YESTERDAY'S POLYPLOIDS AND THE MYSTERY OF DIPLOIDIZATION

Kenneth H. Wolfe

Thirty years after Susumu Ohno proposed that vertebrate genomes are degenerate polyploids, the extent to which genome duplication contributed to the evolution of the vertebrate genome, if at all, is still uncertain. Sequence-level studies on model organisms whose genomes show clearer evidence of ancient polyploidy are invaluable because they indicate what the evolutionary products of genome duplication can look like. The greatest mystery is the molecular basis of diploidization, the evolutionary process by which a polyploid genome turns into a diploid one.

SYNTENY
The property of being located on the same chromosome.

ANEUPLOIDY
Presence of extra copies, or no copies, of some chromosomes.

HOX GENE CLUSTERS
Tandem arrays of homeobox genes that have crucial roles in development. There are four Hox clusters in humans but only one in invertebrates.

Department of Genetics, Smurfit Institute, Trinity College, University of Dublin, Dublin 2, Republic of Ireland. e-mail: khwolfe@tcd.ie

No one would expect a 30-year-old book that is mostly about fish cytogenetics to be of much interest to modern molecular biologists, particularly if that book had received lukewarm reviews at the time of publication[1,2]. Nevertheless, citations of Susumu Ohno's book *Evolution by Gene Duplication*[3] have tripled between the years 1990 and 2000. In this book, written when only a few protein sequences were known, Ohno proposed that it is much easier to make new genes by duplicating old ones than to create them *de novo*, and that genome duplication (polyploidy) was a quick and easy way to produce vast numbers of duplicate genes. Genome duplication could open the door to duplicating whole biochemical pathways. He famously proposed that two (or possibly three) rounds of polyploidy had occurred during the early evolution of the vertebrate lineage, but that further polyploidization then became impossible in mammals owing to the emergence of the X/Y sex-chromosome system. The human genome would thus be a paleopolyploid: an ancient polyploid that had later become diploid again, by means of sequence divergence between the duplicated chromosomes.

The renewed interest in Ohno's ideas stems from two lines of research that began to bear fruit in the late 1980s. The first was what is now called comparative genomics. Genetic map comparisons among mammals confirmed that they contain large segments[4] of conserved SYNTENY with conserved gene order. As early as 1973, Ohno had identified an apparently duplicated chromosomal segment within the human genome, which was delineated by two pairs of duplicated genes on chromosomes 11 and 12 (REF. 5). His proposal that these segments were remnants of ancient polyploidy (or some other form of ANEUPLOIDY) meant that the development of comparative genetic maps between mammalian genomes went hand-in-hand with a search for duplicated regions within them (for example, REF. 6). The second line of research began with the discovery that the four HOX GENE CLUSTERS in mammals had evolved by quadruplication of a prototypic cluster similar to that of *Drosophila*. Schughart *et al.*[7] suggested that this quadruplication could have been associated with polyploidizations of the type envisaged by Ohno. Not only were the orders of the Hox genes in each cluster conserved between human and mouse, but also the gene order was essentially conserved among the four mammalian clusters. The Hox clusters are a quadruplicated chromosomal segment. Subsequent discoveries of other duplicated genes that were linked to the Hox clusters indicated that the duplicated chromosomal regions might be quite large[8].

Ohno's book was not very explicit about the number and timing of the proposed genome duplications, but the most widely accepted form of the hypothesis, which has been called the 2R hypothesis[9], is that there were two rounds of genome duplication in vertebrate ancestry: one immediately before, and one immediately after, the divergence of the lamprey lineage (see FIG. 1 and REF. 10).

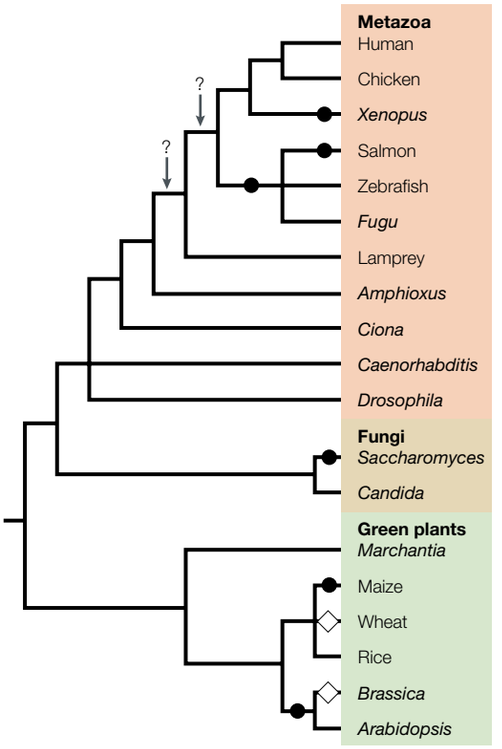


Figure 1 | **Phylogenetic positions of some likely polyploidy events during eukaryote evolution.** Filled circles mark lineages in which genome duplication (AUTO- or ALLOTETRAPLOIDY) has been inferred; open diamonds mark two hexaploid lineages of plants. The question marks show the positions of the two rounds of genome duplication proposed under the 2R hypothesis.

The so-called one-to-four rule[11,12], which states that genes from invertebrates, such as *Drosophila,* should have four ORTHOLOGUES in vertebrates, is often said to be a corollary of the 2R hypothesis, but this depends on the extent to which genes are deleted after each round of duplication (FIG. 2); the only gene families having a one-to-four relationship of the expected type will be those in which no gene deletions or other (non-polyploidy) gene duplications have occurred.

The question of whether the 2R hypothesis for vertebrates is correct has recently become quite controversial, with evo-devo (evolution of development) researchers (for example, REF. 13) tending to favour the hypothesis, and molecular phylogeneticists (for example, REF. 14) tending to dispute it. Consequently, the literature includes some authors who accept the hypothesis and have gone ahead to estimate the dates of the polyploidizations[15] or the extent of gene loss after each round of duplication[16], whereas other authors are still questioning whether the 2R hypothesis is correct at all[9,10,14,17]. It might be hoped that a complete human genome sequence would resolve the question of whether genome duplications have occurred in the vertebrate lineage (and if so, how often and when), and that we might be able to learn something from model organisms. Even in completely sequenced models, however, there has not been complete consensus about how their genomes have evolved. The number of possible genome duplications and their timing has been disputed in both yeast (BOX 1) and *Arabidopsis thaliana* (BOX 2).

In part, the different interpretations of vertebrate genomes by different workers are due to the variety of analytical approaches used. These approaches can be categorized as either map based or tree based (see below), or in some cases a combination of the two. One of the principal difficulties in testing the 2R hypothesis is the lack of consensus about its predictions[10]. Although it was originally assumed that having more sequence data would resolve the 2R question, it is now clear that the problem lies as much on the data interpretation side ("What should the sequence of a paleopolyploid genome look like?") as on the data acquisition side. This point is illustrated keenly by the differing interpretations of the *Arabidopsis* genome that are offered by different groups (BOX 2). To get some idea of what to expect in a paleopolyploid vertebrate, it is worth studying other eukaryotes in which there is (relatively) clear evidence of polyploidy, and to examine the success of map-based and tree-based approaches in untangling their history. I will compare the results from mammals with those from four other eukaryotes: yeast, *Arabidopsis*, maize and zebrafish. My perspective is perhaps unique in that my work has been criticized by paleopolyploidy sceptics in the case of yeast[18] and by paleopolyploidy proponents in the case of vertebrates[19].

Map-based approaches

The map-based approach to identifying ancient polyploids is to study the chromosomal locations of duplicated genes, looking for chromosomes (or sections of chromosomes) that can be paired up because they contain sets of similar genes. Ideally, the duplicated genes should be in the same order on the two chromosomes. The results from this method alone were sufficient to justify strong arguments in favour of polyploidy in the completely sequenced genomes of *Saccharomyces cerevisiae* (BOX 1) and *Arabidopsis* (BOX 2). Neither of these species was suspected to be a polyploid before their genomes were sequenced; indeed, they were chosen as model organisms for genome projects on the basis of having compact genomes. In both yeast and *Arabidopsis*, paired chromosomal regions can be identified that cover more than half the genome. Neither species contains significant triplicated regions, at least in some analyses. These observations point to a single identifiable polyploidy event during the evolution of each of these species. Inter- and intrachromosomal rearrangements later broke up entire duplicated chromosomes into smaller duplicated segments.

Using a map-based approach to identify duplicated chromosomal regions in a genome conceptually amounts to finding the significant diagonals in a dot-matrix plot of BLAST hits when the genome is compared with itself. Finding diagonals can be quite difficult because there are two crucial unknown parameters: the extent to which duplicated genes are later deleted, and the extent to which the order of genes

AUTOPOLYPLOIDY
Doubling the copy number of each chromosome in a species.

ALLOPOLYPLOIDY
The fusion of two distinct parental species to form a hybrid, the genome of which is the sum of the two parental genomes.

ORTHOLOGUES
Homologous genes that originated through speciation (for example, human α-globin and mouse α-globin).

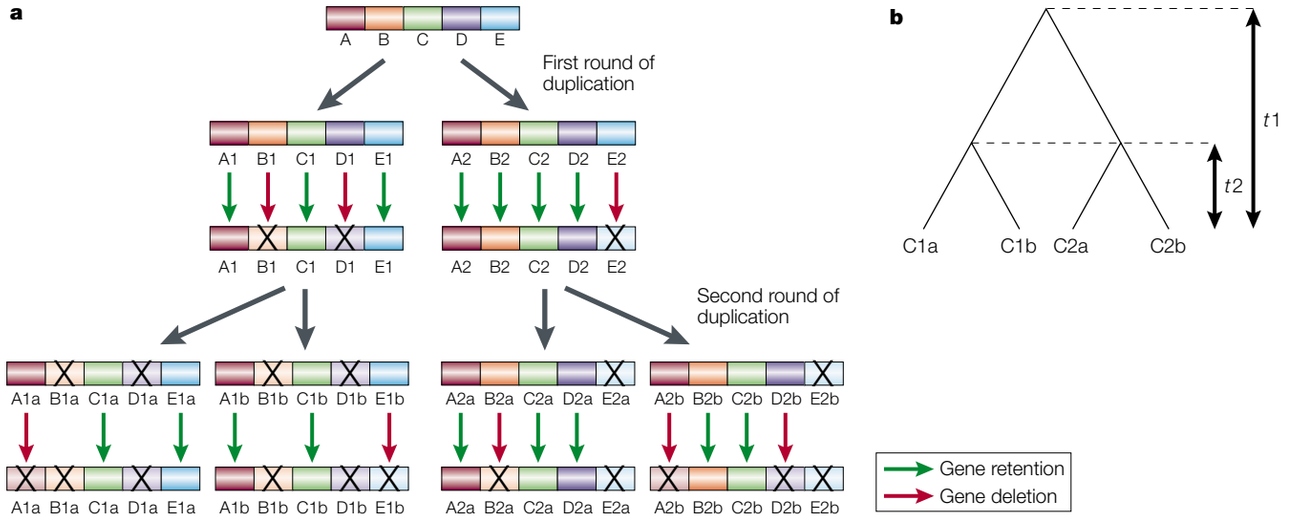
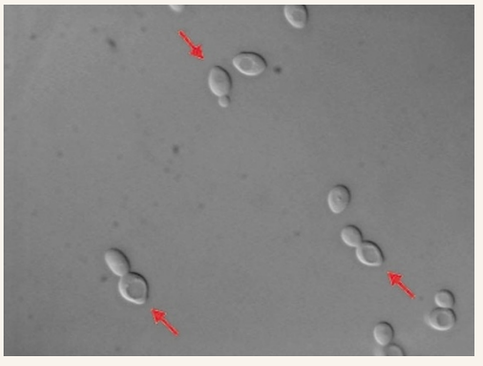


Figure 2 | **Model of the 2R hypothesis and its phylogenetic implications. a** | How to turn five genes into nine by a 2R model. A–E represent five genes on a chromosome in an ancestral organism. At least one copy of each ancestral gene (A–E, top line) is present at all times, but only the descendants of gene C in the paleotetraploid (bottom line) obey the one-to-four rule. **b** | Phylogenetic tree expected from analysis of the four copies of gene C. The branch lengths *t*1 and *t*2 should be proportional to the ages of the two rounds of duplication. The ratio *t*2/*t*1 should be the same whether calculated using the comparison of C1a with C1b, or C2a with C2b. If autopolyploidy has occurred, this ratio should also be constant across the genome for all genes present in four copies.

Box 1 | **Paleopolyploidy in the *Saccharomyces cerevisiae* genome**

Analyses of the yeast genome sequence indicated that it contained duplicated chromosomal regions, in which a group of genes on one chromosome had a group of homologues on another chromosome[32,36,48,49] (FIG. 3). Of yeast's 5,800 genes, ~900 are members of duplicated gene pairs located in duplicated chromosomal regions[50]. Many of these gene pairs have important functions, and are likely to cause significant differences between the physiology of *Saccharomyces cerevisiae* and other yeasts in which the genes are not duplicated. For example, the duplicated proteins Pip2 and Oaf1 are transcription factors of the Zn_2Cys_6 zinc-finger family with roles in the regulation of peroxisomal β-oxidation[51]. Usually, members of this family homodimerize to bind to DNA, but Pip2 and Oaf1 form a heterodimer. The *PIP2* and *OAF1* genes are regulated differently, and only the Oaf1 protein responds to the inducer molecule oleate[51]. These complexities of regulation cannot exist in other yeast species where *PIP2* and *OAF1* are not duplicated and only a homodimer is possible.

The structure of the *S. cerevisiae* genome has been interpreted to be a paleopolyploid[32,52]. Other groups[48,49], including the recent Génolevures project[18,53], have suggested instead that the various duplicated chromosomal regions (blocks) were produced independently at different times by 'segmental' duplications of parts of the chromosomes. The key question is whether the blocks duplicated simultaneously or not. Two pieces of evidence are difficult to reconcile with any model other than polyploidy[32]. First, the blocks do not overlap with one another; under a model of multiple independent duplications one would expect regions that had been duplicated to sometimes become duplicated again, producing three or more copies of the region. This is never seen in *S. cerevisiae* (except at the telomeres, which all viewpoints agree to be a special case). Second, the two copies of most blocks (50 out of 55 of them) have the same centromere-to-telomere orientation. This is compatible with the break-up of duplicated chromosomes into smaller blocks by reciprocal translocation between chromosomes, as envisaged by a polyploidy model[41,54], but to reconcile this observation with a segmental duplication model it is necessary postulate that when sections of chromosome are copied to remote sites in the genome they preferentially integrate (or preferentially survive) a particular way around. One model of genome evolution[18], which proposes independent duplications of chromosomal segments and insertion of the duplicated segment in a random orientation relative to centromeres, does not explain either of these observations. Moreover, the discovery that genes frequently become inverted during ascomycete chromosomal evolution[18,55], and the finding that gene family sizes in other yeasts are similar to those in *S. cerevisiae*[53], are fully compatible with the paleopolyploidy model.



Friedman and Hughes[36] recently made an independent search for duplicated regions of the yeast genome and found 28 blocks in which the genes had uniformly high levels of divergence at sites of silent substitution, which is consistent with paleopolyploidy. Their report that some (or all) genes in 11 other blocks had low synonymous site divergence, and so seemed to be duplicated more recently than the rest of the genome, is almost completely attributable to the presence of telomere-linked genes in these blocks (telomere-proximal sequences are peculiar as they tend to homogenize owing to dynamic DNA-exchange processes[56]). When blocks adjacent to telomeres are omitted, only four gene pairs out of the 280 studied by Friedman and Hughes[36] show anomalous levels of synonymous site divergence.

DICOT
The larger subclass of angiosperms that has two seed leaves (cotyledons) in the embryo.

along chromosomes has become scrambled after polyploidization. The higher these two quantities, the harder it will be to detect evidence of ancient polyploidy. The significant diagonals were immediately apparent in the yeast genome and *Arabidopsis* (FIG. 3), owing to the relatively large number of polyploidy-derived duplicate genes (called OHNOLOGUES[20]) and the relatively low background of other PARALOGOUS hits. In the working draft sequence of the human genome[21,22], diagonals in dot-matrix plots are much harder to discern, and algorithms designed to search for duplicated genomic regions are faced with the problem of distinguishing between mere noise and long, sparse diagonals that might be highly degraded duplicated regions. Moreover, the pattern of diagonals predicted under the 2R hypothesis for vertebrates (as opposed to the single round of duplication suggested for yeast and *Arabidopsis*) becomes complex after only a small number of interchromosomal rearrangements (FIG. 4).

Papers that have reported map-based analyses of mammalian genomes have generally been supportive of the 2R hypothesis. Among the first of these were Nadeau[6] and Lundin[23]. These studies analysed the locations of duplicated genes, or genes that obey the one-to-four rule, looking for chromosomes (or sections of chromosomes) that could be paired up because they contained similar sets of genes. In general, the resolution of gene mapping was only to the level of cytogenetic bands, so gene order could not be studied in detail. This type of research led to the discovery of several potentially quadruplicated regions, notably on human chromosomes (HSA) 1/6/9/19 (REFS 24,25), HSA 4/5/8/10 (REFS 26,27) and HSA 1/2/8/20 (REF. 19), as well as the Hox chromosomes HSA 2/7/12/17 (REF. 8). In each of these regions, at least five unrelated genes have copies on at least three of the four chromosomes. As well as these quadruplicated regions, many map-based studies (for example, REFS 23,28) identified apparently duplicated chromosomal regions in mammalian genomes and indicated that these might be remnants of polyploidy. It has been difficult to evaluate the statistical significance of the proposed duplicated or quadruplicated regions because, with the exception of the Hox clusters, the chromosomal regions involved are quite large, but (at least until very recently[21,22]) only a small fraction of the genes in those regions had been identified and the gene orders were unknown. As pointed out elsewhere[10], the presence of a few HOMOLOGOUS gene pairs on two human chromosomes is not sufficient to indicate that the chromosomes are related by a large duplication, given the large sizes of chromosomes and the prevalence of multigene families.

Tree-based approaches

Phylogenetic trees can be used to test paleopolyploidy hypotheses because genes that were duplicated simultaneously should betray the same history. All the gene pairs that make up a duplicated chromosomal segment should be the same age, and, if a single round of polyploidy occurred, all the duplicated segments in a genome should be the same age. Under the 2R hypothesis for vertebrates, each set of four genes should give a particular TOPOLOGY — called (AB)(CD) — in a phylogenetic tree (FIG. 2), and the estimates of the ratio between the ages of the younger and older rounds of duplication should be consistent both within trees (each tree yields two estimates of this ratio; FIG. 2) and among trees (drawn from different genes). These predictions allow several different tree-based tests of paleopolyploidy hypotheses.

Box 2 | **Paleopolyploidy in the *Arabidopsis thaliana* genome**

Like yeast, the *Arabidopsis* genome contains many duplicated regions (FIG. 3). There has been disagreement as to whether these regions are the remnants of a single polyploidy event, multiple successive polyploidies or multiple independent segmental (parts of chromosomes) duplications. One of the first publications compared the sequence of a tomato bacterial artificial chromosome (BAC) clone with the *Arabidopsis* genome, and concluded that two rounds of polyploidy had occurred in *Arabidopsis* at ~112 and ~180 Myr ago[57]. The more recent polyploidization occurred after the *Arabidopsis* lineage had diverged from the tomato lineage, whereas the older one could have happened in their DICOT ancestor. When the complete sequence of the *Arabidopsis* genome was later published, dot-matrix plots showed that much of the genome (except for the centromeric regions) fell into pairs[42,58]. These plots provide compelling evidence in support of one polyploidy event, and the authors (see link to the *Arabidopsis* Genome Initiative (AGI))[42] suggested that this was the more recent of the two polyploidy events proposed by Ku *et al.*[57]. However, the plots do not provide any sign of the proposed older event.

A significantly different interpretation of the *Arabidopsis* genome was made by Vision and colleagues[34]. They found 103 duplicated chromosomal regions (blocks). These include many overlapping ones, which are indicative of multiple duplication events at different times. This result contradicts the AGI's report[42] that there are no triplicated regions in the *Arabidopsis* genome; the difference must lie in the details of the block-finding methods used by the two groups. Vision *et al.* used a MOLECULAR CLOCK method to estimate the ages of the 103 duplicated blocks they found, and proposed that most of the blocks fell into four age classes, each potentially corresponding to a polyploidization[34]. This part of their analysis must be questioned because it relies on the assumption that the gene pairs making up each of the 103 blocks have the same average rate of protein-sequence evolution, whereas a fundamental observation in molecular evolution is that different proteins evolve at different rates[59]. So, the 'oldest' duplicated blocks might instead just be those that happen to contain the genes with the fastest rates of evolution, and hence the most divergent sequences.

An independent molecular-clock analysis of duplicated genes by Lynch and Conery[60], although it did not consider the chromosomal locations of the genes, found that *Arabidopsis* is unusual among eukaryotes in having a cohort of duplicated genes that all seemed to be approximately the same age (65 Myr). This result is consistent with a single round of polyploidy at that time. Lynch and Conery's age estimates for each gene pair were based on the assumption that all genes have the same synonymous nucleotide substitution rate — a much more reasonable assumption than the use of a single nonsynonymous rate for all groups of genes in both papers by Vision and colleagues[34,57]. The balance of evidence indicates that a single, large-scale duplication event, probably a polyploidization 65 Myr ago, is the dominant feature of the genome's visible history. Determining whether other, older, block duplications also occurred in *Arabidopsis* will require more thorough analysis using phylogenetic trees and sequences from an OUTGROUP, such as rice.

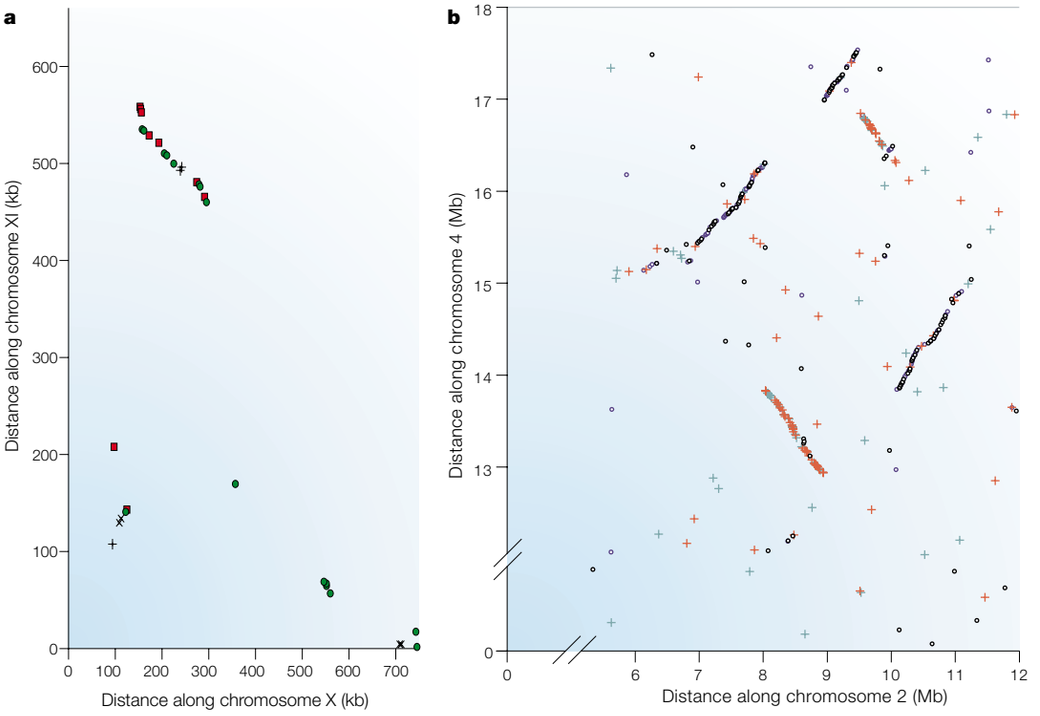


Figure 3 | **Dot-matrix plots of yeast and *Arabidopsis*. a** | Yeast chromosome X compared with chromosome XI (from REF. 32); **b** | *Arabidopsis* chromosome 2 compared with chromosome 4. Each dot represents a single BLASTP hit between proteins that are encoded on the two chromosomes, plotted at the positions of the corresponding genes. Diagonals indicate groups of genes that form duplicated chromosomal segments. Colours and symbols denote different transcriptional orientations of genes.

MOLECULAR CLOCK
The principle that any gene or protein has a near-constant rate of evolution in all organisms, which means that the amount of sequence divergence between two sequences will be proportional to the amount of time elapsed since their shared ancestor existed.

OUTGROUP
A species or sequence that is known to diverge earlier than the other species or sequences being analysed.

OHNOLOGUE
A pair of duplicate genes (paralogues) produced by the process of genome duplication.

PARALOGUES
Homologous genes that originated by gene duplication (for example, human α-globin and human β-globin).

HOMOLOGUES
Genes that share a common ancestor and are usually similar in sequence.

TOPOLOGY
The branching arrangement of a phylogenetic tree.

TELEOST
Bony fish.

TETRAPOD
Four-limbed animal.

In contrast to the map-based results, several tree-based studies have reached negative conclusions about the 2R hypothesis. Hughes[9] applied the (AB)(CD) topology test to a set of nine vertebrate gene families that had been proposed by Sidow[29] to exemplify the one-to-four rule. Only one of the nine gave the expected topology. A similar conclusion was reached by Martin[17] from a set of 35 gene families, and by the International Human Genome Sequencing Consortium[21] from 57 gene families that were in 4:1:1 ratios among human, *Drosophila* and *Caenorhabditis elegans*.

Analyses by Hughes and others[30,31] of gene sets that make up the HSA 1/6/9/19 (major histocompatibility complex (MHC)) and 2/7/12/17 (Hox) regions also did not support the 2R hypothesis. Some of the gene pairs (or triples or quadruples) that are located in these regions are much older than others. Approximate duplication dates for most of these genes could be estimated by reference to the positions of sequences from other organisms, such as *Drosophila* or yeast, on the same trees. This means that the conclusion from these studies, that many of the genes in the MHC and Hox regions are far too old or far too young to be products of genome duplications during early vertebrate evolution, do not depend on any assumption that the molecular clock has been well behaved, which is a criticism often levelled at tree-based estimates of dates. For example, the duplication that gave rise to the cyclin-dependent kinase 7 (*CDK7*) and 3 (*CDK3*) genes, located on the Hox cluster chromosomes HSA2 and HSA17, respectively, occurred before the animal–plant–fungal divergence (over 1,000 Myr ago) and so could not have happened simultaneously with the Hox cluster duplications (about 600 Myr ago)[31].

One tree-based study that supported the 2R hypothesis was that of Gibson and Spring[19]. Although the (AB)(CD) topology was seen in only two of the five genes they examined in a potentially quadruplicated region on HSA 1/2/8/20, they argued that some of the other trees were not significantly different from this topology. They also raised the interesting point that if two rounds of genome duplication happened within a relatively short interval (they suggest 10 Myr), so that the second round occurred before diploidization of the first round was complete, the prediction that all gene families will have an (AB)(CD) topology is not valid.

Wang and Gu[15] used phylogenetic trees in an attempt to estimate the dates of genome duplications in vertebrates. Their study did not test the 2R hypothesis but instead assumed it to be true. They examined 26 three-gene families with invertebrate outgroups, and used the molecular clock to estimate the dates of the two gene duplications that produced these families. The gene families in this study were chosen on the basis of having three members. In addition, the study was based on both duplications occurring within the interval between the TELEOST–TETRAPOD divergence at 430 Myr ago and the *Drosophila* (or *C. elegans*)–vertebrate divergence

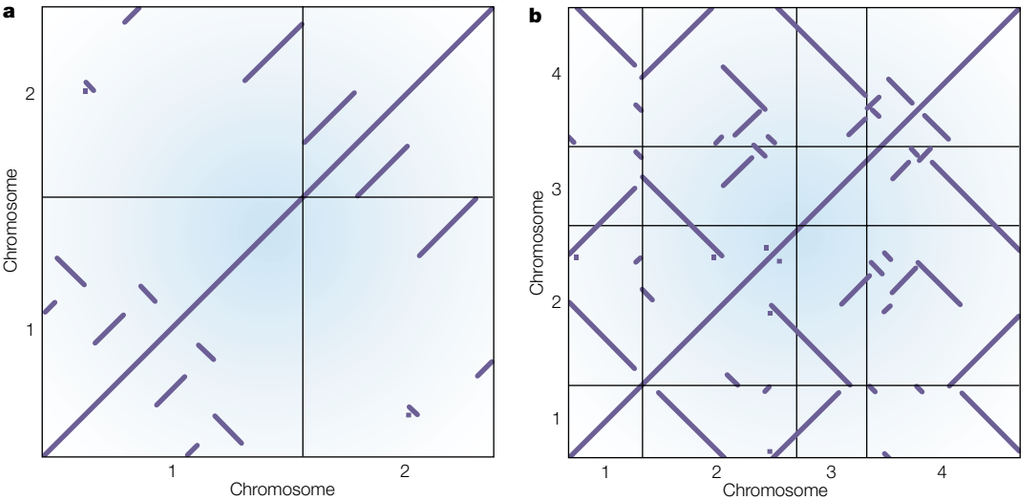


Figure 4 | **Computer simulations of the dot-matrix patterns expected from one or two rounds of genome duplication. a** | One round of genome duplication, and **b** | two rounds of genome duplication; in both cases, after a small number of chromosomal rearrangements and assuming no gene deletion. In **a**, a chromosome containing 200 genes was duplicated, and five interchromosomal translocations were made; each gene appears twice in the genome. In **b**, a chromosome containing 100 genes was duplicated twice, and then five interchromosomal translocations were made; each gene appears four times in the genome.

at 830 Myr ago. This subjective filtering of the data makes Wang and Gu's reasoning circular; the average duplication dates they obtained (488 and 594 Myr ago) were obliged to fall somewhere in the 430–830 Myr ago interval. Moreover, there was no consistent ratio between the estimated first and second duplication dates in the various gene families they studied. In their study of 23 four-gene families, only 5 showed an (AB)(CD) topology.

Tree-based, or molecular-clock-based, approaches to dating gen(om)e duplications in model organisms have given surprisingly poor results, which perhaps calls into question the use of these approaches to reject the 2R hypothesis in vertebrates. In yeast, the lack of good outgroup sequences meant that duplication dates were estimated for only 12 of the 376 gene pairs studied by Wolfe and Shields[32]. These dates were quite variable, which was attributed to resetting of the molecular clock by GENE CONVERSION or some other form of sequence homogenization between the different copies of duplicate genes. Recent analysis of a larger set of yeast gene pairs with *Candida albicans* as an outgroup has produced results more in line with expectation[33]. In *Arabidopsis*, Vision *et al.*[34] made some sweeping assumptions about the molecular clock to produce age estimates for each of the 103 duplicated genomic segments they identified, which they interpreted as indicating that many rounds of polyploidy have occurred in this species (see discussion in BOX 2).

There are several phenomena that could potentially limit the ability of tree-based approaches to detect genome duplications, by throwing up unanticipated phylogenetic results for some genes even where genome duplications have occurred. First, phylogenetic trees are never more than an estimate of history and it is important not to infer conclusions from trees that do not have strong BOOTSTRAP support. Second, gene pairs that seem to be ohnologues on the basis of their chromosomal locations, but that seem to be much older than neighbouring gene pairs in the same duplicated block, could be explained by the presence of tandemly repeated genes in an ancestor followed by loss of different members of the tandem array in different daughter lineages[35,36]. Third, gene conversion (for example, between genes *C1a* and *C2b* in FIG. 2b) could produce misleading topologies in a 2R situation, as well as misleading dates in 1R or 2R situation. Last, the mechanics of diploidization are important, as discussed below.

GENE CONVERSION
Non-reciprocal allelic exchange.

BOOTSTRAP ANALYSIS
Type of statistical analysis to test the reliability of certain branches in the evolutionary tree. The bootstrap proceeds by re-sampling the original data, with replacement, to create a series of bootstrap samples of the same size as the original data. The bootstrap value of a node is the percentage of times that node is present in the set of trees constructed from the new data sets.

QUADRIVALENT
A cytological structure in which four copies of a chromosome are aligned on the meiotic spindle.

BIVALENT
A cytological structure in which two copies of a chromosome are aligned on the meiotic spindle.

Diploidization and segmental allotetraploidy

Genome duplication can occur either through autopolyploidy or allopolyploidy. Spring[28] has argued that allopolyploidy of vertebrates is a more likely scenario than autopolyploidy. If vertebrates are paleo-allopolyploids, tree-based analytical methods might not be an appropriate way to detect this because of the ambiguous outcomes possible from diploidization, as explained below and best illustrated in the case of maize[37].

Diploidization, the evolutionary process whereby a tetraploid species 'decays' to become a diploid (paleotetraploid) with twice as many distinct chromosomes, is one of the most interesting but unclear aspects of genome evolution. The key event is the switch from having four chromosomes that form a QUADRIVALENT at meiosis, to having two pairs of chromosomes each of which forms a BIVALENT. In population-genetics terms, this is the switch from having four alleles at a single locus (tetrasomic inheritance) to having two alleles at each of two distinct loci (disomic inheritance) (FIG. 5). The molecular basis of diploidization is not understood

at all, but it presumably occurs through the accumulation of DNA sequence changes (and/or deletions) between the chromosomes.

In an allotetraploid, each locus will initially have four alleles, two from each parent. If a reasonable length of time elapses before the species becomes diploidized, genetic drift could cause one locus to become fixed for alleles that originate from one parent, whereas some other locus might retain alleles from both parents (FIG. 5).

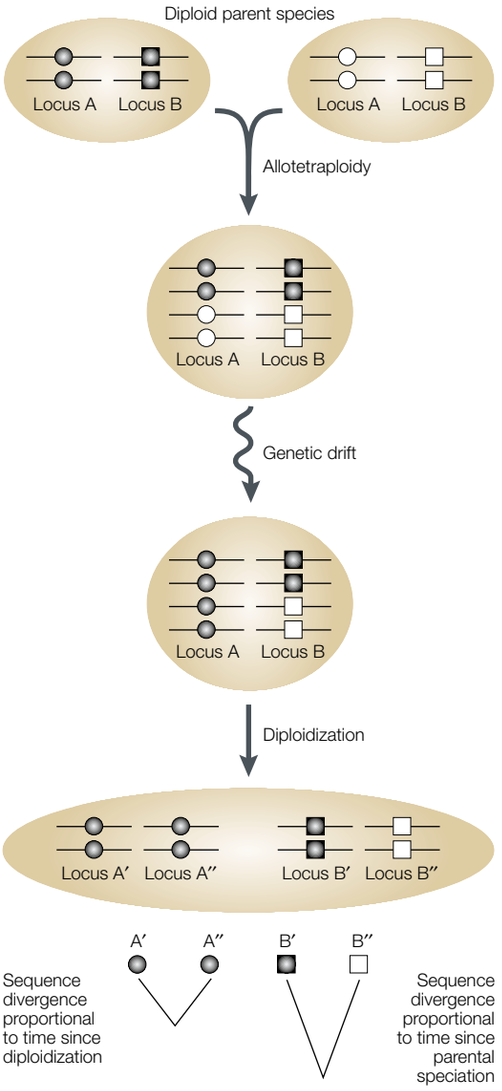


Figure 5 | **Effect of genetic drift and diploidization on inter-locus age estimates in an allotetraploid.** Genetic drift during tetrasomic inheritance of locus A leads to fixation of an allele derived from only one of the parents, so that molecular-clock analysis of the duplicated A′ and A″ loci in the paleopolyploid descendant points to a recent divergence time, corresponding to the diploidization date. Locus B remains polymorphic for the two parental alleles during the tetrasomic phase, so that the molecular estimate of the divergence time between its diploidized daughter loci B′ and B″ might correspond to the speciation date between the two parents (depending on how the alleles segregate at diploidization).

HOMEOLOGUES
Sister chromosomes (or sister loci) resulting from polyploidy in plants.

After diploidization and further sequence evolution, this means that the amount of sequence divergence between some paralogous ('HOMEOLOGOUS' in plant terminology) loci will be proportional to the time elapsed since diploidization, whereas at other loci it will correspond to the time since the speciation between the parents of the allopolyploid. This situation is called segmental allotetraploidy, and the consequence is that phylogenetic trees drawn from some pairs of paralogous loci will point to one divergence date, whereas other pairs of loci will point to a different date (FIG. 5). The maize genome seems to have this structure[37–39].

To make matters worse, in both plants[37] and animals (such as salmonid fish[40]), a single species can harbour a mixture of tetraploid and diploidized loci. In other words, diploidization does not necessarily happen simultaneously for all chromosomes or even for all loci on a particular chromosome. If this is commonplace, it will wreak havoc on the tree-based analysis of paleopolyploids. The consequence of independent diploidization dates for each locus would be a continuum of divergence dates for duplicated loci, ranging from the very recent back to the parental speciation date.

For vertebrates, a 2R hypothesis that involves two rounds of allotetraploidy with genetic drift during tetrasomic inheritance (FIG. 5) would lead to the prediction of four possible sets of ratios among the branch lengths in a phylogenetic tree (two values for both *t*1 and *t*2 in FIG. 2), which might not be distinguishable from statistical noise. Tree-based approaches to identifying paleopolyploids will fare even worse if the date of diploidization can be different for different loci, as mentioned above. Allopolyploidy hypotheses can also upset the (AB)(CD) topology prediction, although different genes in the genome are still expected to yield a consistent tree.

The 'use it or lose it' parameter

The second aspect of diploidization is the loss of genes from sister chromosomes, by mutation or deletion. This is an important factor in the 2R debate because extensive deletion of genes after genome duplication could almost completely obscure the evidence that a duplication had occurred. The fate (retention in duplicate versus loss of one copy) of a duplicated gene pair depends on whether natural selection will act to prevent one of the copies becoming a pseudogene if it is hit by an inactivating mutation. Historically, it has been thought that the duplicate genes must diverge in function for this to happen[3], but more recently many other theoretical arguments have been put forward to explain the persistence of duplicated genes (BOX 3).

Estimating the extent to which ohnologues are retained during subsequent evolution is an area in which model organism genomes might be of some help. In yeast, only about 16% of genes in the genome are members of an ohnologue pair (indicating that about 8% of the original set of duplicated genes were retained in duplicate[32,41]). The number for *Arabidopsis* seems higher, with about 25% of genes being a member of a pair in the analyses of both the *Arabidopsis* Genome Initiative[42] and Vision *et al.*[34].

Box 3 | **What determines whether a duplicated gene is lost or kept?**

Genes are continually being created by duplication and destroyed by mutation. The half-life of duplicated genes in eukaryotes was estimated recently to be only 3–7 Myr[60]. Is this turnover a lottery, or are some genes more likely to survive in duplicate than others? Evidence from both vertebrates[61] and yeast[45] indicates that survival occurs more readily in some gene classes than others, but that the details might differ between taxa (for example, glycolytic genes are often single copy in vertebrates but duplicated in yeast). Ideas as to why some genes might be more likely than others to survive after duplication include the following:

- Subfunctionalization: degenerative mutations in duplicated genes might result in the situation that neither daughter gene alone produces sufficient gene product, which results in natural selection to retain both of the daughter genes[62–64]. Subfunctionalization can be viewed quantitatively, in terms of levels of gene expression[65], or qualitatively, in terms of the partitioning of ancestral functions (see below).
- Partitioning of functions: multifunctional genes might have a relatively high chance of survival after duplication if their functions become parcelled out among the daughter genes[66]. Examples include genes that are expressed in multiple tissues, such as *engrailed* genes[63], and enzymes with multiple substrates.
- Dominant-negative phenotypes: genes that encode multidomain proteins might have an increased chance of survival after duplication if point mutations in those genes tend to be dominant and have deleterious phenotypes[61].
- Dosage effects: some types of gene might be under selection for increased gene expression (for example, some yeast ribosomal protein genes are duplicated but encode identical proteins[45]).

Several studies have indicated that a polyploidization event occurred in an ancestor of teleost fish, shortly after this lineage diverged from the lineage leading to tetrapods[12,43] (FIG. 1). This makes zebrafish and *Fugu rubripes* (the puffer fish) the most relevant model organisms for the 2R hypothesis; it seems reasonable to expect that the frequency of gene deletion (and possibly other outcomes) from the fish-specific polyploidization should be similar to what might have happened in the earlier two rounds of polyploidization proposed for all vertebrates. A recent study by Postlethwait *et al.*[44] indicated that 20% of human genes might be duplicated in zebrafish, which implies that 33% (= 40/120) of zebrafish genes are members of ohnologue pairs. If this low rate of gene survival after duplication were also true of the 2R duplications, and if we make the arguable assumption that the choice of which genes to retain in each round is random, the cumulative effect of low survival in each gene lineage would be that less than 1% of human genes should obey the one-to-four rule[16]. Such a situation would make the 2R hypothesis very difficult to disprove, but it could also be argued that, if so, polyploidization would have had only a minor effect on the vertebrate proteome and was not the powerful evolutionary force envisaged by Ohno.

The future

Strong evidence of paleopolyploidy has so far failed to materialize where it was most anticipated (the human genome[21,22]), although it has mischievously appeared where it was least expected (yeast and *Arabidopsis*). Much work needs to be done on model organisms that are probable paleopolyploids (yeast, *Arabidopsis*, zebrafish and maize) to improve our understanding of their evolution, particularly with regard to the 'black box' of diploidization. Sequencing the genomes of several species that are descended from the same polyploidy event would throw light on diploidization and show, for example, whether the extensive gene deletion that seems to have occurred in all eukaryotic paleopolyploids is random with respect to gene functions. In yeast, some functional categories such as signal transduction are overrepresented among the ohnologues[45], but it is too early to say whether this is an inevitable outcome of diploidization.

The preliminary analysis of the human genome[21,22] does not provide strong support for the 2R hypothesis, but in biology it is notoriously difficult to prove the absence of something. Alternative hypotheses about possible adaptive reasons for the positioning of genes in the eukaryotic genome deserve serious consideration[30,46,47]. Formal tests of the 2R hypothesis in vertebrates will require a null hypothesis, which is not easy to construct (compare REFS 36,39). The simplest null hypothesis is that each gene duplication occurred independently, but where does this leave other concepts such as the regional duplication of parts of chromosomes, or the aneuploidy of single chromosomes? Now that we have most of the cards[10], it is still proving remarkably difficult to play 'snap'.

Links

DATABASE LINKS *CDK7* | *CDK3* | Pip2 | Oaf1 | *engrailed*
FURTHER INFORMATION *Saccharomyces* Genome Database | The *Arabidopsis* Information Resource | Maize Database | The Zebrafish Information Network | *Arabidopsis* Genome Initiative | Génolevures project | UCSC Draft Human Genome Browser | Ken Wolfe's lab | Yeast gene duplications

1. Spofford, J. B. Phylogenetic mechanism. *Science* **175**, 617–618 (1972).
2. Lewin, B. Genes in tandem. *Nature* **230**, 314 (1971).
3. Ohno, S. *Evolution by Gene Duplication* (George Allen and Unwin, London, 1970).
4. Nadeau, J. H. & Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA* **81**, 814–818 (1984).
5. Ohno, S. Ancient linkage groups and frozen accidents. *Nature* **244**, 259–262 (1973).
6. Nadeau, J. H. in *Advanced Techniques in Chromosome Research* (ed. Adolph, K. W.) 269–296 (Marcel Dekker, New York, 1991).
7. Schughart, K., Kappen, C. & Ruddle, F. H. Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proc. Natl Acad. Sci. USA* **86**, 7067–7071 (1989).
8. Ruddle, F. H., Bentley, K. L., Murtha, M. T. & Risch, N. Gene loss and gain in the evolution of the vertebrates. *Development* S155–S161 (1994).
9. Hughes, A. L. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* **48**, 565–576 (1999).
10. Skrabanek, L. & Wolfe, K. H. Eukaryote genome duplication — where's the evidence? *Curr. Opin. Genet. Dev.* **8**, 694–700 (1998).
11. Ohno, S. Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. *Semin. Cell Dev. Biol.* **10**, 517–522 (1999).
12. Meyer, A. & Schartl, M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* **11**, 699–704 (1999).
13. Holland, P. W. H. Introduction: gene duplication in development and evolution. *Semin. Cell Dev. Biol.* **10**,

515–516 (1999).
14. Martin, A. Is tetralogy true? Lack of support for the 'one-to-four' rule. *Mol. Biol. Evol.* **18**, 89–93 (2001).
15. Wang, Y. & Gu, X. Evolutionary patterns of gene families generated in the early stage of vertebrates. *J. Mol. Evol.* **51**, 88–96 (2000).
16. Nadeau, J. H. & Sankoff, D. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* **147**, 1259–1266 (1997).
17. Martin, A. P. Increasing genomic complexity by gene duplication and the origin of vertebrates. *Am. Nat.* **154**, 111–128 (1999).
18. Llorente, B. *et al.* Genomic exploration of the hemiascomycetous yeasts. 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Lett.* **487**, 101–112 (2000).
19. Gibson, T. J. & Spring, J. Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc. Trans.* **28**, 259–264 (2000).
20. Wolfe, K. Robustness — it's not where you think it is. *Nature Genet.* **25**, 3–4 (2000).
21. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
22. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
23. Lundin, L. G. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16**, 1–19 (1993).
A classic paper that identified numerous apparent duplicated chromosomal regions in mammals.
24. Katsanis, N., Fitzgibbon, J. & Fisher, E. M. C. Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel *PBX* and *NOTCH* loci. *Genomics* **35**, 101–108 (1996).
25. Kasahara, M., Nakaya, J., Satta, Y. & Takahata, N. Chromosomal duplication and the emergence of the adaptive immune system. *Trends Genet.* **13**, 90–92 (1997).
26. Pébusque, M.-J., Coulier, F., Birnbaum, D. & Pontarotti, P. Ancient large scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* **15**, 1145–1159 (1998).
27. Wraith, A. *et al.* Evolution of the neuropeptide Y receptor family: gene and chromosomal duplications deduced from the cloning and mapping of the five receptor subtypes in pig. *Genome Res.* **10**, 302–310 (2000).
28. Spring, J. Vertebrate evolution by interspecific hybridisation — are we polyploid? *FEBS Lett.* **400**, 2–8 (1997).
29. Sidow, A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6**, 715–722 (1996).
30. Hughes, A. L. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol. Biol. Evol.* **15**, 854–870 (1998).
One of the first reports of wildly heterogeneous duplication date estimates from genes thought to form a duplicated chromosomal region.
31. Hughes, A. L., da Silva, J. & Friedman, R. Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res.* (in the press).
32. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
33. Pál, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* (in the press).
34. Vision, T. J., Brown, D. G. & Tanksley, S. D. The origins of genomic duplications in *Arabidopsis*. *Science* **290**, 2114–2117 (2000).
35. Smith, N. G., Knight, R. & Hurst, L. D. Vertebrate genome evolution: a slow shuffle or a big bang? *Bioessays* **21**, 697–703 (1999).
36. Friedman, R. & Hughes, A. L. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* **11**, 373–381 (2001).
37. Gaut, B. S. & Doebley, J. F. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl Acad. Sci. USA* **94**, 6809–6814 (1997).
38. Gaut, B. S., Le Thierry d'Ennequin, M., Peek, A. S. & Sawkins, M. C. Maize as a model for the evolution of plant nuclear genomes. *Proc. Natl Acad. Sci. USA* **97**, 7008–7015 (2000).
39. Gaut, B. S. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res.* **11**, 55–66 (2001).
Interesting exploration of methods for statistically testing genome duplication hypotheses.
40. Allendorf, F. W. & Danzmann, R. G. Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics* **145**, 1083–1092 (1997).
41. Seoighe, C. & Wolfe, K. H. Extent of genomic rearrangement after genome duplication in yeast. *Proc. Natl Acad. Sci. USA* **95**, 4447–4452 (1998).
42. *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
Supplementary figure 1 (http://www.tigr.org/~salzberg/MUMmer-index.html) contains dot-matrix DNA comparisons of all pairs of *Arabidopsis* chromosomes.
43. Amores, A. *et al.* Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**, 1711–1714 (1998).
44. Postlethwait, J. H. *et al.* Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* **10**, 1890–1902 (2000).
45. Seoighe, C. & Wolfe, K. H. Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* **2**, 548–554 (1999).
46. Hughes, A. L. *Adaptive evolution of genes and genomes* (Oxford Univ. Press, New York, 1999).
47. Hurst, L. D. The evolution of genomic anatomy. *Trends Ecol. Evol.* **14**, 108–112 (1999).
48. Mewes, H. W. *et al.* Overview of the yeast genome. *Nature* **387**, S7–S65 (1997).
49. Coissac, E., Maillier, E. & Netter, P. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. *Mol. Biol. Evol.* **14**, 1062–1074 (1997).
50. Seoighe, C. & Wolfe, K. H. Updated map of duplicated regions in the yeast genome. *Gene* **238**, 253–261 (1999).
51. Baumgartner, U., Hamilton, B., Piskacek, M., Ruis, H. & Rottensteiner, H. Functional analysis of the Zn(2)Cys(6) transcription factors Oaf1p and Pip2p. Different roles in fatty acid induction of β-oxidation in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **274**, 22208–22216 (1999).
52. Vision, T. J. & Brown, D. G. in *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families* (eds Sankoff, D. & Nadeau, J. H.) 479–491 (Kluwer Academic, Dordrecht, 2000).
53. Llorente, B. *et al.* Genomic exploration of the hemiascomycetous yeasts. 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Lett.* **487**, 122–133 (2000).
54. Keogh, R. S., Seoighe, C. & Wolfe, K. H. Evolution of gene order and chromosome number in *Saccharomyces*, *Kluyveromyces* and related fungi. *Yeast* **14**, 443–457 (1998).
55. Seoighe, C. *et al.* Prevalence of small inversions in yeast gene order evolution. *Proc. Natl Acad. Sci. USA* **97**, 14433–14437 (2000).
56. Louis, E. J. The chromosome ends of *Saccharomyces cerevisiae*. *Yeast* **11**, 1553–1573 (1995).
57. Ku, H. M., Vision, T., Liu, J. & Tanksley, S. D. 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* **97**, 9121–9126 (2000).
58. Paterson, A. H. *et al.* Comparative genomics of plant chromosomes. *Plant Cell* **12**, 1523–1540 (2000).
59. Graur, D. & Li, W.-H. *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, Massachusetts, 1999).
60. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
Global study of the frequency of occurrence and degree of divergence of duplicated genes in several eukaryotic genomes. This paper shows that duplicated genes arise at a high rate but have a short half-life.
61. Gibson, T. J. & Spring, J. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet.* **14**, 46–49 (1998).
62. Hughes, A. L. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond.* B **256**, 119–124 (1994).
63. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
64. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473 (2000).
65. Blanchard, J. L. & Lynch, M. Organellar genes: why do they end up in the nucleus? *Trends Genet.* **16**, 315–320 (2000).
66. Wagner, A. The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene functions. *Genetics* **154**, 1389–1401 (2000).

Acknowledgements

I am grateful to K. Hokamp, A. McLysaght and L. Skrabanek for discussion.