Norwegian University of Life Sciences

Master's Thesis 2021 30 ECTS

Faculty of Chemistry, Biotechnology and Food Science

The regulatory network underlying gene expression divergence of gene duplicates in Atlantic salmon

Shatabdi Deb Proma

Animal Breeding and Genetics

Acknowledgement

It is a great pleasure to express my gratitude to my supervisors, Torgeir Rhoden Hvidsten and Marinus te pas for their guidance and overall insights throughout the journey of this thesis. Without their support I would not have been able to accomplish this work within time.

I am really thankful to Gareth Benjamin Gillard for his willingness to provide necessary information regarding the datasets.

Finally it is my privilege to thank Dr. Carl Gunnar Fossdal for agreeing to evaluate my work.

Abstract

Atlantic salmon, provides excellent opportunities for studying vertebrate genome evolution after whole genome duplication (WGD). This remains congruent with the extreme rate of duplicated gene copies following Ss4R (fourth round of whole genome duplication) in the common ancestor of salmonids. However, little is known about the role of TFs in driving duplicate gene expression divergence. Here we aimed at contributing to the understanding of TF evolvability by modelling a TF-gene regulatory network using the Inferelator algorithm for the first time in Atlantic salmon genome. This was achieved by using ATAC-seq data and RNA-Seq gene expression counts. With this network, we studied the tendency of TFs to evolve towards asymmetric expression of duplicate gene copies, where one copy diverted to expression gain leaving another copy retained, performing ancestral function. Firstly, our network analysis implied that Inferelator modelled a biologically meaningful network. Along with this, TF evolvability indicated, presence of conserved TFs, despite the expression dissimilarities between duplicates. Moreover, our gene ontology results suggested that these TFs were mostly involved in the cell cycle function. In conclusion, we suggest that modification in the co-activators of our TFs could explain their being preserved towards asymmetric expression patterns of the duplicates.

Introduction

Atlantic salmon (*Salmo salar*), holds a remarkable position for serving as an economically valuable fish species globally (Houston and Macqueen, 2019). In addition, it contributes to the wild fisheries and recreational sports fisheries (Lien *et al*. 2016). It belongs to the family Salmonidae, comprising of 11 genera along with many other species (Nelson *et al*., 2016). Atlantic salmon appears to be a significant and interesting source for studying genome evolution owing to its dramatic duplication processes unlinked to the evolutionary patters in other vertebrates (Lien *et al*. 2016).

Duplication of genetic materials is a stochastic event that contributes to the evolutionary changes of an organism. It is mainly governed by single gene duplication or whole genome duplication (WGD). WGD is a more common incident for plants than for the animals or vertebrates (Van de Peer *et al*., 2009). However, compared to the root of all vertebrates where there were two WGDs (1R and 2R) (Dehal and Boore 2005), a third level of subsequent WGD event (3R) (Jaillon *et al*., 2004; Nakatani *et al*., 2007) occurred in teleosts species, followed by a fourth round of WGD (Ss4R) in the common ancestor of salmonids (Macqueen and Johnston 2014). The Ss4R or autotetraploidization happened at around >80 million years ago, after a divergence of salmonids from their closest species pike (Macqueen and Johnston 2014; Gillard, 2019). This leads to an additional interest for Atlantic salmon, among other salmonids, because of its genome having an ongoing process of rediploidization, where tetraploid state or Ss4R is shaped back to a diploid state (Lien *et al*. 2016). Although genome evolution has been previously studied in Atlantic salmon, little is known about the post-WGD driven role of transcription factors (TFs) influencing the expression of genes.

WGD or polyploidization (a major driver for changing the entire genomic configuration of an organism), shapes selectively functional traits by creating post-WGD favoured genes (Gillard *et al*., 2020). Previous research has emphasized on understanding the role of different selection constraints generating underlying changes in genomic composition and leading to adaptive phenotypes (Zhao *et al*., 2020). The development of novel traits and adaptation following WGD is propelled by duplication through sub-functionalization or neo-functionalization (Prince and Pickett, 2002; Conant *et al*., 2008) where duplicated genes evolve either by dividing original functions between copies (Nowak *et al*., 1997) or by gaining new function in one copy (He and Zhang, 2005). These two main models explain the loss or retention of paralogs after duplication (Ohno, 1970; He and Zhang, 2005). Conversely, duplicates can be silenced or lost because of the higher frequency of deleterious mutation than the beneficial ones, directing to the phenomenon called pseudogenization (Mungpakdee *et al*., 2008). Another complex explanation, known as gene balance hypothesis, reckons that selection driving long term retention of duplicated genes, is believed to be caused due to dosage balance constraints against loss (Birchler and Veitia, 2012). Hence, evolution of whole genome has gained attention by researchers, leading different approaches to analyse the concept of duplication mode in relation with regulatory divergence of gene expression (Zhao *et al*., 2020).

Prior studies have focused on understanding regulatory dynamics changing gene expression patterns, in numerous species including both prokaryotes (eg: bacteria (McAdams *et al*., 2004)) and eukaryotes (eg: human (Battle *et al*., 2014)) which briefly encompass active chromatin configuration and binding of transcription factors (TFs) to the gene promoters to initiate transcription (Klemm *et al*., 2019). This binding form is determined by a gene regulatory network (GRN) (Thompson *et al*., 2015) connecting a TF to a specific set of genes or a gene with certain TFs (Jones and Vandepoele, 2020), allowing exploration of the knowledge gap between systems biology and regulatory interactions after WGD.

Binding of TFs to promoters or enhancers determines to what extent the genes would express within a network under the regulatory control (Gillard, 2019). It is governed by a change in cis versus trans

regulatory mutations which influence the functional divergence of genes after evolution (Jones and Vandepoele, 2020). The changes in *cis* regulatory elements, i.e. promoter divergence, at target gene levels are associated with gain or loss of TF binding and expression shifts over evolutionary time (Jones and Vandepoele, 2020). Conversely, mutation or protein sequence change (trans) in a TF, lead to consequential change of activities of the target genes associated with the regulation of that particular TF (Nowick and Stubbs, 2010). Hence, alteration of specific TFs can have impact on their downstream gene expressions within a regulatory network (Nowick and Stubbs, 2010). Therefore, in order to infer evolutionary changes in the biological networks (Nowick and Stubbs, 2010), it is necessary to delineate the functional divergence of the TFs towards expression variation of genes.

Evolution of duplicated genes have been extensively studied in plants and yeasts, however, here we have a potential curiosity in the WGD driven underlying mechanisms in Atlantic salmon. Our interest harmonizes with the extreme retention rate of paralogs in Atlantic salmon (Carmona-Antoñanzas *et al*., 2013). In addition, autotetraploidization in Atlantic salmon has initiated evolutionary forces to have consequence in the expression of genes leading up to adaptive gain (Carmona-Antoñanzas *et al*., 2013). This is supported by Gillard *et al*. (2020) who used a phylogenetic Ornstein- Uhlenbeck (OU) model (Rohlfs *et al*., 2014) and found liver specific gain in the gene expression of Atlantic salmon following WGD. The paper stated that the acquisition of new TFBSs was associated with the increased expression in one of the duplicates (Gillard *et al*., 2020). Depending on the binding of TFs to the regulatory sites, the paralogous copies increase or decrease expression which is activated by specific TFs. To regulate this expression dynamics, which TFs are involved to change expression in the same or different direction still needs to be resolved. In other words, it generates a research question "**Are there some TFs, more prone to evolvability that can lose or gain targets more easily across the genome on a short time scale?"**.

Here we tried to address a part of the aforementioned question by two steps analysis. Initially we build a TF-gene regulatory network by including a prior information table on TF-gene interaction besides RNA counts. For this we used ATAC-seq reads from Atlantic salmon species to link gene expression to TFs. Then for modelling this global regulatory network we used an algorithm called Inferelator (Miraldi *et al*., 2019). Later, we used the upregulated + conserved (up+cons) copies of paralogs with which we wanted to explore TFs that were more inclined to switch towards upregulated copies compared to their conserved partners and generate evolvability over evolutionary time scale. We expect that TFs showing differences in shifting between targets would provide insights about understanding the TF evolvability towards gene expression divergence in Atlantic salmon.

Glossary

Paralogs: Copies of genes that have duplicated within the same genome over evolutionary timescale.

Active chromatin configuration: Open regions of the chromatin where TFs have gained access to bind and regulate gene expression (Klemm *et al*., 2019).

GRN: The gene regulatory network represents the genes and their interactions as nodes and edges respectively (Thompson *et al*., 2015). Nodes are basically comprised of regulators like TFs, mi-RNA, signalling proteins and specific genes whereas edges are the direction toward their targets (Thompson *et al*., 2015).

TFs: Transcription factors are proteins, associated with transcription process of DNA to messenger RNA and expression of specific genes.

ATAC-seq: Assay for Transposase-Accessible Chromatin using sequencing, preloaded with Tn5 transposase, is used to find the open chromatin regions in the DNA (Klemm *et al*., 2019). This enzyme cuts and inserts adaptors into the DNA which facilitates the tracking of the Tn5 transposase to detect the open regions (Klemm *et al*., 2019). An open region/accessible region is important for the TFs to get access to their binding sites in order to initiate transcription.

RNA-seq data: It is obtained by converting the RNA to cDNA, following sanger sequencing or next generation sequencing (NGS) techniques with an addition of adapters to both ends of the cDNA fragments that allows sequencing (Wang *et al*., 2009). As an output, we get the reads of varying length because of different NGS techniques (Wang *et al*., 2009). The reads are mapped to the genome and the number of reads mapped to each gene is counted (Wang *et al*., 2009). These counts, normalized for the total number of reads in the sample, are then used for downstream gene expression analysis .

Inferelator: It is a method that uses standard regression model to infer TF-gene regulatory network (Miraldi *et al*., 2019). Two important datasets are required as an input to this method; the gene expression dataset and the prior information table where we have previous knowledge about the gene-TF interaction (Miraldi *et al*., 2019). When a gene is regulated by TF, it is scored in the prior as 1 or more based on the number of transcription factor binding sites or 0 if there is no interaction. Another element is the gold standard interaction matrix, which can be used to evaluate the accuracy of the inferred network. This contains the predictors for which we can estimate the absence or presence of an interaction with a score of 0 or 1 accordingly. Generally gold standard contains interactions for few TFs whereas prior includes more genome wide data. But here we have gold standard that is not different from the prior.

DNAse-seq: An endonuclease called Deoxyribonuclease I (DNAse I) cleaves the accessible DNA irrespective of the highly dense chromatin region (Klemm *et al*., 2019). After the library sequencing, hypersensitive sites can be easily recognized as the accessible regions.

Asymmetrical evolution: Is the phenomenon when expression of both gene copies increase or decrease together after duplication.

Autotetraploidization: Fourth round of whole genome duplication event that occurred in the salmonids common ancestry (Lien *et al*., 2016).

TFA: Here, TFA represents the profiling of TFs by measuring the expression of genes of the TFs in the prior (Castro *et al*., 2019). This is basically a determination of transcription initiatory or inhibitory factors like chromatin configuration, post-translational regulation of proteins, protein-protein interactions that control the status of TFA in a cell (Castro *et al*., 2019).

Jasper: It contains sequence motifs or binding profile of the TFs to which they bind.

TPM: Transcripts per million is basically a normalization method for the RNA-seq data ("TPM", 2016). In general, this is calculated in three steps: Initially, for each sample, the reads per kilobase (RPK) value is computed as reads divided by the length of each gene in kilobase; then the sum of the RPK values are divided by 10^6 that gives a scaling factor which is used to divide each RPK value in the third step ("TPM", 2016).

Confidence score: This score is estimated to get an idea about the rank of individual edges in the TF- gene regulatory network. To obtain robust predictions, TF-gene interactions are predicted from different subsets of the data (i.e. bootstrap-samples) and the final confidence score for each interactions is based on how high that interaction ranked across all the different subsets.

Precision, Recall, MCC and F1: All of them are the predictors to measure the performance of the model.
To understand these 4 metrices, we drew a confusion matrix table for our datasets:

	Predicted	Predicted
Actual	**True positive** (It happens when an interaction is predicted by the Inferelator that is actually present in the gold standard)	**False negative** (When there is an interaction not predicted by the Inferelator that is actually present in the gold standard)
Actual	**False positive** (When there is no interaction in the gold standard but is predicted by the Inferelator)	**True negative** (When there is no interaction in the gold standard and is not predicted by the Inferelator)

In the context of this table, we measured the performance metrices as following:
Precision: It determined, of the interactions that are predicted by the model, the fraction that are actually in the gold standard. The formula for the calculation is:
Precision = $\frac{TP}{TP+FP}$ ("Precision and recall", 2021)
Recall: It determined, of the interactions that are actually in the gold standard, what fractions are predicted by the model. The formula for the calculation is:
Recall = $\frac{TP}{TP+FN}$ ("Precision and recall", 2021)
F1: It was calculated by combining both precision and recall and known as weighted harmonic mean of both performance metrices ("F-score", 2021). The formula for the calculation is :

F1 = $\frac{2\times(Precision\times Recall)}{Precision+Recall}$ ("F-score", 2021)

(https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/).

MCC: The Matthews correlation coefficient is another performance metric that was measured because of our interest in both positive and negative classes of the confusion matrix. It is basically ranged between -1 and +1, whereas -1 means the absence of an interaction in the gold standard that is predicted by the model: 0 means the model predicts an interaction randomly; +1 means the presence of an interaction in the gold standard that is predicted by the model ("Matthews correlation coefficient", 2021). The coefficient is calculated by using the following formula.

MCC = $\frac{TP\times TN-FP\times FN}{[(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)]^{1/2}}$ ("Matthews correlation coefficient", 2021)

Area under precision-recall curve (AUPRC): The area under PR curve (AUPRC) displays the quality of the method regarding predicting interactions. If it is 1, it means the model has predicted all the interactions from the prior while minimizing predicted interactions outside the prior. In other words, a value of 1 indicates that there are no false positives or false negatives.

Materials and methods

RNA-seq data

We got expression datasets that are available on EBI databases with project codes (PRJEB24480, PRJEB30483, PRJEB34437). Moreover, the data also contained samples, collected during smoltification, that can be found in the github repository page (https://gitlab.com/garethgillard/megaLiverRNA/-/blob/master/data/local_GSF2_SE.tsv). From these projects, only liver tissue samples were used for our analysis. Briefly, they used RNA-seq data for Atlantic salmon, obtained through a dietary based feeding trial (Gillard, 2019) from which we only focused on 146 fresh water samples. Sequence reads were mapped and quantified for estimating counts and Transcripts Per Million reads (TPM) by using STAR (Dobin *et al*., 2013) and RSEM (Li and Dewey, 2011) respectively. Afterwards, we used the TPM counts for the fresh water samples and took the logarithmic values of the counts for further analysis.

ATAC-seq data for network analysis:

The raw ATAC-seq data, available in the database ArrayExpress with accession number: E-MTAB-9001, were mapped and filtered for quality check. Reads were used by TOBIAS tool (Bentsen *et al*., 2020) to perform TF-footprinting analysis, including TF motifs from JASPER (Fornes *et al*., 2020). For footprinting, Tobias uses ATAC reads that undergo bias correction for Tn5 cutsites (Bentsen *et al*., 2020). This allows estimsation of footprinting scores based on the depth and accessibility of the local footprint (Bentsen *et al*., 2020). Tobias, then scrutinizes and generates these scores by using TF motifs in order to account for TF binding sites in the DNA, allowing its capability to differ between bound and unbound regions in the genome (Bentsen *et al*., 2020).

Generating data for running into the Inferelator:

For the downstream network analysis, we considered duplicates having both increased and decreased expression in one copy compared to their conserved copy (up + conserved and down + conserved) that gave us in total of 3184 genes. For each gene, we looked for 3000bp upstream and 200bp downstream of the transcription factor binding sites. Afterwards, we matched the Jasper TFs to salmon TFs through blasting for finding the regulators in salmon for these targets, followed by translation of the UniProt TF sequences. For each TF, a maximum of four genes were selected with an evalue < 1E – 10 and alignment length > 100. Later, to construct TF-gene interaction network, we created a prior matrix with genes in the rows and regulators in the columns, having a value of 1 or more, if TF binds upstream of the genes depending on the footprinting datasets and 0, if there is no interaction. This resulted in the reduction of the number of targets, leaving 2426 number of genes as well as 729 TFs in the prior list.

We then took the liver expression data where we performed a clustering analysis for averaging the expression in samples that are very similar. This was done by setting a correlation threshold of 0.975.

Afterwards, we calculated the expression similarities between TFs and their targets. Initially, we estimated the average correlation of the TFs with their targets in the priors. Moreover, to get the TFs with similar binding profiles (i.e. similar targets), we clustered the prior matrix. TFs having more than 50% of the targets in common ended up in the same cluster.

Eventually we selected 100 TFs from the prior table to run as an input to Inferelator depending on their high correlation with the targets and target redundancy. The selection was performed manually by looking for the functional importance of individual TFs from the literature review. Furthermore, we

tried to keep a majority of the salmon TFs having an average correlation of 0.40 or more with their targets in the prior. Inferelator then ran the regression analysis using both expression and the prior table.

TF-gene regulatory network:

In order to compute the network, the method first tried to get the transcription factor activity (TFA) profile. Hence, it looked at the expression of the targets of the TFs rather than looking at the expression of the TF itself. TF activity is proportional to the expression of the genes in the prior that it regulates (Miraldi *et al*., 2019). So the method used the knowledge from the prior table that contains the information of the TF regulatory interactions with their targets and the expression of the genes which forms the TFA. The regression was then performed to find out which genes the specific TFs regulate using their activity profile.

To infer the network, the regression of the model attempted to calculate the expression of a gene in a sample that is equal to the weighted sum of the activity of all the TFs (Miraldi *et al*., 2019). If the weight was significantly different from 0, this delineated that the TF affected the expression of the specific gene.

$$x_{ij} = \sum_{k \in TFs} b_{ik} a_{kj}$$

Here, x_{ij} stands for the expression level of gene i in condition j; a_{kj} symbolizes the TFA for TF k in condition j and b_{ik} represents the weight of the TF k on gene i (Miraldi *et al*., 2019***).***

Network analysis:

After running the Inferelator, we got a network in the form of a TSV file. The outflow of the algorithm can be found in the github repository (https://inferelator.readthedocs.io/en/latest/results.html). Briefly, the columns named targets and regulators are the genes and TFs correspondingly, which are ranked from highest to lowest according to the combined confidences score. The value of gold standard interactions is predicted as 1 depending on the presence of genes and TFs in the gold standard list and 0 while absent. Here, the gold standard is complementary to the priors in the network where the prior column contains the values in the prior network. The scores in precision, recall, MCC and F1 columns are the values calculated based on all the links predicted to have TF-gene interaction upwards of that specific row. In contrast, scores underneath that row are considered as values without TF-gene interaction. Rows having 0 and 1 indicate that the algorithm predicts interactions which is either absent or present in the gold standard matrix accordingly. Moreover, rows with NA value means that the genes or TFs are not present in the gold standard at all.

GO enrichment analysis:

The gene ontology enrichment analysis was conducted with R package "salmonfisher" (https://gitlab.com/sandve-lab/salmonfisher) and topGO (Alexa and Rahnenführer, 2009). We used salmonfisher to get the GO ids and topGO for obtaining the ontology terms. We tested the significance of the GO terms by performing classic fisher methods that used the threshold for p value < 0.01 (Alexa and Rahnenführer, 2009). We ran this test on each regulator to obtain the GO annotation for their targets individually.

Comparison of P values between true and randomized network:

In order to check the performance robustness of our network, we compared the p-values of the significant GO ids between both networks. Randomization was based on a bootstrapping approach of the targets from our network. We tested if the targets of each 100 TF having enriched gene functions differ between networks in the context of their significant p-values.

GO similarity tree:

The tree has been constructed for all the significant GO ids obtained for each regulator or TF. For this, we made a dissimilarity matrix by using jaccard method. Afterwards, we have performed a hierarchical clustering analysis "hclust" on the matrix that uses the method "ward.D2", resulting in clusters, followed by squaring of the dissimilarities. All of the functions come from the package stats in R (R Core Team, 2013).

Please note that our tree does not reflect phylogenetic relationship among TFs.

Testing for TFs by combining all paralogs:

To test if there were overlapping TFs between paralogs (upregulated and conserved), we made a contingency table for both copies, where the rows presented the upregulated TFs and the columns showed the conserved TFs. Furthermore, we ran fisher test on the contingency table for all the paralogs to get the significant TFs by setting the pvalue threshold to less than 0.01. Later, we manually corrected the statistical output for multiple testing in order to get the final outcome.

Results and discussion

We evaluated the effect of TFs on the expression of duplicated genes using the Inferelator model.

Performance metrices:

The model predicted 56066 interactions among 2426 genes and 100 TFs. However, the prior contained 19108 interactions between TFs and genes. Hence, the performance metrices evaluated how well Inferelator predicts the true interactions from the prior.

We saw that with decreasing precision, recall increases in the precision-recall (PR) curve (Figure 1A). Precision goes down as the confidence score gets lower (Supplementary file). On the other hand, we saw an increase in recall because the method predicted more of the interactions in the prior as it makes more predictions.

In Figure 1B and 1C, both F1 and MCC scores increased with response to lower confidence score. For the confidence of 0.0755 (MCC) and 0.0378 (F1), resulting in 26046(F1) and 17843(MCC) predicted interactions, we observed the highest F1 and MCC score of 0.53 and 0.4787 respectively. From Figure 1D, we observed that most of the interactions had low confidence score whereas few had a very high confidence. This goes parallelly with the selection of interactions for which the predictions had low confidence score.

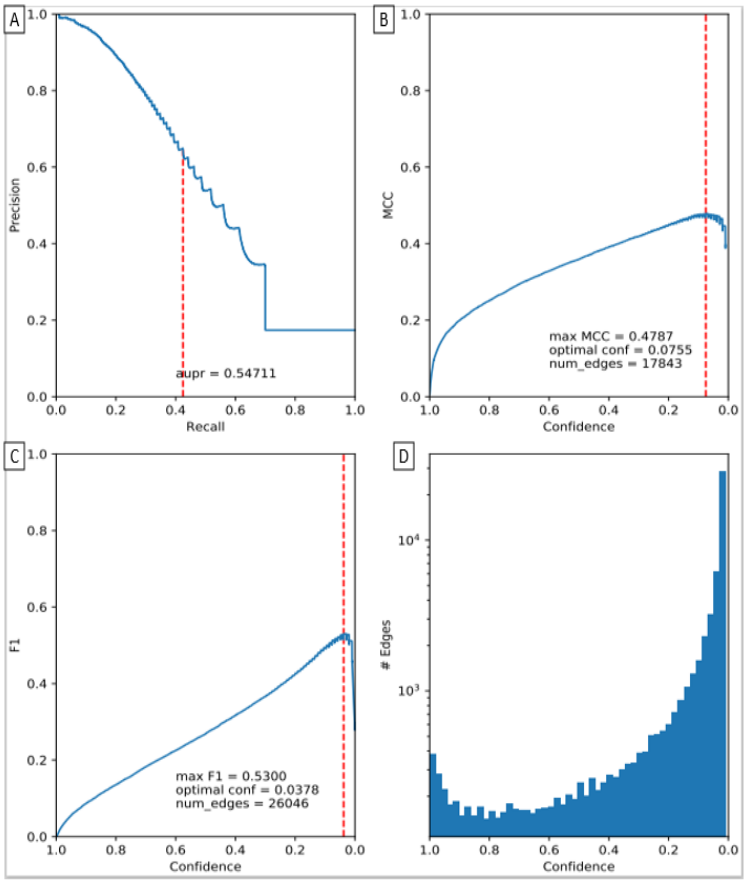

Figure1A, 1B and 1C show the performance metrics on the y- axis increasing from 0 to 1 while x axis displays confidence score decreasing towards the right (except 1A). 1D represents number of edges for the level of confidence score for the average TF-gene interactions.

We also ran the method on the same input data despite selecting for 100 TFs (Figure 2). We included all 729 TFs which gave us a total of 13122 predicted interactions, of which 7649 were in the prior. The performance metrices had very poor scores for both F1 and MCC compared to the previous performance scores. Moreover, The AUPR represented a very low quality (compared to Figure 1A) prediction of the method for predicting interactions. Therefore, we continued with the first network for the rest of our analysis.

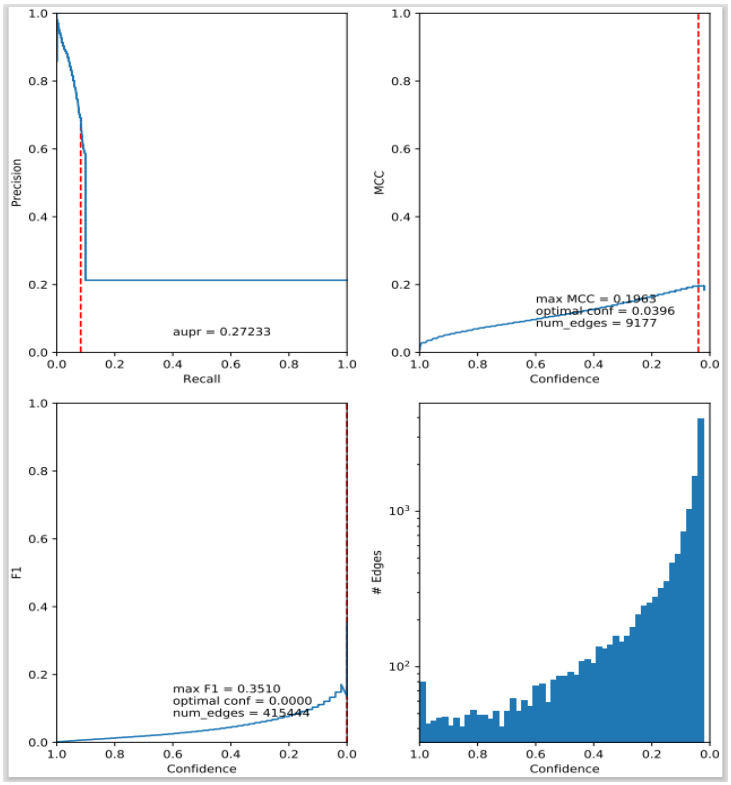

Figure2A, 2B and 2C show the performance metrics on the y- axis increasing from 0 to 1 while x axis displays confidence score decreasing towards the right (except 2A). 2D represents number of edges for the level of confidence score for the average TF-gene interactions.

P-values for true and randomized network

To check the validity of our (true) network, we compared the p-values of our GOs with those from the bootstrap results corresponding to a randomized network. We observed a clear difference between the p-values for both networks where the peak for the true network was at the left side of the plot with maximum values close to 0 (Figure 3). On the contrary, it was mostly at the right side for the randomized network with larger values close to 0.01 (Figure 3). Moreover, for the bootstraps, the randomized network had a highest GO ids count for the peak at around 80. Conversely, for the true network, it had values higher than 80 for the majority of GO ids. The plots showed that our result was not random and therefore Inferelator models a meaningful network.

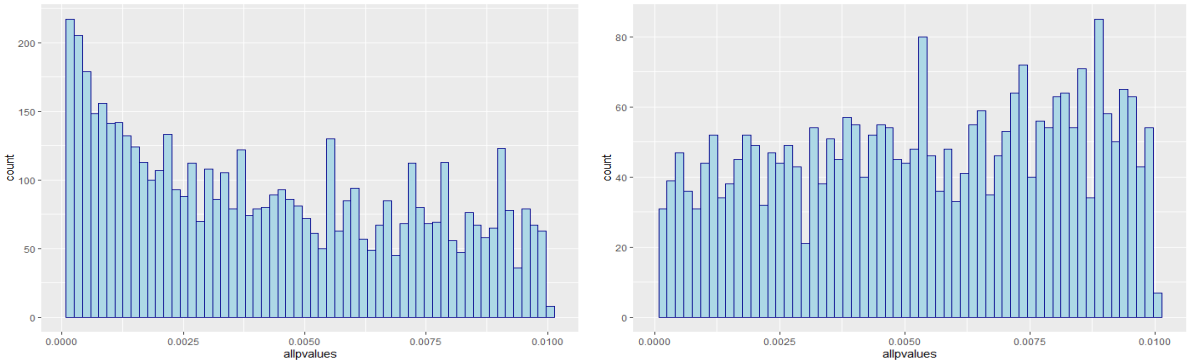

Figure 3. It represents the p-values of all the significant GO ids, for 100 TFs corresponds to a true (left) and a randomized network (right). The y axis shows the counts for all the GO ids and x axis shows significant p values between 0 and 0.01. The left plot presents the true network and right one is for the randomized network.

By comparing both networks we observed that Inferelator worked well when we used a higher number of genes than the number of TFs. The first network (true) contained a 24 times higher number of genes compared to the later network (randomized) that had nearly 3 times more genes than that of TFs. This could be because of the overfitting of our model, however it requires additional approaches to confirm the concept.

Hierarchical clustering for GO similarities

We investigated the 100 TFs that we chose closely to check their modularity at the functional level. In other words, we were interested in exploring if there are groups of TFs sharing the same GOs, therefore, performing similar functions. We observed that TFs that have very small distance, shared GO ids between them (Figure 4). Compared to this, TFs that are very distant from each other do not shared GO ids (Figure 4). While looking at the bootstraps result we found that the bigger clusters were not meaningful at all or happened by chance. This suggests that our TFs are different from each other and performing specialized functions because of the little overlaps regarding GO ids between them. Figure 5 shows the number of TFs per GO id. We used the total number of GO ids for all 100 TFs. For the majority of the GO ids (more than 1500) there was only one TF present and for very few GO ids (around 4) showed a comparatively higher number of TFs (approximately 10) that overlapped, thus supporting our previous argument concerning Figure 4.

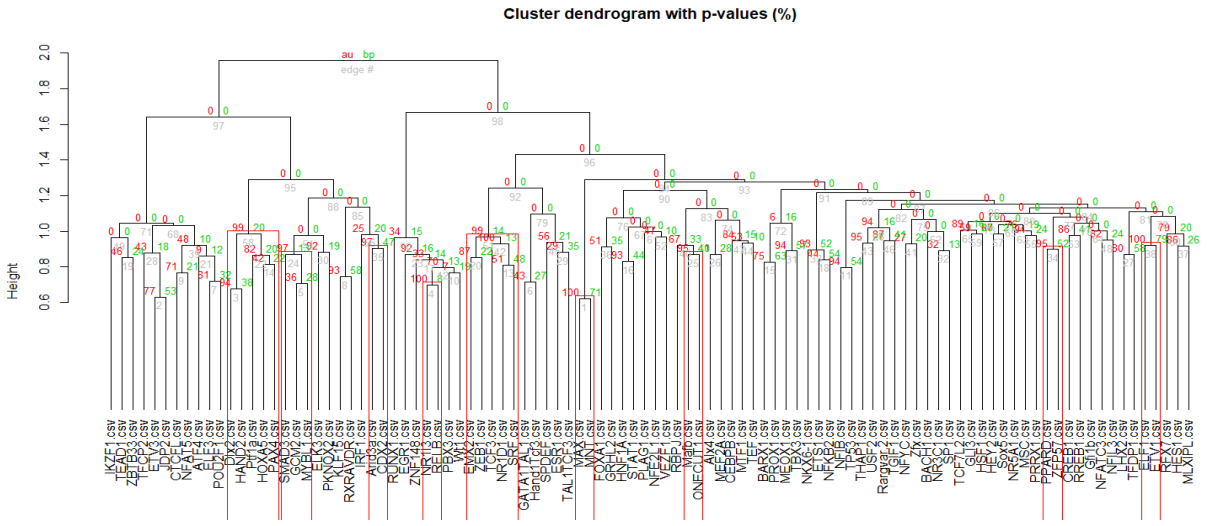

Figure 4. The bootstrap result for the hierarchical clustering of 100 TFs based on their GO similarities. Here, in the y axis the height is the square of the TFs dissimilarities and x axis shows the distance between and among TFs. The numbers are the percentage of times when we get that branch in the bootstrap.

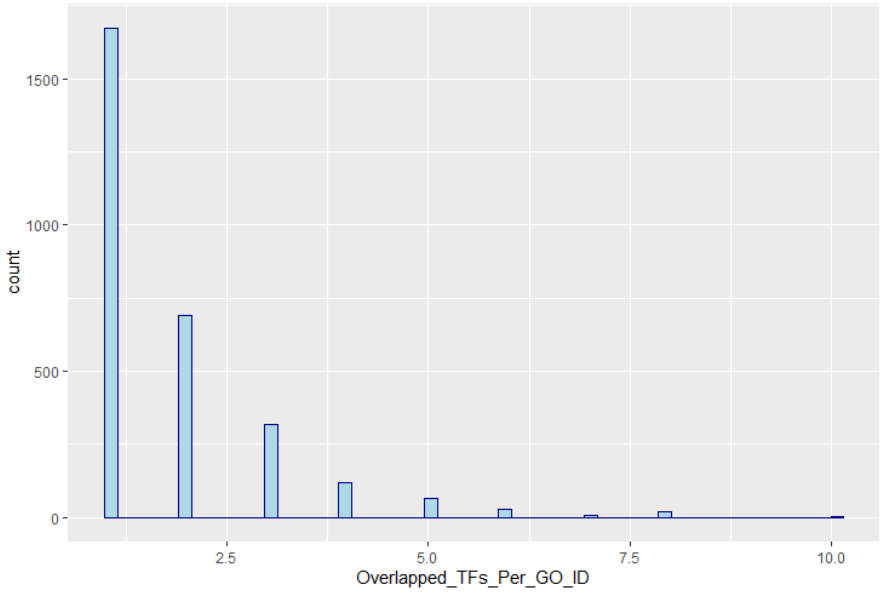

Figure 5. This plot shows the number of overlapped TFs for the GO ids where the y axis is the number of count corresponding to the GO ids and x axis displays the number of TFs.

As the pattern of clustering with 100 TFs shows few shared GO between and among them, this raises the question: can we expect the similar pattern if there was a network with all the TFs instead of 100? In other words, if we do have all the TFs in the network, can there be modularity in the network like what we found for the TFs (Figure 4)? However, as we do not have clusters among the TFs (Figure 4), it could mean that there is no modularity at our network level as well. Alternatively, if there is modularity, it could be because of our incomplete sampling as we used a very small subset of the total TFs. However, we could not check the network modularity because of our time limit which can be measured by using local and global clustering coefficient to capture the degree of modularity of the network (Watts and Strogatz, 1998).

Next, as we took samples from the liver, we were interested to closely look into the GO enrichment for TFs like HNF1A, KLF15 and FOX1A, highly expressed in liver and associated with lipid metabolism (Gillard *et al*., 2020). Also, we wanted to check if our model is biologically meaningful. Hence, we expected the TFs to be involved in the functions that were directly or indirectly related to liver. Our enrichment results showed that HNF1A, KLF15 and FOX1A are involved in several other functions along with their earmarked specificity in liver/lipid metabolism. Please note that for the majority of the results, in order to compare with our GO terms, we used information from human and mice. Therefore, sometimes, it can be uncertain to happen in fish.

Krüppel-like factor 15 (KLF15) has been reported to play a crucial role in cell proliferation (Wang *et al*., 2018). These authors stated that KLF15 inhibits cell cycle process by upregulating CDKN1 A/p21 and CDKN2A/p15 complex in human (Wang *et al*., 2018). CDKN1A or cyclin- dependent kinase inhibitor 1 A, a tumor suppressor , arrests cell cycle event, when binding of p21 to it blocks Cdk2 and Cdk1 activities and blocks entry into S phase from G1 phase (Alberts *et al*., 2002). Moreover, KLF15 arrests cell cycle progression in mouse by inhibiting DNA replication (Ray and Pollard, 2012). On the other hand, there could be a trade-off between liver lipogenesis and gluconeogenesis, initiated by KLF15 (Takeuchi *et al*., 2016). These authors proposed that during fasting in mice, KLF15 could supress lipogenesis by arresting its associated downstream lipogenic genes and initiate gluconeogenesis (Takeuchi *et al*., 2016). These findings coincide with our enrichment status for this TF. We found its involvement in G1/S transition checkpoint, interphase, purine nucleoside triphosphate metabolic

process and in electron transport chain (Table 1). This could mean that KLF15 promotes DNA replication at the S phase when there is enough energy. And so, during fasting phase, due to lack of energy, this replication stops and glucose metabolism is initiated.

Hepatic nuclear factor 1-α (HNF1A), plays an important role in the development of mammalian liver and kidney (Lau *et al*., 2018). Its function in mammals includes, arresting cell cycle event by blocking G2/M phase (Zeng *et al*., 2011). Moreover, it blocks the glucose energy metabolism that initiates anti-proliferative mechanism in mammals (Wang *et al*., 2019).

Conversely, function of HNF1A in the glucose metabolism of fish is not well studied. HNF1A is involved in the secretory function of insulin in regards to glucose in mammals (Beysel *et al*., 2019). Moreover, Kuo *et al*., (2015) reported that steroid hormone like glucocorticoids elevation increases blood glucose level (hyperglycaemia) when there is not sufficient production of insulin from pancreatic Beta cell. However, like mammals, hyperglycaemia in fish is not reported to be a clinical case due to lack of insulin production, although (Moon, 2001) reported it to be a potential reason. There is a controversial topic about teleost being glucose intolerant or not (Moon, 2001). Navarro *et al*., (2002) reported that teleost clears loads of glucose slowly compared to mammals. But it is not clear how do they do that to alleviate increased glucose level in blood. In mammals, in order to ensure gradients for glucose transport, they need to be phosphorylated to glucose-6-phosphate by hexokinase (HK) which requires glucokinase (GK) expression in liver (Niswender *et al*., 1997; Moon, 2001). Their presence is important for maintaining insulin concentration in mammalian blood (Niswender *et al*., 1997). To regulate this glucose-6-phosphate system in mammals, HNF1A plays a key role in glucose homeostasis (Moon, 2001; Lau *et al*., 2018). In Atlantic salmon, GK activities were found to be reported by (Tranulis *et al*., 1996). However, the regulatory mechanism of HNF1A is not clear in fish. Therefore, with the reference from (Kuo *et al*., 2015, Zeng *et al*., 2011 and Wang *et al*., 2019), including our GOs (Table 2), we suggest that in response to increased glucose in blood, HNF1A regulates the steroid hormone receptor signalling pathway followed by decreasing glucocorticoid levels in blood up to a certain point. After that, it shunts glucose energy metabolism to prevent cell proliferation.

After combining both literature studies and enrichment results, we did not find any link between our GOs (Table 3) and the expression of FOXA1 in liver. Hence, we thought this could be because of an artifact of our model. But, after checking for the distribution of the slopes (data not shown), we found that it upregulates many genes. Therefore, this TF could be doing pleotropic functions in both liver and brain in Atlantic salmon.

Table 1 describes the top 5 gene enrichment analysis for KLF15 performed with Fisher's exact test. The GO terms stand for the biological processes associated with the corresponding GO id. Annotated means the number of genes that are annotated for that specific GO id, among which the method finds significant and expected number of genes. Class fisher represents the corresponding p-value score for the most significant GO term.

Serial No.	GO.ID	GO Terms	Annotated	Significant	Expected	Classfisher
1	GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	23	13	4.56	0.00010
2	GO:0044819	mitotic G1/S transition checkpoint	18	11	3.57	0.00014
3	GO:0022900	electron transport chain	30	15	5.95	0.00019
4	GO:0009144	purine nucleoside triphosphate metabolic process	57	23	11.3	0.00025
5	GO:0051325	interphase	19	11	3.77	0.00027

Table 2 describes the top 5 gene enrichment analysis for HNF1A performed with Fisher's exact test. The GO terms stand for the biological processes associated with the corresponding GO id. Annotated means the number of genes that are annotated for that specific GO id, among which the method finds significant and expected number of genes. Classfisher represents the corresponding p-value score for the most significant GO term

Serial No.	GO.ID	GO Terms	Annotated	Significant	Expected	Classfisher
1	GO:0061005	cell differentiation involved in kidney development	13	9	2.54	0.00013
2	GO:0033145	positive regulation of intracellular steroid hormone receptor signalling pathway	5	5	0.98	0.00028
3	GO:0046339	diacylglycerol metabolic process	14	8	2.73	0.00195
4	GO:0030879	mammary gland development	51	19	9.95	0.00216
5	GO:0001990	regulation of systemic arterial blood pressure by hormone	9	6	1.76	0.00262

Table 3 describes the top 5 gene enrichment analysis for FOXA1 performed with Fisher's exact test. The GO terms stand for the biological processes associated with the corresponding GO id. Annotated means the number of genes that are annotated for that specific GO id, among which the method finds significant and expected number of genes. Classfisher represents corresponding p-value score for the most significant GO term

Serial No.	GO.ID	GO Terms	Annotated	Significant	Expected	Classfisher
1	GO:0007033	vacuole organization	36	21	10.2	0.00014
2	GO:0021533	cell differentiation in hindbrain	10	8	2.83	0.00102
3	GO:0021681	cerebellar granular layer development	10	8	2.83	0.00102
4	GO:0021683	cerebellar granular layer morphogenesis	10	8	2.83	0.00102
5	GO:0019827	stem cell population maintenance	52	25	14.73	0.00175

Testing for paralogs

To test for TFs among paralogs, we investigated a total of 612 paralogs (306 upregulated + 306 conserved). We explored if there were TFs that overlapped among paralogs or that appeared in only one of the copies. In other words, we tested if there were TFs with presence in one paralog is independent from the presence in the other paralog. While testing for significance, we found the majority of the TFs to be close to p value 0 (Figure 6). After correcting for multiple testing it resulted in four TFs (Table 4). In this test we were interested in the TFs with the lowest p-value and those who had observed greater than expected for false-true or true-false combinations. Moreover, this means that if we find these four TFs in one paralog, they are most likely to be absent in the other paralog. However, the contingency table (Table 4) showed that all these four TFs (CTCFL, ERG1, EMX2, ETV3) had enrichment in the observed side higher than the expected for false-false and true-true combinations. On the contrary, they did not have enrichment for the other diagonal. Therefore, it means that if we find any of these TFs binding to one paralog, they are most likely to bind to the other copy, despite the fact that one copy is upregulated and other is not.

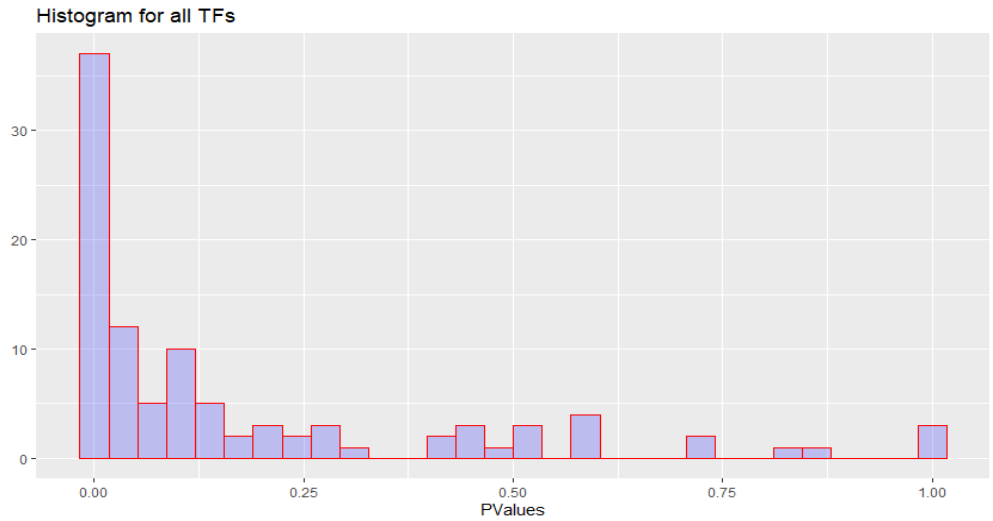

Figure 6. It represents the histogram corresponding to the p-values for the 100 TFs. Y axis is the number of TFs and x axis is the p-values concerning to the number.

Table 4. Contingency table for the TFs that have shown significance in the fisher test where the TFs are presented as a row. We quantified both observed and expected instances to test our hypothesis. Our null hypothesis was that TFs bind to both paralogous copies, and alternate hypothesis was that they bind to one of the copies. The orange colour shows that we have higher number in the observed state than the expected while the green colour represents the opposite pattern.

Contingency Table			Conserved Paralog			
			False		True	
Upregulated Paralog	False		Observed	Expected	Observed	Expected
		CTCFL	191	176.72	37	52.785
		EMX2	177	162.945	40	54.315
		ETV3	195	176.715	42	58.905
		EGR1	199	186	34	46.512
	True		Observed	Expected	Observed	Expected
		CTCFL	45	58.905	33	17.595
		EMX2	52	66.555	37	22.185
		ETV3	35	52.785	34	17.595
		EGR1	45	58.752	28	14.688

Next, as the abovementioned TFs bind to both paralogs or neither and also because the copies had different expression, we investigated the top 5 biological processes that these TFs are involved in. Thereby, we hypothesized that increased expression of one copy but not the other copy could possibly be related to the fundamental functions initiated by these TFs, which can be unbalanced if there is a change in the TFs with respect to the expression patterns of the paralogs.

Table 5 describes the top 5 gene enrichment analysis for CTCFL performed with Fisher's exact test. The GO terms stand for the biological processes associated with the corresponding GO id. Annotated means the number of genes that are annotated for that specific GO id, among which the method finds significant and expected number of genes. Classfisher represents corresponding p-value score for the most significant GO term

Serial No.	GO.ID	GO Terms	Annotated	Significant	Expected	Classfisher
1	GO:0043484	regulation of RNA splicing	46	26	13.72	0.00013
2	GO:0033120	positive regulation of RNA splicing	9	8	2.68	0.00040
3	GO:0035264	multicellular organism growth	83	39	24.76	0.00058
4	GO:0016569	covalent chromatin modification	150	63	44.75	0.00069
5	GO:0016570	histone modification	150	63	44.75	0.00069

Table 6 describes the top 5 gene enrichment analysis for EGR1 performed with Fisher's exact test. The GO terms stand for the biological processes associated with the corresponding GO id. Annotated means the number of genes that are annotated for that specific GO id, among which the method finds significant and expected number of genes. Classfisher represents corresponding p-value score for the most significant GO term

Serial No.	GO.ID	GO Terms	Annotated	Significant	Expected	Classfisher
1	GO:0030097	hemopoiesis	247	86	61.04	0.00010
2	GO:0048705	skeletal system morphogenesis	59	28	14.58	0.00010
3	GO:0002520	immune system development	262	90	64.75	0.00012
4	GO:0019222	regulation of metabolic process	1234	342	304.98	0.00012
5	GO:0001568	blood vessel development	170	63	42.01	0.00013

Table 7 describes the top 5 gene enrichment analysis for EMX2 performed with Fisher's exact test. The GO terms stand for the biological processes associated with the corresponding GO id. Annotated means the number of genes that are annotated for that specific GO id, among which the method finds significant and expected number of genes. Classfisher represents corresponding p-value score for the most significant GO term

Serial No.	GO.ID	GO Terms	Annotated	Significant	Expected	Classfisher
1	GO:0042886	amide transport	452	178	144.54	0.00011
2	GO:0006886	intracellular protein transport	282	118	90.18	0.00012
3	GO:0015031	protein transport	433	171	138.47	0.00014
4	GO:0015833	peptide transport	445	175	142.3	0.00015
5	GO:0007098	centrosome cycle	35	22	11.19	0.00015

Table 8 describes the top 5 gene enrichment analysis for ETV3 performed with Fisher's exact test. The GO terms stand for the biological processes associated with the corresponding GO id. Annotated means the number of genes that are annotated for that specific GO id, among which the method finds significant and expected number of genes. Classfisher represents corresponding p-value score for the most significant GO term

Serial No.	GO.ID	GO Terms	Annotated	Significant	Expected	Classfisher
1	GO:0060968	regulation of gene silencing	8	8	2.58	0.00012
2	GO:0006396	RNA processing	146	68	47.15	0.00014
3	GO:0046822	regulation of nucleocytoplasmic transport	37	23	11.95	0.00016
4	GO:0065002	intracellular protein transmembrane transport	10	9	3.23	0.00026
5	GO:0071806	protein transmembrane transport	10	9	3.23	0.00026

CTCFL is a known paralog of CTCF, also called Brother Of Regulator of Imprinted Sites (BORIS) (Loukinov *et al*., 2002). After WGD in early teleost fish, retention of CTCFL in the genome of stickleback and medeka could be explained by sub-functionalization (Taylor *et al*., 2003). On the other hand, these duplicates evolved by performing distinct functions in mammals after their divergence from monotremes (Hore *et al*., 2008). This suggests that in salmonids, therefore, in Atlantic salmon, both copies could perform similar ancestral gene functions as salmonids belong to the later stage of teleost divergence. CTCF, a highly conserved transcriptional regulator protein (Bell *et al*., 1999), performs by binding to the transcription start sites (TSSs) of many genes in order to control their expression (Nora *et al*., 2017). Its ubiquitous functions corresponds to the finding from Wang *et al*. (2020), who reported

contribution of CTCF in hepatocyte repopulation by activating genes that regulate the cell cycle. They stated that enrichment of CTCF in the accessible chromatin region could be associated with demethylation (Wang *et al*., 2020). Moreover, our GOs (Table 5) confirms the transcriptional and translational control of CTCF in liver that could be consistent with hepatocyte repopulation .

EGR1 is a highly conserved transcriptional regulator across vertebrate evolution (Drummond *et al*., 1994) and it regulates cell proliferation via p21 mediated pathway (Li *et al*., 2020). Our results (Table 6) also supports its involvement in the development of the immune system, blood vessels, blood cells and the skeletal system.

EMX2 has been found to act on cellular proliferation by regulating the Wnt/B catenin signalling pathway (Li *et al*., 2012). Wnt/B catenin is a evolutionarily conserved pathway important for fundamental development (Pennica et al., 1998) as well as tissue homeostasis in adults (Clevers et al., 2014). But the contribution of EMX2 in tissue homeostasis is not well studied. EMX2, has been studied in murine brain where its role has been suggested to control regulation of Beta catenin during Wnt pathway (Muzio *et al*., 2005). Beta catenin in the epithelial cell junctions, undergoes stabilization instead of degradation during Wnt signalling pathway (Dickinson *et al*., 2011). Also, binding of Beta catenin with the TFs appearing in the cell nucleus, activates transcription for the downstream target genes of the pathway (Molenaar *et al*., 1996; Korinek *et al*., 1998). Although the potential mechanism of EMX2 remains unclear throughout the pathway, our GOs (Table 7) suggest its contribution in the transport of Beta catenin into the nucleus to initiate transcription. However, there could be several other nuclear proteins regulating the journey of Beta catenin towards downstream gene expression (Söderholm and Cantù, 2021). Albeit, the whole mechanism remain unsolved, the interplay among transcriptional regulation, nuclear factors and our GOs suggest further investigation with a focus on EMX2 activity throughout the process of Wnt/B catenin signalling pathway.

ETV3, a tumor suppressor in chronic lymphocytic leukaemia (Green *et al*., 2010), is found to be involved in cell cycle arrest, inflammation and protein synthesis (Carlson *et al*., 2011). Protein phosphorylation of ETV3 by ERK1/2 plays a crucial part in downstream ERK activated mechanisms. ERK1/2 (MAPKs-Mitogen-activated protein kinases), conserved in both vertebrates and invertebrates, are integral part of multiple biological processes that include the immune system (Dong *et al*., 2002), development (Aouadi *et al*., 2006), glucose homeostasis (Bost *et al*., 2005) and memory (Govindarajan *et al*., 2006). These kinases phosphorylate various proteins to initiate the regulation of these processes (Yoon and Seger, 2006). Carlson *et al*. (2011) reported that phosphorylation of ETV3 by ERK1/2 promotes activation of genes involved in cell cycle, mRNA processing and translation. On the other hand, while the activity of ERK1/2 ceases, ETV3 (newly translated) instantly represses its target genes, thus allowing temporary blow-up of post ERK activated transcriptional activities (Carlson *et al*., 2011). Supported by our GOs (Table 8), we found the role of ETV3 in the regulation of cell proliferation. However, we could not link how does it relates to protein transmembrane transport.

Seemingly, our TFs are involved in the functions mostly related to the cell proliferation. Gillard *et al*. (2020) reported that the association of adaptive evolution towards cell cycle in salmon have potentially influence on genome stability. This relation was found to be biased towards upregulated copies between up+cons paralogs (Gillard *et al*., 2020). These authors also stated higher promoter divergence with regards to more bound transcription factor binding sites (bTFBSs) in the increased copies (upregulated) (Gillard *et al*., 2020). However, our TFs do not belong to the group of TFs which were found to be associated with adaptive gain in the paper (Gillard *et al*., 2020). Therefore, It suggests that it could be easier to modify the co-activators of our TFs rather than the TF themselves towards the paralogous expression turnover. To test the hypothesis in future, we can check for the TFs by using the same datasets but in the reversed direction. Then we can select for the genes and look into the

TFs that regulate them, followed by running GO enrichment analysis on them. Otherwise, we need to perform the similar chi square test in the whole datasets for which we have similar expression in the paralogs. If we get the same TFs for all (up + up, down + down and cons + cons), it means that these four TFs are very congruent with our hypothesis that they need to be conserved in Atlantic salmon.

Conclusion

In conclusion, measuring the validity of the network including GO enrichment analysis showed that Inferelator modelled a purposive TF-gene regulatory network. However, our results recommended that answering the gap between modularity at the functional level and network level could ameliorate the robustness of the model. On the other hand, our analysis involving WGD driven TF evolvability with respect to paralogous expression variation suggested that CTCFL, EGR1, EMX2 and ETV3 remained preserved in salmon rather than being biased towards the upregulated copies of the paralogs throughout the evolutionary time.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Intracellular control of cell-cycle events. In *Molecular Biology of the Cell. 4th edition*. Garland Science.

Aouadi, M., Binetruy, B., Caron, L., Le Marchand-Brustel, Y., & Bost, F. (2006). Role of MAPKs in development and differentiation: lessons from knockout mice. *Biochimie*, *88*(9), 1091-1098.

Alexa, A., & Rahnenführer, J. (2009). Gene set enrichment analysis with topGO. *Bioconductor Improv*, *27*, 1-26.

Bell, A. C., West, A. G., & Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, *98*(3), 387-396.

Bost, F., Aouadi, M., Caron, L., & Binétruy, B. (2005). The role of MAPKs in adipocyte differentiation and obesity. *Biochimie*, *87*(1), 51-56.

Birchler, J. A., & Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, *109*(37), 14746-14753.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., ... & Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*, *24*(1), 14-24.

Beysel, S., Eyerci, N., Pinarli, F. A., Kizilgul, M., Ozcelik, O., Caliskan, M., & Cakal, E. (2019). HNF1A gene p. I27L is associated with early-onset, maturity-onset diabetes of the young-like diabetes in Turkey. *BMC endocrine disorders*, *19*(1), 1-7.

Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., ... & Looso, M. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature communications*, *11*(1), 1-11.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9: 938–950.

Carlson, S. M., Chouinard, C. R., Labadorf, A., Lam, C. J., Schmelzle, K., Fraenkel, E., & White, F. M. (2011). Large-scale discovery of ERK2 substrates identifies ERK-mediated transcriptional regulation by ETV3. *Science signaling*, *4*(196), rs11-rs11.

Carmona-Antoñanzas, G., Tocher, D. R., Taggart, J. B., & Leaver, M. J. (2013). An evolutionary perspective on Elovl5 fatty acid elongase: comparison of Northern pike and duplicated paralogs from Atlantic salmon. *BMC Evolutionary Biology*, *13*(1), 1-13.

Clevers, H., Loh, K. M., & Nusse, R. (2014). An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control. *science*, *346*(6205).

Castro, D. M., De Veaux, N. R., Miraldi, E. R., & Bonneau, R. (2019). Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS computational biology*, *15*(1), e1006591.

DRUMMOND, I. A., ROHWER-NUTTER, P. A. T. R. I. C. I. A., & SUKHATME, V. P. (1994). The zebrafish egr1 gene encodes a highly conserved, zinc-finger transcriptional regulator. *DNA and cell biology*, *13*(10), 1047-1055.

Dong, C., Davis, R. J., & Flavell, R. A. (2002). MAP kinases in the immune response. *Annual review of immunology*, *20*(1), 55-72.

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *Plos biol*, *3*(10), e314.

Dickinson, D. J., Nelson, W. J., & Weis, W. I. (2011). A polarized epithelium organized by β-and α-catenin predates cadherin and metazoan origins. *Science*, *331*(6022), 1336-1339.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15-21.

Fornes, O., Castro-Mondragon, J. A., Khan, A., Van der Lee, R., Zhang, X., Richmond, P. A., ... & Mathelier, A. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, *48*(D1), D87-D92.

F-score. (2021, June, 26). In *Wikipedia*. https://en.wikipedia.org/wiki/F-score

Govindarajan, A., Kelleher, R. J., & Tonegawa, S. (2006). A clustered plasticity model of long-term memory engrams. *Nature Reviews Neuroscience*, *7*(7), 575-583.

Green, M. R., Jardine, P., Wood, P., Wellwood, J., Lea, R. A., Marlton, P., & Griffiths, L. R. (2010). A new method to detect loss of heterozygosity using cohort heterozygosity comparisons. *BMC cancer*, *10*(1), 1-9.

Gillard, G. B. (2019). Evolution of gene expression following the whole genome duplication in salmonid fish.

Gillard, G. B., Groenvold, L., Røsæg, L., Holen, M. M., Monsen, O., Koop, B. F., ... & Hvidsten, T. R. (2020). Comparative regulomics reveals pervasive selection on gene dosage following whole genome duplication. *BioRxiv*.

He, X., & Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, *169*(2), 1157-1164.

Hore, T. A., Deakin, J. E., & Marshall Graves, J. A. (2008). The evolution of epigenetic regulators CTCF and BORIS/CTCFL in amniotes. *PLoS genetics*, *4*(8), e1000169.

Houston, R. D., & Macqueen, D. J. (2019). Atlantic salmon (Salmo salar L.) genetics in the 21st century: taking leaps forward in aquaculture and biological understanding. Animal genetics, 50(1), 3-14.

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., ... & Crollius, H. R. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, *431*(7011), 946-957.

Jones, D. M., & Vandepoele, K. (2020). Identification and evolution of gene regulatory networks: Insights from comparative studies in plants. *Current opinion in plant biology*, *54*, 42-48.

Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O. & Clevers, H. (1998) Mol. Cell. Biol. 18, 1248–1256.

Kuo, T., McQueen, A., Chen, T. C., & Wang, J. C. (2015). Regulation of glucose homeostasis by glucocorticoids. *Glucocorticoid signaling*, 99-126.

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, *20*(4), 207-220.

Loukinov, D. I., Pugacheva, E., Vatolin, S., Pack, S. D., Moon, H., Chernukhin, I., ... & Lobanenkov, V. V. (2002). BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. *Proceedings of the National Academy of Sciences*, *99*(10), 6806-6811.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics, 12(1), 323.

Li, J., Mo, M., Chen, Z., Chen, Z., Sheng, Q., Mu, H., ... & Zhou, H. M. (2012). Adenoviral delivery of the EMX2 gene suppresses growth in human gastric cancer.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., ... & Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, *533*(7602), 200-205.

Lau, H. H., Ng, N. H. J., Loo, L. S. W., Jasmen, J. B., & Teo, A. K. K. (2018). The molecular functions of hepatocyte nuclear factors–In and beyond the liver. *Journal of hepatology*, *68*(5), 1033-1048.

Li, T. T., Liu, M. R., & Pei, D. S. (2020). Friend or foe, the role of EGR-1 in cancer. *Medical Oncology*, *37*(1), 1-8.

Molenaar, M., van de Wetering, M., Oosterwegel, M., PetersonMaduro, J., Godsave, S., Korinek, V., Roose, J., Destree, O. & Clevers, H. (1996) Cell 86, 391–399.

Moon, T. W. (2001). Glucose intolerance in teleost fish: fact or fiction?. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *129*(2-3), 243-249.

McAdams, H. H., Srinivasan, B., & Arkin, A. P. (2004). The evolution of genetic regulatory systems in bacteria. *Nature Reviews Genetics*, *5*(3), 169-178.

Muzio, L., Soria, J. M., Pannese, M., Piccolo, S., & Mallamaci, A. (2005). A mutually stimulating loop involving emx2 and canonical wnt signalling specifically promotes expansion of occipital cortex and hippocampus. *Cerebral Cortex*, *15*(12), 2021-2028.

Mungpakdee, S., Seo, H. C., Angotzi, A. R., Dong, X., Akalin, A., & Chourrout, D. (2008). Differential evolution of the 13 Atlantic salmon Hox clusters. *Molecular Biology and Evolution*, *25*(7), 1333-1343.

Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1778), 20132881.

Miraldi, E. R., Pokrovskii, M., Watters, A., Castro, D. M., De Veaux, N., Hall, J. A., ... & Bonneau, R. (2019). Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. *Genome research*, *29*(3), 449-463.

Matthews correlation coefficient. (2021, July 26). In *Wikipedia.* https://en.wikipedia.org/wiki/Matthews_correlation_coefficient

Nowak, M. A., Boerlijst, M. C., Cooke, J., & Smith, J. M. (1997). Evolution of genetic redundancy. *Nature*, *388*(6638), 167-171.

Niswender, K. D., Shiota, M., Postic, C., Cherrington, A. D., & Magnuson, M. A. (1997). Effects of increased glucokinase gene copy number on glucose homeostasis and hepatic glucose metabolism. *Journal of Biological Chemistry*, *272*(36), 22570-22575.

Navarro, I., Rojas, P., Capilla, E., Albalat, A., Castillo, J., Montserrat, N., ... & Gutiérrez, J. (2002). Insights into insulin and glucagon responses in fish. *Fish Physiology and Biochemistry*, *27*(3), 205-216.

Nakatani, Y., Takeda, H., Kohara, Y., & Morishita, S. (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome research*, *17*(9), 1254-1265.

Nowick, K., & Stubbs, L. (2010). Lineage-specific transcription factors and the evolution of gene regulatory networks. *Briefings in functional genomics*, *9*(1), 65-78.

Nelson, J. S., Grande, T. C., & Wilson, M. V. (2016). Fishes of the World. John Wiley & Sons.

Nora, E. P., Goloborodko, A., Valton, A. L., Gibcus, J. H., Uebersohn, A., Abdennur, N., ... & Bruneau, B. G. (2017). Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, *169*(5), 930-944.

Ohno S. (1970). Evolution by gene duplication. New York: Springer-Verlag.

Pennica, D., Swanson, T. A., Welsh, J. W., Roy, M. A., Lawrence, D. A., Lee, J., ... & Levine, A. J. (1998). WISP genes are members of the connective tissue growth factor family that are up-regulated in wnt-1-transformed cells and aberrantly expressed in human colon tumors. *Proceedings of the National Academy of Sciences*, *95*(25), 14717-14722.

Prince, V.E., and Pickett, F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. 3: 827–837.

Precision and recall. (2021, August, 2). In *Wikipedia*. https://en.wikipedia.org/wiki/Precision_and_recall

Ray, S., & Pollard, J. W. (2012). KLF15 negatively regulates estrogen-induced epithelial cell proliferation by inhibition of DNA replication licensing. *Proceedings of the National Academy of Sciences*, *109*(21), E1334-E1343.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rohlfs, R. V., Harrigan, P., & Nielsen, R. (2014). Modeling gene expression evolution with an extended Ornstein–Uhlenbeck process accounting for within-species variation. *Molecular biology and evolution*, *31*(1), 201-211.

Söderholm, S., & Cantù, C. (2021). The WNT/β-catenin dependent transcription: A tissue-specific business. *WIREs Mechanisms of Disease*, *13*(3), e1511.

Tranulis, M. A., Dregni, O., Christophersen, B., Krogdahl, Å., & Borrebaek, B. (1996). A glucokinase-like enzyme in the liver of Atlantic salmon (Salmo salar). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *114*(1), 35-39.

Taylor, J. S., Braasch, I., Frickey, T., Meyer, A., & Van de Peer, Y. (2003). Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome research*, *13*(3), 382-390.

Thompson, D., Regev, A., & Roy, S. (2015). Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annual review of cell and developmental biology*, *31*, 399-428.

Takeuchi, Y., Yahagi, N., Aita, Y., Murayama, Y., Sawada, Y., Piao, X., ... & Shimano, H. (2016). KLF15 enables rapid switching between lipogenesis and gluconeogenesis during fasting. *Cell reports*, *16*(9), 2373-2386.

TPM. (2016, August, 4). In *Wikipedia*. http://www.arrayserver.com/wiki/index.php?title=TPM

Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, *10*(10), 725-732.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440-442.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57-63.

Wang, X., He, M., Li, J., Wang, H., & Huang, J. (2018). KLF15 suppresses cell growth and predicts prognosis in lung adenocarcinoma. *Biomedicine & Pharmacotherapy*, *106*, 672-677.

Wang, X., Hassan, W., Zhao, J., Bakht, S., Nie, Y., Wang, Y., ... & Huang, Z. (2019). The impact of hepatocyte nuclear factor-1α on liver malignancies and cell stemness with metabolic consequences. *Stem cell research & therapy*, *10*(1), 1-8.

Wang, A. W., Wang, Y. J., Zahm, A. M., Morgan, A. R., Wangensteen, K. J., & Kaestner, K. H. (2020). The dynamic chromatin architecture of the regenerating liver. *Cellular and molecular gastroenterology and hepatology*, *9*(1), 121-143.

Yoon, S., & Seger, R. (2006). The extracellular signal-regulated kinase: multiple substrates regulate diverse cellular functions. *Growth factors*, *24*(1), 21-44.

Zeng, X., Lin, Y., Yin, C., Zhang, X., Ning, B. F., Zhang, Q., ... & Xie, W. F. (2011). Recombinant adenovirus carrying the hepatocyte nuclear factor-1alpha gene inhibits hepatocellular carcinoma xenograft growth in mice. *Hepatology*, *54*(6), 2036-2047.

Zhao, N., Ding, X., Lian, T., Wang, M., Tong, Y., Liang, D., ... & Xu, C. (2020). The Effects of Gene Duplication Modes on the Evolution of Regulatory Divergence in Wild and Cultivated Soybean. *Frontiers in genetics*, *11*.

Supplementary file

This file contains TF-gene regulatory network that is obtained from Inferelator

https://drive.google.com/file/d/1bBgem1XW48s_WkYQ1iJrM7hOhVOUn5D9/view?usp=sharing

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway