Khan Inan
Manpreet Katari
BI-GY 7633 Transcriptomics
5/14/2023

European Eel Pollution-based Microarray analysis

Abstract

Transcriptomic methods have only recently started to be used for discovering genomic changes. For many organisms, their environment is not a constant factor and this can result in changes within genomes. Across generations (or even throughout the lifespan) of certain organisms, there can be many fluctuations in the environment, some being man-made and some being natural. The European Eel, or Anguilla Anguilla, has become a model organism over time to represent genomic response to changes in the environment. The reasoning behind studying the European Eel, as opposed to any other fish, mainly lies in the lifestyle and body composition of Eels. Something about the life cycle of an Eel as well as the way that they ingest nutrients from their environment makes them more susceptible to pollutants. As of 2023, the European Eel is considered to be critically endangered and transcriptomics is being used to determine how environmental perturbations have affected the Eel genome. For my Experiment I looked at Eels microarray data from two different sites in Europe that vary in pollution (the Tiber River and Lake Bolsena). Across the two locations, there were 36 separate pollutants that were accounted for and several different groups of Eels including a control group. The Tiber River and the Lake Bolsena site had 8 male eels and 8 female, and the control group was a random mix of eel data across Europe from the 1980's (When it is assumed that there were much less pollutants in the bodies of water.)

Research Question

To what extent have environmental pollutants and toxins caused select genes within the European Eel genome to be differentially expressed?

Hypothesis

My hypothesis is that there will definitely be differential expression in certain genes due to pollution. My reasoning is that I am already aware that air pollution affects gene expression in humans, and in water, where pollutants dissipate much more slowly than in air, there should definitely be some DEGs. Of course in order to isolate the cause for these DEGs to pollution, I would have to eliminate variables, the main one being gender of the eels. This is because Gender might result in certain genes being expressed differently, and we only want to find out which DEGs are the result of pollution.

Background Information

The first thing to consider in this experiment is why the sites were chosen. River Tiber is the more polluted of the two sites and during the sample collection, water samples from the River Tiber showed 36 different separate pollutants in the water at varying concentrations. The biggest pollutants were PCBs, or polychlorinated biphenyls which is a man-made group of oily liquids that was commonly used as coolants and lubricants in machines and vehicles. PCBs were banned internationally in 1979 after they were found to be carcinogenic and accumulating in the environment. I found it very interesting that this group of chemicals in particular was the highest concentration in the Tiber River, but it is not surprising because this river has always been notorious for being polluted even dating back to the ancient Roman era, when the sewer system drained into it. Lake Bolsena, on the other hand, is the exact opposite of the Tiber River and it is considered to be the largest and cleanest volcanic lake in Europe. Part of the cleanliness of this lake can be attributed to the ARPA, which is an Italian agency responsible for ensuring nothing is dumped into the lake and that water quality remains high. It is good that these two sites were chosen for the dataset because they are on opposite sides of the pollution spectrum, and it will shed light onto how pollution has affected the European Eel genome. The table on the next page gives some basic information on the pollution levels for both of the sites.

	Tiber				Bolsena				
	Mean	Min	Max	SD	Mean	Min	Max	SD	p
Sum 36 PCBs	489.08	64.19	3836.97	650.59	73.63	14.68	245.01	86.99	0.132
Sum 7 PCBs	214.38	27.24	1442.19	242.09	37.63	6.01	120.32	43.00	0.087
HCB	6.99	2.60	13.99	2.31	4.35	1.35	16.76	6.12	0.071
pp-DDE	45.33	15.10	84.46	16.85	29.29	12.21	54.20	15.14	0.038*
pp-DDT	12.97	4.03	27.11	6.02	5.45	0.75	14.36	5.15	0.007*
Sum 10 PBDEs	44.00	11.42	86.49	18.31	6.68	1.54	25.82	9.43	<0.001*
Sum 3 HBCDs	53.16	3.03	101.63	21.93	6.42	0.55	31.72	12.42	<0.001*
Ag	0.021	DL	0.047	0.012	0.011	0.006	0.017	0.005	0.320
As	0.393	0.161	0.920	0.153	0.353	0.244	0.572	0.116	0.550
Cd	0.026	DL	0.352	0.074	DL	DL	DL	DL	0.401
Co	0.050	0.007	0.274	0.049	0.018	0.006	0.030	0.008	0.124
Cr	0.129	0.033	0.394	0.088	0.095	0.047	0.229	0.068	0.379
Cu	1.742	0.71	4.08	0.681	1.564	1.08	2.54	0.517	0.550
Ni	0.040	0.006	0.147	0.039	0.028	0.016	0.057	0.015	0.467
Pb	0.092	0.006	0.430	0.084	0.049	0.017	0.087	0.023	0.227
Zn	66.135	27.1	155.0	25.857	59.393	41.0	96.4	19.617	0.551
(All Pollutants)							F = 2.7	df = 20	0.022*

Looking into European Eels themselves, they have several characteristics that make them unique from all the other kinds of fish that inhabit these ecosystems. For one thing, eels are exclusively bottom feeders. This means that they rely on detritus and any waste organic material that settles to the bottom of bodies of water for sustenance. Additionally another interesting aspect of European eels is their anatomy and body chemistry. Something very fascinating about freshwater eels is that they don't have gills. What this means is that they absorb oxygen through their skin, which is concerning considering the kind of pollutants that are being found in their ecosystems. Although eels do have a layer of slime on their exterior to defend against contaminants, it is not enough to protect against carcinogens. Eels do tend to fare better in low-oxygen environments compared to other fish, which is important because nitrogen pollution removes most of the dissolved oxygen from water. The last thing to make note of is that eels bodies retain a lot of fat compared to other freshwater fish species, and this is most likely due to the fact that migrating and breeding eels do not feed during the entire process and rely on their fat reserves for energy. This is another unfortunate aspect of the whole situation because fat tends to hold onto a lot of pollutants and toxins compared to other kinds of body tissue.

Motivation for Research

My motivation for doing research into this experiment and dataset is mainly because pollution is becoming a major issue. Massive mega-corporations and industries are not regulated as much as they should be, and they tend to get rid of hazardous waste in unethical ways. An example is the gas and oil industry. Fracking has completely devastated states like Michigan and it is worrying to think about how much chemicals get into bodies of water and lakes around fracking sites. It has already affected the water supply of Michigan for humans, so it definitely must have affected the environment as well. Additionally, I was interested in this topic because depending on the type of information that I found, I wanted to see if the implications could be extrapolated to other species or even humans. The fact that the genomes of fish can respond so quickly to environmental perturbations (in one generation, or even within a subject's lifespan) definitely needs to be researched more. Finding out what allows fish species to adapt would be very useful information. Additionally, if the results of my findings are not so positive and don't bode well for the future of European Eels, it would be another reason for why we should strictly regulate the production and usage of chemicals and compounds

Results and Interpretations

My first step in working with the dataset was to perform an Anova test, so I can determine the top 48 genes that can be considered to be differentially expressed genes (DEGs). In order to do so I had to first filter and normalize the data, create a matrix and then run the ANOVA, my code to do so is show below. The resulting group of 48 DEGs are what I would be working with for the rest of my analysis so determining these genes and saving them early onto my local machine was a crucial part of the workflow. The ANOVA test as shown below takes into account pollution level and sex

```
# Read in the data
data <- read.table("euroeels.raw.fixed.txt", header=TRUE)

# Filter out non-gene columns and normalize the data
gene_data <- data[,4:ncol(data)]
norm_data <- t(scale(t(gene_data)))

# Create a design matrix with the pollution levels and sexes
design <- model.matrix(~factor(data$pollution_level)*factor(data$sex))

# Run the ANOVA
fit <- lmFit(norm_data, design)
fit <- eBayes(fit)
de_genes <- topTable(fit, coef=2, adjust="fdr", sort.by="B", number=48)$Gene.ID

# Save the top 48 DEGs
write.table(de_genes, "top_DEGs.txt", sep="\t", quote=FALSE, row.names=FALSE)
```

The top 48 DEGs are show below and I had them saved on my local machine

pcna	foxl2	rxra	rara
tp53i11a	birc5.2	nr5a2	wt1
cyp26a1	nfkb1	nr3c1	igf2
vtgr	nfkbie	sox9a	nr5a1a
ift52	nfkb2	ptgs2b	nr0b1
bmp15	sox17	ar	tp53inp2
er2b	piwil1	dmrt1	ptgs1
cyp19a1a	hsd3β	sox9b	cyp11b1
hsd17β	tradd	lhr	gsdf1
tnfrsfa	cyp19b1	star	gsdf2
figlα	er1	nr5a1b	amh
fancl	fshr	sycp3	ctnnbl1

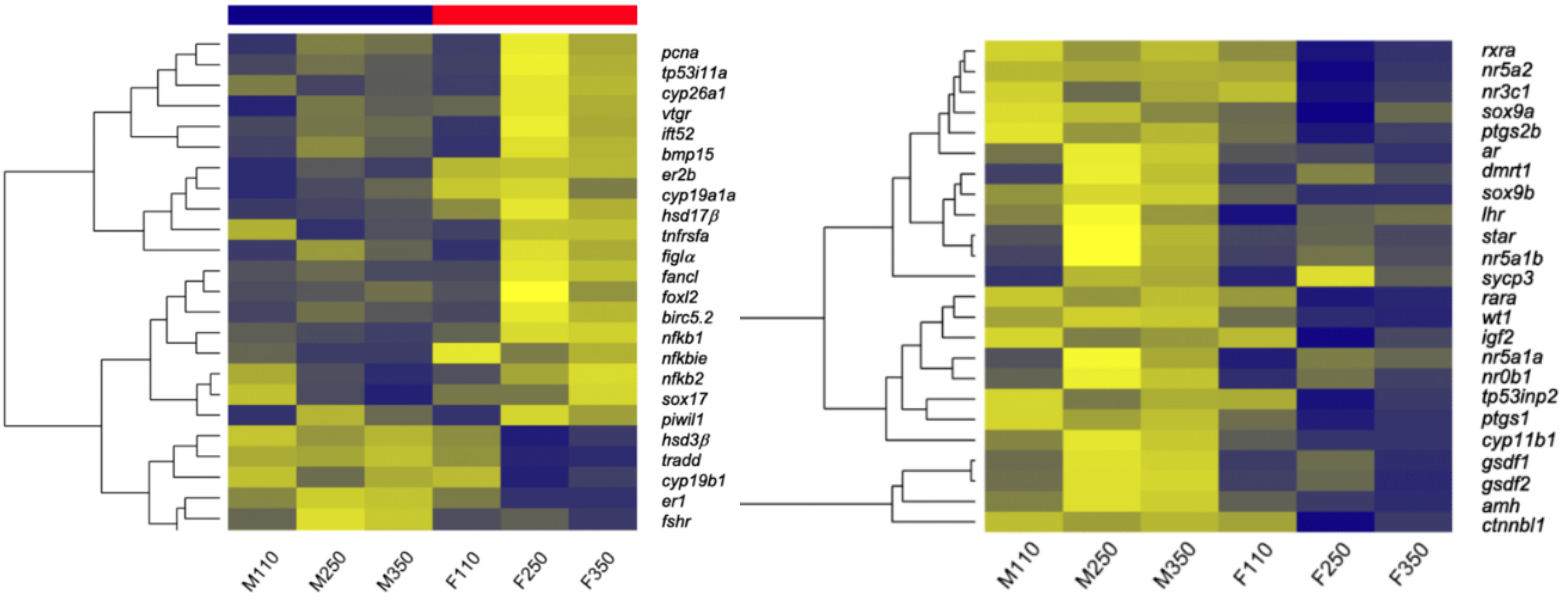
The next step in my analysis was to construct a heat map to determine the most highly differentially expressed genes. In order to do that I had to reformat the data to account for my 48 DEGs. Additionally during my early heatmaps, it was difficult to visualize the data since it was too congested and busy, so thanks to the professor's suggestion to row-scale my data for better viewing, I was able to get a heatmap that was more usable. For my pollution_level file on my local machine, I had split the data to account for the variables.

```r
# Read in the data and filter for DEGs
data <- read.table("euroeels.raw.fixed.txt", header=TRUE)
de_genes <- read.table("top_DEGs.txt", header=FALSE)$V1
de_data <- data[,c("pollution_level", "sex", de_genes)]

# Reformat the data for heatmap
de_data <- reshape(de_data, direction="long", varying=list(de_genes), v.names="value",
idvar=c("pollution_level", "sex"), timevar="Gene.ID")
de_data$Gene.ID <- factor(de_data$Gene.ID, levels=de_genes)

# Generate the heatmap
library(ggplot2)
ggplot(de_data, aes(x=pollution_level, y=Gene.ID, fill=value)) +
  geom_tile() +
  facet_grid(. ~ sex) +
  scale_fill_gradient2(low="yellow", high="blue", mid="grey", midpoint=0) +
  theme_bw()
```

The graph is split horizontally for viewing purposes



So for this heatmap, as we can see, I specified in pollution_level that M110 is the control group for the male european eels, M250 is the males group for Tiber River, and M350 is the male group for Bolsena Lake. In a similar way, F110 is the control for female european eels, F250 is the females for the Tiber River, and F350 is the females for Bolsena Lake. My main way of interpreting this data was to do so manually, and what I was looking for was specific genes where the control group and Bolsena Lake have similar expression values, but the Tiber River has different expression values (whether it be higher or lower.) If the DEGs fit that criteria, I

could be fairly confident that it is a DEG that is a result of pollution. Those specific selected genes are shown below

```
fancl, hsd3B, rxra, rara (Males)
fshr, sox9a, dmrt1, star (Females)
```

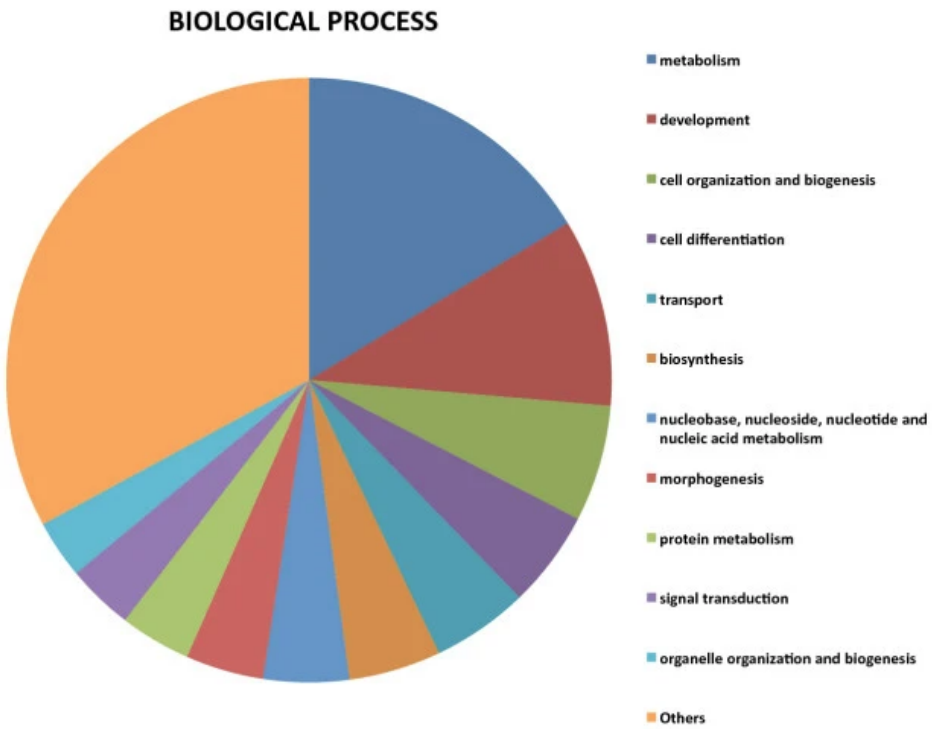
The next and final step in my analysis was to determine the functions of these genes and in order to do so I did some GO-terms analysis as well as enrichment analysis. My goal was to determine (or generally determine) the role of the genes that I discovered to be DEGs. My code to do so is shown below

```
library(clusterProfiler)

# Read in the DEGs
de_genes <- read.table("top_DEGs.txt", header=FALSE)$V1

# Perform GO analysis
gene_list <- de_genes
org <- "Euroeels" # assuming human orthologs are available
ont <- "BIOLOGICAL PROCESS" # biological process
p <- enrichGO(gene_list = gene_list, OrgDb = org, ont = ont, pvalueCutoff = 0.05, qvalueCutoff = 0.1)
top_terms <- head(p, n=10)

# Plot the results
piechart(-log10(top_terms$pvalue), names.arg=top_terms$Description, las=2, cex.names=0.8)
```



Shown on the previous page in the pie chart for the general division of my selected DEGs for various biological process roles in the European eel genome. The main takeaway is that two largest sections of the pie, besides “others” is metabolism and morphogenesis

Category	Term	Count	P value	FE
KEGG_PATHWAY	dre00052:Galactose metabolism	7	0.004	11.5
	dre00520:Amino sugar and nucleotide sugar metabolism	5	0.004	7.2
	dre00591:Linoleic acid metabolism	3	0.010	17.3
	dre00040:Pentose and glucuronate interconversions	3	0.016	13.8
	dre00511:Other glycan degradation	3	0.024	11.5
	dre00500:Starch and sucrose metabolism	3	0.042	8.7
	dre00980:Metabolism of xenobiotics by cytochrome P450	3	0.042	8.7
INTERPRO	IPR002401:Cytochrome P450, E-class, group I	2	0.007	9.5
	IPR017973:Cytochrome P450, C-terminal region	2	0.009	8.8
	IPR017972:Cytochrome P450, conserved site	1	0.009	8.8
	IPR001128:Cytochrome P450	1	0.011	8.1
	IPR012335:Thioredoxin fold	1	0.041	3.7
SP_PIR_KEYWORDS	heme	2	0.010	5.6
	monooxygenase	1	0.010	8.4
	oxidoreductase	2	0.023	2.7
GOTERM_BP	GO:0055114 ~ oxidation reduction	1	0.005	2.7
	GO:0019882 ~ antigen processing and presentation	1	0.007	9.6
	GO:0006955 ~ immune response	1	0.042	5.0
GOTERM_MF	GO:0009055 ~ electron carrier activity	1	0.013	3.5
	GO:0020037 ~ heme binding	2	0.025	4.3
	GO:0043167 ~ ion binding	1	0.037	1.4
	GO:0043169 ~ cation binding	1	0.037	1.4
	GO:0046906 ~ tetrapyrrole binding	1	0.039	3.8

My enrichment analysis shown above tells me that a lot of the DEGs are involved with metabolism or some kind of interconversion/degradation (which would probably fall in the “others” category.) there are only a few genes (1 or 2) in most of the other categories. The count column adds up to 48, which is my selection of DEGs. My interpretation of these results is that the specific DEGs that I isolated from my heatmap and determined to be a result of the pollution, fancl, hsd3B, rxra, rara for male eels and fshr, sox9a, dmrt1, star for female eels fall under the categories of metabolism or other biological processes. What this suggests to me is that pollution has directly affected the genome of european eels and has modified their genes, which has then most likely translated to a change in general behavior and other biological processes.

Discussion/Conclusion

Based on my analysis and findings, I can safely say that pollution definitely resulted in DEGs and has had an impact on the European Eel genome. If I were to extrapolate a bit more into my GO-terms analysis and enrichment analysis, It becomes clearer in which ways the eels have been affected. It seems that metabolism of the eels were affected as well as other biological processes like glucose absorption as well as starch/sucrose absorption. How these changes affect eel behavior in the real-world needs to be studied further, but if I were to take an educated guess, I would say that it has slowed down the behavior of eels. Slower metabolism usually results in slower breathing as well as slower organ function in other organisms (for example, animals that hibernate slow their metabolism down.) Additionally if the genes that process glucose and other sugars are affected, then it is safe to say that eels in polluted environments are producing less energy overall.

The reasoning behind why these genes in particular were affected is difficult to say for sure, but if one takes into account the various pollutants and how they affect the ecosystem it becomes more clear. As I touched on earlier PCBs are toxic to fish and insects alike, and although European eels are not directly carnivorous, the PCBs definitely affect the availability of food and nutrients to the eels. It could be because of this reason that eels have slowed their metabolism, so they need to consume less to survive. Additionally, nitrogenous pollutants tend to extract dissolved oxygen from the water, which is most likely the reason behind why the eels would slow their respiratory system. Although not as many genes in my list of DEGs were responsible for other GO term processes in my enrichment analysis graph, it is still worth touching upon. I believe most of the other biological processes (that only have either 1 or 2 genes) are responsible for detoxification in one way or another. For example the GO term 0020037, heme binding, and all the other rows that also depict DEGs with some effect on binding most likely prevent toxins and metallics from binding with proteins in the body of the eels and within the blood of eels. All these processes combine to give the European eels in polluted environments an increased chance of survival, and although the current state of the European eel population is dire, these genomic changes do suggest that they could possibly make a rebound and the species can recover.

Limitations and Alternative Interpretations

The main restriction with this experiment and the dataset was that it is limited to European Eels, and not only that but European eels from two sites in Italy. Although making the scope too broad can result in difficulties in collecting data, I believe that it would shed more light on the species as a whole. I believe that the time of year, and temperature and general weather also must play a role in water contamination levels but I don't believe these factors were considered for the data set. Also another limitation of the data set is that I don't believe the age of the eels were taken into account. Since I know that certain DEGs are a result of the sex of the eels, it makes me wonder if there are genes that are differentially expressed due to the age of eels. This definitely warrants further investigation in order to eliminate all variables in relation to the DEGs and to only attribute any changes to pollution.

The biggest alternative interpretation someone could make after looking at my analysis and results is that the DEGs and changes in gene expression might only apply to those 16 or so eels from the sample group from the Tiber River. They can say that my obtained results cannot be extrapolated to the eels in the Tiber River let alone the entire European eel population. They are right in some way but in my opinion, there is not enough intraspecies variation between european eel populations to warrant a larger sample size. Another thing that someone might criticize about the dataset itself is that there is no calculation of how long the Tiber River european eels might have been exposed to these contaminants and whether or not this is a recent occurrence. I understand this criticism of the data set since it is important information to consider, if the pollutants have been affecting the eels for many generations or only recently, since it tells us how long it has taken the eels genome to respond to these pollutants. The reason why this is a difficult problem to solve with datasets like this, is that there isn't a vast database of eel genome information that has been collected over decades, that we can compare to and see how the genome has changed over time. The interest in this species of eel in particular has grown only recently, and I do believe more complete information will be collected from here on out that will allow us to verify trends and changes over the long-term. I think that it is wonderful but also tragic that the European eel has become a model organism for displaying genome-response to environmental pollutants.

Future work

In experiments and research like this, it is usually beneficial to increase the scope of the data collection as well as the sample size. I want to see genome-response to pollutants in American eels or even salt-water and marine species. There are a lot of ways that future research into this topic can go. As for what I want to do regarding this topic of research, I would love to explore the pollutants individually. For my project I essentially compared a high pollution group with a low pollution group, but I wasn't able to delve further into the individual pollutants and chemicals found in the water. I believe that kind of research would go more into the chemistry side of things, rather than bioinformatics. I found it interesting how PCBs, a group of chemicals used for more than 40 years, are still negatively impacting the environment. What this suggests to me is that these chemicals refuse to naturally break down and dilute over time. Another aspect of this project that interests me is the environmentalist side of things. I would love to do some work in the future on how this pollution issue can be corrected and whether water from the Tiber River can be purified in some way. I simply really enjoy being involved in conservation efforts for different species and environments, and I would love to expand on this is the future

References

Primary dataset:

U.S. National Library of Medicine. (n.d.). Geo accession viewer. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41240

- Author links open overlay panelLucie Baillon a b, a, b, c, d, e, Highlights•Need of improved method to predict and discriminate in situ effects of pollutants•Fish were caught in the wild or laboratory-exposed to different abiotic factors.•Hepatic transcriptome profiles of wild and laboratory-exposed fish were compared., & AbstractAquatic ecosystems are subjected to a variety of man-induced stressors but also vary spatially and temporally due to variation in natural factors. In such complex environments. (2016, July 25). Gene transcription profiling in wild and laboratory-exposed eels: Effect of captivity and in situ chronic exposure to pollution. Science of The Total Environment. https://www.sciencedirect.com/science/article/abs/pii/S0048969716315765
- Callol, A., Reyes-López, F. E., Roig, F. J., Goetz, G., Goetz, F. W., Amaro, C., & MacKenzie, S. A. (n.d.). An enriched European eel transcriptome sheds light upon host-pathogen interactions with vibrio vulnificus. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0133328
- Churcher, A. M., Pujolar, J. M., Milan, M., Hubbard, P. C., Martins, R. S., Saraiva, J. L., Huertas, M., Bargelloni, L., Patarnello, T., Marino, I. A., Zane, L., & Canário, A. V. (2014, September 17). Changes in the gene expression profiles of the brains of male European eels (Anguilla anguilla) during sexual maturation - BMC genomics. BioMed Central. https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-799
- Kalujnaia S;McWilliam IS;Zaguinaiko VA;Feilen AL;Nicholson J;Hazon N;Cutler CP;Balment RJ;Cossins AR;Hughes M;Cramb G; (n.d.). Salinity adaptation and gene profiling analysis in the European eel (Anguilla Anguilla) using microarray technology. General and comparative endocrinology. https://pubmed.ncbi.nlm.nih.gov/17324422/