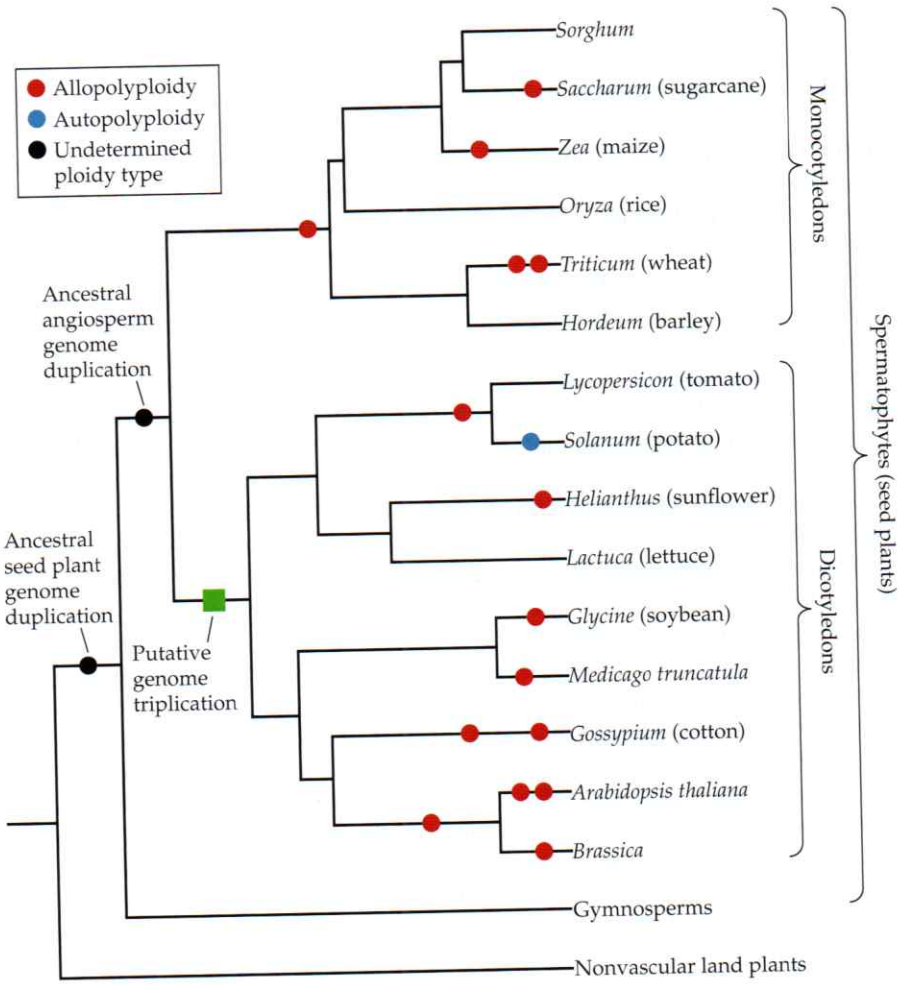

Figure 7.37 Inferred polyploidy events during the evolution of angiosperms. The ancestral angiosperm genome duplication is estimated to have occurred 190–230 million years ago. The ancestral seed plant duplication is estimated to have occurred 320–350 million years ago. (Data from Adams and Wendel 2005 and Jiao et al. 2011.)

S. cerevisiae; and lager, made with the bottom-fermenting yeast *S. pastorianus*. Ale is the older of the two, probably having been produced as early as 7,500 BCE in the Euphrates valley of modern Syria and Iraq (Bamforth 1998; Hornsey 2003). Lager brewing arose in early fifteenth-century Bavaria, gained broad acceptance by the late nineteenth century, and has since become the most popular technique for producing alcoholic beverages. Unlike ales and wines, lagers require slow, low-temperature fermentation that is carried out by cryotolerant *S. pastorianus* strains. Libkind et al. (2011) have shown that the domesticated species *S. pastorianus* was created by the fusion of an *S. cerevisiae* strain used to produce ale and a newly discovered cryotolerant species called *Saccharomyces eubayanus* found in Patagonian forests. The draft genome sequence of *S. eubayanus* is 99.5% identical to the non-*S. cerevisiae* portion of the *S. pastorianus* genome. The 0.5% difference is due to several changes in the genome of *S. eubayanus* following the tetraploidization event, changes almost certainly driven by the stringent artificial selection in the brewing environment.

CONSEQUENCES OF POLYPLOIDY At a phenotypic level, the effects of polyploidization are often mild and idiosyncratic (Otto 2007). In some cases, moreover, polyploidy seems to have almost no effect on the phenotype. For example, diploid, autopolyploid, and allopolyploid *Chrysanthemum* species vary in chromosome number from 18 to 198, yet they are almost indistinguishable from one another. Similar observations have been made in roses (*Rosa*), leptodactylid toads (*Odontophrynus*), and goldfish (*Carassius*).

Cell volume generally rises with increasing genome size (Cavalier-Smith 1978; Gregory 2001), although the exact relationship between ploidy and cell volume var-

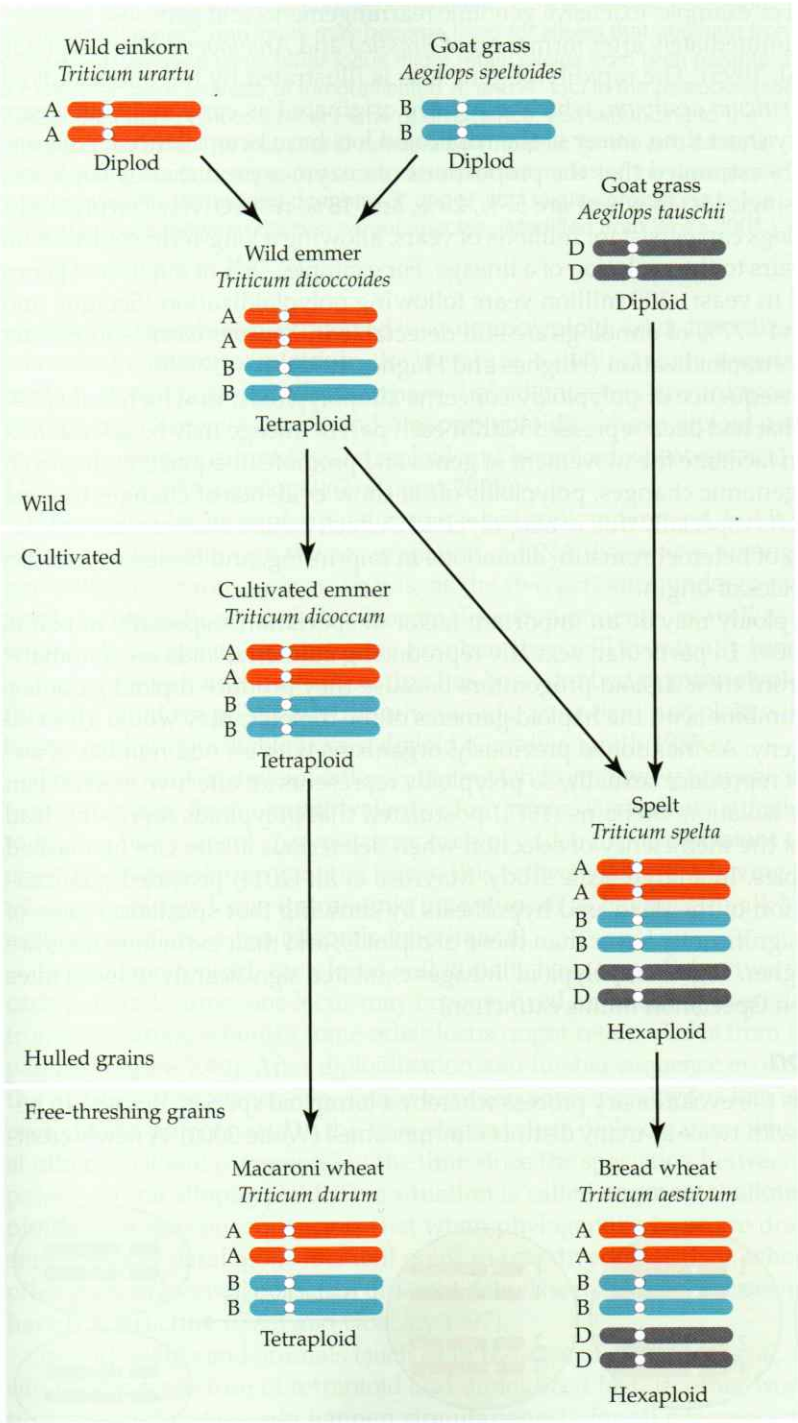

Figure 7.38 The evolution of wheat from diploid grasses to the free-threshing tetraploid macaroni (or durum) wheat and the hexaploid bread wheat. Molecular data and molecular clock considerations indicate that *Triticum urartu* and *Aegilops speltoides* diverged from a common ancestor ~6.5 million years ago. *Aegilops tauschii* turned out to be a homoploid derived through a hybridization between a relative of *T. urartu* and a relative of *Ae. speltoides* ~5.5 million years ago. The tetraploidization event giving rise to wild emmer, *Triticum dicoccoides*, occurred ~800,000 years ago. *Triticum dicoccum* was cultivated as early as 17,000 years ago. Cultivated emmer has larger grains than wild emmer, a fact attributed to selection. The hybridization of *T. dicoccoides* with *Ae. tauschii* that gave rise to the hexaploid *Triticum spelta* occurred less than 400,000 years ago, and *T. spelta* was cultivated ~10,000 years ago. Cultivated emmer and spelt separately acquired a mutation that changed how grains grew in the ears, from being enclosed by hard shells to being looser and easier to thresh. This yielded bread wheat and macaroni wheat, free-threshing wheats that are about 8,500 years old. (Data from Dvorak and Akhunov 2005 and Marcussen et al. 2014.)

ies among environments and among taxa. Interestingly, although average cell size is larger in polyploids, the size of the adult polyploidy organism may or may not be altered; as a rough generalization, polyploidization is more likely to increase adult body size in plants and invertebrates than in vertebrates (Otto and Whitton 2000; Gregory and Mable 2005). The poor correlation between cell size and organismal size was even remarked upon by Albert Einstein, who stated, "Most peculiar for me is the fact that in spite of the enlarged single cell, the size of the animal is not correspondingly increased" (Fankhauser 1972).

An important feature of many newly formed polyploids is that their genomes are unstable and undergo rapid repatterning and segmental loss (Feldman et al. 1997;

Wendel 2000). For example, extensive genomic rearrangements and gene loss were recorded almost immediately after formation of *Brassica* and *Arabidopsis* allopolyploids (e.g., Song et al. 1995). The rapidity of gene loss is illustrated by the allohexaploid bread wheat, *Triticum aestivum*, which may have originated as early as 10,000 years ago. In this very short time, many of the triplicated loci have been silenced. Aragoncillo et al. (1978) estimated that the proportions of enzymes produced by triplicate, duplicate, and single loci in wheat are 57%, 25%, and 18%, respectively. Surprisingly, however, ohnologs can persist for millions of years, allowing a long-term contribution of these gene pairs to the evolution of a lineage. For example, ~8% of duplicated genes have remained in yeast ~100 million years following polyploidization (Seoighe and Wolfe 1999), and ~77% of ohnologs are still detectable in *Xenopus laevis* ~30 million years after allotetraploidization (Hughes and Hughes 1993).

Another consequence of polyploidy concerns allopolyploids, in which transposable elements that had been repressed within each parent lineage may be activated in hybrids and can facilitate the movement of genes and promote unequal crossing over. In addition to genomic changes, polyploids often show evidence of changes in gene expression. This is especially true of allopolyploids, which exhibit changes in methylation, disruption of heterochromatin, alterations in imprinting, and biased expression of genes by species of origin.

Finally, polyploidy may be an important factor in speciation, especially in plants (Wood et al. 2009). In particular, sexually reproducing autotetraploids are automatically isolated from their diploid progenitors because they produce diploid gametes; were these to combine with the haploid gametes of the diploids, they would give rise to triploid progeny. As mentioned previously, organisms with an odd number of autosomes cannot reproduce sexually, so polyploidy represents an effective mechanism of reproductive isolation. Stebbins (1971) postulated that polyploids represent dead ends because of the inefficiency of selection when deleterious alleles can be masked by multiple copies. In a large-scale study, Mayrose et al. (2011) provided quantitative corroboration of the dead-end hypothesis by showing that speciation rates of polyploids are significantly lower than those of diploids, and their extinction rates are significantly higher. Together, polyploid lineages exhibited significantly reduced rates of diversification (speciation minus extinction).

Diploidization

Diploidization is the evolutionary process whereby a tetraploid species "decays" to become a diploid with twice as many distinct chromosomes (Wolfe 2001). A newly creat-

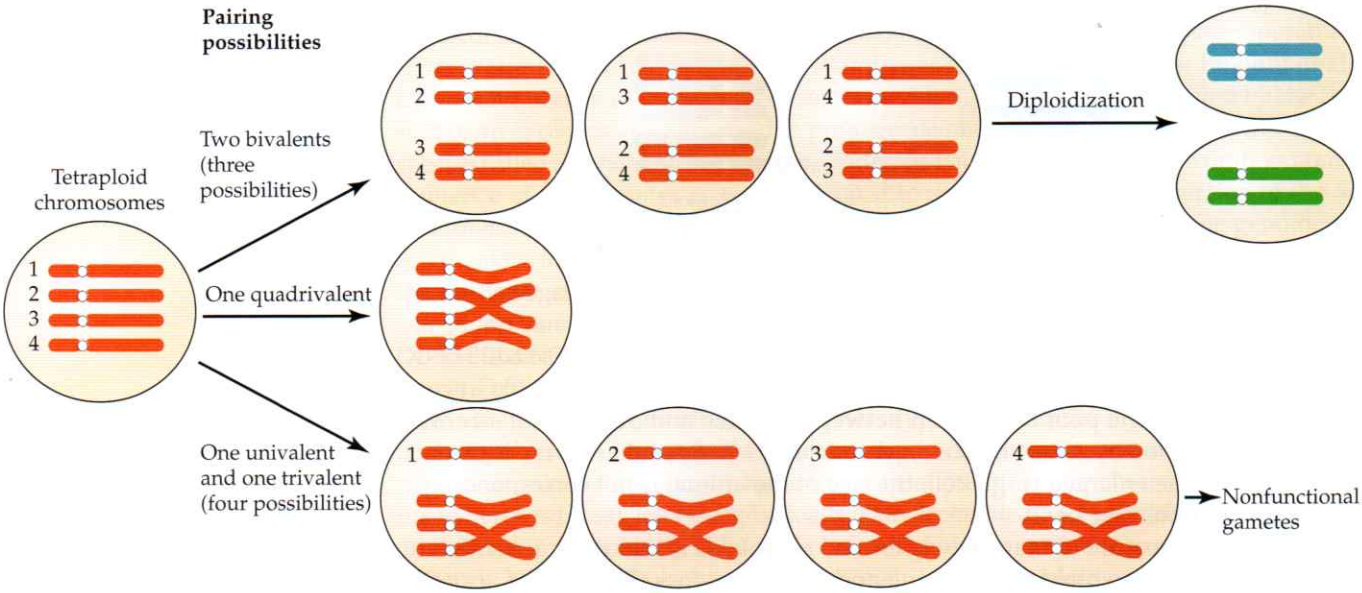

Figure 7.39 Immediately after tetraploidization, the homologous and homeologous chromosomes can exchange genetic material through recombination. The four homologous and homeologous chromosomes (numbered) may pair as two bivalents (three possible combinations) or a quadrivalent. Both pairing types can yield functional gametes. The four chromosomes may also pair as a univalent and a trivalent (four possible combinations), yielding nonfunctional gametes. As the homeologous chromosomes diverge from each other, they can no longer form a quadrivalent, and recombination is restricted to the homologous chromosomes in the bivalents.

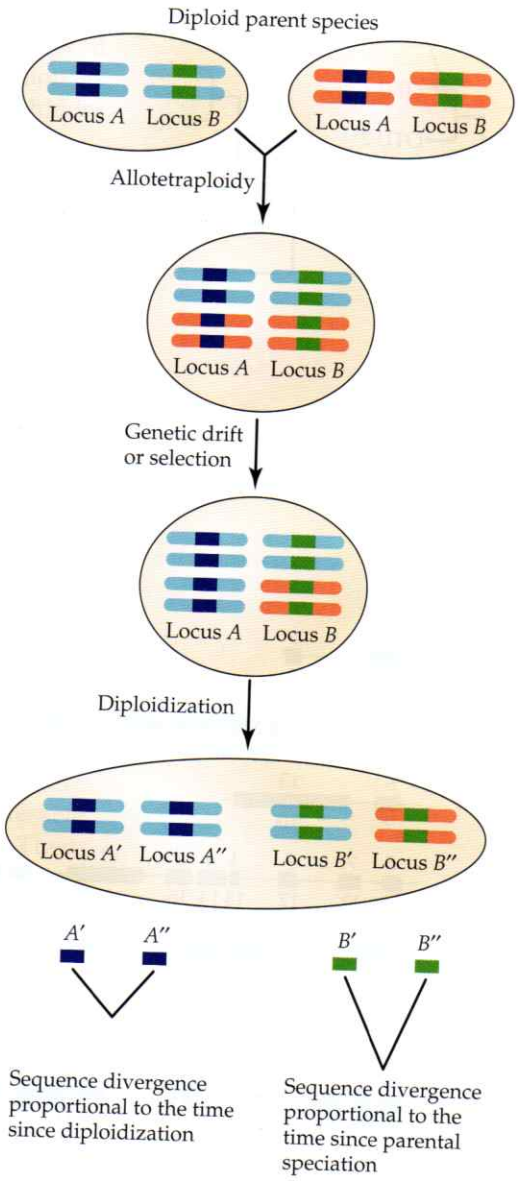

Figure 7.40 Diploidization and its phylogenetic consequences. Following allotetraploidization, one locus may become fixed for alleles that originate from one parent, whereas some other locus might retain alleles from both parents, so a molecular clock analysis of the duplicated *A′* and *A″* loci in the paleopolyploid descendant may point to a recent time of divergence, corresponding to the diploidization date. The locus for *B* remains polymorphic for the two parental alleles during the tetrasomic phase, so molecular estimates of the divergence time between its diploidized daughter *B′* and *B″* loci might correspond to the speciation date between the two parent species. (Modified from Wolfe 2001.)

ed polyploid is sometimes referred to as **neopolyploid**; after diploidization it is called a **paleopolyploid**. In the literature, a distinction is sometimes made between paleopolyploids, whose diploid ancestors are unknown or extinct (e.g., *Xenopus laevis*), and **mesopolyploids**, whose diploid ancestors are known or extant (e.g., tetraploid and hexaploid wheat species) (see Mangenot and Mangenot 1962; Guerra 2008).

The molecular basis of diploidization is not well understood, but it presumably occurs through the accumulation of DNA differences among the homeologous chromosomes. That is, as the two genomes undergo mutations, translocations, and chromosomal rearrangements, as well as unbalanced changes in chromosome number, they will eventually become a single new genome, a situation that has been dubbed **cryptopolyploidy** (literally, a hidden polyploidy). In other words, an ancient polyploid will no longer be distinguishable from a diploid (Cavalier-Smith 1985a).

The key event in diploidization is the switch from having four chromosomes that may form a **quadrivalent** or four types of trivalents at meiosis, to having two pairs of chromosomes each of which forms a **bivalent** (Figure 7.39). 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**).

In an allotetraploid, each locus will initially have four alleles, two from each parent. In time, one locus may become fixed for alleles that originate from one parent, whereas some other locus might retain alleles from both parents (Figure 7.40). After diploidization and further sequence evolution, the amount of sequence divergence between some paralogous loci is expected to 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 when phylogenetic trees are drawn, some pairs of paralogous loci will point to one divergence date, whereas other pairs of loci will point to a different date. The maize genome seems to have this structure (Gaut and Doebley 1997).

In both plants and animals (such as in the case of salmonid taxa), a single species can harbor 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, tree-based analyses of paleopolyploids may yield very confusing results. 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.

Distinguishing between gene duplication and genome duplication

Most genomes contain gene duplications. They can be the result of either (1) gene or segmental duplications or (2) whole-genome duplications. How can one distinguish between the two mechanisms? We note that we are mainly concerned with ancient polyploidizations rather than recent ones (such as that in wheat), in which the

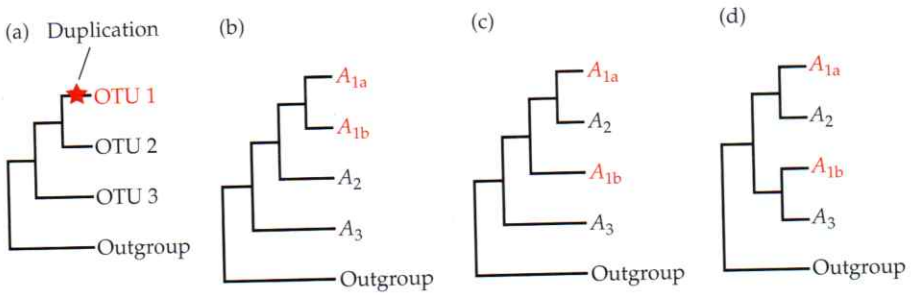

Figure 7.41 (a) Phylogenetic tree for three OTUs and an outgroup, in which a genome duplication (red star) occurred in the lineage leading to OTU 1. (b) The expected phylogenetic positions for duplicated genes A_{1a} and A_{1b} from OTU 1 if indeed they are ohnologs. If, however, A_{1a} and A_{1b} from OTU 1 are not ohnologs, they may be found in different phylogenetic positions, such as in trees (c) or (d).

differences are trivially simple. Several telltale signs have been proposed as evidence of ancient polyploidization.

A good indication of polyploidy would be if all gene trees exhibited the expected topology. For illustration, let us assume that the species tree is as in **Figure 7.41a** and that a genome duplication occurred in the lineage leading to OTU 1, leading to the duplication of all the genes in the genome. In such a case, each of the paralogous genes should yield the same tree (**Figure 7.41b**). If, however, the paralogous genes yield alternative trees (**Figure 7.41c,d**), then it is unlikely that all the gene duplications occurred at the same time. We need, however, to add a caveat, to the effect that the

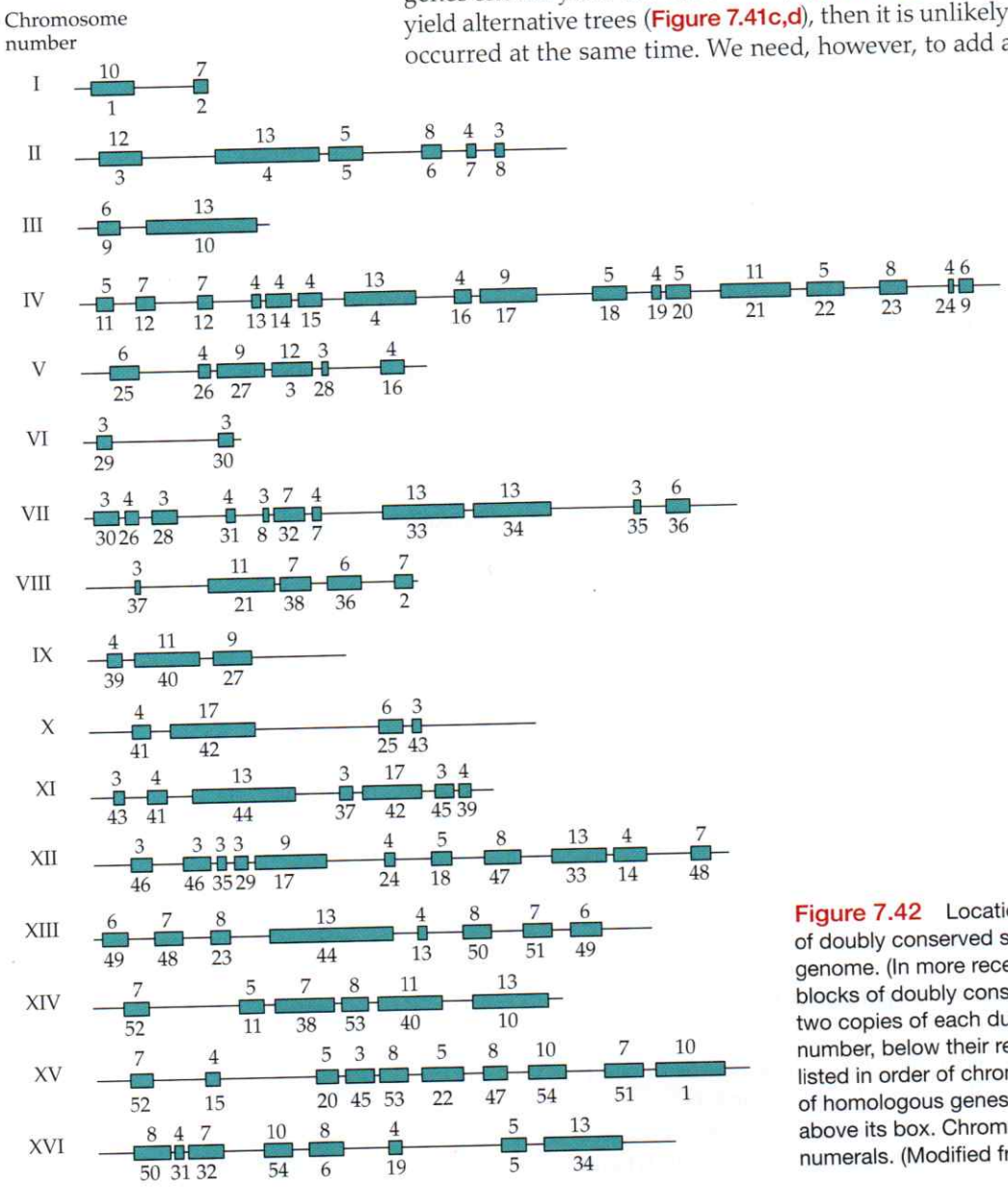

Figure 7.42 Locations of 54 nonoverlapping blocks of doubly conserved synteny (blue boxes) in the yeast genome. (In more recent studies, a greater number of blocks of doubly conserved synteny were found.) The two copies of each duplicated region are given the same number, below their respective boxes. Numbers are listed in order of chromosomal occurrence. The number of homologous genes in each duplicated region is listed above its box. Chromosome numbers are given in roman numerals. (Modified from Wolfe and Shields 1997.)

expectation that all gene pairs yield the same phylogenetic topology assumes that diploidization occurred at the same time in all the genes. Moreover, gene loss and processes of recombination, such as unequal crossing over and gene conversion, as well as gene duplication that occurs subsequent to genome duplication, may hinder the expected result.

Another piece of information that is typically adduced as evidence of genome duplication concerns regions of double synteny, i.e., two or more genomic regions containing paralogous arrays of genes. Here, we present two cases in which the hypothesis of whole-genome duplication was investigated.

THE YEAST GENOME: TETRAPLOIDY OR REGIONAL DUPLICATIONS? *Saccharomyces cerevisiae* has long been suspected of being a cryptotetraploid (Smith 1987). Wolfe and Shields (1997) systematically searched the complete yeast proteome for regions of double synteny. The criteria used for defining two regions as duplicated were (1) a sequence similarity between the two regions associated with a probability of less than 10^{-18} that it was fortuitous; (2) at least three protein-coding genes or open-reading frames in common, with intergenic distances of less than 50 Kb; and (3) conservation of gene order and relative orientation of the genes. According to these criteria, Wolfe and Shields (1997) identified 54 nonoverlapping pairs of duplicated regions spanning about 50% of the yeast genome (Figure 7.42). Of the approximately 5,800 genes in the yeast genome, about 900 were paralogs located in duplicated chromosomal regions (also called **blocks of doubly conserved synteny**, or **paralogons**).

There are two possible explanations for these observations. Either (1) the duplicated regions formed independently by many regional duplications occurring at different times during the evolution of *S. cerevisiae*, or (2) the duplicated regions were produced simultaneously by a single tetraploidization event, followed by massive rearrangements of the genome and loss of many redundant duplicate genes. There are two reasons to favor the latter model. First, 50 of the duplicated regions have maintained the same orientation with respect to the centromere. Second, based on a Poisson distribution, 54 independent regional duplications would be expected to result in about seven triplicated regions (i.e., duplicates of duplicates), but none was observed.

Wolfe and Shields (1997) proposed that *S. cerevisiae* is an ancient tetraploid, formed through the fusion of two ancestral diploid yeast genomes, each containing about 5,000 genes. They estimated the tetraploidization event to have occurred approximately 100 million years ago in the ancestor of four *Saccharomyces* species. The new species then became a cryptotetraploid, and about 90% of the duplicate gene copies were lost through sequence decay or deletion. Some 70–100 subsequent map disruptions (i.e., regional translocations) were inferred to have been required to explain the current chromosomal distribution of the duplicate genes (Seoighe and Wolfe 1998).

These conclusions were almost immediately challenged because they were based on a small set of yeast genes (less than 10%) and could be explained almost as easily by independent duplications of chromosomal segments (e.g., Koszul et al. 2004; Martin et al. 2007).

The clearest way to prove the existence of an ancient polyploidization event would be to find a species that descended directly from a common ancestor along a lineage that diverged before the whole-genome duplication event. Phylogenetic data indicate that *Kluyveromyces waltii* and its paraphyletic congeneric *K. lactis* diverged from the lineage leading to *S. cerevisiae* and the related yeast *S. bayanus* before the polyploidization event (Figure 7.43).

The expected genomic signature of whole-genome duplication is illustrated in Figure 7.44. Following duplication, sister regions would undergo gene loss by deletion; one or the other of the two paralogous copies of each gene would be lost in most cases, with both paralogs being retained only very rarely. Eventually, the only residual signature that two regions arose from ancestral duplication would be the presence of a few paralogous genes in the same order and orientation scattered amidst a multitude

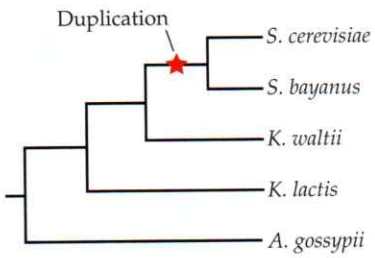

Figure 7.43 Schematic phylogenetic tree for five yeast species: *Saccharomyces cerevisiae*, *S. bayanus*, *Kluyveromyces waltii*, *K. lactis*, and *Ashbya gossypii*. The whole-genome duplication event is marked with a red star. The consequencies of such an event are illustrated in Figure 7.44.

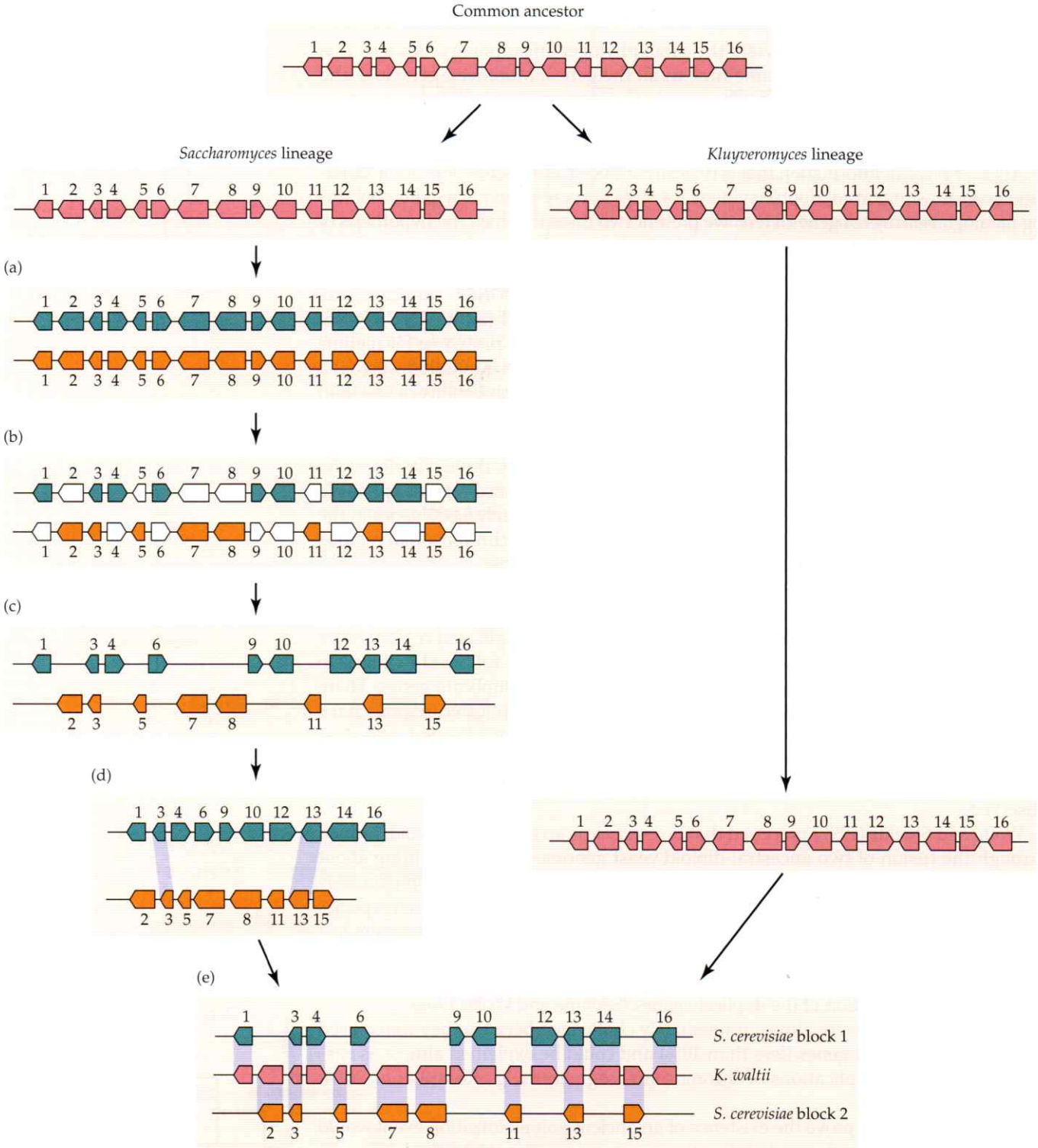

Figure 7.44 Consequences of genome duplication and subsequent gene loss in a comparison between a species that underwent duplication and one that diverged before the duplication event. (a) After divergence from *Kluyveromyces waltii*, the *Saccharomyces* lineage underwent a genome duplication event, creating two copies of every gene and chromosome. (b) The vast majority of duplicated genes underwent mutation and gene loss. (c) Sister segments retained different subsets of the original gene set, keeping two copies for only a small minority of duplicated genes. (d) Within *S. cerevisiae*, the only evidence comes from the conserved order of duplicated genes (numbered 3 and 13) across different chromosomal segments; the intervening genes are unrelated. (e) Comparison with *K. waltii* reveals the duplicated nature of the *S. cerevisiae* genome, interleaving genes from sister segments on the basis of the gene order in *K. waltii*, which is supposed to reflect the ancestral gene order. Spacing between *S. cerevisiae* genes is set to match *K. waltii* chromosomal positions. The pointed side of each gene denotes the direction of its transcription. In the *Saccharomyces* lineages, duplicated blocks are in blue and orange, and nonfunctionalized genes about to decay into pseudogenes or be deleted are in white. (From Kellis et al. 2004.)

of unrelated genes. Paired regions containing such signatures would be relatively short because chromosomal rearrangements would have disrupted global gene order over time, leaving only weak evidence of ancestral duplication.

As *K. waltii* diverged before the whole-genome duplication, it would be expected to display a 1:2 mapping pattern with *S. cerevisiae*. This synteny map should have the following properties: (1) nearly every region in *K. waltii* should correspond to two sister regions in *S. cerevisiae*; (2) nearly every region of *S. cerevisiae* should correspond to one region of *K. waltii*; and (3) the two sister regions in *S. cerevisiae* should each contain an ordered subsequence of the genes in the corresponding region of *K. waltii*, resulting in an interleaving pattern as in Figure 7.44.

Kellis et al. (2004) identified a total of 253 blocks of doubly conserved synteny containing 75% of *K. waltii* genes and 81% of *S. cerevisiae* genes. A detailed view of a region of chromosome 1 of *K. waltii* and the cross-species mapping of genes from two blocks in *S. cerevisiae*, one from chromosome 4 and the other from chromosome 12, are shown in **Figure 7.45**. This picture repeated itself with virtually every block of doubly conserved synteny.

The picture that emerges is that a genome duplication event doubled the number of chromosomes in the *Saccharomyces* lineage, but subsequent gene events led to the current *S. cerevisiae* genome, which is only 13% larger than that of *K. waltii* and contains only 10% more genes. The polyploid genome returned to functional normal ploidy through a large number of deletion events. In principle, gene loss can occur by large segmental deletions or individual gene deletions, and it can either be balanced between the two sisters or act primarily on one of them. Analysis of duplicated blocks revealed that gene loss occurred by many small deletions (the average size of a lost segment is two genes) and was typically balanced between the two sister regions (average balance 57% to 43%). A similar pattern of gene loss was observed in the evolution of the ciliate *Paramecium tetraurelia*, whose nearly 40,000 genes arose through at least three successive whole-genome duplications (Aury et al. 2006).

VERTEBRATE POLYPLOIDY? THE 2R HYPOTHESIS Based on the then common belief that vertebrates possess more genes than invertebrates and his belief that gene duplications are invariably maladaptive, Ohno (1970) suggested that one or more genome duplications occurred in the lineage leading to vertebrates. He was not very explicit about the number and timing of the genome duplications. Following the discovery

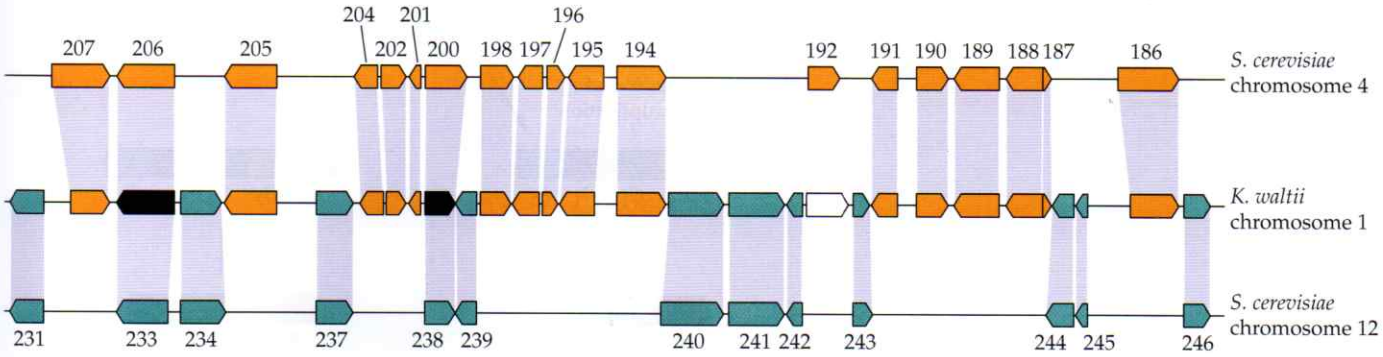

Figure 7.45 A comparison of a portion of chromosome 1 of *Kluyveromyces waltii* with two blocks of doubly conserved synteny in *Saccharomyces cerevisiae*, one from chromosome 4 (orange) and one from chromosome 12 (blue). Genes are arbitrarily numbered as in the *Saccharomyces* Genome Database (Cherry et al. 1997), with the pointed sides indicating the direction of transcription. Each gene of *K. waltii* is shown in red or blue if it has a match in *S. cerevisiae* chromosomes 4 or 12, respectively; in black if it has a match in both *S. cerevisiae* chromosomes; and in white if there is no match in any of the two syntenic blocks. Spacing between *S. cerevisiae* genes is set to match *K. waltii* chromosomal positions. (Modified from Kellis et al. 2004.)

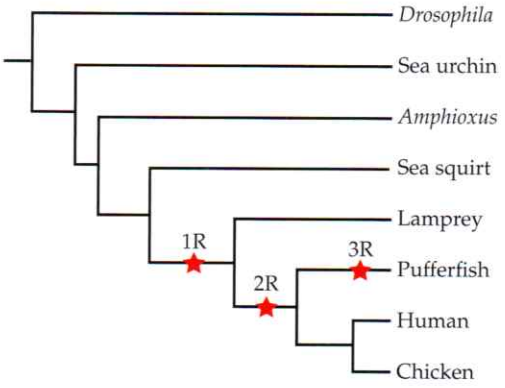

Figure 7.46 The timing of three putative whole-genome duplications in the evolution of deuterostomes, with a protostome (*Drosophila*) as outgroup. Genome duplications 1R and 2R mark the two duplications assumed by the 2R hypothesis. The 3R marks a genome duplication assumed to have occurred in teleost fishes.

that the four Hox gene clusters in mammals had evolved from an ancestral cluster by quadruplication, Kappen et al. (1989) suggested that two polyploidization events occurred in the vertebrate lineage. Holland et al. (1989) placed the first putative duplication event on the lineage leading to the vertebrates after the divergence of the amphioxus (*Branchiostoma lanceolatum*) and sea squirts (genus *Ciona*) (**Figure 7.46**). The second putative duplication was placed on the lineage leading to the jawed vertebrates (Gnathostomata) after the divergence of hagfish and lampreys (Cyclostomata). In time, the assumption of two rounds of tetraploidization on the evolutionary branch leading to jawed vertebrates came to be known as the "2R hypothesis" (Hughes 1999), leading to some memorable puns, including the Hamletian "2R or not 2R" (Hughes and Friedman 2003). A 2R scenario is shown in **Figure 7.47**. In this particular scenario, deletions of paralogs may occur subsequent to genome duplication. Within vertebrates, an additional genome duplication may have occurred in bony fishes (Figure 7.46; Jaillon et al. 2004).

Let us now examine a simple expectation of the 2R hypothesis. In the absence of any gene duplications prior to, in between, or subsequent to the two rounds of genome duplication, the expectation is an (*AB*)(*CD*) topology, where *A*, *B*, *C*, and *D* are paralogous genes (**Figure 7.48a**). Any other topology (e.g., **Figure 7.48b,c**) may be interpreted as a refutation of the 2R hypothesis. Hughes and Friedman (2003) looked for four-member paralogous gene families in the human genome that have a single homolog in *Drosophila* and reconstructed their phylogenetic relationships. Of the 92 resolved phylogenetic topologies, only 22 (24%) supported the 2R hypothesis. Out of the 53 phylogenies in which all internal branches received statistically significant support, only 11 topologies (21%) supported the 2R hypothesis, leading the authors to "decisively" reject the hypothesis. We remind the reader, however, of the very strict assumptions of this test. If one selects only those human paralogs that duplicated before divergence of tetrapods from the bony fishes, the 2R hypothesis cannot

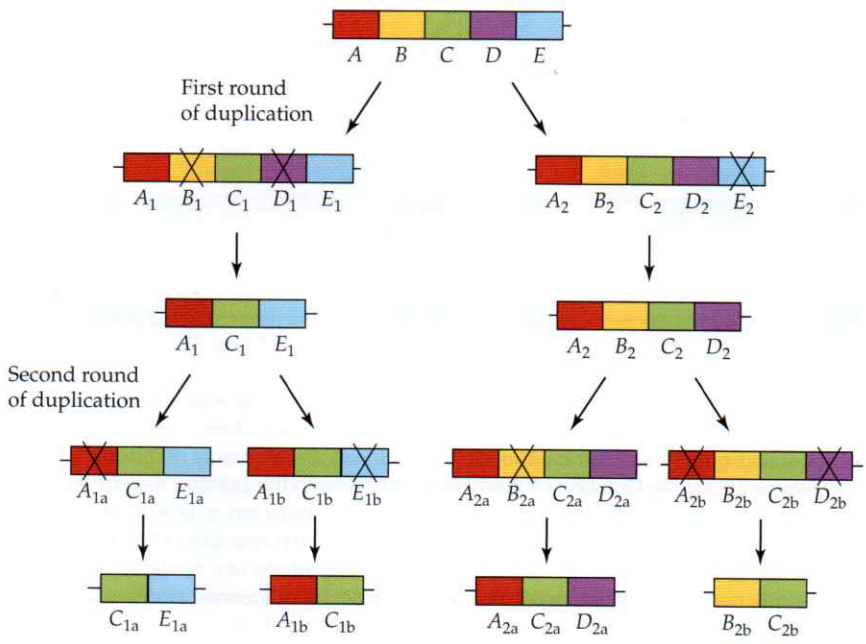

Figure 7.47 Model of the 2R hypothesis (with deletions) and its phylogenetic implications. Starting with five linked genes (*A–E*; top row), two rounds of gene duplication result in nine genes (bottom row). At least one copy of each ancestral gene is present in the genome at all times. Only the descendants of gene *C* retained all four resultant copies. (Modified from Wolfe 2001.)

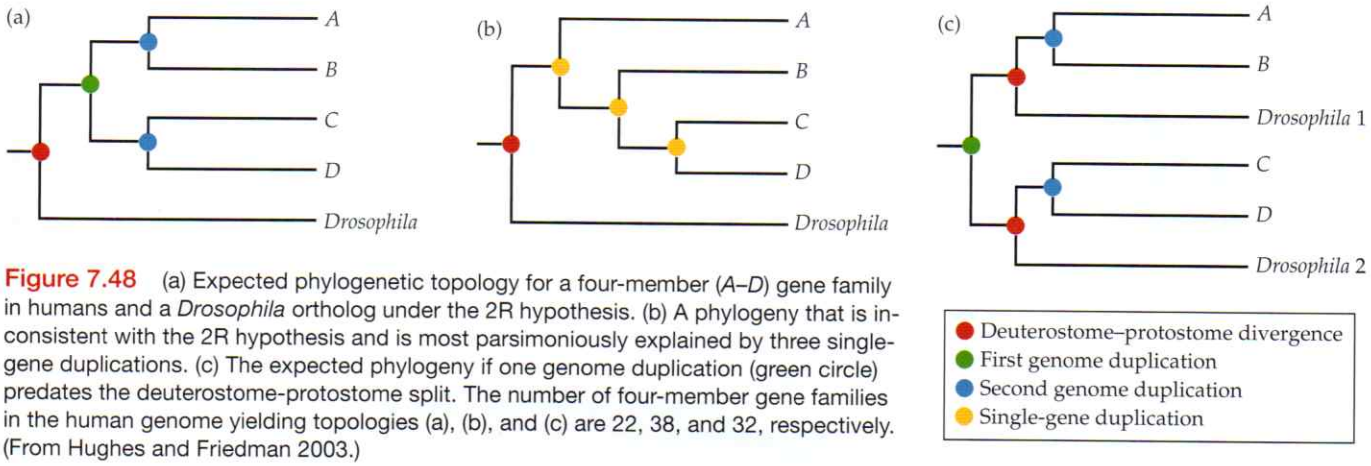

Figure 7.48 (a) Expected phylogenetic topology for a four-member (*A–D*) gene family in humans and a *Drosophila* ortholog under the 2R hypothesis. (b) A phylogeny that is inconsistent with the 2R hypothesis and is most parsimoniously explained by three single-gene duplications. (c) The expected phylogeny if one genome duplication (green circle) predates the deuterostome-protostome split. The number of four-member gene families in the human genome yielding topologies (a), (b), and (c) are 22, 38, and 32, respectively. (From Hughes and Friedman 2003.)

be rejected (Dehal and Boore 2005). McLysaght et al. (2002) and Hokamp et al. (2003) found paralogons covering over 44% of the human genome, many more than would be expected by chance. Moreover, molecular clock analyses of all protein families in humans that have orthologs in the fruit fly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans* indicated that a burst of gene duplication activity took place 350–650 million years ago. These findings together lend support to the hypothesis that one or more polyploidization events occurred early in the vertebrate lineage.

Thus, although the 2R hypothesis remains highly contentious, and may not be solvable solely by phylogenetics, the possibility remains that vertebrates (including the readers of this book) may in fact be cryptooctoploids, and the fish they consume cryptohexadecaploids!