The Quest for Genomic Homology

Klaas Vandepoele, Cedric Simillion and Yves Van de Peer*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Abstract: New initiatives to sequence complete genomes of related organisms have introduced a new era of large-scale evolutionary genomics. The comparative analysis of these genomes allows us to obtain a comprehensive view of many aspects of eukaryotic genome evolution. Consequently, new computational methods and approaches are being developed in order to investigate chromosomal organisation, rearrangements and segmental homology. Here, we review the different techniques currently available to identify homologous chromosomal segments in closely and more distantly related species and highlight some of the difficulties inherent to the statistical validation of putative genomic homology. In addition, advantages of cross-species genome analysis are discussed as well as novel approaches to study large-scale gene duplications.

Received on: 20/12/03 - Accepted on: 27/02/04

Key Words: genome evolution, genomic homology, colinearity, gene duplication.

INTRODUCTION

Comparative genomics provides an efficient way to detect functional elements in genomic sequences. The observation that functional regions are conserved throughout evolution, in contrast to their non-functional counterparts, has triggered the sequencing of (at least parts of) genomes of closely related animals, plants and fungi [1-6]. Such large-scale sequencing projects offer an integrated framework for comparative sequence analysis and greatly enlarge our knowledge about gene structure, function and regulation. Perhaps the most illustrative example is the sequencing of the mouse genome, which, in comparison with the human genome, has allowed the identification of many regulatory elements and has improved gene annotation in both human and mouse [7-14]. Moreover, the detection of signals that are conserved, but cannot be recognised in the absence of a cross-species comparison makes it possible to discover new functional elements, such as non-coding RNAs [15, 16], and hint to their importance in biological systems.

Apart from the improved detection of conserved elements and a better understanding of the complexity embedded in biological processes through the comparative analysis of the genes involved, the availability of an increasing amount of genome sequences from a large variety of organisms makes it possible to study the organisation of genes in a genome. Especially the characterisation of different types of rearrangements (e.g. inversions, translocations and transpositions), duplications and gene loss exposes the actual impact of genome evolution on the complete catalogue of genes encoded by the genome [17-22]. However, in order to study genome organisation and genome evolution, it is essential that conserved regions between and within genomes can be correctly identified. Since these homologous regions, derived from a common ancestral region, may have been extensively rearranged, their identification is not always obvious. In this review, we discuss the different techniques currently applied for the detection of homologous chromosomal regions and their application to the analysis of large-scale duplication events. Furthermore, we highlight some of the advantages of having access to related genomes when unravelling a genome's evolutionary past.

THE DETECTION OF HOMOLOGOUS CHROMOSOMAL SEGMENTS

Choosing the best method for the detection of homology at the genomic level highly depends on the resolution one wants to obtain and on the nature of the genomic information that is available. If complete genomic sequences of closely related species are available, the most straightforward way to detect homology is by comparing the sequences at the DNA level using a standard sequence similarity search tool such as BLAST [23] or FASTA [24]. Similarly, a DNA-based sequence comparison can be applied to identify recently duplicated and thus paralogous chromosomal regions within the same genome. For the comparison of very long stretches of DNA, both pairwise alignment tools (e.g., Smith-Waterman [25], DOTTER [26], MUMmer [27], PipMaker [28], SSAHA [29], BLAT [30], BLASTZ [31], AVID [32], LAGAN [33]) and multiple alignment tools (Multi-LAGAN [33], MultiPipmaker [34]) have been developed. If both input sequences are closely related, large-scale alignments can be generated, which show a detailed base-to-base mapping between the two genomic sequences. Although some of the programs listed above are able to cope with genomic sequences from more distantly related organisms, the increasing amount of sequence dissimilarity between such genomes, or alternatively between anciently duplicated regions within the same genome, seriously complicates the detection of significant homology over long genomic distances (e.g. 100-1,000kb). Rather, small conserved fragments, typically conserved exons or non-coding conserved

*Address correspondence to this author at the Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium; Tel: +32 9 331 3807; Fax: +32 9 33 13 809; E-mail: yvdp@gengenp.rug.ac.be

1389-2029/04 $45.00+.00

sequences might be recovered, but these provide little overall information on the evolution of chromosomes or complete genomes.

When the amount of sequence similarity at the DNA level is too low to determine homology between or within genomes, the inference of conserved gene content and order (i.e. colinearity) provides an elegant alternative to unravel common ancestry of chromosomal regions. The advantage of this method, compared to DNA sequence alignment methods, is that similarity that has faded away at the DNA level still can be detected at the protein level. This is demonstrated in (Fig. **1**) showing a comparison of two highly similar and two degenerated paralogous chromosomal regions in the genome of *Arabidopsis thaliana*, both at the DNA level and at the protein level. Where for recently duplicated (and thus highly similar) regions homology can still clearly be inferred by both methods (i.e. DNA-based alignments and colinearity at the protein level), the homology between the degenerated paralogous regions is only visible through the detection of colinearity at the protein level.

THE MAP-BASED APPROACH: DETECTION OF CONSERVED CONTENT AND ORDER

The identification of homologous chromosomal regions between distantly related organisms is thus usually based on a genome-wide comparison that aims at delineating regions of conserved gene content and order in different parts of the genome. The same is true for the detection of duplicated chromosomal regions within the same genome. Although the map-based approach can be applied on the basis of different types of genomic information (e.g. genes, molecular markers or local DNA similarities, see further), we will explain the general concept of this method with genes as the genomic units of a chromosome. Essential in the map-based approach is that the (absolute or relative) chromosomal locations of all genes (or in general the units describing the chromosome under investigation) are known.

Although the detection of colinearity seems a fairly simple way to detect genomic homology, the dynamic nature of genomes, responsible for the duplication, deletion, and rearrangements of genomic DNA, results in a degraded pattern of colinearity that makes it difficult to detect more ancient homology. Nevertheless, the correct identification of homologous segments remains an important issue. Regarding large-scale gene duplication, several studies already applied a map-based approach for the detection of duplicated segments in fully sequenced genomes [35-38]. Recently, we developed a publicly available software tool, called ADHoRe, for the automatic detection of homologous regions combined with a robust statistical validation [39, 40]. The general concept of ADHoRe makes it possible to use the software tool for the analysis within one genome, i.e. to look for paralogous regions with duplicated genes, or for comparisons between genomes of different organisms, i.e. to look for orthologous regions. Moreover, events such as inversions, deletions and tandem duplications that complicate the detection of homology, can be taken into account. Based on similar principles, Gaut and coworkers recently published the LineUp package that aims at detecting significant chromosomal homology based on molecular marker information, even if substantial rearrangements of marker order have occurred [41].

THE ADHoRe ALGORITHM

In the map-based approach as implemented in ADHoRe, the information on homologous gene pairs is stored in a matrix of (m.n) elements (m and n being the total number of genes on each genomic fragment), each non-zero element (x, y) being a pair of homologous genes (x and y denote the coordinates of these homologous genes or anchor points). Figure **2a** shows a small hypothetical gene homology matrix (GHM). In the matrix, colinear segments are represented as diagonal lines, while tandem duplications form horizontal or vertical lines, inversions can be detected by considering the orientation of the elements, and gaps in diagonal regions refer to gene loss or gene insertions in duplicated blocks. To detect colinearity, one has to find more or less diagonal series of elements (i.e. homologous genes) in the matrix. This way of presenting the organization of genes on genomic segments reduces finding colinearity to a clustering problem. During construction of the GHM, ADHoRe subjects it to a number of procedures. For example, after identification of the homologous genes, irrelevant data points need to be removed, a process we refer to as negative filtering [39]. During this step, all elements that cannot belong to a cluster because they are too far away from other elements in the homology matrix – i.e. homologous genes that most probably have not been created by the block duplication - are removed. Also tandem duplications are removed from the matrix. Since we are looking for diagonal regions in the GHM, purely horizontal or vertical regions due to tandem duplications are remapped by collapsing all tandem duplications. This way it is easier to detect diagonal regions, since they are no longer interrupted by horizontal or vertical elements. The end result is a matrix that has been cleaned up by filtering and a colinear region is now defined in the matrix representation as a number of elements (which we refer to as anchor points) showing clear diagonal proximity (Fig. **2b**). In order to find such diagonals on a mathematical basis, we have developed a special distance function that yields a shorter distance for elements that are in diagonally close proximity than points that are in horizontal or vertical proximity [39]. Figure **3** shows the application of this distance function to a hypothetical example. Briefly, all elements in the GHM that are in close proximity are grouped into clusters. Subsequently, the quality of each cluster is examined and can be used to remove non-colinear homologous regions (see Fig. **3**). Finally, it is investigated whether detected clusters can be combined into larger homologous regions [39].

STATISTICAL SIGNIFICANCE OF COLINEARITY

When all clusters (i.e., colinear regions) have been compiled as described above, colinear segments (or clusters in the homology matrix) that are not statistically significant need to be removed. The goal of this procedure is to determine which colinear regions could occur purely by chance and are therefore not biologically significant. This problem was first recognised by Gaut (2001) who introduced

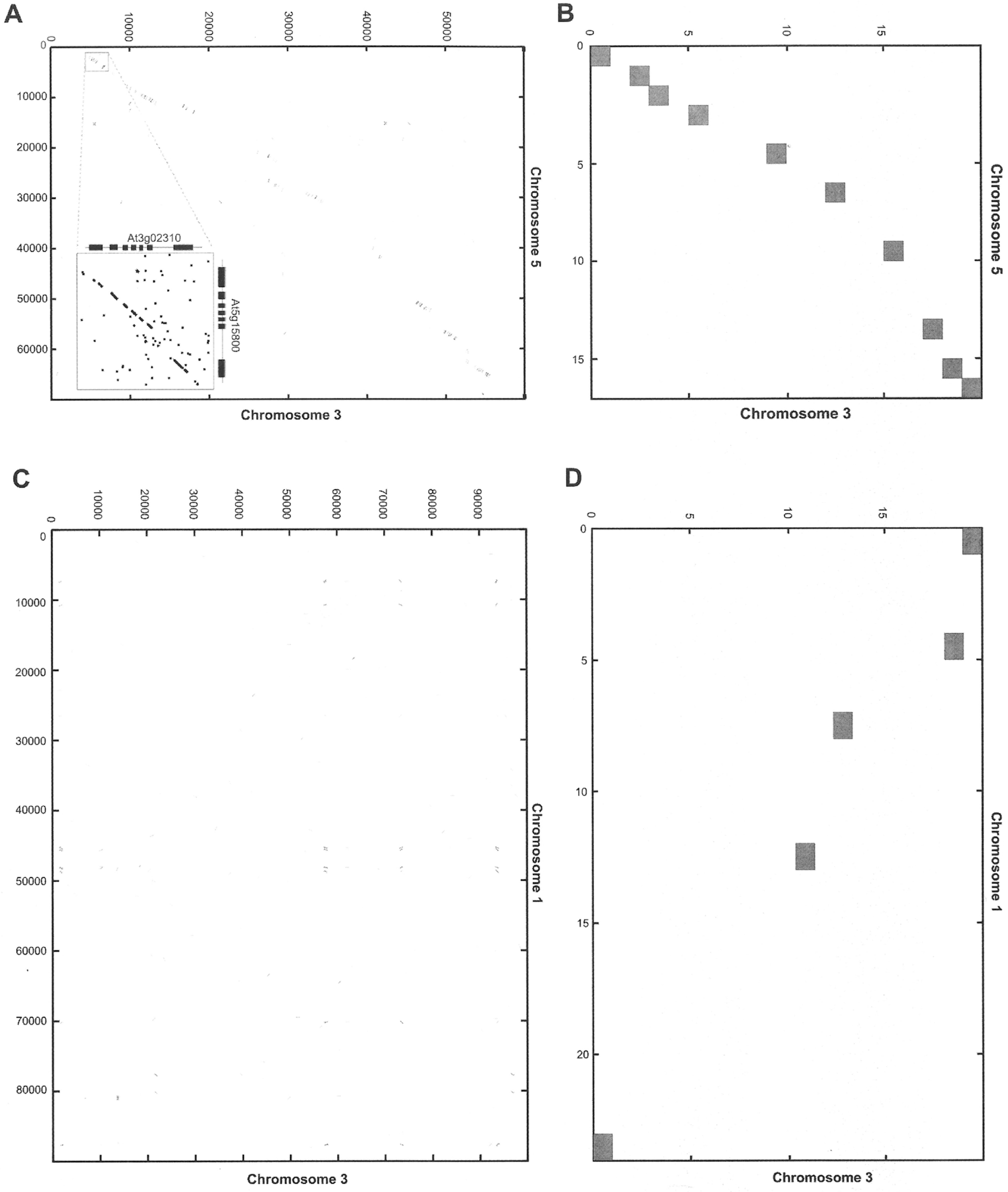

Fig. (1). Comparison of duplicated regions in *Arabidopsis* through both DNA-based alignments and the detection of colinearity (conserved gene content and order). Figs 1A and 1B show a recently duplicated chromosomal segment between chromosome 3 (size 55,6 kb or 21 annotated genes) and chromosome 5 (size 65.5 kb or 20 annotated genes) that can be detected by DNA-based alignments and by colinearity at the protein level, respectively. DNA-based alignments were created using MUMmer (parameters: -l 15 –b –c; Delcher et al., 1999). The zoom-in, created with DOTTER (Sonnhammer and Durbin, 1995), shows the conserved exon-intron structure at the DNA level of a paralogous gene pair. Figs 1C and 1D show an ancient duplication event between chromosome 1 (size 89,7 kb or 26 annotated genes) and chromosome 3 (size 100,4 kb or 24 annotated genes). Whereas colinearity at the protein level enables the detection of this anciently duplicated segment (D), no similarity at the DNA level can be found (C). Note that in panels A and C the axes of the graph represent the base pairs of the chromosome, where in panels B and D the graph represent genes positioned along the chromosome.

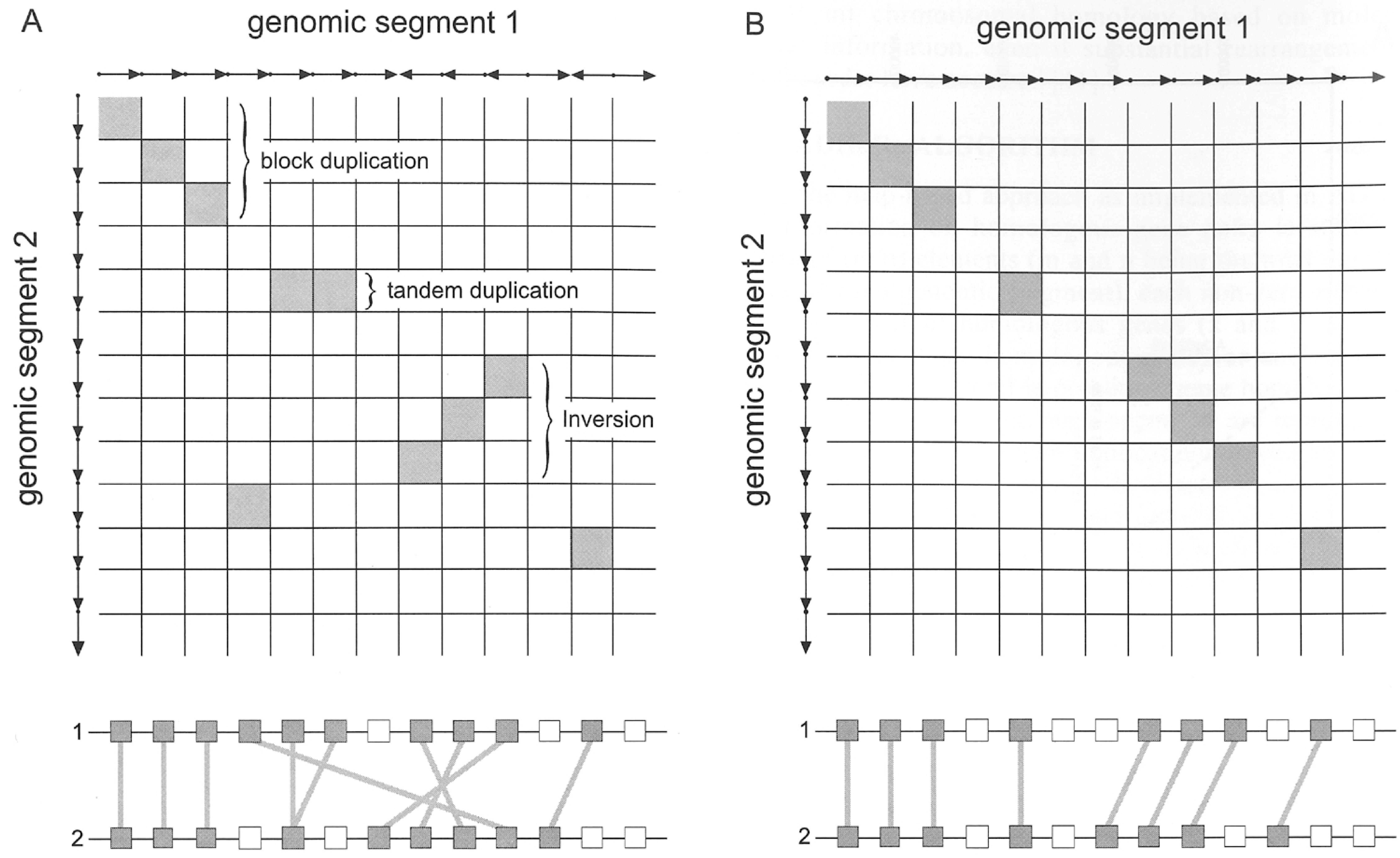

Fig. (2). Hypothetical gene homology matrix. Arrows on the axes of both segments represent genes on the genomic segments. Grey cells illustrate homologous genes (anchor points). In fig 2A, the original organisation of all genes, including tandem duplications and inversions, is shown. Fig. 2B shows the same gene homology matrix after remapping of tandem duplications and the removal of irrelevant single data points, i.e., homologous genes that are most likely not part of the block duplication. In addition, the small inverted colinear segment of 3 anchor points was restored to its original orientation, in order to create a larger colinear region.

a statistical test to validate whether colinearity of genetic markers represented genuine homology or could be expected by chance. To this end, the number of anchor points (i.e. homologous genes) within a colinear segment together with the size of the segment was compared with colinear segments found in a large number of randomised data. If the original colinear segment contains more markers or markers in closer proximity than expected by chance, the conclusion is that both segments are indeed homologous. This is usually implemented as a statistical test (a so-called permutation test or Monte Carlo simulation), sampling a large number of reshuffled data sets and calculating the probability that a colinear region, characterised by a number of conserved genes and an average gap size, can be found by chance.

Several recently published analyses have applied statistical validation through comparisons of observed data with expected data obtained by randomization tests [39, 43-45]. Although frequently done, the selection of colinear segments based solely on the number of anchor points within a colinear region is not entirely correct. This is due to the fact that the significance of colinearity strongly depends on the overall distribution of the homologous genes in a colinear segment, rather than on the total number of homologous genes [46]. One can easily imagine that the significance of 7 homologous genes within a colinear region of 15 genes is much higher than a colinear region of 100 genes with 7 homologous genes. Therefore, taking into account the number of anchor points in a cluster together with the average distance between all anchor points in a cluster (or reciprocal density) provides a more reliable way to calculate the probability that a cluster detected in the real dataset could have been generated by chance. This will result in small but dense clusters being retained, whereas loose small clusters will be rejected, since the chance that they were generated by chance is high.

A major drawback of the validation of colinearity through the comparison with randomised datasets is that the analysis of the large number of permutated datasets (typically 100 or 1000) is computationally expensive and in many cases more time-consuming than analysing the original dataset. Consequently, new methods have been developed for the validation of colinearity that do not require the presence of randomised data [47, 48].

SELECTION AND IDENTIFICATION OF HOMOLOGOUS GENES

In order to identify statistically significant orthologous or paralogous colinear segments based on the gene homology matrix, it is important to use strict criteria before concluding whether two genes (or markers) are anchor points. In the case of genetic maps, information about similar units – applied for describing the chromosome - is derived from markers that cross-hybridize on different chromosomal locations, whereas in sequenced genomes anchor points are

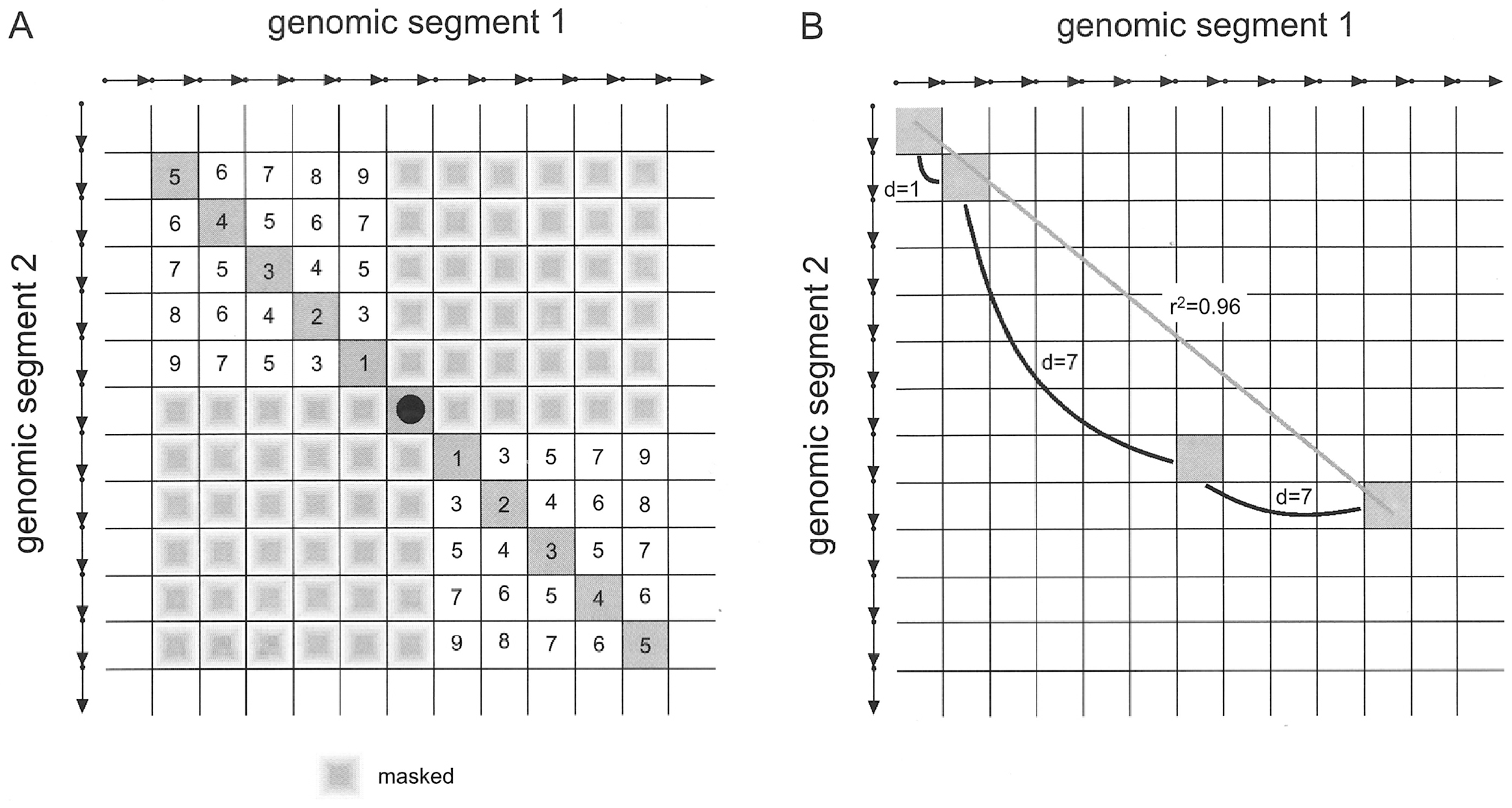

Fig. (3). Application of the diagonal pseudo distance (DPD) function to the detection of elements with diagonal proximity in the gene homology matrix. Fig. 3A shows the DPD for a given cell in the matrix to the central black dot (anchor point). The diagonal pseudo distance is smaller for diagonally orientated elements (grey boxes) than for elements deviating from the diagonal. Shaded boxes represent elements (genes) with an infinite distance to the central dot, since these elements are unlikely to be part of the duplicated segment that contains the anchor point (black dot). Fig. 3B shows the iterative clustering of elements for a colinear region with positive orientation (i.e. from top left to down right) in the homology matrix. All genes lie within a maximum gap distance G (e.g., 30) of each other. The best-fit line and its coefficient of determination (r^2) shows the quality of the cluster, which is clearly above the predefined Q value cut-off, here set to 0.9. As a result, all four homologous genes are considered to have been arisen by a block duplication.

simply homologous DNA or protein sequences. In the map-based approach, usually lists of predicted genes resembling the order of the genes on the chromosome are used for comparing genomic segments. Recently, Pevzner and Tesler (2003) used local similarities at the DNA level to compare the genomes of human and mouse, bypassing problems due to possible erroneous gene annotation. Nevertheless, as discussed above, homology at evolutionary distances where only protein similarity is conserved is missed. A possible solution, not yet implemented as far as we know, would be to identify homology between two segments by combining local similarities both at the DNA level and protein level. This method would have the advantage that it offers higher resolution compared to using only protein sequences and consequently should provide a more accurate view of the actual similarities between genomic sequences, both in coding and non-coding regions.

A first crucial step in applying the map-based approach as described above is the identification of homologous genes. Usually, an all-against-all sequence similarity search (e.g. BLASTP; [23]) is performed to find homologous proteins. Apart from applying an E-value or a similarity score cut-off, additional parameters such as the coverage of the alignable region on both potentially homologous genes can be applied to select 'suitable' homologues (for examples, [36, 37]). A major problem in identifying homologous genes based on sequence similarity is the discrimination between paralogous and orthologous genes, especially if genes belong to large multigene families [49]. For example, finding colinearity considering gene families with only one member in each genome will provide strong evidence to define truly orthologous segments between distantly related genomes. In contrast, the inclusion of large gene families will introduce a large number of homologous anchor points in the GHM of which only a very small fraction represents genuine orthology. Therefore, prior to the construction of the GHM, one should consider to first define all gene families and their sizes using specifically designed cluster algorithms [50-53]. In order to reduce the noise created by paralogy, only small gene families could then be selected and included in the analysis.

LARGE-SCALE GENE DUPLICATION EVENTS AND GENE LOSS

Often, very degenerated block duplications that originated hundreds of millions of years ago cannot be identified as such by directly comparing the duplicated segments. Differential gene loss, which is responsible for the loss of a different, but complementary set of genes on both paralogous genomic segments, makes it impossible to detect significant colinearity by directly comparing anciently duplicated regions. Therefore, two genomic segments in the same genome form a ghost duplication when their homology can only be inferred through comparison with the genome of another species [54]. In Figure **4**, the chromosomal segments A3.1 and A2.1 from *Arabidopsis* clearly show a pattern of differential gene loss when compared with the rice segment R10.1, since a number of genes located on the rice segment have been lost in one of the two paralogous segments of *Arabidopsis* (e.g. genes belonging to gene family 6733 (serine/threonine protein kinase), 4240 (bZIP leucine zipper) and 7796 (palmitoyl-protein thioesterase precursor)). Based on similar principles, hidden duplications can be inferred, which are heavily degenerated block duplications that cannot be identified by directly comparing both duplicated segments with each other, but only through comparison with a third segment of the same genome [55]. Consequently, hidden duplications are important to consider for determining the actual number of duplication events that have occurred over time, as previously demonstrated for *Arabidopsis* [56, 40]. Indeed, by taking into account hidden duplications, one can often group additional segments in a multiplicon (a set of mutually homologous segments), as shown in (Fig. **4**). The number of segments in a multiplicon, referred to as the multiplication level [40, 48, 55], can be used to infer the number of duplication events that must have occurred. For example, the presence of three homologous rice segments in the multiplicon, shown in (Fig. **4**), reveals that 2 duplication events must have occurred.

Apart from combining data of two genomes, Wong and co-workers [57] integrated partial sequence information of 14 related yeast strains in order to find evidence for an entire genome duplication event in *S. cerevisiae*. In their approach, the combination of a large number of chromosomal homologous segments allowed detecting heavily degraded duplicated regions, scattered throughout the genome.

GENOMIC PROFILES: AN EXTENSION TO THE MAP-BASED APPROACH

Although considering transitive homologies such as hidden and ghost duplications allows the identification of many additional, previously undetectable, homologous genomic segments, it still requires that each of the homologous segments show significant colinearity with at least one other homologous segment. However, it is possible that, within a given multiplicon, one or more segments have diverged that much from the others in gene content and gene order, that they no longer show any clear colinearity with any of the other segments. Such segments that are in the twilight zone of genomic homology cannot be detected with any of the currently available methods. Recently, we have been developing a new software to uncover chromosomal segments that are homologous (in respect with having common ancestry) to others, but can no longer be identified as such due to extreme gene loss. This is done by aligning clearly colinear segments and using this alignment as a 'genomic profile' that combines gene content and order information from multiple segments to detect these heavily degenerated homology relationships (see Fig. **5**); [48].

After the initial detection of a level 2 multiplicon with the basic ADHoRe algorithm, an alignment of the two segments that form this multiplicon, is created where the anchor points of the multiplicon are positioned in the same columns. Using this alignment now as a profile, a new type of homology matrix can be constructed in which the gene products of a segment are compared to the gene products of the profile. Once this homology matrix is constructed, it is again presented to the basic ADHoRe algorithm, which will again detect clusters of anchor points applying the same statistical validation method as described before. This time, however, new significant clusters will not reveal homology between two individual segments, but between the two segments inside the profile (i.e. the initial level 2 multiplicon) and a third segment. Because this type of GHM combines gene content and order information of the different segments in the profile, it is possible to detect homology relationships with a third segment that could not be recognised by directly comparing any of the segments of the multiplicon individually with this third segment. If such a third segment is detected, it is added to the multiplicon, thereby increasing its multiplication level, and the corresponding profile is updated by aligning the new segment to it. The entire detection process can now be repeated with the newly obtained profile [48].

BIOLOGICAL IMPLICATIONS OF LARGE-SCALE GENE DUPLICATIONS FOR GENE FUNCTION

The widespread occurrence of large and small-scale duplication events highly complicates the extrapolation of functional relationships between homologous genes of different species. Whereas one-to-one orthologous relationships suggest conservation of gene function, complex many-to-many homologous relationships offer limited information regarding conservation or divergence of gene function [58]; (see Fig. **4**). Although initially duplicated genes harbour redundant gene function, models have been formulated to explain the evolution of new functions (neo-functionalisation) or preservation of both duplicates by sub-functionalisation, where both members of a pair experience degenerative mutations that reduce their joint levels and patterns of activity to that of the single ancestral gene [59]. Some biological examples of sub-functionalisation have been documented [60], but it remains unclear whether this model accounts for the majority of preserved gene duplicates.

One way to further understand the evolutionary mechanisms underlying the expansion of gene families is to combine segmental or tandem duplications with gene phylogenies. Recently, Cannon and co-workers [61] developed a suite of programs for the detailed analysis of gene families combining comparative genomic positional information with phylogenetic reconstructions. As such, the impact of tandem and segmental duplications on gene family evolution can be inferred, allowing scientists to get deeper insights into the evolution of gene sub-families, which might be associated with functional divergence, or the acquisition of extra, potentially redundant gene copies in particular species. Finally, this approach can provide valuable clues about conserved gene function in orthologous genes and functional divergence in paralogous genes.

CONCLUSION

It is clear that large-scale genome sequencing and advanced comparative sequence analysis offer a powerful combination to study the complex evolutionary forces that shape the structure of genomes. The analysis of complete genomes and the comparison of gene organisation in related

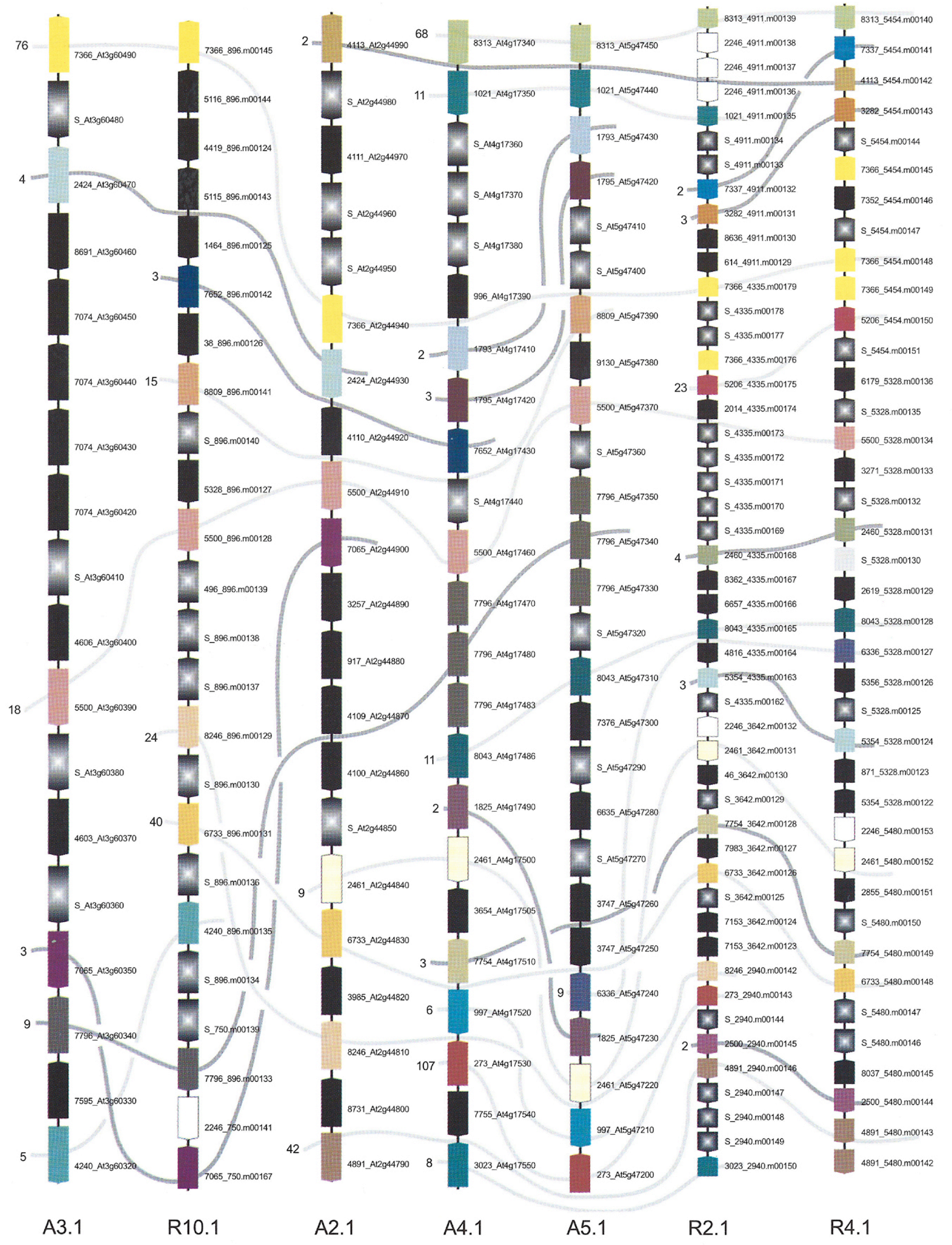

Fig. (4). Set of homologous chromosomal segments (multiplicon) of *Arabidopsis* (segments A) and rice (segments R). Boxes represent the genes on the chromosomal segments whereas connecting lines indicate the anchor points (i.e. homologous genes part of the same gene family). Dark grey connecting lines show gene families of which 50% or more of all genes are present in the multiplicon shown (see text for details). Therefore, these genes provide a particularly strong case for homology. For each genomic segment, the names of the genes preceded by the gene family ID are shown. Grey shaded boxes represent genes with no homologues in *Arabidopsis* and rice (gene family 'S' for singleton) and white boxes represent annotated genes with high similarity to retrotransposons. By considering the colinearity between *Arabidopsis* and rice, a set of, at first sight unrelated, *Arabidopsis* segments can be joined into a multiplicon with multiplication level 4 (i.e. the number of homologous segments in a multiplicon). Vice versa, this colinearity reveals that all three rice segments are linked with each other by two duplication events.

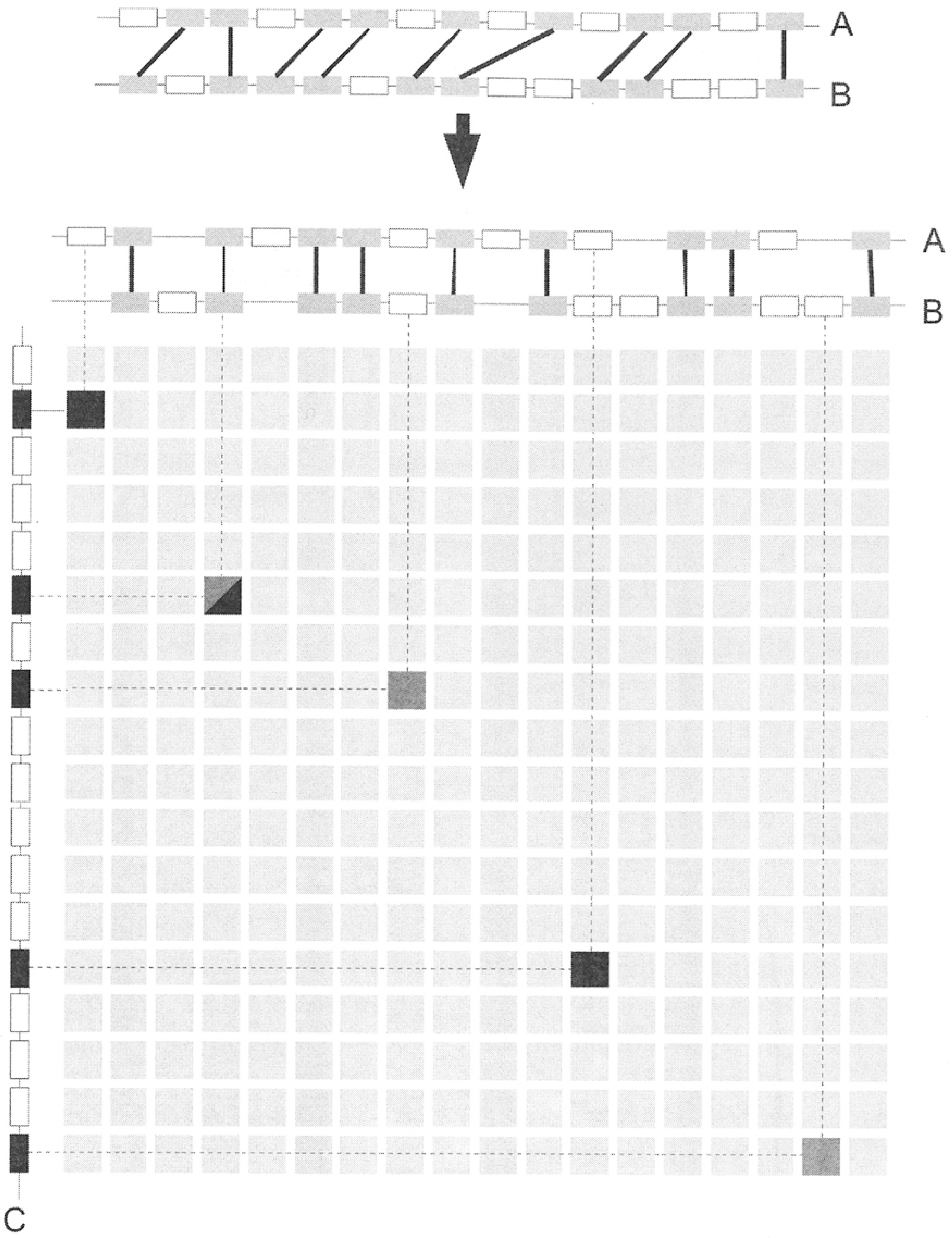

Fig. (5). Detection of homology through a genomic profile. The upper section shows an initially detected level 2 multiplicon (a pair of homologous chromosomal segments). The grey boxes connected by black lines represent pairs of homologous genes (anchor points) between the two segments. The lower section shows the construction of a homology matrix using this multiplicon as a profile. To accomplish this, the multiplicon is first aligned by inserting gaps at the proper positions (depicted by empty spaces in the alignment). The homology matrix can now be constructed by comparing this profile with the genes of a chromosomal segment C (shown on the left of the matrix). Anchor points in the matrix are detected whenever a gene of this chromosomal segment belongs to the same gene family as one of the genes in any of the segments in the profile. The black squares represent homologues between segments A and C, the dark grey between B and C. The black/dark-grey square denotes a gene that has a homologue on both segment A and B. Combining segments A and B in a profile thus results in 5 anchor points with segment C, whereas the individual segments A and B only have 3 anchor points with segment C, which might be too few to decide on statistical significant homology.

species finally allows scientists, at different levels of resolution from large-scale events such as translocations, duplications and segmental deletions to single-base pair differences, to unravel processes that drive gene and genome evolution [62]. Moreover, through the development of novel computational methods that allow the reliable detection of remnants of ancient large-scale gene duplication events, the evolutionary past of many eukaryotic genomes starts to reveal its secrets.

ACKNOWLEDGEMENTS

We would like to thank all members from the Bioinformatics and Evolutionary Genomics group for helpful discussions. K.V. and C.S. are indebted to the Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie for a predoctoral fellowship.

REFERENCES

[1] Hardison, R. C.; Oeltjen, J. and Miller, W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome. Res.*, **1997,** *7*: 959-66.

[2] Thacker, C., Marra, M. A., Jones, A., Baillie, D. L. and Rose, A. M. Functional genomics in Caenorhabditis elegans: An approach involving comparisons of sequences from related nematodes. *Genome. Res.*, **1999,** *9*: 348-59

[3] Waterston, R. H.; Lindblad-Toh, K.; Birney, E.; Rogers, J.; Abril, J. F.; Agarwal, P.; Agarwala, R.; Ainscough, R.; Alexandersson, M.; An, P.; Antonarakis, S. E.; Attwood, J.; Baertsch, R.; Bailey, J.; Barlow, K.; Beck, S.; Berry, E.; Birren, B.; Bloom, T.; Bork, P.; Botcherby, M.; Bray, N.; Brent, M. R.; Brown, D. G.; Brown, S. D.; Bult, C.; Burton, J.; Butler, J.; Campbell, R. D.; Carninci, P.; Cawley, S.; Chiaromonte, F.; Chinwalla, A. T.; Church, D. M.; Clamp, M.; Clee, C.; Collins, F. S.; Cook, L. L.; Copley, R. R.; Coulson, A.; Couronne, O.; Cuff, J.; Curwen, V.; Cutts, T.; Daly, M.; David, R.; Davies, J.; Delehaunty, K. D.; Deri, J.; Dermitzakis, E. T.; Dewey, C.; Dickens, N. J.; Diekhans, M.; Dodge, S.; Dubchak, I.; Dunn, D. M.; Eddy, S. R.; Elnitski, L.; Emes, R. D.;

Eswara, P.; Eyras, E.; Felsenfeld, A.; Fewell, G. A.; Flicek, P.; Foley, K.; Frankel, W. N.; Fulton, L. A.; Fulton, R. S.; Furey, T. S.; Gage, D.; Gibbs, R. A.; Glusman, G.; Gnerre, S.; Goldman, N.; Goodstadt, L.; Grafham, D.; Graves, T. A.; Green, E. D.; Gregory, S.; Guigo, R.; Guyer, M.; Hardison, R. C.; Haussler, D.; Hayashizaki, Y.; Hillier, L. W.; Hinrichs, A.; Hlavina, W.; Holzer, T.; Hsu, F.; Hua, A.; Hubbard, T.; Hunt, A.; Jackson, I.; Jaffe, D. B.; Johnson, L. S.; Jones, M.; Jones, T. A.; Joy, A.; Kamal, M.; Karlsson, E. K.; Karolchik, D.; Kasprzyk, A.; Kawai, J.; Keibler, E.; Kells, C.; Kent, W. J.; Kirby, A.; Kolbe, D. L.; Korf, I.; Kucherlapati, R. S.; Kulbokas, E. J.; Kulp, D.; Landers, T.; Leger, J. P.; Leonard, S.; Letunic, I.; Levine, R.; Li, J.; Li, M.; Lloyd, C.; Lucas, S.; Ma, B.; Maglott, D. R.; Mardis, E. R.; Matthews, L.; Mauceli, E.; Mayer, J. H.; McCarthy, M.; McCombie, W. R.; McLaren, S.; McLay, K.; McPherson, J. D.; Meldrim, J.; Meredith, B.; Mesirov, J. P.; Miller, W.; Miner, T. L.; Mongin, E.; Montgomery, K. T.; Morgan, M.; Mott, R.; Mullikin, J. C.; Muzny, D. M.; Nash, W. E.; Nelson, J. O.; Nhan, M. N.; Nicol, R.; Ning, Z.; Nusbaum, C.; O'Connor, M. J.; Okazaki, Y.; Oliver, K.; Overton-Larty, E.; Pachter, L.; Parra, G.; Pepin, K. H.; Peterson, J.; Pevzner, P.; Plumb, R.; Pohl, C. S.; Poliakov, A.; Ponce, T. C.; Ponting, C. P.; Potter, S.; Quail, M.; Reymond, A.; Roe, B. A.; Roskin, K. M.; Rubin, E. M.; Rust, A. G.; Santos, R.; Sapojnikov, V.; Schultz, B.; Schultz, J.; Schwartz, M. S.; Schwartz, S.; Scott, C.; Seaman, S.; Searle, S.; Sharpe, T.; Sheridan, A.; Shownkeen, R.; Sims, S.; Singer, J. B.; Slater, G.; Smit, A.; Smith, D. R.; Spencer, B.; Stabenau, A.; Stange-Thomann, N.; Sugnet, C.; Suyama, M.; Tesler, G.; Thompson, J.; Torrents, D.; Trevaskis, E.; Tromp, J.; Ucla, C.; Ureta-Vidal, A.; Vinson, J. P.; Von Niederhausern, A. C.; Wade, C. M.; Wall, M.; Weber, R. J.; Weiss, R. B.; Wendl, M. C.; West, A. P.; Wetterstrand, K.; Wheeler, R.; Whelan, S.; Wierzbowski, J.; Willey, D.; Williams, S.; Wilson, R. K.; Winter, E.; Worley, K. C.; Wyman, D.; Yang, S.; Yang, S. P.; Zdobnov, E. M.; Zody, M. C. and Lander, E. S. Initial sequencing and comparative analysis of the mouse genome. *Nature,* **2002,** *420*: 520-62.

[4] Wortman, J. R.; Haas, B. J.; Hannick, L. I.; Smith, R. K.; Jr.; Maiti, R.; Ronning, C. M.; Chan, A. P.; Yu, C.; Ayele, M.; Whitelaw, C. A.; White, O. R. and Town, C. D. Annotation of the *Arabidopsis* genome. *Plant Physiol.,* **2003,** *132*: 461-8.

[5] Cliften, P. F.; Hillier, L. W.; Fulton, L.; Graves, T.; Miner, T.; Gish, W. R.; Waterston, R. H. and Johnston, M. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome. Res.,* **2001,** *11*: 1175-86.

[6] Kellis, M.; Patterson, N.; Endrizzi, M.; Birren, B. and Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature,* **2003,** *423*: 241-54.

[7] Levy, S.; Hannenhalli, S. and Workman, C. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics,* **2001,** *17*: 871-7.

[8] Dermitzakis, E. T.; Reymond, A.; Lyle, R.; Scamuffa, N.; Ucla, C.; Deutsch, S.; Stevenson, B. J.; Flegel, V.; Bucher, P.; Jongeneel, C. V. and Antonarakis, S. E. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature,* **2002,** *420*: 578-82.

[9] Dieterich, C.; Cusack, B.; Wang, H.; Rateitschak, K.; Krause, A. and Vingron, M. Annotating regulatory DNA based on man-mouse genomic comparison. *Bioinformatics,* **2002,** *18*: S84-90.

[10] Alexandersson, M.; Cawley, S. and Pachter, L. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome. Res.,* **2003,** *13*: 496-502

[11] Pedersen, J. S. and Hein, J. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics,* **2003,** *19*: 219-27

[12] Flicek, P.; Keibler, E.; Hu, P.; Korf, I. and Brent, M. R. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome. Res.,* **2003,** *13*: 46-54.

[13] Collins, J. E.; Goward, M. E.; Cole, C. G.; Smink, L. J.; Huckle, E. J.; Knowles, S.; Bye, J. M.; Beare, D. M. and Dunham, I. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome. Res.,* **2003,** *13*: 27-36.

[14] Clamp, M.; Andrews, D.; Barker, D.; Bevan, P.; Cameron, G.; Chen, Y.; Clark, L.; Cox, T.; Cuff, J.; Curwen, V.; Down, T.; Durbin, R.; Eyras, E.; Gilbert, J.; Hammond, M.; Hubbard, T.; Kasprzyk, A.; Keefe, D.; Lehvaslaiho, H.; Iyer, V.; Melsopp, C.; Mongin, E.; Pettett, R.; Potter, S.; Rust, A.; Schmidt, E.; Searle, S.; Slater, G.; Smith, J.; Spooner, W.; Stabenau, A.; Stalker, J.; Stupka, E.; Ureta-Vidal, A.; Vastrik, I. and Birney, E. Ensembl 2002: accommodating comparative genomics. *Nucleic. Acids Res.,* **2003,** *31*: 38-42.

[15] McCutcheon, J. P. and Eddy, S. R. Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. *Nucleic. Acids. Res.,* **2003,** *31*: 4119-28.

[16] Lagos-Quintana, M.; Rauhut, R.; Meyer, J.; Borkhardt, A. and Tuschl, T. New microRNAs from mouse and human. *RNA.,* **2003,** *9*: 175-9.

[17] Nadeau, J. H. and Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA,* **1984,** *81*: 814-8.

[18] Seoighe, C.; Federspiel, N.; Jones, T.; Hansen, N.; Bivolarovic, V.; Surzycki, R.; Tamse, R.; Komp, C.; Huizar, L.; Davis, R. W.; Scherer, S.; Tait, E.; Shaw, D. J.; Harris, D.; Murphy, L.; Oliver, K.; Taylor, K.; Rajandream, M. A.; Barrell, B. G. and Wolfe, K. H. Prevalence of small inversions in yeast gene order evolution. *Proc. Natl. Acad. Sci. USA,* **2000,** *97*: 14433-7.

[19] Bennetzen, J. L. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant. Cell.,* **2000,** *12*: 1021-9.

[20] Ranz, J. M.; Casals, F. and Ruiz, A. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. *Genome. Res.,* **2001,** *11*: 230-9.

[21] Coghlan, A. and Wolfe, K. H. Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. *Genome. Res.,* **2002,** *12*: 857-67.

[22] Pevzner, P. and Tesler, G. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome. Res.,* **2003,** *13*: 37-45.

[23] Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W. and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **1997,** *25*: 3389-402.

[24] Pearson, W. R. and Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA,* **1988,** *85*: 2444-8.

[25] Smith, T. F. and Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.,* **1981,** *147*: 195-7.

[26] Sonnhammer, E. L. and Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene,* **1995,** *167*: GC1-10.

[27] Delcher, A. L.; Kasif, S.; Fleischmann, R. D.; Peterson, J.; White, O. and Salzberg, S. L. Alignment of whole genomes. *Nucleic Acids Res.,* **1999,** *27*: 2369-76

[28] Schwartz, S.; Zhang, Z.; Frazer, K. A.; Smit, A.; Riemer, C.; Bouck, J.; Gibbs, R.; Hardison, R. and Miller, W. PipMaker--a web server for aligning two genomic DNA sequences. *Genome. Res.,* **2000,** *10*: 577-86

[29] Ning, Z.; Cox, A. J. and Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome. Res.,* **2001,** *11*: 1725-9.

[30] Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome. Res.,* **2002,** *12*: 656-64

[31] Schwartz, S.; Kent, W. J.; Smit, A.; Zhang, Z.; Baertsch, R.; Hardison, R. C.; Haussler, D. and Miller, W. Human-mouse alignments with BLASTZ. *Genome. Res.,* **2003,** *13*: 103-7

[32] Bray, N.; Dubchak, I. and Pachter, L. AVID: A global alignment program. *Genome. Res.,* **2003,** *13*: 97-102.

[33] Brudno, M.; Do, C. B.; Cooper, G. M.; Kim, M. F.; Davydov, E.; Green, E. D.; Sidow, A. and Batzoglou, S. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome. Res.,* **2003,** *13*: 721-31

[34] Schwartz, S.; Elnitski, L.; Li, M.; Weirauch, M.; Riemer, C.; Smit, A.; Green, E. D.; Hardison, R. C. and Miller, W. MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.,* **2003,** *31*: 3518-24.

[35] Wolfe, K. H. and Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature,* **1997,** *387*: 708-13.

[36] McLysaght, A.; Hokamp, K. and Wolfe, K. H. Extensive genomic duplication during early chordate evolution. *Nat. Genet.,* **2002,** *31*: 200-4.

[37] Li, W. H.; Gu, Z.; Cavalcanti, A. R. and Nekrutenko, A. Detection of gene duplications and block duplications in eukaryotic genomes. *J. Struct. Funct. Genomics,* **2003,** *3*: 27-34

[38] Blanc, G.; Hokamp, K. and Wolfe, K. H. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.,* **2003,** *13*: 137-44.

[39] Vandepoele, K.; Saeys, Y.; Simillion, C.; Raes, J. and Van De Peer, Y. The Automatic Detection of Homologous Regions (ADHoRe) and Its Application to Microcolinearity Between *Arabidopsis* and Rice. *Genome. Res.,* **2002a**, *12*: 1792-801.

[40] Simillion, C.; Vandepoele, K.; Van Montagu, M. C.; Zabeau, M. and Van De Peer, Y. The hidden duplication past of *Arabidopsis* thaliana. *Proc. Natl. Acad. Sci. USA,* **2002,** *99*: 13627-32.

[41] Hampson, S.; McLysaght, A.; Gaut, B. and Baldi, P. LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res.,* **2003,** *13*: 999-1010.

[42] Gaut, B. S. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res.,* **2001,** *11*: 55-66.

[43] Friedman, R. and Hughes, A. L. Pattern and timing of gene duplication in animal genomes. *Genome. Res.,* **2001,** *11*: 1842-7.

[44] Vision, T. J.; Brown, D. G. and Tanksley, S. D. The origins of genomic duplications in *Arabidopsis. Science,* **2000,** *290*: 2114-7

[45] Cavalcanti, A. O.; Ferreira, R. O.; Gu, Z. O. and Li, W. H. Patterns of Gene Duplication in Saccharomyces cerevisiae and Caenorhabditis elegans. *J. Mol. Evol.,* **2003,** *56*: 28-37.

[46] Durand, D. Vertebrate evolution: doubling and shuffling with a full deck. *Trends Genet,* **2003,** *19*: 2-5.

[47] Calabrese, P. P.; Chakravarty, S. and Vision, T. J. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics,* **2003,** *19*: I74-I80.

[48] Simillion, C.; Vandepoele, K.; Saeys, Y. and Van De Peer, Y. Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome. Res.,* **2004,** in press.

[49] Jensen, R. A. Orthologs and paralogs - we need to get it right. *Genome. Biol.,* **2001,** *2*: Interactions1002.

[50] Tatusov, R. L.; Galperin, M. Y.; Natale, D. A. and Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic. Acids. Res.,* **2000,** *28*: 33-6.

[51] Li, W. H.; Gu, Z.; Wang, H. and Nekrutenko, A. Evolutionary analyses of the human genome. *Nature,* **2001,** *409*: 847-9.

[52] Remm, M.; Storm, C. E. and Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.,* **2001,** *314*: 1041-52.

[53] Enright, A. J.; Van Dongen, S. and Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.,* **2002,** *30*: 1575-84.

[54] Vandepoele, K.; Simillion, C. and Van de Peer, Y. Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends Genet.,* **2002b,** *18*: 606-8.

[55] Vandepoele, K.; Simillion, C. and Van de Peer, Y. Evidence that rice, and other cereals, are ancient aneuploids. *Plant Cell* 15, 2192-2202.

[56] Ku, H. M.; Vision, T.; Liu, J. and Tanksley, S. D. Comparing sequenced segments of the tomato and arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA,* **2000,** *97*: 9121-6

[57] Wong, S.; Butler, G. and Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl. Acad. Sci. USA,* **2002,** *99*: 9272-7.

[58] Doyle, J. J. and Gaut, B. S. Evolution of genes and taxa: a primer. *Plant. Mol. Biol.,* **2000,** *42*: 1-23

[59] Lynch, M. and Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics,* **2000,** *154*: 459-73.

[60] Prince, V. E. and Pickett, F. B., Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet,* **2002,** *3*: 827-37.

[61] Cannon, S. B. and Young, N. D. OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC. Bioinformatics,* **2003,** *4*: 35

[62] Eichler, E. E. and Sankoff, D. Structural dynamics of eukaryotic chromosome evolution. *Science,* **2003,** *301*: 793-7.