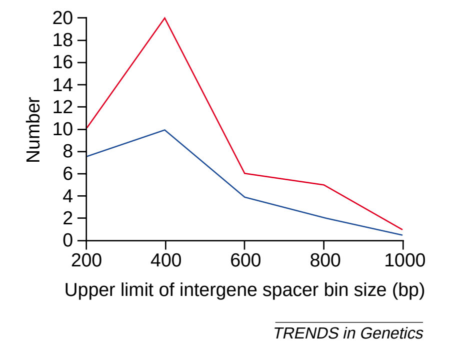
in the co-expressed class, whereas the apparent enrichment of conservation of gene pairs is highly significant.

Prior evidence suggests that divergently oriented genes (←→) are especially likely to belong to a single regulatory unit [11]. Within the highly co-expressed group, the genes in divergent orientation are indeed more common than expected from their overall frequency (we expect 65, but observe 85). Contrary to previous suggestions [11], we find a dearth of gene pairs in which both genes are in the same orientation (87 observed, 115 expected). Overall, there is a significant difference in the proportion of types in different orientations in the highly co-expressed class compared with their frequencies in the dataset as a whole ($\chi^2 = 13.88$, $v = 2$, $P < 0.001$).

Of the 42 pairs that are conserved within the highly co-expressed class, 19 (45%) are in the divergent orientation in yeast, approximately double their frequency within the dataset as a whole (G test of independence, $P < 0.01$), and higher than their frequency within the co-expressed class, although not significantly so (G test of independence, $P > 0.05$). Although the above results suggest that divergent orientation is important for co-regulation and for conservation of pairs, we do not find that the divergent genes retain their orientation at an especially high rate. Of the 19, 14 (74%) have the same orientation in *Candida*, which compares with 62% of conserved pairs that have the same direction in both species (i.e. 103 out of the sample of 166). Nonetheless, our findings are consistent with the observation [12] that between *S. cerevisiae* and *Candida albicans*, divergently transcribed gene-pairs that are conserved in evolution have a higher probability of being co-regulated than divergently transcribed gene pairs that are disrupted in evolution.



TRENDS in Genetics

Fig. 3. The number of conserved highly co-expressed gene pairs observed (red) and expected (blue) as a function of the intergene spacer size. The *x* axis numbers indicate the upper limit to the subgroup size (i.e. 200 represents genes with intergene spacer less than or equal to 200, 400, indicate 201 to 400 etc.).

We conclude that, consistent with the null neutral model, gene pairs that have small intergene spacer are the most likely to be conserved. This result emphasizes the need to control for the length of intergene spacer when testing hypotheses of gene order evolution. However, among the most highly co-expressed gene pairs, a clear signal of selection is evident, with co-expressed genes being retained at about twice the expected rate. Only a small proportion of this signal can be explained as a consequence of reduced intergene spacer. Consequentially the null neutral model cannot be considered an adequate description of gene order conservation.

Acknowledgements

We should like to thank Martijn Huynen and an anonymous referee for comments on an earlier version of the manuscript. C.P. is funded by a Royal Society/Nato visiting fellowship and L.D.H. by B.B.S.R.C

References

1 Wang, P.J. *et al.* (2001) An abundance of X-linked genes expressed in spermatogonia. *Nat. Genet.* 27, 422–426

2 Saifi, G.M. and Chandra, H.S. (1999) An apparent excess of sex- and reproduction-related genes on the human X chromosome. *Proc. R. Soc. London Ser. B* 266, 203–209

3 Lercher, M.J. *et al.* (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* 31, 180–183

4 Spellman, P.T. and Rubin, G.M. (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* 1, 5

5 Cohen, B.A. *et al.* (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* 26, 183–186

6 Blumenthal, T. *et al.* (2002) A global analysis of *Caenorhabditis elegans* operons. *Nature* 417, 851–854

7 Zhang, J.Z. and Nei, M. (1996) Evolution of antennapedia-class homeobox genes. *Genetics* 142, 295–303

8 Huynen, M.A. *et al.* (2001) Inversions and the dynamics of eukaryotic gene order. *Trends Genet.* 17, 304–306

9 Seoighe, C. *et al.* (2000) Prevalence of small inversions in yeast gene order evolution. *Proc. Natl. Acad. Sci. U. S. A.* 97, 14433–14437

10 Pál, C. *et al.* (2001) Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. *Mol. Biol. Evol.* 18, 2323–2326

11 Kruglyak, S. and Tang, H. (2000) Regulation of adjacent yeast genes. *Trends Genet.* 16, 109–111

12 Huynen, M.A. and Snel, B. Exploiting the variations in the genomic associations of genes to predict pathways and reconstruct their evolution. In *Frontiers in Computational Genomics* (Galperin, M.Y. and Koonin, E.V., eds), Horizon Scientific Press (in press)

Laurence D. Hurst*
Elizabeth J.B. Williams
Csaba Pál
Dept of Biology and Biochemistry, University of Bath, Claverton Down, Bath, UK BA2 7AY.
*e-mail: l.d.hurst@bath.ac.uk

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

Klaas Vandepoele, 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

http://tig.trends.com 0168-9525/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(02)02796-8

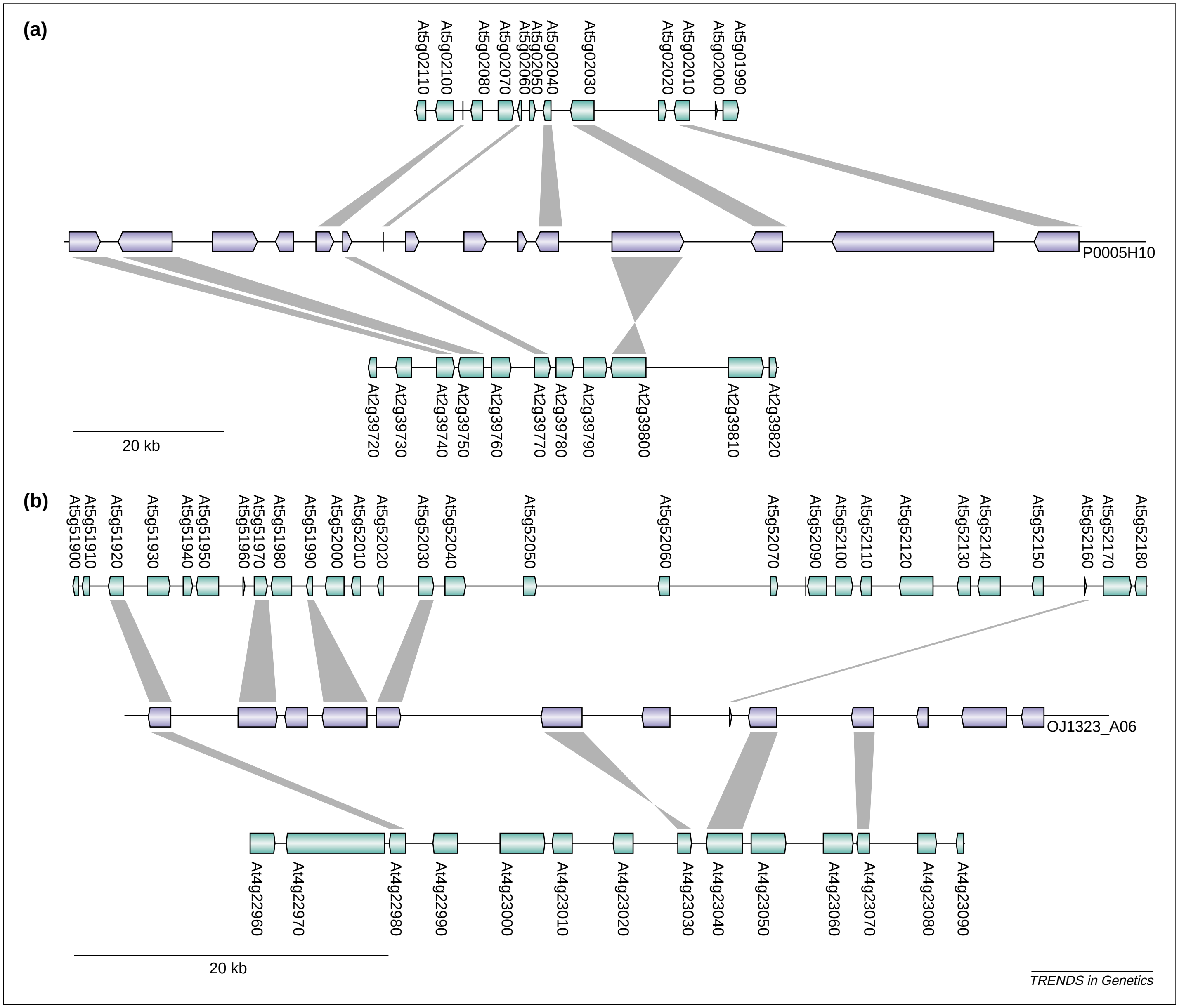


Fig. 1. 'Ghost' block duplications in the *Arabidopsis* genome. Homologous genes between *Arabidopsis* (green) and *Oryza sativa* (purple) are indicated by grey bands. (a) Two genomic segments of *Arabidopsis*, on chromosomes 2 (top) and 5 (bottom), map to the same rice segment. Therefore, these segments are paralogous and result from a duplication event within the *Arabidopsis* genome. Because of differential gene loss, the duplicated *Arabidopsis* segments no longer have any paralogous genes in common. As a result, this duplication can not be detected anymore. (b) 'Ghost' block duplication between *Arabidopsis* chromosomes 4 (top) and 5 (bottom). One anchor point (i.e. the paralogous gene pair At5g51920 – At4g22980) is still present on both segments, but is insufficient to detect microcolinearity between the two segments.

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 polyploidy 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 IRGSP 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 Rouzé for helpful discussions and Jeroen 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 Bevordering van het Wetenschappelijk–Technologisch Onderzoek in de Industrie for a predoctoral fellowship.

References

1 Wolfe, K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 333–341
2 Force, A. *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545
3 Gibson, T.J. and Spring, J. (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet.* 14, 46–49
4 Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473
5 Wagner, A. (2002) Selection and gene duplication: a view from the genome. *Genome Biol.* 3, reviews1012.1–1012.3
6 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155
7 Vision, T.J. *et al.* (2000). The origins of genomic duplications in *Arabidopsis*. *Science* 290, 2114–2117
8 Gaut, B, S. (2001) Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res.* 11, 55–66
9 Friedman, R. and Hughes, A. L. (2001) Gene duplication and the structure of eukaryotic genomes. *Genome Res.* 11, 373–381
10 Vandepoele, K. *et al.* The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.* (in press)
11 Ku, H.M. *et al.* (2000) 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. U. S. A.* 97, 9121–1126
12 Blanc, G. *et al.* (2000). Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell* 12, 1093–1101
13 The *Arabidopsis* Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
14 Simillion, C. *et al.* The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* (in press)
15 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100
16 Yu, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. spp. *indica*). *Science* 296, 79–92
17 Sasaki, T. and Burr, B. (2000). International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.* 3, 138–141
18 Shimamoto, K. and Kyozuka, J. (2002) Rice as a model for comparative genomics of plants. *Annu. Rev. Plant Biol.* 53, 399–419
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
20 Salse, J. *et al.* (2002) Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Res.* 30, 2316–2328
21 Wikström, N. *et al.* (2001). Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. London Ser. B.* 268, 2211–2220

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 Ghent, Belgium.
*e-mail: yvdp@gengenp.rug.ac.be