Chalmers University of Technology

BBT045

Applied bioinformatics

Reproducing trancriptomics analysis from study on the effects of domestication in atlantic salmon

March 9, 2021

Authors:
Linn Hanna
hannlinn@chalmers.se

Stauber Näslund Louise,
lounas@student.chalmers.se

de Juana Fabra Amalia Muskilda,
muskilda@student.chalmers.se

Introduction

The study of how domestication of animals can change the genetics and physiology of the species always have interested humans, and it is relevant in how humans change our environment. Domestication syndrome, as the process of change is called, is often due to an intended or unintended selection of individuals due to the change of environment when a species is domesticated. Such changes can be, e.g., a change of environment or a change of diet. Focusing on the dietary change of a species may lead to an evolution in the metabolism.

The Atlantic salmon (*Salmo salmar*) is interesting to investigate for the domestic syndrome as the species has been used in systematic breeding programs since 1971, which have aimed to improve traits such as delayed sex maturation; higher feed conversion rate; faster growth; and many other traits that boost animal production. The control of diet for the domesticated salmon is hypothesised to have impacted how the fish metabolises compounds of the food. The wild salmon have a diet that often contains substantial amounts of long-chain polyunsaturated fatty acids (LC-PUFAs), but in the wild finding, food can be scarce. The LC-PUFAs are essential for fish because they are vital components of cell membranes, they regulate cell membrane fluidity, and they are essential components of neural tissues. Until two decades ago, the domesticated salmon have had unlimited supplies to a diet consisting of fish oil (FO), which have high levels of LC-PUFAs. Now the domesticated salmon are mostly fed vegetable oil (VO), which have low levels of LC-PUFAs. The change of diet is hypothesised to have changed the domesticated salmon's lipid metabolism [1].

The article *Comparative transcriptomics reveals domestication-associated features of Atlantic salmon lipid metabolism* [1], goes into depth in the question about change of the metabolism in Atlantic salmon due to domestication. The article includes finding differentially expressed genes for both wild and domesticated salmon, the metabolic organs (pyloric caeca and the liver), and each diet (FO and VO diets).

Results in the article

The authors found, among other things, that 230 genes were significantly differentially expressed between domesticated and wild salmon in the pyloric caeca for the FO diet and that a principal component analysis (PCA) of the

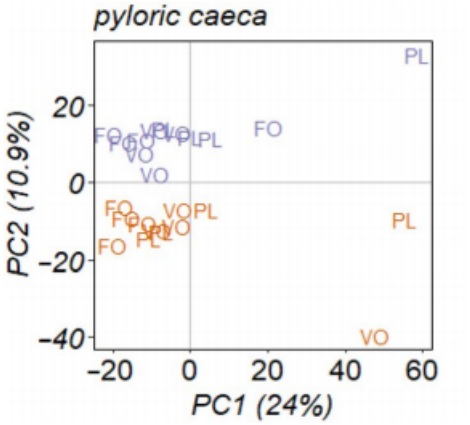

Figure 1: Score plot of PCA on $\log_2$ count per million of the top 1000 most variant genes across all samples.Two salmon strains (domesticated and wild) were fed diets rich in either fish oil (FO), vegetable oil (VO) or phospholipid (PL) from initial feeding. Samples were taken after 94 days of feeding.

top 1000 most variable genes showed a clear separation between domesticated and wild salmon, see figure 1.

Our analysis

In our project, we wanted to investigate whether these results could be reproduced using the original RNAseq-data from the article. However, we did not use the whole data set for our analysis. Only six domesticated and wild samples, respectively, from the pyloric caeca for the FO diet were used.

Method

The code used in this project can be accessed at GitHub repository [`https://github.com/louisenaslund/BBT045_AppliedBioinformatics`].

RNA-seq data

The original RNA-seq data were obtained from ArrayExpress (project accession no. E-MTAB-8306, available at `https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8306/`) and analysed according to Jin et al. [2]. The

read sequences were quality trimmed using `cutadapt` (version 3.2), were Illumina TRUEseq adaptors (obtained from Illumina adapter sequences document) and low-quality bases (Phred score <20) on read ends were removed, and reads were filtered for length (minimum length 40 bases). The subsequent reads were then aligned to the salmon reference genome (ICSASG_v2) using `star` (version 2.7.8a), and the NCBI salmon genome annotation (available for download at `https://www.ncbi.nlm.nih.gov/assembly/GCF_000233375.1/`). Raw gene counts were generated using `htseq-count` (version 0.13.5) and the NCBI salmon genome annotation.

Data analysis

The data analysis was performed using R version 3.6.3.

Differentially expressed genes

For the differentially expressed genes (DEG) analysis, the raw counts were normalised using counts per million and filtered so that each gene had a CPM count of at least 1 in 25% of all the samples. Differential expression was then tested on the samples, using the `exactTest` from R package `edgeR`. P-values were then adjusted for multiple testing using the false discovery rate (FDR). Genes with a q value (FDR adjusted p-values) <0.05 and an absolute $\log_2$ fold change > 1 between domesticated and wild salmon were considered to be differentially expressed.

Principal component analysis

The principal component analysis (PCA) we did was on $\log_2$ CPM of the top 1000 most variant genes of our dataset, similar to what they did in the article. We used the R function `prcomp` that performs a $\log_2$ CPM and a principal components analysis.

Results

The differential expression analysis resulted in 322 significantly DEGs between domesticated and wild salmon for samples from the pyloric caeca under the FO diet. Furthermore, the top three most significant genes were LOC106566856, LOC106603040, and LOC106587663.

Principal component analysis (PCA)

The results from our PCA can be seen in the figure 2.

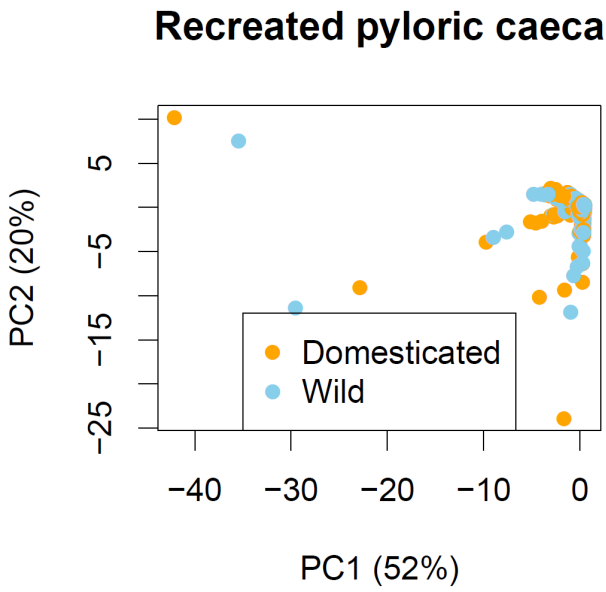

Figure 2: Score plot of the $\log_2$ counts per million of the top 1000 most variant genes.

Discussion

In the original article, the authors found 230 DEGs, where the top three most differentially expressed genes were LOC106560721, LOC106584737, and LOC106562896. The resulting genes are not in agreement with the results from this project, where we found 322 DEGs, with the top three ones being: LOC106566856, LOC106603040, and LOC106587663. The difference in the number of DEGs could be since only six samples from domesticated and wild were used in this project, compared to the 24 samples in total for the pyloric caeca and FO diet analysed in the original article. The low number of samples could also have influenced the top DEGs. The effect of the small number of samples on a DEG analysis workflow in discussed further by Baccarella et al. [3]. Even though our results did not mirror the results of the article we tried

to reproduce, we can still draw the same conclusion as in the article: there is a different in domesticated and wild salmon according to what expressed genes the strains show.

Our PCA results are very different from the one in the article and should not be compared as they analyse quite different datasets. Our dataset is not as large as the one in the article, with us only analysing one diet and just a few samples. We also did not use only samples taken 94 days from feeding, but a mixture of ages. The method is the same, but the data so different so that it is hard to compare the results. In the article, they find a big split in their PCA between wild and domesticated salmon, our results did not show the same split. From our PCA results, we cannot draw the conclusion that there is a big difference between domesticated and wild salmon.

Summary

We succeeded in performing a transcriptomic analysis of six samples of each strain in wild and domesticated Atlantic salmon in the pyloric caeca for fish oil diet. We saw that there was a different in the expressed genes, and domestication has an impact. In addition to this, a principal component analysis (PCA) of the most variable genes was attempted, but the results did not tell us much, probably because the analysed data was too small. Different data gives different results. More time and computational power may lead to more interesting results.

Outlook

We only analysed six samples of domesticated and wild salmon from the pyloric caeca for the FO diet. To fully reproduce the original study, all samples should be analysed, which is something we would do if we were to redo this project. Furthermore, during this project, we learned the importance of reproducible research and how important it is to provide adequate material from your study so that your results can be reproduced and verified. We also learned that only a detailed methods section is not always enough and that you should also provide your code for the sake of reproducibility.

References

[1] Y. Jin, R. E. Olsen, T. N. Harvey, M.-A. Østensen, K. Li, N. Santi, O. Vadstein, A. M. Bones, J. O. Vik, S. R. Sandve, and Y. Olsen, "Comparative transcriptomics reveals domestication-associated features of atlantic salmon lipid metabolism," *Molecular Ecology*, vol. 29, no. 10, pp. 1860–1872, 2020. DOI: https://doi.org/10.1111/mec.15446. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15446. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15446.

[2] Y. Jin, R. E. Olsen, M.-A. Østensen, G. B. Gillard, S. A. Korsvoll, N. Santi, A. B. Gjuvsland, J. O. Vik, J. S. Torgersen, S. R. Sandve, *et al.*, "Transcriptional development of phospholipid and lipoprotein metabolism in different intestinal regions of atlantic salmon (salmo salar) fry," *BMC genomics*, vol. 19, no. 1, p. 253, 2018. DOI: https://doi.org/10.1186/s12864-018-4651-8.

[3] A. Baccarella, C. R. Williams, J. Z. Parrish, and C. C. Kim, "Empirical assessment of the impact of sample number and read depth on rna-seq analysis workflow performance," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–12, 2018.