OHNOLOGS v2: A comprehensive resource for the genes retained from whole genome duplication in vertebrates

Param Priya Singh [1,2,*], and Hervé Isambert [1,*]

[1] Institut Curie, Research Center, CNRS UMR168, PSL Research University, 26 rue d'Ulm, 75005, Paris, France
[2] Present Address: Stanford University, Department of Genetics, 300 Pasture Drive, Stanford, CA 94305, USA.

To whom correspondence should be addressed. Tel: +33 1 56 24 64 74; Email: herve.isambert@curie.fr.

To whom correspondence should be addressed. Tel: +1 408 680 9845; Email: param@stanford.edu.

ABSTRACT

All vertebrates including human have evolved from an ancestor that underwent two rounds of whole genome duplication (2R-WGD). In addition, teleost fish underwent an additional third round of genome duplication (3R-WGD). The genes retained from these genome duplications, so-called ohnologs, have been instrumental in the evolution of vertebrate complexity, developmental patterns and susceptibility to genetic diseases. However, the identification of vertebrate ohnologs has been challenging, due to lineage specific genome rearrangements since 2R- and 3R-WGD. We have previously identified vertebrate ohnologs using a novel synteny comparison across multiple genomes. Here, we refine and apply this approach on 27 vertebrate genomes to identify ohnologs from both 2R- and 3R-WGD, while taking into account the phylogenetically biased sampling of available species. We assemble vertebrate ohnolog pairs and families in an expanded OHNOLOGS v2 database, which also includes non-protein coding RNA genes. We find that teleost fish have retained most 2R-WGD ohnologs common to amniotes, which have also retained significantly more ohnologs from 3R-WGD, whereas a higher rate of 2R-WGD ohnolog loss is observed in sauropsids compared to mammals and fish. OHNOLOGS v2 should allow deeper evolutionary genomic analysis of the impact of WGD on vertebrates and can be freely accessed at http://ohnologs.curie.fr.

INTRODUCTION

Gene duplication provides raw material for evolution of new gene functions (1). Duplication of single genes or genomic segments is a continuous evolutionary process that creates diversity in terms of copy number variations (CNV) across individuals, and paralogs across species. In addition, dramatic evolutionary accidents corresponding to Whole Genome Duplication (WGD) have also occurred in the evolutionary past of most eukaryotic lineages including plans, fungi, and animals (2-4). For example, all extant vertebrates have experienced two rounds of whole genome duplications (2R-WGD) in their evolutionary past (5-8). In addition, a third round of genome duplication have also occurred in the teleost fish lineage (3R-WGD) (9-11). 2R-WGDs likely played important role in the evolution and

diversification of vertebrate specific innovations such as neural crest cells, placodes, and a complex brain (12,13). Many key genes implicated in development of these structures can be traced back to 2R-WGD. Similarly, 3R-WGD likely played an important role in the expansion of the diversity of teleost fish lineage making it the most species rich vertebrate group (14-17). Hence, the genes retained from these three WGD events have been instrumental in the evolution of vertebrates (18).

The genes originated from these ancient polyploidy (paleo-polyploidy) events are now called ohnologs after Susumu Ohno who first hypothesized the two rounds of WGD events in vertebrate ancestors (1,5,19). Ohnologs are known to have distinct evolutionary, genomic and functional properties that distinguish them from small-scale duplicates and singletons (20-23). They also show greater association with diseases and cancer than non-ohnolog genes (24-29), and have been suggested to be dosage balanced (24) , which was subsequently shown to be indirectly mediated by their high susceptibility to dominant mutations (25,27,28).

Given the specific impact WGDs have had on the evolution of vertebrates, a comprehensive database of vertebrate ohnologs is highly desirable. While there are some useful resources available for comparison of synteny across species (30-33) there is no database that reliably identifies ohnologs from both vertebrate 2R-WGDs and fish 3R-WGD. To start filling this gap, we developed in 2015, OHNOLOGS, a repository of ohnologs retained from the 2R-WGD in six amniote vertebrates (human, mouse rat, pig, dog and chicken) (33). OHNOLOGS is based on a novel comparative macro-synteny approach that reliably identifies ohnologs, despite lineage specific genome rearrangement, gene loss and small scale duplication events, by combining macro-synteny information (gene content regardless of exact order) across multiple outgroups and vertebrate genomes (33).

Here, we expand this multiple genome synteny comparison approach to 27 vertebrate species including 4 teleost fish species. We further improve the statistical confidence assessment of each ohnolog pair with a weighted quantitative confidence score (q-score) taking into account the phylogenetically biased sampling of available vertebrate species. In addition, we uncover ohnologs, including in non-protein coding RNA gene classes, from both 2R-WGD in early vertebrates (2R-ohnologs) and 3R-WGD in teleost fish (3R-ohnologs). The expanded OHNOLOGS database is the most comprehensive repository of ohnologs in vertebrates. Using the new OHNOLOGS database we show that the average 2R- and 3R-ohnolog retention rates are 25% and 18% respectively, with sauropsida showing the highest lineage specific loss of 2R-ohnologs, and teleost fish showing the highest lineage specific retention of 2R-ohnologs. We found that 2R-ohnologs are significantly more likely to also retain 3R-ohnologs, in agreement with earlier reports (34). OHNOLOGS v2 should facilitate deeper evolutionary analysis of the unique properties of ohnologs, and their lineage specific retention and loss in different vertebrates.

RESULTS

Data collection and processing

OHNOLOGS v2 includes 2R-ohnolog pairs and families in 27 vertebrates that have a chromosome level assembly with a majority of their genes anchored on chromosomes in Ensembl version 84 (35). This includes 18 mammals, 4 sauropsids (lizards and birds), 4 teleost fish and spotted gar. In addition, we also included 3R-ohnolog pairs and families in 4 teleost fish genomes. We used 5 non-vertebrate outgroups to identify 2R-ohnologs and 7 vertebrate outgroups to identify fish specific 3R-ohnologs (Figure 1 and Supplementary Table S1).

We collected genes (protein coding, micro-RNA, miscellaneous RNA, rRNA, snRNA, and snoRNA), their orthologs, paralogs and relative duplication node for all these organisms from Ensembl v84 using biomaRt (36-38). We chose these six classes of genes because they have information on orthologs and paralogs across many vertebrates, and the genes that had a lot of small-scale duplications (> 30) were excluded from analysis as they inflate the synteny calculations. Genome data for Amphioxus was obtained from JGI and amphioxus orthologs with other organisms were identified using BLASTp (8). To identify duplication time of paralogs consistently, we took the consensus timing across 7 Ensembl versions (v80 – v86).

We adapted the macro-synteny comparison approach, previously developed in (33), to identify ohnologs retained from both 2R-WGD (2R-ohnologs) and 3R-WGD (3R-ohnologs). Briefly, for each pair of outgroup and paleo-polyploid organisms, we first identified blocks of conserved macro-synteny using windows ranging from 100 to 500 genes (outgroup comparison). These macro-synteny blocks have a pattern of doubly-conserved synteny, where a window in the outgroup genome shares orthology with at least two other windows in the paleo-polyploid genome. The paralogs residing on these windows and duplicated at the time of 2R- or 3R-WGD are candidates for being 2R- or 3R-ohnologs, respectively. Similarly, we also identified syntenic windows by comparing each paleo-polyploid genome to itself (self comparison).

To refine these ohnologs further and eliminate spurious synteny patterns, we computed a quantitative score (called q-score) to assess the probability that any ohnolog pair could be identified by chance, following the approach developed in (33). In brief, all q-scores from different windows and outgroups were combined to give a global q-score for each ohnolog pair from outgroup comparison. Using multiple outgroups allowed us to identify ohnologs that may have moved to non-syntenic locations in some of the outgroup genomes. Similarly we obtained a q-score for self comparison to assess the chance of spurious association. In addition, while we used a simple geometric average of q-scores in (33), which cannot capture the gain of statistical power expected from the integration of multiple vertebrate genomes, here we developed a refined weighting scheme of species, which also takes into account the strong phylogenetically biased sampling of included species by using different weights for each vertebrate genome depending on its shared homology with other included genomes (see Supplementary Methods for details, Tables S2 & S3).

Using both self and outgroup averaged q-scores, we generated 3 sets of ohnologs (corresponding to strict, intermediate and relaxed criteria), and combined them into ohnolog families. Finally, we compiled both the 2R- and 3R-ohnolog pairs and ohnolog families for each organism in the interactive OHNOLOGS v2 database using Apache, CGI, Perl, Bootstrap and jQuery.

Navigating the OHNOLOGS database

The home page lists all the organisms that are included in OHNOLOGS for 2R and 3R-WGD along with an introduction on ohnologs and WGDs. The search page allows a user to search for a gene symbol, Ensembl Id GO term or any keyword (Figure 2A). The search page also allows one to generate ohnolog families for any user-defined q-score criteria for a given organism. Upon a keyword or GO term query, all matching genes will be displayed along with their ohnolog status (Figure 2B). If a queried gene is an ohnolog, its ohnolog family will be displayed on the result page (for both 2R and 3R WGD for teleost fish) (Figure 2C and 2D). We show families for our strict q-score filter, and display the intermediate and relaxed families only if additional ohnologs are identified upon relaxing the q-score filter. The result page also includes links to pair page that has all ohnolog pairs that went into constructing that family (Figure 2E). The family result pages also links to the orthologous genes and ohnolog families in other vertebrates, to study the conservation of ohnolog families in other vertebrates.

The ohnolog pairs and families for our three pre-defined q-score filters can be explored and downloaded from the Browse/Download pages (Figure 2F). We link the genes on the browse pages to external databases including Ensembl, NCBI gene, GeneCards (for human), MGI (for mouse) and ZFIN (for zebrafish). The details of our approach, family descriptions and more details on q-score have also been included on the help page.

Summary of the contents of the OHNOLOGS database

Using the expanded OHNOLOGS database we assessed the retention and loss of ohnologs across different vertebrates. We found that vertebrates have retained on average 25% 2R-ohnologs, which include two rounds of WGD, with teleost fish having subsequently retained on average 18% 3R-ohnologs (intermediate criteria) (Figure 3A, 3B and Supplementary Table S4). Sauropsids have retained the lowest numbers of 2R-ohnolog families, while teleost fish have retained the highest numbers of 2R-ohnolog families (Figure 3A). Yet, interestingly, we observe that more compact genomes, such as turkey and tetraodon, which typically contain also fewer genes, have retained about the same numbers of ohnologs than other birds or fish, respectively (Supplementary Table S4). This strong conservation of ohnologs across diverse genome sizes is consistent with their proposed retention mechanism through purifying selection in paleo-polyploid species (25,27,28).

A vast majority of retained ohnologs consists of protein-coding genes, while non-protein coding genes represent only a small fraction of ohnologs (Supplementary Table S5). For example, in human, out of the 7358 2R-ohnolog pairs from the relaxed criteria only 28 are mi-RNA ohnolog pairs and 2 are sno-RNA ohnolog pairs (Supplementary Table S5).

Remarkably, for all analysed vertebrates the size of 2R-ohnolog families rarely exceeds 4 ohnologs (Figure 3C), as expected for two rounds of WGD events. Similarly, virtually all 3R-ohnolog families are of size 2, as they are derived from just a single WGD event (Figure 3D). These family sizes also suggest a low rate of small-scale duplications and genome rearrangements following both 2R and 3R-WGD.

Next, the database can be used to analyse the branch-specific loss and retention of ohnologs. For instance, we found that 1,316 out of 2,373 ohnolog families with relaxed confidence criteria in human had an identical size in nearly all the 18 mammals (*i.e.* corresponding to a variance over mean size ratio lower than 0.1 across all 18 mammals). Then, out of these 1,316 conserved 2R-ohnolog families in mammals, 702 have the same size in teleost fish, including 396 families which also share the same size in sauropsids while the remaining 306 families correspond mainly to additional 2R-ohnolog losses in sauropsids; 119 families are larger in teleost fish and contain fish-specific 2R-ohnologs, while 86 families are smaller in teleost fish and correspond to 29 amniota-specific, 49 mammalia-specific and only 8 sauropsida-specific retentions of 2R-ohnologs.

Finally, we assessed if teleost fish with their additional 3R-WGD event tend to retain more ohnologs from the previous 2R-WGD events. Indeed, we found that in all four analysed teleost species, 2R-ohnologs tend to retain significantly more 3R-ohnologs (Figure 3E), in agreement with earlier reports (34). The retention of 3R-ohnologs is even higher for 2R-ohnologs that have retained 3 or 4 family members, and for the 2R-ohnologs that have been retained in all the 27 vertebrates (Figure 3E). For example zebrafish 2R-ohnologs from the intermediate criteria that have been also retained in all the analysed vertebrates are twice as likely to retain their 3R-ohnologs compared to genome-wide expectation ($p = 5e\text{-}88$, Chi-square test). This suggests that the evolutionary mechanism for the expansion of specific gene families through the retention of 2R-ohnologs (25,27,28) might also explain the biased retention of 3R-ohnologs.

CONCLUSION

The updated OHNOLOGS v2 database is the most comprehensive resource available for the genes retained from both ancestral vertebrate 2R-WGDs and teleost fish specific 3R-WGD. It is based on a robust pipeline that downloads and processes datasets automatically using Ensembl, which makes it amenable to easy updates. We plan to expand and update OHNOLOGS periodically. Algorithmically, it is based on a quantitative comparative macro-synteny approach, which also takes into account the phylogenetically biased sampling of available vertebrate species. This approach assesses the confidence in each ohnolog pair and robustly identify ohnologs, despite lineage specific genome rearrangement, gene loss and small-scale duplication events. Using the datasets in OHNOLOGS we show a greater lineage specific ohnolog loss in sauropids compared to other vertebrate groups, and a high retention of 2R-ohnologs in subsequent 3R-WGD in teleost fish. In the light of the evolutionary significance of ancient WGDs and ohnologs for vertebrate evolution, the expanded and improved OHNOLOGS database should facilitate deeper comparative, evolutionary, genomic and functional analyses of the ohnolog genes in vertebrates.

AVAILABILITY

All the data and code used to construct OHNOLOGS is available at http://ohnologs.curie.fr and its associated GitHub repository at https://github.com/param-p-singh/Ohnologs-v2.0.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We acknowledge technical support from the service informatique of Institut Curie for hosting and maintaining the server infrastructure. We thank Vincent Cabeli and Marcel Ribeiro Dantas for help in updating the server.

FUNDING

PPS was supported by a PhD fellowship from Erasmus Mundus (UPMC) and La Ligue Contre le Cancer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST

Authors declare that they have no conflict of interest.

TABLE AND FIGURES LEGENDS

Figure 1. A schematic phylogeny (not scaled) of the organisms in the OHNOLOGS v2 database. Vertebrates analysed for 2R-WGD are in orange, and teleost fish species analysed for 3R-WGD are underlined. Outgroup species used to identify 2R- and 3R-ohnologs have been highlighted.

Figure 2. Navigating the OHNOLOGS database.
A) Screenshot of the search page. B) Result page for a keyword search of “rat sarcoma viral oncogene” shows the matching genes in human. C) Ohnolog family page for HRAS gene in the human genome. D) From the family page, users can navigate to ortholog families in other vertebrates, e.g. zebrafish HRASA. E) Ohnolog pair page for zebrafish for NRAS gene. F) Browse/Download page for zebrafish showing both 2R and 3R-ohnolog pairs and families for all the three criteria.

Figure 3. Description of the ohnolog genes, pairs and families in the database.
A) Number of retained individual 2R-ohnolog genes, pairs and families in all the 27 vertebrates. Bars represent the numbers from the intermediate criteria. Ohnologs from strict and relaxed criteria are indicated by dots. B) Number of retained individual 3R-ohnolog genes, pairs and families in the 4 teleost fish species. Bars represent the numbers from the intermediate criteria. Ohnologs from strict and relaxed criteria are indicated by dots. C) Size of the 2R-ohnolog families from the intermediate criteria in vertebrates. Note that a vast majority of the families are of size 2, 3 or 4. D) Sizes of the 3R-

ohnolog families from the intermediate criteria in the teleost fish hardly exceed size two. E) The 2R-ohnologs are significantly more likely to retain 3R-ohnologs, compared to genome-average. The retention of 3R-ohnologs is even higher for the 2R-ohnologs that belong to family size 3 or 4, and for the 2R-ohnologs conserved in all the 27 vertebrates. All the p-values are less than 1e-41, Chi-square test. Family counts are from the intermediate criteria.

REFERENCES

1. Ohno, S. (1970) *Evolution by gene duplication*. Springer-Verlag.
2. Van de Peer, Y., Maere, S. and Meyer, A. (2009) The evolutionary significance of ancient genome duplications. *Nat Rev Genet*, **10**, 725-732.
3. Van de Peer, Y., Mizrachi, E. and Marchal, K. (2017) The evolutionary significance of polyploidy. *Nat Rev Genet*, **18**, 411-424.
4. Schwager, E.E., Sharma, P.P., Clarke, T., Leite, D.J., Wierschin, T., Pechmann, M., Akiyama-Oda, Y., Esposito, L., Bechsgaard, J., Bilde, T. *et al.* (2017) The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC biology*, **15**, 62.
5. Ohno, S., Wolf, U. and Atkin, N.B. (1968) Evolution from fish to mammals by gene duplication. *Hereditas*, **59**, 169-187.
6. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. and Inoko, H. (2002) Evidence of en bloc duplication in vertebrate genomes. *Nat Genet*, **31**, 100-105.
7. Dehal, P. and Boore, J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, **3**, e314.
8. Putnam, N.H., Butts, T., Ferrier, D.E., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.K. *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064-1071.
9. Christoffels, A., Koh, E.G., Chia, J.M., Brenner, S., Aparicio, S. and Venkatesh, B. (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol*, **21**, 1146-1151.
10. Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A. *et al.* (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946-957.
11. Meyer, A. and Van de Peer, Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, **27**, 937-945.
12. Cañestro, C. (2012) *Two rounds of whole-genome duplication: evidence and impact on the evolution of vertebrate innovations*. Springer, Berlin Heidelberg
13. Holland, L.Z. (2009) Chordate roots of the vertebrate nervous system: expanding the molecular toolkit. *Nat Rev Neurosci*, **10**, 736-746.
14. Hoegg, S., Brinkmann, H., Taylor, J.S. and Meyer, A. (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol*, **59**, 190-203.
15. Semon, M. and Wolfe, K.H. (2007) Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet*, **23**, 108-112.
16. Glasauer, S.M. and Neuhauss, S.C. (2014) Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol Genet Genomics*, **289**, 1045-1060.
17. Taylor, J.S., Braasch, I., Frickey, T., Meyer, A. and Van de Peer, Y. (2003) Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res*, **13**, 382-390.
18. Marletaz, F., Firbas, P.N., Maeso, I., Tena, J.J., Bogdanovic, O., Perry, M., Wyatt, C.D.R., de la Calle-Mustienes, E., Bertrand, S., Burguera, D. *et al.* (2018) Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature*, **564**, 64-70.
19. Wolfe, K. (2000) Robustness--it's not where you think it is. *Nat Genet*, **25**, 3-4.

20. Huminiecki, L. and Heldin, C.H. (2010) 2R and remodeling of vertebrate signal transduction engine. *BMC biology*, **8**, 146.
21. Roux, J., Liu, J. and Robinson-Rechavi, M. (2017) Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates. *Mol Biol Evol*, **34**, 2773-2791.
22. Brunet, F.G., Volff, J.N. and Schartl, M. (2016) Whole Genome Duplications Shaped the Receptor Tyrosine Kinase Repertoire of Jawed Vertebrates. *Genome Biol Evol*, **8**, 1600-1613.
23. Guo, B. (2017) Complex Genes Are Preferentially Retained After Whole-Genome Duplication in Teleost Fish. *J Mol Evol*, **84**, 253-258.
24. Makino, T. and McLysaght, A. (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A*, **107**, 9270-9274.
25. Singh, P.P., Affeldt, S., Cascone, I., Selimoglu, R., Camonis, J. and Isambert, H. (2012) On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. *Cell Rep*, **2**, 1387-1398.
26. Dickerson, J.E. and Robertson, D.L. (2012) On the origins of Mendelian disease genes in man: the impact of gene duplication. *Mol Biol Evol*, **29**, 61-69.
27. Malaguti, G., Singh, P.P. and Isambert, H. (2014) On the retention of gene duplicates prone to dominant deleterious mutations. *Theor Popul Biol*, **93**, 38-51.
28. Singh, P.P., Affeldt, S., Malaguti, G. and Isambert, H. (2014) Human dominant disease genes are enriched in paralogs originating from whole genome duplication. *PLoS Comput Biol*, **10**, e1003754.
29. Tinti, M., Dissanayake, K., Synowsky, S., Albergante, L. and MacKintosh, C. (2014) Identification of 2R-ohnologue gene families displaying the same mutation-load skew in multiple cancers. *Open Biol*, **4**, 140029.
30. Catchen, J.M., Conery, J.S. and Postlethwait, J.H. (2009) Automated identification of conserved synteny after whole-genome duplication. *Genome Res*, **19**, 1497-1505.
31. Muffato, M., Louis, A., Poisnel, C.E. and Roest Crollius, H. (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, **26**, 1119-1121.
32. Jandzik, D., Garnett, A.T., Square, T.A., Cattell, M.V., Yu, J.K. and Medeiros, D.M. (2015) Evolution of the new vertebrate head by co-option of an ancient chordate skeletal tissue. *Nature*, **518**, 534-537.
33. Singh, P.P., Arora, J. and Isambert, H. (2015) Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Comput Biol*, **11**, e1004394.
34. Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noel, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A. *et al.* (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*, **5**, 3657.
35. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C.G. *et al.* (2018) Ensembl 2018. *Nucleic acids research*, **46**, D754-D761.
36. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439-3440.
37. Durinck, S., Spellman, P.T., Birney, E. and Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, **4**, 1184-1191.
38. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M., Amode, R., Brent, S. *et al.* (2016) Ensembl comparative genomics resources. *Database : the journal of biological databases and curation*, **2016**.

Figure 1

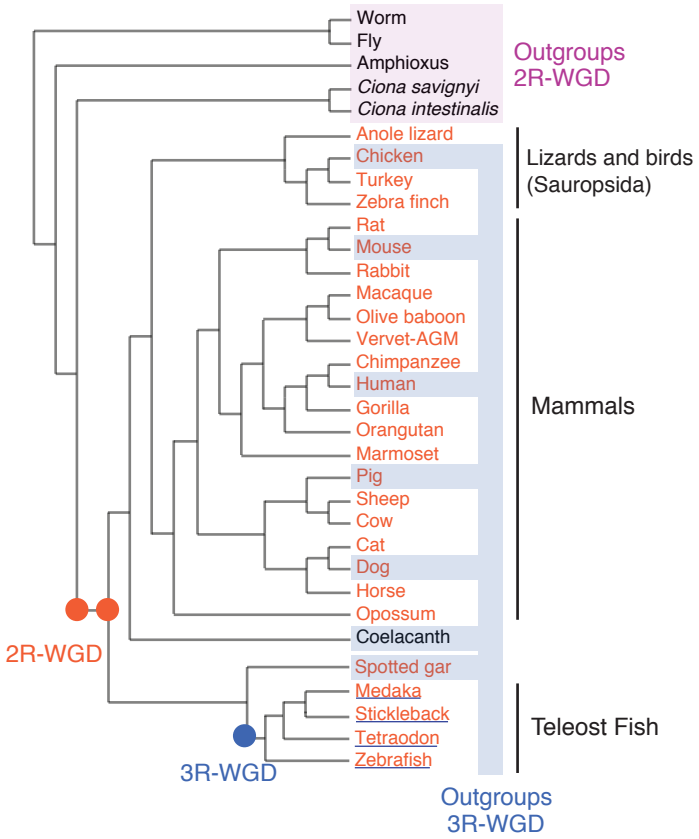
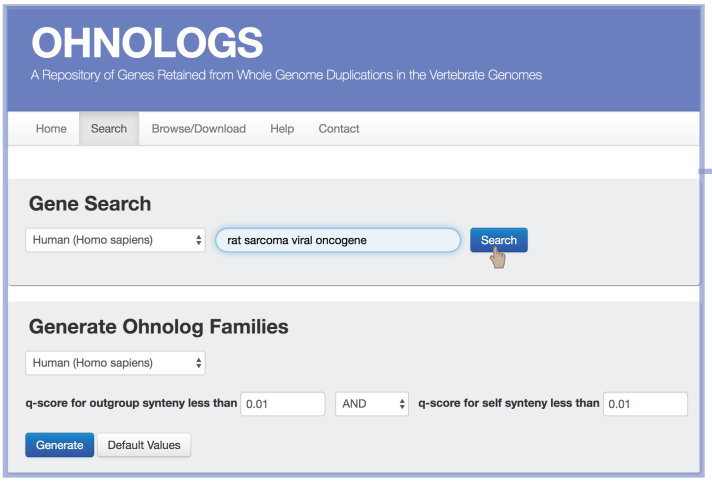
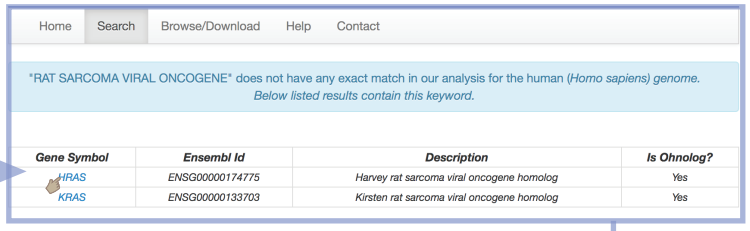

Worm
Fly
Amphioxus
Ciona savignyi
Ciona intestinalis
Outgroups
2R-WGD
Anole lizard
Chicken
Turkey
Zebra finch
Lizards and birds
(Sauropsida)
Rat
Mouse
Rabbit
Macaque
Olive baboon
Vervet-AGM
Chimpanzee
Human
Gorilla
Orangutan
Marmoset
Pig
Sheep
Cow
Cat
Dog
Horse
Opossum
Mammals
Coelacanth
2R-WGD
Spotted gar
Medaka
Stickleback
Tetraodon
Zebrafish
Teleost Fish
3R-WGD
Outgroups
3R-WGD

Figure 2

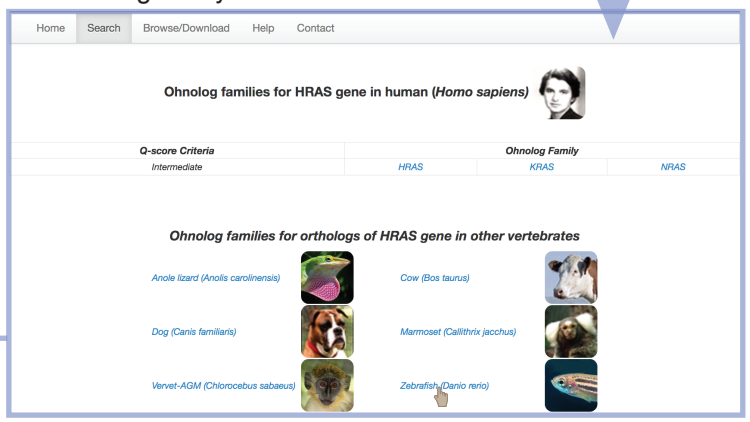
A. Search page on OHNOLOGS.



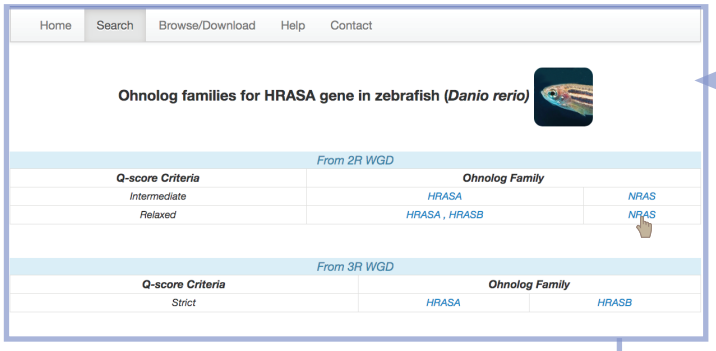
B. Search results for a keyword search.



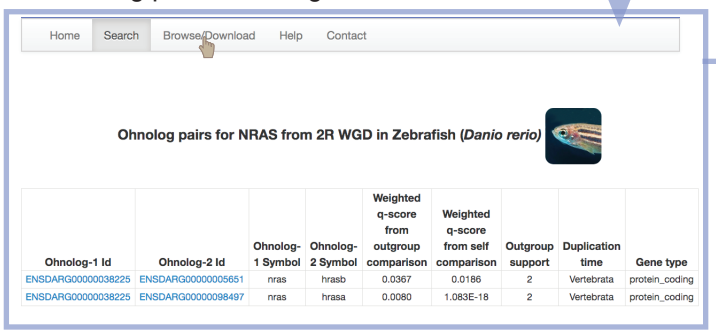
C. Ohnolog family for HRAS in human.



D. Ohnolog family for HRASA in zebrafish.



E. Ohnolog pairs including NRAS in zebrafish.



F. An example Browse/Download page for zebrafish.

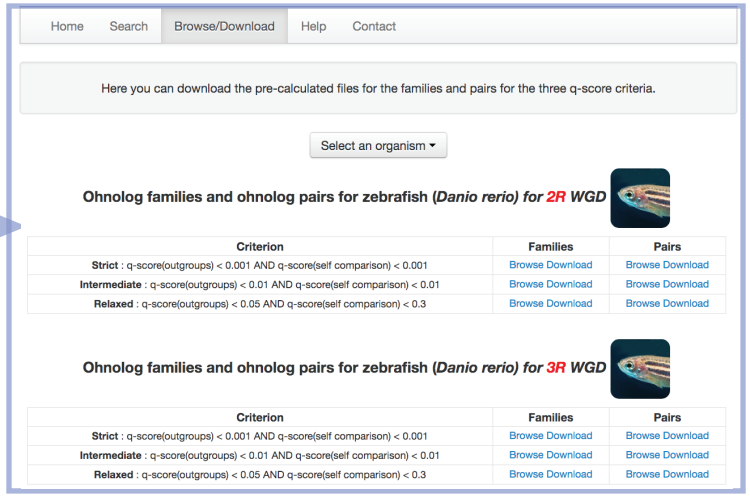
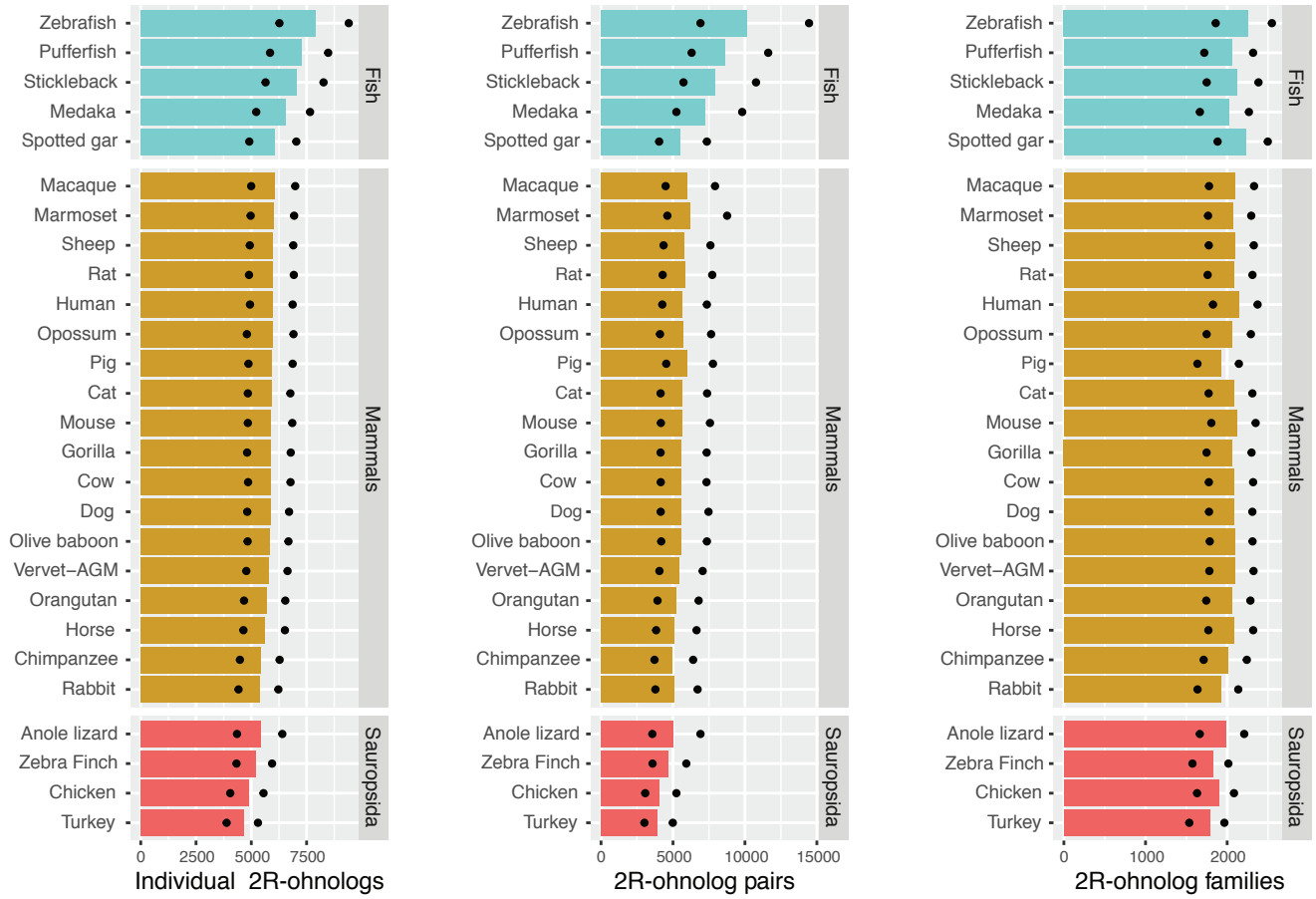
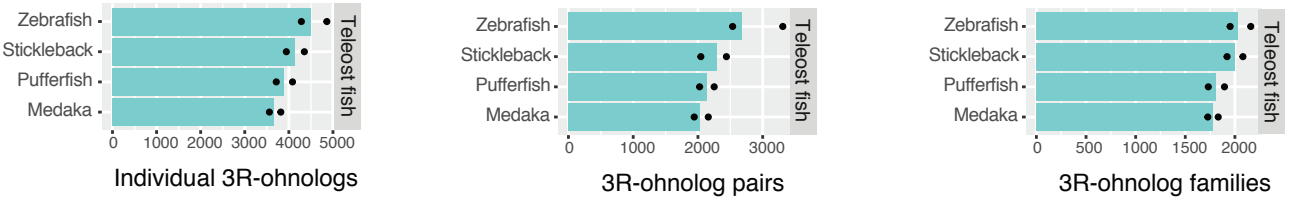


Figure 3 A. Individual 2R-ohnologs, pairs and families in vertebrates

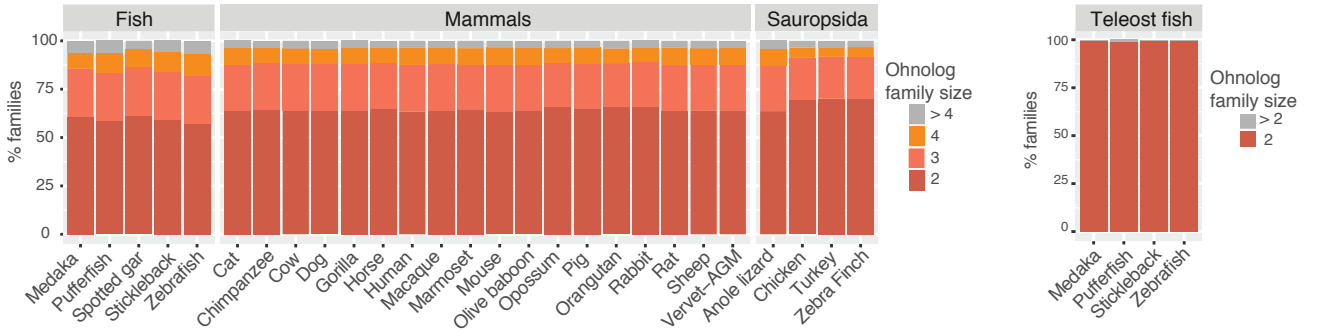


B. Individual 3R-ohnologs, pairs and families in the teleost fish

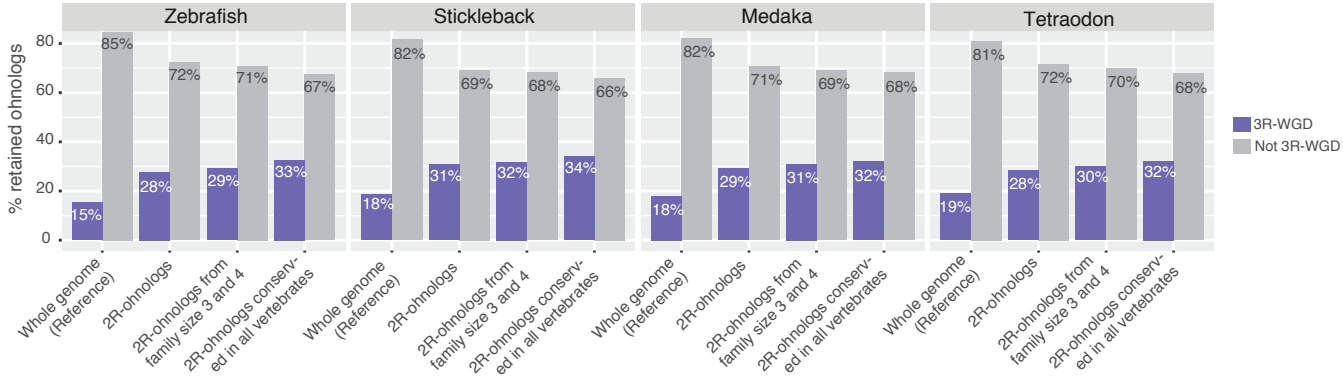


C. 2R-ohnologs family sizes

D. 3R-ohnologs family sizes



E. 2R-ohnologs retain more 3R-ohnologs



Supplementary Methods

for manuscript

OHNOLOGS v2: A comprehensive resource for the genes retained from whole genome duplication in vertebrates

Param Priya Singh[1,2] & Hervé Isambert[1]

[1]Institut Curie, Research Center, CNRS UMR168, PSL Research University, 26 rue dUlm, 75005 Paris, France
[2]Present Address: Stanford University, Department of Genetics, 300 Pasture Drive, Stanford, CA 94305, USA

Contents

1 Confidence assessment of ohnolog pairs: combining q-scores

1.1 OHNOLOGS v1 database: a simple average of log q-scores over vertebrate species

The confidence assessment of individual ohnolog pairs in the original OHNOLOGS v1 database (*1*) relied on the definition of quantitative outgroup and self-synteny scores (q-scores) for each vertebrate species, *i.e.* Q^k_{outgroup} and Q^k_{self}, where k =human, mouse, rat, pig, dog and chicken. See Singh *et al.* PLoS Comput Biol 2015 paper (*1*) for a detailed computation of q-scores from synteny comparison.

Then, to circumvent the difficulty to identify ohnolog pairs in each vertebrate genome due to lineage specific rearrangement, gene loss and small scale duplication events, geometric averages of outgroup and self-synteny q-scores were taken over the six amniote species included in the OHNOLOGS v1 database:

$$\log \bar{Q}_{\text{outgroup}} = \sum_{k=1}^{6} \frac{1}{6} \log Q^k_{\text{outgroup}} \tag{S1}$$

$$\log \bar{Q}_{\text{self}} = \sum_{k=1}^{6} \frac{1}{6} \log Q^k_{\text{self}} \tag{S2}$$

Using such averaged q-scores was shown to improve the statistical significance of the inferred ohnologs by allowing to identify ohnolog pairs that are no longer in significant synteny in a particular vertebrate genome, if their respective orthologs form a high confidence ohnolog pair in other vertebrates.

However, simple q-score averages fall short of (i) assessing the gain in statistical power expected from the integration of multiple vertebrate species (as the weights 1/6 in Eqs. S1 and S2 sum to 1), as well as, (ii) taking into account the phylogenetically biased sampling of vertebrate species by using equal weights (1/6), while the more recently diverged mouse and rat genomes are expected to bring rather redundant information on ohnolog retention as compared to phylogenetically more distant species such as chicken.

1.2 OHNOLOGS v2 database: a weighted sum of log q-scores over vertebrate species

The expanded OHNOLOGS v2 database addresses the shortcomings on the statistical confidence of ohnolog pairs from the original OHNOLOGS v1 database.

To this end, we modified the definitions of outgroup and self-synteny q-scores, from Eqs. S1 and S2, as weighted sums of log q-scores over all $N = 27$ vertebrate species included in the OHNOLOGS v2 database for 2R-ohnologs and all $N = 4$ included teleost fish species for 3R-ohnologs (see Table S1),

$$\log \bar{Q}_{\text{outgroup}} = \sum_{k=1}^{N} w_k \log Q^k_{\text{outgroup}} \tag{S3}$$

$$\log \bar{Q}_{\text{self}} = \sum_{k=1}^{N} w_k \log Q^k_{\text{self}} \tag{S4}$$

where the weights (w_k) are meant to (i) capture the gain in statistical power expected from the integration of 27 vertebrates including 4 teleost fish species (*i.e.* $\sum_k w_k > 1$) and (ii) take into account the strong phylogenetically biased sampling of included species by using different weights for each vertebrate genome depending on its shared homology with other included genomes.

The computation of the individual weights, $w_k^{\rm 2R}$ for 2R-ohnologs and $w_k^{\rm 3R}$ for 3R-ohnologs, are detailed in the following section. It is based on the times of divergence between each pairof vertebrate genomes included in the study (Table S2) and the values of $w_k^{\rm 2R}$ and $w_k^{\rm 3R}$ are listed in Table S3.

The overall gain of statistical power is estimated as $\sum_k w_k^{\rm 2R} \simeq 4.52$ for 2R-ohnologs and $\sum_k w_k^{\rm 3R} \simeq 2.41$ for 3R-ohnologs. This corresponds to an effective number of "independent species" of about 4.5 out of the 27 included vertebrates for assessing the confidence of 2R-ohnologs and to an effective number of "independent species" of about 2.4 out of the 4 included teleost fish for assessing the confidence of 3R-ohnologs.

In addition, as anticipated, recently diverged species of overrepresented vertebrate subgroups are assigned very small weights, which only amount to a very small fraction of the total weight. In particular, each of the 8 included primates has an individual weight around 0.01-0.02, while the sole representatives of long diverged subgroups have proportionally very large weights, such as Spotted Gar ($w \simeq 0.72$) or Anole Lizard ($w \simeq 0.57$).

A consequence of the gain of statistical power between OHNOLOGS v1 and v2 databases is that we could define more stringent confidence criteria for ohnolog pairs and generated ohnolog families as,

- strict $\quad \bar{Q}_{\rm outgroup} < 0.001$ AND $\bar{Q}_{\rm self} < 0.001$
- intermediate $\quad \bar{Q}_{\rm outgroup} < 0.01$ AND $\bar{Q}_{\rm self} < 0.01$
- relax $\quad \bar{Q}_{\rm outgroup} < 0.05$ AND $\bar{Q}_{\rm self} < 0.3$

2 Weighting scheme for phylogenetically related sequences

As discussed above, the effective number N' of "independent species" is smaller than the actual number N of phylogenetically related species included in the analysis.

One way to estimate N' and the corresponding weights w_k for each phylogenetically related species (with $\sum_k^N w_k = N'$) is through the apparent increase of variance of an ordinal character x (such as the number of genome rearrangements) across N non-independent species. The result is quite general and the increase of variance can be used to infer consistent weights for a generic dataset of N non-independent samples, as discussed in the next section.

2.1 Generic increase of variance due to non-independent samples

The generic increase of variance between the N non-independent samples can be illustrated on the example of a theoretical dataset with N' independent samples, each repeated $n_{k'}$ times (or not repeated if $n_{k'} = 1$) to yield a larger dataset of $N = \sum_{k'=1}^{N'} n_{k'}$ non-independent samples.

The variance obtained for the larger non-independent dataset of size N reads:

$$\begin{aligned} V_N &= \frac{1}{N^2}\sum_{k}^{N}\sum_{\ell}^{N}\langle \delta x^{(k)}\delta x^{(\ell)}\rangle \\ &= \frac{1}{N^2}\sum_{k'}^{N'}\sum_{\ell'}^{N'} n_{k'}n_{\ell'}\langle \delta x^{(k')}\delta x^{(\ell')}\rangle \\ &= \frac{1}{N^2}\sum_{k'}^{N'} n_{k'}^2\langle \delta x^{(k')^2}\rangle \\ &= \frac{1}{N^2}\sum_{k}^{N} n_{k}\langle \delta x^{(k)^2}\rangle \end{aligned} \tag{S5}$$

as $\langle \delta x^{(k')}\delta x^{(\ell')}\rangle = \delta_{k',l'}\langle \delta x^{(k')^2}\rangle$ for independent samples and using $\sum_{k'}^{N'} n_{k'}f(k') \equiv \sum_{k}^{N} f(k)$ with $n_k = n_{k'}$ for each of the $n_{k'}$ samples k that are duplicates of sample k'.

When all samples are independent, that is, if $n_k = 1$ for all N samples, one recovers the well known results (adopting the rescaling $\langle \delta x^{(k)^2}\rangle = 1$ for all k),

$$V_N = \frac{1}{N^2}\sum_{k}^{N} n_{k}\langle \delta x^{(k)^2}\rangle = \frac{1}{N} \tag{S6}$$

By contrast when the samples are not all independent, that is, if $n_k > 1$ for some of the N samples ($\sum_{k}^{N} n_k > N$), one gets

$$V_N = \frac{1}{N^2}\sum_{k}^{N} n_{k}\langle \delta x^{(k)^2}\rangle = \frac{1}{N_{\rm app}} > \frac{1}{N} \tag{S7}$$

as if the apparent number of independent samples was smaller, $N_{\rm app} < N$.

This suggests to weight each non-independent sample k with a probability weight, $w_k = 1/n_k \leqslant 1$ with $\sum_{k}^{N} w_k = N'$ and to define the corrected variance for effective sample size as,

$$\begin{aligned} V_{N'} &= \frac{1}{N'^2}\sum_{k}^{N}\sum_{\ell}^{N} w_k w_\ell\langle \delta x^{(k)}\delta x^{(\ell)}\rangle \\ &= \frac{1}{N'^2}\sum_{k'}^{N'}\sum_{\ell'}^{N'} w_{k'}n_{k'}w_{\ell'}n_{\ell'}\langle \delta x^{(k')}\delta x^{(\ell')}\rangle \\ &= \frac{1}{N'^2}\sum_{k'}^{N'} w_{k'}^2 n_{k'}^2\langle \delta x^{(k')^2}\rangle \\ &= \frac{1}{N'^2}\sum_{k'}^{N'}\langle \delta x^{(k')^2}\rangle = \frac{1}{N'} \end{aligned} \tag{S8}$$

using $w_k n_k = 1$. Note that $V_{N'}$ can also be expressed in the actual sample space with N non-independent samples, instead of the effective sample space with N' independent samples (which are not typically known), as,

$$\begin{aligned} V_{N'} &= \frac{1}{N'^2} \sum_k^N \sum_\ell^N w_k w_\ell \langle \delta x^{(k)} \delta x^{(\ell)} \rangle \\ &= \frac{1}{N'^2} \sum_{k'}^{N'} w_{k'}^2 n_{k'}^2 \langle \delta x^{(k')^2} \rangle \\ &= \frac{1}{N'^2} \sum_k^N w_k^2 n_k \langle \delta x^{(k)^2} \rangle \\ &= \frac{1}{N'^2} \sum_k^N w_k \langle \delta x^{(k)^2} \rangle = \frac{1}{N'} \end{aligned} \tag{S9}$$

using $\sum_{k'}^{N'} n_{k'} f(k') \equiv \sum_k^N f(k)$ and $\forall k, w_k n_k = 1$ and $\langle \delta x^{(k)^2} \rangle = 1$.

2.2 Sample weighting scheme by inversion of the variance-covariance matrix

The above results show that the sample weights $\{w_k\}$ are solutions of the following equation,

$$\sum_k^N \sum_\ell^N w_k w_\ell \langle \delta x^{(k)} \delta x^{(\ell)} \rangle = \sum_k^N w_k \langle \delta x^{(k)^2} \rangle \tag{S10}$$

While Eq. S10 with N unknown weights is underdetermined, one can easily show that this equation also applies to individual summand for each k as,

$$\forall k, \quad \sum_\ell^N w_\ell \langle \delta x^{(k)} \delta x^{(\ell)} \rangle = \sum_{\ell'}^{N'} w_{\ell'} n_{\ell'} \langle \delta x^{(k)} \delta x^{(\ell')} \rangle = w_k n_k \langle \delta x^{(k)^2} \rangle = \langle \delta x^{(k)^2} \rangle, \tag{S11}$$

using $\sum_\ell^N f(\ell) \equiv \sum_{\ell'}^{N'} n_{\ell'} f(\ell')$ and $\forall k, w_k n_k = 1$.

Eq. S11 can be written in the following matrix form, after rescaling $\delta x^{(k)}$ by its mean deviation as $\delta x^{(k)} / \sqrt{\langle \delta x^{(k)^2} \rangle}$,

$$\Sigma \begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tag{S12}$$

where $\Sigma = [\Sigma_{k\ell}]$ with $\Sigma_{k\ell} = \langle \delta x^{(k)} \delta x^{(\ell)} \rangle / \sqrt{\langle \delta x^{(k)^2} \rangle} \sqrt{\langle \delta x^{(\ell)^2} \rangle}$ is the rescaled variance-covariance matrix between samples, which leads to the following weight solution whenever the variance-covariance

matrix is invertible,

$$\begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix} = \Sigma^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tag{S13}$$

While Eq. S13 seems to give the straightforward solution of the generic sample weighting problem, in practice, the variance-covariance matrix Σ is typically not invertible. In particular, straightforward averages of variances and covariances over the available samples, which imply $\sum_k^N \delta x^{(k)} = 0$, yields a singular variance-covariance matrix (as all rows and columns sum to zero).

Yet, in some particular cases, the form of the variance-covariance matrix can be conjectured independently from the specific data of interest and used to solve Eq. S13.

This is the case for time series of dynamical systems with exponential relaxation over time (*2, 3*) or for phylogenetically related sequences (*4, 5*), as discussed in the next section.

2.3 Application to weighting phylogenetically related sequences

The form adopted for the variance-covariance matrix of phylogenetically related sequences is directly inspired by the form proposed by Altschul *et al* in ref. (*5*) to estimate weights of sequence data related by a tree.

Following these authors, we reason that as genome rearrangements and gene loss events accumulated in ancestral vertebrate genomes after each WGD event, the distance of their alignment with the reference paleoploid genome progressively shift. At first approximation, one expect a *linear* accumulation of some finite number (X) of genome rearrangements and gene loss events over time, as these evolutionary changes are essentially non-reversible (small scale duplication events might even lead to exponential growth of gene families over time (*6*)). This is to be contrasted with an unbiased reversible random walk in sequence space, which would lead to a purely diffusive dynamics with a sublinear (square root) accumulations of changes over time.

Hence, due to this progressive shift in genome space, the variance of accumulated changes, X_k, of a given vertebrate genome, G_k, is expected to increase *quadratically* with time t_k since a WGD event, *i.e.* $\langle \delta x^{(k)^2} \rangle \sim \sigma^2 t_k^2$, instead of linearly with time for a perfectly diffusive dynamics.

Similarly, the covariance of accumulated changes in two genomes, G_k and G_ℓ, having diverged after some time $t_{k\ell}$ after a WGD event is expected to increase *quadratically* as, $\langle \delta x^{(k)} \delta x^{(\ell)} \rangle \sim \sigma^2 t_{k\ell}^2$, assuming that subsequent changes after the two genomes have diverged were completely independent and could not therefore further increase the covariance.

All in all, this leads to the following form for the rescaled variance-covariance matrix, $\Sigma = [\Sigma_{k\ell}]$, where $\Sigma_{k\ell} = \langle \delta x^{(k)} \delta x^{(\ell)} \rangle / \sqrt{\langle \delta x^{(k)^2} \rangle} \sqrt{\langle \delta x^{(\ell)^2} \rangle} = t_{k\ell}^2 / t_k t_\ell$, that is independent for the prefactor σ^2.

In the application to compute the weight w_k of each species, the times of 2R-WGD and 3R-WGD were estimated by averaging recent estimates as $t_{2R} = 535$ MY (*7–10*) and $t_{3R} = 328$ MY (*11–15*), respectively, and the times since divergence of each pairs of species, $d_{k\ell} = t_{WGD} - t_{k\ell}$, were taken from the TimeTree database (*16*) and listed in Table S2. The final values for 2R- and 3R-WGD weights are listed in Table S3.

3 References

1. Singh PP, Arora J, Isambert H (2015) Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Comput. Biol.* 11(7):e1004394.

2. Jones RH (1975) Estimating the variance of time averages. *J. Appl. Meteor.* 14(2):159–163.

3. Verny L, Sella N, Affeldt S, Singh PP, Isambert H (2017) Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* 13(10):e1005662.

4. Felsenstein J (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* 25(5):471–492.

5. Altschul SF, Carroll RJ, Lipman DJ (1989) Weights for data related by a tree. *J. Mol. Biol.* 207(4):647–653.

6. Evlampiev K, Isambert H (2008) Conservation and topology of protein interaction networks under duplication-divergence evolution. *Proc. Natl. Acad. Sci. U.S.A.* 105(29):9863–9868.

7. Putnam NH, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453(7198):1064–1071.

8. Holland LZ, et al. (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* 18(7):1100–1111.

9. Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10(10):725–732.

10. Smith JJ, Keinath MC (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. *Genome Res.* 25(8):1081–1090.

11. Jaillon O, et al. (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 431(7011):946–957.

12. Hoegg S, Brinkmann H, Taylor JS, Meyer A (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* 59(2):190–203.

13. Christoffels A, et al. (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol.* 21(6):1146–1151.

14. Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci. U.S.A.* 101(6):1638–1643.

15. Hurley IA, et al. (2007) A new time-scale for ray-finned fish evolution. *Proc. Biol. Sci.* 274(1609):489–498.

16. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S (2015) Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* 32(4):835–845.