Chapter 14

Detection and Phylogenetic Assessment of Conserved Synteny Derived from Whole Genome Duplications

Shigehiro Kuraku and Axel Meyer

Abstract

Identification of intragenomic conservation of gene compositions in multiple chromosomal segments led to evidence of whole genome (WGDs) duplications. The process by which WGDs have been maintained and decayed provides us with clues for understanding how the genome evolves. In this chapter, we summarize current understanding of phylogenetic distribution and evolutionary impact of WGDs, introduce basic procedures to detect conserved synteny, and discuss typical pitfalls, as well as biological insights.

Key words: Whole genome duplication, Conserved synteny, Chromosome rearrangement, Differential gene loss, Hidden paralogy

1. Introduction

Whole genome duplications (WGDs), which resulted in new copies of existing genes, are considered to have provided possibilities of adaptive evolution (1, 2). The first indication of WGD dates back to 1970s (3). Later in 1990s, its direct evidence, supported by molecular sequences, emerged (4–6). After DNA sequences of several whole genome-scale became available, many studies revealed similar arrays of genes on different chromosomes within a single mammalian genome (conserved synteny; (7–11)). This large-scale intragenomic redundancy originated from the so-called "two-round whole genome duplications" (2R-WGDs) implicated at the base of all extant vertebrates, including jawless fishes (12, 13). Whole genome sequencing highlighted that the actinopterygian fish lineage experienced an additional WGD before the radiation of all extant teleost fishes (14, 15). More recently, it was reported that several plant lineages also experienced WGDs (16). Including the lineages leading to ciliates and the yeast, large-scale

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*,
Methods in Molecular Biology, vol. 855, DOI 10.1007/978-1-61779-582-4_14, © Springer Science+Business Media, LLC 2012

genome sequence resource have allowed us to detect WGDs in many different eukaryotic lineages (1).

The term "synteny," initially coined by a geneticist (17), originally stood for "presence of multiple genes on the same chromosome." The contemporary use of the term is extended to "conservation of similar arrays of genes between different chromosomes in a genome" (see ref. (18)). In this chapter, we keep the original definition of the term "synteny," and call the conservation of similar gene orders in multiple genomic regions "conserved synteny." Focusing on practical uses of publicly available resources, we present basic procedures to detect conserved synteny and to evaluate it referring to general patterns of gene family evolution.

2. Detection of Conserved Synteny

Conserved synteny containing color opsin genes is shown in Fig. 1 (modified from ref. (12)). Using this as an example, below we describe a basic procedure to detect conserved synteny.

2.1. Retrieving Sequences with Positional Information from Public Databases

In analyzing an already sequenced genome, public databases such as Ensembl (URL: http://www.ensembl.org; (19)) provide ready-to-use information of gene annotation and their chromosomal positions. To retrieve such information in Ensembl, the BioMart interface (URL: http://www.biomart.org; (20)) is convenient.

Analyzing multiple genomes may provide more information, especially when different organisms have retained different sets of paralogs after WGD. For example, in Fig. 1, human chromosome 1 does not harbor any color opsin gene because of a secondary loss of an opsin paralog in the eutherian lineage (21, 22), while paralogs of other gene families, such as *Lrrn2*, *Nfasc*, *Mapkapk2*, and *PlxnA2*, are retained (Fig. 1). In the chicken genome, the opsin paralog missing on human chromosome 1, is retained as a green opsin gene on chicken chromosome 26, allowing us to detect higher conservation of synteny than in human (12).

2.2. Preparing Sequences from Nonannotated Genomes

Even if no genome annotation database is available for the species of interest, one can perform a compact survey of conserved synteny, as long as a handful of genome sequences with certain lengths are available. Using available genome sequences as input, gene prediction programs can identify putative protein-coding genes, and report their sequences and positions [see Chapter 6 of this Volume (23) and also (24) for overviews of gene prediction tools]. Some gene prediction programs, such as Augustus (25), are capable of training themselves to adapt parameters to the species of interest, which is expected to result in improvement in identifying genes.

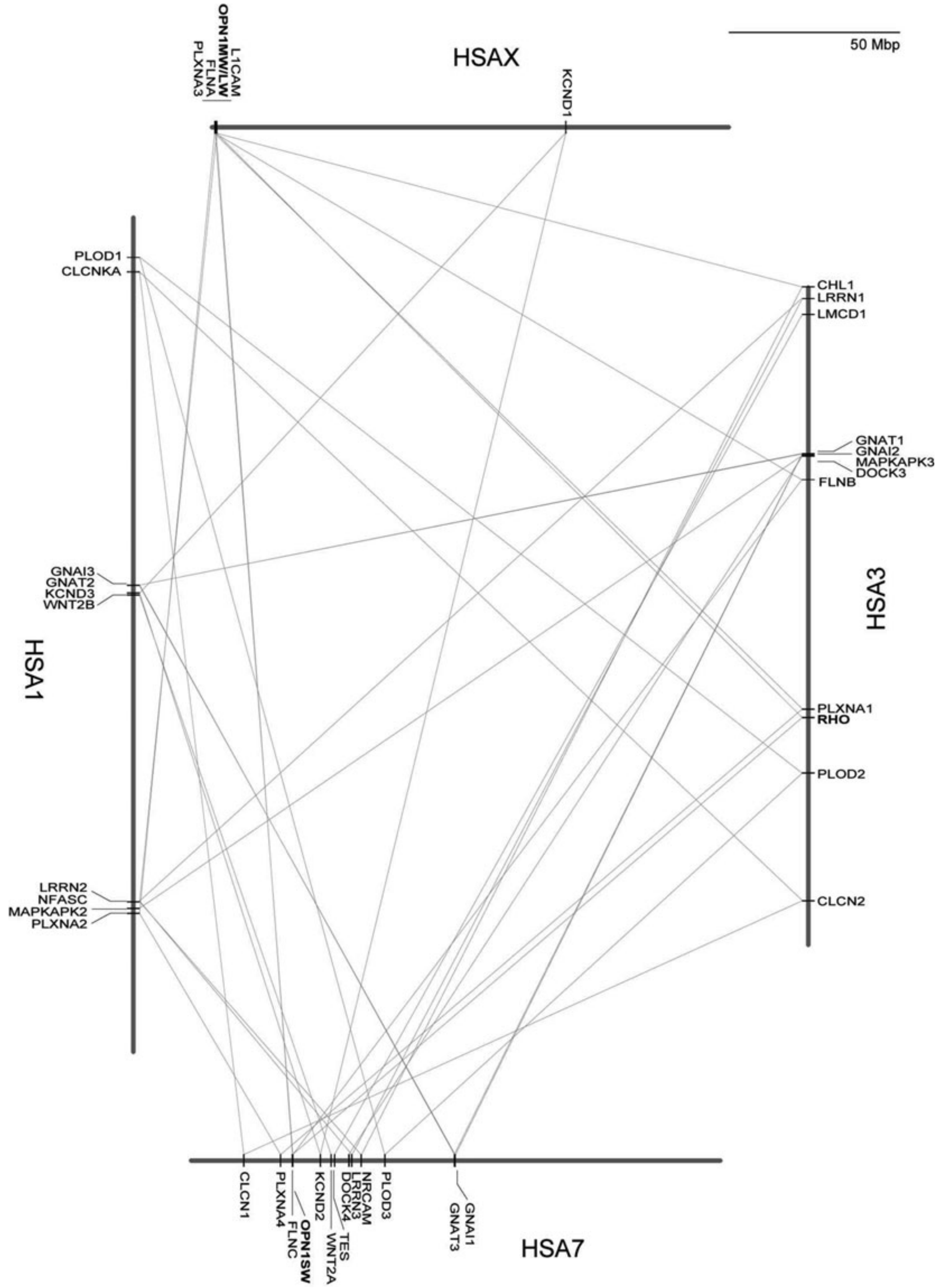

Fig. 1. Conserved synteny in the human genome containing vertebrate color opsin genes. This conserved synteny was detected by all-against-all homology search between these four chromosomes bearing color opsin genes based on the procedure introduced in Subheadings 2.1 and 2.3, followed by phylogenetic assessment explained in Subheading 2.5 (see ref. 12 for details). Color opsin genes are highlighted in *bold*. Paralogous gene pairs located on chromosomes next to each other were connected with *gray lines*. Gene names are shown as symbols specified by HUGO Gene Nomenclature Committee (HGNC).

2.3. Identifying Gene-by-Gene Homology Between Genomic Regions

In principle, initial clues of conserved synteny between two genomic regions can be detected by all-against-all homology search using Blastp (26) (see Fig. 3 in ref. (27); also see ref. (28)). In this process, some pairs of genes may exhibit weak similarity, and if they are not significantly similar or not similar enough because of too ancient gene duplication (for example, if they are two distantly related genes in a large gene family), this case will be noise in detecting conserved synteny derived from a recent WGD. Based on the so-called bi-directional best hit [BBH; or reciprocal best hit (RBH)] principle introduced in Chapter 9 of this Volume (29), this type of noise can be removed. When the selected genome is supposed to have more than two duplicated regions, as in the tetraploidized vertebrate genomes (10), this procedure requires the closest attention. It is because too stringent criterion in RBH can result in false-negatives (30).

Before the all-against-all homology search mentioned above, it is also recommended paying attention to repetitive elements possibly scattered throughout the input genome sequences. For example, in analyzing vertebrate genomes, the presence of many copies of long interspersed nucleotide element-1 (LINE1) usually results in noise masking real signals of conserved synteny, especially when the gene set is prepared according to the procedure described above in Subheading 2.2. In contrast, such repetitive elements are not annotated as protein-coding genes in Ensembl. Such repeats can be identified and masked in advance by RepeatMasker (http://www.repeatmasker.org) using a repeat library publicly available at RepBase (http://www.girinst.org/repbase/index.html; (31)). Moreover, to detect repeats in genomes of organisms with little genomic resources, tools such as RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) can facilitate the detection of species-specific ("*de novo*") repeats, in addition to those in RepBase. Gene families with a large number of members with similar sequences in the input genome (e.g., genes encoding Zn finger proteins, solute carrier proteins, and olfactory receptors in mammalian genomes) can also lead to noise. It should be noted that even after removing these potential sources of noise, the detected gene-by-gene homology spanning a certain range of chromosomes may still retain more noise resulted from small-scale evolutionary events (e.g., secondary insertion of genes or translocation of chromosomal segments). These are removed later in Subheading 2.5.

2.4. Identifying Large-Scale Conserved Synteny Using Publicly Available Tools

There are a few useful tools to detect conserved synteny available in public. The program i-ADHORe provides a possibility to detect conserved synteny within and between genomes (32). One of the advantages of this tool is that by incorporating information of gene orders of multiple organisms, one can more reliably identify conserved synteny through ancestral reconstruction of gene order. The performance of this tool, compared with a search based on only a single species, should be evident when relevant genomic regions experienced a considerable amount of secondary changes (see above Subheading 2.1 for an example of conserved synteny containing color opsin genes).

A more convenient resource accessible online is Synteny Database (URL, http://teleost.cs.uoregon.edu/synteny_db; (33)). In using this tool, one can select an organism of interest from a short list, currently containing only bony vertebrates. On the other hand, one can get sophisticated graphical output. Ensembl Genome Browser, introduced above, also contains orthology and paralogy information for every gene entry. If the organism to be analyzed is found in Ensembl (see the list of species at http://www.ensembl.org/info/about/species.html), one can retrieve a list of Ensembl gene entries paralogous to genes harbored in a selected genomic region through the BioMart interface. In the retrieved list, genes located in a relatively short genomic region or more may be detected, which could be a possible duplicate of the selected genomic region. Again, the results obtained in this step still contain possible noise. In the next step, the signal of conserved synteny is purified by assessing phylogenetic timing of gene duplications resulting in the detected gene-by-gene homology in Subheading 2.5.

The online browser Genomicus (URL, http://www.dyogen.ens.fr/genomicus; (34)) allows users to explore orthologous and paralogous conserved synteny in an interactive graphic interface. Genomicus functions based mostly on gene position information as well as molecular phylogeny in Ensembl. In this sophisticated resource, one can also search for intergenic conserved elements which may be responsive for transcriptional regulation of neighboring genes.

2.5. Phylogenetic Confirmation of Coincident Gene Duplications

The approaches mentioned above facilitate identification of similar arrays of genes, but do not provide information about timing of WGD. The only solution to provide time scale is a phylogenetic approach. Technical details of modern framework of molecular phylogenetics are introduced in Chapter 4 of this Volume (35). Reconstruction of molecular phylogenetic trees allows us to refine the gene-by-gene homology caused by genome duplication by removing homologous gene pairs introduced by small-scale events and to estimate the timing of genome duplication. For the former purpose, it is strongly recommended exploring all public databases to collect as many similar sequences as possible for reconstructing phylogenetic trees. An example highlighting the importance of this step is the conserved synteny between four genomic regions containing *Hox* clusters that duplicated in the 2R-WGDs (36). Members of many gene families are shared between those four genomic regions (37), but paralogous gene sets duplicated at different phylogenetic timings are also found frequently between those regions, such as *Wnt1*, *-2 (2A)*, and *-3* genes (Fig. 2). Although this gene set is sometimes documented as part of Hox-bearing conserved synteny (38), gene duplications giving rise to *Wnt1*, *-2*, and *-3* occurred before the origin of bilaterians

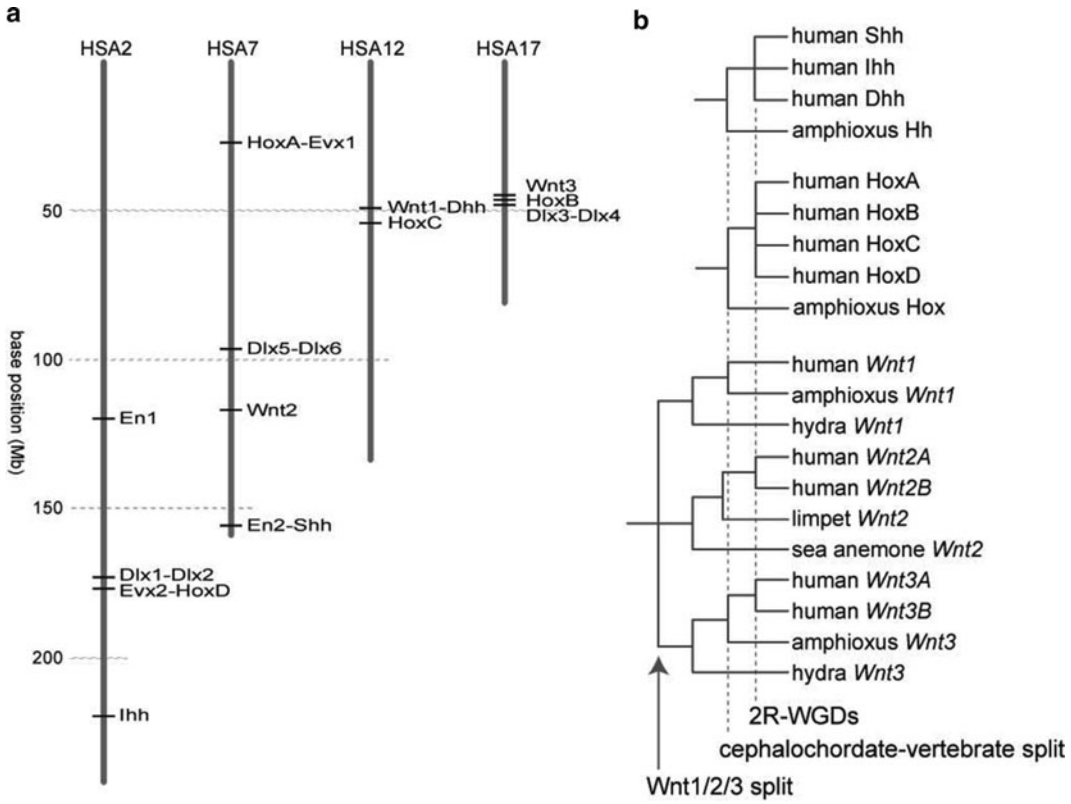

Fig. 2. Phylogenetic assessment of syntenic gene orders focusing on *Hox* clusters and *Wnt1/2/3* genes as a test case. (**a**) Chromosomal locations of some selected genes involved in vertebrate development (adopted from Fig. 4.3 of the ref. 38) in the human genome. When multiple genes are found in a short genomic region, their names are connected with a hyphen. (**b**) Timings of gene duplications giving rise to multiple paralogs in (**a**). See ref. 36 for Hox gene phylogeny and ref. 39 for the Wnt gene family. The trees presented here show that the gene duplication between *Wnt1*, *-2 (2A)*, and *-3* occurred more anciently than the split between *HoxA*, *-B*, *-C*, and *-D* clusters and between *Shh*, *Ihh*, and *Dhh* genes. For simplicity, other *Wnt* subtypes are not included in this schematized tree.

because each of *Wnt1*, *-2*, and *-3* has invertebrate orthologs (Fig. 2; (39)). This is not compatible with the timing of duplications between HoxA, -B, -C, and -D clusters (Fig. 2; (36)). As emphasized above, without phylogenetic assessment, similar gene arrays detected between chromosomes cannot serve as pure evidence of conserved synteny derived from large-scale duplication.

3. Interpretation of Conserved Synteny

Several possible sources of noise are already explained above, but in interpreting conserved synteny, some more factors should be taken into account. Especially in analyzing ancient genome duplications, there are more misleading factors that can mask genuine evolutionary history.

3.1. Statistical Validation of Conserved Synteny

There are many factors varying gene arrays—for example, secondary gene gains/losses, compaction/expansion of genes and intervals, and alteration of transcriptional orientation. How can we be sure that similar gene arrays we detect now really originated from WGD in the past? It is obvious that in a species with a small number of chromosomes in its karyotype (e.g., fly, fission yeast), we can more frequently find particular orders of genes on the same chromosome by chance. Thus, a window size in comparing gene orders is another important parameter. It may be optimal to set the window size at 50–100 neighboring genes (33). Although there is no sophisticated model available taking these factors into account, a typical approach is to randomize certain times the positions of members of the detected similar arrays of genes with permutation, and monitor how frequently the detected gene arrays in real data appear (40). In the Synteny Database introduced above, based on this approach, only conserved gene arrays that are significantly supported are shown in output (33).

3.2. Differential Patterns of Duplicate Loss After WGDs

When the number of homologs are compared between pre-WGD species and post-WGD species, the latter always have a less number of genes than the number estimated with the time of duplications—for example, in spite of the 2R-WGDs, the *Ciona intestinalis* genome is thought to contain as many as approximately 14,000 genes, compared to approximately 22,000 genes in the human genome. After the 2R-WGDs, gene families with only two (but not three or four) duplicates are more frequently observed, in spite of the 1:4 relationship estimated by the “two-round” WGDs (41). This suggests that a considerable number of duplicates derived from WGDs are destined to become extinct immediately after WGDs. Whether a new duplicate arose in a small-scale event (for example, tandem duplication) or large-scale event (for example, WGD), the fate of the new duplicate largely depends on the pattern in functional differentiation between the paralogs. If a duplication acquired new functions (neofunctionalization) or a subset of functions possessed by the original gene before the duplication (subfunctionalization), the gene should have had a higher chance to be retained in the genome (42, 43). In the case of the WGD in the *Paramecium* lineage, it was proposed that dosage compensation played a role in this process (44).

As depicted in Fig. 1, the influence of the loss of duplicates after WGD acts differently between gene families—some gene families lost duplicates on chromosome 1 while many others lost those on chromosome X. To confirm the differential pattern of duplicate loss between gene families, genome sequences of species that diverged before the WGD event should provide convincing evidence. For instance, genomes of papaya, *Kluyveromyces lactis*, and amphioxus provided unambiguous evidence confirming WGDs in the lineages of *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and vertebrates, respectively (1). Nonetheless, it should be noted that genomes that phylogenetically serve as a pre-WGD condition have also experienced a certain amount of chromosome rearrangement in their independent evolutionary lineages, and have not necessarily retained intact pre-WGD condition.

Differential patterns in retention of duplicates are also observed between different species—different lineages of species could have retained different sets of duplicates from each other. This largely confuses orthology/paralogy identification, causing the so-called "hidden paralogy" (45). Impact of hidden paralogy caused by differential gene loss has been emphasized especially in assigning orthology to genes of cyclostomes that are thought to have diverged immediately after 2R-WGDs (46).

Interestingly, some analyses of functions of retained duplicates have led to the understanding that particular groups of protein-coding genes are more frequently retained after WGDs. In the plants, it was shown that genes categorized as transcription factors, signal transducers, and developmental genes are more frequently retained, and 90% of the increase in gene number is accounted for by retention of these groups of genes after three rounds of WGDs in this lineage (47). This type of enrichment analyses can be performed based on Gene Ontology (GO) categorization of molecular functions, cellular components, and biological processes. Overrepresentation of particular GO terms can be revealed by publicly available tools, such as DAVID (48), GOSTAT (49), and FatiGO (50).

3.3. Rearrangement of Conserved Synteny After WGDs

In Fig. 1, we can detect three segments of chromosome 1 sharing paralogs with the three other chromosomes. This is thought to be caused by intrachromosomal rearrangement after the WGD. In fact, in the chicken genome, at least *Kcnd3* and *Wnt2B* in the segment in the middle of this chromosome are located in a 2.7 Mb segment of chicken chromosome 26 together with *Lrrn2*, *Nfasc*, *Mapkapk2*, and *PlxnA2* located in a different segment on human chromosome 1 (Fig. 1; (12)). This intrachromosomal rearrangement should have occurred in the mammalian lineage after the separation of the sauropsida (reptiles and birds) lineage. Another well-studied example is a rearrangement of Hox-containing conserved regions (51). Regarding the conserved synteny derived from the 2R-WGDs, it should be noted that genomic regions with conserved synteny documented since 1990s harbor only a small fraction of the entire gene repertoire. In other words, many more genes are buried in regions which do not exhibit obvious signals of conserved synteny. This suggests that, during more than 500 million years of evolution, conserved synteny has decayed through successive chromosomal rearrangement.

3.4. Applying Conserved Synteny to Addressing Different Types of Questions

Large-scale duplication events result in multiple gene families whose members duplicated at the same time and thus are in an array in the genome. Based on this assumption, timings of gene duplications in gene families whose members encode too short genes to reconstruct reliably the evolutionary history or experienced unusual secondary events preventing phylogenetic reconstruction can be estimated by analyzing other gene families in the same conserved synteny. One example recently reported by us is the timing of gene duplication between *Pax4* and *Pax6* genes. Coexistence of rapid-evolving *Pax4* gene and highly conserved vertebrate *Pax6* and invertebrate *eyeless* genes had prevented a reliable reconstruction of evolutionary history, but phylogenetic analysis on neighboring gene families suggested that *Pax4* and *Pax6* duplicated in the 2R-WGDs (52). In this example, the use of conserved synteny provided more insights into evolutionary transition of transcriptional regulation in this group of genes. Utility of the same approach has also been demonstrated for dating the timing of duplication of short genes that do not yield sufficient resolution in phylogenetic tree reconstruction (53, 54).

4. Exercises

1. Find research articles in NCBI PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez) containing the term “synteny.” Find out which of the two different usages of the term (see Subheading 1) the authors of those articles employ.
2. Following the procedure introduced in the text, identify conserved synteny containing genes encoding fibroblast growth factor receptor (FGFR) 1, -2, -3, and -4 in the human, chicken, and zebrafish. Discuss which of the three species have the most conserved synteny within the genome. What kind of genomic changes gave rise to the difference in gene orders between these species?
3. As explained in the text, some genomic regions have rigidly retained ancestral gene order, while other regions have not. What are the possible factors that may have caused this difference?

References

1. Van de Peer, Y., Maere, S., Meyer, A. (2009) The evolutionary significance of ancient genome duplications. *Nat Rev Genet*, **10**, 725–32.
2. Kuraku, S., Meyer, S. (2010) “Whole Genome Duplications and the Radiation of Vertebrates in Evolution after Gene Duplication. Pp. 299–311.” Katharina Dittmar and David Liberles, Eds. Wiley-Blackwell, NY.
3. Ohno, S.: Evolution by gene duplication. New York: Springer-Verlag; 1970.
4. Lundin, L. G. (1993) Evolution of the vertebrate genome as reflected in paralogous

chromosomal regions in man and the house mouse. *Genomics*, **16**, 1–19.

5. Holland, P. W., Garcia-Fernandez, J., Williams, N. A., Sidow, A. (1994) Gene duplications and the origins of vertebrate development. *Dev. Sppl.*, 125–133.
6. Sidow, A. (1996) Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev*, **6**, 715–22.
7. Endo, T., Imanishi, T., Gojobori, T., Inoko, H. (1997) Evolutionary significance of intra-genome duplications on human chromosomes. *Gene*, **205**, 19–27.
8. Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T., Ishibashi, T. (1996) Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proc Natl Acad Sci U S A*, **93**, 9096–101.
9. Katsanis, N., Fitzgibbon, J., Fisher, E. M. (1996) Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics*, **35**, 101–8.
10. Pebusque, M. J., Coulier, F., Birnbaum, D., Pontarotti, P. (1998) Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol Biol Evol*, **15**, 1145–59.
11. Thornton, J. W. (2001) Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc Natl Acad Sci U S A*, **98**, 5671–6.
12. Kuraku, S., Meyer, A., Kuratani, S. (2009) Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? *Mol Biol Evol*, **26**, 47–59.
13. Dehal, P., Boore, J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, **3**, e314.
14. Meyer, A., Schartl, M. (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr Opin Cell Biol*, **11**, 699–704.
15. Meyer, A., Van de Peer, Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, **27**, 937–45.
16. Fawcett, J. A., Maere, S., Van de Peer, Y. (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A*, **106**, 5737–42.
17. Renwick, J. H. (1971) The mapping of human chromosomes. *Annu Rev Genet*, **5**, 81–120.
18. Passarge, E., Horsthemke, B., Farber, R. A. (1999) Incorrect use of the term synteny. *Nat Genet*, **23**, 387.
19. Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., et al. (2009) Ensembl 2009. *Nucleic Acids Res*, **37**, D690–7.
20. Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., Kasprzyk, A. (2009) BioMart Central Portal – unified access to biological data. *Nucleic Acids Res*, **37**, W23–7.
21. Jacobs, G. H. (1993) The distribution and nature of colour vision among the mammals. *Biol Rev Camb Philos Soc*, **68**, 413–71.
22. Davies, W. L., Carvalho, L. S., Cowing, J. A., Beazley, L. D., Hunt, D. M., Arrese, C. A. (2007) Visual pigments of the platypus: a novel route to mammalian colour vision. *Curr Biol*, **17**, R161–3.
23. Alioto, T. (2012) Gene prediction. In Anisimova, M., (ed.), Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business media, LLC.
24. Picardi, E., Pesole, G.: Computational methods for *ab Initio* and comparative gene finding. In: *Data Mining Techniques for the Life Sciences* Edited by O Carugo, F Eisenhaber, vol. 609: Springer Verlag; 2010.
25. Stanke, M., Waack, S. (2003) Gene prediction with a Hidden-Markov model and a new intron submodel. *Bioinformatics*, **19**, Suppl. 2, pages ii215-ii225.
26. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–402.
27. Wolfe, K. H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet*, **2**, 333–41.
28. Van de Peer, Y., Meyer, A.: Large-scale gene and ancient genome duplications. In: *The Evolution of the Genome* Edited by R Gregory: Elsevier; 2005.
29. Altenhoff, A. M., Dessimoz, C. (2012) Inferring orthology and paralogy. In Anisimova, M., (ed.), Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business media, LLC.
30. Gabaldon, T. (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biol*, **9**, 235.

31. Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, **110**, 462–7.

32. Simillion, C., Janssens, K., Sterck, L., Van de Peer, Y. (2008) i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics*, **24**, 127–8.

33. Catchen, J. M., Conery, J. S., Postlethwait, J. H. (2009) Automated identification of conserved synteny after whole-genome duplication. *Genome Res*, **19**, 1497–505.

34. Muffato, M., Louis, A., Poisnel, C. E., Roest Crollius, H. (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, **26**, 1119–21.

35. Aris-Brosou, S., Rodrigue, N. (2012) The essentials of computational molecular evolution. In Anisimova, M., (ed.), Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business media, LLC.

36. Kuraku, S., Meyer, A. (2009) The evolution and maintenance of Hox gene clusters in vertebrates and the teleost-specific genome duplication. *Int J Dev Biol*, **53**, 765–73.

37. Larhammar, D., Lundin, L. G., Hallbook, F. (2002) The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res*, **12**, 1910–20.

38. Carroll, S. B., Grenier, J. K., Weatherbee, S. D.: From DNA to diversity: molecular genetics and the evolution of animal design. Malden, Mass.: Blackwell Science; 2001.

39. Kusserow, A., Pang, K., Sturm, C., Hrouda, M., Lentfer, J., Schmidt, H. A., Technau, U., von Haeseler, A., Hobmayer, B., Martindale, M. Q., et al (2005) Unexpected complexity of the Wnt gene family in a sea anemone. *Nature*, **433**, 156–60.

40. Deonier, R. C., Tavaré, S., Waterman, M. S.: Computational genome analysis: an introduction. New York: Springer; 2005.

41. Furlong, R. F., Holland, P. W. (2002) Were vertebrates octoploid? *Philos Trans R Soc Lond B Biol Sci*, **357**, 531–44.

42. Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531–45.

43. Lynch, M., O'Hely, M., Walsh, B., Force, A. (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics*, **159**, 1789–804.

44. Hughes, T., Ekman, D., Ardawatia, H., Elofsson, A., Liberles, D. A. (2007) Evaluating dosage compensation as a cause of duplicate gene retention in Paramecium tetraurelia. *Genome Biol*, **8**, 213.

45. Daubin, V., Gouy, M., Perriere, G. (2001) Bacterial molecular phylogeny using supertree approach. *Genome Inform*, **12**, 155–64.

46. Kuraku, S. (2010) Palaeogenomics of the vertebrate ancestor—impact of hidden paralogy in hagfish and lamprey gene phylogeny. *Integr Comp Biol*, **50**, 124–129.

47. Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*, **102**, 5454–9.

48. Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., Lempicki, R. A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, **4**, P3.

49. Beissbarth, T., Speed, T. P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–5.

50. Al-Shahrour, F., Minguez, P., Tarraga, J., Medina, I., Alloza, E., Montaner, D., Dopazo, J. (2007) FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res*, **35**, W91–6.

51. Lynch, V. J., Wagner, G. P. (2009) Multiple chromosomal rearrangements structured the ancestral vertebrate Hox-bearing protochromosomes. *PLoS Genet*, **5**, e1000349.

52. Manousaki, T., Feiner, N., Begemann, G., Meyer, A., Kuraku, S. (2011) Co-orthology of *Pax4* and *Pax6* to the fly *eyeless* gene: molecular phylogenetic, comparative genomic, and embryological analyses. *Evol Dev*, **13**, 448–459.

53. Braasch, I., Volff, J. N., Schartl, M. (2009) The endothelin system: evolution of vertebrate-specific ligand-receptor interactions by three rounds of genome duplication. *Mol Biol Evol*, **26**, 783–99.

54. Kuraku, S., Takio, Y., Sugahara, F., Takechi, M., Kuratani, S. (2010) Evolution of oropharyngeal patterning mechanisms involving *Dlx* and *endothelins* in vertebrates. *Dev Biol*, **341**, 315–23.