COMPUTATIONAL APPROACHES TO UNVEILING ANCIENT GENOME DUPLICATIONS

Yves Van de Peer

Abstract | Recent analyses of complete genome sequences have revealed that many genomes have been duplicated in their evolutionary past. Such events have been associated with important biological transitions, major leaps in evolution and adaptive radiations of species. Here, we consider recently developed computational methods to detect such ancient large-scale gene duplication events. Several new approaches have been used to show that large-scale gene duplications are more common than previously thought.

SYMPATRIC SPECIATION
Genetic divergence that leads to species formation in the same habitat.

BLOCK OR SEGMENTAL DUPLICATIONS
A duplication of several genes at the same time. The result is two genomic segments that share a similar set of genes and are therefore homologous.

TANDEM DUPLICATION
Duplication of (single) genes that create tandem repeats in the genome. The *HOX* genes are a well-known example of a gene family generated through tandem duplication.

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium.
e-mail: Yves.VandePeer@psb.ugent.be

doi:10.1038/nrg1449

In his widely cited book '*Evolution by Gene Duplication*', Susumu Ohno was one of the first to suggest that gene duplication might have been more important in shaping the evolution of biological novelty and complexity than natural selection acting on point mutations[1]. Many now share this view. In particular, large-scale gene duplication events might explain major leaps in evolution or adaptive radiations of species. For example, had large-scale gene duplication events in early vertebrate evolution[2–7] not occurred, vertebrates as we know them today might not have existed[1,8,9]. Moreover, genome duplication is probably the main way in which SYMPATRIC SPECIATION and adaptation occurs in plants, owing to its large-scale effects on gene regulation and developmental processes[10,11]. Similarly, in ray-finned fishes, where genome doubling has occurred after they diverged from the land vertebrates, but before their own massive diversification[7,12–17], the more complex genomic architecture that duplications permit might have allowed faster responses to changing environments through adaptation and speciation[12,18–20].

Because of its putative impact on evolution, the search for traces of ancient genome duplications has recently received much attention. However, despite the growing body of evidence for the importance of large-scale gene duplications, the number, age and biological significance of such events in eukaryote evolution are still vigorously debated. Much of the recent controversy stems from differences in the results obtained from the various approaches applied to the identification of BLOCK OR SEGMENTAL DUPLICATIONS. The high level of gene loss after duplication, as well as translocations, chromosomal rearrangements and recombination, complicates the identification of duplicated segments, particularly when these duplications are ancient[21–23]. However, new bioinformatics approaches allow remnants of block duplications to be detected, even when apparent homology is undetectable.

Here, these new computational approaches to identifying ancient large-scale duplications are reviewed. In particular, I focus on comparisons of protein-coding sequences in structurally annotated genomes, which have proved to be most useful for identifying ancient genome duplications. First, approaches to identifying regions of homology within a genome are discussed. The various methods for dating duplication events and the comparative approach to the identification of duplications are then summarized. Finally, I discuss the future outlook for this rapidly expanding field.

Uncovering genomic homology

Gene-homology matrices. Remnants of large-scale gene duplication events, where large segments, whole chromosomes or even whole genomes have been duplicated — as opposed to smaller-scale duplications, such as TANDEM DUPLICATIONS — can be detected by the delineation of 'blocks' or 'segments' in the genome that contain a set of homologous genes. This involves the identification of a number of homologous gene pairs, usually referred to as 'ANCHOR POINTS', in relatively close proximity between two different segments in the genome, either on the same chromosome, or on different chromosomes. The evidence for a block duplication is strengthened if the order in which the homologous genes are found on the chromosomes is conserved. When the similarity between two such genomic segments is statistically significant — that is, unlikely to be the result of chance — the duplicated genes are assumed to be the result of a single duplication event. The statistics that determine COLLINEARITY usually depend on the number of gene pairs that still can be identified as homologous and the distance over which they are found. This, in turn, usually depends on the number of 'single' genes that interrupt collinearity.

In practice, to detect homology between different chromosomal segments, chromosomes are represented as lists of genes, which are sorted according to their position on that chromosome. From these gene lists, genes that have homologues in other chromosomal segments are identified, usually on the basis of an all-against-all similarity search with tools such as BLAST[24] (see BOX 1 for more information on homology and the delineation of gene families).

Next, the information on homologous genes is stored in a so-called gene homology matrix (GHM), a hypothetical example of which is shown in FIG. 1. In a GHM, collinear regions appear as diagonal lines of dots or squares representing homologous genes, whereas tandem duplications form either horizontal or vertical lines, depending on which genomic segment has the additional copies. Inversions are identified as diagonals with opposite orientation (not shown) and gaps in diagonal regions correspond to insertions (that have arisen through translocation, not duplication) or losses of genes.

The big challenge in interpreting GHMs is, therefore, the identification of diagonals. The number of anchor points that are found in close proximity (referred to as clusters) usually depends on a preset gap size, which is the number of unique intervening genes that are allowed between two anchor points. The gap size parameter value should ideally be adjustable, because this distance can differ considerably between recent and older block duplications. A second parameter that might be considered for the automatic delineation of 'clusters' of homologous gene pairs is a quality parameter that determines the extent to which the elements of a cluster actually fit on a diagonal line and are, therefore, probably the result of the same duplication event[25,26]. When a putative collinear region has been detected as a diagonal in a GHM, a permutation test can be used to evaluate its statistical significance. In such tests, GHMs are built from randomized data sets to calculate the probability that a group of homologous gene pairs is found in close proximity owing to chance (BOX 2).

One of the first genomes that GHMs were applied to, in order to find evidence for large-scale gene duplications, was yeast[27]. The analysis of GHMs showed that non-overlapping duplicated regions covered at least 50% of the yeast genome. Furthermore, the orientation of all detected duplicated segments with respect to the centromere was generally conserved. These two observations led the authors to conclude that a single large-scale duplication event, most probably a POLYPLOIDY event, was the only explanation for the detected duplication pattern. Vision and colleagues[28] used a similar approach to develop an algorithm that searches for duplicated regions in the *Arabidopsis thaliana* genome. They treated the dots of the resulting GHM as nodes in a graph and assigned weights to the connecting edges; diagonal series of dots were then detected as

ANCHOR POINT
A pair of homologous genes in a duplicated segment. Several anchor points in close proximity form strong evidence for a block duplication.

COLLINEARITY
Conservation in gene order and gene content between two genomic segments.

POLYPLOIDY
A polyploid organism has more than two sets of chromosomes.

E-VALUES
The expect value (E) is a parameter that describes the number of hits one can expect to see by chance when searching a database of a particular size. The lower the E-value, the more significant the match is, and the more probable that two sequences are homologous.

PARALOGUES
Homologous genes that have originated through gene duplication events; that is, by tandem, block or whole-genome duplication events.

GRAPH-CLUSTERING ALGORITHM
This is applied to separate, sparsely-connected, dense subgraphs (here gene families). This means that the graph is partitioned in such a way that the distance between the subgraphs (clusters) is maximized, whereas the sum of the distances within each subgraph is minimized.

Box 1 | Homology and the delineation of gene families

A gene family with n members in a genome adds $n(n-1)/2$ dots to a gene homology matrix (GHM), which means that a family of a hundred members adds a few thousand dots. Consequently, the presence of several of these large gene families in a genome can obscure true genomic homology, in particular when methods that do not take into account gene order are used to uncover duplicated regions[90,91].

To cope with large gene families and the possible noise they introduce, a threshold can be put on the maximum gene-family-size to be included in the analysis. Of course, to do this, all homologous genes in the dataset have first to be grouped into gene families. Absolute criteria, such as E-VALUES, can be used to decide whether genes are true homologues (or in this case PARALOGUES)[24]. However, sequences have often diverged to the extent that their common origin is questionable, when based on direct sequence comparison. For example, if sequences are only 20 to 30% identical, it is difficult to decide whether they are homologous or not. To address this problem, Rost[92] used an empirical formula, in which the cut-off sequence identity increases with decreasing length of the alignable regions. This strategy takes into account the greater chance of having a high identity for a short alignable region.

Short proteins that share one or more domains with longer proteins pose another problem. Relying only on significant (local) similarity will possibly result in gene families that contain non-homologous proteins that only share a similar domain. Ideally, genes should be ascribed to the same gene family only if they also have highly similar domain architectures. To distinguish local homology from global homology, additional parameters can be considered, such as coverage of the alignable region on both potentially homologous genes[93]. To actually group homologous genes into families, the single linkage algorithm can be used. The principle is simple; if protein A hits protein B and B hits C, then proteins A, B and C are put in the same family, regardless of whether protein A hits protein C or not. Recently, a new and promising approach has been developed (TRIBE–MCL) for clustering protein sequences into families[94]. The method applies a GRAPH-CLUSTERING ALGORITHM on a weighted graph that represents precomputed sequence similarity relations where nodes equal proteins and connections equal similarity relationships.

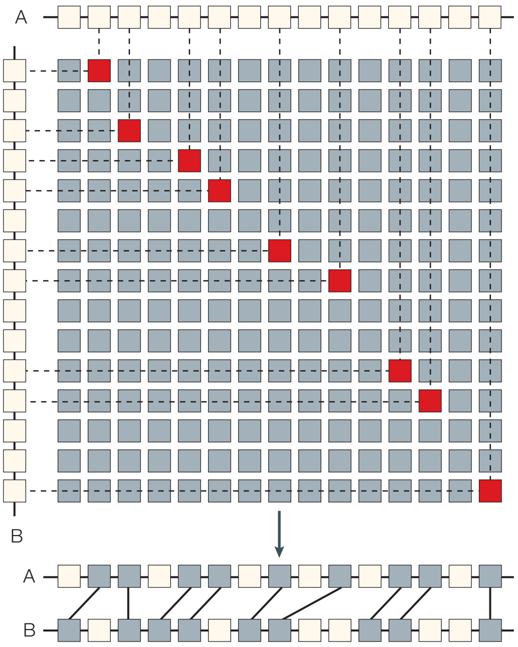

Figure 1 | **Gene homology matrix (GHM).** A and B represent two hypothetical genomic segments that are being evaluated for homology. Each white square represents a gene. Red squares in the GHM indicate homologous genes, which are indicated in grey and connected by black lines in the multiplicon underneath. Red squares form a cluster of anchor points. When these form a diagonal in the GHM, they delineate a block or segmental duplication (see text for details).

minimum-weight paths. These series of dots were then combined to delineate duplicated regions. Their results indicated that as much as 81% of the *A. thaliana* genome was duplicated and that numerous regions probably underwent multiple duplications[28].

Entering the twilight zone of homology. The identification of recent gene duplications is relatively straightforward, but ancient duplications pose more of a problem, particularly because of the increasing possibility that duplicate genes have been lost. The frequency of gene-copy preservation after duplication over large evolutionary periods is unexpectedly high[29,30] and several models have been put forward to explain the high retention of duplicates[31–33]. Nonetheless, the most probable fate of a gene duplicate is NON-FUNCTIONALIZATION and, consequently, gene loss[30]. Large segmental deletions or individual gene deletions can cause gene loss. This process can be balanced between two homologous segments or act primarily on one of them. However, analyses of homologous segments in yeast and plants seem to indicate that gene loss is primarily the result of small deletions and is typically balanced between the homologous segments[34–37].

So, whereas block duplications can easily be identified because they form clear diagonals in GHMs, extensive gene loss and gene translocations can obscure homology between two segments. In particular, after tens or hundreds of millions of years of evolution, too few homologous gene pairs might remain (in close proximity) to detect statistically significant collinearity. In such cases, comparison of both genomic segments with a third segment can still allow homology to be detected (FIG. 2a). Heavily degenerated segments in the genome that share little similarity with each other can often be shown to have been derived from the same common ancestral segment, because they both still show sufficient collinearity with a third segment. The principle is simple: if A and B are homologous, and A and C are homologous, then B and C must be homologous.

Such TRANSITIVE HOMOLOGY analyses can provide important information about the number of duplication events that have occurred over time. For example, in *A. thaliana*, taking into account such transitive homology relationships has considerably contributed to solving a long-standing controversy about the actual number of large-scale gene duplications that the *A. thaliana* genome has undergone. In this case, Simillion and colleagues[34] identified many sets of homologous segments (so-called MULTIPLICONS), spread all over the genome, with multiplication levels of between five and eight (FIG. 2b,c). This finding, and the fact that dating based on synonymous substitutions (see below) clearly revealed three significantly different age classes, strongly indicated that *A. thaliana* had undergone three, but not more, whole-genome doublings, a finding that was confirmed later by phylogenetic analyses[38].

Transitive homologies have proven important for uncovering many previously undetected duplications[34]. However, a genomic segment must still show clear collinearity with at least one other segment in the genome to be able to determine whether the two are homologous (FIG. 2a). Frequently, this is not the case: homologous segments have diverged so much in both gene content and order that they no longer show detectable collinearity with any of the other segments. Such segments could not be detected with any of the methods previously available. Recently, software has been developed to build so-called 'GENOMIC PROFILES'[26]. These profiles combine gene content and order information from multiple segments and can then be used as more sensitive probes to sweep the rest of the genome to uncover additional homology (FIG. 3).

The different approaches described above work particularly well for identifying ancient duplications when more recent duplication events have also occurred. This is best illustrated for *A. thaliana*, where remnants of the most recent genome duplication were used to uncover the older duplications. Indeed, based on the identification of recently duplicated homologous segments, transitive homology relationships were able to be considered[34], genomic profiles built[26] and the ancestral gene order reconstructed[38,39]. This example illustrates that all these different approaches can be valuable for uncovering additional, more ancient duplications.

Content without order. So far, we have discussed methods that look for collinearity between genomic segments; that is, conservation of gene content and order. Although in many cases these methods are sufficiently sensitive to find remnants of ancient duplication events,

NON-FUNCTIONALIZATION
When one of the two duplicate genes accumulates deleterious mutations in coding or regulatory sequences that ultimately renders the gene non-functional.

TRANSITIVE HOMOLOGY
When homology between two genes or genomic segments can only be inferred through a third gene or segment.

MULTIPLICON
A set of homologous genomic segments. The multiplication level of a multiplicon refers to the number of homologous segments that the multiplicon consists of.

GENOMIC PROFILES
A genomic profile is formed by genomic segments that combine gene order and gene content. These profiles can then be used to identify more concealed forms of homology.

Box 2 | Statistical validation of block duplications

Different statistical tests have been developed to discriminate between true PARALOGONS and genomic segments that share some homologous genes because of the accidental organization of these genes in the genome. The most straightforward way of doing this is to compare the clusters of anchor points, obtained from a real genome, with those obtained from a large number of randomized datasets, generated by reshuffling the order of all genes in the genome of interest (permutation testing). The simplest way to discriminate between false positive clusters and real ones is to set a threshold for the minimum size; that is, the number of anchor points that a cluster must contain. This threshold is obtained by considering the size distribution — in other words, the number of anchor points in duplicated segments — of the clusters from the randomized datasets. Above a certain size, the probability that clusters arose by chance becomes sufficiently low for any cluster of that size or bigger to be safely regarded as a true positive[2,41]. More advanced tests have been developed that not only take into account the number of anchor points, but also the spacing between these anchor points. A cluster will then be regarded as a true positive when, for a given size, its anchor points are in closer proximity than a minimum density found in randomized datasets. Although randomization tests are easy to implement, they are computer intensive. Typically 1,000 or more randomized datasets are generated, after which each of these datasets must be searched for groups of anchor points in close proximity. Recently, several analytical tests have been developed that readily estimate the significance of a cluster of anchor points without the need for performing randomizations. Generally, these methods compare the local density of dots in gene-homology matrices (GHMs) to the overall density of dots. By comparing these values, the probability that the cluster was generated by chance rather than a segmental duplication can be calculated[26,90,95,96].

the requirement for conserved gene order can sometimes be too strict. For example, when analysing the human genome, little evidence for ancient large-scale gene duplication events is found based on conservation of both gene content and order. However, by releasing the constraint of conserved gene order and only considering conservation in gene content, substantial additional evidence is found for duplicated segments in the human genome, such as those containing the major histo-compatibility complex loci[40].

One way to evaluate conserved gene content independent of gene order in an automated manner is to compare two genomic windows and to count the number of homologous gene pairs in these windows. Such a strategy has been adopted to analyse the PARANOMES of *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*, with a genomic window defined as a region containing eight genes that have at least one homologue somewhere in the genome[41]. All possible non-overlapping windows were compared with one another. By comparing the results from real genomes or, even better, by comparing paranomes with windows from randomly shuffled genomes (BOX 2), all three genomes investigated showed a significantly larger number of windows that share at least four homologous gene pairs than were expected from a random gene distribution. In addition, Cavalcanti and colleagues[42] developed an alternative window-based strategy that, instead of dividing the dataset into non-overlapping windows, considers all possible genomic windows in a genome.

In a large-scale analysis of the human genome, McLysaght and colleagues[2] used a similar approach that is able to both identify and delineate individual pairs of homologous segments. Starting from two homologous genes, each at a different chromosomal location, the software looks for two other homologous genes that are each located at a predefined distance from the former two genes. When such a pair is found, it is added to the first pair to form a cluster of homologous genes. Additional pairs of genes that are in the vicinity of the cluster are searched for and subsequently added until no more pairs can be found. The resulting clusters are used to delineate pairs of homologous segments. Using this strategy, 44% of the human genome was covered by homologous segments with six or more pairs of duplicated genes. Combined with phylogenetic dating (see below), this observation indicated that at least one polyploidy event occurred early in the evolution of vertebrates[2].

The chromosomes that carry the *Hox* genes provide a nice example of conservation of gene content but not gene order (FIG. 4). Although the Hox loci themselves are conserved in gene order[43], the regions flanking the Hox clusters of different vertebrate genomes are conserved in gene content only[4]. Using GHMs as described above, these regions would not be identified as being homologous. However, using a window-based approach that shows overall conservation of gene content, SYNTENIC regions can be clearly recognized, which provides evidence that not only Hox clusters, but much larger regions have been duplicated during large-scale duplication events.

What sort of duplication? When many homologous segments have been uncovered within the same genome, the question remains as to how these segments arose. Is the observed distribution the result of one or more successive large-scale gene duplication events, followed by intense rearrangements and deletions, or rather, is it the result of several smaller independent block duplications[44–47]? The lack of overlapping block duplications provides important evidence in favour of a single duplication event. For example, the absence of such overlaps was one of the most compelling arguments in favour of the hypothesis that *S. cerevisiae* was an ancient polyploid that had its

PARALOGON
Homologous genomic segments created by partial or complete genome duplication.

PARANOME
The complete set of duplicated genes in a genome. The paranome can be formed by both small-scale and large-scale gene duplication events.

SYNTENY
Two loci are called syntenic when they are located on the same chromosome.

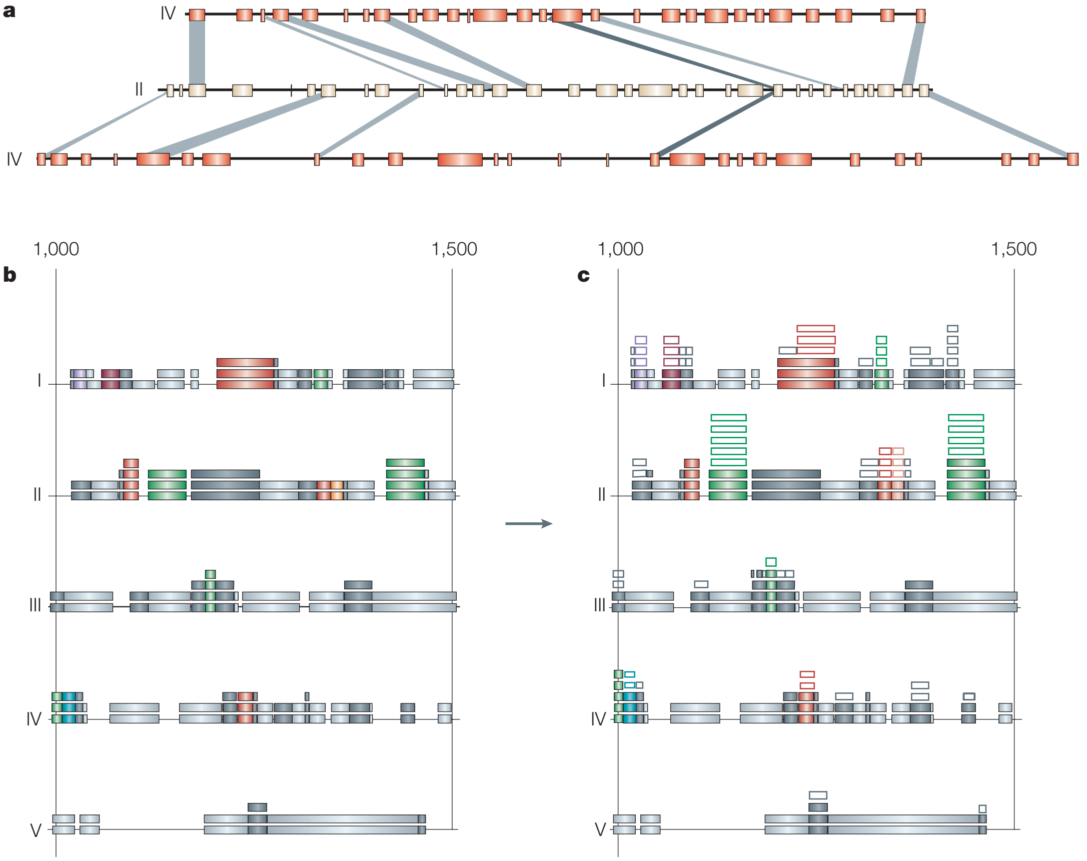

Figure 2 | **Hidden duplications and transitive homology. a** | Example of a set of homologous segments where homology between the two segments of chromosome IV in *Arabidopsis thaliana* can only be inferred through homology with (part of) chromosome II (homology is indicated by the grey bands). Both segments on chromosome IV have only one homologous gene in common (dark grey band). However, both segments have several homologous genes in common with a segment on chromosome II. Therefore, it can be concluded that both segments on chromosome IV are also homologous. **b**,**c** | Considering transitive homology relationships can considerably increase the multiplication level of genomic segments, as shown for a fragment of the five chromosomes of *A. thaliana*. Baselines (black) represent genes 1,000 to 1,500 on the five chromosomes. Boxes on the baselines indicate segments that are part of a group of homologous segments (referred to as multiplicons). The number of boxes above the baselines indicates the number of additional segments that are homologous to the segment on the baseline. Filled boxes represent clear duplications, whereas empty boxes denote so-called concealed or hidden duplications (see **a**), compared with the chromosome segment on the baseline (see text for details). For all multiplicons with a multiplication level greater than four, a different colour was used. Multiplicons with multiplication levels of three or four are marked in dark grey, whereas a multiplication level of two is marked in light grey. In **b**, transitive homology is not considered. In **c**, transitive homology is considered, which results in the identification of many additional homologous segments.

whole genome duplicated[27,48]. Similarly, initial reports on the *A. thaliana* genome sequence used the apparent lack of overlapping duplications to argue for a single duplication event[49–51]. However, in this case, more sophisticated analyses uncovered many overlapping block duplications, increasing the number of ancient polyploidy events in *A. thaliana* from one to three[28,34,38].

A more convincing way to determine whether block duplications are the result of a limited number of large-scale gene duplication events is by dating the block duplications. In fact, the methods that can be used to infer block duplication dates are not different from those that are used to date individual gene duplications (see below), and block duplication origins are usually inferred by averaging the ages of individual anchor points in duplicated blocks[7,28,52].

Dating duplication events

Although the identification of segmental duplications is usually considered strong evidence for large-scale gene duplications, block duplications do not necessarily have to be identified to infer genome-wide duplication events. If many gene duplicates can be shown to have been created at about the same time, this can also be considered as strong evidence that most of these paralogous genes have been created by a single event. Several methods are commonly used to date gene duplication events, the most notable of which are absolute dating based on third codon or synonymous substitution rates; absolute dating based on non-synonymous substitution rates or protein-based distances; and relative and absolute dating by the construction and analysis of phylogenetic trees (FIG. 5).

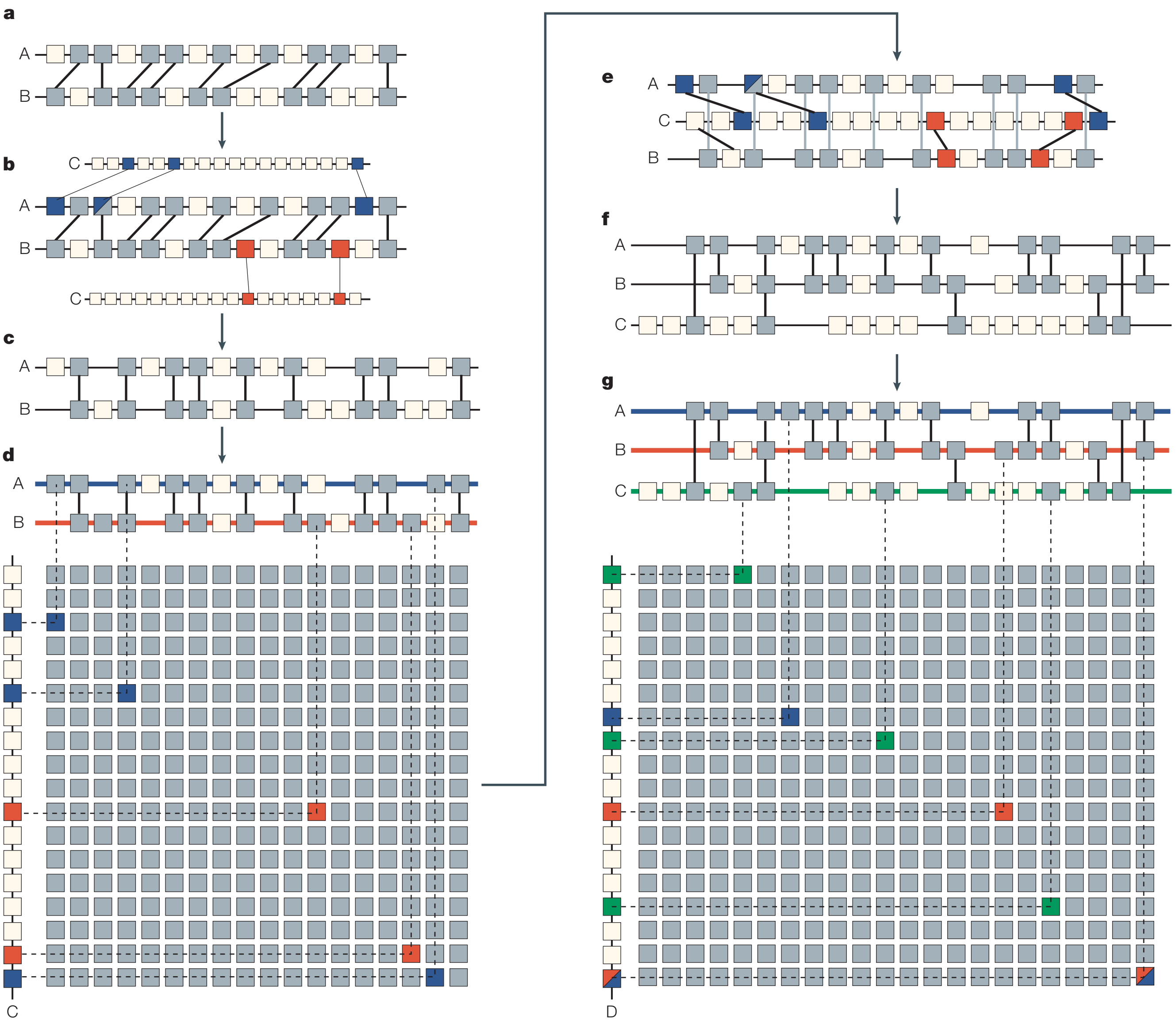

Figure 3 | **Detection of genomic homology using a genomic profile. a** | Segments A and B are homologous segments (see FIG. 1). **b** | A third segment, C, is being compared with both segment A and segment B, and shares some homologous genes with both segments, but too few to infer statistically significant homology. **c** | Segments A and B are aligned to form a profile. Note that, as a consequence of the alignment procedure[26], sets of non-homologous genes (empty boxes) can be placed at the same position in the profile. **d** | A homology matrix can now be constructed by comparing this profile with the genes of another chromosomal segment (segment C on the left of the matrix). Anchor points in the matrix are detected whenever a gene of this chromosomal segment is homologous to one of the genes in any of the segments in the profile. The blue squares represent anchor points between segments A and C, the red between B and C. Note that segment A shares three anchor points with segment C, whereas segment B shares two anchor points with segment C, but when combined in a profile they together have five anchor points with C. **e**,**f** | The new segment is aligned against the existing profile and consequently added to it. Half-shading of the third gene on segment A indicates homology with two other genes on different segments (see also panels **f** and **g**). **g** | This new profile can again be compared against another segment, D. Again, anchor points with segments A en B are shown in blue and red, respectively, whereas anchor points with segment C are shown in green. Note that segment D has only two (segments A and B) or three (segment C) anchor points with each segment in the profile individually, but a total of six anchor points with the profile as a whole. A red-blue square denotes an anchor point between segments A, B and D. Half-shaded squares indicate homology between genes of segments A, B and C.

Absolute dating based on synonymous substitutions. Because most substitutions in third-codon positions do not result in amino-acid replacements, the rate of fixation of these substitutions is expected to be relatively constant in different protein-coding genes[53] and to reflect the overall mutation rate[54]. The time of divergence, or duplication (T), between two sequences can be calculated from this as $T = K_S/2\lambda$, where K_S is the fraction of synonymous substitutions per synonymous site and λ is the mean rate of synonymous substitution[53].

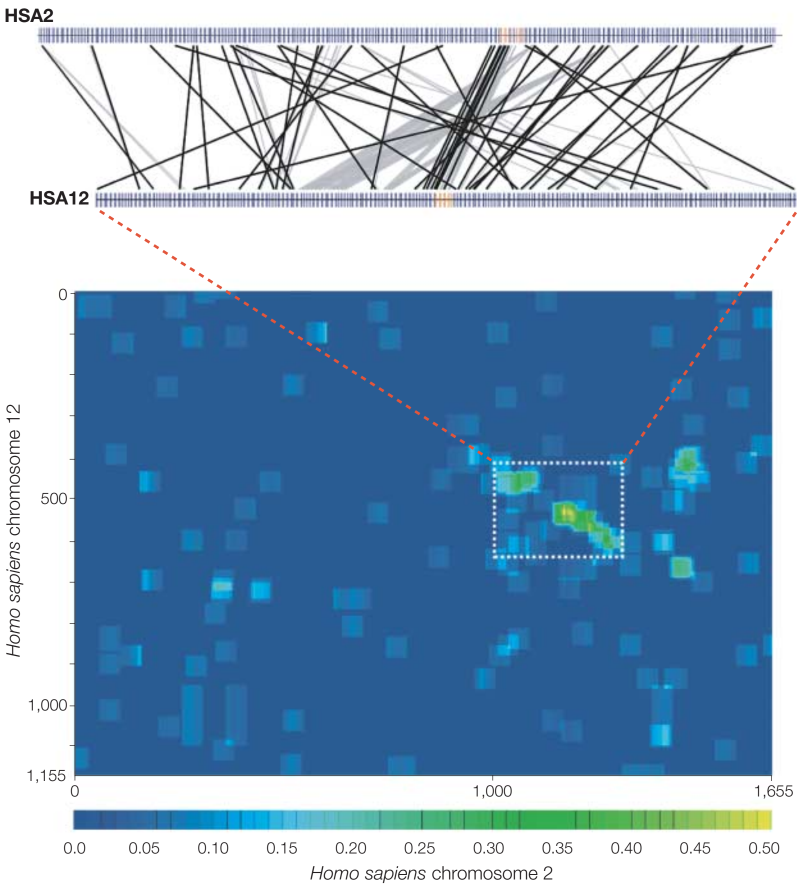

Figure 4 | **Construction of a gene duplication landscape.** The upper panel compares two segments of the human genome, from chromosomes 2 and 12. Homologous genes are connected by black lines; if genes have been tandemly duplicated on one chromosome and are homologous to genes on the other chromosome, they are indicated by grey connecting lines. Genes in red indicate the *HOX* genes. The panel below shows the result of a two-dimensional sliding window approach that was applied to determine the amount of gene content conservation between the two chromosomes. A two-dimensional sliding window of size *p* by *p* (one square in the figure) is defined, where *p* is the number of genes in the compared segments (here this is set to 25). During every sliding step, the percentage of homologous genes between both chromosomes for a particular window is calculated. This value is then represented by a colour, where the intensity is a measure for the number of homologous genes between two genomic segments; blue indicating low conservation in gene content, and green and yellow indicating moderate to high conservation in gene content.

SILENT SUBSTITUTIONS
Nucleotide substitutions that do not lead to amino-acid replacements. They are considered to be neutral and to occur in a clocklike manner.

SYNONYMOUS SITE
One at which a nucleotide change does not alter the amino acid encoded.

The value for λ differs for various organisms; in dicots, for example, the estimate is 1.5 synonymous substitutions per 10^8 years (REF. 55), for monocots 6.5 synonymous substitutions per 10^9 years (REF. 56), and for mammals it is considered to be 2.5 substitutions per 10^9 years (REF. 30).

Although SILENT SUBSTITUTIONS have been used extensively to infer duplication dates[27,30,34,39,57,58], there is one important caveat that applies to this method, namely that dating based on such substitutions can be applied only when K_S is relatively small. Higher values of K_S point to saturation of SYNONYMOUS SITES and should, therefore, be used with caution when drawing any conclusions regarding the date of duplication events. Zhang and colleagues[60] also noted the sometimes large variation among K_S estimates for contemporaneously duplicated genes in *A. thaliana*. However, Vandepoele and colleagues[57] showed that removal of outliers greatly reduced the variation of the final K_S estimates for duplicates of the same age. There are different ways to compute the number of synonymous substitutions per synonymous site, depending on which method is used to correct for multiple mutations at these sites[61–65].

Recently, Blanc and Wolfe[58] provided an elegant way to uncover evidence for large-scale gene duplication events based on K_S dating of ESTs. These authors used ESTs to identify pairs of duplicated genes in 14 plant species. By plotting the number of pairs of homologous ESTs against the time since duplication, they obtained age distributions of duplicated genes for the different plant species. Where a temporal peak of duplicates is observed, disturbing the normally observed exponential decay of a number of duplicated genes through time[30,66], it was concluded that these have probably been created by large-scale gene duplication or polyploidy events. With this approach, even without having complete genome information, it is possible to show that a majority of paralogues were created during a short time-interval and are, therefore, probably the result of a complete genome duplication (TABLE 1).

Absolute dating from protein-based distances. Although protein-based distances (distances based on amino-acid differences) are known to vary considerably among proteins, several attempts have been made to use such distances to date duplication events. For example, Vision and colleagues[28] have used amino-acid replacement rates (K_A) to date block-duplication events in *A. thaliana* and concluded that at least four age classes could be defined. These authors assumed that, although the amino-acid replacement rate of different proteins might vary considerably, the overall distribution of amino-acid substitution rates is the same throughout the genome. If that assumption were valid, then any contemporaneously duplicated block that contains several homologous pairs would provide a more or less independent sample of the distribution. Furthermore, the average values of K_A for blocks duplicated at the same time must necessarily be less variable around the true mean than the individual protein values themselves. Nevertheless, it has been shown that protein distances are not reliable for dating duplicated blocks containing heterogeneous classes of proteins. For example, different block duplications in *A. thaliana*, estimated to be of similar age based on mean protein distance[28], actually turned out to be heterogeneous in age when compared to dating based on synonymous substitution rates[52]. The use of synonymous and, consequently, neutral substitutions for evolutionary distance calculations would, therefore, be the more reliable way of estimating duplication events, unless there is no alternative because the duplications are too old.

Dating by phylogenetic means. Another way of dating duplication events is by mapping them onto phylogenetic trees. In relative terms, this approach helps to determine whether duplications have occurred prior

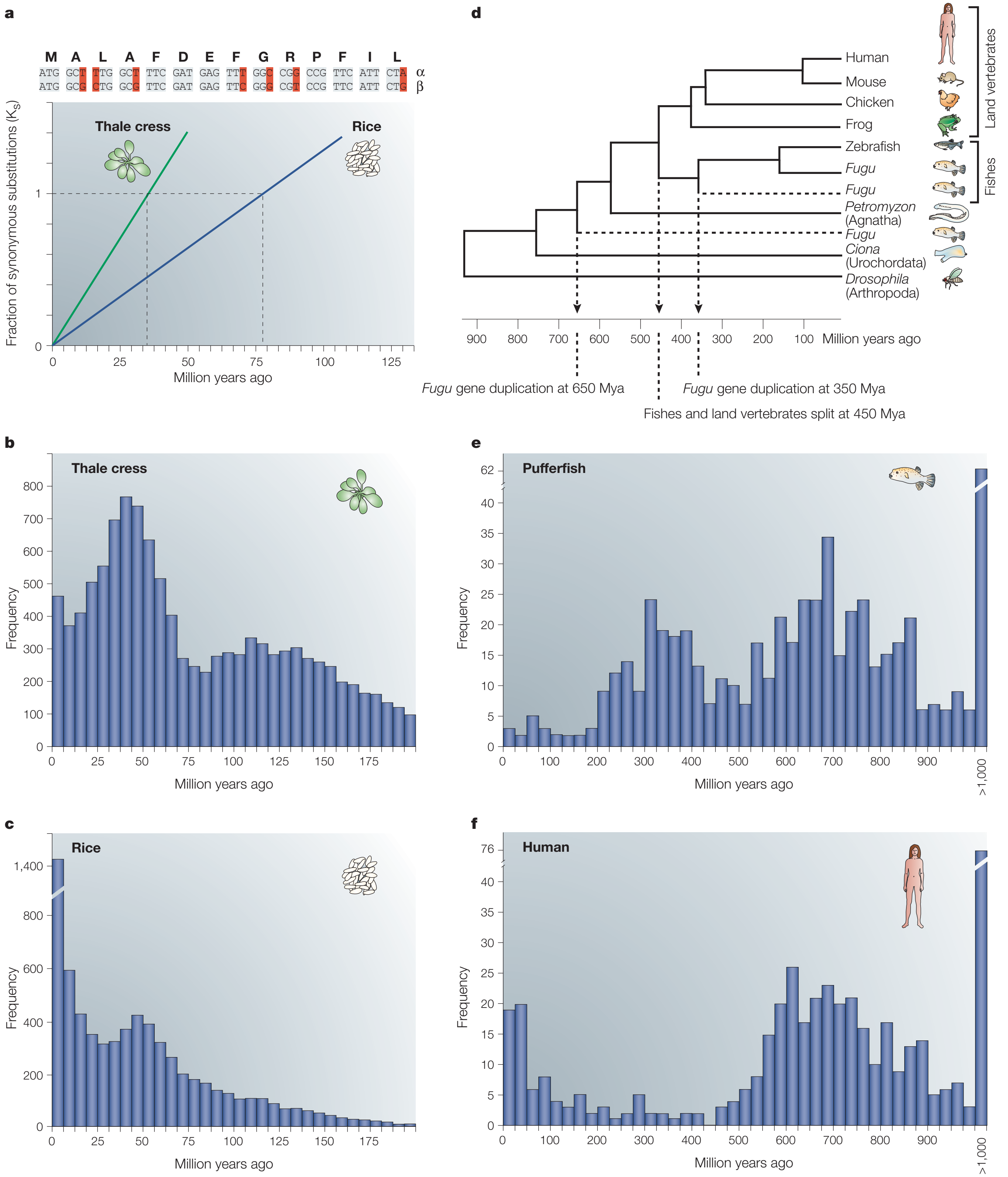

Figure 5 | **Age distributions of duplicated genes.** Absolute dates of duplication events (that is, age of duplicates) can be obtained by estimating the fraction of synonymous substitutions per site (K_S) between duplicated genes (part **a**) or by the constructing linearized trees (part **d**). Parts **b** and **c** show age distributions for the paranomes of *Arabidopsis thaliana* and rice, respectively, based on K_S values (part **a**). Parts **e** and **f** show age distributions for the paranomes of pufferfish and human, respectively, based on linearized phylogenetic trees of vertebrates (part **d**). Due to the fact that many phylogenetic trees cannot be converted into linearized trees[7], because of deviations of the molecular clock, the number of duplicates for which a reliable date can be inferred is much smaller than with K_S-based dating. On the other hand, K_S-based dating can only be used for relatively recent duplication events, owing to saturation of synonymous substitutions (see text for details).

Table 1 | **Overview of the computational approaches applied to eukaryotic genomes and the main conclusions***

Investigated genome	Computational method	Conclusion	Reference
Yeast (*Saccharomyces cerevisiae*)	GHM/collinearity	Whole-genome duplication	27
	Comparative analysis of related genomes		3,36,82,88
Candida glabrata	GHM/collinearity	Whole-genome duplication	88
Thale cress (*Arabidopsis thaliana*)	Inference of ancestral gene order and content	Whole-genome duplication	39
	GHM/collinearity/transitive homology relationships	3 whole-genome duplications	34
	Phylogeny		38
Rice (*Oryza sativa*) and other cereals	GHM/K_S-based age distribution/ relative dating	Partial-genome duplication	57
	GHM/collinearity/phylogeny	Whole-genome duplication	101,102
Maize	K_S-based age distribution for EST data	Whole-genome duplication	58
Poplar‡	K_S-based age distribution for EST data	Whole-genome duplication	Y.V.de P. unpublished results
Tomato and potato‡	K_S-based age distribution for EST data	Whole-genome duplication	58
Soybean‡	K_S-based age distribution for EST data	Whole-genome duplication	58
Cotton‡	K_S-based age distribution for EST data	Whole-genome duplication	58
Vertebrates§	Collinearity/absolute dating based on phylogeny	Whole-genome duplication	2,3,6
	Quadruplicate paralogy (collinearity/gene content)	2 whole-genome duplications	5,40
	Absolute dating based on phylogeny		7
Ray-finned fishes	Relative dating based on phylogeny	Whole-genome duplication	17
	Relative/absolute dating based on phylogeny/GHM		7,103
Drosophila melanogaster	Collinearity/windows-based gene content	Only a few block duplications	41,91
Caenorhabditis elegans	Collinearity/windows-based gene content	Only a few block duplications	41,91

*Mapping data were not considered. ‡Poplar, tomato, potato, soybean, and cotton also shared the two older duplications of *A. thaliana*. §In all vertebrates, also lineage specific segmental duplications have been reported[47,83,84]. GHM, gene homology matrix.

to, or after, a speciation event. If it can be shown that a majority of duplicated genes have been created after one speciation event, but before another one, this could point to a large-scale gene duplication event that has occurred between both speciation events[67,68]. Relative dating has been applied successfully to study large-scale gene duplication events in yeast[69], *A. thaliana*[38], rice[57,70] and ray-finned fishes[17,71]. However, tree topology has also been used as evidence against whole-genome duplications, in particular regarding the proposed genome doublings in early vertebrate evolution (BOX 3). If regions of quadruplicate paralogy are historical remnants of two whole-genome duplications, this should be reflected in the shape of phylogenetic trees drawn from their constituent genes. Indeed, if two TETRAPLOIDY events had occurred in early vertebrate evolution, one would, at first sight, expect symmetrical tree-topologies of the form ((A + B)(C + D)), with the age of the AB split the same as the age of the CD split[21,72]. The fact that many, or even most, tree topologies that are based on duplicated vertebrate genes do not show a 2 + 2 topology is therefore to be considered evidence against the 2R hypothesis[73–75] (BOX 3). However, Furlong and Holland[76] argue that incongruent tree topologies are not in disagreement with sequential genome duplication, but are to be expected when two AUTOTETRAPLOIDY (but not ALLOTETRAPLOIDY) events have taken place in close succession. Gene trees will then simply reflect the random order of DIPLOIDIZATION of chromosomes, rather than the order of chromosome duplication, and tree topologies will, in general, be asymmetrical. There are other reasons why one might infer asymmetrical tree topologies. Gibson and Spring[77] argue that the period between both duplication events that are proposed to have taken place in early vertebrate evolution could have been as short as, or shorter than, 10 million years. In such cases, in particular for sequences that have been duplicated more than about 600 million years ago (see below), gene quartets will not contain adequate phylogenetic signals to resolve internal branches, and inferred tree topologies will be essentially random. In addition, many genes have unequal rates of evolution after duplication[11,36,78,79], making them particularly susceptible to tree reconstruction artefacts[80].

If the timing of a speciation event is known with confidence, gene trees can also be used to infer absolute dates. This is usually performed by the construction of LINEARIZED TREES[81], which assumes equal rates of evolution in different lineages of the tree. To create such linearized trees, relative-rate and branch-length tests for rate heterogeneity are usually applied to these trees to check for deviations from the assumption of a constant MOLECULAR CLOCK. Faster or more slowly evolving sequences are removed, so that the data set contains only sequences evolving at a similar rate. By comparing the divergences of duplicated genes with a fixed calibration point — that is, the date of a speciation event — the absolute date of origin of paralogous genes can be inferred.

TETRAPLOID
A tetraploid organism has four sets of homologous chromosomes, instead of the usual two.

AUTOTETRAPOLYPLOID
Tetraploidy, in which all the chromosomes come from the same species; that is, a tetraploid is formed by the doubling of its own genome.

ALLOTETRAPOLYPLOID
An allotetrapolyploid originates by the fusion of the genomes of two different, but closely-related species.

DIPLOIDIZATION
The evolutionary process whereby a polyploid species becomes a diploid again. The molecular basis of diploidization is not known yet.

LINEARIZED TREE
Linearized trees are phylogenetic trees that assume equal evolutionary rates in different lineages since their divergence from a common ancestor. As such trees assume a clocklike behaviour of the underlying molecular marker, a timescale can be superimposed on them.

Box 3 | 2R or not 2R?

Although based on rather inaccurate indicators such as genome size and ISOZYME complexity, Ohno[1] suggested that the genomes of vertebrates have been shaped by two complete genome duplications, one on the shared lineage leading to both cephalochordates and vertebrates and a second one on the 'fish or amphibian' line. Later, important indications for two rounds (2R) of large-scale gene duplications in early vertebrate evolution came from the analysis of *Hox* genes and Hox-gene clusters[97]. The observation that protostome invertebrates, as well as the deuterostome cephalochordate *Amphioxus*, possess a single Hox cluster, whereas the amphibians, reptiles, birds, mammals and lobe-finned fish (such as the coelacanth and lungfishes) have four clusters[98], supported the 2R hypothesis. Since then, evidence for and against the 2R hypothesis has been put forward, and several modifications have been proposed, assuming a diversity of small- and large-scale gene duplication events[99,100]. Based on QUADRUPLICATE PARALOGY between different genomic segments, some have strongly argued for two rounds of genome duplications[5,40], whereas others, often analysing the same data but using different techniques, found clear evidence for only one genome-doubling event[2,3,6]. Still others reject whole-genome duplications in vertebrates altogether and only accept a continuous rate of gene duplication[41,73]. Recently, it was shown that about 70% of the duplicates found in pufferfish are between 500 and 900 million years old, while duplicates that evolved between 250 and 450 million years ago only account for 30%. Using the fish-specific genome duplication as a benchmark, and assuming equal rates of gene loss throughout the evolution of vertebrates, this again indicates that two genome duplications, rather than one, occurred at the dawn of vertebrate evolution[7].

MOLECULAR CLOCK
The hypothesis that, in any given gene or DNA sequence, mutations accumulate at an approximately constant rate in all evolutionary lineages as long as the gene or the DNA sequence retains its original function.

ISOZYME
Different forms of the same enzyme (synonymous with allozymes), which were used as some of the first biochemically-based genetic markers.

QUADRUPLICATE PARALOGY
Quadruplicate homology, where homology is found between four different genomic segments, is often considered to be evidence for two rounds of large-scale gene duplication (for example, Hox clusters).

Recently, absolute dating has been applied to the human and pufferfish (*Fugu rubripes*) genomes, by comparing the divergence of duplicated genes in both genomes with the date of speciation between ray-finned fishes and land vertebrates[7]. Both relative and absolute dating of the complete predicted set of protein-coding genes indicate that initial genome duplications, estimated to have occurred at least 600 million years ago, shaped the genome of all vertebrates[2,3]. In addition, absolute dating and analysis of more than 150 block duplications in the *F. rubripes* genome clearly supports a fish-specific genome duplication, about 320 million years ago, that coincided with the vast radiation of most modern ray-finned fishes[20] (FIG. 5 d).

The comparative approach

Another strategy for uncovering duplicated segments that have become unrecognizable because of differential gene loss is to compare gene order information of one genome with that of the genome of another, related species[35]. Recently, two papers have been published that provide compelling evidence for the genome duplication in *S. cerevisiae* through a comparative analysis with the sequences of related species. Comparison of the *S. cerevisiae* genome with the recently sequenced genomes of the yeast *Kluyveromyces waltii*[36] and the filamentous ascomycete *Ashbya gossypii*[37] clearly showed that almost every region in the two newly sequenced genomes corresponds with two regions of *S. cerevisiae*. Previously, Wong and colleagues[82] reached similar conclusions based on comparing the *S. cerevisiae* genome with partial gene order information from 13 hemiascomycete genomes. They devised so-called proximity plots, which resemble the previously discussed GHMs, but with the difference that in a proximity plot a dot signifies the fact that X and Y are neighbouring genes in another genome. The rationale behind this approach is that for a pair of segments that has undergone considerable differential gene loss, pairs of genes that were neighbours in the ancestral sequence will also show up as diagonal patterns on the proximity plot. Superimposing a proximity plot on a classical GHM can further enhance this diagonal pattern[82]. By combining a proximity plot of partial sequence data from 13 other hemiascomycete yeasts and a GHM, 82% of the *S. cerevisiae* genome was found in duplicated regions, which was a dramatic increase in sensitivity when compared with the previously reported 50% (REF. 27).

Despite its elegance, the comparative approach can also be combined with the other approaches discussed above; this method is limited to related species, because it relies on the assumption that gene order is largely conserved between the genomes in the dataset. In other words, the method is not applicable to genomes that have undergone extensive rearrangements since their divergence or to genomes that have undergone genome duplication a long time ago.

Future outlook

Complete genome sequencing has revealed that many eukaryotic organisms are paleopolyploids that have had their genome duplicated, sometimes more than once. Currently, genome sequences are being determined from many species that are probably also descendants from the same polyploidy events. Also, they might have experienced lineage-specific segmental or whole-genome duplications. Therefore, it is anticipated that more large-scale gene duplication events in eukaryotic genomes will be unveiled and that the detection of such events will soon become standard procedure[44,45,47,83,84]. Reliable identification of duplicated regions is also imperative for understanding the evolution of genome structure and processes such as gene loss, gene retention and gene evolution. It also forms the basis for being able to address questions such as whether gene retention after large-scale gene duplication events is biased with respect to gene functions[11]. Dating of large-scale gene duplication events can bring about possible correlations with biological innovations or speciation events[9]. However, as with all ancient events, reconstruction of what has happened during hundreds of millions of years of genome evolution is not straightforward and demands inventive approaches. This has been demonstrated clearly for the *S. cerevisiae* genome. Although contested for a long time [46,85], the use and combination of different computational and comparative approaches[36,37,82,86–88] has ultimately led to the now generally accepted view that *S. cerevisiae* is indeed an ancient polyploid. The ever-increasing amount of genomic data, the use of robust gene trees[89] and the continuing progress in computational tools to identify genomic homology will probably also strengthen the evidence for other ancient polyploidy events, such as those that are thought to have taken place in ray-finned fishes and early vertebrates.

1. Ohno, S. *Evolution by Gene Duplication* (Springer, New York, 1970).
2. McLysaght, A., Hokamp, K. & Wolfe, K. H. Extensive genomic duplication during early chordate evolution. *Nature Genet*. **31**, 200–204 (2002).
3. 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. *Nature Genet.* **31**, 205–209 (2002).
4. Larhammar, D., Lundin, L.-G. & Hallböök, F. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res*. **12**, 1910–1920 (2002).
5. Lundin, L.-G., Larhammer, D. & Hallböök, F. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. *J. Struct. Funct. Genomics* **3**, 53–63 (2003).
6. Panopoulou, G. *et al.* New evidence for genome-wide duplications at the origin of vertebrates using an *Amphioxus* gene set and completed animal genomes. *Genome Res.* **13**, 1056–1066 (2003).
7. Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A. & Van de Peer, Y. Major events in the genome evolution of vertebrates: paranome age and size differs considerably between fishes and land vertebrates. *Proc. Natl Acad. Sci. USA* **101**, 1638–1643 (2004).
8. Holland, P. W. More genes in vertebrates? *J. Struct. Funct. Genomics* **3**, 75–84 (2003).
9. Aburomia, R., Khaner, O. & Sidow, A. Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail. *J. Struct. Funct. Genomics* **3**, 45–52 (2003).
   These authors devised a method to estimate the amount of change in morphological complexity during vertebrate evolution and noticed that increase in complexity coincided with postulated whole-genome duplication events in early vertebrate evolution.
10. Otto, S. P. & Whitton, J. W. Polyploid incidence and evolution. *Annu. Rev. Genet*. **34**, 401–437 (2000).
11. Blanc, G. & Wolfe, K. H. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16**, 1679–1691 (2004).
12. Amores, A. *et al.* Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**, 1711–1714 (1998).
13. Naruse, K. *et al.* A detailed linkage map of medaka, *Oryzias latipes*: comparative genomics and genome evolution. *Genetics* **154**, 1773–1784 (2000).
14. Elgar, G. *et al.* Generation and analysis of 25 Mb of genomic DNA from the pufferfish *Fugu rubripes* by sequence scanning. *Genome Res*. **9**, 960–971 (1999).
15. Postlethwait, J. H. *et al.* Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* **10**, 1890–1902 (2000).
16. Woods, I. G. *et al.* A comparative map of the zebrafish genome. *Genome Res.* **10**, 1903–1914 (2000).
17. Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. & Van de Peer, Y. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.* **13**, 382–390 (2003).
18. Wittbrodt, J., Meyer, A. & Schartl, M. More genes in fish? *BioEssays* **20**, 511–512 (1998).
19. 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).
20. Postlethwait, J., Amores, A., Cresko, W., Singer, A. & Yan, Y.-L. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet.* (in the press).
21. Wolfe, K. H. Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet*. **2**, 333–341 (2001).
22. Gu, X. & Huang, W. Testing the parsimony test of genome duplications: a counterexample. *Genome Res*. **12**, 1–2 (2002).
23. Seoighe, C. Turning the clock back on ancient genome duplication. *Curr. Opin. Genet. Devel*. **13**, 636–643 (2003).
24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol*. **215**, 403–410 (1990).
25. Vandepoele, K., Saeys, Y., Simillion, C., Raes, J. & Van de Peer, Y. The automatic detection of homologous regions (ADHoRe) and its application to microcollinearity between *Arabidopsis* and rice. *Genome Res*. **12**, 1792–1801 (2002).
26. Simillion, C., Vandepoele, K., Saeys, Y. & Van de Peer, Y. Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res*. **14**, 1095–1106 (2004).
27. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
28. Vision, T. J., Brown, D. G. & Tanksley, S. D. The origins of genomic duplications in *Arabidopsis*. *Science* **290**, 2114–2117 (2000).
29. Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
30. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
    Seminal paper that describes the birth and death rate of genes in eukaryotic genomes. The study describes a continuous mode of gene duplication, the rate of which is similar to nucleotide substitutions.
31. Gibson, T. J. & Spring, J. Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc. Trans*. **28**, 259–264 (2000).
32. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473 (2000).
33. Wagner, A. Selection and gene duplication: a view from the genome. *Genome Biol.* **3**, 1012.1–1012.3 (2002).
34. Simillion, C., Vandepoele, K., Van Montagu, M., Zabeau, M. & Van de Peer, Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **99**, 13627–13632 (2002).
    The *A. thaliana* genome was shown to contain numerous segments that seemed to have been duplicated between five and eight times. This observation can be explained by inferring three, but no more, genome-wide duplication events.
35. Vandepoele, K., Simillion, C. & Van de Peer, Y. Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* through rice. *Trends Genet*. **18**, 606–608 (2003).
36. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617–624 (2004).
    This paper (see also reference 37) uses a non-duplicated genome sequence of a related yeast species to prove the existence of an ancient genome duplication in *S. cerevisiae*.
37. Dietrich, F. S. *et al.* The *Ashbya* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**, 304–307 (2004).
38. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).
39. Blanc, G., Hokamp, K. & Wolfe, K. H. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res*. **13**, 137–144 (2003).
40. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. & Inoko, H. Evidence of *en bloc* duplication in vertebrate genomes. *Nature Genet*. **31**, 100–105 (2002).
41. Friedman, R. & Hughes, A. L. Gene duplication and the structure of eukaryotic genomes. *Genome Res*. **11**, 373–381 (2001).
42. Cavalcanti, A. R., Ferreira, R., Gu, Z. & Li, W.-H. Patterns of gene duplication in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. *J. Mol. Evol*. **56**, 28–37 (2003).
43. Gehring, W. J. *Master Control Genes in Development and Evolution: the Homeobox Story* (Yale Univ. Press, New Haven, 1998).
44. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
45. Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res*. **14**, 789–801 (2004).
46. Koszul, R., Caburet, S., Dujon, B. & Fischer, G. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J*. **23**, 234–243 (2004).
47. Tuzun, E., Bailey, J. A. & Eichler, E. E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res*. **14**, 493–506 (2004).
48. Seoighe, C. & Wolfe, K. H. Updated map of duplicated regions in the yeast genome. *Gene* **238**, 253–261 (1999).
49. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12**, 1093–1101 (2000).
50. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
51. Paterson, A. H. *et al.* Comparative genomics of plant chromosomes. *Plant Cell* **12**, 1523–1540 (2000).
52. Raes, J., Vandepoele, K., Saeys, Y., Simillion, C. & Van de Peer, Y. Investigating ancient duplication events in the *Arabidopsis* genome. *J. Struct. Funct. Genomics*. **3**, 117–129 (2003).
53. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, New York, 2000).
54. Hughes, A. L. *Adaptive Evolution of Genes and Genomes* (Oxford Univ. Press, New York, 1999).
55. Koch, M. A., Haubold, B. & Mitchell-Olds, T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol. Biol. Evol*. **17**, 1483–1498 (2000).
56. Gaut, B. S., Morton, B. R., McCaig, B. C. & Clegg, M. T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl Acad. Sci. USA*. **93**, 10274–10279 (1996).
57. Vandepoele, K., Simillion, C. & Van de Peer, Y. Evidence that rice, and other cereals, are ancient aneuploids. *Plant Cell* **15**, 2192–2202 (2003).
58. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678 (2004).
    An elegant approach to uncover large-scale gene duplication-events based on age distributions of paralogous expressed sequence tags.
60. Zhang, L., Vision, T. & Gaut, B. S. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol*. **19**, 1464–1473 (2002).
61. Li, W.-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**, 96–99 (1993).
62. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
63. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
64. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
65. Conery, J. S. & Lynch, M. Nucleotide substitutions and the evolution of duplicate genes. *Pac. Symp. Biocomput.* **6**, 167–178 (2001).
66. Lynch, M. & Conery, J. S. The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics* **3**, 35–44 (2003).
67. Hughes, A. Phylogenetic tests of the hypothesis of block duplication of homologous genes on chromosomes 6, 9, and 1. *Mol. Biol. Evol*. **15**, 854–870 (1998).
68. Robinson-Rechavi, M., Boussau, B. & Laudet, V. Phylogenetic dating and characterization of gene duplications in vertebrates: the cartilaginous fish reference. *Mol. Biol. Evol*. **21**, 580–586 (2004).
69. Langkjaer, R. B., Cliften, P. F., Johnston, M. & Piskur, J. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* **421**, 848–852 (2003).
70. Chapman, B. A., Bowers, J. E., Schulze, S. R. & Paterson, A. H. A comparative phylogenetic approach for dating whole genome duplication events. *Bioinformatics* **20**, 180–185 (2004).
71. Van de Peer, Y., Taylor, J. & Meyer, A. Are all fishes ancient polyploids? *J. Sruct. Funct. Genomics* **2**, 65–73 (2003).
72. Skrabanek, L. & Wolfe, K. H. Eukaryote genome duplication — where's the evidence? *Curr. Opin. Genet. Dev*. **8**, 694–700 (1998).
73. 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).
74. Martin, A. Is tetralogy true? Lack of support for the 'one-to-four rule'. *Mol. Biol. Evol.* **18**, 89–93 (2001).
75. Hughes, A. L. & Friedman, R. Testing hypotheses of genome duplication in early vertebrates. *J. Struct. Funct. Genomics* **3**, 85–93 (2003).
76. Furlong, R. F. & Holland, P. W. H. Were vertebrates octoploid? *Phil. Trans. R. Soc. Lond. B* **357**, 531–544 (2002).
77. Gibson, T. J. & Spring, J. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet*. **14**, 46–49 (1998).
78. Van de Peer, Y., Taylor, J. S., Braasch, I. & Meyer, A. The ghost of selection past: rates of evolution and functional divergence in anciently duplicated genes. *J. Mol. Evol*. **53**, 436–446 (2001).
79. Zhang, P., Gu, Z. & Li, W.-H. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol*. **4**, R56 (2003).
80. Taylor, S. T. & Brinkmann, H. 2R or not 2R? *Trends Genet*. **17**, 488–489 (2001).
81. Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* **12**, 823–833 (1995).
82. Wong, S., Butler, G. & Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl Acad. Sci. USA* **99**, 9272–9277 (2002).
    A clever approach to use of partial gene order information of related species to demonstrate genome duplication in *S. cerevisiae*.

83. Cheung. J. *et al.* Recent segmental and gene duplications in the mouse genome. *Genome Biol*. **4**, R47 (2003).
84. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol*. **4**, R25 (2003).
85. Llorente, B. *et al.* Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of redundancy compared to *Saccharomyces cerevisiae*. *FEBS Lett.* **22**, 122–133 (2000).
86. Li, W.-H., Gu, Z., Cavalcanti, A. R. O. & Nekrutenko, A. Detection of gene duplications and block duplications in eukaryotic genomes. *J. Struct. Funct. Genomics* **3**, 27–34 (2003).
87. Wolfe, K. Evolutionary genomics: yeasts accelerate beyond BLAST. *Curr. Biol*. **14**, R392–394 (2004).
88. Dujon *et al.* Genome evolution in yeasts. *Nature* **430**, 35–44 (2004).
89. Van de Peer, Y., Frickey, T., Taylor, J. S. & Meyer, A. Dealing with saturation at the amino acid level: a case study involving anciently duplicated zebrafish genes. *Gene* **295**, 205–211 (2002).
90. Durand, D. & Sankoff, D. Tests for gene clustering. *J. Comput. Biol*. **10**, 453–482 (2003).
91. Gu, Z., Cavalcanti, A., Chen, F. C., Bouman, P. & Li, W.-H. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol. Biol. Evol*. **19**, 256–562 (2002).
92. Rost B. Twilight zone of protein sequence alignments. *Protein Eng*. **12**, 85–94 (1999).
93. Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A. Evolutionary analyses of the human genome. *Nature* **409**, 847–849 (2001).
94. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
95. Calabrese, P. P., Chakravarty, S. & Vision, T. J. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* **19** (Suppl. 1), i74–i80 (2003).
96. Durand, D. Vertebrate evolution, doubling and shuffling with a full deck. *Trends Genet*. **19**, 2–5 (2003).
97. Holland, P. W., Garcia-Fernandez, J., Williams, N. A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Dev. Suppl.* 125–133 (1994).
98. Holland, P. W. Vertebrate evolution: something fishy about Hox genes. *Curr. Biol.* **7,** R570–572 (1997).
99. Spring, J. Vertebrate evolution by interspecific hybridization — are we polyploidy? *FEBS Lett*. **400**, 2–8 (1997).
100. Makalowski, W. Are we polyploids? A brief history of one hypothesis. *Genome Res*. **11**, 667–670 (2001).
101. Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA* **101**, 9903–9908 (2004).
102. Guyot, R. & Keller, B. Ancestral genome duplication in rice. *Genome* **47**, 610–614 (2004).
103. Christoffels, A. *et al.* (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol*. **21**, 1146–1151 (2004).

Acknowledgements

I would like to thank all members of the Bioinformatics and Computational Biology Division for their interest and stimulating discussions. I am particularly grateful to Klaas Vandepoele, Cedric Simillion, Dirk Gevers, Jeroen Raes and Kathleen Marchal for critical readings of the manuscript, and to the Department of Plant Systems Biology for continuous support. I also gratefully acknowledge the constructive comments of the three anonymous reviewers.

Competing interests statement

The author declares that he has no competing financial interests.

Online links

FURTHER INFORMATION

BLAST: http://www.ncbi.nlm.nih.gov/BLAST/

Author's laboratory: http://www.psb.ugent.be/bioinformatics/

Access to this interactive links box is free online.