

Genome duplication strikes back

Jürg Spring

Institute of Zoology, University of Basel, Biocenter/Pharmazentrum, Klingelbergstrasse 50, CH-4056 Basel, Switzerland. e-mail: j.spring@unibas.ch

The human genome contains up to four paralogs of many *Drosophila* genes. Two rounds of whole-genome duplication followed by substantial gene loss could explain this pattern easily, but this hypothesis has often been questioned. Mounting evidence from the human genome sequence now confirms at least one genome duplication during early chordate evolution.

Gene and genome duplications have been discussed as prerequisites for further evolution since the publication in 1970 of Susumu Ohno's book, *Evolution by Gene Duplication*. Intense interest in the *Hox* clusters, and the fact that invertebrates have only one whereas mice and humans have four each on different chromosomes, has helped to spread the idea that two whole-genome duplications could have been instrumental in vertebrate evolution. However, many investigators have taken issue with this simple view. In 1998, Ken Wolfe, a pioneer in yeast and plant genome duplication studies, spoke for the skeptics when he asked: where's the evidence?

In May's issue and this issue of *Nature Genetics*, there are three complementary studies that seem to provide some

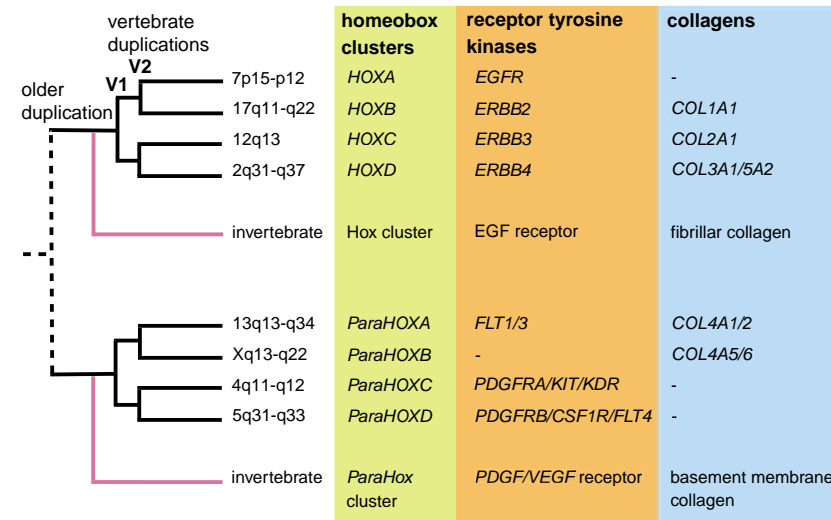
answers. On page 200 of this issue, Aoife McLysaght and colleagues¹ document extensive genomic duplication during early chordate evolution. They suggest that at least one round of polyploidy was involved and conclude that humans (13%), like yeast (16%) and *Arabidopsis thaliana* (25%), are paleopolyploids, meaning that after ancient genome duplications, 13% of human genes are still recognizable as duplicates. On page 205 of this issue, Xun Gu and colleagues² show that both large and small-scale duplications are required to explain the age-distribution of human gene families. And Laurent Abi-Rached and colleagues³ found evidence of *en bloc* duplication by comparison of amphioxus and human genes in the major histocompatibility complex (MHC) regions.

Duplications abound

When it became clear that many invertebrates have a single *Hox* cluster corresponding to four human equivalents, it was expected that the duplication of developmentally important genes would be shown to be important for the appearance of novelties in vertebrate developmental pathways⁴. In 1996, it became clear that, in addition to *HOX* clusters and their neighbors, the MHC genes are found in four large paralogous genome regions⁵. Moreover, it appeared that genes of all classes, from those encoding aldolases to those encoding zinc-finger proteins, could be found in remnants of tetrapacks on all human chromosomes⁵. Although it seemed obvious that this happened by genome duplication, it was unclear whether it occurred by endoduplication (autopolyploidy) or involved hybridization (allopolyploidy). Much of the debate over gene-by-gene versus whole-genome duplication surrounded the expectations that phylogenetic trees of human paralogs should take the symmetrical form (AB)(CD) as idealized in the Figure, and that duplications should have happened at the same time^{6,7}. This would only be true, however, for autopolyploids, and would require equal mutation rates.

The complete genome will tell

One of the big disappointments of the first analysis of the complete human genome sequence^{8,9} was that it could not be decided whether individual gene duplications or two whole-genome duplications were the basis of its observed fourfold complexity. Unfortunately, the sequence is still incomplete, and even the most thorough study by McLysaght *et al.*¹ is based on only about 24,000 genes. From this set, about 6,000 are found in 1,642 paralogous regions, but only 191 families could be used for molecular clock analysis. Similarly, the age-distribution of human gene families described by Gu *et al.*² was only calculated for 1,739 duplicates belonging to 749 families. The comparison of human and amphioxus MHC regions by



More genome duplications? *HOX* clusters are just the best known examples of 1,642 paralogous regions detectable in the human genome¹. The similarity of the four *HOX* clusters with the EGF receptor family and fibrillar collagens as neighbors to the *ParaHox* clusters with the PDGF/VEGF receptor family and basement membrane collagens as neighbors suggests the existence of even more ancient genome duplications. *D. melanogaster* and *C. elegans* are highly derived organisms and of limited value for genome comparison. The genome sequence of amphioxus, which would be the preferred invertebrate model organism for genome comparison, has not yet been completed. Most gene families also underwent tandem duplications and deletions, which is one of the problems when invertebrate genes have to be assigned to vertebrate orthologs. *HOX* clusters contain 9–11 genes and the *ParaHoxA* cluster consists of 3 genes, *GSH1*, *IPF1* and *CDX2*, while in *ParaHoxB*, *C*, *D* only one gene is left.



Abi-Rached *et al.*³ concentrates on 31 gene families in relatively well-covered human genome regions; however, the incompleteness of the amphioxus data reminds us how much there is still to be done. Interestingly, the MHC and the *HOX* regions are still the best examples of gene duplication, with regions of 20 to 40 Mb containing 26 to 29 duplicates¹. Completely sequenced and annotated genomes of humans and amphioxus will be necessary to fill the gaps.

The problem

Why is it so difficult to prove the obvious? Everyone can identify candidate paralogous regions such as the Hox and MHC, but the invertebrate orthologs from which they arose are not always easy to identify. *Drosophila melanogaster* and *Caenorhabditis elegans* are highly derived organisms even within their phyla, and are often too remote as an invertebrate outgroup. Urochordates or amphioxus, as primitive members of the phylum Chordata, would be perfect for comparisons with vertebrates. The genomes of some urochordates are ideally small¹⁰, although it will be important to show that this did not cause too many aberrations. Unfortunately, the genome of amphioxus is quite large, and not even the MHC region has been completely sequenced³.

The rate of individual gene duplication is so high that whole-genome duplications are not necessary to explain the human gene number¹¹. But tandem duplications are also associated with a high rate of gene silencing within a few million years of their occurrence, and much of the confusion is probably owing to the two mechanisms working simultaneously. Whole-genome duplications should be rare events followed by massive gene loss, whereas small-scale duplications and deletions are happening continuously². Originally, the *Hox* cluster was a product of tandem duplications of one member of this much larger family of homeobox transcription factor genes. In amphioxus, the single *Hox* cluster with 14

genes has an older paralogous equivalent, the *ParaHox* cluster with 3 genes¹². It should not come as a surprise that the *ParaHox* cluster has four human paralogs on four different chromosomes. For the analysis of the human genome, vertebrate paralogs and prevertebrate paralogs have to be distinguished, which was not yet possible with *D. melanogaster* and *C. elegans* as outgroups¹.

Animal-specific gene clusters

The human *HOX* clusters are in regions that also contain many other animal-specific duplicated genes, such as those encoding receptor tyrosine kinases of the EGF receptor family or fibrillar collagens (see Figure). Interestingly, the immediate neighbors of the human *ParaHox* genes are members of the PDGF and VEGF receptor tyrosine kinases, and the genes encoding the basement-membrane collagens are in the same chromosomal regions. Genes encoding the receptor tyrosine kinases duplicated early in animal evolution into subfamilies such as EGF, VEGF or PDGF receptor families, with additional duplications during the vertebrate radiation¹³. Collagens are one of the oldest recognized characteristics of animals, and the presence of fibrillar and basement-membrane collagens even in the most simple animals, the sponges, supports the idea that a genome duplication could have been important at the origin of the animal radiation.

Gene or genome duplications *per se* will not lead to an increase in the complexity of life forms. *Xenopus laevis*, a recognized tetraploid, is not much different from *Xenopus tropicalis*, a closely related diploid frog. Most fish might be ancient tetraploids¹⁴; salmon and some sturgeon have undergone even further polyploidizations. The study of animal genome evolution could eventually profit from the findings of plant research, where it is well accepted that 70% of species are of polyploid origin¹⁵. This is also true in the light of recent developments in the field of gene silencing by mechanisms

referred to as 'homology effects', which were also noticed first in plants. An increase in mutation rate may obviate the silencing of paralogs with identical stretches of sequences. The amphioxus MHC region suggests that one of the human paralogous regions is more like the ancestral version, whereas the other three seem to have mutated more quickly and lost more individual genes².

One down, one to go

The reconstruction of phylogenies and the dating of divergence times are methods dependent on statistical reproducibility. However, the few major transitions in evolution, such as the origin of multicellular animals or the radiation of vertebrates, are very peculiar events that probably occurred during catastrophic times in the history of life. As such, they might have followed rules that are far different than those established during times that are accessible to us. Nonetheless, we can see that there are four paralogous regions in the human genome with an age-distribution compatible with vertebrate-specific duplications. Obviously, a single duplication is not sufficient to explain four paralogs or the older paralogies already present in amphioxus, indicating that we will hear more about this. □

1. Mclysaght, A., Hokamp, K. & Wolfe, K.H. *Nature Genet.* **31**, 200–204 (2002).
2. Gu, X., Wang, Y. & Gu, J. *Nature Genet.* **31**, 205–209 (2002).
3. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. & Inoko, H. *Nature Genet.* **31**, 100–105 (2002).
4. Holland, P.W., Garcia-Fernandez, J., Williams, N.A. & Sidow, A. *Development suppl.* 125–133 (1994).
5. Spring, J. *FEBS Lett.* **400**, 2–8 (1997).
6. Wolfe, K.H. *Nature Rev. Genet.* **2**, 333–341 (2001).
7. Friedman, R. & Hughes, A.L. *Genome Res.* **11**, 1842–1847 (2001).
8. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
9. Venter, J.C. *et al.* *Science* **291**, 1304–1351 (2001).
10. Seo, H.C. *et al.* *Science* **294**, 2506 (2001).
11. Lynch, M. & Conery, J.S. *Science* **290**, 1151–1155 (2000).
12. Brooke, N.M., Garcia-Fernandez, J. & Holland, P.W. *Nature* **392**, 920–922 (1998).
13. Miyata, T. & Suga, H. *Bioessays* **23**, 1018–1027 (2001).
14. Taylor, J.S., Van de Peer, Y., Braasch, I. & Meyer, A. *Phil. Trans. R. Soc. Lond. B* **356**, 1661–1679 (2001).
15. Matzke, M.A., Scheid, O.M. & Matzke, A.J. *Bioessays* **21**, 761–767 (1999).