APPLIED BIOINFORMATICS

PROJECT REPORT

Mario L. Martin

Student ID: 130020616

BS32010: Applied Bionformatics

March 2014

During the present project, analyses of two different gene expression experiments data have been performed. The first experiment was a microarray expression profiling, and the second one was another expression profiling performed using RNA-sequencing. The data obtained from these experiments consisted in gene expression levels from 4 replicates of untreated and untreated tissues.

The different analyses consisted on, given the files for each experiment, identify, using RStudio-based programming, the most differentially expressed genes. Once these had been indentified, a search for homologous genes in other species was done and a phylogenetic analysis, including the building of a phylogenetic tree, was performed, in order to see the relationship between all the homologous genes and identify which homologues of which species were more closely related to our experimental specie (*Gallus gallus*). Finally, one of the more closely related homologues for each most differentially expressed gene was taken and then a BLAT search against the *Gallus gallus* genome and some closely related species genomes was carried out using the ENsembl online tools (BLAT) to find the relevant area of the chicken genome and the orthologous regions in the other genomes. Finally, a comparison was done between these two genome regions in order to see the differences in the chromosome localization, presence of more than one copy of the gene per genome and other relevant data.

Experimental Data Background

One of the most exciting discoveries in the last decades is the detection, at molecular level, of the different components of what have been called “molecular clocks”. Molecular clocks are biochemical mechanisms presents in all known cells (prokaryotes, eukaryotes and archaea) that regulate different cellular events by defining a constant rhythm in which a given process is cyclically performed. The duration of each cycle depends on the process and the molecular clock which regulates that process.

The circadian clock, for example, is a mechanism that oscillates with a period of 24 hours, and it is highly important in all organisms, from bacteria to higher mammals, as it coordinates the day-night cycles. In some photosynthetic bacteria, this is very useful in order to coordinate the expression of photosynthesis-related genes during the day and nitrogen-fixing genes during the night, in order to save energy (as photosynthesis can be done only during the day) and also avoid unwished cross-reaction between both processes (as the oxygen generated during the photosynthesis easily deactivates the nitrogenase complex used for nitrogen fixation). In higher mammals, like humans, the circadian clocks are highly important to regulate the sleeping, brain processes, metabolic pathways, body temperature and other biological activities; and its deregulation is related with a lot of different illness, such as obesity or sleeping disorders.

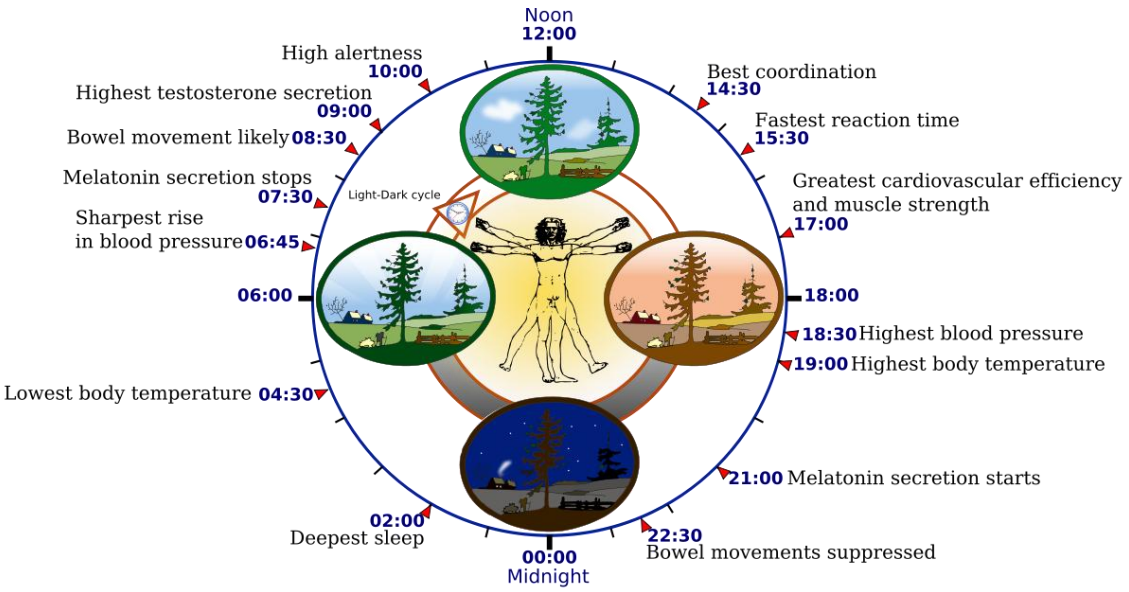


Figure 1: Schematic diagram showing some of the relevant processes regulated by the circadian cycle in humans. Source: http://en.wikipedia.org/wiki/Circadian_rhythm

During the formation of the vertebrate body, at the same time the mesoderm, ectoderm and endorderm are being formed; the paraxial mesoderm (which is how the mesoderm at the side of the neural tube is called) starts to separate into bilateral blocks called **somites**, which will give rise to skeletal muscle, dermis, tendons, cartilage and endothelial cells.

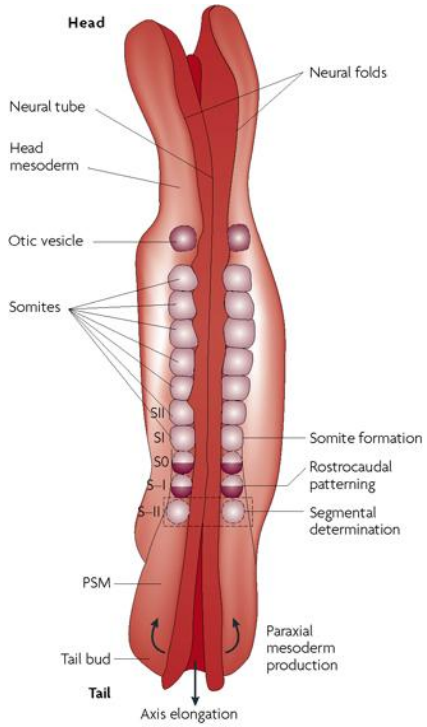


Figure 2: Skematic Dorsal view of a 4-week-old human embryo showing somites and the presomitic mesoderm (PSM) forming the paraxial mesoderm that flanks the axial neural tube.

Source: Mary-Lee Dequéant & Olivier Pourquié. (2008) Segmental patterning of the vertebrate embryonic axis. *Nature Reviews Genetics* 9, 370-382

It has been shown, from previous experiments in chicken (*Gallus gallus*), that the formation of each somites (somitogenesis) it is a highly regulated process that follows a 90 minutes cycle. This hour and a half oscillation relies in the negative feedback of mRNA and proteins of the different genes involved directly (as a part of the somitogenesis) or indirectly (as regulators) in the process, as well as the short half lives of the key components (1).

The research where the data comes from is focused in one of the key components which regulate the timing of this cycle, a CdK (Cyclase Dependent Kinase) inhibitor called roscovitine. CdKs are the core of the cell cycle regulation and important transcriptional regulators. Roscovitine inhibits these kinases, altering the expression and regulation of other key cell components and therefore altering the duration of the cell cycle (1). The activity of this molecule plays an important role during the morphogenesis of the vertebrate body and the somitogenesis. Roscovitine, as cell cycle disruptor, has also been shown to play key roles in apoptosis, and that is why it is being studied as a drug candidate for the treatment cancerous cells in different types of cancer.

In the present study, 8 chicken early-stage embryos were treated or untreated with roscovitine (4 replicates x treated + untreated samples). After that, two different techniques were used to obtain the gene expression profiles: a microarray (Affymetric GeneChip array) and RNA sequencing. The data from these analyses was the starting point for the present project.

Overview of the project

The aim of the present project is integrate the different bioinformatics techniques learned and used during the module in a practical exercise using real data from real experiments where all the skills learned can be applied.

The first part of the project will consist in, given the data from the microarray/RNA Sequencing experiments, use R programming to write the proper scripts that will allow to obtain the most differentially expressed genes between the treated and untreated samples for each set of data. For each technique set of data, different packages specifically designed for this kind of bioinformatical analysis, which will be explained in detail in the corresponding sections.

From these analyses, two lists of genes will be obtained, one for each technique, corresponding to the most differentially expressed. As the techniques used were different, as well, as the RStudio packages, normalization processes and R scripts, the genes obtained can be very different between the two techniques and even between two analyses of the same technique data.

Two genes from these lists were selected for the next steps, in order to work with more than just one variable and be able to work with more than one point of view, but at the same time using a reasonable low number of genes, as some of the scripts in the next steps took a long time to run. In future projects or real-work analyses, more genes should be selected (as many as possible) in order to have a wide point of view of the results, as well several different points of view to compare.

For each of the two genes, the mRNA (the reason why mRNA sequences were used will be explained below in the corresponding section) was searched in NCBI Nucleotide database and a MEGABLAST search against the same Nucleotide database was performed. Repetitive sequences (such as different splicing versions of the same gene) were removed and the rest of the sequences were downloaded in a FASTA file.

The sequences of each gene were aligned using two different multiple sequences alignment algorithms: MUSCLE and Clustal Omega. These alignments were phylogenetically analyzed using RStudio and the proper packages and scripts in order to build two different phylogenic trees (one for each alignment) for each gene. Each phylogenetic analysis included fitting each tree with the most suitable nucleotide substitution model, as well as a bootstrapping in order to know how good the final fitted tree for each alignment of each gene was.

Finally, the trees generated were compared with the ones that the NCBI and ENSEMBL generate by themselves, in order to observe the main differences and similarities between them and discuss them and their causes.

Lastly, to close the project, a close homologue of each of the two selected genes was selected, and a BLAT search against some closely related to chicken genomes, as well as the own chicken genome, in the ENSEMBL database was performed. This retrieved a genomic map for each of the genomes indicating regions with homology for our sequence (the closely related homologue gene), as well as the region in the genome were the gene's homologues are located. A comparison between the different genomes was performed in order to see if the location of the selected genes was conserved between the different species selected (syntenia) or not, and discuss these results.

A summary map of the project can be seen in figure 3.

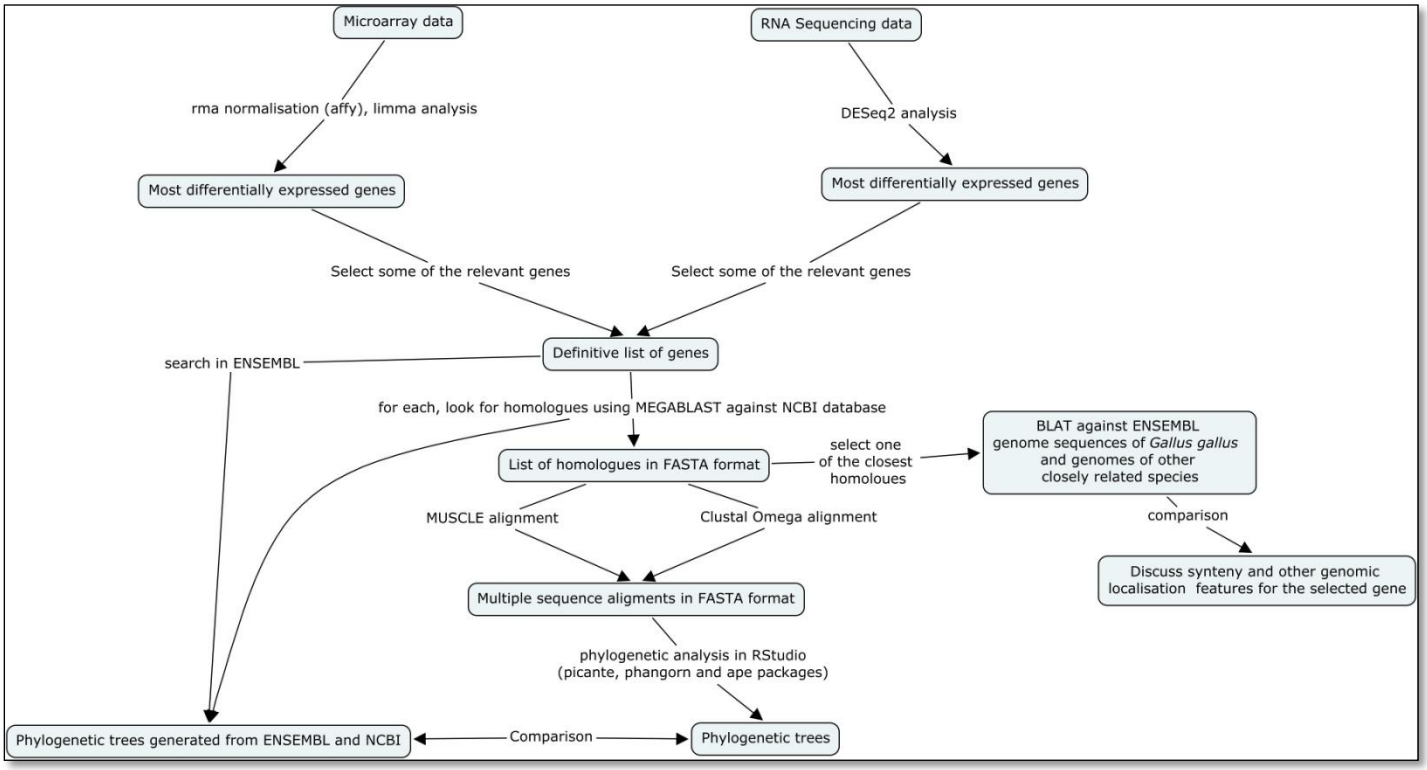


Figure 3: Summary map of the Project. Includes all the relevant points and a summary of the analyses performed in each step.

Identification of the most differentially expressed genes

The first step for the data analysis was identifying the most differentially expressed genes (between treated and untreated samples) for each set of samples from each technique (microarray and RNA Sequencing). For both set of samples, programming in RStudio was used to extract the desired data but, as the two techniques are pretty different, as it is the format of the output files for those, different scripts and packages were used for each technique data, which will be described below in the corresponding sections.

RStudio (2) is a free open source software that provides an accessible environment for programming using R language (Figure 4). It is used for statistical computing and graphics, being very useful for its intuitive interface, flexibility and the high number of different packages available to perform almost any statistical analysis in any area, from economy to biology.

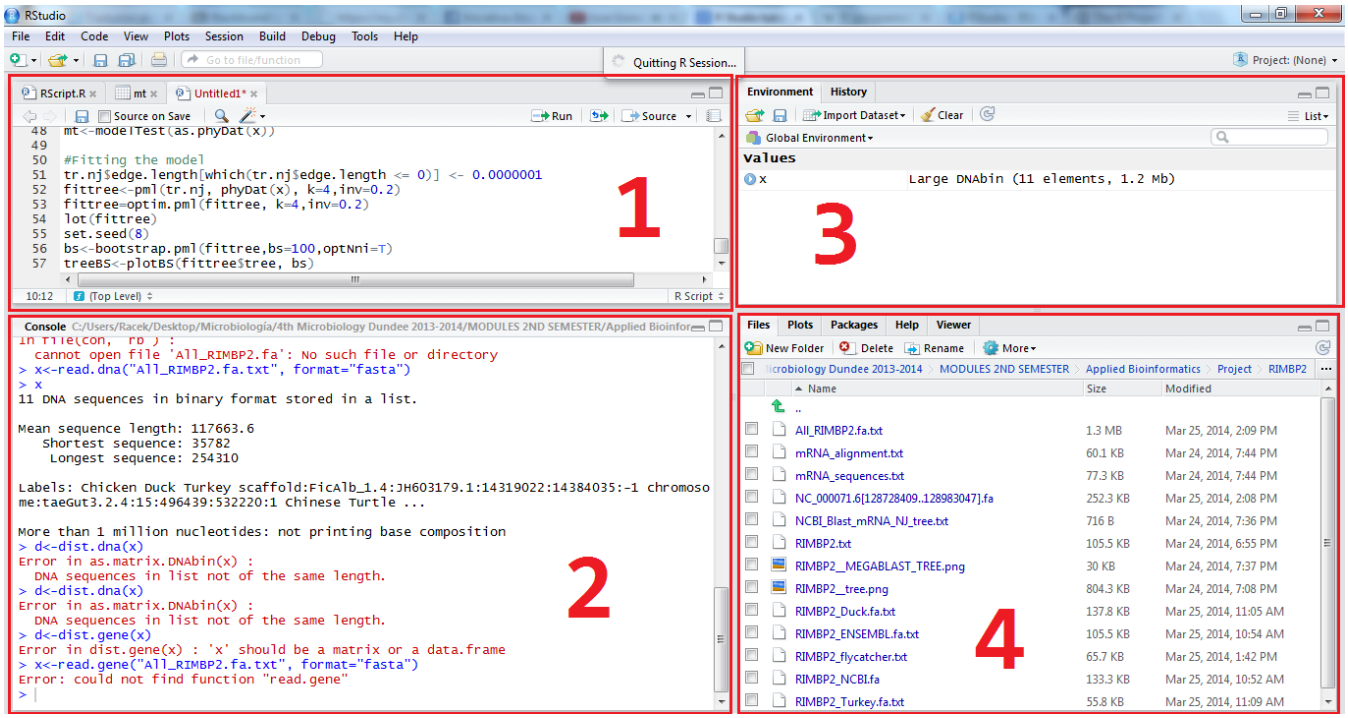


Figure 4: RStudio interface in Windows. **1.-** Data window (where data can be visualized and scripts can also be opened/created and edited and then be sent to the console to be executed. **2.-** Console (whercomplete definition of their content. The historye the code is executed, either from the console, from a R file or directly writing in the console). **3.-** Environment/History (The environment contains all the objects, matrix, lists and values created. The history contains all the commands executed in the console from the start of the session). **4.-** Files (shows the files present in the current working directory), plots (contains all the plots created during the session), packages (all the packages being used) and help (where help about the different packages/commands can be found).

R (3) is a free software programming language specially developed for statistical computing and graphics. It allows the user an easy data manipulation, calculation and graphical display, and can be easily expanded using the different packages developed by different users around the world, that provides specific tools for specific data analyses.

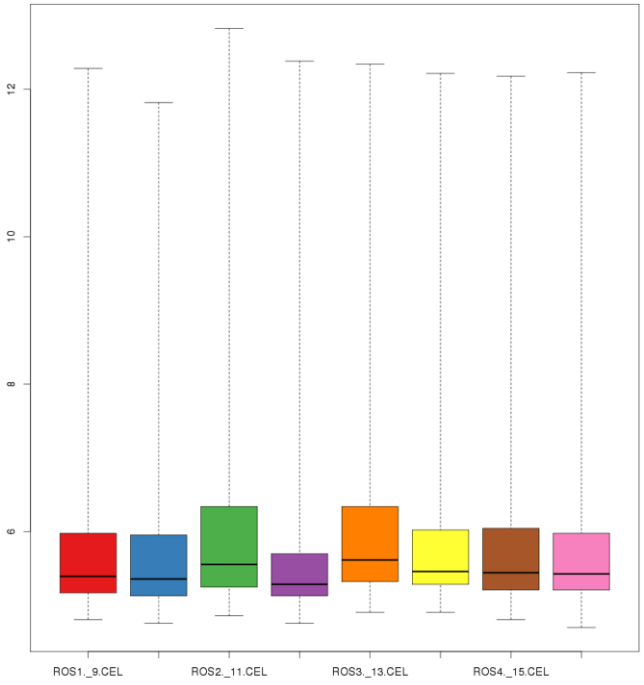
Microarray

Microarray is a technique to measure the expression levels of a high mnumber of genes at the same time, but can also be used to genotype genome regions. The technique was first described in the 80's, but it was popularized in their current form (as miniaturized microarrays or chips) in the last 90's. The principle is easy to understand: each chip contains thousands of spots, each of them containing single stranded DNA with a specific sequence (for example, the complementary strain of a mRNA for a gene of interest). After that, cDNA from our sample (mRNA is extracted and retrotranscribed into cDNA) from all the genes being expressed at a given time under certain conditions is spotted. If the gene which complementary strand is located in the spot is being expressed, its cDNA will hybridize with the DNA in the spot. Then, a fluorophor, is used to detect the spots where there is hybridization and also detect the expression level, according to the level of fluorescence (which is correlated with the number of hybridized DNA). As each spot can contain DNA for one gene, the expression level for each of them can be detected and stored in a file as a numeric value, which it is what the files that would be used contains.

The data obtained from the microarray was stored in eight .CEL files, one for each replica (four) and treatments (treated with roscovitine or untreated). The processing of these files to obtain the data can be summarized in the following points:

1. Box plotting the expression values of each file from the raw data.
2. Normalization using the rma algorithm (Robust Multi-array Average).
3. Box plotting the expression values of each file from the normalized data.
4. Limma analysis to obtain the mostly differentially expressed genes between treated and untreated samples.

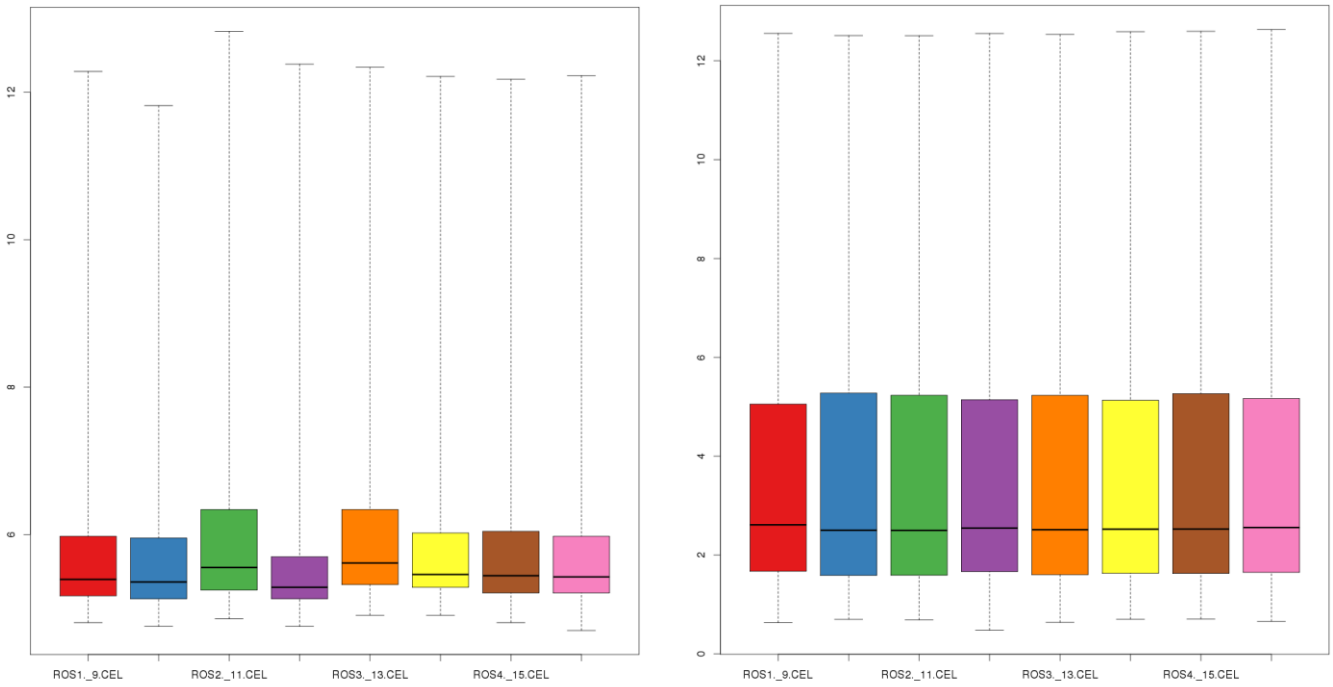
As said, the first step consisted in box plotting the raw data in order to compare this box plot with the one that would be generated after the normalization, and confirm that the normalization had worked as expected (Plot 1).



Plot 1: Box plot of the unnormalized data. As it can be seen, the values distribution is highly different from one sample to another.

The following step was normalizing the data. For this, the affy package for RStudio was used. Affy (Affymetric microarray analysis) is a specific package to analyze data coming from microarrays performed using Affymetric chips. This package forms part of a set of different packages focused in the analysis of genomic data developed under the name of Bioconductor (4). Among the different tools included in the affy package, normalization can be performed

using different algorithms. RMA was chosen to perform the normalization, and the output normalized values were box plotted to be compared with the unnormalized ones (Plot 2)



Plot 2: Unnormalized data box plot against normalized data box plot. As it can be observed, the normalized data (in the right) expression values distribution follows the same scale and are equitably distributed.

After that, another correction process was performed, consisting in removing from the analysis genes with low expression values and/or low differential expression values. This was done to remove values too close to the background noise that could interfere in the analysis, as well as to remove genes that would not have a variation in their expression and therefore will not be differentially expressed (which is what is wanted to be detected). To perform this, as well as the next step (the identification of the mostly differentially expressed genes), another Bioconductor package was used: limma (Linear Models for Microarray Data). Limma is a package specially designed for differential expression analysis of data from microarray experiments using Bayesian and other statistical methods. It must be said that any gene was removed from the dataset using this analysis.

The last step in the microarray data processing consisted on identifying the most differentially expressed genes. Using the limma package, the genes were sorted by their differential expression. This table was then correlated with the chicken database (chicken.db) to get the ENSEMBL ID and gene SYMBOL for all the genes (as in the files those genes contains an own identification codename). After that, the top 5 were picked out and written in a file (table 1).

ENSEMBLID	SYMBOL	logFC	AveExpr	P. Value	adj.P. Value
ENSGALG00000006719	GNAZ	2.2369697	6.3674884	0.05498	0.98028249
NA	NA	2.9475005	3.397566	0.00615	0.98028249
NA	NA	3.1375353	-2.23544	0.04762	0.98028249
NA	NA	4.6917198	-1.71278	0.11541	0.98028249
NA	NA	5.2099695	-1.826574	0.09564	0.98028249

Table 1: Most differentially expressed genes for the microarray. As it can be seen, the adjusted p value for all them is the same, so the genes were sorted by the difference in the average expression, being GNAZ the most differentially expressed with a 6 fold difference most expressed in the treated sample.

As it can be seen in the table, four of the five genes had no ENSEMBL ID or SYMBOL associated. The complete script was checked and the analysis was performed again, but no errors were found and the problem remained after the second analysis.

RNA Sequencing

RNA sequencing, also known as WTSS (Whole Transcriptome Shotgun Sequencing) is a new generation technique that takes advantage of the potential of the next generation sequencing techniques to detect the quantity of each RNA from a genome at a given time under certain conditions. The main advantage over microarrays is the higher coverage of RNA-Seq, as no complementary strands to hybridize are needed, which makes that, if there is no spot in the array for one of the genes being expressed, it will not be detected. Furthermore, the sensibility of RNA-Seq is also higher, allowing detecting mutations and amounts of RNA that cannot be detected in microarrays. There are several methods to perform this technique, from direct sequencing to a previous library of poly(A)RNA building, in any case, the output data obtained from the different methods is of the same kind, and the variations lies in the sensibility and other features.

For the analysis of the RNA-Seq data, another Bioconductor package was used: DESeq2, an R package to analyze the count data from RNA-Seq and identify the mostly differentially expressed genes. BioMart, another package that allows querying huge amounts of data from different databases, was also used to query information about the genes in our dataset to the ENSEMBL database.

As it was done in the microarray analysis, the data was normalized, the normalization was checked, and then, after the analysis, another list of the top expressed genes was obtained (Table 2).

ESNEMBL ID	SYMBOL	Base mean	Log2foldchange	P value	Adj P value
ENSGALG00000002579	RIMBP2	144.133904127055	2.76208228473817	2.3796259813202626e-15	2.3182316310022e-11
ENSGALG00000001115	MMEL1	36.6008832799172	-2.19181334357884	9.23842185870153e-15	4.50003528737352e-11
ENSGALG00000025884	NABP1	198.549937094987	-1.56225815080273	5.88645514155538e-12	1.91152819963442e-08
ENSGALG00000025884	NABP1	198.549937094987	-1.56225815080273	5.88645514155538e-12	1.91152819963442e-08
ENSGALG00000025884	NABP1	198.549937094987	-1.56225815080273	5.88645514155538e-12	1.91152819963442e-08
ENSGALG00000020492	MUC6	34.0359894722591	2.48869057252896	3.45990651662423e-11	8.42660232123832e-08

Table 2: Most differentially expressed genes for the RNA-Sequencing. The genes are sorted by fold change (in log2) adnd adjusted p value (the lowest the value, the most confidence in the result), as both are correlated (the biggest fold changes correspond to the lowest adjusted p values).

Surprisingly, none of the top 6 genes identified was the also identified in the microarray analysis. As the other four genes from the microarray had no SYMBOL or ENSEMBL ID associated, it cannot be known if any of them were also in the RNA-Seq list.

For each gene in the list, some brief general information was searched in order to suggest if any of them could play an important role in the somatogenesis.

GNAZ

Guanine nucleotide-binding protein G(z), subunit alpha. The protein encoded by this gene is a member of the family of the G proteins. It is related with the maintaining of the ionic balance in the cochlear fluids, as well as interact ins everal G-protein mediated pathways, which are one of the major regulatory systems in the body. Furthermore, it has been shown that the levels of this protein are higher in the fetal brain, which may indicate that this protein can develop some function in the embryonic development.

Source: http://www.genecards.org/cgi-bin/carddisp.pl?gene=GNAZ

RIMBP2

RIMS binding protein 2. Present in several parts of the body, and related with brain problems such as Alzheimer in humans. Interacts with RIMS (Regulating synaptic membrane exocytosis) proteins, which are involved in the regulation of vesicle traffic in neurons.

For the following steps, two of the genes identified (from both lists) were selected at random: MUC6 and MMEL1.

Source: http://www.genecards.org/cgi-bin/carddisp.pl?gene=RIMBP2

http://en.wikipedia.org/wiki/RIMS1

MMEL1

Membrane metallo-endopeptidase 1. Degrades small peptides smaller than 3kDa which contains neutral aromatic or aliphatic residues. It is related to sperm function, regulating the fertilization and early embryonic development, so tht would be another good candidate to study for the research.

Source: http://www.genecards.org/cgi-bin/carddisp.pl?gene=MMEL1

NABP1

Nucleic acid binding protein 1. It binds ssDNA, and play important roles in a lot of different cell processes. It is associated with the SOSS system (a multimeric complex which repairs DNA), which is also associated with the mitosis cell cycle checkpoint , and therefore it could play some role in somatogenesis by interacting with the CdKs and other components involved in the cell cycle checkpoints that also regulate the somites formation process.

Source: http://www.genecards.org/cgi-bin/carddisp.pl?gene=NABP1

MUC6

Mucin 6. May be involved in the modulation of the composition of the protective mucus layer in response to the presence/absence of bacteria dn other agents in the lumen. It is also relates with thecytoprotection of epithelial surfaces, and it is used as a tumor marker in some cancers.

It is also suggested that may play a role in epithelial organogenesis. As this is also a part of the developments, roscovitine, involved in the cell cycle, may have been regulating also this process trough this gene regulation.

Source: http://www.genecards.org/cgi-bin/carddisp.pl?gene=MUC6

Homologues search

Once the lists of most expressed genes for each technique were obtained and the two genes to perform the rest of the analyses were selected (MUC6 and MMEL1), the next step was identifying, for each of them, homologous genes to perform a phylogenetic analysis which includes a tree building to visualize the phylogenetic relationship between all the homologues.

First approaches and problems

The first idea to identify the homologues was perform a search using the gene symbol in the Nucleotide NCBI database. Nucleotide is a huge database that includes sequences from several different sources, including GenBank, RefSeq, TPA (Third-Party Annotation) and others. In other words, any kind of DNA and RNA sequences in all the databases associated to NCBI. The aim of this search was find the DNA sequence for MUC6 and MMEL1 genes to then perform a MEGABLAST search with that sequence against the own Nucleotide database to find homologues.

The problem came with the results for this search, as there was no sequence for the gene, only the whole genome sequence for the *Gallus gallus* chromosome 21 and 5 (where MMEL1 and MUC6 are located, respectively), and the only sequence that could be obtained from them was the whole loci sequence, which implies working with sequences of more than 100.000 base pairs, which make the posterior analyses (alignments, tree building) virtually impossible to perform in the available computers, as well as retrieving different results than the ones that would be found working only with the gene sequence, as the whole loci may contain unrelated regions that can change more during the evolutionary process, changing the final results from the ones that would be obtained working only with the gene sequence. The other sequence found with this search was the mRNA predicted sequence for each gene, but that was initially discarded, as the main objective was working with the original gene DNA sequences.

The second idea was search the gene in the ENSEMBL database, which does contain the DNA gene sequence. The sequences for both genes were downloaded and, using them MEGABLAST search against Nucleotide was performed (just in case that the sequences were in the database, but with different names as they could have not been classified yet), but, as expected, the only sequences found were the *Gallus gallus* whole genome chromosomes sequences (as the loci for each gene in those sequences obviously matched with the MMEL and MUC6 sequences from ENSEMBL) and mRNA sequences for the MUC6 and MMEL1 of *Gallus gallus* and other related species.

As none of them were DNA gene sequences, a second search was performed in the ENSEMBL database. The entry for each gene (in our case, MUC6 and MMEL1) contains a lot of information about the sequence, location, variants, expression and also phylogenetic relationships with homologues genes found in the ENSEMBL database (Figure 5). Using this information, some of the homologues were searched and picked up from the ENSEMBL database and stored in a FASTA file in order to perform the following steps.

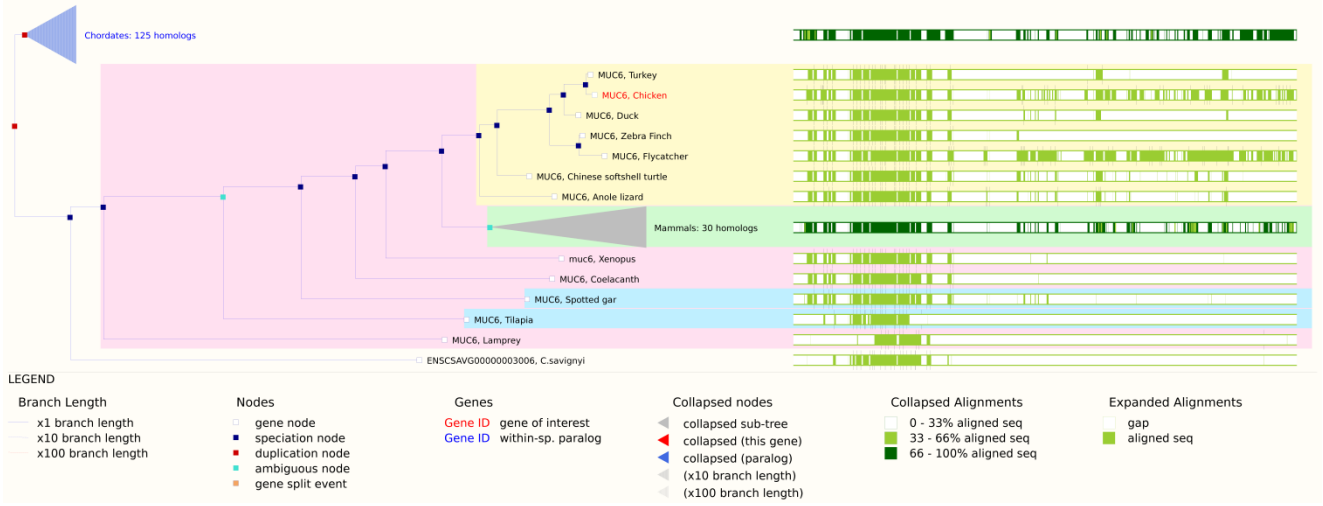


Figure 5: Phylogenetic tree built by the ENSEMBL site for MUC6 using the homologous genes contained in their database, which shows the evolutionary relationship between the. As the algorithms and techniques used to build this tree were not specified, it was going to be used only as a reference to pick homologues and then perform an own multiple sequence alignment and tree building and confirm or not these phylogenetic relationships. Unluckily, as said, the sequences in the database were not suitable for the alignment methods available.

Nevertheless, when the next step was tried to be carried out, it was seen that these sequences were not also suitable for the phylogenetic analyses. It was tried to perform a multiple alignment of the two set of homologous sequences using two different algorithms (MUSCLE and Clustal Omega, that will be detailed below), but the analyses were retrieving an error message, indicating that the process could not been completed. After some trials and errors, it was realized that some of the sequences downloaded were uncompleted (in other words, some regions in the sequence contained regions with repetitive “NNNNNNNNNN”, which indicates that this part of the sequence was not available yet) and that was the reason the multiple alignment software crashed when they tried to align the sequences. The first idea to solve this problem was remove those sequences, but that implied working with a very low number of homologues, which would have made the analyses , under the point of view of the student, not enough statistically significant.

Working with mRNA sequences

The solution chosen was working with the mRNA sequences found during the first search against Nucleotide. Those sequences were all completed, with no gaps, and, as they are the RNA version of the gene sequence, the phylogenetic relationships should be highly similar, if not identical.

The mRNA sequence of MUC6 and MMEL1 was MEGABLASTed against the Nucleotide database, and a list of homologous mRNA of MUC6 and MMEL1 in other species was retrieved. The repetitive sequences from the same species were removed (as in some cases there were slightly different splice variants –but highly related between them, so all the splices would be together in the phylogenetic tree-) and the rest were downloaded in a single FASTA file for each gene. This file would be used for the following phylogenetic analyses. NCBI has an own tools to generate unfitted Neighbour–Joining (NJ) trees using the MEGABLAST data (which includes %identity, e-scores, alignment scores and other relevant data). This tree was also downloaded in order to be compared with the trees that would be generated in the following steps.

Phylogenetic analyses

Multiple sequence alignment

Once we had the homologues list in FASTA format, the main objective was performing a phylogenetic analysis for each set of sequences in order to see and discuss the evolutionary relationships between them.

The first step for this was performing a multiple sequence alignment (MSA). There are several ways to do this, including a large set of different algorithms. For the present project, two of the most popular algorithms were used: MUSCLE (multiple sequence comparison by log-expectation) and Clustal Omega. The decision of using two different algorithms was done because they are the most popular methods for each of the two major alignment approaches used: progressive alignment (Clustal) and iterative method alignment (MUSCLE). This would also useful to check if the relationship between the different homologues is “strong” or “clear” enough to be the same or very similar using both methods, or if the two methods retrieved alignments different enough to change the final phylogenetic tree (as the method used to build the phylogenetic tree would be the same).

Clustal is probably the most popular method used for multiple sequence alignments, which uses a progressive alignment construction, also known as hierarchical or tree method. This is a heuristic method that has two main stages: the building of a initial guide tree using methods such as UPGMA or NJ; and after that, the multiple alignment process, where each sequence is added to the growing MSA according to the guide tree (in other words, one sequence is used

as starting point, and it is aligned with the most closely relates according to the guide tree, then the next sequence is added and aligned, and the same for the rest of them). The main problem of this method is that, as they use a guide tree as a reference, any error in the building of this tree would be extended to the rest of the alignment. The version used for this project is Clustal Omega, the newest version available, which contains secondary methods to ensure the quality of the alignment.

MUSCLE is becoming most and most popular in the last times. It is a iterative method, which means that it realigns the initial sequences at the same time new sequences are added to the MDA, which allows a better optimization of the final result. The strong point of MUSCLE is that it has a very accurate distance measurement, which makes the final relation value between the different sequences more reliable. These are the reasons why MUSCLE (as well as other iterative methods) is considered, in a general way, a more accurate method than Clustal and other progressive alignment methods.

The alignments for both set of data and both methods were performed using the EMBL-EBI website (4), as it has a very intuitive interface and supports a lot of different formats for the input and output files. It is also possible to change the default options, allowing the user to personalize the alignment. For this project, the default options were used, as it was considered that the number of sequences and the size of each of them did not required any kind of special treatment to improve the alignment.

From this step, 4 files in FASTA format were retrieved, containing the different alignments for each method and set of homologues. These files would be used to perform the phylogenetic analysis and tree building in the following steps.

Distance matrix and tree building and optimization

The final step of the phylogenetic analysis was the tree building. For this, the guidelines given by the professor David Booth were followed in order to build a proper phylogenetic tree, find the most suitable nucleotide substitution model for each alignment of each set of homologues and build a new tree fitting that model. Finally, a bootstrapping for each tree was performed in order to check the quality of the final fitted tree.

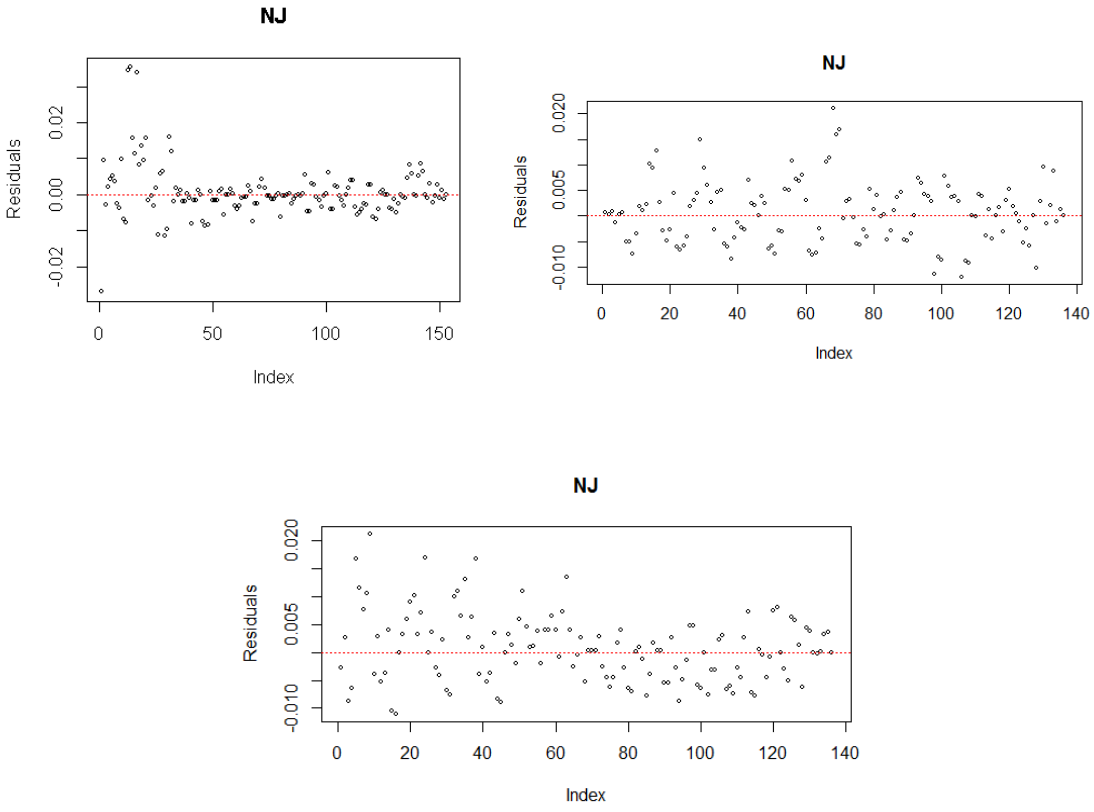
Once all the trees were built, all of them were compared: the ones generated by the process that would be explained (from the MUSCLE and Clustal alignments) and the ones generated by NCBI and ENSEMBL sites.

Three packages were used for this section: ape, phangorn and picante.

These three packages are designed to read, plot, manipulate and test phylogenetic trees and perform analysis of phylogenetic distances. **Ape** is focused in the tree building, plotting and manipulation. **Phangorn** is more focused in analyzing the phylogenetic trees and test the models and distances to optimize and fit the trees to the best model. **Picante** is another package used for phylogenetic analysis that contains a set of tools to phylogenetic diversity metrics, perform comparative analyses and other analyses focused in distributions community structure and species interaction (more focused to be used in ecology).

For all the trees, the same schematic procedure was used:

First of all, the data was read from the alignment FASTA files, and a distance matrix was calculated using the default K80 (Kimura) model, which uses an algorithm that distinguishes between transitions and transversions, as they have different probabilities to occur (6) .Using this matrix, an tree using the NJ algorithm was built, and then the distorsion between the matrix values and the tree was also calculated in order to confirm that the NJ algorithm being used was good enough. As it can be seen in plot 3, the distortion for all the trees built was minimum, which indicates that the trees were highly representative of the data in the matrix.



Plot 3: Distortion plots for the different unfitted trees built in the first step of the phylogenetic analysis. In the top left, the distortion values for the MUC6 MUSCLE alignment, in the top right, the values for the MMEL1 Clustal; in the bottom, the values for the MMEL1 MUSCLE alignment. In all the cases, the values were really small, indicating that the NJ was a good representation of the data in the matrix.

Unluckily, the tree for the Clustal alignment for MUC6 could not be performed, as the distance matrix contained missing values, which are not allowed to build the tree. The alignment was checked in Jalview, looking as good as the other alignments, and the sequences were realigned again, retrieving the same error.

After that, the objective was fitting the tree to the most suitable substitution model, rather than use the K80 default model. For this purpose, a model test was performed using the phangorn package, which analyze the alignment with each model and determines which of them is most suitable given the sequences present in our alignments. In all the cases, the most suitable algorithm suggested by the model testing was the GTR model (Generalised Time-reversible), which is the most neutral and general model available. It considers each possible nucleotide substitution to have a certain probability to occur, so 6 different substitution rates are used, and also 4 equilibrium base frequency parameters are also considered (7).

Knowing that, all the trees were fitted to a new matrix generated using the GTR model, and a new tree was built in each case. The final step for the phylogenic analysis would have been performing a bootstrap for each fitted gene and then plot the resulting values in the tree.

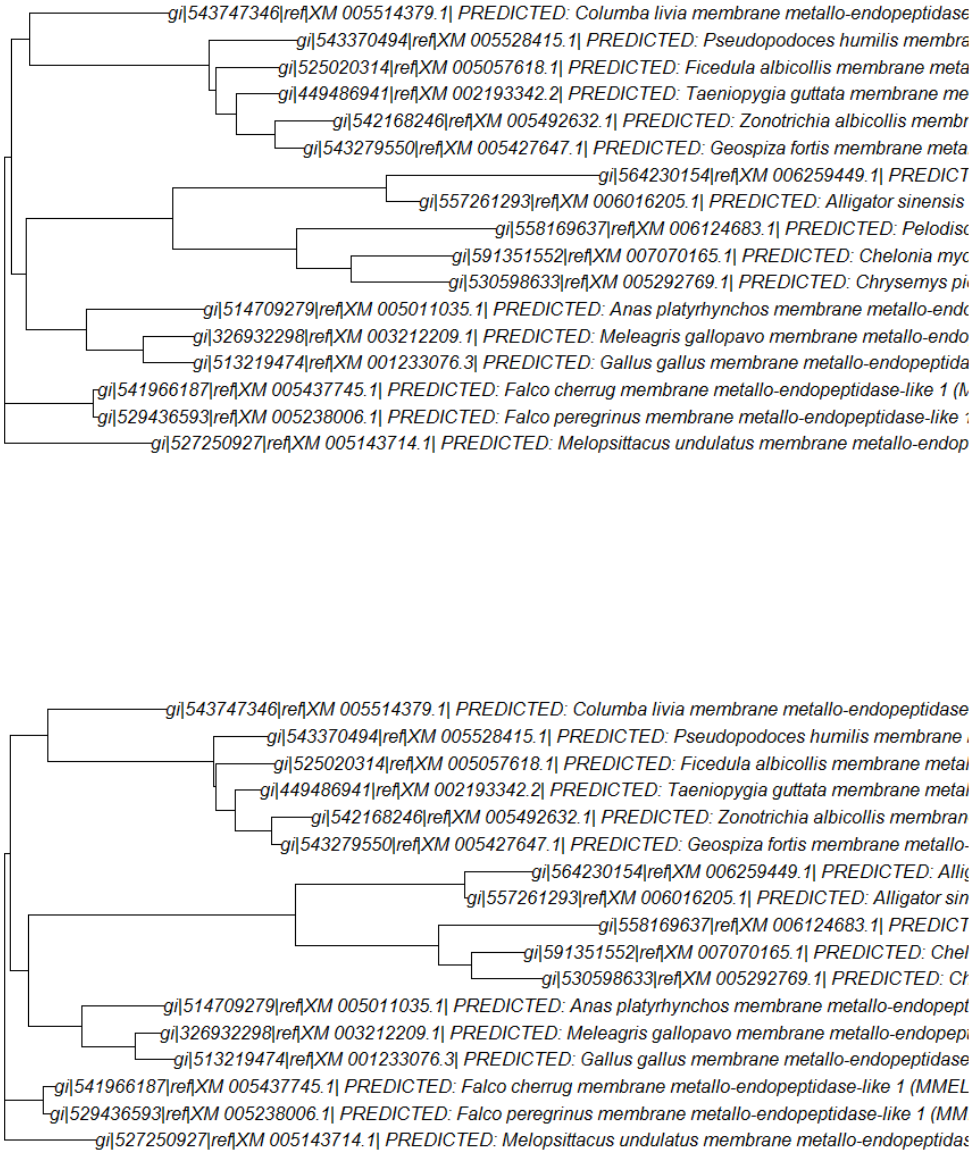
Bootstrapping is a statistical method used to measure the confidence/accuracy of a given estimation. In phylogenetic trees, boostrapping measures what proportion of the data supports each node of a given tree. In other words: how confident can we be that a given relationship in a phylogenetic tree is true or not. To perform this, some of the data is removed from our alignment, and a tree is plotted following the same procedure used for the ditted tree being checked. This tree is compared with the fitted tree, node by node, checking how many of the nodes are shared or not. Then another piece of alignment is removed, and the process is removed. The resulting values for each node indicate how many of the trees built during the bootstrapping shared each node with the original fitted tree, and therefore indicating how supported by the data each node is (for example, if, from 100 bootstrapping cycles, a node have a score of 100, that indicates that no matter what piece of data is removed, the relationship indicated by this node remains, which is an indication that that particular node is highly supported by the data).

The main problem in this process was when the fitted tree with the bootstrapping values for each node was tried to be plotted, as the resulting plot was an unrooted, branch-ioverlapped tree impossible to read. It was tried to root the tree, but for some reason it was absolutely impossible.

Tree comparison and discussion

Some comparisons and discussions can be done from all the available trees built and downloaded from NCBI and ENSEMBL.

The first and easiest comparison that would be nice to discuss would be the one between the unfitted and the fitted tree built during the phylogenetic analysis. The trees from MMEL MUSCLE will be used as an example, as all the pairs of unfitted-fitted trees have the same kind of differences.



Tree 1: Unfitted (K80) MMEL1 MUSCLE NJ tree (top) and fitted (GTR) MMEL1 MUSCLE NJ tree (bottom)

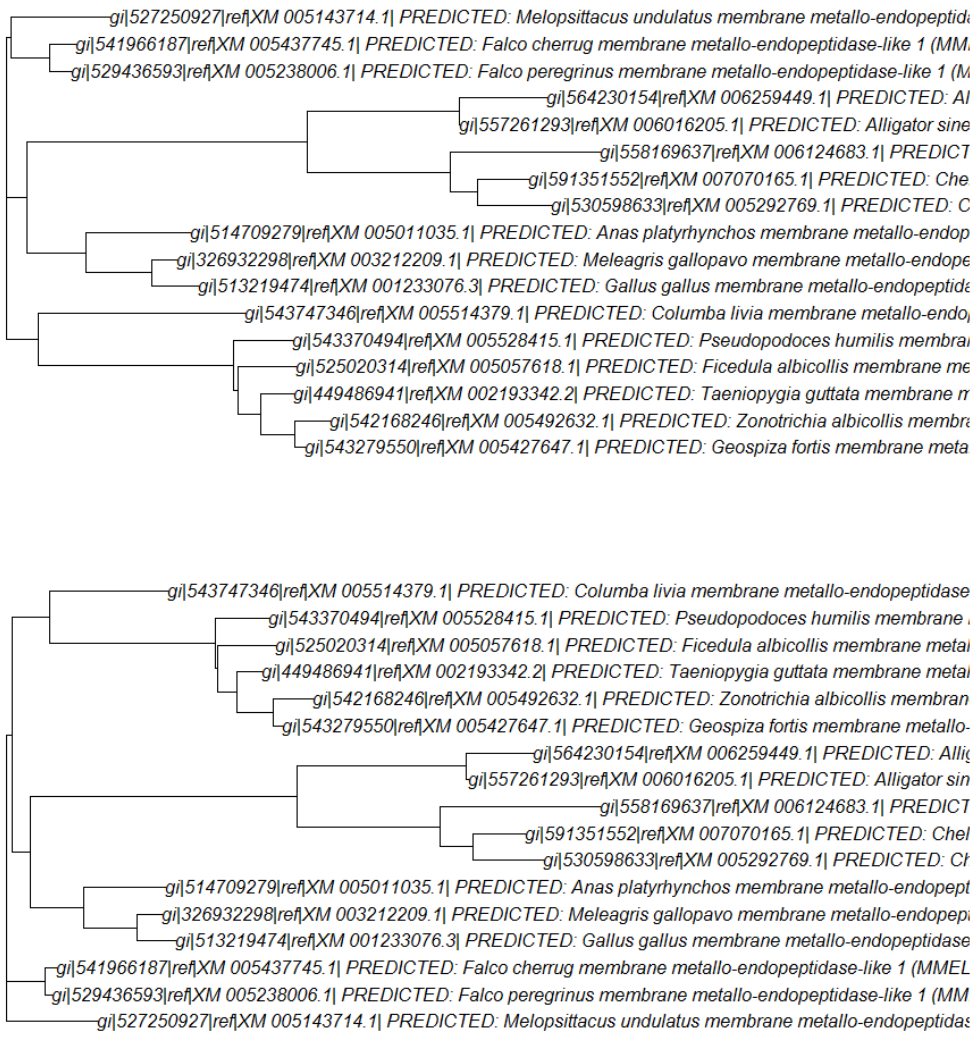
As it can be seen in Tree 1, there is no big differences between the unfitted and the fitted trees (as expected), but there is a correction in the length of most of the branches, which modifies the evolutionary distance between most of the homologues. It is important to notice that, even though there is not any reorganization in any of the branches in the tree, using different models can change the phylogenetic distance between species, and that can be extremely important for the accuracy of the evolutionary history, as slightly differences between two trees (as the ones that can be seen in Tree 1) can suppose hundreds of thousands of years, changing completely details in the phylogenetic history. It can be concluded, then, that choosing the right model is extremely important in this kind of works, and test as many models as possible to find the best one for each set of data to be analyzed is crucial to perform a good phylogenetic analysis.

A second interesting comparison would be between the same tree of a MUSCLE alignment and a Clustal Omega Alignment, in order to observe how the alignment method chosen can influence in the final result. To do this comparison, the fitted trees of MMEL1 for both alignment methods would be compared (Tree 2).

In this case, more significant differences can be observed. Some of the groups of branches only show a slightly difference in their branch length between one tree and the other, but other groups present major differences, with big differences in the branch length. There is also a change in the way the different groups are rooted, as well as a different tree root, but it seems that this is not a difference between the two trees, only a different way to plot the trees.

In all the cases, the differences between the different trees built during the project are only in the length of the branches, which, as it has been said, it is very important, as can suppose a big difference in the phylogenetic history, but those differences do not involve a change in the relationship between close species. In other words: the shape of all the trees compared until now is the same, and there is no major reorganization in the different branches. This is useful, as may indicate that the relationship showed by theses trees between the homologues analyzed is highly supported by the data.

A third and last comparison between the trees built which will be interesting to be performed will be between one tree of each gene (Tree 3).



Tree 2: Clustal Omega NJ fitted MMEL1 tree (top) and MUSCLE NJ fitted MMEL1 tree (bottom)

This analysis is great to analyze whether different genes support the relationship between the species in the trees or if using different genes change completely the phylogenetic relationship between them.

As it can be observed, the two trees present major differences. Nevertheless, seems that the most closely related species to the chicken (turkey and duck) maintain the same relationship between their homologues, in both genes, which may indicate that those species are more closely related between them than with any other in the tree.

On the other hand, some of the relationships present in one of the trees change completely in the other one (for example: in the MUC6 tree *Columbia* is closely related with *Melospittacus,* but in the MMEL1 tree they are more separated). This suggests that in some cases, different genes can be conserved among different species with a different mutation ratio, and sometimes two species may have very similar homologues because, for some reason, they have conserved the gene from the common ancestor, or their gene sequences have mutated in a similar way, while other genes in the same species have suffered completely different mutations, resulting in a high phylogenetic difference between them.

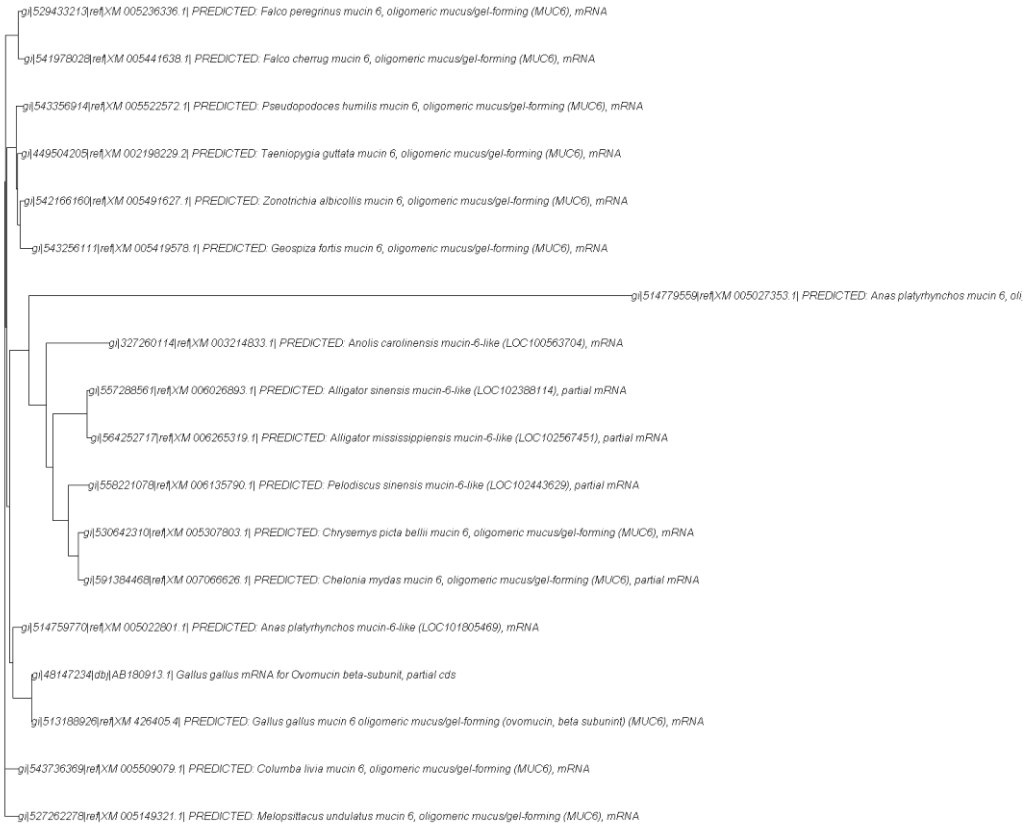
Nevertheless, most of the major groups and phylogenetic relationships were conserved in both trees, which is useful to confirm the closely relationship between these species.

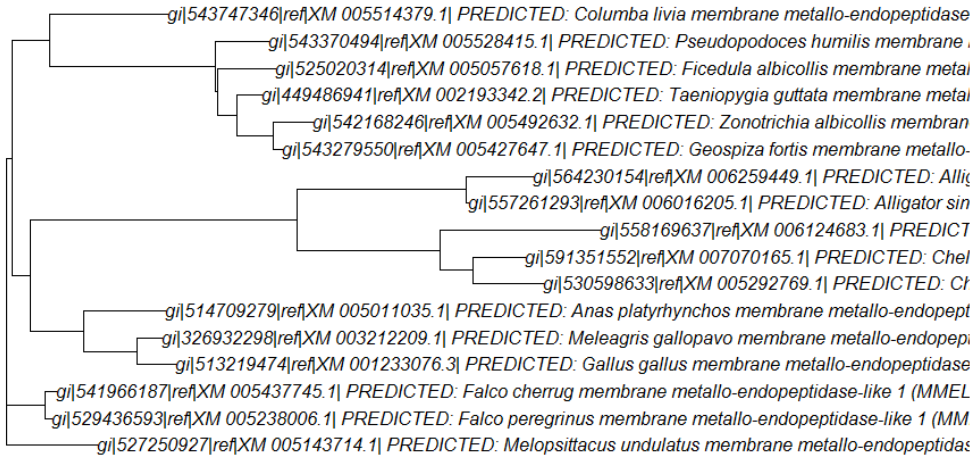
The last comparison to be performed would be between one of the own built trees and the corresponding trees for that gene generated by NCBI and ENSEMBL (tree 4).

As it can be seen, the two most closely related species to chicken (turkey and duck) are common in all three trees, which again indicate a strongly phylogenetic relationship between these species. Nevertheless, the three trees show important differences: in the own built tree, the chicken MMEL1 seems to be more closely related to some alligator and turtle species than other other birds, but in the trees from NCBI and ENSEMBL all the birds genes are together sharing the same common node, independent from non-bird species.

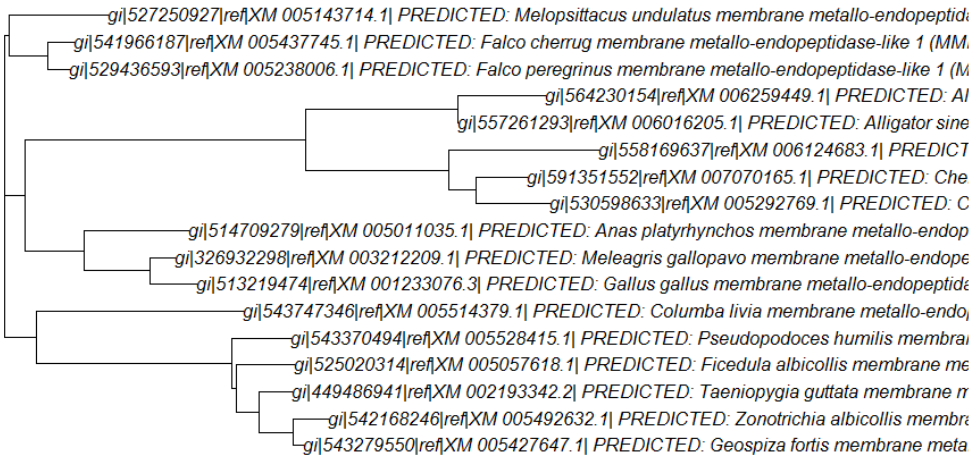
Another important difference lies in the organization of the turtles and the alligators: in our tree, they are grouped together, sharing a common node, but in the NCBI tree, the turtle MMEL1s are shown to be more closely related to the ones belonging to the bird species rather than the alligators.

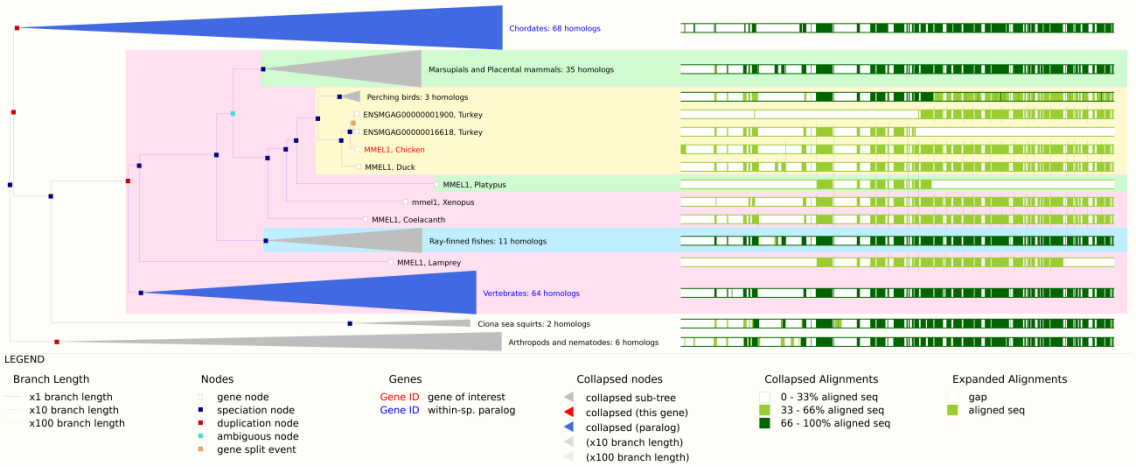
Once again, it is shown that the methodology used to perform the phylogenetic analysis is very important, as different methods lie in completely different results. Ii is true that the ENSEMBL tree uses DNA sequences and the NCBI and our own tree use mRNA, but even between the NCBI and our own tree the differences are big enough to consider completely different phylogenetic histories depending which tree you look at.

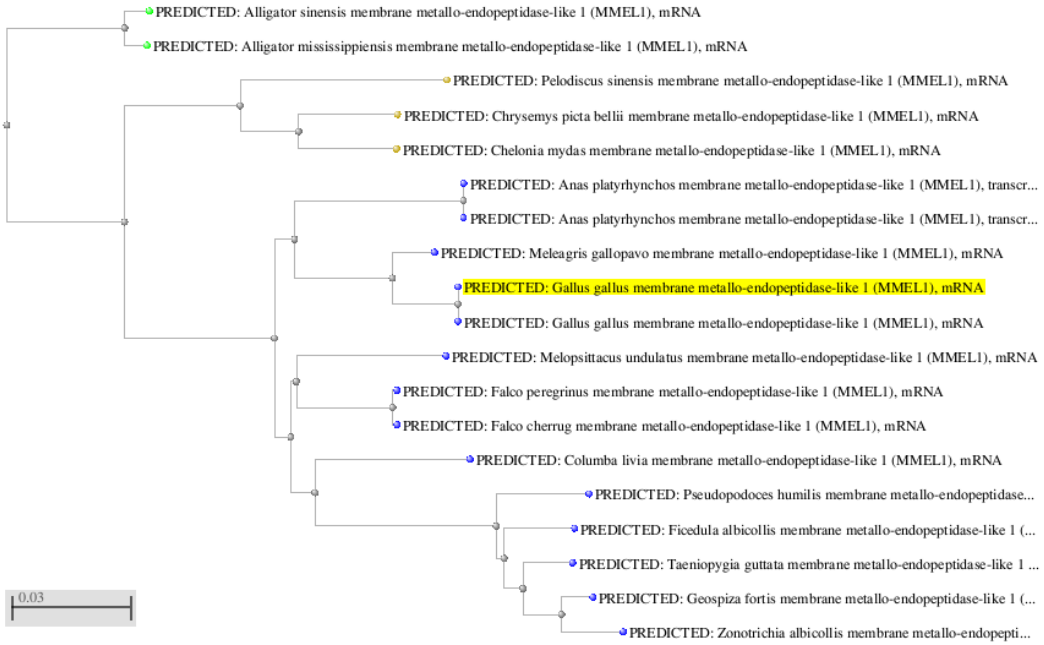




Tree 3: MUSCLE fitted MUC6 NJ tree (top) and MUSCLE fitted MMEL1 NJ tree (bottom)


gi|527250927|ref|XM 005143714.1| PREDICTED: Melopsittacus undulatus membrane metallo-endopeptid
gi|541966187|ref|XM 005437745.1| PREDICTED: Falco cherrug membrane metallo-endopeptidase-like 1 (MM
gi|529436593|ref|XM 005238006.1| PREDICTED: Falco peregrinus membrane metallo-endopeptidase-like 1 (M
gi|564230154|ref|XM 006259449.1| PREDICTED: Al
gi|557261293|ref|XM 006016205.1| PREDICTED: Alligator sine
gi|558169637|ref|XM 006124683.1| PREDICT
gi|591351552|ref|XM 007070165.1| PREDICTED: Che
gi|530598633|ref|XM 005292769.1| PREDICTED: C
gi|514709279|ref|XM 005011035.1| PREDICTED: Anas platyrhynchos membrane metallo-endop
gi|326932298|ref|XM 003212209.1| PREDICTED: Meleagris gallopavo membrane metallo-endope
gi|513219474|ref|XM 001233076.3| PREDICTED: Gallus gallus membrane metallo-endopeptid
gi|543747346|ref|XM 005514379.1| PREDICTED: Columba livia membrane metallo-endo
gi|543370494|ref|XM 005528415.1| PREDICTED: Pseudopodoces humilis membra
gi|525020314|ref|XM 005057618.1| PREDICTED: Ficedula albicollis membrane me
gi|449486941|ref|XM 002193342.2| PREDICTED: Taeniopygia guttata membrane m
gi|542168246|ref|XM 005492632.1| PREDICTED: Zonotrichia albicollis membra
gi|543279550|ref|XM 005427647.1| PREDICTED: Geospiza fortis membrane meta


Chordates: 68 homologs
Marsupials and Placental mammals: 35 homologs
Perching birds: 3 homologs
ENSMGAG00000001900, Turkey
ENSMGAG00000016618, Turkey
MMEL1, Chicken
MMEL1, Duck
MMEL1, Platypus
mmel1, Xenopus
MMEL1, Coelacanth
Ray-finned fishes: 11 homologs
MMEL1, Lamprey
Vertebrates: 64 homologs
Ciona sea squirts: 2 homologs
Arthropods and nematodes: 6 homologs
LEGEND
Branch Length
x1 branch length
x10 branch length
x100 branch length
Nodes
gene node
speciation node
duplication node
ambiguous node
gene split event
Genes
Gene ID gene of interest
Gene ID within-sp. paralog
Collapsed nodes
collapsed sub-tree
collapsed (this gene)
collapsed (paralog)
(x10 branch length)
(x100 branch length)
Collapsed Alignments
0 - 33% aligned seq
33 - 66% aligned seq
66 - 100% aligned seq
Expanded Alignments
gap
aligned seq



Tree 4: MUSCLE fitted NJ MMEL1 tree (top), ENSEMBL tree (middle) and NCBI NJ tree (bottom).

Gene loci comparison between <u>Gallus gallus</u> and closely related species for MUC6 and MMEL1

A last analysis to be performed for this project was a comparison of the location of MUC6 and MMEL1 in the chicken genome and other genomes of closely related species. The species selected for this analysis were the two of the most closely related according to all the phylogenetic trees: turkey (*Meleagris gallopavo*) and zebra finch (*Taeniogypia guttata*). The first intention was use the duck instead of the zebra finch, as it is phylogenetically closer to the chicken, but there was no karyotype available for it in ENSEMBL, which make the results more easily understandable and easy to interpret. Moreover, after the analysis, it was considered that having a species which a bit more evolutionary separated from *Gallus gallus* would be more useful to discuss the synteny between them.

The analysis was performed as follows: one the closest homologous sequence for each gene was downloaded from ENSEMBL, and then a BLAT against the three species genomes in ENSEMBL was performed. The result page contained the information that would be discussed.

Both cases were very similar: the genes selected were highly conserved in the genome of the three different species. In the case of MMEL1, the homologue selected was the duck MMEL1, which was BLATed against the chicken, zebra finch and turkey genomes in ENSEMBL database in order to find the location of MMEL1 in chicken and the corresponding orthologous regions in the other species (Figure 6). As can be observed, the location of MUC6 among the different species is highly conserved in the top region of chromosome 5. Taking into account that the three species contain different number of chromosome, and most of them are very different from one species to another, the conservation of the structure of chromosome 5 and the location of the MUC6 loci is quite interesting.

For the MMEL1 gene, a different homologous was selected, in this case, the MMEL1 of turkey, and, as it was done in MUC6, it was BLATed against chicken, zebra finch and turkey genomes in ENSEMBL database. The results were slightly different, but this gene also shown a high level of conservation in the location in the genome: in chicken and turkey, they are located in the same chromosome (21) in the same loci, and in the case of the zebra finch, even though it seems to be located in a different chromosome, that is not true, as the chromosome 23 of the zebra finch corresponds to the chromosome 21 of chicken and turkey.

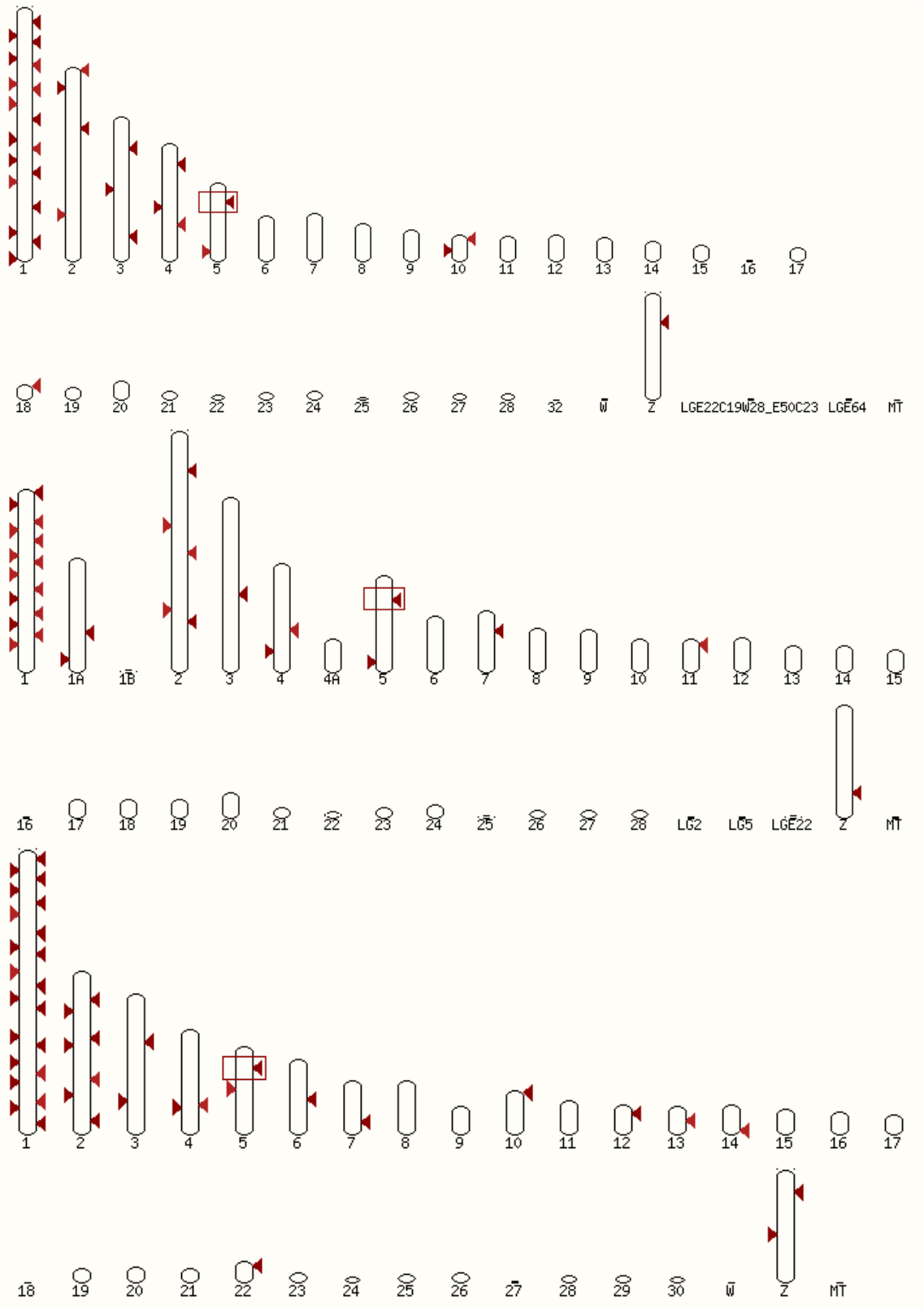


Figure 6: Karyotype of chicken (top), zebra finch (middle) and turkey (bottom). The arrows indicate zones of homology for the duck MUC6 sequence, and the red rectangle shows the region of maximum homology, corresponding to the location of the MUC6 gene for each species.

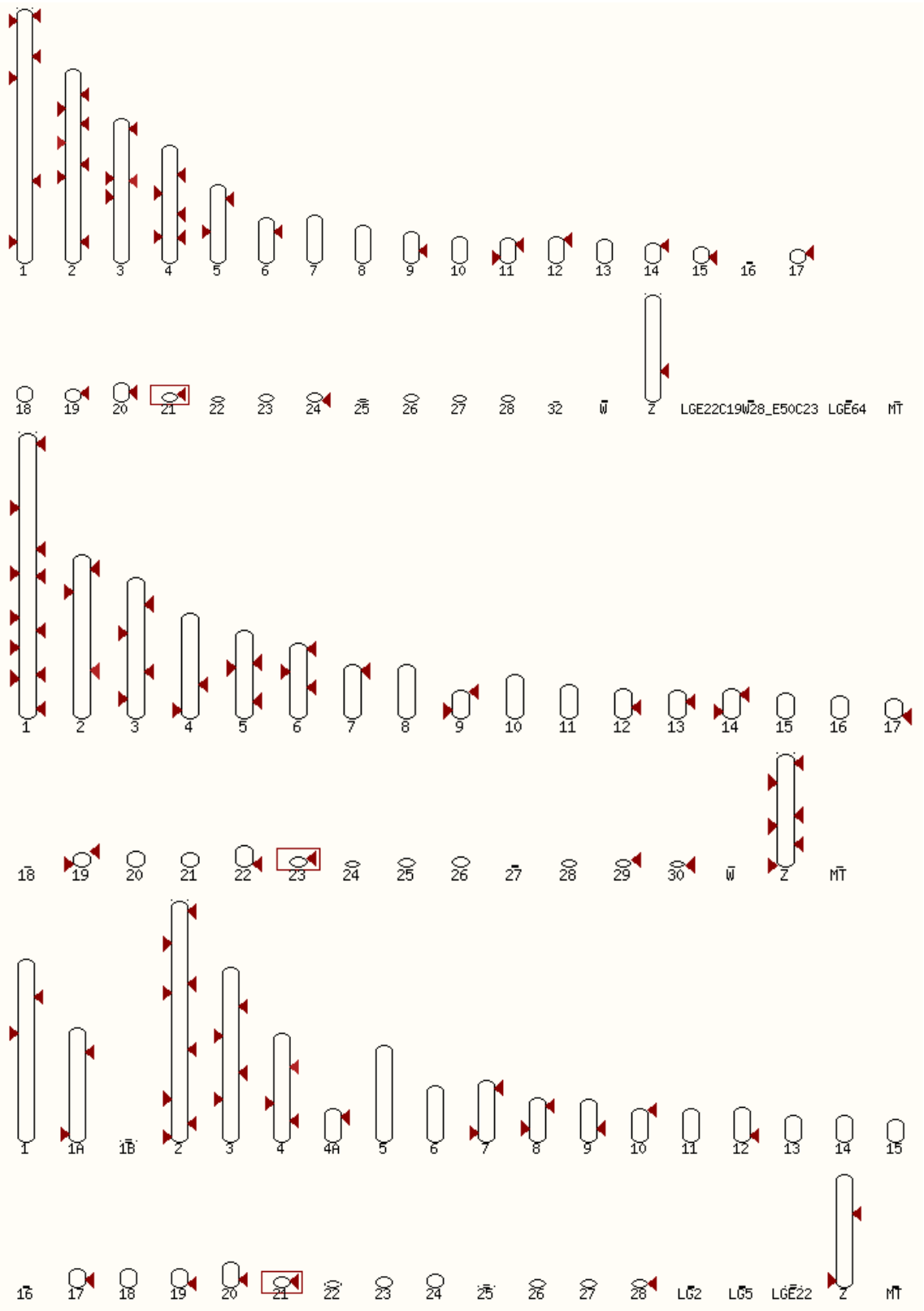


Figure 7: Karyotype of chicken (top), turkey (middle) and zebra finch (bottom). The arrows indicate zones of homology for the duck MMEL1 sequence, and the red rectangle shows the region of maximum homology, corresponding to the location of the MMEL1 gene for each species.

Conclusions

During the present project, the different techniques and tools learned during the module were applied to the resolution of a real practical exercise, which started from raw data from a microarray and a RNA-Sequencing, and step by step, different procedures were performed to extract relevant information (most differentially expressed genes), use proper search tools, perform alignments, building trees, analyze genomic data and discuss the results.

Probably, the most important conclusion that can be obtained from the present work is that the methodologies used are the most important thing to be considered when working, not only in bioinformatics, but in any science field. Different methodologies can retrieve completely different results, and using the wrong methods can lie in trusting results that are not a good representation of the reality, and as science consists in describing natural phenomenon and explain why they happen in the most accurate way possible, it crucial to select the right methodology in any case. That is also the reason why the modeling and testing is so important. A phylogenetic tree can look very nice, but if after a bootstrapping process all the nodes have a very low value, how nice the tree looks is absolutely insignificant.

Another important conclusion that can be obtained from this work is that the understandings of the processes that are being performing are also crucial to have good results. If you understand what are you doing in each step and why are you doing it, you would be able to change some details from the process and adjust the process to your data, as well as solve the different problems that can occur during the process.

More related with the data analyzed, the relationship among some of the species was clearly identified and check by comparing different trees performed with different methods, but there also some significant differences between some of the trees that might be studied more deeply.

About the most differentially expressed genes, one of the most important results in the work, as it was what is important to identify what genes were being more affected by the roscovitine treatment, unluckily the list of differentially expressed genes of microarray and RNA-Sequencing were completely different. It is known that MMEL1 was also obtained by a classmate as one of the most differentially expressed genes, but until the presentation, were all the results of all the classmates will be discussed, it will not be know how different the lists of genes are.

Nevertheless, some of the genes found are involved in the cell cycle regulation and embryonic development, and therefore could be involved in the somatogenesis process and be considered in the future research in this field.

Problems and future objectives

During the project lots of problems appeared, most of them could be solved, such as problems fitting and building trees, retrieving the differentially expressed genes, etc. But some of them could not been solved, and here they will be explained and discussed.

The main problem was the homologous search and alignment. The first idea was working with the gene sequence (DNA), but, as it was said, was completely impossible to find the specific sequence for each gene that did not contain non-sequenced fragments. The only sequences available were the whole loci sequences in NCBI, which, as said, involved working with DNA fragments of 100.000bp. There must be a way to obtain these sequences from some database, and that is one of the future objectives for future works: improve the searching skills in order to be able to obtain the desired sequences in each situation.

The second main problem was building one of the trees. For some reason, when the distance between the sequences in the alignment was performed, some of them were null values. As the alignment was checked and looked nice, there is no reason found that explains why this happened.

The third problem was plotting the fitted tree with the values from the bootstrapping. It was the problem that took more time to be tried to solve, but it was not possible, in all the cases the resulting trees were unrooted, overlapped and impossible to interpret. Nevertheless, the bootstrapping values could be checked, and in all the cases, from 100 bootstrappings, more than 60 were supporting all the nodes, in most of them being values over 90, including a lot of 100. Even though the result could not be “visualized” in a tree, the process was performed and revised, and worked as expected.

Finally, as an auto-critique, the genomic comparison performed, looking the gene localization in the chicken genome and the orthologous regions in other species genomes was not as extended and detailed as it was first thought. For future projects the time for each section should be more equally distributed in order to perform better analysis in all the sections, and not to focus too much in just one of them.

As a future approach, it would be interesting to perform a phylogenetic analysis using the protein sequences for all the genes used in the phylogenetic analysis in this project, and compare the trees and relationships between the species for both DNA and proteins, as gene sequences contains lots of intronic sequences that did not affect the final amino acidic sequence, and therefore two genes may look very different when comparing their DNA sequences, but be identical or very similar in their function, as the protein sequence is almost the same; or just the opposite, two genes may look very similar, but the nucleotide substitution may be affecting key points of the protein, changing completely its function.

Data repository

All the scripts, matrix, trees, plots, figures and other relevant information generated and related with this project can be found at the following github repository:

https://github.com/mlmartin/Project_assesment.git

References

1. **Schofield, Pietà.** [Online] http://www.compbio.dundee.ac.uk/user/pschofield/Teaching/Bioinformatics/Slides/Assignment.pptx.

2. **Allaire, JJ., Kawak, Tareef and al, et.** http://www.rstudio.com/. *RStudio.* [Online]

3. **Gentleman, Robert and Ihaka, Ross.** http://www.r-project.org/. *R-Project.* [Online]

4. **Institute, European Bioinformatics.** http://www.ebi.ac.uk/Tools/msa/. *EMBL-EBI Multiple Sequence Alignment.* [Online]

5. **Dublin, UCD.** http://www.clustal.org/. *Clustal.* [Online]

6. **drive5.** http://www.drive5.com/muscle/. *MUSCLE.* [Online]

7. http://www.bioconductor.org/. *Bioconductor.* [Online]