1 Evolutionary patterns of 64 vertebrate genomes (species) revealed by phylogenomics
2 analysis of protein-coding gene families

3 Jia Song^{1,2#}, Xia Han¹, Kui Lin^{1*}

4 ¹ State Key Laboratory of Earth Surface Processes and Resource Ecology and Ministry of
 5 Education Key Laboratory for Biodiversity Science and Ecological Engineering, College of
 6 Life Sciences, Beijing Normal University, Beijing 100875, China

7 ² Institute of Molecular Medicine, Renji Hospital, Shanghai Jiao Tong University School of
 8 Medicine, Shanghai, 200127, China

9 *Corresponding author

10 #present address

11 E-mail:

12 [Kui Lin: linkui@bnu.edu.cn](mailto:linkui@bnu.edu.cn)

13 [Jia Song: songjia2013123@mail.bnu.edu.cn](mailto:songjia2013123@mail.bnu.edu.cn)

14 [Xia Han: hanxia@mail.bnu.edu.cn](mailto:hanxia@mail.bnu.edu.cn)

15

16 **Abstract**

17 **Background:** Recent studies have demonstrated that phylogenomics is an important basis for
18 answering many fundamental evolutionary questions. With more high-quality whole genome
19 sequences published, more efficient phylogenomics analysis workflows are required urgently.

20 **Results:** To this end and in order to capture putative differences among evolutionary histories
21 of gene families and species, we developed a phylogenomics workflow for gene family
22 classification, gene family tree inference, species tree inference and duplication/loss events
23 dating. Our analysis framework is on the basis of two guiding ideas: 1) gene trees tend to be
24 different from species trees but they influence each other in evolution; 2) different gene
25 families have undergone different evolutionary mechanisms. It has been applied to the
26 genomic data from 64 vertebrates and 5 out-group species. And the results showed high
27 accuracy on species tree inference and few false-positives in duplication events dating.

28 **Conclusions:** Based on the inferred gene duplication and loss event, only 9~16% gene
29 families have duplication retention after a whole genome duplication (WGD) event. A large
30 part of these families have ohnologs from two or three WGDs. Consistent with the previous
31 study results, the gene function of these families are mainly involved in nervous system and
32 signal transduction related biological processes. Specifically, we found that the gene families
33 with ohnologs from the teleost-specific (TS) WGD are enriched in fat metabolism, this result
34 implying that the retention of such ohnologs might be associated with the environmental status
35 of high concentration of oxygen during that period.

36 **Keywords:** Gene family tree; Species tree; Phylogenomics; Vertebrate; Duplication
37 retention/preservation.

38

1. Background

With the recent advances in next-generation genome sequencing technologies, a large amount of high-quality genomes covering diverse taxa have been published[1]. The development and application of efficient and practical computational methods, such as comparative genomics[2], are very helpful for scientists to use these data to understand the underlying genetic mechanisms[3]. As one kind of comparative genomics strategies, phylogenomics[4] was firstly raised by Eisen JA in 1998. At first it had been exclusively defined as the prediction of protein functions from a phylogenetic view[4]. While in molecular systematics, phylogenomics is usually used to infer the evolutionary relationship of species using genome-scale sequencing data[5]. Uniting these two disparate definitions, phylogenomics is now widely regarded as the molecular phylogenetic analysis of genome-scale data sets[6], which can be used for predicting gene function[7-10], inferring evolutionary patterns of macromolecules[11-13], establishing the relationships and divergence times of genes/species[14, 15], exploring the genome duplications[16-19], and so on.

Phylogenomics data are available in several databases, such as EnsemblCompara[20], PhylomeDB [21] and Panther[22]. But high-quality phylogenomics data is still indispensable. On the one hand, these databases are known to contain many errors and uncertainties[23]. Directly using them in orthology detection or genome dynamics study could lead to erroneous results[24]. The causes of these errors are variable. As far as concerned, these databases considered little in the following two aspects: 1) the differences in histories of genes and species because of a hierarchy of evolutionary processes[25]; 2) the different selection stress on duplication/loss events in different gene families. On the other hand, most of these databases only contain data from model species, which is a limitation to the new sequenced genomes. Therefore, we believed an integrative and universal phylogenomics workflow, which is able to capture more differences among the evolutionary processes of different gene families and species, is imperative.

Here, we constructed a phylogenomics workflow mainly based on OrthoFinder[26], BEAST[27], Guenomu[28], RAxML[29], Notung[30], IQ-TREE[31] and SiCIE[32] aiming to include the following two guiding concepts: 1) gene trees tend to be different from species trees but they influence each other in evolution; 2) different gene families have undergone different evolutionary mechanisms. In detail, an efficient species tree inference method and a parameter-learning method were proposed to model the evolutionary differences among

different gene families and species trees. Based on protein sequences and CDSs (coding DNA sequences) from certain species, our workflow was designed to conduct species tree inference and duplication/loss dating following gene families' classification and gene family tree inference/modification. As a case study, we applied our workflow to get the gene duplication history of 64 vertebrates' genomes.

Duplications are of great significance as they would affect single gene, a stretch of several genes, whole chromosomes or even whole genomes and they are considered as the major driving forces for evolution of genetic novelty[33, 34]. However, many basic features of the evolution by gene duplication remain unknown[33, 35]. We applied our workflow on the genomic data of 64 vertebrates and 5 other eukaryotic species from Ensembl v84[36]. A species tree and 9,767 reconciled gene family trees were obtained. These results were then used to explore the WGD retention patterns and features, long-term local duplication preservation events and relative gene functions on vertebrate genomes.

2. Results and Discussion

2.1 An efficient gene tree–species tree phylogenomics workflow

2.1.1 Introduction to our phylogenomics workflow

A phylogenomics workflow was constructed for multi-species genome evolutionary history exploration. As shown in Figure 1, the whole workflow could be divided into four processes. Under the guidance of the first guiding concept that we have mentioned above, the initial species tree was inferred based on the posteriors of gene families trees under a bayesian supertree model, which take both the gene duplication-loss and multispecies coalescent events into consideration. Meanwhile, inspired by supertree methods, whole genome-wide gene family trees were then used to revise the initial species tree based on the incongruent clades between the initial species tree and the available public species tree. In this way, it is able to efficiently reduce computational complexity by using the available species tree information and guarantee the accuracy by using genome-wide data. Then under the guidance of our second guiding concept, the fourth process in our workflow applied a parameter-learning process, which was designed to conduct gene tree modification and gene duplication/loss events dating. During the duplication/loss dating process, the parameters (event-costs: costdup and costloss) setting makes great influences[37]. In the previous studies[11, 20, 21], event-costs were usually set to the same values for all families. Here, we designed a parameter-learning process to find out the optimal parameter set for each gene family, which

105 may help to capture the difference of selection pressures on gene duplication/loss in different
106 gene families.

107 *2.1.2 Comparison with other similar works*

108 In order to quantify the accuracy of our phylogenomics workflow, we compared the
109 inferred species tree with the mammals species tree published by Song *et al.* 2012[38] (Figure
110 S1 in additional file 1) and compared the inferred reconciled gene family trees with
111 EnsemblCompara in ancestral genome content metric, ancestral chromosome linearity
112 metric[24] and duplication consistency score[20]. Here, ancestral genome content metric is
113 based on the assumption that the ancestral genome content sizes should be close to the extant
114 genomes. Ancestral chromosome linearity metric assumed that each gene on ancestral
115 genomes should have zero, one or two neighbors, with a peak at two while genes with three
116 or more neighbors are the errors from the inferences. And duplication consistency score
117 measures the intersection of the number of species post duplication over the union. It's based
118 on the assumption that most duplication should have the gene persisting at least in an equally
119 likely manner in subsequent lineages[20].

120 Firstly, we compared the inferred final species tree (Figure 2) with Song's mammals tree.
121 Among the totally 31 shared species, only the tree shrew showed a incongruent evolutionary
122 location between the two species trees. The correct location of tree shrew along the species
123 tree is still under controversy[39]. Thus, our final species tree shows high accuracy in the
124 mammals' clade.

125 Secondly, according to our reconciled gene family trees, there were 50,916 duplication
126 events occurred in the evolutionary history of 9,767 gene families. For the related 8,514 gene
127 family trees from EnsemblCompara, there were 132,396 duplications. Then, as shown in
128 Table 1, ancestral genome size inferred from our results shows closer average size to extant
129 genomes than EnsemblCompara. As shown in Figure 3A, results from our workflow include
130 much more ancestral genes with two neighbors and less genes with three or more neighbors
131 compared with EnsemblCompara. Figure 3B shows clearly that the vast majority of the
132 duplications from our workflow have a higher duplication consistency score compared with
133 EnsemblCompara. Above all, EnsemblCompara output much more duplication nodes
134 compared with our workflow. The vast majority of these duplications from EnsemblCompara
135 perform worse on the three metrics mentioned above. Furthermore, we inferred another
136 phylogenomics result by following our workflow but without the reconciliation
137 parameter-learning in process 4. Results improved a little by the reconciliation

parameter-learning according to the three metrics. Actually, the reconciliation parameter-learning process might have bring in more improvement on accuracy. Because we can only compare the results based on the 9,767 gene families in our core set while parameter-learning have already helped us to filter out the gene families which easy to receive wrong reconciliation results.

2.1.3 Limitations and future development

In the species tree inference process, only 527 gene families were used to infer the initial species tree to avoid costing too much computational time to get the gene family tree posteriors. Theoretically, most important information reflected by other gene families will be lost. So we revised the less supported clades on the initial species tree based on genome-wide gene family trees. However, there are two problems. First, algorithms that can deal with genome-wide gene families directly are more preferred. Second, there is no available species tree like the Ensembl species tree at most times.

Algorithms able to directly infer the species tree based on all gene families are more preferred. However, as the representations of the two main categories of such methods, Phyldog and *BEAST are not suitable for big scale family data. Firstly, methods as *BEAST cannot deal with paralogous genes which are common in gene families. Secondly, Phyldog[40] is limited by the sample size. Phyldog was designed to co-estimate genes and species trees under a DL model in a maximum likelihood framework, which get results in a short running time theoretically. Under our test, however, it was out of memory (our computational resource: 4T in memory) when we applied Phyldog on all the 11,698 families by default parameters. Then the family number was reduced to about 130, it can infer a species tree and 130 the gene family trees. From the Phyldog species tree (Figure S2 in additional file 1), we can see some obvious mistakes. Perhaps, the MSAs of the selected 130 gene families were not enough to reflect the real relationships of species. We will try to seek or develop an efficient species tree inference algorithm, which is able to co-estimate gene and species tree basing on genome-wide gene families for our workflow in the near future. Currently for our limited computational resources and large-scale data, our workflow may be a good choice. In addition, there is no available species tree like the Ensembl species tree at most times. To overcome this, it could be a proper way to choose two or more gene family sets randomly to get two or more initial species trees and compare the initial trees with each other to get the incongruent clades (Figure S3 in additional file 1).

In addition, read-through genes might also cause problems. A read-through/conjoined[41-43] gene is formed at the time of transcription by combining at least part of one exon from each of two or more distinct (parent) genes. In the gene family classification process, read-through genes/proteins will result in some nesting gene families including their parents. Such situations have not been considered in most phylogenomics datasets. In our case study, we seek these nesting gene families (Figure S4 in additional file 1) based on the read-through genes annotated in the GENCODE annotation file of HUMAN (V24) [44] and filtered out 454 such families. However, annotations of read-through/conjoined genes on other genomes are lacking or in low accuracy. It merits further attention to find a better way to deal with these families.

2.2 The features of whole genome duplication on vertebrate evolution

It is now clear that there have been three major WGDs in vertebrate genomes evolutionary history. Two (named 1R WGD and 2R WGD respectively) occurred near the base of the vertebrates' evolutionary history and the third (named TS WGD) occurred at the base of the teleost fishes' evolutionary history [45-49]. Although WGDs are often credited with great evolutionary importance, the processes governing the retention of ohnologs (paralogs generated by WGD) and their biological significance remain unclear. In this section, we explored the patterns of ohnologs retention and the relative function based on our reconciliation results of 9,767 gene families.

We got the gene families with ohnologs retention by seeking the duplications on the reconciled gene family trees and then mapped these duplications onto the species tree. Similar with previous studies[50-53], these three WGD-affected ancestral branches show about 9%~16% gene duplication retention (additional file 2, supplementary material) which are significantly higher than other ancestral branches (P-value = 0.00193, Wilcox test, Table S1 and Figure S5 in additional file 1). Protein-protein interactions (PPIs) are enriched among the members of these gene families according to the human genes and their PPI data (Table S2 in additional file 1). This might reflect the gene dosage selection effects[54] after WGDs. Ohnologs retention from the 2R WGD might have undergone the weakest dosage selection among these three WGDs. Then based on duplication overlap rates (defined in Materials and Methods), we found that compared with other branches on the species tree, gene families with duplication retentions on the three WGD-affected branches are significantly more overlapped (P-value < 0.05, Fisher exact test). As shown in Figure 4A, 68 gene families retained ohnologs after all the three WGDs and 588 families after at least two WGDs.

203 According to human gene ontology information, gene families with ohnologs retention
 204 after these three WGDs are mainly involved in development, signaling and gene regulation
 205 (Figure S6 in additional file 1), which are consistent with the previous studies[55-59]. Then
 206 we divided these families into seven classes according to their ohnologs retention pattern after
 207 the three WGDs (Figure 4B). We found the 68 gene families with ohnologs retention after all
 208 of the three WGDs (class 1) are mainly involved in functional categories related to neuron,
 209 axon, signal and cell growth. Class 2 consists of gene families with ohnologs retention after
 210 both the TS and the 1R WGD and class 3 consists of gene families with ohnologs retention
 211 after both the TS and the 2R WGD. These two classes show similar GO enrichment results
 212 and they are both enriched in functional categories related to neuron, axon and cell-junction.
 213 Class 4, which consists of gene families with ohnologs retention after both the 1R and the 2R
 214 WGDs, are mainly involved in signal transduction. Above all, combined with the results from
 215 published studies[60, 61], nervous system and signal transduction related gene families are
 216 highly expanded on all vertebrate genomes through these three WGDs. Combined with the
 217 PPI enrichment results, this retention pattern may be a result of gene dosage selection.

218 The other three classes that consist of gene families with all ohnologs from one WGD are
 219 enriched in different functional categories and might reflect different retention mechanisms.
 220 We found that gene families in class 7 are enriched in fat metabolism. Further, we used the
 221 gene ontology data of zebra fish to redo the GO enrichment analysis, and the results (Figure
 222 4C) showed more GO terms involved in fat metabolism, including anabolism and catabolism.
 223 As is well known, fat releases much more energy than other nutrients such as carbohydrate
 224 and protein in exhaustive oxidation and this process costs much more oxygen at the same
 225 time. More specifically, acyl-CoA and fatty-acyl-CoA, which included in many enriched
 226 biological processes, are essential products in metabolic process with oxygen consumption.
 227 Interestingly, this TS WGD happened at the period that the earth has its highest content of
 228 oxygen level (up to 33%) during the evolutionary history of vertebrates (Figure S7 in
 229 additional file 1). All of these lead to a suggestion that the high content of oxygen might be a
 230 kind of selection to the duplication retention after the TS WGD to promote fat as a main way
 231 to store energy. This might be one reason that fish have more unsaturated fatty and it is worth
 232 more discussion in future works.

233 **2.3 The features of local duplications on vertebrate genomes**

234 It should be noticed that many paralogs in current gene families were not originated by
 235 the WGD events mentioned above, but by extensive local duplications[62]. So we also

identified the local duplications from our reconciliation results to explore such retention pattern in vertebrates. We firstly found that there were many more duplications occurred on the extant species-specific branches than the ancestral ones on the species tree (Kolmogorov-Smirnov test in R, $p\text{-value} < 2.2e-16$, Table S1 in additional file 1). As previous studies[63, 64] indicated that three steps are responsible for the generation of preserved gene duplications: origin through mutation (duplication), a fixation/spreading phase and a preservation/maintenance phase when the fixed change is maintained. The majority of duplications on extant species might still be under the fixation/spreading phase. While most duplications on ancestral genomes might already be under the preservation/maintenance phase for the most recent ancestral genome on our species tree existed 6.5 million years ago, which has already exceeded the average half-life of a gene duplication (approximately 4 million years) provided by previous studies[33].

Previous studies always focused on the duplication mutation rates and duplication fixation rates. Different from these studies, we estimated the duplication preservation/maintenance rates based on the duplications annotated on the ancestral genomes and their origin time (see Materials and Methods). After removing the duplications resulted from WGDs, we estimated the duplication preservation rates for 9,581 gene families. 7,075 gene families have no local duplications on ancestral genomes, which indicated that about 74% gene families in our core data have no long-term duplication preservation and most gene families kept singleton status during the evolutionary history of vertebrate genomes. We then got non-zero duplication preservation/maintenance rates for 2,506 gene families and 95% duplication preservation rates of these gene families are distributed between 0.0009 and 0.016 (Figure S8 in additional file 1). According to the gene and GO information from human, the gene families with long-term local duplications preservation are mainly involved in ion transport and some important signaling pathways. In addition, we also found that some local gene duplications might be retained through the natural selection caused by oxygen-level changes (Figure 5).

3. Conclusions

Based on two guiding concepts, we developed an integrative phylogenomics workflow by integrating an efficient species tree inference workflow, which adopt advantages from co-estimation and supertree methods, and a parameter-learning process to account for more about the relationship and differences among species and gene trees. It was designed for gene family classification, gene family tree and species tree inference and duplication/loss dating.

269 Then, we analyzed the genomic data of 64 vertebrates and 5 out-groups from Ensembl as a
270 case study to demonstrate a complete application of our workflow on the accurate inference
271 of the evolutionary history of genome-wide gene families and species. Based on our
272 phylogenomics results, we captured evolutionary traces from two different duplication
273 retention mechanisms. We found that dosage selection might play an important role on
274 ohnologs retention after WGDs and the changing environmental oxygen content might be a
275 kind of natural selection affecting paralogs from both WGDs and local duplications. Above
276 all, we expected that our workflow will facilitate further studies aiming to explore genome
277 evolutionary histories.

278

279 **4. Methods**

280 **4.1 Gene family classification**

281 In order to get genomic sequences and annotations with high quality, we used the data of
282 69 species from Ensembl v84 as a case study to introduce our workflow. We downloaded all
283 protein sequences and CDS sequences of these species from FTP site of Ensembl
284 (<http://www.Ensembl.org/>, Build 84)[36] and chose their longest protein and CDS to be the
285 representation for each gene. The too short (shorter than 10aa) and too simple (stop codons
286 percent greater than 20) genes were filtered out then.

287 Here, OrthoFinder-0.4[26] was used to identify homology relationships between these
288 sequences. OrthoFinder is a very efficient algorithm, which can overcome gene length bias
289 and phylogenetic distance problems in gene family classification. After this step, we got
290 totally 54,808 gene families. We then removed too simple (members from a unique species)
291 and too complex gene families, which including known read-through genes (according to
292 gene annotation file (v24) of human in ENCODE[44]) or with more than 1,000 members.
293 17,025 gene families were left for following analysis (additional file 2)

294 **4.2 Gene family tree inference**

295 In this step, protein sequences of gene families were aligned in MAFFT v7[65](--auto)
296 and then translated into CDS alignments by translatorX[66]. The poorly aligned regions were
297 removed from these CDS MSAs by trimAl[67]. Here, we removed some gene families with
298 specific labels in its sequences (such as X) or with very poor alignment quantity (additional
299 file 2). For the left 14,037 CDS MSAs, we inferred the gene family trees in RAxML
300 v8.2.9[29] under GTRGAMMI sequence evolution model. For some MSAs including less

than four members and some MSAs including too much gaps, we finally only got reliable phylogenetic trees for 11,698 gene families.

4.3 Species tree inference

There are 579 gene families including members from all of the 69 species. We filtered out the gene families with members' distribution various largely ($CV > 0.5$) on different species to avoid information asymmetry. So 527 gene families were left for species tree inference.

BEAST[27] (parameters: a gamma-distributed model of rate variation with four discrete categories and an HKY substitution model with a strict clock, 10,000,000 generations, sampling every 5000 generations) was used to infer the posterior distributions of these 527 gene family trees. The results possessed a good convergence under these parameters setting (with effective independent sample size greater than 200 for each parameter). Guenomu was used to infer species tree by considering gene duplication, loss and multispecies coalescent simultaneously (10,000,000 generations, sampling every 10,000 generations) based on these tree posteriors. It outputted two species trees and we used the one with 99.9% probability as the initial species tree.

In addition, we downloaded the Ensembl species tree inferred by EnsemblCompara to find out the possible errors on the initial tree. Firstly, we compared the initial species tree with Ensembl species tree to find out the incongruent clades. We found seven species (including ancestral species) bearing different phylogenetic sister-branches between these two trees (Figure S9 in additional file 1). Secondly, in order to find out the true phylogenetic sister-branches of these seven species, SiCIE v1.2[32] was used to extract phylogenetic supports from the 11,698 gene family trees. The results (Table S3 in additional file 1) show that three clades on initial species tree got significantly higher supports than the respective clades on the Ensembl species tree. Conversely, other three clades on Ensembl species tree got significantly higher supports. Unfortunately, the rest incongruence couldn't find a clear relationship from these 11,698 gene family trees. We then improved the initial species tree by modifying its three weaker supported incongruent clades. The final species tree is displayed in Figure 2.

4.4 Species/gene trees reconciliation

Inspired by the "Felsenstein equation" [68], we put forward a parameter-learning method to find out the optimal event-costs for each gene family based on two optimal principles. Firstly, the modified gene family tree should have largest ML (maximum likelihood) value based on the corresponding MSA (multiple sequence alignment) of CDS. Secondly, the

334 optimal reconciled results should contain the fewest number of events to explain the
335 incongruences between the gene family tree and species tree. Then, based on the optimum
336 event-costs pairs of each gene family, we modified the low supported clades on the gene
337 family tree and further dated the evolutionary events (duplication and loss) by reconciliation.
338 In our case study, 11,698 gene family trees were used as inputs. Finally, 9,767 gene families
339 got uniquely reconciled gene family tree under their optimal event-costs pairs. More details
340 are described below.

341 We used Notung v2.8.1.7, IQ-TREE v1.5.2 and our parameter-learning scripts to finish
342 species/gene trees parameter-learning and reconciliation. After the event-costs pairs (costdup
343 and costloss) assignment, Notung is able to modify the gene family tree and date gene
344 duplications/losses under a DL model in a parsimony strategy. In order to seek the optimal
345 event-costs pairs set for each gene family, 15 event-costs pairs (costdup, costloss) with
346 different cost ratios (costdup/costloss) were used to parameter-learning and reconciliation in a
347 cycle process (Figure 1). The detailed steps are described as follow:

348 Step 1. Rearrange the gene family tree: the gene family tree was rearranged under the
349 ‘Rearrange mode’ of Notung. We rearranged the weakly supported regions (edges with
350 bootstrap less than 50) in the gene family tree to produce alternate gene family trees with
351 minimum DL score based on the current event-costs pair. Here, at most 100 eligible
352 alternate gene trees will be outputted. IQ-TREE was then used to pick out the most
353 optimal one with maximal maximum likelihood based on the respective CDS MSA.

354 Step 2. Root the gene family tree: the gene family tree was rooted under the ‘Rooting
355 mode’ of Notung by minimizing DL score based on current event-costs pair.

356 Step 3. Reconcile the species/gene tree: duplication/loss events were assigned on the
357 gene family tree under the ‘Reconcile mode’ of Notung by minimizing DL scores based
358 on current event-costs pair. Then current event-costs pair was set to the next pair and the
359 analysis jumped to step 1 if the current event-costs pair wasn’t the last one. Otherwise,
360 analysis jumped to step 4.

361 Step 4. Construct the optimal event-costs pairs set for each gene family: we used
362 IQ-TREE to calculate the ML (maximum likelihood) for each resulted gene family tree,
363 which was inferred under different event-costs pairs. In this way, we obtained the
364 optimal event-costs pairs set I, which consists of event-costs pairs resulting in maximal
365 ML trees. Meanwhile, we constructed the optimal event-costs pairs set II, which consists
366 of event-costs pairs resulting in minimal DL events in the reconciled results. The

intersection of optimal event-costs pairs set I and II was considered as optimal event-costs pairs III. The final optimal set for the family was empty if the optimal event-costs pairs set III contains more than one member and the reconciled results are inconsistent under members. Otherwise, the final optimal event-costs pairs set is the optimal event-costs pairs set III.

In our analysis, we totally used 15 (costdup, costloss) pairs (additional file 2). Finally, we obtained optimal reconciliation results for 9,767 gene families, which called core gene families in this study.

4.5 Comparison to EnsemblCompara

We downloaded the gene family trees inferred by EnsemblCompara from its FTP site. For our workflow integrated gene family classification, our gene families are not consistent with Ensembl's. Here we selected 8,514 Ensembl gene families with more than four gene members and overlapped with the 9,767 gene families in our core results to do the comparison.

Based on the gene adjacencies extracted from annotation files of the 69 extant species, DeCo[69] was used to infer the ancestral genome contents and ancestral gene adjacencies according to our gene trees and Ensembl gene trees, respectively. Then, we calculated the duplication consistency score[20] for each duplication on these two gene tree sets.

Duplication consistency score

$$= \frac{\text{The species intersection between the left and right sub-trees}}{\text{The species union between the left and right sub-trees}}$$

4.6 Others

4.6.1 PPI and GO enrichment analysis

We used protein-linked information from STRING (v10.5) to finish the PPI enrichment analysis. We firstly abstracted the human protein-protein interaction network with combined score greater than 700. Then we abstracted the sub-network, whose nodes consisting of genes in our core gene families and edges linking members from different gene families. We found that there were 118,028 edges out of 68,641,957 gene pairs from different gene families on this sub-network. Then, we counted edges and such gene pairs among different WGD-affected gene family classes. As Table S2 (additional file 1), we found the PPIs were enriched in these classes (Fisher exact test).

The GO (gene ontology) enrichment analysis in this study was conducted by R package named 'clusterProfiler'[70] basing on annotation data from 'org.Hs.eg.db' and 'org.Dr.eg.db'.

4.6.2 Local duplication preservation rate

397 The local duplication preservation rates were inferred based on the gene duplications on
 398 ancestral genomes and their origin time. Firstly, in order to get the approximate existing time
 399 of each ancestors on the species tree, we downloaded the dated species tree (Figure S7 in
 400 additional file 1) for the 69 species from TIMETREE (www.timetree.org)[71] and use its time
 401 information to date our species tree. We dated the ancestral nodes with consistent sub-trees
 402 between these two trees. In this way, we got approximate existing time for ancestral nodes
 403 where 20 or more families originated. Finally, we dated 23 such ancestral nodes and got the
 404 origin time of 9,581 (total: 9,767) gene families.

Gene family duplication preservation rate

$$= \frac{\text{The number of duplications happened on the ancestral genomes}}{\text{The approximate origin time of this family}}$$

405 4.6.3 Duplication overlap between two branches on species tree

406 For each ancestral branch, we got a gene family set consisting of gene families expanded
 407 at this branch, and we labeled this set as D_i . In this work, we defined a measure named
 408 ‘duplication overlap’ to describe the overlap rate of expanded gene families between two
 409 branches on the species tree.

$$\text{Duplication overlap between branch a and b} = \text{Intersection}(a, b) = \frac{D_a \cap D_b}{D_a \cup D_b}$$

410

411 **List of abbreviations**

412 WGD: Whole Genome Duplication

413 TS: Teleost-Specific

414 DL: Duplication-Loss

415 PPI: Protein-Protein Interaction

416 GO: Gene Ontology

417 CDS: Sequence coding for amino acids in protein

418 MSA: multiple sequence alignment

419 CV Coefficient of Variance

420 ML maximum likelihood

421 **Declarations**

422 **Ethics approval and consent to participate**

423 Not applicable.

424 **Consent for publication**

425 Not applicable.

426 **Acknowledgements**

427 We thank HaiLing Fang and Bing Liu for their assistance in data preparation and figure
428 modification.

429 **Authors' contributions**

430 KL conceived of this project and improved the manuscript. JS designed the experiment,
431 performed the analysis and wrote the manuscript. XH downloaded the data and performed
432 some analysis. All authors read and approved the final manuscript.

433 **Competing interests**

434 The authors have declared no competing interests.

435 **Funding**

436 This work was supported by the State Key Basic Research and Development Plan
437 (2017YFA0605104) and a project of the State Key Laboratory of Earth Surface Processes and
438 Resource Ecology.

439 **Availability of data and materials**

440 Data in the study and parameter-learning related scripts are freely available via the website
441 <http://cmb.bnu.edu.cn/69vertebrates/>.

442

443 **References**

- 444 1. Pagani I, Liolios K, Jansson J, Chen IMA, Smirnova T, Nosrat B, Markowitz
445 VM, Kyrpides NC: **The Genomes OnLine Database (GOLD) v.4: status of**
446 **genomic and metagenomic projects and their associated metadata.**
447 *Nucleic Acids Research* 2012, **40**(D1):D571-D579.
- 448 2. Hardison RC: **Comparative genomics.** *Plos Biology* 2003, **1**(2):156-160.
- 449 3. Bundalovic-Torma C, Parkinson J: **Comparative Genomics and**
450 **Evolutionary Modularity of Prokaryotes.** *Advances in Experimental*
451 *Medicine and Biology* 2015, **883**:77-96.
- 452 4. Eisen JA: **Phylogenomics: Improving functional predictions for**
453 **uncharacterized genes by evolutionary analysis.** *Genome Research*
454 1998, **8**(3):163-167.
- 455 5. Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the**
456 **reconstruction of the tree of life.** *Nature Reviews Genetics* 2005,
457 **6**(5):361-375.

- 458 6. Kumar S, Filipski AJ, Battistuzzi FU, Pond SLK, Tamura K: **Statistics and**
459 **Truth in Phylogenomics.** *Molecular Biology and Evolution* 2012,
460 **29(2):457-472.**
- 461 7. Mi H, Muruganujan A, Casagrande JT, Thomas PD: **Large-scale gene**
462 **function analysis with the PANTHER classification system.** *Nature*
463 *Protocols* 2013, **8(8):1551-1566.**
- 464 8. Fang G, Bhardwaj N, Robilotto R, Gerstein MB: **Getting Started in Gene**
465 **Orthology and Functional Analysis.** *Plos Computational Biology* 2010,
466 **6(3).**
- 467 9. Lee DA, Rentzsch R, Orengo C: **GeMMA: functional subfamily**
468 **classification within superfamilies of predicted protein structural**
469 **domains.** *Nucleic Acids Research* 2010, **38(3):720-737.**
- 470 10. Rappoport N, Karsenty S, Stern A, Linial N, Linial M: **ProtoNet 6.0:**
471 **organizing 10 million protein sequences in a compact hierarchical**
472 **family tree.** *Nucleic Acids Research* 2012, **40(D1):D313-D320.**
- 473 11. Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R: **Gene**
474 **Duplicability of Core Genes Is Highly Consistent across All**
475 **Angiosperms.** *Plant Cell* 2016, **28(2):326-344.**
- 476 12. Morel G, Sterck L, Swennen D, Marcet-Houben M, Onesime D, Levasseur A,
477 Jacques N, Mallet S, Couloux A, Labadie K *et al*: **Differential gene**
478 **retention as an evolutionary mechanism to generate biodiversity**
479 **and adaptation in yeasts (vol 5, 11571, 2015).** *Scientific Reports* 2015,
480 **5.**
- 481 13. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y:
482 **The gain and loss of genes during 600 million years of vertebrate**
483 **evolution.** *Genome Biology* 2006, **7(5):R43.**
- 484 14. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC,
485 Nabholz B, Howard JT *et al*: **Whole-genome analyses resolve early**
486 **branches in the tree of life of modern birds.** *Science* 2014,
487 **346(6215):1320-1331.**
- 488 15. Capella-Gutierrez S, Marcet-Houben M, Gabaldon T: **Phylogenomics**
489 **supports microsporidia as the earliest diverging clade of sequenced**
490 **fungi.** *BMC Biology* 2012, **10(1):47.**

- 491 16. Ascencio D, Ochoa S, Delaye L, DeLuna A: **Increased rates of protein**
492 **evolution and asymmetric deceleration after the whole-genome**
493 **duplication in yeasts.** *BMC Evolutionary Biology* 2017, **17**(1):40.
- 494 17. Marcet-Houben M, Gabaldon T: **Beyond the Whole-Genome Duplication:**
495 **Phylogenetic Evidence for an Ancient Interspecies Hybridization in**
496 **the Baker's Yeast Lineage.** *Plos Biology* 2015, **13**(8).
- 497 18. Ambreen S, Khalil F, Abbasi AA: **Integrating large-scale phylogenetic**
498 **datasets to dissect the ancient evolutionary history of vertebrate**
499 **genome.** *Molecular Phylogenetics and Evolution* 2014, **78**:1-13.
- 500 19. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE,
501 Tomsho LP, Hu Y, Liang H, Soltis PS *et al*: **Ancestral polyploidy in seed**
502 **plants and angiosperms.** *Nature* 2011, **473**(7345):97-U113.
- 503 20. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E:
504 **EnsemblCompara GeneTrees: Complete, duplication-aware**
505 **phylogenetic trees in vertebrates.** *Genome Research* 2009,
506 **19**(2):327-335.
- 507 21. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Marcet-Houben M,
508 Gabaldon T: **PhylomeDB v4: zooming into the plurality of**
509 **evolutionary histories of a genome.** *Nucleic Acids Research* 2014,
510 **42**(D1):D897-D902.
- 511 22. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD:
512 **PANTHER version 11: expanded annotation data from Gene Ontology**
513 **and Reactome pathways, and data analysis tool enhancements.**
514 *Nucleic Acids Research* 2017, **45**(D1):D183-D189.
- 515 23. Boeckmann B, Robinson-Rechavi M, Xenarios I, Dessimoz C: **Conceptual**
516 **framework and pilot study to benchmark phylogenomic databases**
517 **based on reference gene trees.** *Briefings in Bioinformatics* 2011,
518 **12**(5):423-435.
- 519 24. Noutahi E, Semeria M, Lafond M, Seguin J, Boussau B, Gueguen L,
520 El-Mabrouk N, Tannier E: **Efficient Gene Tree Correction Guided by**
521 **Genome Evolution.** *Plos One* 2016, **11**(8):e0159559.
- 522 25. Szoellosi GJ, Tannier E, Daubin V, Boussau B: **The Inference of Gene**
523 **Trees with Species Trees.** *Systematic Biology* 2015, **64**(1):E42-E62.

- 524 26. Emms DM, Kelly S: **OrthoFinder: solving fundamental biases in whole**
525 **genome comparisons dramatically improves orthogroup inference**
526 **accuracy**. *Genome Biology* 2015, **16**:157.
- 527 27. Bouckaert R, Heled J, Kuehnert D, Vaughan T, Wu C-H, Xie D, Suchard MA,
528 Rambaut A, Drummond AJ: **BEAST 2: A Software Platform for Bayesian**
529 **Evolutionary Analysis**. *Plos Computational Biology* 2014,
530 **10**(4):e1003537.
- 531 28. De Oliveira Martins L, Mallo D, Posada D: **A Bayesian Supertree Model**
532 **for Genome-Wide Species Tree Reconstruction**. *Systematic Biology*
533 2016, **65**(3):397-416.
- 534 29. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and**
535 **post-analysis of large phylogenies**. *Bioinformatics* 2014,
536 **30**(9):1312-1313.
- 537 30. Chen K, Durand D, Farach-Colton M: **NOTUNG: A program for dating**
538 **gene duplications and optimizing gene family trees**. *Journal of*
539 *Computational Biology* 2000, **7**(3-4):429-447.
- 540 31. Lam-Tung N, Schmidt HA, von Haeseler A, Bui Quang M: **IQ-TREE: A Fast**
541 **and Effective Stochastic Algorithm for Estimating**
542 **Maximum-Likelihood Phylogenies**. *Molecular Biology and Evolution*
543 2015, **32**(1):268-274.
- 544 32. DeBlasio DF, Wisecaver JH: **SICLE: a high-throughput tool for extracting**
545 **evolutionary relationships from phylogenetic trees**. *Peerj* 2016,
546 **4**:e2359.
- 547 33. Lynch M, Conery JS: **The evolutionary fate and consequences of**
548 **duplicate genes**. *Science* 2000, **290**(5494):1151-1155.
- 549 34. S O: **Evolution by Gene Duplication**. *Springer-Verlag* 1970.
- 550 35. Kondrashov FA, Kondrashov AS: **Role of selection in fixation of gene**
551 **duplications**. *Journal of Theoretical Biology* 2006, **239**(2):141-151.
- 552 36. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D,
553 Cummins C, Clapham P, Fitzgerald S, Gil L *et al*: **Ensembl 2016**. *Nucleic*
554 *Acids Research* 2016, **44**(D1):D710-D716.
- 555 37. Bansal MS, Alm EJ, Kellis M: **Reconciliation Revisited: Handling**
556 **Multiple Optima when Reconciling with Duplication, Transfer, and**

- 557 **Loss. *Journal of Computational Biology* 2013, 20(10):738-754.**
- 558 38. Song S, Liu L, Edwards SV, Wu S: **Resolving conflict in eutherian**
- 559 **mammal phylogeny using phylogenomics and the multispecies**
- 560 **coalescent model (vol 109, pg 14942, 2012).** *Proceedings of the*
- 561 *National Academy of Sciences of the United States of America* 2015,
- 562 **112(44):E6079-E6079.**
- 563 39. Foley NM, Springer MS, Teeling EC: **Mammal madness: is the mammal**
- 564 **tree of life not yet resolved?** *Philosophical Transactions of the Royal*
- 565 *Society B-Biological Sciences* 2016, **371(1699):pii: 20150140.**
- 566 40. Boussau B, Szoellosi GJ, Duret L, Gouy M, Tannier E, Daubin V:
- 567 **Genome-scale coestimation of species and gene trees. *Genome***
- 568 *Research* 2013, **23(2):323-330.**
- 569 41. Roginski RS, Raj BKM, Birditt B, Rowen L: **The human GRINL1A gene**
- 570 **defines a complex transcription unit, an unusual form of gene**
- 571 **organization in eukaryotes. *Genomics* 2004, 84(2):265-276.**
- 572 42. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J,
- 573 Alioto T, Manzano C, Chrast J *et al*: **Prominent use of distal 5 '**
- 574 **transcription start sites and discovery of a large number of**
- 575 **additional exons in ENCODE regions. *Genome Research* 2007,**
- 576 **17(6):746-759.**
- 577 43. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek
- 578 R: **Transcription-mediated gene fusion in the human genome. *Genome***
- 579 *Research* 2006, **16(1):30-36.**
- 580 44. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F,
- 581 Aken BL, Barrell D, Zadissa A, Searle S *et al*: **GENCODE: The reference**
- 582 **human genome annotation for The ENCODE Project. *Genome Research***
- 583 **2012, 22(9):1760-1774.**
- 584 45. Fuerst R: **EVOLUTION BY GENE DUPLICATION - OHNO,S. *Social Biology***
- 585 **1972, 19(1):89-90.**
- 586 46. Dehal P, Boore JL: **Two rounds of whole genome duplication in the**
- 587 **ancestral vertebrate. *Plos Biology* 2005, 3(10):1700-1708.**
- 588 47. Panopoulou G, Poustka AJ: **Timing and mechanism of ancient**
- 589 **vertebrate genome duplications - the adventure of a hypothesis.**

- 590 *Trends in Genetics* 2005, **21**(10):559-567.
- 591 48. Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y: **Major events**
592 **in the genome evolution of vertebrates: Paraneome age and size differ**
593 **considerably between ray-finned fishes and land vertebrates.**
594 *Proceedings of the National Academy of Sciences of the United States of*
595 *America* 2004, **101**(6):1638-1643.
- 596 49. Pasquier J, Cabau C, Thaovi N, Jouanno E, Severac D, Braasch I, Journot L,
597 Pontarotti P, Klopp C, Postlethwait JH *et al*: **Gene evolution and gene**
598 **expression after whole genome duplication in fish: the PhyloFish**
599 **database.** *Bmc Genomics* 2016, **17**.
- 600 50. Smith JJ, Keinath MC: **The sea lamprey meiotic map improves**
601 **resolution of ancient vertebrate genome duplications.** *Genome*
602 *Research* 2015, **25**(8):1081-1090.
- 603 51. Brunet FG, Crollius HR, Paris M, Aury J-M, Gibert P, Jaillon O, Laudet V,
604 Robinson-Rechavi M: **Gene loss and evolutionary rates following**
605 **whole-genome duplication in teleost fishes.** *Molecular Biology and*
606 *Evolution* 2006, **23**(9):1808-1816.
- 607 52. Nakatani Y, Takeda H, Kohara Y, Morishita S: **Reconstruction of the**
608 **vertebrate ancestral genome reveals dynamic genome**
609 **reorganization in early vertebrates.** *Genome Research* 2007,
610 **17**(9):1254-1265.
- 611 53. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T,
612 Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K *et al*: **The amphioxus**
613 **genome and the evolution of the chordate karyotype.** *Nature* 2008,
614 **453**(7198):1064-U1063.
- 615 54. Saitou N, Nei M: **THE NEIGHBOR-JOINING METHOD - A NEW METHOD**
616 **FOR RECONSTRUCTING PHYLOGENETIC TREES.** *Molecular Biology and*
617 *Evolution* 1987, **4**(4):406-425.
- 618 55. Makino T, McLysaght A: **Ohnologs in the human genome are dosage**
619 **balanced and frequently associated with disease.** *Proceedings of the*
620 *National Academy of Sciences of the United States of America* 2010,
621 **107**(20):9270-9274.
- 622 56. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de

- 623 Peer Y: **Modeling gene and genome duplications in eukaryotes.**
624 *Proceedings of the National Academy of Sciences of the United States of*
625 *America* 2005, **102**(15):5454-5459.
- 626 57. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y:
627 **The gain and loss of genes during 600 million years of vertebrate**
628 **evolution.** *Genome Biology* 2006, **7**(5).
- 629 58. Freeling M, Thomas BC: **Gene-balanced duplications, like tetraploidy,**
630 **provide predictable drive to increase morphological complexity.**
631 *Genome Research* 2006, **16**(7):805-814.
- 632 59. Semon M, Wolfe KH: **Consequences of genome duplication.** *Current*
633 *Opinion in Genetics & Development* 2007, **17**(6):505-512.
- 634 60. Holland LZ: **Evolution of new characters after whole genome**
635 **duplications: Insights from amphioxus.** *Seminars in Cell &*
636 *Developmental Biology* 2013, **24**(2):101-109.
- 637 61. Roux J, Liu J, Robinson-Rechavi M: **Selective Constraints on Coding**
638 **Sequences of Nervous System Genes Are a Major Determinant of**
639 **Duplicate Gene Retention in Vertebrates.** *Molecular Biology and*
640 *Evolution* 2017, **34**(11):2773-2791.
- 641 62. Canestro C, Albalat R, Irimia M, Garcia-Fernandez J: **Impact of gene**
642 **gains, losses and duplication modes on the origin and diversification**
643 **of vertebrates.** *Seminars in Cell & Developmental Biology* 2013,
644 **24**(2):83-94.
- 645 63. Innan H, Kondrashov F: **The evolution of gene duplications: classifying**
646 **and distinguishing between models.** *Nature Reviews Genetics* 2010,
647 **11**(2):97-108.
- 648 64. Levasseur A, Pontarotti P: **The role of duplications in the evolution of**
649 **genomes highlights the need for evolutionary-based approaches in**
650 **comparative genomics.** *Biology Direct* 2011, **6**(1):11.
- 651 65. Katoh K, Standley DM: **MAFFT Multiple Sequence Alignment Software**
652 **Version 7: Improvements in Performance and Usability.** *Molecular*
653 *Biology and Evolution* 2013, **30**(4):772-780.
- 654 66. Abascal F, Zardoya R, Telford MJ: **TranslatorX: multiple alignment of**
655 **nucleotide sequences guided by amino acid translations.** *Nucleic Acids*

656 *Research* 2010, **38**:W7-W13.

657 67. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T: **trimAl: a tool for**

658 **automated alignment trimming in large-scale phylogenetic analyses.**

659 *Bioinformatics* 2009, **25**(15):1972-1973.

660 68. Felsenstein J: **PHYLOGENIES FROM MOLECULAR SEQUENCES -**

661 **INFERENCE AND RELIABILITY.** *Annual Review of Genetics* 1988,

662 **22**:521-565.

663 69. Berard S, Gallien C, Boussau B, Szollosi GJ, Daubin V, Tannier E: **Evolution**

664 **of gene neighborhoods within reconciled phylogenies.** *Bioinformatics*

665 2012, **28**(18):I382-I388.

666 70. Yu G, Wang L-G, Han Y, He Q-Y: **clusterProfiler: an R Package for**

667 **Comparing Biological Themes Among Gene Clusters.** *Omics-a Journal*

668 *of Integrative Biology* 2012, **16**(5):284-287.

669 71. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S: **Tree of Life Reveals**

670 **Clock-Like Speciation and Diversification.** *Molecular Biology and*

671 *Evolution* 2015, **32**(4):835-845.

672

673 **Figure Legends**

674 **Figure 1 Flowchart illustrating our workflow**

675 Our workflow mainly consists of four processes. The third and forth are the most important
676 processes in our workflow. The inputs are displayed in green rectangles. The intermediate
677 results are displayed in red rectangles while the final results are displayed in blue rectangles.
678 The software, operation and some parameters used in this workflow are marked on the arrows
679 in grey, blue and black font respectively.

680 **Figure 2 Final species tree**

681 The common names of species are displayed in parentheses following the Latin names. And
682 the common name of the common species between this species tree and the mammals species
683 tree published by Song *et al.* 2012[38] are in blue font.

684 **Figure 3 Comparison between our workflow and EnsemblCompara**

685 ‘our workflow 1’ represents the standard workflow we have described in Materials and
686 Methods section and ‘our workflow 2’ represents the same workflow but without
687 parameter-learning in duplication/loss dating. **a.** Ancestral chromosome linearity metric.
688 Extant 1 represents the genes neighborhoods status on extant genomes based on our 9,767
689 core gene families. Extant 2 represents the genes neighborhoods status on extant genomes
690 based on Ensembl 8,514 gene families. The rest three represent the genes neighborhoods
691 status on ancestral genomes inferred from different phylogenomics results. **b.** Duplication
692 consistency score.

693 **Figure 4 WGD-affected gene family classes and related gene function**

694 **a .** Intersection among the three WGDs. Gene families with ohnologs retention are highly
695 overlapped among the tree WGDs. We divided these ohnologs retention gene families into
696 seven classes. 1R represents the first round WGD occurred on vertebrate genomes. 2R
697 represents the second round WGD occurred on vertebrate genomes. TS represents the teleost
698 fish specific WGD. **b.** Enriched functional categories comparison among the seven classes.
699 The ‘A^B’ represents the intersection of ‘A’ and ‘B’. **c.** The biological processes enrichment
700 results of class 7 which consist of gene families with ohnologs retention after TS WGD only.
701 This analysis conducted based on gene ontology data of zebra fish.

702 **Figure 5 GO enrichment results of gene families with long-term local duplications** 703 **retention**

704 The oxygen levels response related biological processes are labeled in red font.

705 **Tables**

Table 1 Average genome sizes comparison

Pipelines	Ancestral Genomes ^c	Extant Genomes ^d	Ancestral/extant Ratio ^e
Our pipeline ^a	10290.25	9994.07	1.03
Our pipeline ^b	10457.22	9994.07	1.05
EnsemblCompara	22389.23529	13329.38	1.68

^aOur standard workflow according to the processes as we described in Materials and Methods.

^bInstead of the optimal parameters with default parameters (1.5,1) when reconciliation by Notung in our workflow.

^cAverage ancestral genomes size according to our core data/Ensembl 8,514 gene families.

^dAverage extant genomes size according to our core data/Ensembl 8,514 gene families.

^eRatio of average ancestral genomes size and average extant genomes size.

Supplementary materials

additional file 1: supplemental figures and tables.

additional file 2: supplemental methods and materials.

Gene families identification

All protein and CDS sequences from 69 species

One protein sequence as representation for each gene

orthomclFilterFasta + OrthoFinder

54,808 gene families

Gene trees inference

11,698 gene trees

RAxML

11,698 MSAs (CDS)

TranslatorX

11,698 MSAs (protein)

Mafft + trimAl

11,698 gene families

Selected

Selected

Species tree inference

SICe

Revised species tree

Incongruent relationship set

Guenomu

Species tree 1

Posterior distributions of 527 random selected gene family trees

Mafft + trimAl + TranslatorX + BEAST

Exist Information of species relationship: Species tree 2

Gene family tree correction in duplication/loss events dating

The gene tree for each gene family + the species tree + a parameter pair (costdup, costloss)

15 parameters (costdup, costloss) pairs

Notung (rearrange)

Gene trees with minimum reconcile scores

iqtree

The gene tree with maximal likelihood based on CDS MSA

Notung (root)

Rooted Gene tree

Notung (reconcile)

Reconciled gene trees

The optimum event-costs pairs
The consistent gene family tree and
duplication/loss events annotation

The corrected gene trees with maximal likelihood based on CDS MSA

The corrected gene trees with minimum reconcile events

iqtree

15 reconciled gene trees

Vertebrates

Mammals

A)

B)

Biological processes