Phylogenetic analyses of human 1/2/8/20 paralogs suggest segmental duplications during animal evolution

Farhan Haq¹ · Usman Saeed² · Rida Khalid¹ · Muhammad Qasim³ · Maryam Mehmood¹

Received: 25 September 2017 / Accepted: 16 May 2019 / Published online: 24 May 2019
© King Abdulaziz City for Science and Technology 2019

Abstract

Susumu Ohno hypothesized that the diversity of vertebrate gene families and genome is due to two rounds of whole genome duplications (also referred as 2R hypothesis). The quadruplicate paralogous blocks present on 1/2/8/20 chromosomes are taken as one of the evidences in favor of the 2R. In this study, we investigated that whether 2R has shaped the vertebrate evolution using gene families residing on chromosomes 1/2/8/20. Evolutionary history of 22 gene families (11 from the current study and 11 from the previous study) was evaluated by the phylogenetic analysis with triplicated or quadruplicated distribution on these human chromosomes 1/2/8/20. The phylogenetic analysis was performed using high-quality whole genomic sequence data of multiple species with neighbor-joining (NJ) and maximum likelihood (ML) methods. The phylogenetic tree topology of these gene families revealed variable duplication time points during invertebrate–vertebrate evolution. Topology comparison approach categorized 22 gene families into three groups. Tree topologies of ten gene families fell into Group 1 (duplications prior to invertebrate–vertebrate split), four in Group 2 (i.e., (AB) (C) (D), topology incongruent with 2R) and eight in Group 3 (((AB) (CD)), 2R congruent topology). Therefore, taken together the current and previous data of 1/2/8/20 paralogs, we propose that, in addition to whole genome duplications events, current developmental, morphological and genomic complexity of the vertebrate genomes may also have originated through segmental duplications occurring at varying time points during the course of animal evolution.

Keywords Phylogenetics · 2R hypothesis · Segmental duplications · Paralogous blocks · Tree topology

Introduction

The exciting findings of developmental and morphological complexities to the elucidation of complex cellular processes (e.g. immune system), all can be best described in the light of the evolutionary history of genomes (Gregory 2002; Meyer and Van de Peer 2005). In past decades, whole

genome duplications (WGDs) were considered the key players bringing increased genome complexity (in terms of its size, organization, and function) from invertebrates to vertebrates (Ohno 1970, 1973; Zhang 2003). According to WGD concept, the current day complexity of multiple gene families arose from two rounds of WGD events which occurred at the time of vertebrate origin (i.e., prior to fish–tetrapods split), famously known as a 2R hypothesis (Kasahara 2007; Wolfe 2001). Later, third round of WGD (3R hypothesis) was also reported specifically in fish lineage (Meyer and Van de Peer 2005).

Though controversial, many scientists were in view of polyploidization in the early vertebrate lineage (Abi-Rached et al. 2002; Furlong and Holland 2002; Kasahara 2007; McLysaght et al. 2002; Meyer and Van de Peer 2005; Wolfe 2001). This 2R hypothesis, proposed by Ohno, is mainly based on genome content, chromosome topologies and the high frequency of tetraploidy in fish and amphibians (Abbasi 2008; Gibson and Spring 2000). The presence of four HOX gene clusters on four fish chromosomes while one HOX gene cluster on

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-019-1768-7>) contains supplementary material, which is available to authorized users.

✉ Farhan Haq
farhan.haq@comsats.edu.pk

¹ Department of Biosciences, COMSATS University Islamabad, Park Road, Chak Shehzad, Islamabad, Pakistan

² Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Munich, Germany

³ Ajou University School of Medicine, Suwon, South Korea

the invertebrate chromosome was provided as an evidence in favor of 2R hypothesis (Martin 2001). Of note, in support of this argument, adjacent gene families present on the four HOX-bearing chromosomes (i.e., human chromosomes 2, 7, 12, and 17) were also considered to be evolved in the same manner along with HOX gene clusters (Larhammar et al. 2002; Sundström et al. 2008). Overall, phylogenetic studies on human chromosomes including 1/6/9/19, 4/5/8/10, 1/2/8/20 and 2/7/12/17 are taken as a supporting evidence in favor of two rounds of polyploidization events (Hokamp et al. 2003; Lundin et al. 2003; Sundström et al. 2008).

Interestingly, later studies opened new challenges to the evidences given in favor of WGDs in early vertebrate evolution (Furlong and Holland 2002). The detailed genomic analyses in these studies using diverse set of invertebrate and vertebrate species opposed the 2R hypothesis (i.e., variable time points of duplication events in different gene families and inconsistent tree topologies) (Abbasi and Grzeschik 2007; Ajmal et al. 2014; Asrar et al. 2013; Hafeez et al. 2016). In this context, a recent study based on map self-comparison approach and phylogenetic analysis revealed small-scale duplication events at different invertebrate–vertebrate time points using quadruplicate paralogous blocks present on human chromosomes 1/2/8/20 (Abbasi and Hanif 2012). Additionally, in our recent study, we demonstrated that multiple gene families present on HOX-bearing chromosomes 2/7/12/17 showed ancient segmental duplications and rearrangements as an evolutionary events producing triplicated and quadruplicated paralogous blocks (Asrar et al. 2013). Based on these observations, segmental duplications and rearrangements are also seeking attention to explain the evolutionary changes that took place at different time points in shaping up of vertebrate genomes (Abbasi 2008). Although according to some researchers, there is incontrovertible evidence supporting the two rounds of whole genome duplication (Kasahara 2007).

In the light of the previous study, using quadruplicate paralogous blocks present on human chromosomes 1/2/8/20, we further inspected small-scale segmental duplications as an additional model of vertebrate evolution. To achieve this goal, a detailed phylogenetic analysis of 11 gene families was performed using high-quality whole genomic sequence data of multiple species with neighbor-joining (NJ) and maximum likelihood (ML) methods (Saitou and Nei 1987; Whelan and Goldman 2001).

Methodology

Extraction of dataset

We performed a comprehensive phylogenetic study of gene families present at chromosomes 1(p36–p34)/9q34,

2p24/6(q21–q23)/8p13, 8(q12–q24)/18p11 and 20(q11–q13). The gene families and their protein sequences were collected after scanning the genomes present at UCSC (<https://genome.ucsc.edu/>) and ENSEMBL genome browsers (<http://asia.ensembl.org/index.html>). We identified 11 genes families present on these four paralogous blocks (Table 1). Out of these four blocks, three gene families were present on all four chromosomes while eight were present on three out of four paralogous blocks (Table 1). The list species selected for the analyses are available in supplementary information (ESM_3).

Identification of orthologs

The identification of all possible putative orthologs for human protein sequences of identified gene families (Table 1) was done using BlastP available at Ensembl genome browser. The orthologous sequences which were not available at Ensembl genome browser were searched in NCBI protein database (<http://www.ncbi.nlm.nih.gov/protein>). The invertebrate sequences which showed higher sequence identity and score compared to other ancestral invertebrate sequences were retained since our major goal was to identify duplication events occurred in vertebrate evolution. Furthermore, the verification of the ancestral orthologous sequences was done with phylogenetic analysis.

Phylogenetic analysis

Multiple sequence alignment of amino acid sequences was done using CLUSTALW (Thompson et al. 1994). Molecular Evolutionary Genetic Analysis (MEGA) v 6.06 tool was used for evolutionary analyses (Tamura et al. 2013). Neighbor-Joining (NJ) method was used for constructing phylogenetic trees. P-distance model for amino acid substitution implemented in MEGA was selected. Complete deletion option was selected to treat gaps in the alignments. The sequences which disrupted the whole phylogenetic trees were excluded from the analysis. Bootstrapping of the trees was done with 1000 replicates (Felsenstein 1985). Furthermore, to validate the phylogenetic trees generated by NJ, maximum likelihood method with Whelan and Goldman (WAG) model and gamma distributed with invariant sites (G+I) was used (Yang 1994). The results are presented in the electronic supplementary information.

Results

The 11 selected gene families present at 1(p36–p34)/9q34, 2p24/6(q21–q23)/8p13, 8(q12–q24)/18p11 and 20(q11–q13) chromosomal regions include collagen protein family (*COL*), disks large (*Drosophila*)

Table 1 List of gene families used in the phylogenetic analysis

Gene members	Chr. location	Human protein accession no
Collagens		
COL16A1	1p35.2	Q07092
COL22A1	8q24.3	Q8NFW1
COL9A2	1p34.2	Q14055
COL9A3	20q13.33	Q14050
Disks, large (Drosophila) homolog-associated protein		
DLGAP1	18p11.31	O14490
DLGAP2	8p23.3	
DLGAP3	1p34.3	O95886
DLGAP4	20q11.23	Q9Y2H0
Erythrocyte membrane protein band 4.1		
EPB41	1p35.3	Q4VB86
EPB41L1	20q11.23	Q9H4G0
EPB41L2	6q23.2	O43491
EPB41L3	18p11.31	Q9Y2J2
Mitochondrial fission regulator		
MTFR1	8q13.1	Q15390
MTFR1L	1p36.11	Q9H019
MTFR2	6q23.3	Q6P444
Na ⁺ /K ⁺ + transporting atpase interacting		
NKAIN1	1p35.2	Q4KMZ8
NKAIN2	6q22.31	Q5VXU1
NKAIN3	8q12.3	Q8N8D7
NKAIN4	20q13.33	Q8IVV8
Regulating synaptic membrane exocytosis		
RIMS2	8q22.3	
RIMS3	1p34.2	Q9UJD0
RIMS4	20q13.12	Q9H426
R-spondin		
RSPO1	1p34.3	Q2MKA7
RSPO2	8q23.1	Q6UXX9
RSPO3	6q22.33	Q9BXY4
Serine incorporator		
SERINC1	6q22.31	Q9NRX5
SERINC2	1p35.2	Q96SA4
SERINC3	20q13.12	Q13530
Stratifin		
SFN	1p36.11	P31947
YWHAB	20q13.12	P31946
YWHAZ	8q22.3	P63104
Transmembrane protein 200		
TMEM200A	6q23.1	Q86VY9
TMEM200B	1p35.3	Q69YZ2
TMEM200C	18p11.31	A6NKL6
YTH N(6)-methyladenosine RNA-binding protein		
YTHDF1	20q13.33	Q9BYJ9
YTHDF2	1p35.3	Q9Y5A9
YTHDF3	8q12.3	Q7Z739

homolog-associated protein family (*DLGAP*), erythrocyte membrane protein band family (*EPB*), mitochondrial fission regulator protein family (*MTFR*), Na⁺/K⁺ transporting ATPase-interacting protein family (*NKAIN*), regulating synaptic membrane exocytosis protein family (*RIMS*), R-spondin protein family (*RSPO*), serine incorporator protein family (*SERINC*), transmembrane protein 200 family (*TMEM200*), YTH N(6)-methyladenosine RNA-binding protein family (*YTHDF*) and stratifin (*SFN*) (Table 1). These gene families are present in at least three out of four quadruplicate paralogous blocks (Fig. 1). As discussed in our previous study (Asrar et al. 2013), the evolutionary history of these functionally important gene families was investigated by constructing phylogenetic trees and applying a topology comparison approach.

Phylogenetic analysis and topology comparison approach

Initially, phylogenetic trees of each gene family were generated using neighbor-joining (NJ) method (Fig. 1; Electronic Supplementary Material ESM_1). After checking the reliability of NJ trees using bootstrap method (1000 pseudoreplicates), the phylogenetic trees were also validated using maximum likelihood (ML) method (Electronic Supplementary Material ESM_2). Furthermore, the bootstrap of 1000 pseudoreplicates was applied on trees generated using ML method. We found that the phylogenies generated by NJ and ML methods reconciled well with each other. The tree topology of each gene family was constructed using phylogenetic trees (Fig. 3). According to the topologies, phylogenetic trees of *NKAIN*, *RIMS*, *TMEM200*, *YTHDF* and *DLGAP* gene families were rooted with the orthologous genes from

Fig. 1 The representation of gene families present on at least three of the four paralogous blocks on chromosomes 1, 2, 8 and 20. The gene families analyzed in this study including collagen protein family (COL), disks large (Drosophila) homolog-associated protein family (DLGAP), erythrocyte membrane protein band family (EPB), mitochondrial fission regulator protein family (MTFR), Na⁺/K⁺ transporting ATPase-interacting protein family (NKAIN), regulating synaptic membrane exocytosis protein family (RIMS), R-spondin protein family (RSPO), serine incorporator protein family (SER-

INC), transmembrane protein 200 family (TMEM200), YTH N(6)-methyladenosine RNA-binding protein family (YTHDF) and tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein family (YWHA) are enclosed in rectangles. The rest of the gene families mentioned are analyzed in two previous studies (Abbasi and Hanif 2012; Gibson and Spring 2000). The putative break points are represented with the symbol 'X'. None of the features of this figure are drawn to scale

invertebrate species (Fig. 3a, b), whereas *COL*, *RSPO*, *EPB41*, *SFN*, *MTFR* and *SERINC* gene families were rooted with the orthologous genes from both invertebrates and vertebrate species (Fig. 3c). The phylogenetic analysis for *COL* and *EPB41* gene families suggested that they diversified by three duplication events which occurred mainly during invertebrate lineage; two of these duplications occurred prior to invertebrate–vertebrate split (ESM_1B and ESM_1C). Tree topologies for *SFN*, *MTFR* and *SERINC* gene families suggested that they diversified by two duplication events; one occurred before and one after the invertebrate–vertebrate split (Fig. 2; ESM_1F and ESM_1G). The analysis for *RSPO* gene family suggested that it diversified by two duplication events and both of these duplication events occurred prior to invertebrate–vertebrate split (ESM_1E). The duplication events in the rest of five gene families including *NKAIN*, *RIMS*, *TMEM200*, *YTHDF* and *DLGAP* occurred within a vertebrate lineage (Fig. 2; ESM_1A, ESM_1D, ESM_1H and ESM_1I).

As reported in our previous finding (Asrar et al. 2013), the gene families are divided into three major groups based

on their tree topologies (Fig. 3). Group 1 includes *COL*, *RSPO*, *EPB41*, *SFN*, *MTFR* and *SERINC* gene families, in which few duplication events occurred before the vertebrate origin (Fig. 3a). Group 2 includes *DLGAP* gene family, in which duplication events occurred during chordate lineage but showed incongruent topology (i.e., (AB)(C)(D)) (Fig. 3b). Group 3 includes *NKAIN*, *RIMS*, *TMEM200* and *YTHDF* gene families which duplicated within a vertebrate lineage and showed congruent topologies (i.e., (AB) (CD)) (Fig. 3c). In group 3, the congruent topologies of *RIMS*, *TMEM200* and *YTHDF* gene families can be explained by three gene loss events.

A similar study with 11 gene families (including *EYA*, *SDC*, *HCK*, *MATN*, *MYC*, *PTPRT*, *SRC*, *KCNS*, *WISP*, *MYC* and *LPTM*) on 1/2/8/20 paralogous chromosomes also evaluated the possibility of WGDs at the time of vertebrate origin (Fig. 1) (Abbasi and Hanif 2012). They also found three distinct groups based on tree topologies. Together with the previous and current study, a total of 22 gene families can be classified into three different topology groups. Out of 22, 10 gene families correspond to the Group 1, 4 out of

Fig. 2 The NJ phylogenetic genetic trees of MTFR and DLGAP gene families. The evolutionary distances were computed using the uncorrected p-distance method. Complete deletion option was used to treat gaps and missing data. The percentage of replicate trees in which

the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches; only values $\geq 50\%$ are presented here. Scale bar shows amino acid substitution per site

Fig. 3 The NJ tree topologies of gene families present on at least three of the four paralogous blocks on chromosomes 1, 2, 8 and 20. The three different tree topologies are ((AB)(CD), concordant with 2R hypothesis; ((AB)(C)(D)), incongruent with 2R hypothesis; and prior to inver-vert split. The term 'Absent' represents a gene loss event in a particular chromosome. The duplications in NKAIN, RIMS, TMEM200 and YTHDF gene families showed (AB)(CD) tree topologies with three gene loss events. The duplications DLGAP gene family showed (AB)(C)(D) tree topology. The duplications in COL, RSPO, EPB41, SFN, MTFR and SERINC gene families occurred prior to inver-vert split

22 correspond to Group 2 and the remaining 8 correspond to Group 3. Both Groups 1 and 2, with 14 of 22 gene families showed duplication history prior to vertebrate-invertebrate split and are not in concordance with the 2R topology (i.e., (AB)(CD)).

Discussion

Long-term evolutionary divergence, rearrangements, and breakage tend to obscure the genome-shaping events occurring in distant vertebrate history (Abbasi 2015). However, genome-wide pairwise comparisons and map self-comparisons provide significant information of the recent vertebrate history (Asrar et al. 2013; Gibson and Spring 2000). Genomic blocks ranging up to 1 mb in size while being present in more than one locations with over 90% sequence identity are characterized as segmentally duplicated blocks. The presence of these duplicated blocks

has been confirmed by multiple researchers (Bailey et al. 2002; Bailey and Eichler 2006; Cheng et al. 2005; Kehrer-Sawatzki and Cooper 2008). It has been estimated that only these primate segmental duplications account up to 5% of the human genome (Bailey and Eichler 2006). The role of these segmental duplications has not only been in creating novel primate genes and shaping primate genome, but also for large-scale hominoid-specific chromosomal rearrangements (Kehrer-Sawatzki and Cooper 2008). Considering the translational events, chromosomal breakage and rearrangement events, it becomes a tedious task to predict the exact evolutionary events in distant vertebrate history using intergenomic and intragenomic map comparisons (Asrar et al. 2013). Phylogenetic analysis with special emphasis on topology comparisons can be used as a more reliable and efficient approach to determine the history of vertebrate paralogs (Abbasi 2008; Hughes 1998, 1999).

Gibson and Spring were first to test the 2R hypothesis using gene families present on 1/2/8/20 chromosomes.

They believed that due to limited genetic information tree topologies were incongruent with 2R topology (Gibson and Spring 2000). Later, Abbasi and Hanif (2012) performed phylogenetic analysis of similar gene families analyzed by Gibson and Spring, with maximum genetic information available. Interestingly, they suggested that incongruence of duplication histories with the 2R pattern is not because of paucity of the data, but because of independent small-scale duplications and rearrangements happening throughout invertebrate–vertebrate lineage (Abbasi and Hanif 2012).

In this study, we extended the work done by Abbasi and Hanif (2012), and tested the hypothesis that whether two rounds of WGD events occurred at the time of vertebrate origin using gene families present on 1/2/8/20 human chromosomes. For that purpose, we examined the phylogenetic trees of 22 gene families located on human chromosomes 1/2/8/20. The tree topologies of the gene families were compared with the topology congruent with 2R (i.e., (AB) (CD)). The detailed comparison led to the classification of duplication histories into three major groups (Fig. 3). The gene families which were classified into single group shared similar duplication history. The identification of these groups indicated that all gene families were not evolved through the same duplication pattern. Of note, the majority of these gene families, 14 out of 22, might have evolved through independent segmental duplications that either occurred before or after the origins of vertebrates (encompassing Groups 1 and 2).

Taken together the current and previous data of 1/2/8/20 paralogs, we suggest that vertebrate genome duplication events have occurred at variable time points during animal evolution. Of note, this is the first study covering maximum number of species and gene families using 1/2/8/20 paralogs. Furthermore, the data suggest that present day developmental, morphological and genomic complexity of the vertebrate genome occurring at varying time points during the course of animal evolution may have originated through different events including independent small-scale duplications/rearrangements and whole genome duplications. These observations are also in harmony with the data showing recent small-scale duplication and rearrangements in different primate species. Therefore, based on the latest biological information available, we propose that the small-scale segmental duplications and rearrangements throughout animal evolution may also be considered a realistic model of animal evolution, in addition to 2R model.

Acknowledgements Special thanks to the teams behind UCSC, NCBI and Ensembl genome browsers, for making all the data publically available for different analyses.

Compliance with ethical standards

Conflict of interest Authors have no conflict of interests.

References

- Abbasi AA (2008) Are we degenerate tetraploids? More genomes, new facts. *Biol Direct* 3:50
- Abbasi AA (2015) Diversification of four human HOX gene clusters by step-wise evolution rather than ancient whole-genome duplications. *Dev Genes Evol* 225:353–357
- Abbasi AA, Grzeschik K-H (2007) An insight into the phylogenetic history of HOX linked gene families in vertebrates. *BMC Evol Biol* 7:239
- Abbasi AA, Hanif H (2012) Phylogenetic history of paralogous gene quartets on human chromosomes 1, 2, 8 and 20 provides no evidence in favor of the vertebrate octoploidy hypothesis. *Mol Phylogenet Evol* 63:922–927
- Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H (2002) Evidence of en bloc duplication in vertebrate genomes. *Nat Genet* 31:100–105
- Ajmal W, Khan H, Abbasi AA (2014) Phylogenetic investigation of human FGFR-bearing paralogs favors piecemeal duplication theory of vertebrate genome evolution. *Mol Phylogenet Evol* 81:49–60
- Asrar Z, Haq F, Abbasi AA (2013) Fourfold paralogy regions on human HOX-bearing chromosomes: role of ancient segmental duplications in the evolution of vertebrate genome. *Mol Phylogenet Evol* 66:737–747
- Bailey JA, Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* 7:552–564
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. *Science* 297:1003–1007
- Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S et al (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437:88–93
- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783
- Furlong RF, Holland PWH (2002) Were vertebrates octoploid? *Philos. Trans R Soc Lond B Biol Sci* 357:531–544
- Gibson TJ, Spring J (2000) Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem Soc, Trans*, p 28
- Gregory TR (2002) Genome size and developmental complexity. *Genetica* 115:131–146
- Hafeez M, Shabbir M, Altaf F, Abbasi AA (2016) Phylogenomic analysis reveals ancient segmental duplications in the human genome. *Mol Phylogenet Evol* 94:95–100
- Hokamp K, McLysaght A, Wolfe KH (2003) The 2R hypothesis and the human genome sequence. *J Struct Funct Genom* 3:95–110
- Hughes AL (1998) Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol Biol Evol* 15:854–870
- Hughes AL (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J Mol Evol* 48:565–576
- Kasahara M (2007) The 2R hypothesis: an update. *Curr Opin Immunol* 19:547–552
- Kehrer-Sawatzki H, Cooper DN (2008) Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol* 16:41–56

- Larhammar D, Lundin L-G, Hallböök F (2002) The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res* 12:1910–1920
- Lundin L-G, Larhammar D, Hallböök F (2003) Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. In: Meyer A, Van de Peer Y (eds) *Genome evolution*. Springer, Netherlands, pp 53–63
- Martin A (2001) Is tetralogy true? Lack of support for the “one-to-four rule”. *Mol Biol Evol* 18:89–93
- McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. *Nat Genet* 31:200–204
- Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays* 27:937–945
- Ohno DS (1970) Introduction in evolution by gene duplication. Springer, Berlin, pp 1–2
- Ohno S (1973) Ancient linkage groups and frozen accidents. *Nature* 244:259–262
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Sundström G, Larsson TA, Larhammar D (2008) Phylogenetic and chromosomal analyses of multiple gene families syntenic with vertebrate Hox clusters. *BMC Evol Biol* 8:254
- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 30:2725–2729
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680
- Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691–699
- Wolfe KH (2001) Yesterday’s polyploids and the mystery of diploidization. *Nat Rev Genet* 2:333–341
- Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 39:306–314
- Zhang J (2003) Evolution by gene duplication: an update. *Trends Ecol Evol* 18:292–298