İSTANBUL UNIVERSITY PRESS

# What Reference Genome Assemblies Tell Us and How to Detect the Best Available Version: A Case Study in Trout

**Münevver Oral[1,2,3]**

## ABSTRACT

Genomic studies have largely been accelerated by the advances of next generation sequencing technologies since the beginning of the millennium. This, in turn, has motivated the generation of more reference genome assemblies not only in model organisms but also in species of scientific interest. In the present study, we employed a comparison study between the two different reference genome assemblies available for the same species, *Salmo trutta*, in GenBank. The results indicated an overall 90% similarity index between the two assemblies. Furthermore, the inversion regions of which assembly needs corrections were detected. Taking into account the whole genome duplication origin of the *Salmonidae* family, both assemblies were of good quality. However, the updated version of the Wellcome Sanger Institute assembly (*fSalTru_1.2*) outperformed the Norwegian assembly and was detected as the best available reference genome assembly in *Salmo trutta*.

**Keywords:** Reference genome, genome comparison, brown trout, *Salmo trutta*

**ORCID IDs of the author:**
M.O. 0000-0001-7318-6641

[1]Recep Tayyip Erdogan University, Faculty of Fisheries and Aquatic Science, Rize, Turkiye

[2]ISEM, CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France

[3]Département Biologie-Ecologie, Université de Montpellier, Montpellier Cedex 5, France

**Correspondence:**
Münevver Oral
E-mail:
munevver.oral@erdogan.edu.tr

## INTRODUCTION

In the early 2000s, the world experienced one of the most important breakthrough events in genomic science. (Liu, 2011; Goodwin et al., 2016; Whibley et al., 2021). The development of Next Generation Sequencing (NGS) technologies was motivated by the Human Genome (HG) project. Although the consortium announced that the first draft would be ready in 2005, advances in NGS technologies accelerated the process to a speed that could not have been previously estimated and the first draft of the HG project was made available two years ahead of delivery time, in 2003 (Roushan et al., 2014; Wu et al., 2016). Since then, researchers have been motivated to generate more sequence data in a short period of time and in a cost-effective manner, while a significant effort has been made to not compromise accuracy (Jian & Schneeberger, 2017; Enguita & Leitão, 2022).

As predicted almost two decades ago by Mardis (2006), accessing personalised genome assemblies is becoming a reality for humans in 100 minutes at 99.9% accuracy and 30x high coverage (Chin & Khalak, 2019). Furthermore, the cost of sequencing is estimated to become even more affordable for personalised medicine in the near future. A review by Wu et al. (2016) highlights the collaborative efforts in human genomics that have significantly led the way towards impactful achievements in clinical and practical medical applications.

Over the past two decades, advances in sequencing platforms enabled reference genome assemblies to be generated not only in model organisms but also in non-model organisms, including economically important fish species, e.g., Atlantic salmon (Lien et al., 2016), brown trout (Hansen et al., 2021), rainbow trout (Berthelot et al., 2014), European seabass (Tine

et al., 2014), common carp (Xu et al., 2014) and zebrafish (Howe et al., 2013). The previously announced consortium Genome 10K aimed to sequence 10,000 vertebrate species' genomes corresponding almost one species for each genus (Koepfli et al., 2015). In addition, since 2017, the Vertebrate Genome Project (VGP) has aimed to sequence the whole genome of 71,657 vertebrate species, including agnatha, cartilaginous fishes, abmphibia, osteichthyes and reptilia (Rhie et al., 2021).

Genome assemblies serve as a starting point from which a catalogue of reference DNA sequence is provided in species of interest (Kersey, 2019). These are of particular interest for exploring genome-wide variations and evolutionary histories as well as understanding species biology, biodiversity and conservation (Whibley et al., 2021).

The present study was motivated by the availability of two different reference genome assemblies on the public server, which is uncommon and can be confusing for most researchers. Working as a part of the international research community on genetics, it is of critical importance to detect the best available version of the reference genomes in any species of interest. By doing so, intra- and interpopulation variation can be better captured from a better-quality catalogue sequence due to increased alignment rate. Therefore, taken all together, the aim of the present study was (i) to decide the best available version of the reference genome for *Salmo trutta* that will be of use for alignment of multiple NGS data generated by Illumina technologies and (ii) serve as a pilot study for researchers to detect the best available reference genomes in any species of interest.

## MATERIALS AND METHODS

The National Center for Biotechnology Information (NCBI) was used to download the available reference genome assembly for brown trout (URL – 1). There were two available versions of the reference genome for species of interest. These were generated by (i) the Wellcome Sanger Institute and (ii) the Norwegian University of Life Science (see Table 1 for the details of the assemblies). Both reference genome assembly metrics were initially compared based on quality so as to decide the best available version; both assemblies were assessed by visualising similarities and flag ups.
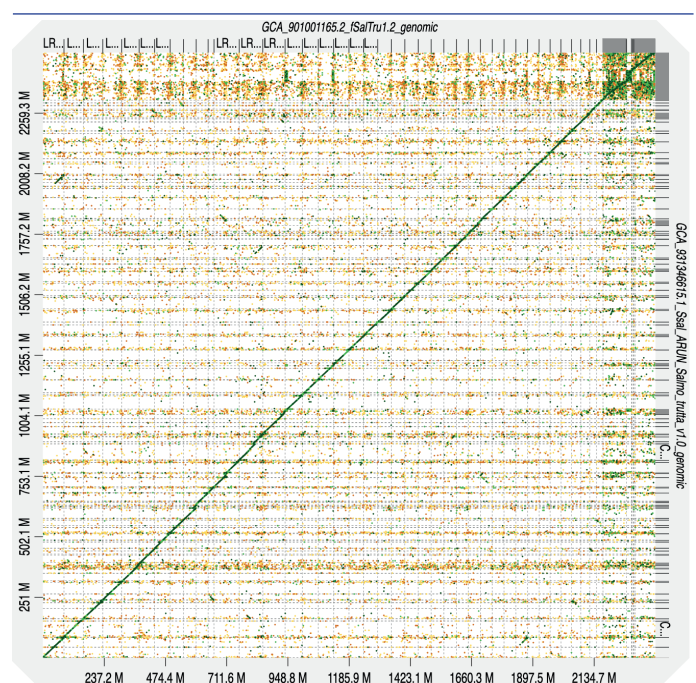
### Comparison of two reference genomes
Downloaded and zipped sequence FASTA files were uploaded to the online server (https://dgenies.toulouse.inra.fr/run) of D-genies version 1.4 (Cabanettes & Klopp, 2018). From the new alignment tool, the following parameters were chosen: (i) Target: GCA_901001165.2_fSalTru1.2_genomic and (ii) Query: GCA_931346615.1_Ssal_ARUN_Salmo_trutta_v1.0_genomic. Aligner and repeat options were kept as default for the analysis, and the task was submitted. The analysis time depends on the size of assemblies and it was completed in a day.

## RESULTS AND DISCUSSION

The D plot generated has shown relatively high similarity between the two different genome assemblies being compared (Table 1). The alignment match was supported over 75% identity,

indicated with continuous green dots (Fig 1). First, the noise was masked to visualise the higher percentage similarities. Overall, the summary identity resulted in over 36.36% similarity supported for >75%, followed by 54.37% for <75% similarity, 3.65% for <50% similarity, 0.45% for <25% similarity and 5.17% for no match at all between the two assemblies (Fig 2) in which, overall, the analysis supported statistically significant similarity over 90% (indicated as shades of green bars in Fig 2). When compared from the dropdown menu for each chromosomal region belonging to the target and query assemblies, respectively, this analysis also has resulted in high association supported with >75% identity index (Fig 3). There were flag ups indicating which regions of the assemblies need polishing using higher depth of coverage and longer sequencing for better quality. Furthermore, there were cases of inversions, structural rearrangement in the form of deletions, repeats and/or translocations, which were highlighted with opposite direction lines (see in details Fig 1 for such regions). These were end results of the same sequence but were represented in a different order. Similarly, these regions indicated a need for polishing to achieve a better-quality reference genome assembly for the available versions. The shaded grey bars on the upper right of the graph indicate sequences that are being merged in the form of a contig representing less than 0.2 % of the total assembly length. These regions signal that unassigned parts of the genome need to be assigned to the right positions in the genome.



**Figure 1.** The D plot was generated indicating the similarity index of two reference genome assemblies compared for *S.trutta*. The bottom and left sides of the graph demonstrate the size of the sequence in nucleotide position while the right and upper sides of the graph indicate the comparison assemblies of query and target, respectively.

| No match: | 5.17 % |
| < 25 %: | 0.45 % |
| < 50 %: | 3.65 % |
| < 75 %: | 54.37 % |
| > 75 %: | 36.36 % |
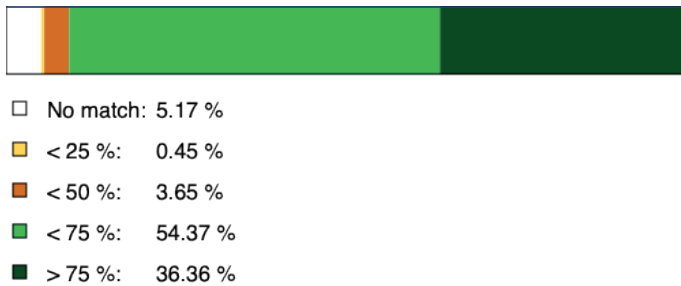
**Figure 2**. Similarity summary graph between the two reference genome assemblies.

Individual comparison between the target and query assemblies along the chromosome and matching contigs accordingly revealed high correspondence, thus a set of best matching D plot graphs are demonstrated in Fig 3. As the same colour code panel was applied to this analysis, the majority of these D plot graphs were supported over 75% similarity, indicated as dark green continuous lines, and in some cases, noise was detected, indicating assembly similarity less than 50% between the two assemblies, which appears as orange dots in the detailed graphs.

Reference genome assemblies have significantly improved the understanding of biological pathways and associations. However, genome assembly workload is an ongoing process which aims to serve the high-quality sequence archive for species of interest with gapless alignment (Whibley et al., 2021). While advances in next generation sequencing technologies multiply the capacities and the quantities of the high throughput data being generated in a short period of time, in parallel, the evolving era of bioinformatic analysis offers new tools to analyse such a large volume of data in an efficient manner (Jung et al., 2020). In the present study, we utilised an online tool developed for the comparison of two genomes available in GeneBank for *Salmo trutta*.

*Salmo trutta* is of scientific and economic interest as well as being popular among anglers. Thus, the species has been introduced to several countries (references therein Lobón-Cerviá, 2018). The phylogeny of *Salmo* has long been the subject of international debate due to the lack of understanding of the interaction between genotypic and phenotypic variation, and to its being regarded as a complex species, as opposed to designation of a single species (Ferguson 2004; Kottelat & Freyhof 2007). However, a recent opinion paper by Guinand, Oral and Thougard (2021) suggested a multiple species direction for the genus. The Salmonidae family has been though an additional round of whole genome duplication, which have caused one of the most complex genomes in vertebrates (Danzmann et al., 2008; Allendorf et al., 2015; Ohno et al., 1967). As indicated from the 2.37 Gb genome size and 41 chromosomes, *Salmo trutta* has a larger genome compared to most diploid fish species (Hansen et al., 2021). In the present study, D-genies enabled us to compare the similarity between the two large reference genomes of *Salmo trutta*. Both assemblies are of good quality, which is indicated with the quality metrics listed in Table 1 and a green line of correspondence demonstrating over 75% similarity in the colour coded identity panel (see Fig.1). The D-genies software utilises
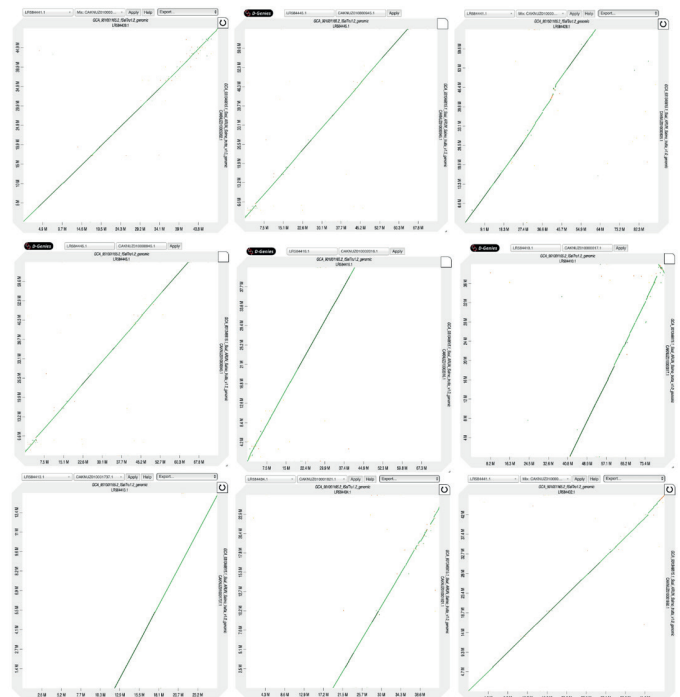


**Figure 3**. The best matching chromosome/contig graphs between the assemblies. The bottom and left sides of the graph demonstrate the size of the sequence in nucleotide position while the right and upper sides of the graph indicate the comparison assemblies of query and target, respectively.

a Blast-similar identity approach by applying the formula of I=M/N, in which I refers to Identity matrix, M indicates the quantity of matching nucleotides in the reference genome and N indicates the quantity of nucleotides, including gaps. Additionally, there were cases of inversions indicated as small chunks of opposite direction diagonal lines on the graph (Fig.1). These sequences exist in both assemblies, yet not in the same order. These are potential flag ups for reordering assemblies to improve quality. Fig 3 represents the best matching chromosomes/contigs between the assemblies, as follows: LR584441.1 versus CAKNUZ010005362.1, LR584445.1 versus CAKNUZ010000945.1, LR584416.1 versus CAKNUZ010002016.1, LR584413.1 versus CAKNUZ010001737.1 and LR584434.1 versus CAKNUZ010001821.1 in the Sanger Institute assembly versus the Norwegian assembly, respectively. Overall, taking into account the multisequence approach and high coverage of the Sanger Institute assembly (Table 1) as well as the chromosome level stage as opposed to the contig level unit in the Arun assembly, *SalTru_1.2.* generated by the Sanger Institute was selected for the downstream bioinformatic analysis of *Salmo trutta* sequenced via the Illumina platform.

One of the biggest limitations of the reference genome assemblies is the accuracy of the genomes and their annotations (Rhie et al., 2021). This is particularly the case in eukaryotic genomes, which contain high repeat content and duplication level (Elliot & Gregory, 2015). Regardless of the quantities of the reference ge-

**Table 1.**  General metrics of both reference genome assemblies of brown trout

| | Source: Wellcome Sanger Institute fSalTru | | Source: Norwegian University of LifeScience |
| --- | --- | --- | --- |
| | Previous version_1.1 | Current version_1.2 | Ssal_ARUN_Salmo_trutta_v1.0 |
| Accession number | GCA_901001165.1 | GCA_901001165.2 | GCA_931346615.1 |
| Assembly level | Contig | Chromosome | Contig |
| Sample (tissue) | Female (spleen) | Female (spleen) | Male (n/a) |
| Sample diploidy | Doubled Haploid | Doubled Haploid | n/a |
| Total seq. length | 2.371.880,186 | 2.371.880,186 | 2.510.277,823 |
| Total ungapped length | 2.298.279,497 | 2.298.279,497 | 2.510.277,823 |
| Genome coverage | 68x | 68x | 32x |
| Number of scaffolds | 1.441 | 1.441 | n/a* |
| Scaffold N50 | 52.209,666 | 52.209,666 | n/a* |
| Scaffold L50 | 18 | 18 | n/a* |
| *Contig count | n/a | n/a | 5,616 |
| *Contig N50 | n/a | n/a | 31.004,729 |
| *Contig L50 | n/a | n/a | 29 |
| Number of contigs | 5.378 | 5.378 | 5.616 |
| Total number of chromosomes and plasmids | n/a | 41 | n/a |
| Number of component sequences (WGS or clone) | 1.441 | 1.441 | 5.616 |
| Registration date | 02.06.2019 | 24.04.2021 | 19.02.2022 |

n/a: data is not available

nomes being available on public servers (1,348,815 genomes as of September 2022) the main focus should be directed towards improving the quality of the available reference genomes, including closing gaps with high coverage, as well as moving forward primary contig or scaffold level assemblies to the chromosome levels. To do so, hybrid sequencing approaches are applied by taking into consideration the advantages and disadvantages of the sequencing platforms. The Sanger Institute has applied the combination of the PacBio and 10X Genomics Chromium platform as well as BioNano and Hi-C data, achieving higher coverage, while the Norwegian assembly involved PromethION data from Oxford Nanopore technology combined with the Illumina platform. Given the quality metrics of both assemblies, the hybrid sequencing approach resulted in better quality assemblies as well as helping to close gaps while dealing with such an extended heterogenous genome.

The second most challenging task during reference genome assembly workload is dealing with the heterozygosity of the specimen being sequenced. In order to eliminate these cases, most assemblies utilise doubled haploid (DH) individuals, as these are theoretically 100% homozygotes thus helping eliminate the complications of duplicated genomes due to high heterozygosity (Whibley et al, 2021). Typically, eggs (n) of diploid female are fertilised using an irradiated sperm (n) from a diploid male. Although the genetic content of the sperm is inactive, the irradiated sperm is still motile and capable of initiating fertilisation. As there will be no genetic contribution from the sire, shock (chemical, physical or heat treatment) needs to be applied so as to ensure viability (2n) during the first mitosis. The resultant mitotic gynogenetics are produced by fully maternal ge-

nome transmission (Arai, 2001; Komen & Thorgaard, 2007; Oral, 2016; Manan et al., 2022). DH genomes offer the possibility of generating a more straightforward workflow, as fully a homozygous genome increases the chances of detecting any artefacts and/or sequencing errors in the form of variation. Therefore, DH individuals are preferred and have been utilised widely for reference genome assembly procedures in several aquatic species (Howe et al., 2013; Brawand et al., 2014; Berthelot et al., 2014; Xu et al., 2014; Lien et al., 2016; Hansen 2021). In the present study, the Sanger Institute's assembly was based on a DH female (Hansen et al., 2021) while no further information was provided in the Norwegian Arun assembly other than the sequencing of a male brown trout specimen (Table 1).

## CONCLUSION

In an effort to determine the best reference genome assembly for *Salmo trutta* for downstream data analysis, we compared the two available genomes. Overall, the recent version from the Wellcome Sanger Institute (*fSalTru_1.2*) was determined to be the highest quality reference available for *Salmo trutta* in terms of the multisequence approach applied and coverage achieved, as well as consisting the fact that it contains the chromosome level assembly as opposed to the contig level in the Norwegian Arun assembly. Taken all together, *fSalTru_1.2* will be utilised for the upcoming downstream genomic analysis of *Salmo trutta*, which involves a short sequencing approach applied using Illumina technologies.

**Conflict of Interest:** The author declares no conflicts of interest.

**Ethics committee approval:** As the present study was carried out in silico, ethics committee approval was not necessary.

## REFERENCES

Allendorf, F.W., Bassham, S., Cresko, W. A, Limborg, M. T., Seeb, L. W., & Seeb, J. E. (2015). Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *The Journal of Heredity*, 106(3), 217–27. https://doi.org/10.1093/jhered/esv015.

Arai, K. (2001). Genetic improvement of aquaculture finfish species by chromosome manipulation techniques in Japan. *Aquaculture*, 197(1–4), 205–228. https://doi.org/10.1016/S0044-8486(01)00588-9.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P. et al. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, 5, 3657. https://doi.org/10.1038/ncomms4657.

Brawand, D., Wagner, C. E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O. et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, vol 513, 7518, pp. 375-381. https://doi.org/10.1038/nature13726.

Cabanettes, F. & Klopp, C. (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958. https://doi.org/10.7717/peerj.4958.

Chin, C.S. & Khalak, A. (2019). Human Genome Assembly in 100 Minutes. *BioRxiv*. https://doi.org/10.1101/705616.

Danzmann, R.G., Davidson, E.A., Ferguson, M.M., Gharbi, K., Koop, B.F., Hoyheim, B., Lien, S., Lubieniecki, K.P., Moghadam, H.K., Park, J., Phillips, R.B., Davidson, W.S. (2008). Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (rainbow trout and Atlantic salmon). *BMC Genomics*, 9:557. https://doi.org/10.1186/1471-2164-9-557.

Elliott, T.A. & Gregory, T.R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society Biological Science*: 370: 20140331. http://dx.doi.org/10.1098/rstb.2014.0331.

Enguita, J.F. & Leitão, A.L (2022) in *New Frontiers and Applications of Synthetic Biology* Edited by Vijai Singh, Chapter 4 - *Advances, challenges, and opportunities in DNA sequencing technology*. Academic press. p. 31-43, ISBN 9780128244692, https://doi.org/10.1016/B978-0-12-824469-2.00022-1.

Ferguson, A. (2004). The importance of identifying conservation units: Brown trout and pollan biodiversity in Ireland. *Biology and Environment: Proceedings of the Royal Irish Academy*, 104B, 33– 41. https://www.jstor.org/stable/20500223.

Goodwin, S., McPherson, J. & McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews in Genetics* 17, 333–351. https://doi.org/10.1038/nrg.2016.49.

Guinand, B., Oral, M. & Tougard C. (2021). Brown trout phylogenetics: A persistent mirage towards (too) many species. *Journal of Fish Biology*, 99:298–307. https://doi.org/10.1111/jfb.14686.

Hansen, T., Fjelldal, P. G., Lien, S., Smith, M., Corton, C., Oliver, K., Skelton, J. et al. (2021). The genome sequence of the brown trout, Salmo trutta Linnaeus 1758. Wellcome Open Research, 6, 108. https://doi.org/10.12688/wellcomeopenres.16838.1.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E. et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496:498–503. https://doi.org/10.1038/nature12111.

Jiao, W.B. & Schneeberger, K (2017). The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, 36:64–70. http://dx.doi.org/10.1016/j.pbi.2017.02.002.

Jung, H., Ventura, T., Sook, J., Chung, W.J., Kim, B.H., Nam, Kong, H.J., Kim, Y.O., Jeon, M.S. & Eyun, S. (2020). Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Computational Biology*, 16, 10.1371/journal.pcbi.1008325. https://doi.org/10.1371/journal.pcbi.1008325.

Kersey, P.L. (2019). Plant genome sequences: past, present, future. *Current Opinion in Plant Biology*, 48:1–8. https://doi.org/10.1016/j.pbi.2018.11.001.

Koepfli, K.P., Paten, B., Antunes, A., Belov, K., Bustamante, C., Castoe, T.A., Clawson, H., et al. (2015). The Genome 10K Project: a way forward. *Annual Review of Animal Bioscience*; 3:57-111. doi: 10.1146/annurev-animal-090414-014900.

Komen, H. & Thorgaard G. (2007). Androgenesis, gynogenesis and the production of clones in fishes: A review. *Aquaculture*, 269, 150–173. https://doi.org/10.1016/j.aquaculture.2007.05.009.

Kottelat, M. & Freyhof, J. (2007). Handbook of European freshwater fishes. Cornol, Switzerland. https://doi.org/10.1643/OT-08-098a.1.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T.R. et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, (6020). https://doi.org/10.1038/nature17164.

Liu, Z.J. (2011). *Next Generation Sequencing and Whole Genome Selection in Aquaculture*. Wiley-Blackwell. DOI: 10.1002/9780470958964.

Lobón-Cerviá, J. (2018). Princess of the streams: The brown trout Salmo trutta L. as aquatic royalty. In J. Lobón-Cerviá & N. Sanz (Eds.), Brown trout – Biology, ecology and management (pp. 1–13). Hoboken, NJ: Wiley. https://doi.org/10.1002/9781119268352.ch1.

Manan, H., Hidayati, A. B. N., Lyana, N. A., Safwan. (2022). A review of gynogenesis manipulation in aquatic animals. *Aquaculture and Fisheries*, (7): 1,1-6. https://doi.org/10.1016/j.aaf.2020.11.006.

Mardis, E. R. (2006). Anticipating the 1,000 dollar genome. Genome Biology, 7(7), 112. DOI: 10.1186/gb-2006-7-7-112.

Ohno, S., Wolf, U. & Atkin, N. (1967). Evolution from fish to mammals by gene duplication. *Hereditas*, 59(6). DOI: 10.1111/j.1601-5223.1968.tb02169.x.

Oral M. (2016). Insights into isogenic clonal fish line development using high-throughput sequencing technologies. [PhD thesis] University of Stirling, Scotland, UK, available online.

Rhie, A., McCarthy, S., Fedrigo, O., Damas, J, Formenti, G., Koren, S, Uliano-Silva, M. et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592, 737-746. https://doi.org/10.1101/2020.09.08.285395.

Roushan, T., Ahmed, D., & Ali, M. R. (2014). Human Genome Project- A Review. *Medicine Today*, 26(1), 53–55. https://doi.org/10.3329/medtoday.v26i1.21315.

Tine, M., Kuhl, H., Gagnaire, P.A., Louro, B., Desmarais, E., Martins, R. S. T., Hecht, J. et al. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, 5, 5770. https://doi.org/10.1038/ncomms6770.

Whibley, A., Kelley, J.L. & Narum,S.R. (2021). The changing face of genome assemblies: Guidance on achieving high-quality reference genomes. *Molecular Ecology Resources*, (21): 641–652. https://doi.org/10.1111/1755-0998.13312.

Wu, J., Wu, M., Chen, T. & Jiang, R. (2016). Whole genome sequencing and its applications in medical genetics. *Quantitative Biology*, 4(2): 115-128. https://doi.org/10.1007/s40484-016-0067-0.

Xu, P. Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., Xu, J. et al. (2014). Genome sequence and genetic diversity of the common carp, Cyprinus carpio. *Nature Genetics*, 46, 11. https://doi.org/10.1038/ng.3098.

URL – 1: https://www.ncbi.nlm.nih.gov/data-hub/taxonomy/1/ [Accession date: 19.08.2022, 10.40 am]