An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement

Ignacio Maeso, Manuel Irimia, Juan J. Tena, Esther González-Pérez, David Tran, Vydianathan Ravi, Byrappa Venkatesh, Sonsoles Campuzano, José Luis Gómez-Skarmeta, and Jordi Garcia-Fernàndez

Supplementary Discussion S1

Evolutionary scenario for the origin of the amphioxus Irx cluster

The presence of four paralogous arrays of CNRs, together with data from phylogenetic analysis (Irimia et al 2008) and CNR similarity (Supplementary Fig. 4) indicating that A duplicate is more closely related to C and that B is more similar to D, provide a more detailed picture of the evolutionary origin of the amphioxus *Irx* cluster (Supplementary Fig. S5). First, a single tandem duplication of *Irx*, *Sowah* and their associated CNRs generated the ancestors of *IrxA–C* and *IrxB-D* (and their neighbouring *Sowah* genes). Then, this pair duplicated again, giving rise to the four *Irx* genes (*A*, *B*, *C* and *D*) with their respective *Sowah* genes and CNRs. Finally, *IrxC* and its downstream CNRs translocated from its ancestral position next to *SowahC* to a more upstream location, whereas *IrxD* and its downstream CNRs were completely lost. The fact that *IrxC* is now exactly in the place where the lost *IrxD* would be expected to lie suggests that these two events (translocation and loss) are probably related (perhaps through non-homologous recombination between *IrxC* and *IrxD*). In parallel to these processes, nearly all the exonic sequences from three of the *Sowah* duplicates were erased whereas most of their intronic CNRs have been maintained (Fig. 2).

Supplementary Discussion S2

Ancestral organization of the vertebrate clusters and orthology of *Irx7*

Sowah1 is linked to a solitary and divergent member of the *Irx* family in teleosts, *Irx7*, which is the only paralog that has not been confidently assigned to any of the four *Irx* complexes yet, despite extensive efforts and debate (Lecaudey et al. 2001; Itoh et al. 2002; Dildrop and Rüther 2004; Feijóo et al. 2004; Lecaudey et al. 2005). Although *Irx7* seems to be more related to *Irx1/3* (Itoh et al. 2002; Lecaudey et al. 2005), only

Irx2b and *Irx6b* seem to be missing out of the 12 expected *Irx* genes generated after the teleost-specific round of WGD (Dildrop and Rüther 2004; Feijóo et al. 2004) (Fig. 6b). Our results on the relative orientation of *Sowah1* and *Irx7* may help to solve this question. *Sowah2* is located next to *Irx5b*, in a head to head orientation. This is probably the ancestral orientation of the *Sowah-Irx* block in bilaterians since it has been maintained in almost all lineages in at least one *Iroquois* gene (Supplementary Fig. S1).

Thus, the original configuration of the vertebrate *Irx-Sowah* cluster before the WGDs was likely: >*Irx1/3*>, <*Sowah*<, >*Irx2/5*>, >*Irx4/6*>. Only *Irx1/3* would have changed the ancestral orientation to *Sowah*, a change that is also supported by the inverted orientation of the duplicated Ultra Conserved Regions (UCRs) located near *Irx1*/*Irx3* respect the UCRs next to *Irx4*/*Irx6* (Fig. 6). If *Irx7* were one of the "missing" teleost *Irx* genes (*Irx2b* or *Irx6b*)(Dildrop and Rüther 2004; Feijóo et al. 2004), the expected orientation between *Sowah1* and *Irx7* should be head to head. Instead, they are oriented tail to head, just as expected for a *Irx1/3* gene (Fig. 6b). This, together with the higher sequence similarity observed with *Irx1* and *Irx3*, suggests an alternative hypothesis: *Irx7* would be the third copy of the *Irx1/3* paralogous group or, in other words, the only remains of a third *Irx* cluster that was still present in the last common ancestor of tetrapods and teleost fishes, and was completely loss in the former lineage.