C, Wolfe KH: **Yeast genome evolution in the post-genome era**. *Curr Opin Microbiol* 1999, **2**:548-554.
61. Papp B, Pal C, Hurst LD: **Dosage sensitivity and the evolution of gene families in yeast**. *Nature* 2003, **424**:194-197.
62. Vandepoele K, Simillion C, Van De Peer Y: **Evidence that rice and other cereals are ancient aneuploids**. *Plant Cell* 2003, **15**:2192-2202.