Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice

Klaas Vandepeele, Cedric Simillion and Yves Van de Peer

Genome analysis shows that large-scale gene duplications have occurred in fungi, animals and plants, creating genomic regions that show similarity in gene content and order. However, the high frequency of gene loss reduces colinearity resulting in duplicated regions that, in the extreme, no longer share homologous genes. Here, we show that by comparison with an appropriate second genome, such paralogous regions can still be identified.

Published online: 30 October 2002


genome sequencing projects reveal that genomes vary tremendously in size and organization, even among closely related organisms. This seems to be the result of a very dynamic process involving many different factors, such as recombinations, horizontal gene transfer, transposon...
activity, gene duplication and gene loss. In particular, duplications are being identified as important factors in the evolution of most genomes. Apart from small-scale tandem duplications, larger block duplications and even duplications of entire chromosomes or genomes are now postulated to have shaped the genomes of various animals, fungi and plants [1]. From a population genetics point of view [2], the frequency of gene preservation over a large evolutionary period after duplication is unexpectedly high and several models have recently been put forward to explain the retention of duplicates [3–5]. However, the most likely fate of a gene duplicate is nonfunctionalization and consequent gene loss [6].

This observation has consequences for the detection of duplicated regions in genomes. Identifying duplicated regions is usually based on a within-genome comparison that aims to define colinear regions (regions of conserved gene content and order) in different parts of the genome. In general, one tries to identify duplicated blocks of homologous genes that are statistically valid (i.e. that are probably not generated by chance). The statistics that determine colinearity usually depend on two factors, namely the number of pairs of genes that still can be identified as homologous (usually referred to as ‘anchor points’), and the distance over which these gene pairs are found, which usually depends on the number of ‘single’ genes that interrupt colinearity. When a putative colinear region has been detected, its statistical significance is usually evaluated by some sort of permutation test in which a large number of randomized datasets are sampled to calculate the probability that a cluster...
detected could have been generated by chance [7–10]. However, the high level of gene loss – together with phenomena such as translocations and chromosomal rearrangements – often renders it very difficult to find statistically significant homologous regions in the genome, particularly when the duplication events are ancient [11].

The search for traces of (ancient) large-scale gene duplications has received much attention lately, and hypotheses about the number and age of polyploidization events in eukaryotes are actively being discussed. Partly, this is because of the fact that the detection of homologous (paralogous) regions in genomes is not self-evident, for the reasons discussed above and, in consequence, the number of duplicated regions is likely to be underestimated. In plants, the systematic analysis of the Arabidopsis thaliana genome sequence has shown that this genome contains a large number of duplicated regions and that about 60% of the Arabidopsis genes occur in duplicated blocks [12–14]. Here, we show that additional duplicated regions can be discovered in Arabidopsis when its genome is compared with that of rice.

Recently, the draft genome sequences have been reported for two subspecies of rice [15,16], in addition to data being made available by the International Rice Gene Sequencing Project [17]. We used the IGSP data to compile a large set of BAC sequences for which the map position information is available and used these, where possible, to build longer rice contigs. This resulted in a dataset of 453 overlapping BACs, forming continuous genomic stretches of 62 Mb, and a remaining set of 821 individual BACs (representing 104 Mb). We compared these with the Arabidopsis genome to find statistically significant regions of colinearity between the genomes, using a new software tool called ADHoRe (for ‘automatic detection of homologous regions’) [10].

The comparison of rice, the major food source for billions of people and a model for larger cereal crop genomes [18] with Arabidopsis, a model plant organism for dicotyledons, revealed numerous examples of (short) genomic segments that shared conserved gene content and order, as reported previously [14,19,20]. In several cases, two (or more) regions of the Arabidopsis genome showed clear homology with a single region in rice. This is not surprising, because the Arabidopsis genome has undergone at least one [6,13], and probably more [7,14], polyploidizations. However, some of the duplicated regions escape detection in a within-genome comparison of Arabidopsis. More detailed analysis shows that each of these regions in Arabidopsis has lost a different set of genes (see Fig. 1a). This phenomenon, which we refer to as ‘differential gene loss’, turns the originally identical duplicated regions into two nonredundant sets of genes, divided over two distinct genome locations. Differential gene loss thus reduces the number of paralogs that can be identified by a within-genome comparison. For a few genes, both duplicates might have been retained, but in that case the number of anchor points is usually too small to detect significant colinearity when permutation tests are applied (Fig. 1b). Therefore, the use of intergenomic comparisons can help to recover block duplications that had seemingly disappeared.

By considering only a small amount of the rice genome sequence, we were able to detect several examples of such ‘ghost’ duplications in Arabidopsis. Once a completely assembled and well-annotated rice genome sequence is available, comparisons between rice and Arabidopsis, which diverged from one another ~200 million years ago [21] will probably reveal many more of such regions. Furthermore, most probably, many other examples of such ‘ghost’ duplications are waiting to be discovered in other eukaryotic genomes as well.

Acknowledgements
We thank Pierre Rouze for helpful discussions and Jerry Raes and John S. Taylor for a critical reading of the manuscript. K. V. and C. S. are indebted to the Vlaams Instituut voor de Bewerding van het Wetenschappelijk–Technisch Onderzoek in de Industrie for a predoctoral fellowship.

References
8 Gaut, B. S. (2001) Patterns of chromosomal duplication in maize and their implications for comparative maps of the genomes. Genome Res. 11, 55–66
10 Vandepoele, K. et al. (2001). The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. Genome Res. (in press)
19 Mayer, K. et al. (2001) Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of Arabidopsis thaliana. Genome Res. 11, 1167–1174

Klaas Vandepoele
Cedric Simillion
Yves Van de Peer*
Dept of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium.
*e-mail: yvdp@gengenp.rug.ac.be