Evolution of the Vertebrate ParaHox Clusters

Sonja J. Prohaska [a,*] and Peter F. Stadler [a,b]

[a] Bioinformatics Group, Dept. of Computer Science, University of Leipzig, Germany
[b] Institute for Theoretical Chemistry, University of Vienna, Austria.

ABSTRACT

Motivation: The ParaHox cluster contains three *Hox*-related homeobox genes. The evolution of this sister of the *Hox*-gene clusters has been studied extensively in metazoans with a focus on its early evolution. Its fate within the vertebrate lineage, and in particular following the teleost-specific genome duplication, however, has not received much attention.

Results: Three of the four human ParaHox loci are linked with PDGFR family tyrosine kinases. We demonstrate that these loci arose duplications in an ancestral vertebrate and trace the subsequent histroy of gene losses. Surprisingly, teleost fishes have not expanded their ParaHox repertoire following the teleost-specific genome duplication, while duplicates of the associated tyrosine kinases have survived, supporting the hypothesis of a large scale duplication followed by extensive gene loss.

Contact: sonja@bioinf.uni-leipzig.de

ParaHox genes form a group of homeodomain transcription factors that is closely related to the well-known Hox genes. Just as the Hox gene clusters, ParaHox genes form clusters and are linked to receptor tyrosine kinases (Spring, 2002). ParaHox genes are crucial in the development of gut, neutral tube, and brain (Brooke *et al.*, 1998). Like their more famous sister group, the ParaHox genes are arranged in a single tightly linked cluster in the cephalochordate *Branchiostoma floridae* (Brooke *et al.*, 1998), while the human genome contains ParaHox genes on four different chromosomes, see e.g. the recent review by Garcia-Fernàndez (2005) and the references therein. The 2R hypothesis (Holland *et al.*, 1994) explains this fact by two rounds of whole-genome duplications early in the vertebrate lineage.

All vertebrates for which genomic information is currently available share the same system of three *Cdx*, two *Gsx*, and a single *Xlox* gene, with occasional missing genes in different species that are more likely problems with the unfinished genome assemblies rather than true losses. In tetrapoda, three of these genes (*Gsh-1* (gsx), *Ipf-1* (xlox), and *Cdx-2* (cdx)) are arranged in an uninterrupted cluster, while the other three genes are located each on a different chromosome. Three of the four mammalian loci are adjacent to members of the PDGFR tyrosine kinase family. A comparison with the chicken and frog genomes shows that the tyrosine kinases of the B-cluster (on the mammalian X chromosomes) have been lost, probably in conjunction with the origin of the X-chromosome. In contrast, there is no intact ParaHox cluster in teleost fishes, but we observe that most fish ParaHox genes have neighbors from the PDGFR family. Phylogenetic trees of the *Gsh* and *Cdx* proteins show that the redundancy resulting from the duplication was resolved completely: no paralogs from this most recent duplication have survived.

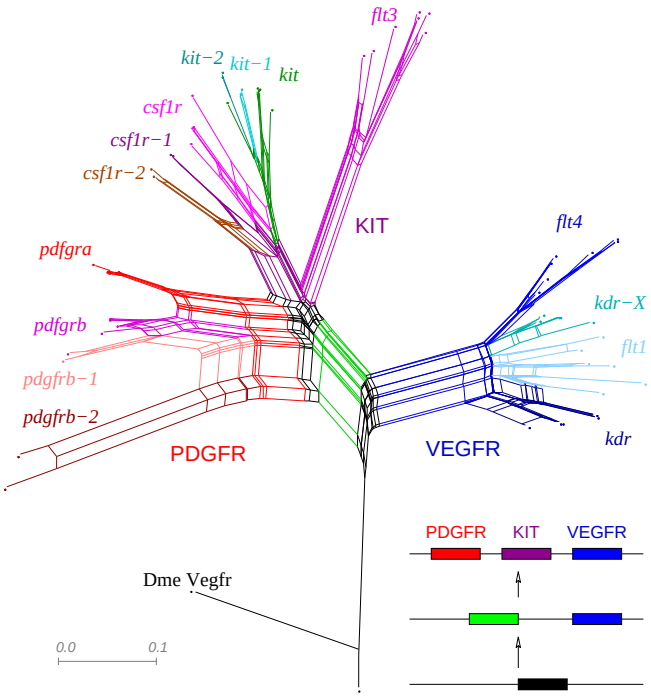

Fig. 1. Neighbor net of PDGFR family receptor tyrosine kinases calculated using `SplitsTree` with the ML protein distance option. The phylogenetic reconstruction implies two tandem-duplication events preceeding the vertebrate-specific duplications (inset). The 2R duplications then produced 2 paralog groups for PDGFR, three paralogs (*csf1r*, *kit*, and *flt3*) of the KIT group and four paralog groups of the VEGFR group: *kdr*, *flt1*, *flt4*, *kdr-X*, the latter having been deleted in mammals. The teleost-specific duplication has left two first-order paralogs of *pdgfrb*, *csf1r*, and *kit*. See electronic supplement for details.

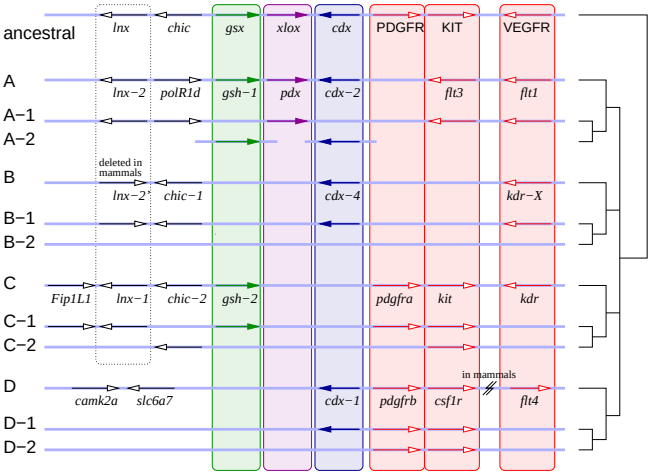

Fig. 2. Structure and evolution of the vertebrate ParaHox loci.

The PDGFR family of receptor tyrosine kinases consists of three subgroups, *pdgfr*, *kit*, and *vegfr*, each of which in turn contains multiple paralogs in vertebrate genomes that are involved in the regulation of gastrointestinal and vascular development (Shibuya, 2002; Gu & Gu, 2003). These genes are of particular medical interest since they are frequently activated in cancer (Dibb *et al.*, 2004). In Fig. 1 we summarize a phylogenetic analysis of PDGFR family tyrosine kinases from Human, Dog, Chick, Xenopus, Fugu, Tetraodon, and Zebrafish genomes. The data imply a history, where the three subgroups arose by local (tandem) duplications from a single invertebrate ancestor prior to the genome duplications.

The region with conserved linkage extends beyond the ParaHox and tyrosine kinase genes: *lnx*, for instance is a RING finger and PDZ domain containing protein that interacts with the cell fate determinant *Numb* (Nie *et al.*, 2002). It appears in three of the four tetrapod clusters. Strikingly, almost all genes at the ParaHox loci are important developmental regulators.

Combining the phylogenetic analysis of the tyrosine kinase family and the ParaHox gene families with a close inspection of the synteny information across all available genomes (see Supplement) implies that the ancestral gnathostome contained four clusters of ParaHox and tyrosine kinase genes that arose by duplications of a single locus (Fig. 2). This is consistent with the 2R hypothesis. In telosts, a further round of duplications has taken place. With the exception of the B-cluster, remnants of the ancestral gene sequence can be found in at least some of the investigated teleost species (zebrafish, fugu, tetraodon, and medaka), see also (Williams *et al.*, 2002). Fig. 2 summarizes our data, which support the hypothesis of a single duplication event in the teleost lineage.

The case study presented here highlights a much more general problem. The annotation of orthologous genes in multi-gene families is often inaccurate based on a pairwise sequence comparisons and clustering alone (Remm *et al.*, 2001; Li *et al.*, 2003), in particular when, as was the case here, in most species only automatically generated gene models are available. Multi-gene famlies, such as the receptor tyrosine kinases, often are subject to strong directional selection in individual paralogs in individual lineages, a fact that makes orthology annotation from distance data even harder. The systematic combination of gene phylogenies with synteny information in multiple species appears to be a promising remedy, even, as the example of the ParaHox clusters shows, in the presence of additional ancient duplications and extensive gene loss.

Acknowledgements. This work was supported by the German *DFG* Bioinformatics Initiative BIZ-6/1-2.

REFERENCES

Brooke, N. M., Garcia-Fernàndez, J. & Holland, P. W. H. (1998). The *ParaHox* gene cluster is an evolutionary sister of the *Hox* gene cluster. *Nature*, **392**, 920–922.

Dibb, N. J., Dilworth, S. M. & Mol, C. D. (2004). Switching on kinases: oncogenic activation of BRAF and the PDGFR family. *Nat. Rev. Cancer*, **4**, 718–727.

Garcia-Fernàndez, J. (2005). Hox, ParaHox, ProtoHox: facts and guesses. *Heredity*, **94**, 145–152.

Gu, J. & Gu, X. (2003). Natural history and functional divergence of protein tyrosine kinases. *Gene*, **317**, 49–57.

Holland, P. W. H., Garcia-Fernàndez, J., Williams, N. A. & Sidow, A. (1994). Gene duplication and the origins of vertebrate development. *Development*, **(Suppl.)**, 125–133.

Li, L., Stoeckert Jr., C. J. & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.

Nie, J., McGill, M. A., Dermer, M., Dho, S. E., Wolting, C. D. & McGlade, C. J. (2002). Lnx functions as a ring type E3 ubiquitin ligase that targets the cell fate determinant numb for ubiquitin-dependent degradation. *EMBO J.*, **21**, 93–102.

Remm, M., Storm, C. E. V. & Sonnhammer, E. L. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

Shibuya, M. (2002). Vascular endothelial growth factor receptor family genes: when did the three genes phylogenetically segregate? *Biol. Chem.*, **383**, 1573–1579.

Spring, J. (2002). Genome duplication strikes back. *Nat. Genet.*, **31**, 128–129.

Williams, H., Brenner, S. B. & Venkatesh, B. (2002). Identification and analysis of additional copies of the platelet-derived growth factor receptor and colony stimulating factor 1 receptor genes in fugu. *Gene*, **295**, 255–264.