Nuclear and Mitochondrial Genes Shed Light On The Evolution Of Salmon

Aniket Mane[1], Alice Yue[2], Zahra Zohrevand[2], and Joseph Lucero[3]

[1]Department of Mathematics, SFU, Burnaby, BC, V5A1S6 Canada
[2]Department of Computer Science, SFU, Burnaby, BC, V5A1S6 Canada
[3]Department of Physics, SFU, Burnaby, BC, V5A1S6 Canada

December 13, 2016

Abstract

Prompted by recent research efforts in elucidating the systematic relationships among salmonid fishes, we investigate a narrow, but important, persisting uncertainty in the current Salmonidae phylogeny. Specifically, we attempt to ascertain the placement of *Onchorynchus (O.) Masou* in the phylogenetic tree of Salmonidae. This uncertainty has remained due to either, a lack of available data, biological factors, or in-comprehensive methodologies. We attempt, with new available data, to create a robust phylogeny beyond what has been done in the past, using a combination of newly available tools such as MUSCLE (Multiple Sequence Comparison by Log-Expectation), BEAST (Bayesian Evolutionary Analysis Sampling Trees), and ASTRAL (Accurate Species Tree Algorithm). We also utilize a novel technique in this field, aligning without consideration of the third codon, that has granted clearer, more supported, phylogenetic trees. From our analysis, we find evidence that the current placement of *O. masou* in the phylogeny of Salmonidae is correct. It is our hope that by creating a more robust phylogeny that we can aid efforts in understanding the remaining uncertainties within this tree, as well as in other areas such as adaptation analysis, comparative genomics, and allocation of conservation priorities.

1 Introduction

Creating a robust phylogeny for the family Salmonidae has been an ongoing effort over the past few decades [38, 39, 40, 41]. Understanding the evolutionary relationships among salmonid fishes offers a gateway into understanding a multitude of evolutionary and ecological concepts, particularly the mechanisms of speciation, hybridization, chromosomal evolution, and introgression. Despite the large quantity of work already done on constructing phylogenetic relationships between the salmonid species there remains, still, a large set of important, unanswered questions regarding the family's evolutionary history. Many of these issues are often attributed to one of two causes: a lack of genomic data for many of the members of this family to analyze, and limitations that are imposed by biological factors including, but not limited to, rapid speciation, frequent hybridization, and local adaptation. Our main objective is to posit a phylogeny of Salmonidae using a greater amount of data than was previously available. Using this new phylogeny, we attempt to address a very narrow, yet unanswered, question regarding the placement of one of the species within the family. Specifically, we seek to establish evidence that will determine the placement of the species *O. masou* within the family and to add to the discussion of whether this particular species should be labeled a trout or a salmon.

2 Previous Work

2.1 Work by Crespi and Fulton

In 2003, Crespi and Fulton performed the first, novel, phylogenetic analysis of all available salmonid data from the National Institutes of Health (NIH) Genbank database. In order to perform their analysis, Crespi and Fulton utilized maximum likelihood and maximum parsimony methods from PAUP* (Phylogenetic Analysis Using Parsimony *and other methods) and bayesian analysis methods from MrBayes in order to determine the phylogeny with the best overall resolution and support.

Of particular importance to the current project at hand is the relationship that Crespi and Fulton derived for *O. masou*. Their analysis, placed *O. masou* as a monophyletic group with *O. clarki* and *O. mykiss*. *O. clarki*, the cutthroat trout, is, as the name suggests, accepted to be a trout. Similarly, *O. mykiss*, the rainbow trout, is accepted to also be a trout. The placement of *O. masou* into a monophyletic group with these two other species suggested, quite strongly, that *O. masou* is also a trout as well.

2.2 Work by Crête-Lafrenière, Weir, and Bernatchez

It was some years after the work of Crespi and Fulton before a major reconsideration of the salmon phylogeny occurred. Influenced by the same considerations that this paper is attempting to address, Crête-Lafrenière et. al, in 2012, performed a new analysis of the proposed phylogenetic relationships that had been present in the Salmonidae family in the years preceding their work. This analysis involved

- Using a more complete data set than was previously available, containing nearly all described salmonid species
- A larger amount of replicates/bootstraps than was previously possible, using a greater amount of computational power that is now available
- Similar techniques utilized by Crespi and Fulton but framed through a supermatrix

- New analytic techniques involving molecular markers.

in an attempt to both test the pre-existing relationships in Salmonidae, as well as ascertain a more resolved phylogeny than was previously possible.

Their analysis yielded many interesting results; however, the relevant result, at least to the scope of the current project, is that they derived phylogenetic relationships within *Oncorynchus* which, on average, contradicts the findings of Crespi and Fulton. Specifically, only one SH test resulted in significant support for the sister taxa relationship between the Pacific trout and the Japanese salmon; however, the authors do admit that, despite a greater number of taxa and an increased amount of gene sampling, discerning relationships of the genus *Oncorhynchus* is still difficult. They attribute the likely reason for this difficulty to be the rapid species radiation of this genus that was evident in the constructed phylogenies that they examined.

Thus, despite the large body of work that has already been done, ascertaining relationships in the rest of the Salmonidae family, there is still a large amount of uncertainty that continues to persist regarding the species-level relationship of the genus *Oncorhynchus*.

3 Methods

3.1 Sequence Alignment

Phylogenetic sequence data usually consist of multiple sequence alignments; the individual, aligned-base positions are commonly referred to as “sites.” These sites are equivalent to “characters” in theoretical phylogenetic discussions, and the actual base (or gap) occupying a site is the “character state”[33].

Aligned sequence positions subjected to phylogenetic analysis represent a priori phylogenetic conclusions because the sites themselves (not the actual bases) are effectively assumed to be genealogically related, or homologous. Sites at which one is confident of homology and that contain changes in character states useful for the given phylogenetic analysis are often referred to as “informative sites”. Steps in building the alignment include selection of the alignment procedure(s) and extraction of a phylogenetic data set from the alignment. The latter procedure requires determination of how ambiguously aligned regions and insertion/deletions (referred to as indels, or gaps) will be treated in the tree-building procedure.

3.1.1 Related Works and Background

In Pais et. al [25] which, is one of the most comprehensive research regarding the accuracy and efficiency of Multiple Sequence Alignments (MSA), the programs Probcons, T-Coffee, Probalign, and MAFFT [26] have been shown to outperform other programs in terms of accuracy. Whenever sequences with large N/C terminal extensions were present in the

BAliBASE suite, Probalign, MAFFT, and also CLUSTAL OMEGA, outperformed Probcons and T-Coffee. The fastest of these programs are CLUSTALW and MUSCLE [27] with CLUSTALW being the least, RAM, memory intensive program.

Based on Crête-Lafrenière et. al [24], using an "all-against-all" BLAST [28] in GenBank (Release 160), 52 genes, including 22 mitochondrial tRNAs, were identified. Then, a FASTA file has been generated and edited for each of the 52 genes using Geneious 3.0.5. Then, for each gene, a total of 45 alignments were carried out in ClustalW [29] using a range of parameter values (Gap Open: 315; Gap Extension: 37). The 45 alignments of a given gene have been then compared in SOAP [30] to retain only stable nucleotide positions in the final alignments.

3.1.2 MUSCLE

The MUSCLE algorithm proceeds in three stages: the draft progressive, improved progressive, and refinement stages [27]. In the draft progressive stage, the algorithm produces a draft multiple alignment, emphasizing speed over accuracy. In the improved progressive stage, the Kimura distanceis used to re-estimate the binary tree used to create the draft alignment, in turn producing a more accurate multiple alignment. The final refinement stage refines the improved alignment made in step two. Multiple alignments are available at the end of each stage.

In the first two stages of the algorithm, the time complexity is $\mathcal{O}(N^2L + NL^2)$, the space complexity is $\mathcal{O}(N^2 + NL + L^2)$. The refinement stage adds to the time complexity another term, $\mathcal{O}(N^3L)$ [25].MUSCLE is often used as a replacement for CLUSTAL. Since it often, though not always, gives better sequence alignments, depending on the chosen options. Morevover, MUSCLE is significantly faster than CLUSTAL, and even more so for larger alignments. A high-level flow of the algorithm is depicted in Figure 1.

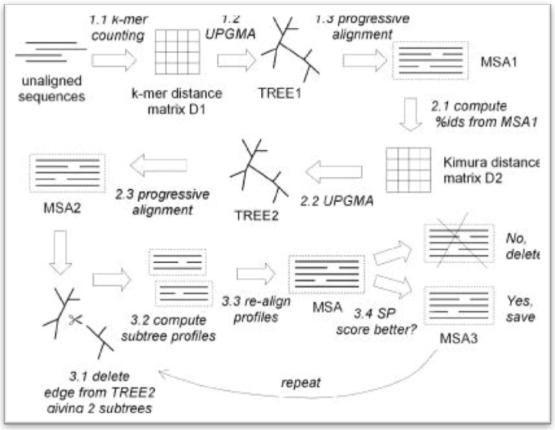
1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
No, delete
3.2 compute subtree profiles
3.3 re-align profiles
MSA
3.4 SP score better?
Yes, save
MSA3
3.1 delete edge from TREE2 giving 2 subtrees
repeat

Figure 1: High level flow of MUSCLE algorithm

3.1.3 MAFFT

MAFFT is a multiple sequence alignment program for unix-like operating systems. It offers various multiple alignment strategies, classified into three types:

a. the progressive method

b. the iterative refinement method with the WSP score

c. the iterative refinement method using both the WSP and consistency scores.

In general, there is a tradeoff between speed and accuracy. The order of speed is $a > b > c$, whereas the order of accuracy is $a < b < c$. In this approach, the accuracy of an alignment of a few distantly related sequences is considerably improved when they are aligned together with their close homologs. The reason for the improvement is probably the same as that for PSI-BLAST. Specifically, the positions of highly conserved residues, ie. those with many gaps, and other additional information are provided by close homologs. According to Katoh et. al, the improvement by adding close homologs is, approximately, 10% which, is comparable to the improvement achieved by incorporating the structural information of a pair of sequences [27]. MAFFT-homologs in the MAFFT server operate like so:

1. Collect a number (50 by default) of close homologs ($E = 1 \times 10^{-10}$ by default) of the input sequences.
2. Align the input sequences and homologs together using the L-INS-i strategy.
3. Remove the homologs.

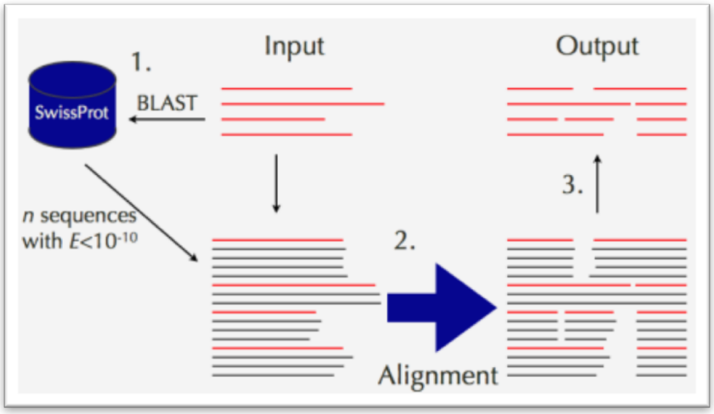

Figure 2: High level flow of MAFFT algorithm

A service based on a similar idea is also available in PRALINE. Note that MAFFT-homologs aim only to improve the alignment accuracy, but not to comprehensively collect the full-length sequences of homologs. Thus, the resulting alignment with homologs acquired by

MAFFT-homologs is not appropriate for evolutionary analyses. For evolutionary analysis, services with complete sequences, such as PipeAlign, are more appropriate.

3.2 Tree construction

Utilizing the original ClustalW sequence alignments given to us and the new MUSCLE sequence alignments that we performed, we created a phylogeny for each gene, the concatenated alignment of all nuclear genes, the concatenated alignment of all mitochondrial genes, as well as for all genes that we have data for. The original alignment we were given inserted gaps (-) instead of missing (?) for specimens that do not have particular genes. In most cases, such as the specimen OMasou100, would not have data for all nuclear genes or mitochondrial genes (Table 1). Therefore, we have also built phylogenies that are based on alignments where these gaps have been appropriately denoted as being missing.

It should also be noted that no out-groups were used to root the tree, as what was important to this analysis are the species-level relationships and not the genus-level relationships. Thus, whether or not we have a rooted tree, ultimately, makes no difference between in our results. For each phylogeny, we compared two phylogenetic inference tools: RAxML (Randomized Axelerated Maximum Likelihood) [1], and BEAST 2 [2].

3.2.1 RAxML

RAxML is a maximum likelihood (ML) phylogenetic inference methodology that attempts to find a phylogenetic tree that maximizes the probability of the data, or the given alignment. Similar to other ML methods, RAxML consists of two repeating steps:

1. Create alternative trees (tree optimization)
2. Test the likelihood scores of these trees (tree testing)

This is done in order to find the best ML tree. To obtain a consensus tree with confidence values, we would repeat these steps by conducted rapid bootstrapping, or sampling, of 1000 trees. Rapid boosting has been shown to perform comparably with regular bootstrapping [10], and bootstrapping 1000 trees is generally more than enough to meet the requirements of most bootstrap stopping criterion [10, 17].

More specifically, RAxML starts with a maximum parsimony tree which, usually yields better likelihood scores when compared to neighbour joining, or random starting trees [1]. RAxML then uses the tree optimization and testing procedure to compute a ML tree and then, subsequently, optimizes the model parameters. This tree is then used to bootstrap ten replicate trees, taking the final tree from the previous bootstrap as the starting tree of the next. The result of this bootstrap iteration, in conjunction with the original alignment, is then used to create another starting MP tree. Once again, this starting tree will initiate ten bootstrap replicates. This process repeats until we have our tree 1000 replicates, from which we can search for our best tree and final likelihood score.

Interestingly, RAxML is a fast algorithm due to the way it conducts its tree optimization: Lazy Subtree Rearrangements (LRS) [11]. LRS is where a subtree is removed from the current best tree, and then reattached to all of the neighbour branches up to $r = [5, 25]$ nodes away. LRS then calculates a likelihood score for each alternative tree topology made, optimizing a radius of up to 3 branches away from the reattachment point to improve efficiency. If an insertion is found to improve likelihood scores, this tree immediately becomes the base tree for the following rearrangements. To make this lazy procedure even faster, we use rapid bootstrapping. This method sets a dynamic threshold, that will decrease from infinity, to stop further LRSs at larger r if increasing the tree search space seems unlikely to yield a higher likelihood score.

In addition to the best tree, we also create a majority rule consensus tree from our 1000 bootstraps. The resulting tree will only consist of branch splits agreed upon by a majority of our bootstrap trees. The other branches would then be collapsed while the remaining ones are assigned a score indicative of the percent of bootstrap trees it is supported by.

The parameters in RAxML comes from our choice of using the generalized time reversible (GTR) model of nucleotide substitution [4, 5, 6]. GTR assumes that each nucleotide substitution appears at a different rate and that nucleotides occur at different frequencies. Namely, it uses the frequency of each base as the stationary distribution of each state or nucleotide $\pi^* = (\pi_A, \pi_T, \pi_C, \pi_G)$. The rate of change of substitution between states is represented by transition rate matrix, conditioned that it will reach π^* after a long period of time and evolution:

$$Q = \begin{pmatrix} -(x_1 + x_2 + x_3) & x_1 & x_2 & x_3 \\ \dfrac{\pi_1 x_1}{\pi_2} & -\left(\dfrac{\pi_1 x_1}{\pi_2} + x_4 + x_5\right) & x_4 & x_5 \\ \dfrac{\pi_1 x_2}{\pi_3} & \dfrac{\pi_2 x_4}{\pi_3} & -\left(\dfrac{\pi_1 x_2}{\pi_3} + \dfrac{\pi_2 x_4}{\pi_3} + x_6\right) & x_6 \\ \dfrac{\pi_1 x_3}{\pi_4} & \dfrac{\pi_2 x_5}{\pi_4} & \dfrac{\pi_3 x_6}{\pi_4} & -\left(\dfrac{\pi_1 x_3}{\pi_4} + \dfrac{\pi_2 x_5}{\pi_4} + \dfrac{\pi_3 x_6}{\pi_4}\right) \end{pmatrix}$$

where x_i , $i \in [1, 6]$ are 6 parameters that need to be defined and π_j , $j \in [A, C, G, T]$ are the frequencies of π^*. The 6 specific parameters, x_i, of our GTRGAMMA model can be estimated by ML which, will converge to their true parameter values as the sample size becomes large.

We also used a GAMMA distribution for the rate heterogeneity model to represent rate variation among sites, in part because GTR is the only model implemented for nucleotide data sets in RAxML. RAxML also offers GTRCAT which, puts site-specific evolutionary rates into a distinct number of categories for more efficient computation. GTRCAT has been shown to yield equivalent phylogeny likelihood values for large data sets but for smaller data sets, mainly in our case (i.e. some genes have less than ten taxas), GTRGAMMA would still help us obtain the better likelihood scores [9].

3.2.2 BEAST

BEAST, on the other hand, is a tool for conducting Bayesian inference on phylogeny. Bayesian inference tries to maximize the posterior which is built up of both the likelihood (as in ML) and the prior. Bayesian methods are becoming more and more widely used, materializing into various popular tools such as MrBayes [12], Batwing [13] and the one used here, BEAST 2. However, we chose BEAST 2 because it still focuses heavily on trees that incorporate time scales, or different rates of evolution on different branches of the tree, a key feature in analyzing specialization events [16]. In terms of the algorithm, they all take advantage of the stochastic statistical sampling-based algorithm, Metropolis-Hastings Markov Chain Mote Carlo (MHG) to estimate a distribution of the posterior probabilities of trees [14, 15, 18]. The MHG algorithm produces a Markov Chain where its current state is defined by $T = \{t, \mathbf{b}, \mathbf{s}, g\}$, a specific tree t, branch lengths $\mathbf{b}$, substitution parameters $\mathbf{s}$, and GAMMA shape parameter g. It then defines its next state T' with a probability of $P(T'|T)$ via the Markov property (i.e. the next state only depends on the current state). The probability of actually moving to the next state is then:

$$\begin{aligned} R &= \left(1, \frac{P(T'|X)}{P(T|X)} \cdot \frac{P(T|T')}{P(T'|T)}\right) \\ &= \left(1, \frac{P(X|T')P(T')/P(X)}{P(X|T)P(T)/P(X)} \cdot \frac{P(T|T')}{P(T'|T)}\right) \\ &= \left(1, \underbrace{\frac{P(X|T')}{P(X|T)}}_{\text{Likelihood Ratio}} \cdot \underbrace{\frac{P(T')}{P(T)}}_{\text{Prior Ratio}} \cdot \underbrace{\frac{P(T|T')}{P(T'|T)}}_{\text{Proposed Next Ratio}}\right) \end{aligned}$$

Where $P(T|T')$ is the probability of a reverse move, which is never done, and X is the given nucleotide multiple sequence alignments. Since the probability $R \in [0, 1]$, HMG then draws a uniform random variable $z \in [0, 1]$ and if $z < R$, HMG will move to state T', otherwise no move is made. These states put together creates a Markov Chain, out of which BEAST 2 samples trees to estimate the posterior distribution.

For the substitution model, again, we used GTR with 4 GAMMA rate heterogeneity rates, a sufficient number of rates for most data our size [22]. We also adjusted starting model parameters using the model test provided in MEGA.

For the clock parameter, we tested in BEAST 2, the relaxed clock log normal model assumes the substitution rates on each branch follow a log normal distribution. This allows each branch to have their own evolutionary rates. BEAST 2 implements a computationally efficient version of this model in that the probability density of rates is discretized into bins, with each bin corresponding to a branch in the computed phylogenetic tree. The strict clock, however, are robust in cases when time of evolution is short, or the tree is shallow. This is because it assumes a strict rate of evolution among all branches, which usually occurs within the same species. Nevertheless, [23] has noted an abundence in introgression among the mitochondrial genes. This can be attributed to intermating between salmonoid species in a common spawning area.

As there is no agreed upon species tree already available, we started BEAST 2 with a random tree prior on the yule process. The Yule process assumes that each salmon independently gives birth a constant rate, a good assumption for analyzing speciation [19, 20, 21]. This assumption works well with our scenario since salmonoid fish are migratory and go to particular spawning areas to mate, lay eggs and subsequently die.

3.3 Supertree Construction and Introgression analysis

3.3.1 ASTRAL

ASTRAL is a statistically consistent coalescent-based method which helps in the estimation of species phylogenies. The underlying motivation behind ASTRAL is the discordance between gene trees generated from different genes. This happens, partly due to horizontal gene transfer, duplications and incomplete gene sorting.

Hence, the estimation of a species tree is generally based on gene trees based on multiple genes, also referred to as loci. One of the possible approaches to achieve the desired estimate is to concatenate the alignments for multiple loci and estimating the species tree based on the new alignment, thus ensuring that the data corresponding to all the genes has been considered. As suggested in ([36]), this method, using concatenation based analyses might lead to statistically inconsistent results. Each gene evolves at a different rate and in a manner that might not be identical to each other. Thus, various parts of the newly created concatenated alignment, corresponding to different genes, evolve at a nonuniform rate. Hence, resultant tree is a statistically inconsistent indicator of the species tree.

In order to provide a brief idea of how ASTRAL works, we try to explain the computation of Weighted Quartet Score (WQ). Given a set G of k binary input gene trees on n taxa, there is a multi-set of

$$\frac{k.n!}{(n-4)!4!}$$

quartet trees induced by the input. The WQ score of a given tree as the number of quartet trees from this multi-set that the given tree also induces.The optimization problem solved by ASTRAL is to find the species tree that maximizes the WQ score.

ASTRAL uses a dynamic programming algorithm to solve the problem. Each internal node of an unrooted tree divides the set of leaves into three parts, defining a tripartition, denoted by $X|Y|Z$. Each tripartition also defines some number of quartet topologies that will be induced by any tree that includes that tripartition as a node. In order to calculate the number of shared induced quartet trees between two given tripartitions, Mirarab et al [37] gave the following formula:

$$Q(T,T') = Q(C) = \sum_{a,b,c \in G_3} F(C_{1,a}, C_{2,b}, C_{3,c})$$

where $T = A_1|A_2|A_3$ and $T' = A'_1|A'_2|A'_3$ are the two tripartitions and $C - ij = |A_i \cap A'_j|$ is the intersections of two partitions, one each from T and T' for all possible combinations i and j. G_3 is the set of permutations of the tripartitions while F can be calculated as:

$$F(a,b,c) = \frac{abc(a+b+c-3)}{2}$$

where a,b and c give the sizes of each of the tripartitions.

The overall score for a tripartition is defined as

$$w(T) = \sum_{g \in G} \sum_{T' \in N(G)} Q(T,T')$$

Mirarab et al [37] proved that the estimation of species trees under this model is statistically consistent.

The input and output of ASTRAL are both unrooted trees in Newick format. Input trees are allowed to contain polytomies; however, higher degrees of interior nodes in the input trees can significantly decrease the speed. The branch lengths in ASTRAL are measured in coalescent units and are a direct indicator of how different the gene trees are. Due to the statistical noise provided by certain genes, it is sometimes desirable to leave the noise inducing loci out of the calculation. However, for the purpose of our research, partly due to sparseness of data for the nuclear genes, we have considered all the genes provided to us for the estimation of the species trees. ASTRAL provides a normalized quartet score for the tree, which is a number between 0 and 1. Typically, this number reflects the proportion of input gene tree quartet trees satisfied by the resultant species tree. This number can be treated as a measure of confidence in a given estimate of the species tree. Hence, the closer this number is to 1, the better it represents the general consensus of the gene trees.

3.3.2 Reconciliation and Introgression analysis

Gene families evolve over time due to gene duplications, lateral transfers, speciations and gene losses. Accounting for these events, the species trees very often differ from the gene trees. In such cases, it becomes necessary to obtain a mapping from the species trees to the gene trees. This can be done using various reconciliation methods. Given a cost of each of the evolutionary events mentioned above, parsimony reconciliation methods help us build the sequence of these events at an optimal cost.

Considering the evolutionary events most likely to happen in case of a particular group of species, various reconciliation models can be used. One of the model used to map the nodes of the gene tree to respective nodes in the species tree is the Duplication-Transfer-Loss

(DTL) model. This model considers horizontal gene transfers as major evolutionary events. Maintaining the time consistency of the transfer is a major issue in this case. A transfer can occur only between contemporary species. Also, two locally consistent transfer should be globally consistent. In other words, given a locally consistent transfer, any other transfer, even if locally consistent, can not violate the time frame of the given transfer. In general, maintaining the time consistency is an NP-hard problem. However, in case the divergent dates (dates for which the species diverges) for the nodes of the species tree are available to us, this can be done in polynomial time.

Introgression [42] a type of hybridization, resulting from backcrossing or transfer of genes from one species into the genomes of another species. Hence, introgression is also referred to as Horizontal Gene Transfer (HGT). In cases where introgression takes place, there is still a portion of the genomes that remains unchanged. The unchanged portions make it possible to recognize that 2 different genomes exist.

Introgression is found to be more prevalent in Mitochondrial DNA. The discordance between the gene trees for mitochondrial and nuclear DNA can attributed to the hybridization between Salvelinus species [43]. In our analysis, we intend to use the species trees obtained using ASTRAL as our reference species trees. We perform the reconciliation analysis using a software *ecceTERA* [44]. EcceTERA implements a parsimony reconciliation algorithm that accounts for DTL and speciation irrespective of the tree being dated. In addition to this, ecceTERA fairs better than most other reconciliation softwares, such as, RANGER-DTL and Notung due to its ability to handle non-binary gene trees.

The discordance between two gene trees or a gene tree and a species tree is the result of introgression. The transfer of a particular gene from a species to another might lead to the gene tree of that particular gene showing an incorrect representation of the evolution of the species tree. In order to detect introgression between species using ecceTERA, we use a higher cost for HGT as compared to duplications and loss. The results from the reconciliation of gene trees, belonging to mitochondrial genes in particular, with the nuclear species trees shed light on the extent of introgression of the gene in question. This, in turn, enables us to identify which gene trees are responsible for the difference between the mitochondrial and nuclear species trees.

4 Results

4.1 Improving the alignments

Based on related work research, we decided to apply two other well-known approaches, MAFFT and MUSCLE, on the given sequences and compared the resulting MSA from each approach with the original MSA using the log-likelihood value of the phylogenetic trees that MEGA (Molecular Evolutionary Genetics Analysis) can build [31].

We applied MUSCLE on the given gene sequences using different parameters for some genes and two specific parameter settings. First of all, we set the parameters of affine gap as opening-gap=400 and gap-extension=0 which can be considered as kind of ideal situations that may happen in coherent genes in the nature. As well, we applied MUSCLE on all of the gene sequences using parameters of opening-gap=15 and gap-extension=7, which have been presented in [24] to compare the final result with given alignments. In order to compare the obtained alignments from MUSCLE and the given alignments, we constructed the maximum likelihood phylogeny tree of all of alignments and compared the pair of alignments based on log-likelihood of the phylogeny trees.

In addition, we applied MAFFT with the same parameters as in [24], using the -auto flag for the length of sequences. Table 2 shows our obtained log-likelihood for different performed alignments. The red rows highlights the subsequences where MUSCLE, with smaller opening gap penalty, gave us an error and we could not get their alignment. The blue rows highlights the genes that their MUSCLE alignment even using a very high opening gap penalty can lead to a very good answer. The green column highlights the best answer which, it turns out, is the MUSCLE alignment. It is worth noting that even phylogenies constructed based on the alignments done by MUSCLE, using BEAST and RAxML which, will be attached as running documents, confirm our obtained result.

4.2 Saturation

Sometimes, nucleotide and amino acid data can provide high support for conflicting relationships. That can be because of the third codon position can get saturated and phylogenetic analysis of this position alone supports a completely different, potentially misleading sister group relationship [32]. Based on this idea, we tried to examine how codon position and saturation might influence resolution and node support among 107 taxa considering 32 nuclear and mitochondrial genes. Obtained result have been presented in the following table. In case of some of the genes, we acquire different branches and different confidence values for species which, illustrate that the Salmonidae genes may be exposed to saturation in third codon position. While we posit that this may be the case, we would need to confirm with someone who is much more versed than we in this field. The created trees for triples of (complete genes, genes without third position, and just third position) have been attached with documents.

It is worth noting again, in this step, for each of phylogenic trees, we did model test using MEGA and applied the suggested parameters in phylogeny tree construction. The result of model test for each gene can be found in the supplemental material.

4.3 Species Tree Construction

We constructed the species trees based on the gene trees output by MEGA and RAxML. Note that for each species tree, we considered only the best or most likely tree in each group. For instance, for computing the species tree for mitochondrial genes, based on gene trees output by raxML, we considered the most likely tree in the RAxML output for each of the 16 mitochondrial genes provided. In doing so, ASTRAL assumes that each of the gene trees used for the estimation of the species tree has a high confidence measure. Thus, if we provide

ASTRAL with a "bad" input (with very less confidence), it will still assume the tree to be the best representation of the corresponding gene tree.

The following table displays the support for each species tree constructed:

Normalized quartet scores (in %)	RAxML	MEGA	MEGA (w/o 3rd codon)
Nuclear genes	86.19	89.15	86.24
Mitochondrial genes	94.10	94.12	92.32
All genes	93.99	93.99	92.18

Quite clearly, each species tree shows a high quartet score of above 80%. This indicates that the trees provided by ASTRAL are indeed a good estimate of the species tree.

The analysis of the placement of *O. masou*, in accordance with the species trees will be discussed in the subsequent sections.

5 Discussion

5.1 Sequence Alignments

Based on the final result, we figure out the two following interesting outcomes:

1. Using the result of MUSCLE alignments with the same parameters as proposed alignments in [24], the log-likelihood of trees are much better which, can be considered as a remarkable improvement and superiority of MUSCLE in comparison with MAFFT and CLUSTALW.

2. The alignment on the under-investigation sequences will often lead to statistically significant more accurate alignments if the difference between penalties of opening a gap and extension of gap would be smaller. Having smaller penalty for opening gap (semi-linear gap penalty) schema assume every single nucleotide mutation is independent of the others which appears to be not logically significant for single genes. While, intuitively affine gap penalty score schemes assume single nucleotide mutation might depend on its neighboring nucleotide mutations. Therefore, the obtained result lead us to two hypotheses:

   - Over time, the mutations have happened in different parts of genes, so there are a lot of subsequence of genes that they match together better than the overall. Thereupon, maybe trying to improve the gene structure from the whole sequence would be a precious and promising research for future. (It should be noted that defined task for our project was just trying to improve alignment for each profile instead of extracting genes from scratch)

   - On the other hand, we are doubtful that whether having phylogenetic tree with bigger log-likelihood can be consider as a right measure to decide between different settings of penalties for affine gap penalty. In fact, we did a thorough research to find the best measure for comparing proposed alignments of different algorithms and the findings were disappointing. In most of serious research the comparison have been done based on the previous knowledge or efficiency of the algorithm like [25] and finally some information in scientific discussion groups and the defined relation between phylogenetic tree and the role of sequence alignment in construction of trees process [33], which have been mentioned in background lead us to take log-likelihood as a reasonable comparison measure between alignments algorithms with same parameters. But, we are not sure that which parameter can be the best to find the most appropriate parameters.

5.2 Masou's position in the family

We analyzed the trees obtained from ASTRAL and, based on prior knowledge about the species sharing the clade, or monophyletic group, with *O. masou*, we attempted to deduce if *O. masou* is a salmon or a trout.

5.2.1 Species trees obtained from nuclear genes

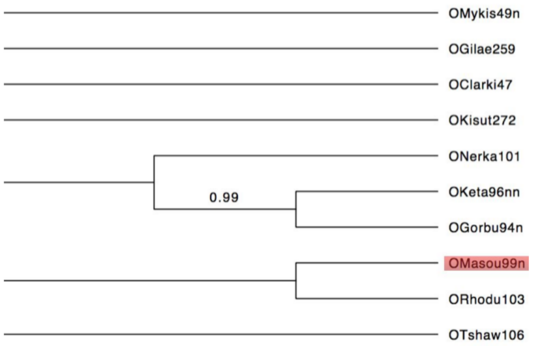

Figure 3: The clade containing *O. Masou* in the species tree for nuclear gene trees obtained using MEGA (without 3rd codon).

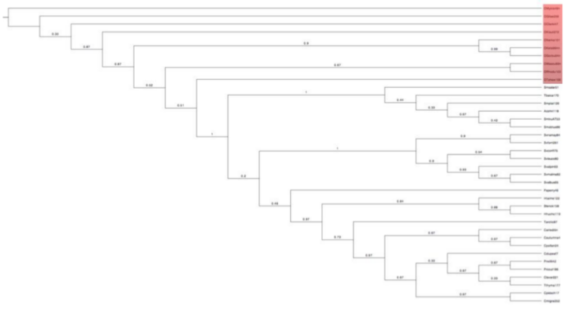

Figure 4: The entire nuclear species tree. The clade containing *O. masou* has been highlighted. It is possible to trace the nearest relatives of the species of interest along the branches of the tree.

Gene name	Gene Position	Current Alignments	MUSCLE 1	MUSCLE 2	MAFFT
			(O. gap= 400)	(O. gap=15)	(O. gap=15)
			(E. gap= 0)	(E. gap=7)	(E. gap=7)
Cytb	1-1141	-13463.83	-13463.83	-11169.6	-13461.24
mtDNA_Coding	1-11413	-71407.5200	-71631.2400	-68347.48	-71464.6499
Mitochondrial	1-15597	-83466.0099	-85263.61	-75333.58	-84944.74
CO1	1142-2692	-13847.75	-13953.48	-12210.38	-13918.05
CO2	2693-3378	-2292.06	-2292.06	-2119.78000	-2292.06
CO3	3379-4164	-3093.56	-3093.56	-2774.02	-3095.44
ATP6	4165-4848	-5302.98	-5302.98	-4465.47	-5302.98
ATP8	4849-5016	-407.4	-407.4	-375.86	-407.4
ND1	5017-5980	-6196.8	-6196.8	-5293.4	-6196.8
ND2	5981-7030	-5054.47	-5054.47	-4260.57	-5054.47
ND3	7031-7381	-2085.46	-2085.46	-1820.54	-2085.46
ND4	7382-8762	-6356.28	-6356.28	-5603.46	-6356.28
ND4L	8763-9059	-1105.3800	-1105.3800	-992.85	-1105.3800
ND5	9060-10891	-8598	-8491.43	**-8598**	-8491.43
ND6	10892-11413	-2673.95	-2673.95	-2284.9499	-2673.95
12S	11414-12353	-2292.63	-2292.63	-2101.14	-2292.63
16S	12354-14036	-4418.99	-4416.09	**-3982.6**	-4411.17
mtDNA_tRNA	14037-15597	-3811.14	-3873.75	-3668.51	-4021.63
Nuclear	15598-17420	-3020.94	-3020.94	-3052.27	-3022.64
18S	15598-29426	-39940.85	-39940.85	-3014.12	-50944.73
CT	17421-17516	-152.93	-152.93	-145.87	-152.93
Epend	17517-17978	-865.64	-865.64	-796.16	-871.23
GH1c	17979-18789	-3442.42	-3442.42	-2524.17	-3452.95
GH1d	18790-19940	-4846.76	-4846.76	-3656.72	-5005.03
GH2c	19941-20580	-2060.2199	-2060.2199	-1560.68	-2133.0700
GH2d	20581-21779	-4954.7	-4954.7	-3685.17	-5188.91
HMG1	21780-22326	-880.01	-880.01	-829.57	-880.01
ITS1	22327-22788	-2319.39	-2319.39	-1700.29	-2446.54
ITS2	22789-23121	-1483.65	-1483.65	-1041.47	-1560.21
LDH	23122-23505	-828.03	-826.67	**-690.63**	-827.38
MetA	23506-23996	-1272.3	-1272.3	-1050.26	-1309.48
MetB	23997-24931	-2528.21	-2528.21	-2038.32	-2712.01
RAG	24932-26487	-2765.03	-2765.03	-2640.14	-2765.03
Tnfa	26488-27016	-945.35	-945.35	876.61	-945.35
Transf	27017-27938	-1771.16	-1771.16	-1552.23	-1787.04
VIT	27939-29426	-3443.39	-3453.32	-2978.87	-3469.7

Table 2: Log-likelihood scores of using different tools on given sequence alignments

Gene Name	Gene Position	Current	MUSCLE 2	MUSCLE (W/O 3rd)
Cytb	1-1141	-13463.83	-11169.6	-7564.36
mtDNA_Coding	1-11413	-71407.5200	-68347.48	-40151.49
Mitochondrial	1-15597	-83466.0099	-75333.58	
CO1	1142-2692	-13847.75	-12210.38	-7504.05
CO2	2693-3378	-2292.06	-2119.7800	-1482.83
CO3	3379-4164	-3093.56	-2774.02	-1819.22
ATP6	4165-4848	-5302.98	-4465.47	-2894.82
ATP8	4849-5016	-407.4	-375.86	-242.9
ND1	5017-5980	-6196.8	-5293.4	-3995.8
ND2	5981-7030	-5054.47	-4260.57	-2865.65
ND3	7031-7381	-2085.46	-1820.54	-1145.1199
ND4	7382-8762	-6356.28	-5603.46	-3755.95
ND4L	8763-9059	-1105.3800	-992.85	-685.73
ND5	9060-10891	-8598	**-8598**	-6790.5
ND6	10892-11413	-2673.95	-2284.9499	-1664.06
12S	11414-12353	-2292.63	-2101.14	-1462.27
16S	12354-14036	-4418.99	**-3982.6**	-2679.36
mtDNA_tRNA	14037-15597	-3811.14	-3668.51	-2486.5500
Nuclear	15598-17420	-3020.94	-3052.27	-2059.94
18S	15598-29426	-39940.85	-3014.12	
CT	17421-17516	-152.93	-145.87	-89.31
Epend	17517-17978	-865.64	-796.16	-551.5800
GH1c	17979-18789	-3442.42	-2524.17	-1710.5
GH1d	18790-19940	-4846.76	-3656.72	-2472.17
GH2c	19941-20580	-2060.2199	-1560.68	-1037.3900
GH2d	20581-21779	-4954.7	-3685.17	-2511.27
HMG1	21780-22326	-880.01	-829.57	-567.85
ITS1	22327-22788	-2319.39	-1700.29	-1115.02
ITS2	22789-23121	-1483.65	-1041.47	-677.82
LDH	23122-23505	-828.03	**-690.63**	-471.69
MetA	23506-23996	-1272.3	-1050.26	-694.32
MetB	23997-24931	-2528.21	-2038.32	-1354.02
RAG	24932-26487	-2765.03	-2640.14	-1831.11
Tnfa	26488-27016	-945.35	876.61	-581.66
Transf	27017-27938	-1771.16	-1552.23	-1045.8399
VIT	27939-29426	-3443.39	-2978.87	-1962.07

Table 3: Log-likelihood comparison between the given alignment, the MUSCLE alignments with all positions considered, and the MUSCLE alignment without consideration of the 3rd codon position

In Fig. 4, *O. masou* has been clearly placed next to *O. rhodu*. It is important to note here that despite being known as the *Biwa trout*, *O. rhodu* is considered to be a sub-species of *O. masou*. As the species *O. masou* itself is under scrutiny here, we did not deem it appropriate to use the classification of *O. rhodu* as a trout to be a sufficient indicator for similar classification of *O. masou*. As a result, we are forced to look further along the tree in search of its closest relatives.

According to Fig. 4, the next closest relatives to *O. masou* are *O. tshawytscha*, *O. nerka*, *O. gorbuscha* and *O. keta*, which are Chinook, Sockeye, Pink and Chum salmon, respectively.

The nuclear species trees obtained using raxML output displayed a similar trend. Despite the sparseness of data, the nuclear genes show overwhelming support for *O. masou* being a salmon, rather than a trout.

5.2.2 Species trees obtained from mitochondrial genes

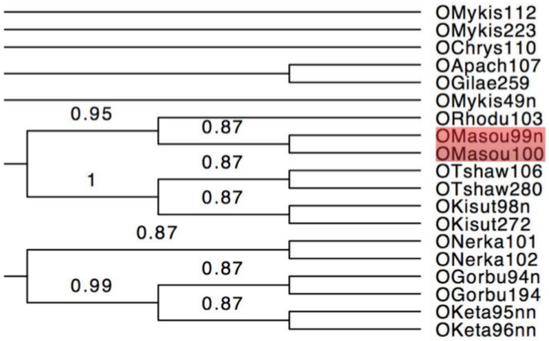

Figure 5: The clade containing *O. Masou* in the species tree for mitochondrial gene trees obtained using MEGA (without 3rd codon).

The Mitochondrial genes provided a more comprehensive and hence, more confident picture of the species tree due to the relative abundance of data as compared to the nuclear genes. Similar to the nuclear species tree, the mitochondrial species tree provides a similar classification for *O. masou*. The closest relatives of *O. masou* in this species tree are *O. tshawytscha* and *O. kisutsch*, Chinook and Coho salmon, respectively. Notice that the support values which, indicate the quartet scores for each branch, are greater than 80%. It is implied that most mitochondrial gene trees consider *O. masou* to be closer to salmon. The species classified as trout, such as *O. mykiss* and *O. gilae* appear as distant relatives of *O. masou* as compared to those classified as salmon.

Thus, the evidence based on both the nuclear and mitochondrial genes indicate that *O. Masou* is most likely a salmon.

6 Conclusions

The species within the family Salmonidae hold great ecological importance as they are often a central part to the ecosystems in which they belong. In particular, salmon species play a key part in ecosystems along the Pacific rim. Their presence provides an indication of the health of rivers and many different species, from grizzly bears to orca whales, depend on salmon to provide the marine-rich nutrients that they require. As such, it is important that we understand their ecological relationships, as well as their life-history.

Building on previous work done in the field, we attempt to address the persisting uncertainty within species-level relationships of the genus *Oncorhynchus*. In particular, we attempt to contribute to the ongoing discussion in regards to the designation *O. masou*: is it a trout or salmon?

Our phylogenetic analysis takes inspiration from previous work; however, it is markedly different in multiple ways. Specifically, we utilize

- A larger amount of data than was previously available
- A greater amount of computational power, resulting in better resolved phylogenetic trees
- New methods that were unavailable at the time of previous analysis. These methods include
    - Bayesian analysis through the program BEAST
    - Removal of 3rd codon position from consideration when doing sequence alignments that results in greater support for the resulting phylogenetic trees
    - Supertree construction

in order to analyze the current established phylogeny of *Oncorhynchus* and observe whether the relationships that exists in that genus persist under these novel methods.

Our investigations find that the current established relationships persist, despite being subjected to new tools and methods. This suggests that the current, established, phylogeny is robust. Moreover, in regards to *O. masou*, we find that, in agreement with Crête-Lafrenière et. al and contrary to results from Crespi and Fulton, *O. masou* is indeed a salmon and not a trout.

It is important to note however that, despite having a larger data set than was previously available, the genetic data set that we used for this analysis still remained severely constrained by a lack of nuclear gene data. Salmonoid mitochondrial genes have large amounts of introgression within them, in part, due to frequent intermating amongst differing species [23]. A key factor that allows this to happen are the mechanisms within the nuclear genes. Therefore, the nuclear genes could derive a derive a whole new phylogeny than the one we have analyzed in this paper. Hence this insufficiency leaves room for future work and the phylogeny of this genus, as well as the phylogeny established for the family Salmonidae as a whole, should be re-examined once more data is available.

7 Acknowledgments

The authors would like to thank Dr. Leonid Chindelevitch (SFU Computing Science), Dr. Bernard Crespi (SFU Biological Science), and Dr. Cedric Chauve (SFU Mathematics) for both helpful discussions, as well as their guidance throughout this project. Cluster time was provided by Compute Canada (Calcul Canada) and WestGrid.

8 Supplemental Material

Attached to this submission is the folder of all codes and acquired data that we created and utilized during this project. The link to the data is http://goo.gl/2um74j

References

[1] Stamatakis, A. (2014). *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 30(9), 1312-1313.

[2] Bouckaert, R., Heled, J., Khnert, D., Vaughan, T., Wu, C-H., Xie, D., Suchard, MA., Rambaut, A., & Drummond, A. J. (2014). *BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.* PLoS Computational Biology, 10(4), e1003537. doi:10.1371/journal.pcbi.1003537

[3] Liu, K., Linder, C. R., & Warnow, T. (2011). *RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation.* PLoS One, 6(11), e27731.

[4] Rodriguez, F. J. L. O. J., Oliver, J. L., Marin, A., & Medina, J. R. (1990). *The general stochastic model of nucleotide substitution.* Journal of theoretical biology, 142(4), 485-501.

[5] Lanave, C., Preparata, G., Sacone, C., & Serio, G. (1984). *A new method for calculating evolutionary substitution rates.* Journal of molecular evolution, 20(1), 86-93.

[6] Tavar, S. (1986). *Some probabilistic and statistical problems in the analysis of DNA sequences.* Lectures on mathematics in the life sciences, 17, 57-86.

[7] Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). *jModelTest 2: more models, new heuristics and parallel computing.* Nature methods, 9(8), 772-772.

[8] Fungiflora, O. S., & Gascuel, O. (2003). *A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood.* Syst Biol, 52, 696704Hjortstam.

[9] Stamatakis, A. (2006, April). *Phylogenetic models of rate heterogeneity: a high performance computing perspective.* In Proceedings 20th IEEE International Parallel & Distributed Processing Symposium (pp. 8-pp). IEEE.

[10] Stamatakis, A., Hoover, P., & Rougemont, J. (2008). *A rapid bootstrap algorithm for the RAxML web servers.* Systematic biology, 57(5), 758-771.

[11] Stamatakis, A., Ludwig, T., & Meier, H. (2005). *RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.* Bioinformatics, 21(4), 456-463.

[12] Huelsenbeck, J. P., & Ronquist, F. (2001). *MRBAYES: Bayesian inference of phylogenetic trees.* Bioinformatics, 17(8), 754-755.

[13] Wilson, I. J., Weale, M. E., & Balding, D. J. (2003). *Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 166(2), 155-188.

[14] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). *Equation of state calculations by fast computing machines.* The journal of chemical physics, 21(6), 1087-1092.

[15] Hastings, W. K. (1970). *Monte Carlo sampling methods using Markov chains and their applications.* Biometrika, 57(1), 97-109.

[16] Drummond, A. J., & Rambaut, A. (2007). *BEAST: Bayesian evolutionary analysis by sampling trees.* BMC evolutionary biology, 7(1), 1.

[17] Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M., & Stamatakis, A. (2009, May). *How many bootstrap replicates are necessary?* In Annual International Conference on Research in Computational Molecular Biology (pp. 184-200). Springer Berlin Heidelberg.

[18] Huelsenbeck, J. P., Ronquist, F., & Hall, B. (2001). *An introduction to Bayesian inference of phylogeny.*

[19] Steel, M., & McKenzie, A. (2001). *Properties of phylogenetic trees generated by Yule-type speciation models.* Mathematical biosciences, 170(1), 91-112.

[20] Gernhard, T. (2008). *The conditioned reconstructed process.* Journal of theoretical biology, 253(4), 769-778.

[21] Yule, G. U. (1925). *A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS.* Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character, 213, 21-87.

[22] Bevan, R. B., Bryant, D., & Lang, B. F. (2007). *Accounting for gene rate heterogeneity in phylogenetic inference.* Systematic biology, 56(2), 194-205.

[23] Crespi, B. J., Fulton, M. J. (2004). *Molecular systematics of Salmonidae: combined nuclear data yields robust phylogeny.* Molecular Phylogenetics and Evolution, 31, 658-679

[24] Crête-Lafrenière, A., Weir, L. K., Bernatchez, L. (2012). *Framing the Salmonidae Family Phylogenetic Portrait: A More Complete Picture from Increased Taxon Sampling.* PLoS One 7(10): e46662. doi:10.1371/journal.pone.0046662

[25] Pais, Fabiano Sviatopolk-Mirsky, et al. *Assessing the efficiency of multiple sequence alignment programs.*Algorithms for Molecular Biology9.1 (2014): 1.

[26] Edgar RC. *MUSCLE: multiple sequence alignment with high accuracy and high throughput.Nucleic Acids Research.* 2004;32(5):1792-1797. doi:10.1093/nar/gkh340.

[27] Katoh, Kazutaka, and Daron M. Standley. *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.*Molecular biology and evolution30.4 (2013): 772-780.

[28] Altschul, Stephen F., et al. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.*Nucleic acids research25.17 (1997): 3389-3402.

[29] 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: 46734680.

[30] Loytynoja A, Milinkovitch MC (2001) *SOAP, cleaning multiple alignments from unstable blocks.* Bioinformatics 17: 573574.

[31] http://www.megasoftware.net/web_help_7/helpfile.htm#hc_first_time_user.htm

[32] Breinholt, Jesse W., and Akito Y. Kawahara. *Phylotranscriptomics: saturated third codon positions radically influence the estimation of trees based on next-gen data.*Genome biology and evolution5.11 (2013): 2082-2092.

[33] Baxevanis, Andreas D., and BF Francis Ouellette.*Bioinformatics: a practical guide to the analysis of genes and proteins.* Vol. 43. John Wiley & Sons, 2004.

[34] Warnow, Tandy, and Binhai Zhu.*Computing and Combinatorics: 9th Annual International Conference, COCOON 2003, Big Sky, MT, USA, July 25-28, 2003,* Proceedings. Vol. 9. Springer Science & Business Media, 2003.

[35] http://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html

[36] Roch and Steel. *Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent* Theoretical Population Biology 100 (2015),pg 5662.

[37] Mirarab, Riaz, Bayzid, Zimmermann, Svenson and Warnow. *ASTRAL: genome-scale coalescent-based species tree estimation.* Vol. 30 ECCB 2014,pages i541i548.

[38] Domanico, M.J., Phillips, R.B., Oakley, T.H. (1997). *Phylogenetic analysis of Pacific Salmon (genus Oncorhynchus) using nuclear and mitochondrial DNA sequences.* Can. J. Fish. Aquat. Sci. 54, 1865 1872.

[39] McPhail, J.D., 1997. *The origin and speciation of Oncorhynchus revisited. In: Stouder, D.J., Bisson, P.A., Naiman, R.J. (Eds.), Pacific Salmon and their Ecosystems: Status and Future Options.* Chapman and Hall, New York, pp. 2938.

[40] Norden, C.R., 1961. *Comparative osteology of representative salmo- nid fishes, with particular reference to the grayling (Thymallus arcticus) and its phylogeny.* J. Fish. Res. Bd. Canada 18, 679791.

[41] Oakley, T.H., Phillips, R.B., 1999. *Phylogeny of Salmonine fishes based on growth hormone introns: Atlantic (Salmo) and Pacific (Oncorhynchus) Salmon are not sister taxa.* Mol. Phylogenet. Evol. 11, 381393.

[42] Harrison and Larson. *Hybridization, Introgression, and the Nature of Species Boundaries.* Journal of Heredity 2014:105(Special Issue):pg 795809

[43] Phillips and Domanico *Phylogenetic analysis of Pacific salmon (genus Oncorhynchus) based on mitochondrial DNA sequence data.* Mol. Phylogenet. Evol. 4,pg 366371.

[44] Jacox, Chauve, Szöllosi, Ponty and Scornavacca *ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony.* Bioinformatics, 2016,pg 13.