Computational analyses of ancient polyploidy

Kevin P. Byrne[1] and Guillaume Blanc[2*]

[1] Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland.
[2] Laboratoire Information Génomique et Structurale, Centre National de la Recherche Scientifique UPR 2589, 13402 Marseille Cedex 20, France

Keywords: polyploidy, genome duplication

Running title: Computational analyses of ancient polyploidy

Corresponding author: Guillaume Blanc
E-MAIL g_blanc@univ-perp.fr
FAX +33-4-91164549

ABSTRACT

Whole genome duplication has played a major role in the evolution of many eukaryotic lineages. Polyploidy has long been postulated as a powerful mechanism for evolutionary innovation, and recent analyses have provided convincing evidence that independent ancient genome duplications occurred in the ancestors of yeast, plants, vertebrates and fish. It is the growing availability of whole genome sequences that has facilitated the detection and analysis of these polyploidizations. However, because polyploidy is often followed by massive gene loss and chromosomal rearrangements, identifying such events is not always easy. Here we review a wide array of computational methods of ever-increasing sophistication developed to identify the obscured traces of ancient polyploidy in genomic sequences. These methods use a variety of analytical approaches, including comparative genomics, phylogenetics and molecular clock analyses. We also review recent research on the long-term evolution of genes and genomes duplicated by polyploidy. This has emerged as a fruitful field, utilizing genome-wide functional information and genomic sequence data to further our understanding of the impact of polyploidy on organismal biology and evolution.

INTRODUCTION

Modern interest in polyploidy is rooted in Ohno's [1] proposal that the easiest way to create new genes is to duplicate old ones, but ideas about the role and significance of polyploidy go back right through the 20th century [2]. Ohno's proposal that two or more rounds of polyploidy occurred during early vertebrate evolution still arouses much debate and active research. Polyploidy is thought to be important because it results in the duplication of all genes, making it a potentially powerful engine of evolutionary novelty.

The detection of crytic polyploidy has been one of most productive areas of genome evolution research over the last ten years and is likely to continue to help deciphering the complex nature of eukaryotic genomes. Put simply polyploidy is when a nucleus contains more than two copies of each chromosome, as a result of a whole genome duplication event (WGD). Polyploidy can be one of two different types depending on the origin of the duplicated genomes [2]: Autopolyploids result from somatic doubling or the fusion of unreduced gametes within a species, and therefore possess duplicated sets of undifferentiated homologous chromosomes. In autotetraploids, chromosomes form tetravalents during meiosis. Each individual locus exhibits tetrasomic inheritance and contains four alleles. Allopolyploids arise from hybridization between different but often closely related species. When the genomes of the diploid parental species are sufficiently differentiated, the duplicated chromosome sets in the allotetraploid form bivalents at meiosis and loci exhibit disomic inheritance like for diploids. One particular type of allopolyploidy, segmental allotetraploidy, arises from the hybridization of species with only partially differentiated chromosome sets. They thus exhibit a mixture of bivalent and tetravalent formation during meiosis [3].

Polyploidy has been observed and documented in a wide range of eukaryotic species, because microscopes have enabled scientists to observe, count and compare chromosomes [2]. The polyploidy events that most interest bioinformaticians are undetectable using classical microscopy approaches because they occurred several million years ago. During the evolutionary periods that separate polyploid ancestors from their extant progeny, ancient polyploid genomes generally undergo extensive chromosomal rearrangements (including inversion, insertion, translocation and massive gene loss) blurring the traces of the WGD event. These rearrangement processes are

collectively referred as diploidization [4], because the polyploid genome progressively returns to a diploid state. Thus, although organisms that exhibit evidence of ancient WGD events are called paleopolyploids, their genomes can and often do behave like diploids. Although historically genetic mapping data and isozyme electrophoresis were first used to infer paleopolyploidy [2], it is important to note that it is sequencing and post genomic data that has lead to the recent explosion of this domain of research. Early computational approaches focused on detecting evidence of polyploidy in genome sequences, but more recently studies focusing on post-polyploidy genome and gene evolution have become areas of interest as well.

Polyploidy was long considered likely in the vertebrate lineage [1, 5], but was unexpected for small genomes like those of *Saccharomyces cerevisiae* [6] and *Arabidopsis thaliana* [7, 8] . The detection of polyploidy in these lineages was one of the first surprises resulting from the sequencing of small eukaryotic genomes, and bioinformaticians have been central to establishing the growing consensus that polyploidy has occurred and is important in many lineages. Surprisingly it is in the lineages where polyploidy was most expected that it has proved hardest to conclusively show evidence for genome duplication events. This includes in particular the vertebrates, where the idea of two rounds (2R) of WGD early in the vertebrate lineage is a long established, if still contentious, theory, known as the 2R hypotheses [1, 5]. Meanwhile lineages where polyploidy was unexpected (such as the hemiascomycete yeasts) have now been conclusively shown to be polyploid. The fact that many model organisms are now clearly demonstrated to be degenerate polyploids lets researchers see what the evolutionary products of genome duplication look like, which should in time help reveal what contribution genome duplication has made to the evolution of the vertebrate genome. The ever increasing quality and quantity of genomic data is allowing for the detection and study of polyploidy in ever more lineages, including vertebrates, where it has been shown that a WGD occurred in the lineage leading to the teleost fish after its divergence from mammals [9, 10].

Polyploidy has been suggested as being responsible for species radiations in the fish and vertebrate lineages. As polyploidy events are identified in more and more lineages, they may prove to have been responsible for other radiations too [11]. Werth & Windham [12] and Lynch & Force [13] have given a very clear and concise theoretical framework as to how polyploidy may lead to such radiations, and how it does so in a

passive manner that can be non-adaptive. The process is rooted in the massive random gene loss that invariably follows a polyploidy event, as most loci return to single copy. This results in the loss of different copies of the duplicate at some loci, a process called divergent resolution or reciprocal silencing. A relatively small number of these reciprocal silencings is sufficient to ensure reproductive isolation, creating the potential for speciation and radiation. Computational approaches have been developed to trace the fate of duplicated genes after a polyploidization and these comparative genomic methods are well placed to cast light on the fundamentals of post-polyploidy evolution. One aspect of post-polyploidy evolution that computational methods may be able to address is the basis of diploidization [4]. It is currently not well understood, but presumably involves changes in DNA sequences and deletions between chromosomes.

This review begins by presenting the specific features of paleopolyploid genomes in terms of structure and duplicated genes. These characteristics are at the heart of the strategies employed both to identify the traces of genome duplication and to study polyploid genomes themselves. We will review the different computational techniques that are currently available, their domain of competence and limitations and some potential routes to improve the sensitivity of detection. A number of recent reviews [14, 15] have addressed most current detection methods in some detail, so we will focus primarily on recent developments since then, while giving an overview of all approaches. The majority of the review will then discuss new developments in the analysis of paleopolyploidy, the process of diploidization and post polyploidy evolution in general. We will look at the impact of functional data on our understanding of these areas and will examine computational methods revealing evolutionary insights from the study of the gene order and content of modern polyploid genomes.

SOME SPECIFIC FEATURES OF PALEOPOLYPLOID GENOMES

Analyses of yeast, plant and vertebrate complete genome sequences have revealed common features of genome organization shared by paleopolyploid genomes in different eukaryotic kingdoms. We will highlight here the most important features

because they represent the hallmarks of paleopolyploidy and are at the heart of the strategies developed to identify ancient WGD events.

Firstly, the most obvious instantaneous result of polyploidy is the doubling of virtually each gene. The structure of modern paleopolyploid genomes indicates that having twin copies is an unstable state for most genes in the long term. Typically, only 10 to 30% of gene duplications arising from polyploidy are still retained in sequenced paleopolyploid genomes tens of millions of years after the WGD event [16, 17]. The remaining genes have in a majority of cases returned to a single copy state. Secondly, the surviving duplicates delineate pairs of chromosome regions where duplicated genes are organized in colinear order, as seen for example in *Arabidopsis thaliana* in Fig. (**1**). Within these regions, the duplicated genes are interspersed with single copy genes, which in most cases, were also duplicated during polyploidization with one of the copies latter lost in one of the regions. The colinearity should initially extend over entire duplicated chromosomes. Over evolutionary time, the ancestral duplicated genomes are scrambled. Inter and intra chromosomal rearrangements, including chromosome fusion, translocation and inversion break up duplicated chromosomes into smaller duplicated segments. So the third hallmark of paleopolyploidy, and the classical schematic representation of paleopolyploid genomes is a mosaic of megabase sized duplicated blocks covering the majority of the chromosomes. Blocks resulting from the same polyploidy event do not overlap with one another. This feature is an important signature that distinguishes the traces of paleopolyploidy from multiple independent duplications of individual chromosomal regions because one would expect regions that had been duplicated once to sometimes become duplicated again, producing three or more copies of the region. Finally, another specific feature of large-scale duplications such as polyploidy is that all genes were duplicated simultaneously. Thus, a signature of paleopolyploid genomes is that they contain an over representation of duplicated genes created at approximately the same time.

DETECTING ANCIENT POLYPLOIDY EVENTS

Detection methods fall under three main headings: tree-based, age-based and map-based. Tree-based methods look for the symmetric gene tree topologies expected after

polyploidization. Age-based methods estimate the age distribute of duplicates in the knowledge that polyploidy derived gene pairs are all formed at the same time. Map-based methods use the genomic locations of paralogs and orthologs in outgroup species to identify duplicated regions.

Within these headings the approaches group into two main classes. Intraspecific methods use genomic data from the species under scrutiny and most early developments fell under this heading. These approaches provide a number of ways to uncover a past polyploidization event, but are dependent on the presence of ohnologs (paralogs arising due to a polyploidization). The map-based approach involves the matching up of chromosomes, or parts of chromosomes, which can be linked by paralogs located in each sister region. Another strategy involves retracing the origin of duplicated genes using tree-based methods or age distributions of paralogs. Interspecific methods take advantage of genomic data from species related to the one under study. The recent explosion in the number of genome projects has opened up the potential of such approaches.

Tree-based approaches and the 2R hypothesis

Historically, the tree-based approach was first used to test the 2R hypothesis, which postulates two rounds of WGD early in the vertebrate lineage [5]. More generally, tree based methods can potentially be used to test any case where successive events of polyploidy are suspected. This approach is based on the expectation that there should be 2^n orthologs in a paleopolyploid genome for every gene in a genome that diverged before the *n* polyploidy events. In addition, the 2R hypothesis predicts that the 4 duplicated genes derived from polyploidy delineate a symmetric phylogenetic tree topology (i.e. (A,B)(C,D); A,B,C,D representing a four-member gene family in the post-polyploidy genome, see Fig. (**2**)). The alternative hypothesis, i.e. that of sequential gene duplication, will not always predict a symmetric topology. In the case of a four-member family, the sequential duplication model predicts giving rise to a symmetric (A,B)(C,D) topology with a proportion of 1/3 and an asymmetric (A(B(C,D))) topology with a proportion of 2/3 [18]. Thus, the test of the null hypothesis that the symmetric topology should be found with a probability of 1/3 in a sample of rooted four-member gene family trees can be used to test for two WGD events in succession.

The question as to whether two rounds of WGD occurred early in the vertebrate lineage is still under debate. Tree-based methods have in general not been in favor of the 2R hypothesis [19-23], whereas map-based approaches (discussed further below) have provided arguments supporting it [24-28]. The first extensive examination of the one-to-four (1:4) rule expected after two rounds (2R) of WGD, used the *Drosophila melanogaster* (pre-2R), *Caenorhabditis elegans* (pre-2R) and human (post-2R) genomes and showed no excess of four-member vertebrate gene families [23, 29, 30]. In addition, 76% of the 92 four-member gene families did not exhibit the symmetrical (A,B)(C,D) topology as predicted by the 2R hypothesis [23]. Given the massive duplicated gene loss that seems to invariably accompany diploidization, the one-to-four rule is probably too conservative, as a majority of genes may have returned to a 1:1, 1:2, or 1:3 ratio after two round of genome duplications. Furthermore, for large gene families, additional single gene duplication events may have occurred after polyploidy, which would result in a one-to-many ratio. Gibson and Spring [31] argued that if the second round occurred before the diploidization of the first round was complete, then this would result in some tetrasomic loci and some octosomic loci in the quadruplicated genome. Gene trees will then simply reflect the random order of diploidization of octosomic loci, rather than the order of chromosomal duplication, and tree topologies will, in general, be asymmetrical [15]. In line with this view, Furlong and Holland [32] made the proposal that two autotetrapolyploidy events occurred in quick succession in the vertebrate ancestor.

More recently Dehal and Boore [33] reconstructed the relationships of all gene families from the full gene sets of the basal chordate outgroup *Ciona intestinalis* (a tunicate) and three vertebrates *Takifugu rubripes* (a pufferfish), mouse and human. The authors determined when each gene duplicated by comparing gene family trees with the evolutionary tree of the organisms. They confirmed the results of previous studies that there exists no strong signal of WGD in the analyses of gene tree topology (like those discussed above) or gene family membership, with no peak at four genes per family in any of the vertebrates. However when the genomic map positions of only the subset of paralogous genes that were duplicated prior to the fish-tetrapod split was plotted, their global physical organization shows clear patterns of four-way paralogous regions (tetra-paralogons) covering a large part of the human genome (25% after 450 Mya). This pattern, with each genomic region corresponding in gene arrangement to sets of

paralogs in three other genomic regions, provides some of the most convincing evidence yet for two distinct genome duplication events early in vertebrate evolution. The fact that paralogous human genes generated by duplications after the split of fish and tetrapods appear to result largely from tandem duplications further reinforces the authors' case. The concept of paralogons [25] is discussed further in the map-based section below, but it is worth noting here that by using both tree-based and map-based data this study offers strong support for the 2R hypothesis, and demonstrates a powerful way of detecting ancient and obscured polyploidy events.

Age distribution of duplicates

This category of computational approaches relies on the fact that polyploidy derived gene pairs have been formed at the same time. Although a majority of the duplicated genes are lost after an ancestral polyploidization, a substantial number of duplicated genes remain in modern paleopolyploid genomes. For example in *Arabidopsis*, the youngest polyploidy event (20-40 Mya) left at least 5168 duplicated genes [34] out of the ~16000 paralogs found in the genome [7]. Thus, if it can be shown that a substantial number of duplicated genes have been created at about the same time, this can be considered as strong evidence that they have been created in a single event such as a polyploidization.

These approaches require the ages of duplication of each gene pair to be estimated. In practice the age of duplication can only be approximated by the age of divergence. These two dates can sometimes be different [35]. In allotetraploids the age of divergence of duplicated genes corresponds to the separation of the two parental diploid genomes, somewhat before the polyploidization event itself and all duplicated genes should have the same age of divergence. In contrast, if the duplicated chromosomes form multivalents during meiosis as in the case of autotetraploidy or segmental allotetraploidy, the ages of divergence of gene pairs will reflect the time of the shift from tetrasomic to disomic inheritance [3], which occurred after the polyploidy event. If this switch is not well coordinated among chromosomes, then the age of divergence of gene pairs formed by polyploidy may be scattered over a broad range. Other mechanisms of sequence homogenization between duplicates such as gene conversion may also delay sequence divergence [17, 35].

Dating of duplicated gene pairs generally relies on the molecular clock hypothesis that is the number of substitution between the compared sequences is proportional to the time of divergence [36]. Analysis of synonymous codon positions has been a method of choice because these sites are generally largely free from selection and so are thought to accumulate change at similar rates among genes [37]. The strategy generally employed is to estimate the level of synonymous substitution (Ks) between each pair of duplicated genes in a genome or for a subset of gene pairs residing in large duplicated chromosomal segments and to plot the number of gene pairs against Ks. The signal of a large-scale duplication event can be observed when a temporal peak of gene duplication is observed in the distribution (Fig. (**3**)). This approach has provided evidence for the polyploid origins of many model plants and teleost fish [3, 16, 38-41]. Schlueter et al. [38] and Maere et al. [16] have developed evolutionary models that can simulate the population dynamics of duplicated genes created by continuous small-scale and periodic large-scale duplication events based on their age distribution in a genome. Models that account for different numbers of large scale events can be fitted to the observed age distribution and likelihood comparison between models allow us to infer the number of large-scale events on a statistical basis. The advantage of age-distribution methods over map-based methods (see below) is that the starting material is not restricted to complete genome sequence as gene pairs can be constructed from EST data as well. In addition, when the rate of synonymous substitution per year is known, the modal Ks value that represent the temporal peak of gene duplication can be translated in absolute time. One of the biggest drawbacks of using Ks to measure divergence dates is that synonymous sites are rapidly saturated due to multiple substitutions, so that Ks becomes impossible to estimate with reliability. Thus, large-scale duplication can only be inferred when Ks is small (i.e. Ks<1-2 but sometimes more; see [16]), which limits the time frame in which paleopolyploidy can be efficiently detected by this method [39].

Dating of divergence can also be carried out using protein-based distance. This method is particularly useful when synonymous codon positions are saturated because protein sequences are known to diverge more slowly. However, the rate of protein divergence varies considerably among proteins so that in contrast to synonymous substitution, a global molecular clock cannot be applied to all pairs of duplicated proteins. The approach requires building protein families, which includes proteins

encoded by duplicated genes as well as at least two orthologous proteins and then reconstructing their phylogeny. One of the orthologs (O in Fig. (**3B**)) must be sufficiently outside the clade of interest so that it may serve as an outgroup to root the phylogeny while the other (B in Fig. (**3B**)) is used to calibrate the protein-specific molecular clock. A test of molecular clock (such as the two cluster test [42], the relative rate test [43] or a likelihood ratio test [44]) can be applied to the reconstructed phylogenies beforehand to exclude those protein families that diverge significantly from the protein-specific molecular clock model. Once a protein family has passed successfully the test of molecular clock, a linearized tree can be computed to re-estimate the branch lengths under the assumption of constant rate of evolution [42]. Then the relative ages of the duplicated protein pairs can be expressed as a proportion of the age of divergence between the species B and the species containing the duplicates (P). If the date of separation between the species B and P is known, the relative ages of duplicated proteins can be converted to absolute time. As for the Ks analysis, gene pairs can be then distributed according to their ages to look for temporal peaks of gene duplication. This method is particularly suitable when one wishes to seek evidence for very old polyploidy events. It has notably provided supportive evidence for paleopolyploidy in *Arabidopsis* [45] and vertebrates [9, 46-48].

Map-based approaches

Intraspecific methods

Map-based approaches are aimed at identifying the remnants of homology between the duplicated chromosomes. Over evolutionary time, the initial perfect colinearity between duplicated chromosomes is progressively degraded by point mutations, insertions, deletions and inversions. Ancient, degraded homology relationships like the duplicated regions in a polyploid genome are best identified by comparing gene content and most commonly gene order as well. Chromosomal segments are thought to be homologous if they share a significant number of homologous genes (identified as such using a tool such as BLAST), which are often organized in colinear order.

The first map-based computational implementation to detect a polyploidization was the 1997 study by Wolfe and Shields [6] in *Saccharomyces cerevisiae*. By assessing the locations of duplicated genes, the authors identified duplicated chromosomal regions in

which at least three pairs of duplicated genes were organized in the same order with intergenic distances of <50 kilobases. Approximately 50% of the genome could be paired into sister regions; the large sister regions did not overlap each other, and the overall orientation of duplicated regions with respect to centromeres and telomeres had remained largely the same. The authors concluded that this duplication pattern could only be caused by a WGD event. Following this pioneering study, several other eukaryotic genomes were analyzed using this approach with some methodological adaptation.

A map-based analysis on the *Arabidopsis* genome [8] concluded that 80% was duplicated and that many regions had undergone multiple duplications, suggesting a series of polyploidy events in its lineage. The detection of the older polyploidy event was largely obscured by the more recent polyploidy event, because duplicated blocks from the recent polyploidization overlap the old polyploidization's blocks. To facilitate identifying the older polyploidy event Blanc et al. [34] and Bowers et al. [49] reconstructed the approximate gene order of the ancestral genome that existed before the recent polyploidy event took place. This was done by walking along the entire genome and merging each duplicated block with its sister region, keeping the longest copy of each ohnolog pair and keeping genes in unduplicated regions of the genome unchanged in location. The polyploidy detection method was then carried out on the pseudo ancestral genome to identify old duplicated blocks from the earlier polyploidy events. Ignoring gene order, Friedman & Hughes [50] compared pairs of genomic windows in three eukaryotic genomes and counted the number of homologous gene pairs between them. They found all the genomes had significantly more windows sharing two or more homologous gene pairs, when compared to randomized genomes, suggesting en-block duplications. This same method was applied as well on the *Arabidopsis* genome to recover duplicated regions [51]. A slightly different approach taking some account of gene order was used in McLysaght [25] and co-workers' analysis of the human genome. Starting with a complete list of similarity hits for all genes in the genome it begins with two homologous genes from different chromosomes and looks for two other homologous genes within a set distance of the first two. These are added to the first to create a cluster and the process continues until it can add no more genes to the cluster. These clusters define paired sister regions in the genome called paralogons; 44% of the human genome was found to be covered by

paralogons with six or more pairs of homologous genes, strongly suggesting a polyploidy event in the early vertebrate lineage.

The different labs have developed their own implementations of intra-specific map-based approaches. Three of them have made their programs available to the scientific community, which can be installed and run on personal computers. Note that these programs can be used to identify synteny relationship between genomes as well. Hampson and colleagues [52] developed the program LineUp (http://titus.bio.uci.edu/lineup/) to identify homologous chromosomal regions in maize using gene/genetic maker order information. The method allows for rearrangements among duplicated genes (inversions) as well as gene deletion or insertion and evaluates the statistical significance of the identified homologous regions. Vandepoele and co-workers [53] developed the ADHoRe program (http://bioinformatics.psb.ugent.be/software.php), which is an interesting improvement of the map-based approach for identifying highly degenerate homologous segments. This program has proven very powerful identifying duplicated regions in the *Arabidopsis* [54] and rice [55] genomes as well as detecting homologous regions between the two [56]. It uncovers chromosomal segments that are homologous to other regions, but cannot be recognized as such because of extreme gene loss. First clearly colinear segments are aligned into a 'genomic profile' that combines information on gene order, strand localization and content from two (or more) segments. Inversions, deletions or insertions are tolerated. A homology matrix of the degenerated segment mapped against this profile can reveal homology that could not be identified directly by comparing individually with any of the segments forming the profile. The revealed homology can be tested for significance and if significant can itself be aligned into the genomic profile to help with revealing homology in further potentially homologous but degenerated segments.

Most map-based methods, including ADHoRe and LineUp, strongly emphasize gene order information. However the treatment of these data is very time consuming. In addition, these algorithms may fail to identify highly jumbled regions. Recently, Hampson et al. [57] questioned the utility of using gene order and strand information for detecting efficiently homologous regions under reasonable application conditions. They noted that if homologous regions were frequently rearranged through inversions or translocations, shared gene density might be more informative than gene order or

strand information. This prompted them to develop the program CloseUp (http://contact14.ics.uci.edu/closeup/) that detects significant chromosomal homology using shared-gene density alone. CloseUp was found to compare favorably in terms of runtime and efficiency against ADHoRe and LineUp using both artificial and real data [57].

Interspecific methods

So far the map-based approaches discussed have utilized intraspecific data - of necessity due to the lack of genome sequences from closely related organisms. Differential gene loss, where two sister regions lose a complementary set of genes, can obscure their common origin and make it challenging to identify them as duplicated segments using only intraspecific data. Sister regions in polyploids are interspersed with 'singletons' – genes that were duplicated but have subsequently returned to single-copy. These have little information value in intraspecific comparative mapping since only ohnologs are used as anchor points. However, singletons can be harnessed by using genomic data from an ancestral species that diverged before polyploidization. It was suggested [6, 58] that the clearest way to prove the existence of an ancient WGD would be to find another species (a pre-WGD species) that diverged from the purportedly polyploid lineage (leading to the modern post-WGD species) before the WGD event. Immediately after genome duplication, every ancestral chromosomal region corresponds to two duplicated blocks in the polyploid genome. In terms of gene order every pair of neighboring genes is also duplicated. Due to the nature of gene loss after polyploidization, a pair of previously adjacent genes may end up as singletons on different chromosomes, although still within the same duplicated block. Without nearby ohnologs as anchors, the pairing of the region would have been impossible to detect intraspecifically, but the gene adjacency relationship is preserved in the pre-WGD genome. Therefore, ancestral gene order information can provide the missing connection between sister regions.

Wong and co-workers [59] were one of the first to use gene content and gene order data from closely related species to improve the resolution of a polyploidy event (in this case in *S. cerevisiae*). Using initial sequence data from 13 other hemiascomycete yeasts a proximity plot was generated with a dot at the co-ordinate (*x*,*y*) if the *S. cerevisiae* genes *x* and *y* are neighboring genes in any of the other 13 genomes, overcoming the loss of gene order information due to differential gene loss. Including

dots for all ohnologs as well showed that over 80% of the genome is duplicated, up from 50% using only intraspecific genomic data, and strongly supporting the case that *S. cerevisiae* is a degenerate polyploid.

Using an interspecific map based approach, it has now been demonstrated conclusively that polyploidy events occurred in the lineages of the hemiascomycete yeasts [17, 60, 61] and teleost fish [40]. Kellis and co-workers suggested [17] that to convincingly demonstrate the existence of an ancient polyploidy event, these pre-WGD and post-WGD species should be related by a 1:2 mapping: where almost every region in the pre-WGD species corresponds to two sister regions in the post-WGD species; the two post-WGD sister regions should contain an ordered subset of the genes in the corresponding pre-WGD region, and nearly every region in the post-WGD species would correspond to one pre-WGD region and so be paired with a post-WGD sister region. In nearly simultaneous studies Kellis et al. [17] and Dietrich et al. [60], respectively using *K. waltii* and *A. gossypii* as the pre-WGD species, both showed conclusively that *S. cerevisiae* meets these criteria, and thus is a paleopolyploid. The sister regions in the post-WGD species were described as blocks of double conserved synteny (DCS). Fig. (**4**) illustrates how convincing this method is. Using a different pre-WGD species, *K. lactis* [61], the 1:2 mapping of the regions from *S. cerevisiae* chromosomes (colored by chromosome) in a DCS pattern is striking. 64% of all the genes in the *K. lactis* genome are in a DCS block and DCS blocks can be identified confidently even in the absence of any remaining ohnologs, with evidence instead coming from gene interleaving and 2:1 mapping with orthologous segments in the pre-WGD species. The inset in Fig. (**4**) shows a close up of a complete DCS block on *K. lactis* chromosome 3 as viewed through the Yeast Gene Order Browser (http://wolfe.gen.tcd.ie/ygob) [62]. The *S. cerevisiae* chromosome 5 region appears as the upper dark-blue track and the chromosome 4 region is seen as the lower light-blue track. The orthologous *K. lactis* region is shown in orange. The example is typical of DCS blocks, with almost all pre-WGD genes having matches in at least one of the two post-WGD sister regions, and genes from the two post-WGD sister regions interleaving onto the pre-WGD species while preserving order and orientation and a small number of remaining ohnologs. Recent work [62] reinforces the conclusion that a polyploidy event took place in the lineage of the hemiascomycetes by showing the level of double conserved synteny to be consistently high in all pair-wise comparisons between three pre-WGD and three post-WGD yeast genomes.

Jaillon and co-workers [40] applied this method to the genome of the teleost fish *Tetraodon nigroviridis* using the human genome as the pre-WGD species. Again the DCS pattern (associating two regions in *Tetraodon* with one in human) was immediately apparent across the entire genome and showed conclusively that a WGD event took place in the teleost fish lineage subsequent to its divergence from mammals. While intraspecific methods give the first signal of WGD, they depend on a minority of duplicated genes (the ohnologs), while the interspecific DCS signature considers all genes with orthologs in the pre-WGD species. This greatly improves the ability to resolve a WGD. In the case of *Tetraodon* it makes the difference between using 3% of the genome to try detect a WGD, and using 80% to prove a WGD took place.

POST-POLYPLOIDY EVOLUTION

Two crucial questions for biologists are how genetic complexity arises and what is the consequence of genetic redundancy. It is now well established that gene duplication, including through genome duplication in eukaryotes, is the main engine of the creation of genetic redundancy. Yet, the evolution of duplicated genes and how it connects with genetic complexity are less well understood. Many computational studies have addressed various aspects of the evolution of duplicated genes, often without regard to the origin of duplicates. However the timing of gene duplication is always a critical parameter when comparing evolutionary attributes between unrelated duplicated genes. The advantage of analyzing polyploidy-derived duplicated genes is that they all have been created at the same time, fixing this parameter for good. Another aspect that differentiates single gene duplication processes from polyploidy is that models of pathway evolution suggest that diversification of developmental and physiological functions depends on many genes acquiring novel protein functions and that this is most likely to occur if many genes are duplicated simultaneously [63-65]. Most recent bioinformatics analyses of the evolution of duplicated genes formed by polyploidy have focused on the patterns of gene loss and function/sequence divergence. Here, we will review computational analyses that address the evolution of duplicates on a genome-wide scale. These analyses have benefited hugely from the increasing amount of large-

scale functional and sequence data. Nowadays researchers have at their disposal various types of data that describe or characterize functional attributes of most genes in a genome. These include protein-protein interaction, proteomics (protein expression, post-translational modifications), and transcriptomics (gene transcription) data as well as various ontology systems and annotation databases that organize genes into functional categories. The improvement of sequencing technologies and the reduction of their costs make the sequencing of several related eukaryotic genomes more and more accessible. The availability of several genomes sharing the same polyploid ancestor allows for the analysis of the fate of the same duplicated genes in different lineages.

Pattern of duplicated gene loss

Interesting experimental work with neo-synthesized allotetraploids of *Arabidopsis* and wheat has shown gene elimination and epigenetic silencing take place almost immediately [66] and that in wheat the patterns of loss are to some degree reproducible [67, 68]. Gene loss is the fate of most duplicated genes and can occur rapidly [12, 13]. Walsh [69] predicted that almost all redundant duplicated copies of genes would become pseudogenes: one of the duplicates is required to maintain the function provided by the ancestral gene and the other is free to accumulate deleterious mutations. However, a substantial fraction of duplicated genes formed by polyploidy are actually maintained in the genome [16] raising the questions as to why and how duplicated genes escape deletion. Researchers have therefore investigated several aspects of the gene loss process and tried to identify which factors determine the loss or retention of duplicates.

Duplicated gene retention vs. function

The development of annotation databases and standardized vocabularies to annotate genomes has offered new opportunities to classify genes into broad functional categories and analyze the function of large set of genes automatically. Using the Yeast Proteome Database annotations [70], Seoighe and Wolfe [71] analyzed the function of the duplicated genes formed by polyploidy in *Saccharomyces cerevisiae*. They found that duplicated genes are not distributed evenly among functional categories, which indicates the fate of duplicated genes is influenced by the function of

the protein they encode. Cyclin genes, cytosolic ribosomal protein genes, heat shock protein genes, and genes involved in glucose metabolism and in the signal transduction apparatus were found to be preferentially retained in duplicate, while all the 44 mitochondrial ribosomal protein genes returned to single copy state. They also showed that selection for increased levels of gene expression was a significant factor determining which genes were retained in duplicate and which were returned to a single copy state.

An analysis of the function of the duplicated genes formed by polyploidy in *Arabidopsis* [72] using the Gene Ontology [73] and MIPS [74] annotations also reached the conclusion that duplicates were significantly over-represented in some functional categories (including transcription factors, ribosomal proteins, 26S proteasome and signal transduction) while they were significantly under-represented in others (including DNA repair proteins, defense related proteins and tRNA synthetases). Interestingly, transcription factors, which are the functional category most preferentially retained in duplicate in *Arabidopsis*, are also over-represented among duplicates after polyploidy in vertebrates [75] and fishes [76] , suggesting a universal route for post-polyploidy evolution in higher eukaryotes. In addition, Seoighe & Gehring [77] found that genes retained in duplicate following one round of genome duplication in *Arabidopsis* are significantly more likely to be retained again after a subsequent genome duplication. Maere et al. [16] made the striking observation that many functional categories that are highly retained in duplicate after polyploidy in *Arabidopsis* tend to be poorly retained in duplicate after small-scale duplication and *vice versa*. These results have shown that the massive gene loss that follows polyploidy is not the result of a mere random deactivation of duplicated genes but instead that the fate of duplicated genes is somewhat tied to their function.

What could cause some functional categories of genes to be preferentially retained or lost after duplication? It has been suggested that genomic redundancy of developmental genes may be selectively maintained to mask the consequences of null homozygotes or errors in transcription and translation [78, 79]. However, theoretical models suggest that one member of a redundant duplicate pair is always eventually lost by random genetic drift [80]. Gibson and Spring [81] argued that genes that encode multidomain proteins might have an increased chance of survival after duplication if point mutations in those genes tend to be dominant and have deleterious phenotypes.

Another perspective has been put forward by Veitia (summarized in [82]), who suggested that a significant cause for the retention of functional duplicates is the requirement for the preservation of stoichiometry within complexes or pathways. The survival (or loss) of dosage-sensitive duplicated genes may constrain the retention (or loss) of paralogs encoding other stoichiometric interactors [83].

Comparative genomics

Comparing two or more genomes can often provide valuable information about one or more of them. In particular, comparing the gene order (synteny) and content of closely related genomes can be very informative. Comparative genomics has given rise to computational methods based on gene order that are very different from techniques for comparing either nucleotide or amino acid sequences. Interspecific map based polyploidy detection methods revealed many interesting structures and features of polyploid genomes and suggested the comparative genomic analyses that followed. Central to the issues comparative genomics has been able to address are the patterns of the loss and retention of duplicated genes in post-WGD species, by comparison both to pre-WGD species and also other post-WGD species. The former comparison allows for the confident assignment of pre-WGD outgroup orthologs to duplicates and the study of the fate of those duplicates, while the latter comparison allows for divergently resolved (in particular, reciprocally silenced) loci to be identified which helps cast light on post-polyploidy speciation and species specific evolution. The hemiascomycete yeasts offer such a set of genomes, offering a unique opportunity to resolve post-polyploidy duplicate gene fate.

The utility of comparative genomics has only been fully realized with the availability of these fully annotated pre-WGD yeast genomes in the last two years, though Wong and co-workers [59] earlier used partial genome sequences and the expected disruption between adjacent genes in a paleopolyploid and a pre-WGD genome to show the polyploidy event in the hemiascomycetes occurred after the divergence of the *S. cerevisiae* and *K. lactis* lineages. Another earlier study [84] reporting the full genome sequences of three post-WGD species which are almost co-linear with *S. cerevisiae*, provided very significant annotation improvements to that genome and identified fast evolving and species specific genes. However the speciations were not close enough to the WGD to provide many cases of reciprocal gene loss. Cliften et al. [85] also sequenced, to 4x coverage, a number of very closely related post-WGD species, as

well as one distant post-WGD species, *S. castellii*, and one pre-WGD species, *S. kluyveri*. However in 2004 the fully annotated genome sequences for three pre-WGD yeast species [17, 60, 61] and a third post-WGD species [61] became available. Used initially to confirm a WGD took place in the *S. cerevisiae* lineage (as discussed earlier) they also opened to the door to using comparative genomics to study post-polyploidy evolution.

Kellis and co-workers [17] carried out evolutionary analyses on the polyploid yeast *S. cerevisiae* and the pre-WGD yeast *K. waltii*, noting that the relationship between them offered the first comparison across an ancient WGD event. They calculated that 88% of duplicates have been lost, via many small deletions (with an average size of two genes), typically balanced between the paired sister regions. With synteny now establishing the ancestry of duplicates (ohnologs) with certainty, Ohno's theory that after a WGD one gene copy is free to diverge while the other retains the ancestral function was put to the test. The 76 ohnologs with accelerated protein evolution relative to their pre-WGD ortholog were found to be biased towards protein kinases and regulatory proteins. Supporting the model that one paralog retained the ancestral function and the other was free to evolve rapidly, in 95% of these loci only one paralog experienced accelerated evolution, hence allowing inferences about newly evolved functions to be made.

A recent study takes a systematic approach to using synteny to study post-polyploidy evolution, utilizing a synteny scoring framework and curated homology sets across four pre-WGD and three post-WGD yeast genomes to trace gene fate in the polyploid species [62]. The resulting Yeast Gene Order Browser (YGOB; Fig. (**5A**)) by showing homology in its correct syntenic context, and taking into account in particular the polyploid nature of some of these genomes, makes genomic regions or loci of interest immediately apparent. It also allows curators to assess homology in the correct syntenic context, allowing for the confident identification of fast evolving loci and overcoming some known limitations of BLAST [86]. However the real power of this approach is its ability to systematically examine the patterns of duplicate gene loss (Fig. (**5B**)) among paleopolyploid yeasts. At each ancestral locus in each post-WGD species two, one or zero copies of the gene have been retained. These losses can proceed differently in different post-WGD species, a process called differential gene loss [87-90] (for an example see the legend discussion of the labelled columns in Fig.

(**5A**)). To systematically study differential gene loss, each pre-WGD was used as a scaffold on which the YGOB software scored the synteny of the gene presences and absences at each ancestral locus in each post-WGD species. By merging the results from all the pre-WGD "scaffolds" and considering only loci with unambiguously scored synteny, the nature of those loci was described in each pairwise comparison of the post-WGD genomes. The majority (74-80%) of traceable loci had single orthologous copies of the gene being retained in both species, in line with expectation [66, 91]. The remaining loci feature both syntenic copies of the gene and were therefore still present in two copies at speciation, with many fewer (8-11%) retained in duplicate now. This analysis offered the first genome-scale analysis of differential loss at an evolutionary depth, and proximity to the WGD, sufficient to capture a significant number of differential gene losses. Of particular interest was the identification of genes that were duplicated at the WGD, remained two-copy at speciation, but have since been differentially inactivated in different post-WGD species, each one losing a different, paralogous, copy of the gene. Between any two post-WGD species, 4-7% of the scorable loci are reciprocal gene losses of this type. These loci will likely be informative about post-polyploidy speciations [2, 13].

As more polyploid genome sequences and related pre-WGD genome sequences become available in more lineages, the utility of comparative genomics will continue to increase, both for the study of post-polyploidy evolution within those lineages and perhaps more interestingly in a further level of comparison between the various evolutionary clades featuring a polyploidy event. This should allow bioinformaticians to discern both the common and distinct elements of post-polyploidy evolution, casting light on the fundamentals of what happens to a genome after polyploidy, and perhaps even dissecting the details (auto- versus allo-polyploidy) of the WGD events themselves. As these fundamentals become clear, new computational strategies for addressing open questions like the 2R hypothesis in vertebrates may present themselves.

Lessons for homology

Given the importance of BLAST to homology assignment it is worth noting that one of the paleopolyploid studies mentioned in the previous section [62] shows that some common assumptions about homology assignment are not well founded. For example, it was discovered that 4-7% of the single-copy homologs between any pair of post-

WGD species are paralogs, confounding the widespread assumption that single-copy homologs shared by two genomes are always orthologous and revealing an important feature of polyploid genomes. Given that very many bioinformatics studies begin with the assembly of datasets within or across genomes based on homology, it is important to be as accurate as possible. This result shows the necessity of appraising BLAST results (especially between closely related polyploid species) in the light of syntenic context. As the quantity and quality of available genomic data increases in the lineages of many model organisms, many of which have now been shown to be polyploid, computational approaches that take advantage of gene order information will become increasingly important when assigning homology. The utility of such methods is particularly obvious in the case of single copy orthologs, but also for confidently distinguishing ohnologs (paralogs arising from a WGD) from members of gene families.

Chromosomal gene context in yeasts reveals many genes with very weak or indirect (via a mutual homolog) BLAST hits are in fact orthologs, paralogs or ohnologs [62]. Rather than being ignored these are perhaps some of the most interesting ohnolog, paralog and ortholog pairs, since they have most sequence (and often functional) divergence. The *S. cerevisiae* ohnolog pair *SPO21-YSW1* illustrates the point. With only 13% sequence identity and no direct BLASTP hit, their chromosomal gene contexts show unambiguously that they are ohnologs and their identical lengths (609 amino acids) and orientation reinforces the point. In short, the contextual, syntenic view is important to accurately examine genome structure and evolution.

Pattern of divergence between duplicated genes

A widespread view is that complete functional redundancy among duplicated genes cannot be evolutionary stable (but see [79]). The theoretical models described above provide explanations as to why some functionally redundant duplicated genes may have gene loss “delayed” or be selected for gene-dosage. However a fundamental assumption is that for both copies of duplicated genes to be stably fixed (i.e. maintained by selection) in the population, they must diverge in some way to carry out distinct functions. Two models of functional divergence are generally considered. In one model, neo-functionalization, one of the redundant copies evolves a new function [1] while the other retains the ancestral function. In the other model, sub-

functionalization, the two gene copies acquire complementary loss-of-function mutations in independent sub-functions, so that both genes are required to produce the full complement of functions of the ancestral gene [92]. The recent papers reviewed in this section address empirically how duplicated genes have evolved.

Exploration of functional data and functional divergence between duplicates

Functional divergence among duplicated genes is difficult to quantify. Different genes play different biological roles in many different ways. Some gene products are part of subcellular structures, other engage in protein-protein interactions, interact with DNA or RNA, or catalyze the transformation of small molecules. Genes with the same biochemical activities may be expressed at different times or in different places. Because the integration of the various aspects of gene functionality is complex, it is impossible to use a single simple measure to summarize them. Recent advances in post-genomic technologies have however allowed for the analysis of various aspects of gene function on a genome-wide scale.

In the context of post-polyploidy evolution, one of the most studied types of large-scale functional data is transcription intensity. Frequently, expression intensities are measured for several thousands of genes under different environmental conditions and tissues. Using expression data generated by microarray or MPSS (Massive Parallel Signature Sequencing) technologies, Blanc & Wolfe [72] and Haberer et al. [93] analyzed the divergence in expression pattern among pairs of duplicates formed by polyploidy in *Arabidopsis*. Both studies showed that a majority of duplicated genes experienced a significant divergence in their expression patterns. A similar conclusion was reached for 40 polyploidy-derived gene pairs examined in cotton [94]. The expression of duplicated genes has also been studied using large-scale transcription data in yeast [95-97], human [98] and plants [38, 99]. The general consensus emerging from these studies is that a large proportion of duplicated genes diverge in expression rapidly after duplication, and the vast majority of gene pairs eventually become divergent in expression. Interestingly, Blanc and Wolfe [72] found several cases where groups of duplicated gene pairs formed by polyploidy have diverged in concert, forming two parallel co-regulated networks, each containing one member of each gene pair. This observation has important implications for divergence in metabolic pathways and confirms previous assumptions [63-65].

Other types of functional data have been used to study the divergence of duplicated genes. For example, Wagner [95] analyzed the fitness effects of null mutation on 45 polyploidy-derived duplicated genes on yeast chromosome 5. His results indicate the spectrum of mutant phenotypes seen in duplicates is not significantly different from that seen in other genes. Nor do the phenotypes resulting from mutations in duplicates become more severe the more they diverge in gene sequence. He concludes that polyploidy has not contributed any lasting genetic redundancy to the yeast genome. Instead, he suggests that whole-genome duplication generated a transient wave of redundancy, which was quickly resolved by either deletion of sequences or their acquisition of new functions. However, other studies [100-102] provide several lines of evidence showing the significant role of duplicate genes in genetic robustness, where the loss of function of one copy is compensated by the other duplicated copy, resulting in no fitness effect. This kind of compensation may be mediated in the majority of cases by recent duplicates before they disappear through deletion or diverge in function. Using protein-protein interaction data from yeast, Wagner [96, 103, 104] found that duplicated gene products do not remain associated with the same interacting proteins, implying that the addition and elimination of interactions between proteins occurs shortly after duplication. This result also points to the rapid functional divergence of duplicated genes. Brun et al. [105] developed a computational method, PRODISTIN, that clusters proteins with respect to their common interactors identified from protein-protein interaction data. Using this method, they analyzed 41 pairs of duplicates formed by polyploidy in *S. cerevisiae* [106] and found that for both gene products in 26 pairs, the lists of interactors are very similar between the duplicates. For the remaining 15 gene pairs, the duplicates were interacting with different partners and therefore exhibited evidence of functional divergence.

Molecular evolution

Although duplication followed by functional diversification is widely believed to be the main source of molecular novelty during evolution, the details of the underlying molecular mechanisms are not well understood. Molecular evolution and phylogenetic approaches can be used to shed light on the process of divergence between duplicated genes. The aim of these approaches is to characterize the rate and the nature of changes in sequences, and the history of past evolutionary events as well as inferring functional shifts.

Functional changes can leave signatures in the sequences of a protein family, which may then be detected with a well-constructed history of their relationships and replacements. The challenge is to identify this record from the background noise of molecular evolution. Divergence of protein function is often revealed by a rate change in those amino acid residues of the protein that are most directly responsible for its new function [107-110]. One simple way to detect rate change is to construct a phylogenetic tree including the two protein duplicates and an outgroup sequence and to test for asymmetrical sequence divergence between the duplicates (i.e. one of the duplicates has evolved at a rate significantly higher than the other) using a relative rate test. Using this approach, it has been estimated that ~20% or more of duplicated genes formed by polyploidy evolved asymmetrically in *Arabidopsis* [72, 111], yeast [17] and fishes [76]. Similar proportions were observed for pairs of duplicated genes in *S. cerevisiae*, *D. melanogaster*, *C. elegans* [112], *S. cereviease* and mammals [110]. More elaborate computational methods for detecting functional shifts in protein or gene family alignments have been recently developed. For relatively recent events, these tests usually rely on comparisons of the nonsynonymous (replacement - Ka) to synonymous (silent - Ks) substitution rates for coding DNA [41, 111, 113]. The ratio of the two measures (ω=Ks/Ka) gives an indication of the strength of natural selection acting to constrain (purifying selection, ω<1) or accelerate (positive selection, ω>1) the fixation of non-synonymous mutations in the sequences. The analysis of the ω ratio for 242 duplicated genes formed by the most recent polyploidy event in *Arabidopsis* showed that they were all under purifying selection and that none exhibited evidence of positive selection [111]. However, this approach is limited by the relatively rapid saturation of synonymous substitutions by multiple hits. In addition, the ω ratio estimated between a pair of sequences can only be calculated as an average over all codons. Generally few codons are subject to positive selection, with the rest of the sequence evolving under purifying selection. Hence criteria such as an average ω greater than one are very conservative for detecting positive selection [113]. Other approaches to study older protein subfamilies rely on the amino acid replacement rates alone to identify sites that are most likely responsible for their divergent, as well as conserved, functions [97, 107, 109, 114-118]. Future studies, applying these methods to analyze the evolution of duplicated genes on a genome-wide scale, may yield interesting new results.

Reconstruction of ancestral genomes

Interesting work has been done to provide generic and abstracted solutions to the problem of reconstructing ancestral genomes as they appeared just before polyploidization [119]. Most approaches are however rooted in the analysis of specific genomes and the tackling of specific problems related to detecting and studying polyploidy, while the availability of genomic sequences for outgroup pre-WGD species now reduces the need for such intraspecific methods.

The earliest use of a reconstructed ancestral genome to examine post-polyploidy evolution was the 1998 study by Seoighe & Wolfe [58] which aimed to estimate properties of the yeast genome prior to the WGD and to reconstruct subsequent gene order evolution. The authors first reversed reciprocal translocations to bring the genome back to a symmetrical configuration, as it would be expected just after WGD. Simulations showed this approach could not regenerate the original block order when the number of translocations is large (as in the real genome), with the fraction of the genome being placed in duplicated blocks decreasing, smaller blocks not being detected, and symmetry being recovered in less reversals than the actual number of translocations. The relatively large minimum number of reversals needed to return the real yeast data to symmetry implied that many small duplicated blocks were then undetected in the *S. cerevisiae* genome, and this was subsequently confirmed [17, 59]. The large number of equally parsimonious paths returning the yeast genome to symmetry complicate the reconstruction of the ancestral gene order, but the authors show how a pre-WGD outgroup species would allow this degeneracy to be resolved. Expanding their model by varying both the number of fixed translocations and the number of retained duplicates (ohnologs) in simulations, the authors generated degenerate polyploid genomes similar to the real modern yeast genome, with about 8% retained duplicate and 70-100 translocations giving results similar to the real data. They note that in the simulated genomes, as in the real genome, the number of ohnologs recovered in duplicated blocks is less than the actual number present.

Of note recently, Jaillon and co-workers [40] used double conserved synteny (DCS) blocks (see Fig. (**4**) for examples), which were identified with respect to the human genome and used to prove a polyploidy event took place in the teleost fish lineage, and to reconstruct the ancestral osteichthyan (bony vertebrate) genome. The DCS blocks

define *Tetraodon* regions that arose from the duplication of a common ancestral region, and notably the blocks fall mainly into simple patterns interleaving either just two or three *Tetraodon* chromosomes. Using the distribution of *Tetraodon* orthologs in the human genome allowed for the partial reconstruction of the history of rearrangements in both lineages. Modeling the possible scenarios of genome duplication followed by recent and ancient fusions and breaks led the authors to conclude that ten large scale interchromosomal events was sufficient to explain the data, linking an ancestral genome of 12 chromosomes to the *Tetraodon* genome with 21 chromosomes. The authors showed that previously established genomic evidence (such as known rearrangements) fitted well with the mosaic of ancestral segments in the human and *Tetraodon* genomes, offering support for their reconstructed ancestral genome, with the higher frequency of rearrangements in the human genome underlying the more complex mosaic of ancestral segments in that lineage. The results also cast light on human genome evolution and show major differences in the evolutionary forces shaping the two genomes, for while only one human chromosome underwent no interchromosomal exchange, 11 *Tetraodon* chromosomes were intact.

It will be interesting to see if similar reconstruction efforts in other lineages reveal further species specific and lineage specific aspects of genome evolution. Both as a practical way to help identify nested polyploidies (as discussed in the detection section, see [34]) and also to garner new insights into genome structure and evolution, the reconstruction of ancestral genomes as they were prior to polyploidy continues to be an area where novel computational approaches will prove useful.

THE FUTURE

We are currently heading for a deluge of comparative genomic data, both interspecific data from closely related genomes, and also, excitingly, intraspecific data from multiple genomes of the same species (e.g. 100 *S. cerevisiae* genomes). While no major methodological improvements are expected in the area of polyploidy detection, further simplifications to reduce computation times are likely (e.g. the recently developed CloseUp program) and there will be a continuing need for new bioinformatics tools to manage and fully utilize the growing quantities of available genomic data.

As regards post-polyploidy evolution, there is still no concise picture of what precisely is occurring after polyploidy. This “mystery of diploidization” [4] provokes a number of open questions which bioinformaticians are well placed to address. While most duplicates are lost after polyploidy, the process of gene loss is far from random. Further investigation is needed into what determines whether a gene is preserved in duplicate or single copy. Many studies strongly suggest that there is selection for duplicates to be retained either because of redundancy/dosage or functional divergence, but what is the balance between these? Does the rate of chromosome rearrangement accelerate after polyploidy? If it does, is it an adaptive or neutral response, is it under genetic control, i.e. is there a genetic response to genome doubling? And while yeast has become the workhorse of the functional genomics community, and therefore most advances have so far been made on it, it will be interesting to see if other lineages, particularly those of multicellular organisms, show the same trends or not.

Thirty-five years after Susumo Ohno [1] first popularized the idea that gene duplication allows one gene copy to diversify in function (with the other kept 'safe' as a backup), the exact relationship between gene duplication and the evolution of new functions is still unresolved. His proposal that polyploidy, by duplicating all genes in a genome, is a powerful engine of evolutionary novelty, is still controversial, particularly, in the vertebrates. Today the design of computational methods, to confirm and explore the consequences of his theories, remains a fruitful area of research for bioinformaticians, and it will continue to be a fast moving and groundbreaking field in the years ahead.

ACKNOWLEDGEMENTS

We thank M. Woolfit and K.H. Wolfe for critical comments on the manuscript and G.C. Conant, J.S. Conery, J.L. Gordon and D.R. Scannell for discussion. KPB is supported by Science Foundation Ireland.

REFERENCES

1. Ohno, S., *Evolution by Gene Duplication*. 1970, London: George Allen and Unwin.
2. Taylor, J.S. and J. Raes, *Duplication and divergence: the evolution of new genes and old ideas.* Annu Rev Genet, 2004. **38**: p. 615-43.
3. Gaut, B.S. and J.F. Doebley, *DNA sequence evidence for the segmental allotetraploid origin of maize.* Proc Natl Acad Sci U S A, 1997. **94**(13): p. 6809-6814.
4. Wolfe, K.H., *Yesterday's polyploids and the mystery of diploidization.* Nat Rev Genet, 2001. **2**(5): p. 333-41.
5. Holland, P.W., J. Garcia-Fernandez, N.A. Williams, and A. Sidow, *Gene duplications and the origins of vertebrate development.* Dev Suppl, 1994: p. 125-33.
6. Wolfe, K.H. and D.C. Shields, *Molecular evidence for an ancient duplication of the entire yeast genome.* Nature, 1997. **387**(6634): p. 708-13.
7. AGI, *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature, 2000. **408**(6814): p. 796-815.
8. Vision, T.J., D.G. Brown, and S.D. Tanksley, *The origins of genomic duplications in Arabidopsis.* Science, 2000. **290**(5499): p. 2114-7.
9. Vandepoele, K., W. De Vos, J.S. Taylor*et al.*, *Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates.* Proc Natl Acad Sci U S A, 2004. **101**(6): p. 1638-43.
10. Christoffels, A., E.G. Koh, J.M. Chia*et al.*, *Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes.* Mol Biol Evol, 2004. **21**(6): p. 1146-51.
11. Amores, A., A. Force, Y.L. Yan*et al.*, *Zebrafish hox clusters and vertebrate genome evolution.* Science, 1998. **282**(5394): p. 1711-4.
12. Werth, C.R. and M.D. Windham, *A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression.* Am. Nat., 1991. **137**: p. 515-526.
13. Lynch, M. and A.G. Force, *The origin of interspecies genomic incompatibility via gene duplication.* Am. Nat., 2000. **156**: p. 590-605.
14. Simillion, C., K. Vandepoele, and Y. Van de Peer, *Recent developments in computational approaches for uncovering genomic homology.* Bioessays, 2004. **26**(11): p. 1225-35.
15. Van de Peer, Y., *Computational approaches to unveiling ancient genome duplications.* Nat Rev Genet, 2004. **5**(10): p. 752-63.
16. Maere, S., S. De Bodt, J. Raes*et al.*, *Modeling gene and genome duplications in eukaryotes.* Proc Natl Acad Sci U S A, 2005. **102**(15): p. 5454-9.

17. Kellis, M., B.W. Birren, and E.S. Lander, *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae.* Nature, 2004. **428**(6983): p. 617-24.
18. Hokamp, K., A. McLysaght, and K.H. Wolfe, *The 2R hypothesis and the human genome sequence.* J Struct Funct Genomics, 2003. **3**(1-4): p. 95-110.
19. Martin, A., *Is tetralogy true? Lack of support for the "one-to-four rule".* Mol Biol Evol, 2001. **18**(1): p. 89-93.
20. Hughes, A.L., *Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1.* Mol Biol Evol, 1998. **15**(7): p. 854-70.
21. 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, 1999. **48**(5): p. 565-76.
22. Hughes, A.L., J. da Silva, and R. Friedman, *Ancient genome duplications did not structure the human Hox-bearing chromosomes.* Genome Res, 2001. **11**(5): p. 771-80.
23. Friedman, R. and A.L. Hughes, *Pattern and timing of gene duplication in animal genomes.* Genome Res, 2001. **11**(11): p. 1842-7.
24. Abi-Rached, L., A. Gilles, T. Shiina, P. Pontarotti, and H. Inoko, *Evidence of en bloc duplication in vertebrate genomes.* Nat Genet, 2002. **31**(1): p. 100-5.
25. McLysaght, A., K. Hokamp, and K.H. Wolfe, *Extensive genomic duplication during early chordate evolution.* Nat Genet, 2002. **31**(2): p. 200-4.
26. Larhammar, D., L.G. Lundin, and F. Hallbook, *The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications.* Genome Res, 2002. **12**(12): p. 1910-20.
27. Lundin, L.G., D. Larhammar, and F. Hallbook, *Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates.* J Struct Funct Genomics, 2003. **3**(1-4): p. 53-63.
28. Vienne, A., J. Rasmussen, L. Abi-Rached, P. Pontarotti, and A. Gilles, *Systematic phylogenomic evidence of en bloc duplication of the ancestral 8p11.21-8p21.3-like region.* Mol Biol Evol, 2003. **20**(8): p. 1290-8.
29. Lander, E.S., L.M. Linton, B. Birren*et al.*, *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
30. Venter, J.C., M.D. Adams, E.W. Myers*et al.*, *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
31. Gibson, T.J. and J. Spring, *Evidence in favour of ancient octaploidy in the vertebrate genome.* Biochem Soc Trans, 2000. **28**(2): p. 259-64.
32. Furlong, R.F. and P.W. Holland, *Were vertebrates octoploid?* Philos Trans R Soc Lond B Biol Sci, 2002. **357**(1420): p. 531-44.
33. Dehal, P. and J.L. Boore, *Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate.* PLoS Biol, 2005. **3**(10): p. e314.

34. Blanc, G., K. Hokamp, and K.H. Wolfe, *A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome.* Genome Res, 2003. **13**(2): p. 137-44.
35. Langkjaer, R.B., P.F. Cliften, M. Johnston, and J. Piskur, *Yeast genome duplication was followed by asynchronous differentiation of duplicated genes.* Nature, 2003. **421**(6925): p. 848-52.
36. Graur, D. and W. Li, in *Fundamentals of Molecular Evolution*. 1999, Sinauer Associates: Sunderland MA. p. 139-153.
37. Li, W.H., *Molecular Evolution*. 1997, Sunderland, MA.: Sinauer Associates.
38. Schlueter, J.A., P. Dixon, C. Granger*et al.*, *Mining EST databases to resolve evolutionary events in major crop species.* Genome, 2004. **47**(5): p. 868-76.
39. Blanc, G. and K.H. Wolfe, *Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes.* Plant Cell, 2004. **16**(7): p. 1667-78.
40. Jaillon, O., J.M. Aury, F. Brunet*et al.*, *Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype.* Nature, 2004. **431**(7011): p. 946-57.
41. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.
42. Takezaki, N., A. Rzhetsky, and M. Nei, *Phylogenetic test of the molecular clock and linearized trees.* Mol Biol Evol, 1995. **12**(5): p. 823-33.
43. Tajima, F., *Simple Methods for Testing the Molecular Evolutionary Clock Hypothesis.* Genetics, 1993. **135**(2): p. 599-607.
44. Felsenstein, J., *Evolutionary trees from DNA sequences: a maximum likelihood approach.* J Mol Evol, 1981. **17**(6): p. 368-76.
45. Ermolaeva, M.D., M. Wu, J.A. Eisen, and S. Salzberg, *The age of the Arabidopsis thaliana genome duplication.* Plant Mol Biol, 2003. **51**(6).
46. Panopoulou, G., S. Hennig, D. Groth*et al.*, *New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes.* Genome Res, 2003. **13**(6A): p. 1056-66.
47. McLysaght, A., P.F. Baldi, and B.S. Gaut, *Extensive gene gain associated with adaptive evolution of poxviruses.* Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15655-60.
48. Gu, X., Y. Wang, and J. Gu, *Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution.* Nat Genet, 2002. **31**(2): p. 205-9.
49. Bowers, J.E., B.A. Chapman, J. Rong, and A.H. Paterson, *Unravelling angiosperm genome evolution by phylogenetic analysis of`chromosomal duplication events.* Nature, 2003. **422**: p. 433-438.
50. Friedman, R. and A.L. Hughes, *Gene duplication and the structure of eukaryotic genomes.* Genome Res, 2001. **11**(3): p. 373-81.

51. Ziolkowski, P.A., G. Blanc, and J. Sadowski, *Structural divergence of chromosomal segments that arose from successive duplication events in the Arabidopsis genome.* Nucleic Acids Res, 2003. **31**(4): p. 1339-50.
52. Hampson, S., A. McLysaght, B. Gaut, and P. Baldi, *LineUp: Statistical Detection of Chromosomal Homology With Application to Plant Comparative Genomics.* Genome Res., 2003. **13**(5): p. 999-1010.
53. Vandepoele, K., Y. Saeys, C. Simillion, J. Raes, and Y. Van De Peer, *The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice.* Genome Res, 2002. **12**(11): p. 1792-801.
54. Simillion, C., K. Vandepoele, M.C.E. Van Montagu, M. Zabeau, and Y. Van de Peer, *The hidden duplication past of Arabidopsisthaliana.* PNAS, 2002. **99**(21): p. 13627-13632.
55. Vandepoele, K., C. Simillion, and Y. Van de Peer, *Evidence that rice and other cereals are ancient aneuploids.* Plant Cell, 2003. **15**(9): p. 2192-202.
56. Vandepoele, K., C. Simillion, and Y. Van de Peer, *Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice.* Trends Genet, 2002. **18**(12): p. 606-8.
57. Hampson, S.E., B.S. Gaut, and P. Baldi, *Statistical detection of chromosomal homology using shared-gene density alone.* Bioinformatics, 2005. **21**(8): p. 1339-48.
58. Seoighe, C. and K.H. Wolfe, *Extent of genomic rearrangement after genome duplication in yeast.* Proc Natl Acad Sci U S A, 1998. **95**(8): p. 4447-52.
59. Wong, S., G. Butler, and K.H. Wolfe, *Gene order evolution and paleopolyploidy in hemiascomycete yeasts.* Proc Natl Acad Sci U S A, 2002. **99**(14): p. 9272-7.
60. Dietrich, F.S., S. Voegeli, S. Brachat*et al.*, *The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome.* Science, 2004. **304**(5668): p. 304-7.
61. Dujon, B., D. Sherman, G. Fischer*et al.*, *Genome evolution in yeasts.* Nature, 2004. **430**(6995): p. 35-44.
62. Byrne, K.P. and K.H. Wolfe, *The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species.* Genome Res, 2005. **15**: Epub ahead of print.
63. Wagner, A., *Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators.* Biol Cybern, 1996. **74**(6): p. 557-67.
64. Wagner, A., *Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization.* Proc Natl Acad Sci U S A, 1994. **91**(10): p. 4387-91.
65. Fryxell, K.J., *The coevolution of gene family trees.* Trends Genet, 1996. **12**(9): p. 364-9.
66. Kashkush, K., M. Feldman, and A.A. Levy, *Gene loss, silencing and activation in a newly synthesized wheat allotetraploid.* Genetics, 2002. **160**(4): p. 1651-9.

67. Ozkan, H., A.A. Levy, and M. Feldman, *Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops-Triticum) group.* Plant Cell, 2001. **13**(8): p. 1735-47.
68. Feldman, M. and A.A. Levy, *Allopolyploidy--a shaping force in the evolution of wheat genomes.* Cytogenet Genome Res, 2005. **109**(1-3): p. 250-8.
69. Konopka, J.B., C. DeMattei, and C. Davis, *AFR1 promotes polarized apical morphogenesis in Saccharomyces cerevisiae.* Mol Cell Biol, 1995. **15**(2): p. 723-30.
70. Hodges, P.E., A.H. McKee, B.P. Davis, W.E. Payne, and J.I. Garrels, *The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data.* Nucleic Acids Res, 1999. **27**(1): p. 69-73.
71. Seoighe, C. and K.H. Wolfe, *Updated map of duplicated regions in the yeast genome.* Gene, 1999. **238**(1): p. 253-61.
72. Blanc, G. and K.H. Wolfe, *Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution.* Plant Cell, 2004. **16**(7): p. 1679-91.
73. Ashburner, M., C.A. Ball, J.A. Blake*et al.*, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
74. Schoof, H., P. Zaccaria, H. Gundlache*t al.*, *MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome.* Nucleic Acids Res, 2002. **30**(1): p. 91-3.
75. Spring, J., *Vertebrate evolution by interspecific hybridisation--are we polyploid?* FEBS Lett, 1997. **400**(1): p. 2-8.
76. Van de Peer, Y., J.S. Taylor, I. Braasch, and A. Meyer, *The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes.* J Mol Evol, 2001. **53**(4-5): p. 436-46.
77. Seoighe, C. and C. Gehring, *Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome.* Trends Genet, 2004. **20**(10): p. 461-4.
78. Krakauer, D.C. and M.A. Nowak, *Evolutionary preservation of redundant duplicated genes.* Semin Cell Dev Biol, 1999. **10**(5): p. 555-9.
79. Nowak, M.A., M.C. Boerlijst, J. Cooke, and J.M. Smith, *Evolution of genetic redundancy.* Nature, 1997. **388**(6638): p. 167-71.
80. Lynch, M., M. O'Hely, B. Walsh, and A. Force, *The probability of preservation of a newly arisen gene duplicate.* Genetics, 2001. **159**(4): p. 1789-1804.
81. Gibson, T.J. and J. Spring, *Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomains proteins.* Trends Genet, 1998. **14**: p. 46-49.
82. Veitia, R.A., *Paralogs in polyploids: one for all and all for one?* Plant Cell, 2005. **17**(1): p. 4-11.
83. Papp, B., C. Pal, and L.D. Hurst, *Dosage sensitivity and the evolution of gene families in yeast.* Nature, 2003. **424**(6945): p. 194-7.

84. Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander, *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-54.
85. Cliften, P., P. Sudarsanam, A. Desikan*et al.*, *Finding functional features in Saccharomyces genomes by phylogenetic footprinting.* Science, 2003. **301**(5629): p. 71-6.
86. Wolfe, K., *Evolutionary genomics: yeasts accelerate beyond BLAST.* Curr Biol, 2004. **14**(10): p. R392-4.
87. Fischer, G., C. Neuveglise, P. Durrens, C. Gaillardin, and B. Dujon, *Evolution of gene order in the genomes of two related yeast species.* Genome Res, 2001. **11**(12): p. 2009-19.
88. Lundin, L.G., *Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse.* Genomics, 1993. **16**(1): p. 1-19.
89. Paterson, A.H., J.E. Bowers, and B.A. Chapman, *Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics.* Proc Natl Acad Sci U S A, 2004. **101**: p. 9903-8.
90. Seoighe, C. and K.H. Wolfe, *Yeast genome evolution in the post-genome era.* Curr Opin Microbiol, 1999. **2**(5): p. 548-54.
91. Nadeau, J.H. and D. Sankoff, *Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution.* Genetics, 1997. **147**(3): p. 1259-66.
92. Force, A., M. Lynch, F.B. Pickett*et al.*, *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.
93. Haberer, G., T. Hindemitt, B.C. Meyers, and K.F. Mayer, *Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis.* Plant Physiol, 2004. **136**(2): p. 3009-22.
94. Adams, K.L., R. Cronn, R. Percifield, and J.F. Wendel, *Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing.* Proc Natl Acad Sci U S A, 2003. **100**(8): p. 4649-54.
95. Wagner, A., *Robustness against mutations in genetic networks of yeast.* Nat Genet, 2000. **24**(4): p. 355-61.
96. Wagner, A., *Asymmetric functional divergence of duplicate genes in yeast.* Mol Biol Evol, 2002. **19**(10): p. 1760-8.
97. Gu, Z., D. Nicolae, H.H. Lu, and W.H. Li, *Rapid divergence in expression between duplicate genes inferred from microarray data.* Trends Genet, 2002. **18**(12): p. 609-13.
98. Makova, K.D. and W.H. Li, *Divergence in the spatial pattern of gene expression between human duplicate genes.* Genome Res, 2003. **13**(7): p. 1638-45.
99. Hughes, A.L. and R. Friedman, *Expression patterns of duplicate genes in the developing root in Arabidopsis thaliana.* J Mol Evol, 2005. **60**(2): p. 247-56.

100. Conant, G.C. and A. Wagner, *Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans.* Proc Biol Sci, 2004. **271**(1534): p. 89-96.
101. Kamath, R.S., A.G. Fraser, Y. Dong*et al.*, *Systematic functional analysis of the Caenorhabditis elegans genome using RNAi.* Nature, 2003. **421**(6920): p. 231-7.
102. Gu, Z., L.M. Steinmetz, X. Gu*et al.*, *Role of duplicate genes in genetic robustness against null mutations.* Nature, 2003. **421**(6918): p. 63-6.
103. Wagner, A., *How the global structure of protein interaction networks evolves.* Proc Biol Sci, 2003. **270**(1514): p. 457-66.
104. Wagner, A., *The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes.* Mol Biol Evol, 2001. **18**(7): p. 1283-92.
105. Brun, C., F. Chevenet, D. Martin*et al.*, *Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.* Genome Biol, 2003. **5**(1): p. R6.
106. Baudot, A., B. Jacq, and C. Brun, *A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein-protein interaction network.* Genome Biol, 2004. **5**(10): p. R76.
107. Gaucher, E.A., M.M. Miyamoto, and S.A. Benner, *Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors.* PNAS, 2001. **98**(2): p. 548-552.
108. Wang, Y. and X. Gu, *Functional Divergence in the Caspase Gene Family and Altered Functional Constraints: Statistical Analysis and Prediction.* Genetics, 2001. **158**(3): p. 1311-1320.
109. Gu, X., *Statistical methods for testing functional divergence after gene duplication.* Mol Biol Evol, 1999. **16**(12): p. 1664-1674.
110. Dermitzakis, E.T. and A.G. Clark, *Differential Selection After Duplication in Mammalian Developmental Genes.* Mol Biol Evol, 2001. **18**(4): p. 557-562.
111. Zhang, L., T.J. Vision, and B.S. Gaut, *Patterns of nucleotide substitution among simultaneously duplicated gene pairs in Arabidopsis thaliana.* Mol Biol Evol, 2002. **19**(9): p. 1464-73.
112. Conant, G.C. and A. Wagner, *Asymmetric sequence divergence of duplicate genes.* Genome Res, 2003. **13**(9): p. 2052-8.
113. Yang, Z. and J.P. Bielawski, *Statistical methods for detecting molecular adaptation.* Trends in Ecology and Evolution, 2000. **15**(12): p. 496-503.
114. Knudsen, B., M.M. Miyamoto, P.J. Laipis, and D.N. Silverman, *Using Evolutionary Rates to Investigate Protein Functional Divergence and Conservation: A Case Study of the Carbonic Anhydrases.* Genetics, 2003. **164**(4): p. 1261-1269.
115. Knudsen, B. and M.M. Miyamoto, *A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins.* PNAS, 2001. **98**(25): p. 14512-14517.

116. Pupko, T. and N. Galtier, *A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes.* Proc Biol Sci, 2002. **269**(1498): p. 1313-6.
117. Gu, X., *Maximum-Likelihood Approach for Gene Family Evolution Under Functional Divergence.* Mol Biol Evol, 2001. **18**(4): p. 453-464.
118. Nam, J., K. Kaufmann, G. Theißen, and M. Nei, *A simple method for predicting the functional differentiation of duplicate genes and its application to MIKC-type MADS-box genes.* Nucl. Acids Res., 2005. **33**(2): p. e12-.
119. El-Mabrouk, N., *Recovery of ancestral tetraploids.* Comparative Genomics, 2000: p. 455-477.

FIGURE LEGENDS

Figure 1. Typical representation of a palopolyploid genome: the case of *Arabidopsis thaliana*. (*A*) Pairs of sister duplicated regions resulting from the most recent polyploidy events are represented with color boxes on the five chromosomes of *A. thaliana*. The fact that the majority of the genome is covered by duplicated blocks and that adjacent duplicated regions do not overlap between each other is a strong indication that the ancestor of *A. thaliana* experienced a WGD event followed by chromosomal rearrangements. (*B*) A close up on a pair of duplicated regions on chromosomes 4 and 5. Black and white boxes represent duplicated and single copy genes respectively. Genes shared by the two regions are joined by lines and are organized in the same order.

Figure 2. Alternative topologies of four membered families resulting from sequential gene duplication or genome duplication. Figure modified from Hokamp et al. (2003). (*A*) Topologies resulting from duplication of one member of a three-membered gene family. The three different duplication scenarios result in three different trees.The tree from the duplication of gene C and that from the duplication of gene D have asymmetric topologies. (*B*) Topologies resulting from two genome duplications in succession. All genes are duplicated at each step, resulting in a symmetric tree topology.

Figure 3. Inference of ancient genome duplication using the duplicate age distribution approach. (*A*) Ks distribution of pairs of duplicated genes in banana (Figure kindly provided by Magali Lescot). In this example, the relative age of divergence of duplicate pairs was estimated using the level of synonymous substitutions (Ks). A conspicious peak centered around Ks=0.5 (indicated by a double star symbol) indicates that a high number of gene duplication occurred within a short period of time in the *Musa* ancestor. This burst of gene duplication is likely to be the result of an ancient WGD event. An initial high density of duplicates (indicated by a single star) is contained within the youngest age classes (0<Ks<0.1). This peak is the product of ongoing single gene duplication processes (see Blanc and Wolfe 2004 for further details). (*B*) Age distribution of human duplicate pairs infered from protein distances (Modified from McLysaght *et al.* (2002)). To estimate the age of divergence of two duplicated protein relative to the age of a reference speciation event (fly-human in this example) a

phylogenetic tree including the two duplicates (P1 and P2), the fly ortholog, B, and an outgroup sequence, O, is constructed assuming a constant rate of evolution. O is chosen to be the most evolutionary distant sequence, to allow rooting of the tree. The relative age of P1 and P2 is then calculated as the ratio of the two distances X/D and is expressed as a fraction of the fly-human divergence age, D. The distribution of relative ages of human duplicates exhibits an excess of gene duplication in the age class 0.4-0.7 D (indicated by a double star symbol). This indicates that a burst of gene duplication activity took place in the period 350-650 Mya, which is compatible with at least one round of polyploidy.

Figure 4. (*A*) Gene correspondence with *K. lactis* chromosomes reveals *S. cerevisiae* sisters regions in double conserved synteny (DCS) blocks. Each region of *K. lactis* has conserved gene order with two regions in *S. cerevisiae* (colored by chromosome number) clearly displaying the 1:2 mapping expected in a polyploid genome when compared with a species that diverged before the WGD. (*B*) Close up on a pair of duplicated regions in *S. cerevisiae* matched to *K. lactis* chromosome 3, as visualized using the Yeast Gene Order Browser (wolfe.gen.tcd.ie/ygob). Note the massive loss of duplicated genes, the small number of remaining ohnologs that act as anchor points and the interleaving genes. This is described as a double conserved synteny (DCS) block.

Figure 5. (*A*) Yeast Gene Order Browser (YGOB; wolfe.gen.tcd.ie/ygob) screenshot focused on the *A. gossypii* gene *ABR086W* with a window size of 6. Each box represents a gene and each color a chromosome. The middle four tracks represent pre-WGD genomes and the top and bottom three tracks represent the A and B tracks of post-WGD genomes. Each genome uses a different color palette (blues for *S. cerevisiae* etc.). The “b” buttons (in the top right corner of each gene box) open a window with BLASTP results against YGOB’s database, “S” buttons (at the bottom of each column) click through to a pillar’s protein sequences, and “T” buttons (at the top of columns) draw approximate phylogenetic trees on the fly. The column marked as “a” is an example of a locus where the gene remains in duplicate in all post-WGD species, column “b” is an example of the most common locus type featuring single copy orthologs in all post-WGD species, while column “c” is an example of a locus that has undergone reciprocal gene loss with syntenic context revealing single copy paralogs in the post-WGD species. (*B*) Schematic diagram of the YGOB approach to tracing and

representing polyploidy. The loci pictured represent the evolution of the 9 central loci in (A), in the three species: *A. gossypii* (green), *S. cerevisiae* (blue) and *S. castellii* (red). The diagram shows how a pre-WGD genome can be used as a scaffold to identify post-WGD intra-genomic sister regions within a genome, how post-WGD inter-genomic regions can be confidently aligned and how reciprocal gene loss can be identified at a locus.

Figure 1

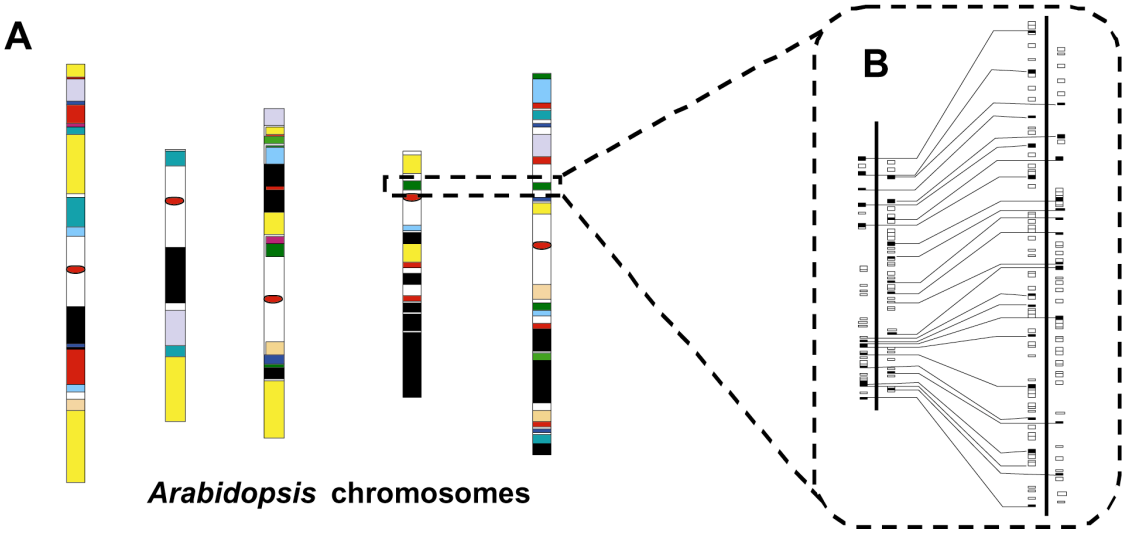


Figure 2

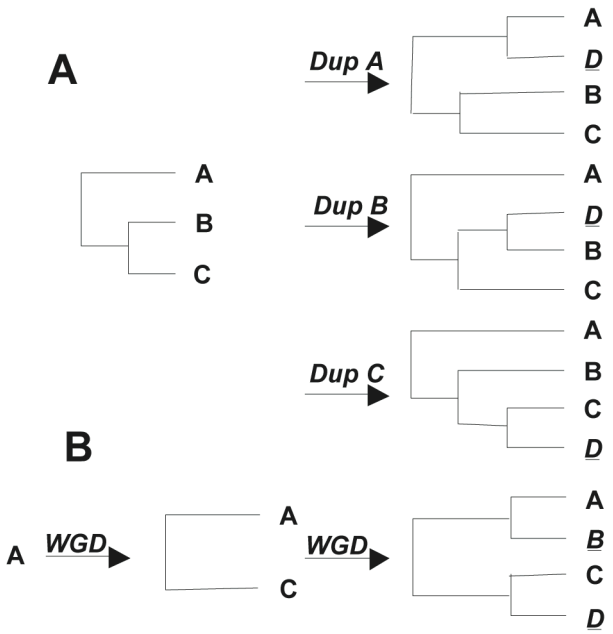


Figure 3

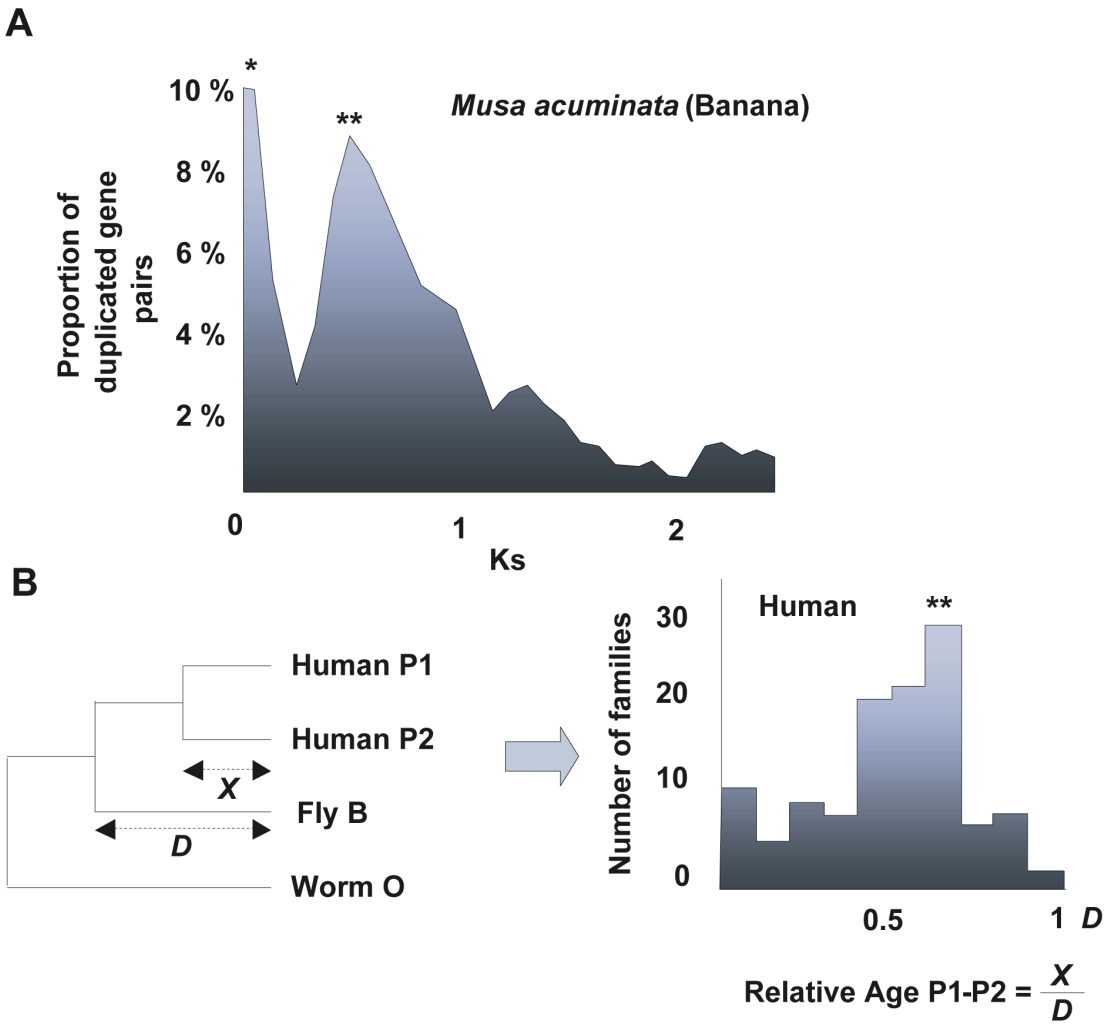

A
*
**
10 %
8 %
6 %
4 %
2 %
Proportion of duplicated gene pairs
Musa acuminata (Banana)
0
1
2
Ks
B
Human P1
Human P2
X
Fly B
D
Worm O
Human
**
30
20
10
0
Number of families
0.5
1 D
Relative Age P1-P2 = X/D

Figure 4

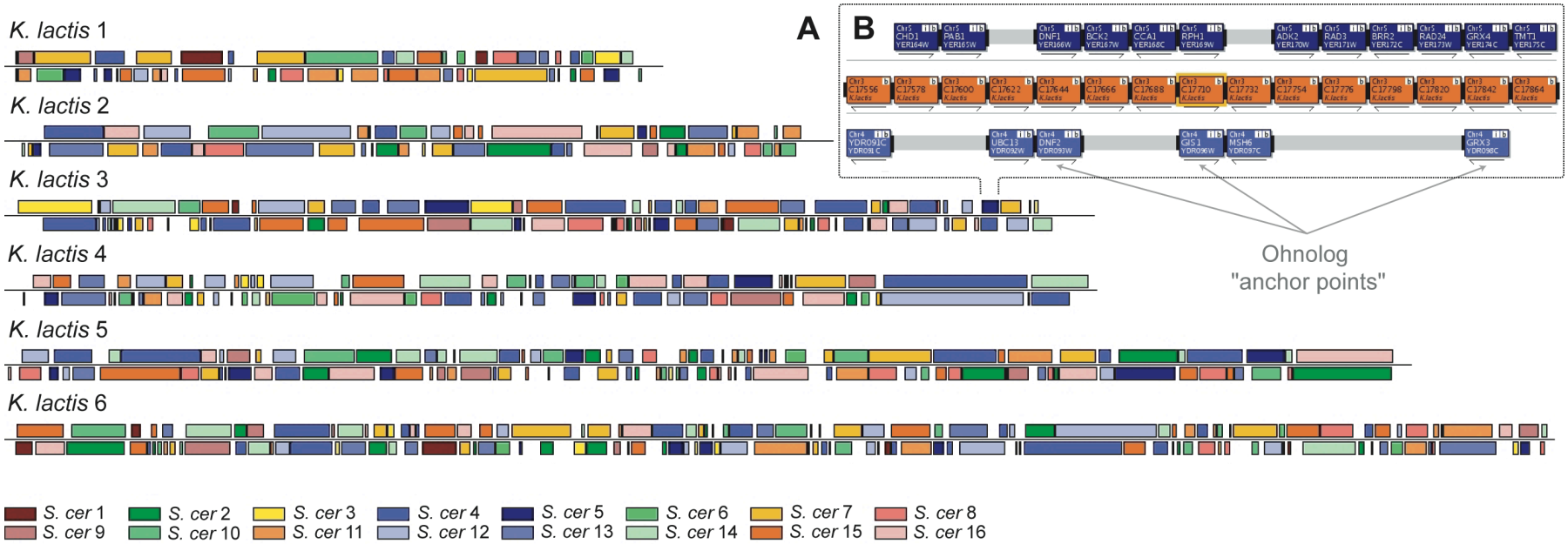
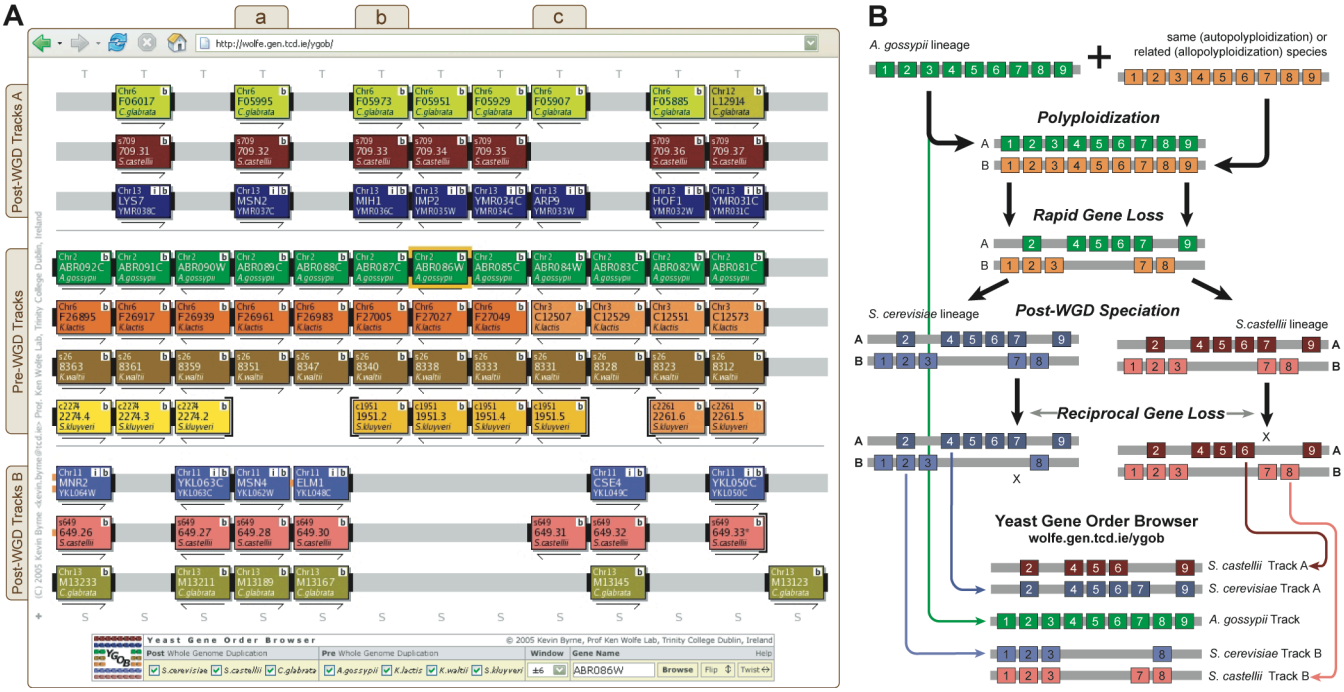

K. lactis 1
K. lactis 2
K. lactis 3
K. lactis 4
K. lactis 5
K. lactis 6
A
B
Ohnolog
"anchor points"
S. cer 1
S. cer 2
S. cer 3
S. cer 4
S. cer 5
S. cer 6
S. cer 7
S. cer 8
S. cer 9
S. cer 10
S. cer 11
S. cer 12
S. cer 13
S. cer 14
S. cer 15
S. cer 16

Figure 5


A
a
b
c
Post-WGD Tracks A
Pre-WGD Tracks
Post-WGD Tracks B
Yeast Gene Order Browser
B
A. gossypii lineage
same (autopolyploidization) or related (allopolyploidization) species
Polyploidization
Rapid Gene Loss
S. cerevisiae lineage
Post-WGD Speciation
S.castellii lineage
Reciprocal Gene Loss
Yeast Gene Order Browser
wolfe.gen.tcd.ie/ygob
S. castellii Track A
S. cerevisiae Track A
A. gossypii Track
S. cerevisiae Track B
S. castellii Track B