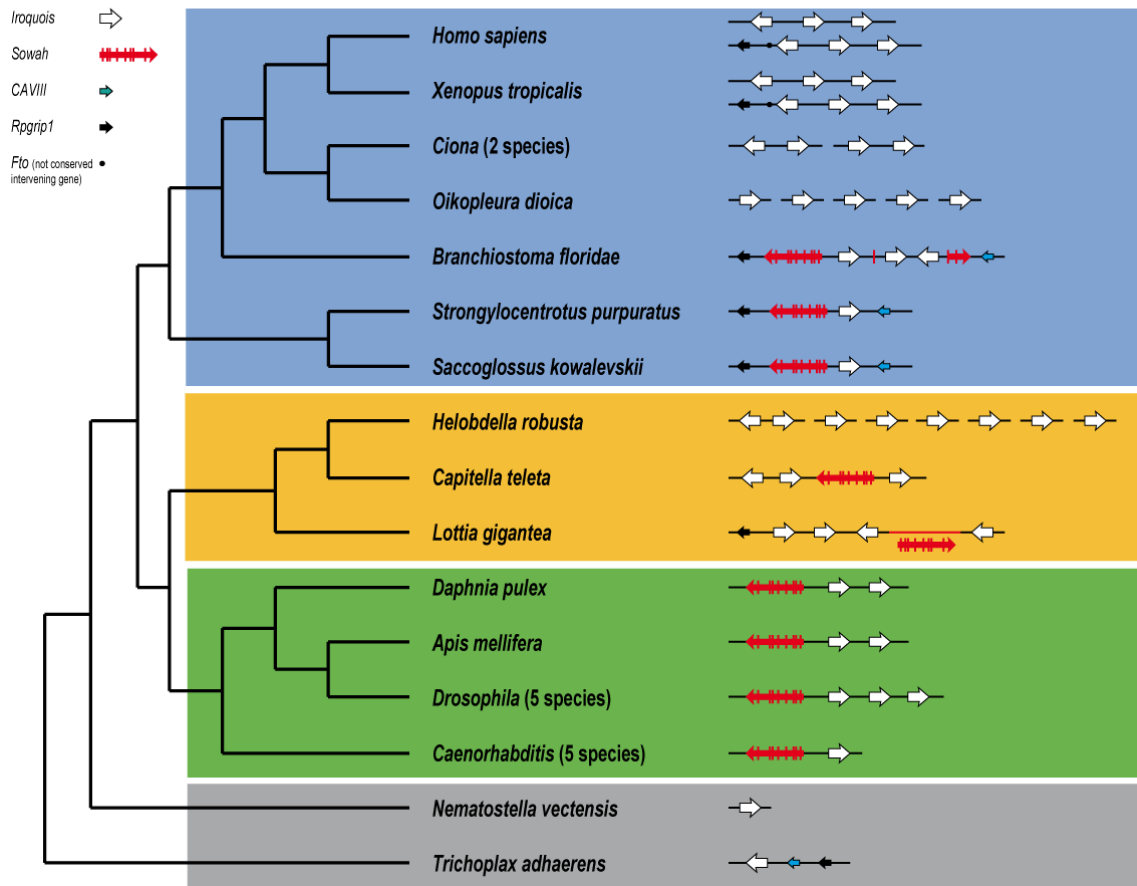


Supplementary Figures

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

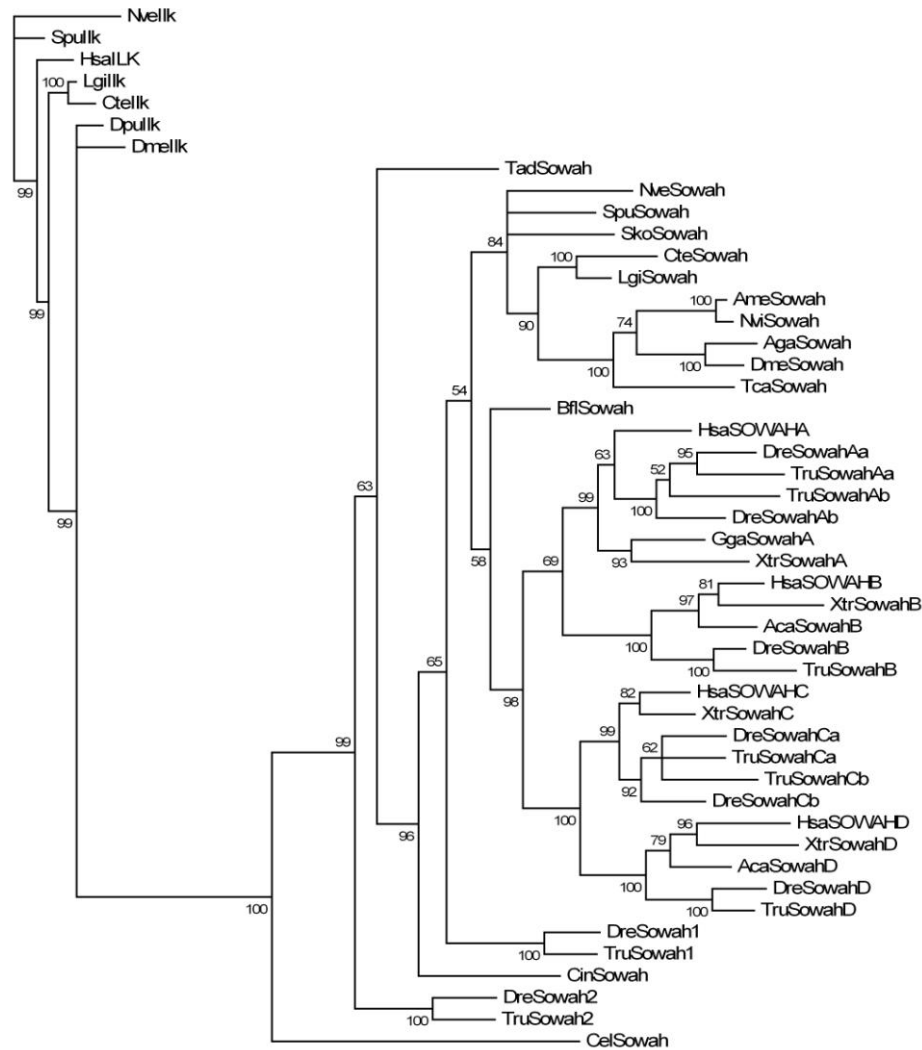
Ignacio Maeso, Manuel Irimia, Juan J. Tena, Esther González-Pérez, David Tran, Vydianathan Ravi, Byrappa Venkatesh, Sonsoles Campuzano, José Luis Gómez-Skarmeta, and Jordi Garcia-Fernández



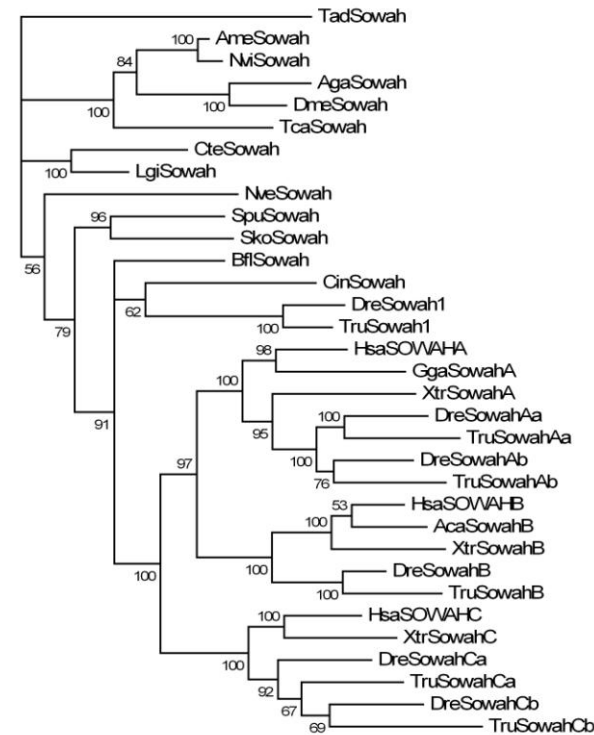
Supplementary Figure S1 – Genomic organization and conserved synteny of *Iroquois* and *Sowah* genes in bilaterians

Iroquois and *Sowah* genes are represented by white and red arrows, respectively, showing transcriptional orientation. Vertical bars in *Sowah* indicate the complex intron-exon structure of these genes. Two other genes showing conserved synteny between different metazoan groups are depicted by smaller black or blue arrows (*Rpgrip1* and *CAVIII*, respectively). A non-conserved (outside vertebrates) intervening gene (*Fto*) is indicated by solid black circles. Continuous black lines represent genomic linkage. The red line in *Lottia* indicates the unavailable genomic region where *Sowah* is expected to be. Coloured rectangles correspond to major metazoan supergroups: blue, deuterostomes; yellow, lophotrochozoans; green, ecdysozoans; grey, non-bilaterians (paraphyletic outgroup).

A



B



Supplementary Figure S2 – Bayesian phylogenetic reconstruction of the Sowah orthologs identified in different metazoan genomes.

(a) 2,300,000-generation tree under the JTT+G model generated with the Ankyrin repeat domains. Several Ilk proteins were used as outgroups. *Sowah* orthology is supported by a posterior probability of 1. (b) 3,000,000-generation tree under the JTT+I+G model of selected *Sowah* members using the Ankyrin domains plus a conserved region in the N-term of the proteins. The increase in the number of positions included and the exclusion of the pseudogenized *Sowah2* and other divergent sequences allowed the recovery of monophyly of chordate members ($p > 0.9$). Aca, *Anolis carolinensis*; Aga, *Anopheles gambiae*; Ame, *Apis mellifera*; Bfi, *Branchiostoma floridae*; Cel, *Caenorhabditis elegans*; Cin, *Ciona intestinalis*; Cte, *Capitella teleta*; Dme, *Drosophila melanogaster*; Dre, *Danio rerio*; Gga, *Gallus gallus*; Hsa, *Homo sapiens*; Lgi, *Lottia gigantea*; Nve, *Nematostella vectensis*; Nvi, *Nasonia vitripennis*; Sko, *Saccoglossus kowalevskii*; Spu, *Strongylocentrotus purpuratus*; Tad, *Trichoplax adhaerens*; Tca, *Tribolium castaneum*; Tru, *Takifugu rubripes*; Xtr, *Xenopus tropicalis*.

NveSowah MS-ANEPLTLALVREFILKNGSRVRNHDLI THFRTQLNDPVRNGKIKVRADFKEYVHQLTTVQTIDGQ-EKYFVLKPKP | STHSTPKMGR
SpuSowah MADAEFEFSADAVQKFLVSNNGKATNVEVVGHFVSKYLNDEFEFVILNRQKFKDYINEVATVVKSEKGQ-EKFLVLKNEF | PATTATSATG
SkoSowah MA--EDEFSEAGIRDFLVQNGGKTNEHLVKRFKIYFNDADNKAL----KFKEYVNAVAVSKTEDGQ-EKYLVLKPKY | ?????????
BflSowah MADY--EFSEFDGVKQFIADRGGRVRNVELVTHFRQHLNDPARKAIAEARELFKDFVNNVAVIRQEEGQ-EKFLVLKPKY | SELPRSDSSS
CinSowah MT---EISVESVCEFITSHGGKVKKNITLVSNFKKLLNDPARDIAARQQFKCECVNTVAVVKTEGQ-EKYVQLKPKY | YTVSVERARQ
DreSowah7 M--SEISEASLLDYIRSSGGKVKNSDLSTKYKHFINHSDQLRAIKYREEFKLIIDRIAVVKSNGQ-EKYLILKPKY | GALSILKAEHQ
CteSowah MG-DSL-FNQDVRDFMISKGGKVRNHDLVSHFSSFLNDPQRVIANRQKFKTFVNTLSSVKLDESQGEKVLVLKRFK | TPISVKEKRAK
LgiSowah MA-ATC-FTLEAVKEFMINNGGRVTNQQLVTHFAFLNDPVSQVITREKFKDYVNELATIKLEEGQ-SKILVLKPKH | SVISVKDRAK
AmeSowah MATP-SELSLEEIRQYLLENGGTARNHVDVVKHFKFLTDPETRVTEARNQFKEYVNTLATIKNEEGQ-EKYLVLKPKY | EQLSFRERMQ
NviSowah MASP-SELSLEEIRKYLLDNGGSARNQDLVKHFKFLTDPESTRVTEARNRKFKEYINTLATIKTEEGQ-EKYLVLKPKF | EQLSVVERMQ
TcaSowah MAS--SDLTLEQIYNYFKENGNVVKNRELVHHFKHYLTDEASKELDARVFKFKFVNTLAQTKE-G*DEKFLILKTRY | TPVSVKERLE
AgaSowah MDTEPKELSMEEIRKFMVLNNYKLTNHTLVKHFRSFLTNKDTQDEARRRFKEYVNNKIASVKHE-G*NEKYLVLRRY | NKL SVKEKMM
DmeSowah MEL-PKELSVAEIRNYMLANECKVTNHALVKHFKFLTHFQSNTEARRKFKTYVTLSTIKNE-N*NQKFLILRKY | NPISVKEATR

1

2

NveSowah KISQPKRGQDGLERN-----SQALSEDV---FEDPDIDGQ-DDLMNIPTIQLO-----DAIEKEWLLTTARGNRAHIMCLLEQEPL
SpuSowah DEAAAPQPDIDIPAPVEIQVCFIEQRYSHSLGG-????????????????????LQ-----EPLEKQWMLSAKGNVPLVIGLLDQEPG
SkoSowah ???LQ-----SETEKQWMLAASAGETPTPLSKILLIKEPH
BflSowah RMETGSSRSDDTVSIRS-----DTASISEK-----GQ-EAE-DD*-GYGSNVSSVAKQ-----DVLEKEWMLKAAVGNVSALRRLSLQDPN
CinSowah NFQSKPKNGEDTTSVS-----EKSMDSR-----GQ-ENEQDD*-EYDRYPVSVSLQ-----SANEKAWMLKSCQDMVEMKLLDNDST
DreSowah7 SEE-ISSSDREEDL-----DKDSGSK-----SSESEQDEE*SGGSLGSTVALQ-----DPLEKEWIFSAASGKLCHLTQLLKQDSS
CteSowah HLNRMTS-DVDVQKPL-----AKHIK-----KQ-----*EYDSSIEVISLQ-----SAADREWITAAAADYHPMNKLLVSNPQ
LgiSowah HLNKIDS-ESDILLPN-----QFKKKS-----GQ-REKSGG*-EDDDTHFISYQ-----GAEKEWMLTCKADYHEMNRLLSKDSS
AmeSowah RFNRMAS-ETDLQGRPNGTITPTKKRSDK-----GQ-----A*-DEDDASVASQ-----LDGKSREWLVRAAQGYQALAKIAAEPR
NviSowah RFNRMAS-ETDLPARPNASIPSVKKRTDK-----GQ-----N*-EDDDRASVASQ-----LMDGKSREWLVRAAQGYQALAKIAAEPR
TcaSowah KFNRMASMEDELSPRQ-----AKSAEKKKEK-----GQ-----S*-EEDDGASVTSLLQLLEPKKCMEWVTASKGDIQELLKLAQDEPR
AgaSowah KFNRMAS-EEEAIPIS-----PIGKKKPKD-----GQ-----A*-EETLTNENLM-----QHPNAKEWILAAKSDYQALAKLSTHEPN
DmeSowah KFNRMAS-EEEAIIIS-----PPAKKKPEK-----GQ-----LI*-EEKDSPEVTL*-AHPKAKEWILMAKANYQELAKMASEYPE

3

4

NveSowah LAKRKDOFTSGQYTOTLALHWAAKHGRDDVAMVAKAGA----DVNSKTHGCGGYTPLHLAAMHGNDRVIKVLVDDFKIADINARDFSGR
SpuSowah LATKKDOFIT????*SALHWAAGKHGKEMIRLVVDRI-----DVNLR????*GYTPLHLVAAAMHNKAIQMQLVSEYHIADINKRDFSGR
SkoSowah LAKRKDOFTS????*TALHWAAGKHGKEMILKLFADRPD-----LDINVKSHGCGGYTPLHLIASMYGNEAAIQMLLSDFKIANVDARYSGK
BflSowah LASKKDOFVSGQYTOTLALHWAAGKHGNGEIVAMADAGL-----DVNLRSHGCGGYTPLHLAAMYGHDIKVLIDAYDADVTIRDYSGR
CinSowah LLEKKDOFSGQYTOTLALHWAAGKHGKNDMLHMLDLKV-----EVNVKSNQG*GYTPLHLAAMNGHEDIISLLQ-HDIANPKMRDYSGR
DreSowah7 LANKKDOFTSV*F*TOALHWAAGKHGKEDMTALVEAGA-----DINTKAGHCGGYTPLHLAALHGHRIHVDLLVRTYGAKENLRDYSGR
CteSowah LAKRKDOFIT????*TOPLHWACKQKTEVVVKLICAYPE-----VNVNARSFQCGGYTPLHLAAMQGHEEVIEMLVQAYGLADSNLRDYSGR
LgiSowah LAKIRDOFTNGQYTOTLALHWAAGKHGKGEVILKLVGKPG-----VNIQRSYGCGGYTPLHLAAMGHEDIIELVQNYKADSNIRDYSGR
AmeSowah LTRLKDOFTSGQYTOTLALHWAAGKHGDENIVKLIAGTYKDYIKSVNETSNQG*GYTPLHLIAMQFHDENIFNLVQVYALANQEI RDYSGK
NviSowah LARTKDOFSS????*TALHWAAGKHGDENIVKLIAGTYKDYIKSVNETSNQG*GYTPLHLIALQFDHENIFNLVQVYALANQEI RDYSGK
TcaSowah LVNRK????*TOVLHWAGKHGKNPKIIQMFAGKYK-----VDVNGKTNQG*GYTPLHLAAQFGHRDIYDMLINEYHADSIRDYSGR
AgaSowah LVKLQDOFITSGQYTOTLALHWAAGKHGSEDVVKLVAGKLL-----ADVNRATNQG*GYTALHIAQFGRNDIFELLCNVYKADRDLLDWSGK
DmeSowah LVKLQDOFATGQYTOTLALHWAAGKHGNEVDVKLIAGTYK-----ADVNRATNQG*GYTPLHLATQFGRNDIFELLVNVYKIANRDIMDWSGN

5

6

6

6

7

7

7

8

NveSowah KPRQVAK-DTLTIEA---QRIRLEHVLT---LSD---SSSESQYSSQG---QOTPKYGLNSIV-----|TPSAALMGYGGGSS
SpuSowah KPRHYLQ-SSSSNYI---KQIMMLGQ---STA---EVIQPHVLMSLAEPKQEKHRHSKIGSLRSS---|---FRHSFRIWGS-A-E
SkoSowah KPKQHMK-KNCANYI---QQIMLQGS---TKS---YGIQRAR-SNSMEKKRGS-GL-----|---FFNKPLKWSGA-D
BflSowah RPKQRLRPSSNTNVDEVVQQLKLVHAT---KQ---FKWA-RGK---TQQRKS-VP-AIFQR---|GLSRDITFAQGGKTPP
CinSowah LPIHYLK-LDMSLPL---QKLLFGSEQMIVKG---FAAGLR-PH---ETDSRKTNRLLSSAISG---|GLT---SVGALMSRSSE
DreSowah7 LPYYLYL-LRDTHSEHDLS*---EVSECHSLTPAGERRNRKISSLF-----|---QSMKRWGSA-E
CteSowah KAKQYLR-NSASTRA---QQLLQNRPPSAASA---GAAPFRRTGSYSSKIR-STGHFGIAAGTAVRNS | KPSPKQSPQTRWKQS-
LgiSowah KAKQYLR-NSASRT---QQLMLLSRRLAAGSVKQQLSDDSFIRT---TLRKSNAKAISSLIQATQSKV | RSKSTLLRSSWGSSE
AmeSowah KARQYLVSEAAVSQDTFRKIKARKK---HAE---KDLGFLRIGSLNVRVKRTTEAFSFLGVA---TS | SSNHEKIHKSWGSA-D
NviSowah KPRQYVSQKAAVSQDTMRKIKARKK---HAE---KDLGFLRIGSLNVRVKRTTEAFNQLGVTTTS | NNSSEKHLKTSWGSA-D
TcaSowah LPAYYLSQDKQKIRKDNIRKIKGRKK---PTD---KDLGFLRIGSLNVRVKRTTEAFSNFLGVGSGSV | ETTSEKVKHKGWSA-D
AgaSowah KALEYQK-QMTSVSESTFSKIKARKK---HSE---KDSGFLRIGSLNVRVKRTTEAFSNFLGVGTNTQ | TLFIDKLKHSWGSA-D
DmeSowah KPLDYSR-QRSSVSASTCSKIKARKK---HAT---EKDLGFLRIGSLNVRVKRTTEAFSNFLGVGNGSG | RNIYDGVHKSWSGA-D

9

10

NveSowah EAPLSV-----AQPPAESGDKHRDRTS---SLGK-----FLHKKKSKDHRQPMPTPYPSNSSAG-----
SpuSowah NLEDKNHKRDRTPPVTPPPSPKNRPKTSHSVNSQSLLQ-----MPPPSGYTVPNR-GRKLRTDKDR-----
SkoSowah SLRST-----TPPESPTFRNLTSDKDS---MPPPK-F-VPSKKGKKRKESMKENV-----
BflSowah STPETT-----PPVLRKDSPPMLRRASA-SKV-----MPPPK-FPAPKTK-KDKKDKKSR-----
CinSowah DYSKLD-----EPNRPGSGS---FRNK---MS---RKKHCKSTKERPTITLVTD-----
DreSowah7 ELA-----PVPEERT-----*---VPHQLMLPAFRPRKFSR-----
CteSowah H-----SGSGQ-ELT-MPPPRDLPVRHRRQSS-RDNKENHS-----
LgiSowah HVDDQD-----PTPPGSSHSPTTSRKGSGF---SDSSRDKTLMPPPSAPTQRRNKSAS-RDNLQSTQSASAI-----QTAS
AmeSowah NVQL-----QDKM-----MPPPKYAPIKKRRS---RAAQDFSSSLREH-----
NviSowah NLPA-----QDKM-----MPPPKYAPIKKRRS---RRATDFVSSRSNNQQQQQQCQQTAS
TcaSowah NVNQE-----Q-NI-----NPGPKINTVKKK-S---KRP---Y-SSGMN-----S
AgaSowah NIQK-----EDSA-----MPPPKYGPVKRRRQ---KRDIEY-----DPSNDHFS-----
DmeSowah NIPHRS-----EDL-----MPPPKAVEYISKRNKSKRSS-YASNTTDSRPDSI-CSSNSS

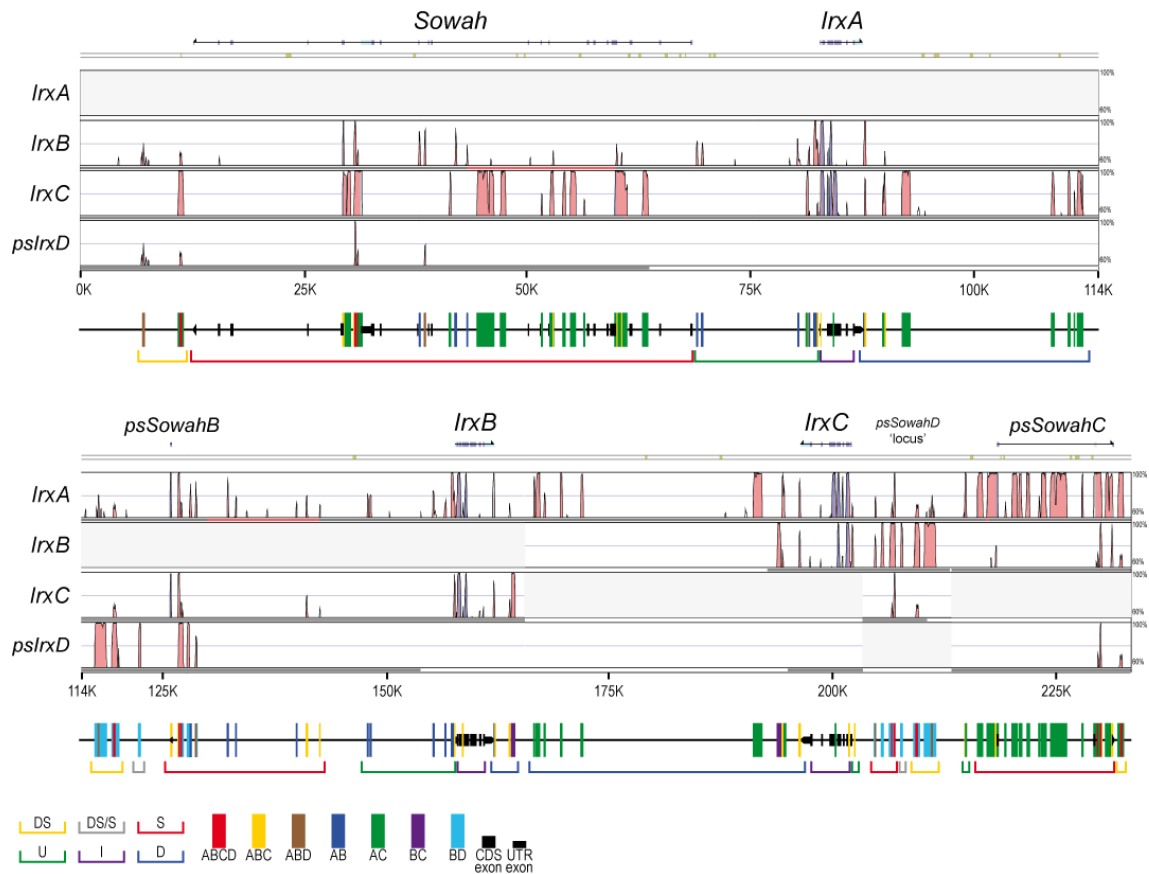
11

NveSowah -----2---LLIGSPIELQRMDEERRENELDMKRTQSDPNLLHDEV-----
SpuSowah -----2---TENMRKSESNPDMIVETPTTFV-----
SkoSowah -----2---GVELRRTSEEPDLGLS-----
BflSowah -----2---GL---RKAESDPDVTGGFA-----
CinSowah -----2---ESDS---GFGSIRSSVSLPLSASTPNLQSRSTP |
DreSowah7 -----*-----
CteSowah -----2SELGAGGLSGSETDLTVGMDSAAASYPNISVQKASYV-----
LgiSowah -----*DPRLSGARSGSDSALSCKNTYID-----
AmeSowah QPTTPLLQKVSRSSSHQOMHRRPTSTTAVSSAA-----SRTQSSNDSDSDTACGFDASAWGSAQL-----
NviSowah QPSTPTLLKGSATAGRRPASVSALTGGLLAQTLSSGISTGSI*QKDHQQQQNDSDSDGACGFDASAWGSA-----
TcaSowah TPGTTPRQMSR-----NFSSFHT-----NDSDSDSAAAGFDSQWKH-----
AgaSowah -----QSVPTTPSQIRAPVGLSENPDGEM-----GDSDSDTACGFDASNWRPTYI-----
DmeSowah SNLNGGYSSMPTTPNQLRAPKG-----IAASFA-----VDSDSDSACGFDSTWSVNCRGSSSSNPSQS-----

12

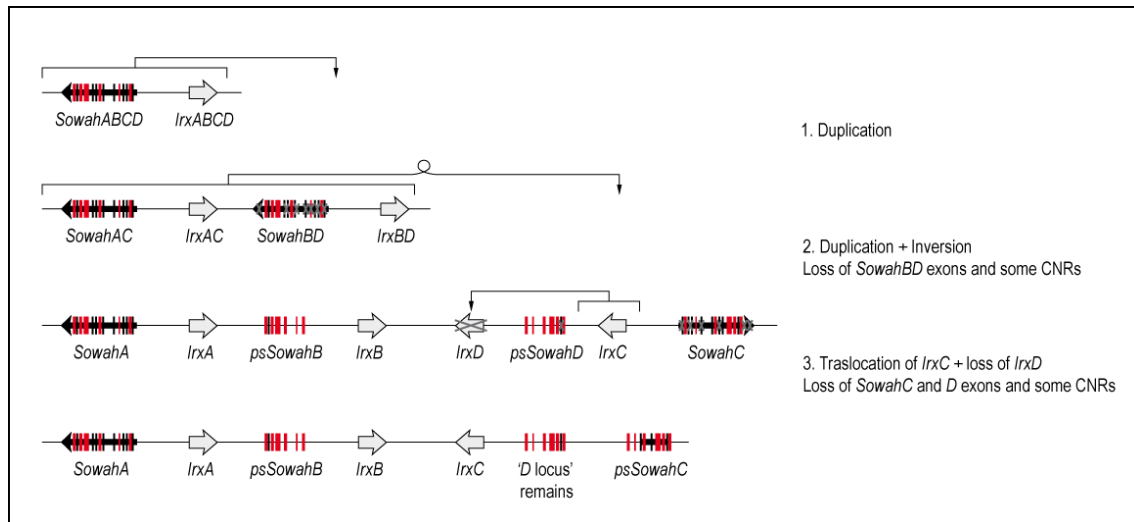
Supplementary Figure S3 - Conservation of intron-exon structures in *Sowah* genes

Sequence alignment for *Sowah* genes of representative bilaterians and one non-bilaterian showing conserved intron positions and phases. Highlighted positions correspond to 50% of similar aminoacid types as generated by BioEdit. Phase 0 is indicated in red, 1 in blue and 2 in green. Non-conserved introns are shown with asterisks. Ancestral introns are numbered in yellow below the alignment. Introns flanked by microexons (except the largest one of 12 nucleotides) are numbered as the previous intron. Vertical bars mark the limits of divergent regions not shown in the alignment and question marks regions that could not be annotated.



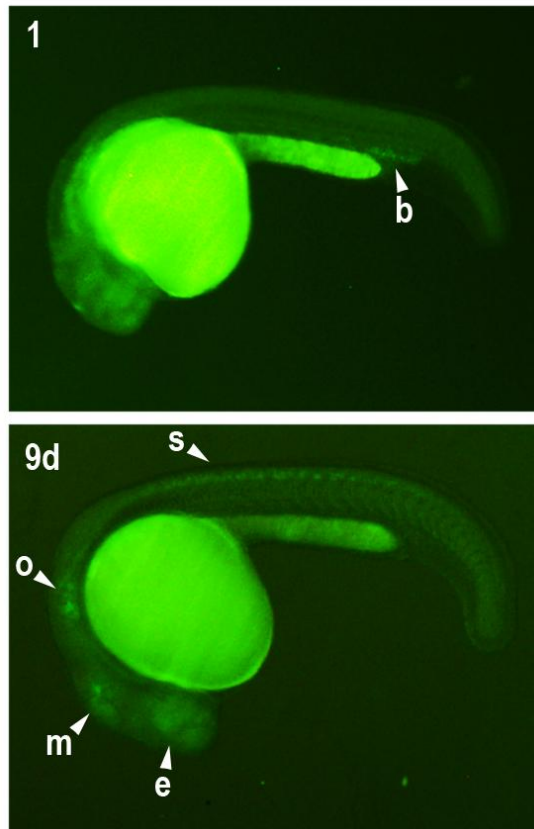
Supplementary Figure S4 – Detailed internal organization of the *Irx-Sowah* complex in *B. floridae*

VISTA plot of the alignments between each of the three *Irx* genes (plus a fourth region corresponding to a putative *IrxD* locus lost during amphioxus evolution) and their respective surrounding non-coding regions, including *Sowah* for the case of *IrxA*. Coloured peaks (blue, coding; turquoise, UTR; pink, non-coding) indicate regions of at least 100 bp and 70% similarity. High copy number elements (such as repeats and mobile elements) are masked, their presence is indicated by khaki segments above the VISTA plot. In contrast to Fig. 2, vertical bars of different colours represent the different conserved repeated blocks, indicating the different *Irx-Sowah* loci in which they are present (i. e. A, B, and D; B and D; etc). Coloured brackets enclose CNRs to show their respective location to *Sowah* and *Irx* loci (DS, Downstream *Sowah*; S, within *Sowah*; U, Upstream *Irx*; I, Introns of *Irx*; D, Downstream *Irx*. DS/S indicate elements of uncertain identity (DS or S); colours for each type are the same as in Fig. 2). Black block arrows and rectangles indicate the exon sequences of the *Irx* genes, of *Sowah* and of the remains of *Sowah* duplicates.



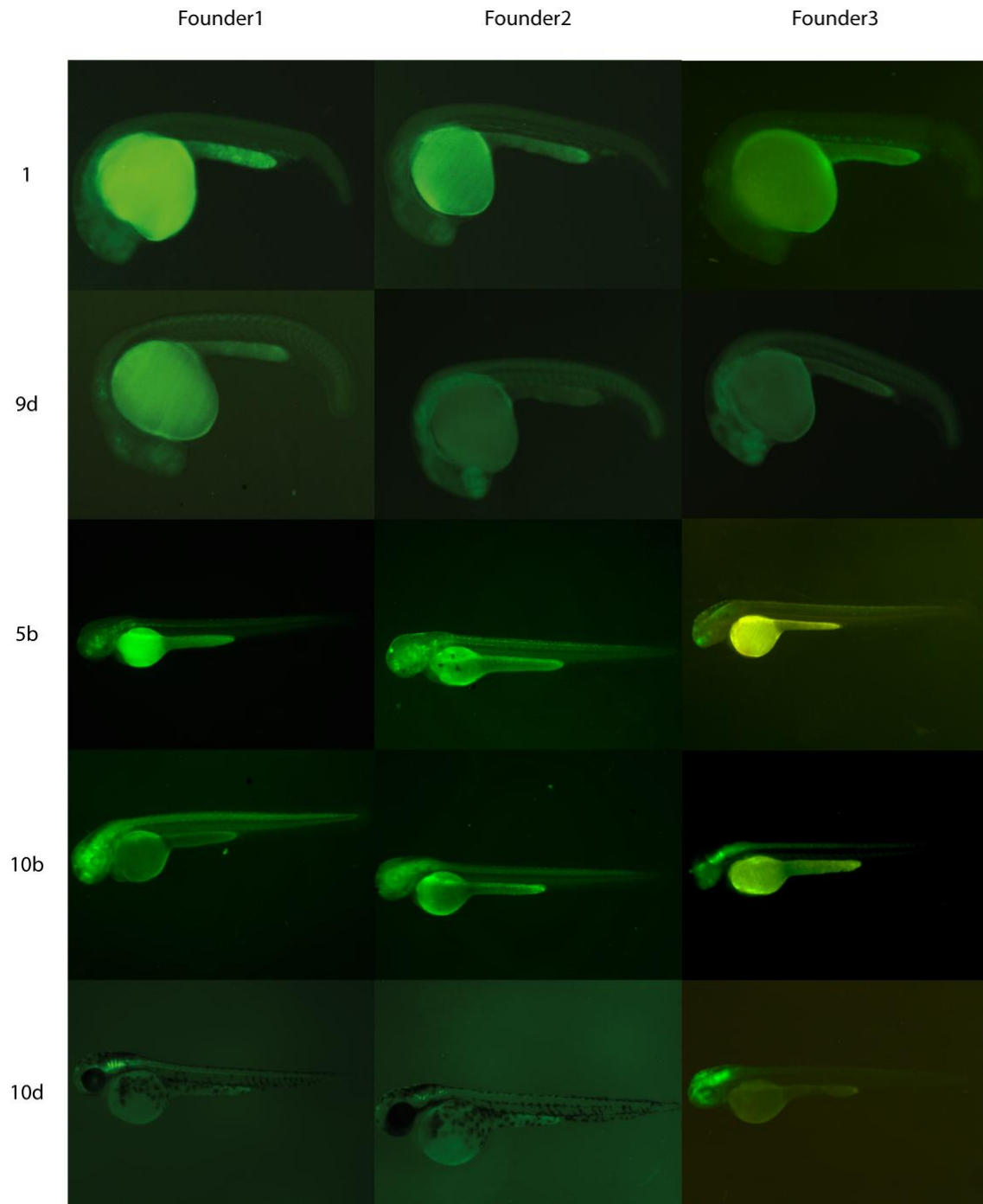
Supplementary Figure S5 - Putative evolutionary scenario for the origin of the *B. floridae* *Irx* cluster

First, ancestral *IrxABCD* and *SowahABCD* genes and their respective conserved non-coding sequences duplicated in tandem, giving rise to an ancestral *IrxBD* gene, an ancestral *IrxAC* gene and their respective *SowahAC* and *SowahBD* copies (1) (*IrxA* and *IrxC*, as well as the associated *Sowah* loci, are more related to one another than to *IrxB*). Some of the CNRs and almost all the exons of the duplicated *SowahBD* were subsequently lost. Then, the two *Irx-Sowah* (or *psSowah*) pairs duplicated and reversed their orientation, originating *IrxA*, *IrxB*, *IrxC* and *IrxD* and *SowahA*, *psSowahB*, *psSowahD* and *SowahC* (2). Finally, *IrxC* translocated and *IrxD* was lost (probably due to the insertion of *IrxC*), and some of the conserved blocks, almost all the exonic sequences of *SowahC* and all the exon remains of *psSowahD* were lost (3). *Sowah*, black arrows; *Irx*, grey arrows. Red bars represent CNRs within *Sowah* introns. Grey crosses indicate losses. For simplicity, only schematic intron-exon structures are indicated for *Sowah*.

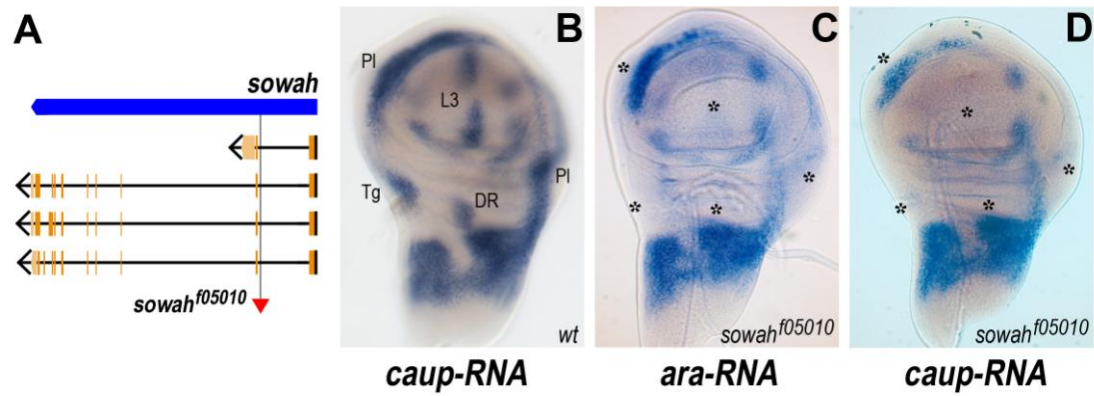


Supplementary Figure S6 - Transcriptional enhancer activity of *B. floridae* sequences from the *Irx-Sowah* complex

GFP expression driven by the CNRs 1a and 9d at 48hpf in side views. In 1a expression is detected in the blood islands. In 9d, expression is detected in the midbrain, the eye and the otic placode. Anterior is to the right. b, blood islands; e, eye; m, midbrain; o, otic placode; s, spinal cord.



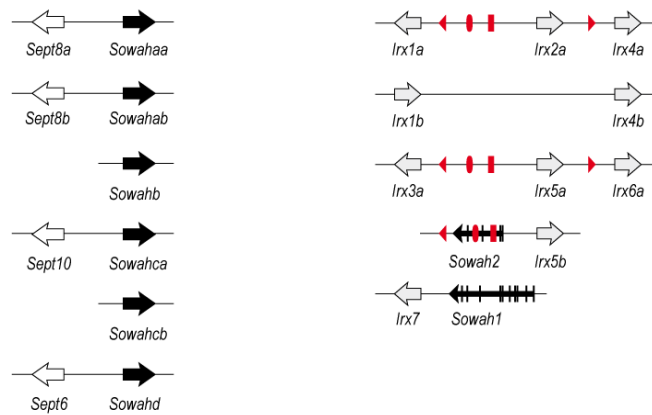
Supplementary Figure S7 – Zebrafish embryos corresponding to three independent transgenic lines generated for each of the *B. floridae* elements.



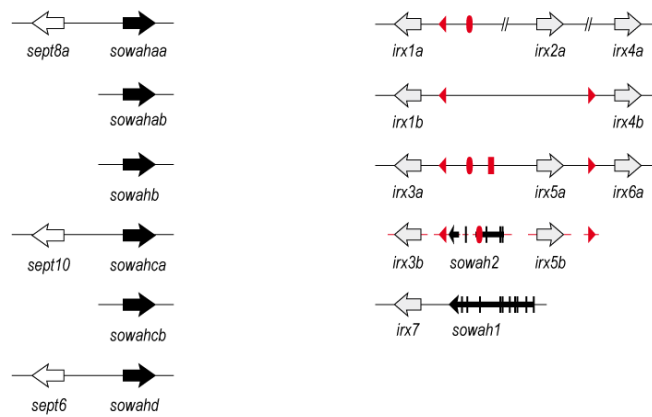
Supplementary Figure S8 - Expression of *ara* and *caup* in *sowah^{f05010}* imaginal wing discs

(a) Physical map of the *sowah* locus indicating the position of the *sowah^{f05010}* insertion (triangle). Under the genes in blue, transcripts are shown as black arrows. Exons are shown in orange with the protein coding regions coloured darker. (b-d) Note the disappearance of *ara* and *caup* expression from the vein L3 and dorsal radius (DR) domains and its strongly reduction in pleura (PI) and tegula (Tg) regions in *sowah^{f05010}* wing discs (asterisks, c and d), compared to a wild type disc (d).

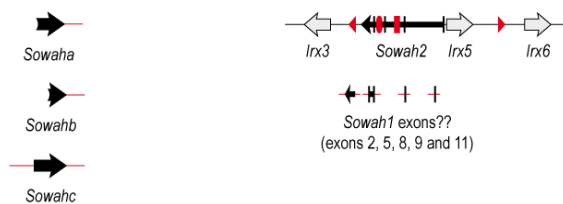
A *Takifugu rubripes*



B *Danio rerio*

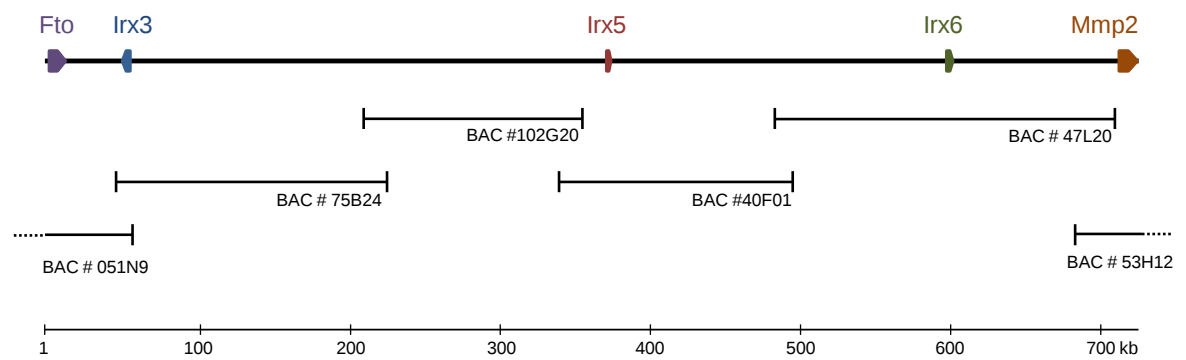


C *Callorhinchus milii*



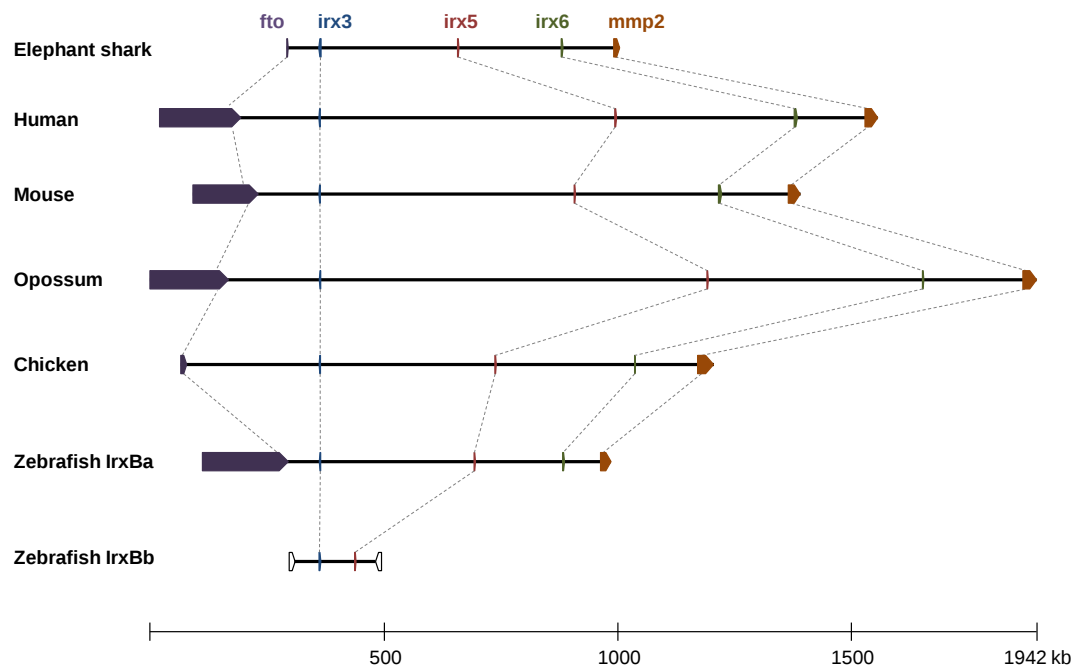
Supplementary Figure S9 – Genomic organization of *Sowah* and *Irx* genes in teleosts and elephant shark

Schematic representation of the genomic organization of *Sowah* (black arrows), *septin* (white arrows) and *Irx* (grey arrows) genes in fugu (a), zebrafish (b) and elephant shark (c). Red figures represent CNRs within *Irx* clusters as in Fig. 6. For simplicity, only schematic intron-exon structures are indicated for *Sowah1-2* genes. Broken arrows and red lines indicate that the fragmented nature of the assemblies impeded a complete annotation of the genes or synteny analyses, respectively.



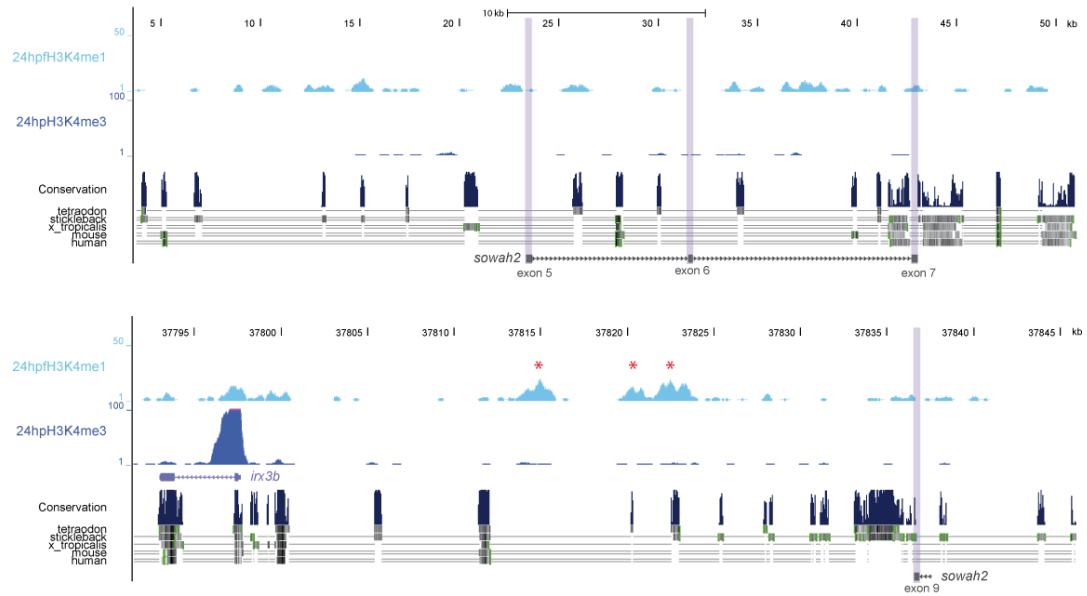
Supplementary Figure S10. - *IrxB* locus in the elephant shark.

BAC clones used for generating the sequence are shown.



Supplementary Figure S11. - Comparison of *IrxB* locus in elephant shark and other vertebrates

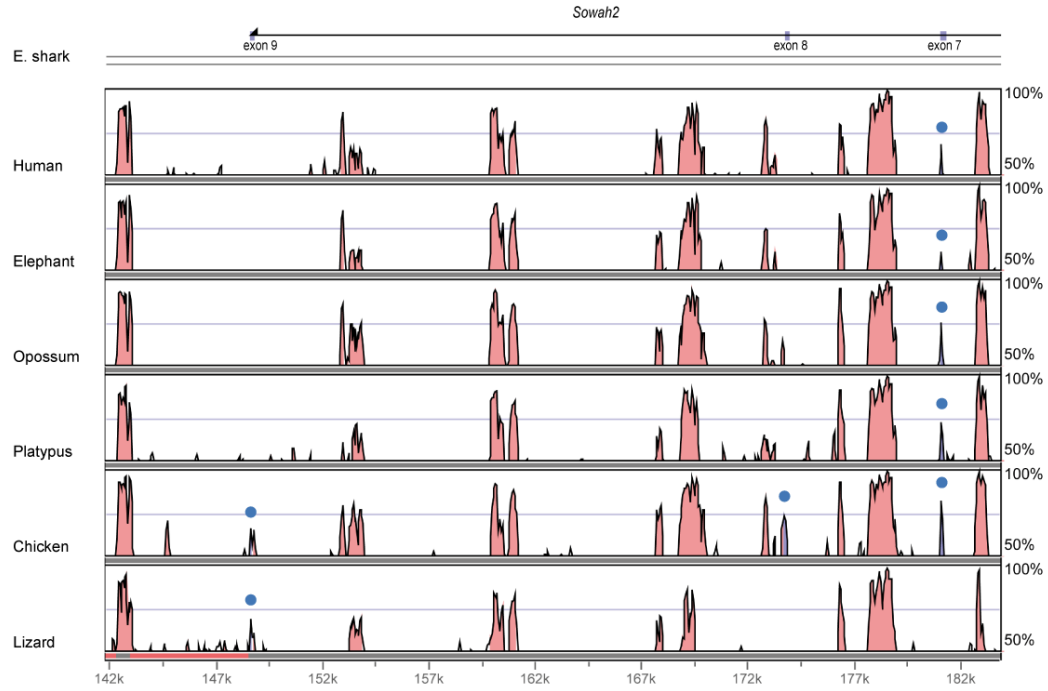
A



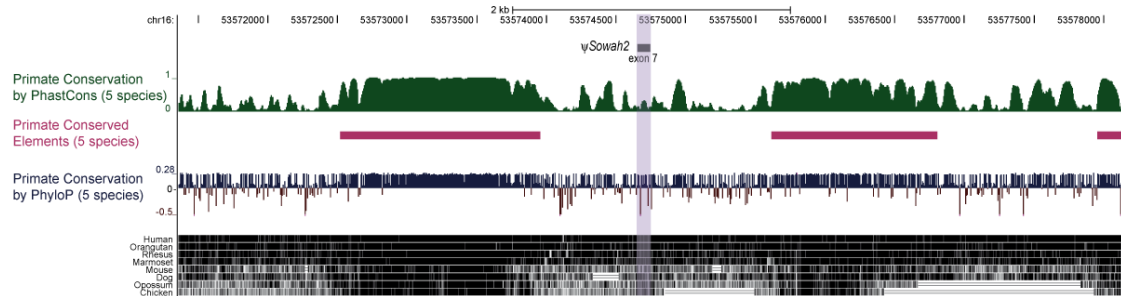
B

	exon 5 (e. shark, zebrafish, fugu, stickleback)	exon 7 (e. shark, zebrafish, fugu, medaka)	exon 8 (e. shark, zebrafish, fugu, medaka)	exon 9 (e. shark, zebrafish, fugu, medaka)
ω	0.05085	0.11630	0.02824	0.00560

C

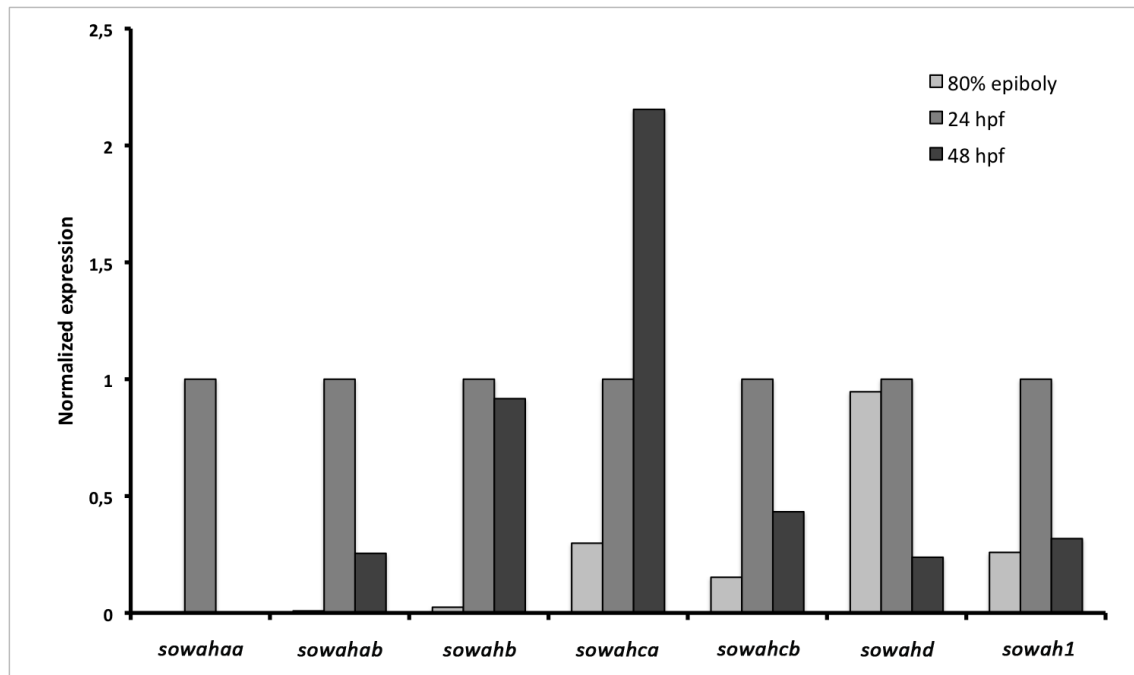


D



Supplementary Figure S12. - Functional status of *Sowah2* in different vertebrates.

(A) *sowah2* exons are not associated with enhancers in zebrafish embryos. UCSC genome browser view of H3K4me1 (enhancers) and H3K4me3 (promoter) tracks generated for 24hpf zebrafish embryos (Aday, 2011). *sowah2* exons (purple bars) are not associated with the deposition of the enhancer-associated epigenetic mark H3K4me1. Red asterisks mark several *irx3b* potential enhancers. The *irx3b* promoter is positive for H3K4me3 and H3K4me1 marks. (B) ω estimations for *Sowah2* exons in teleosts and elephant sharks. In brackets, list of the species included in the alignments used to generate the estimations. (C) VISTA plots of the *Sowah2* region of several tetrapod species, using elephant shark as a reference sequence. Coloured peaks (blue, coding; pink, non-coding) indicate regions of at least 60 bp and 60% similarity. Red dots indicate the location of *Sowah2* pseudogenic exons. We used less stringent parameters to detect *M. domestica* exon 9 and therefore is not represented in this plot. (D) UCSC browser conservation tracks in the *Sowah2* region of the chimpanzee genome, showing low evolutionary constraint on exon 7 pseudogenic remnants



Supplementary Figure S13. Developmental expression of *Rsowah* and *Sowah1* in zebrafish embryos

RT-qPCRs of *sowah1* and *rsowah* paralogs. Total RNA was isolated from 20 embryos of 24 and 48 hpf stages, and 30 embryos of 80% of epiboly stage. cDNA was synthesized from this RNA by retro-transcription, and the relative amount of the different gene products was measured by real time-PCR. All the data were normalized using the housekeeping gene *elongation factor 1-alpha*. The 24 hpf stage sample was taken as reference.