Submitted 28th March 2014

Gallus Gallus gene expression analysis

BS32010 project report

Fiona Macfarlane
110010712

Summary

Gene expression analysis and comparative genomic techniques were used along with multiple software programmes to investigate the gene expression levels of the chicken, *Gallus Gallus.* The data used was taken from a wet lab experiment investigating the effects of the drug, Roscovitine on chicken embryo cells.[2] The results from this experiment were in the form of a microarray and a RNA sequence count data, which were then used to identify the most significantly differentially expressed genes. Further comparisons were then made between the chicken and other eukaryotic species, using phylogenetic and comparative genomic analysis. All files mentioned can be found within the following Github repository:

https://github.com/fmacfarlane/Project.git [35]

Introduction

Roscovitine is a cyclin-dependant kinase (CDK) inhibitor, which represses the enzyme targets within the cell. Suppression of the enzyme targets alters the speed of the growth phase of the cell cycle [1] .This can delay or stunt the growth and development of the chicken embryos. The data analysed comes from a wet lab experiment which compares chicken embryonic cells treated with Roscovitine and cells which remained untreated, the experiment was replicated four times.[2] When working with high density arrays it is useful to learn how genes differ in expression in response to a stimulus or treatment. The genes that are most effected by the drug are the most significantly differentially expressed genes. There are several gene expression analysis techniques that can be used to find the most significant genes from the data.

To identify the most significant genes, R-studio packages can be used to analyse data from the microarray and the RNA sequence count information. R-studio is a graphical environment, released in 2012, which uses the programming language of R [3] .The development of R, by R.Ihaha and R.Gentleman in 1993, was influenced by the programming language S, created by J.Chambers. R is used in many aspects of statistical computing and contains many packages that can be implemented in the field of Bioinformatics, including bioconductor, limma and DESeq2. [34]

Limma Analysis

Bioconductor is a package within R-Studio that was designed to be used for the analysis of genomic data.[4] Bioconductor contains the sub-package affy which can be used to normalise the data from the sampled used. [5]Affy allows the user to normalise data from a microarray, with a choice of normalisation methods. Normalisation of the array data removes the variations caused by treatment effects, signal, and non-treatment effects, noise. Without normalisation the analysis of the array data can give misleading results, however normalising data can be a risk as important data may be recognised as an outlier and smoothed out, also giving false results.

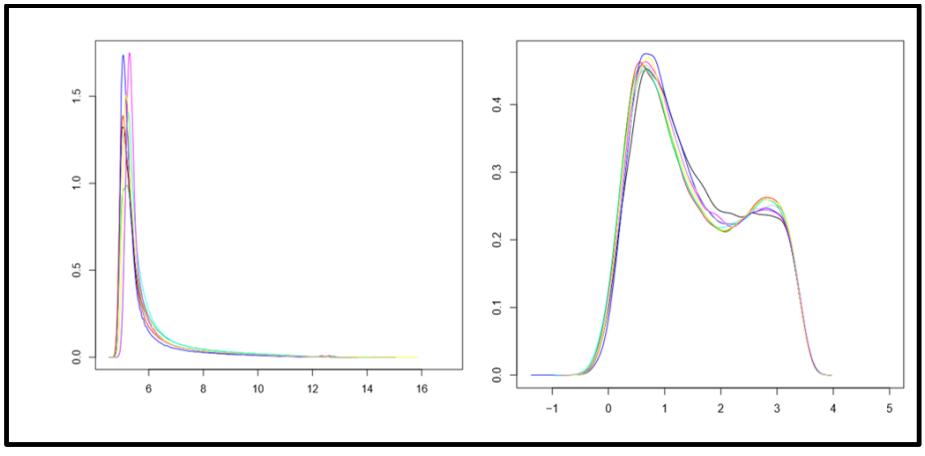
RMA (Robust Multi-array Average) normalisation, described by R.Irizarry, allows for background correction, across array- normalisation and log transformed perfect match values to be implemented.[6] RMA normalises the data at the probe level, and using a model for the distribution of probe intensities, corrects the intensities of the probes. RMA then uses a median polish to account for outlying probe data. The median polish estimates the data, then this estimate is used as the log scale measure of expression. This estimation of the data, is the normalised sample probe data. The RMA normalisation model does not account for the mismatch values, however it was shown, by R.Irizarry, that gene expression is best measured using the log transformed perfect match values. Once the data has been normalised, limma analysis can be used to find the most significantly differentially expressed genes.

The package limma is used in R studio to identify the most significant differentially expressed genes. [7] Developed by G.K.Smyth in 2003, the programme uses the existing models for the expression for the posterior odds of differential expression, in a replicate two colour experiment.[7,8] The posterior odds expression, developed by Lonnstedt and Speed, was found to be useful in ranking the genes in terms of evidence for differential expression. Limma uses a linear model for microarray data, which can be used with both single channel and two colour microarrays, in terms of this project, a single channel microarray was analysed. The incorporation of empirical Bayes methods in the limma method can prevent the reduction in power available to find changes in expression, that can be caused by normalisation. The empirical Bayes models are a method of inferring statistical significance where prior distributions are estimated from the data used.[9] The samples from the microarray are usually pre-processed, by applying normalisation, this could be RMA normalisation. Once limma has been implemented, several statistical results can be viewed. The “AveExp” value is the average log2 expression level for the gene across all of the arrays in the experiment. The “F” value, is an overall test of significance for the gene, which combines the results of the t-test for the contrast. The p-value is the associated value of the p-test, which can be used a test of statistical significance. The adjusted p-value is the p-value adjusted for multiple testing. A low adjusted p-value results in a high likelihood of significant evidence for differential gene expression of the particular gene, therefore the genes with the lowest adjusted P -values are the most significantly differentially expressed.[10]

Implementation

The R script, “Limma.R,” was applied using the .CEL files that contain the experimental data.[35] The script was based on that used in the differential gene expression analysis workshop.[2,36] RMA normalisation was used to pre-process the data before limma was implemented. The distributions of intensities can be visualized before and after normalisation. The plot on the right is the raw density and the plot on the left is the density after RMA normalisation.

Plot 1: Distribution of Intensities



The plots indicate that the normalisation, of the sample data, has smoothed the signal intensities, allowing for a better estimation of the data. The raw plot shows very high levels of variation between intensities, once normalised the intensities are more similar. The statistical results of running limma were then sorted in accordance to adjusted p-value, and shown in the following table.

Table 1 : Sorted Limma Results

	row.names	probeId	ensemblId	geneSymbol	minus	plus	AveExpr	F	P.Value	adj.P.Val
1	618	Gga.10964.1.S1_s_at	ENSGALG00000003265	PTP4A2	11.45889	11.37089	11.41489	6282.495	3.280226e-17	1.408347e-14
2	731	Gga.1118.1.S1_at	NA	NA	11.51057	11.92599	11.71828	5589.447	6.142891e-17	1.408347e-14
3	1061	Gga.11679.1.S1_at	ENSGALG00000007315	PSMD6	10.65748	10.43804	10.54776	5302.840	8.148253e-17	1.408347e-14
4	1176	Gga.11816.1.S1_a_at	NA	NA	10.58822	10.81992	10.70407	5280.109	8.338298e-17	1.408347e-14
5	1313	Gga.11993.1.S1_s_at	ENSGALG00000010539	RNF11	10.74116	10.76658	10.75387	5615.531	5.991292e-17	1.408347e-14
6	1447	Gga.12163.1.S1_at	ENSGALG00000016461	DDX1	10.08235	10.12623	10.10429	5330.819	7.921345e-17	1.408347e-14

The results indicate that the chicken PTP4A2 gene is the most significantly differentially expressed gene, as the gene has a low adjusted p-value. The PTP4A2 gene encodes for the Protein Tyrosine Phosphate Type IVA2 enzyme, which is a vital enzyme within the cell. PTP4A2 is a member of the PTP (Protein Tyrosine Phosphate) family, which consists of cell signalling molecules that have regulatory roles in multiple processes within the cell.[11]

DESeq2 Analysis

There are alternate techniques to find the most significantly differentially expressed genes. DESeq2 is another package used in R which allows the analysis of differential gene expression based on the negative binomial distribution. [12] By basing the method on a negative binomial distribution the model gives a more likely estimation of the data. [13] The algorithms within the program use a generalised linear model to perform statistical assumptions on variations in gene expression. Count outliers are removed to allow the analysis to take place. The DESeq2 method estimates variance mean dependence in the count data and tests for differential expression. The Wald test and Benjamini-Hochberg model are both involved in the DESeq2 algorithms. The Wald test is a parametric statistical test which tests the true value of a parameter based on the sample estimate.[14] The Benjamini- Hochberg (BH) procedure is a method of controlling the false discovery rate by reducing the percentage of false results, allowing a more accurate estimation of the data.[15,16] Applying the DESeq2 method to the data gives various statistical results. The “BaseMean” value is the base mean over all samples. The “log2foldChange” valued gives the log 2 fold change in expression for the treated samples versus the untreated samples. “lfcSe” gives the standard error of the log fold change for the treated versus untreated samples. “Stat” is the Wald statistic, and p-values are the Wald test p values. The values of “padj” are the BH adjusted p-values. Similarly to limma analysis, the adjusted p-values can be compared to find the most significant genes.

The R package, bioMart which is a sub-package of bioconductor can be used along with DESeq2. BiomaRt, developed by S.Durinick, can be used to retrieve genomic data from various genomic browsers, such as ENSEMBL. [17,18]

Implementation

The R script, “DEseqscript.R”, analyses the count sequence data from the RNAseqcounts.txt file, using DESeq2.[20,35] The script was based on an example given during the course of the module.[36] BiomaRt was used to find the relevant information for the significant genes, such as the ensemble gene ids and the external gene-ids. The results were ordered according to adjusted p value (“Padj”) and displayed as the following table. [17]

Table 2 : Sorted DESeq2 results

	ensembl_gene_id	external_gene_id	affy_chicken	baseMean	log2FoldChange	lfcSE	stat	pvalue	Padj
1	ENSGALG00000001115	MMEL1	GgaAffx.751.1.S1_s_at	40.16013	-2.439221	0.2833033	-8.609928	7.310616e-18	8.239795e-14
2	ENSGALG00000004907	KCTD15	Gga.10500.1.S1_at	332.53500	-1.213644	0.1574144	-7.709868	1.259475e-14	7.097774e-11
3	ENSGALG00000012171	GPR39	GgaAffx.7719.1.S1_at	186.46268	-1.183384	0.1633267	-7.245502	4.308417e-13	1.618672e-09
4	ENSGALG00000003518	AGPAT6	GgaAffx.20364.1.S1_at	785.62766	-1.152719	0.1641260	-7.023376	2.165710e-12	6.102428e-09
5	ENSGALG00000003518	AGPAT6	GgaAffx.2193.2.S1_s_at	785.62766	-1.152719	0.1641260	-7.023376	2.165710e-12	6.102428e-09
6	ENSGALG00000003518	AGPAT6	GgaAffx.20364.1.S1_s_at	785.62766	-1.152719	0.1641260	-7.023376	2.165710e-12	6.102428e-09

By comparing the adjusted p-values it can be argued that the MMEL1 gene was the most significantly differentially expressed gene using the RNA sequence count data .The MMEL1 gene encodes for the protein Membrane Metallo Endopeptidase-like 1. MMEL1 is a member of the neutral endopeptidase (NEP) family, which are important in pressure regulation homeostasis and the perception of pain. [19]

Comparison of Limma and DESeq2 Results

R packages, Limma and DESeq2 give very different results, this is due to the different approaches used to analyse the data provided. Both methods include background correction and normalisation, but use different models to achieve this. In both methods the

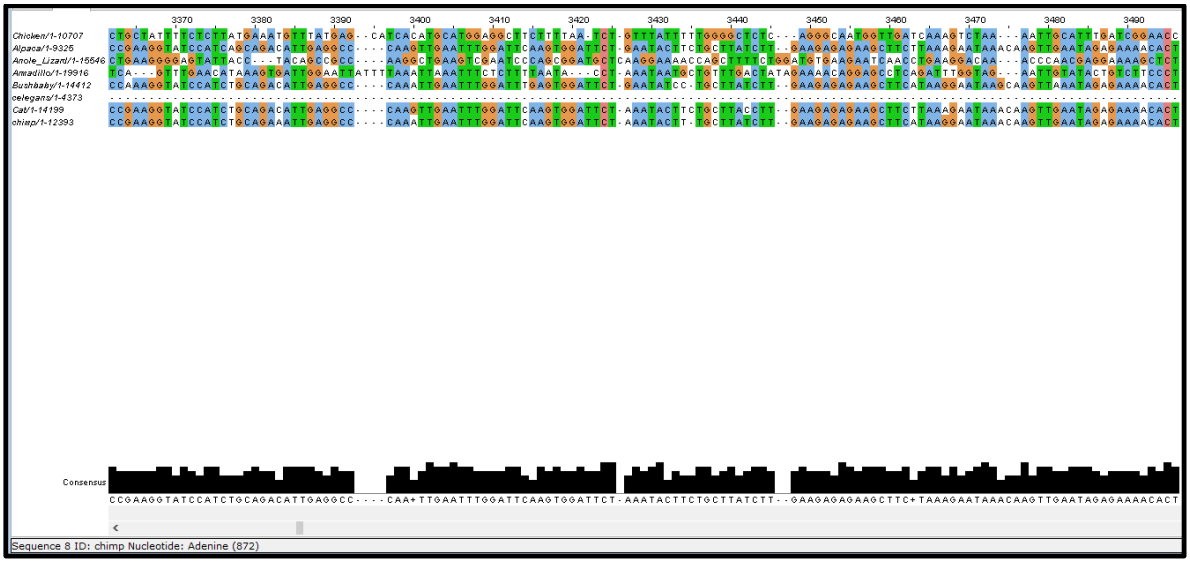
data is smoothed to account for outliers, therefore the resulting data will be different between the methods. The format of the data used may also account for variation between the results, as limma uses microarray data while DESeq2 uses RNA sequencing count data.

Further genomic analysis can be undertaken using the most significant genes. The Chicken PTP4A2 gene was used with comparative genomic and phylogenetic techniques to investigate the chicken's relationship with other organisms, in relation to the gene sequence similarity.

Finding the homologues

Homologous genes are genes with a shared common ancestor in terms of evolutionary development. The genome web browser ENSEMBL was used to find the homologous genes of the chicken PTP4A2 gene. ENSEMBL has the option to view orthologues, genes separated by a speciation event, from the gene that is being viewed, this gave multiple options for the chicken PTP4A2 gene.[18] Seven homologues were selected randomly and their PTP4A2 gene sequences downloaded in fasta format.[35] The eight PTP4A2 gene sequences were then aligned using Jalview's ClustalWS aligning tool, ClustalW is one of the more general and widely used alignment tools. [21,22] This alignment was saved as the file, "homologuesclustalalign.fna.mfa".[35] The species identifiers were altered, to the species names, to allow the information to be easily visualised.

Image 1: Alignment of homologues



The alignment shows many regions of similarity between the PTP4A2 genes, however it is easier to visualise the alignment using phylogenetic trees.

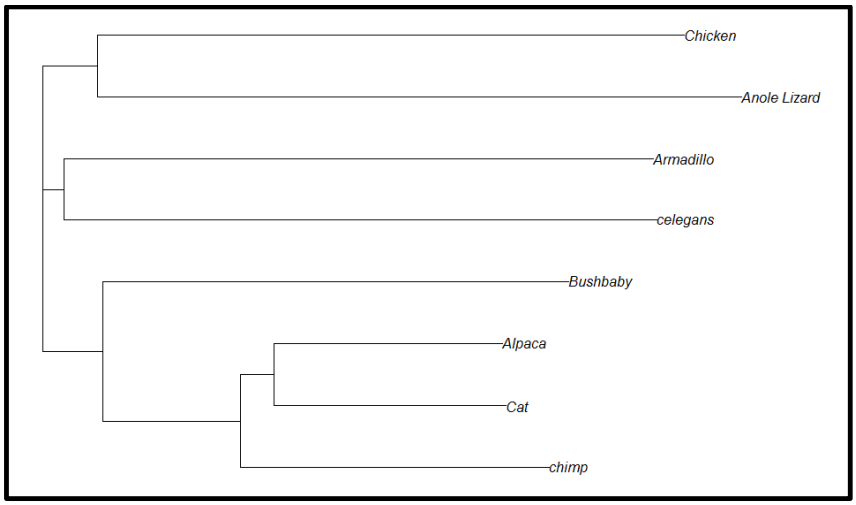
Phylogenetic tree analysis

The homologue alignment that was created was then used to investigate the phylogenetic relationships between the homologous genes. The R script, "plot_trees.R" uses the R packages; ape, phangorn and picante to created phylogenetic trees of the alignment and find the evolutionary distinctiveness scores.[35] Ape, developed by E.Pardis in 2004, provides multiple phylogenetic tree drawing and reading methods.[23] The tree construction within ape is based on the user specified tree construction models, an example of these models is BIONJ. The construction model, BIONJ was based on another tree construction method, neighbour-join (NJ), both models minimises the variance between distances within the phylogeny to construct a tree.[26] BIONJ was selected to be used for tree construction,

due to the residuals calculated using this method being low. Phangorn, can be used to create trees using distance methods.[24]

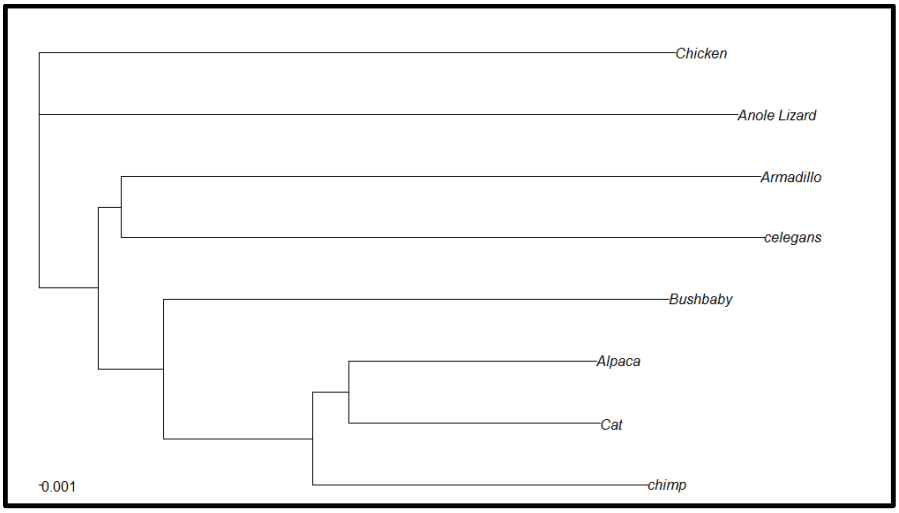
The evolutionary distinctiveness (ED) score can also be a good indicator of similarity between sequences. The ED score indicates how distinct a species is from the others in the phylogeny. Some species are more distinct than others, these distinct species generally represent a large amount of evolution that is unique to them.[27] ED scores can be calculated using the package, picante.[25] The first tree produced was an un-rooted tree created using the distances and BIONJ method of tree construction.

Plot 2:BIONJ style un-rooted tree



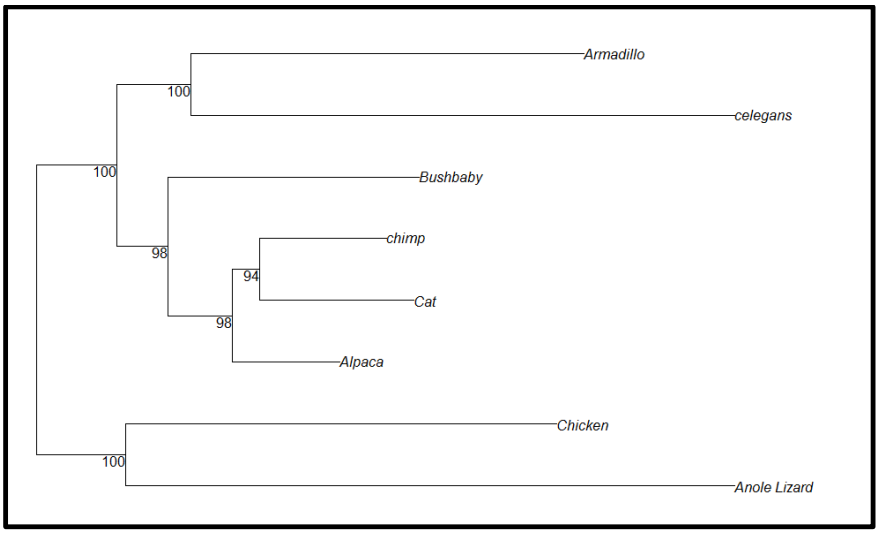
The same tree was then rooted at the chicken PTP4A2 sequence.

Plot 3: BIONJ style rooted tree



Bootstrapping the data allows a measure of confidence in the phylogeny to be calculated:

Plot 4: Bootstrapped tree

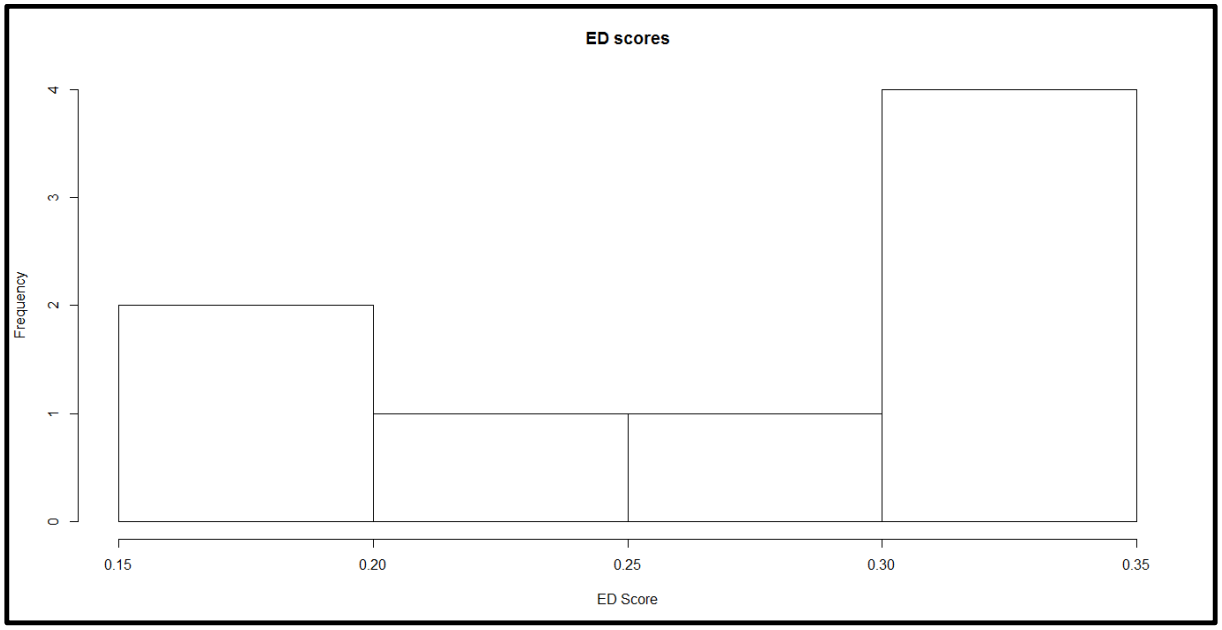


All three trees agree in that the Chicken PTP4A2 gene is most similar to the Anole Lizard PTP4A2 gene. The bootstrapped tree also gives a confidence value of 100 for the clade, containing the lizard and the chicken, this means that all 100 of the bootstrapped trees created support this inferred relationship. The tree also shows a relationship between the Armadillo and the *C.elegans* worm, which is a surprising clade, however the relationship between the chicken and lizard will be the focus of further analysis. The ED scores were also calculated for the sequence alignment.

Table 3: ED Scores

	Species	W
1	Chicken	0.3108858
2	Alpaca	0.1670120
3	Anole_Liz-ard	0.3416262
4	Armadillo	0.3229450
5	Bushbaby	0.2597047
6	celegans	0.3249466
7	Cat	0.1688063
8	chimp	0.2007837

Plot 5: ED score frequencies



The taxa in the phylogeny have low ED scores, this suggests relationships between the taxa as there are no unique organisms. The Anole lizard genome region and the chicken genome region can be compared further using comparative genomic techniques.

Comparative Genomics

Features of genomes or genomic regions can be compared using multiple comparative genomic techniques. Mummer, ACT and BLASTN are techniques which can be used on Xming to determine the similarity of two genomes, or genome regions. [32]

Mummer is system, developed in 1999, that can be used to quickly align genomes or genome regions. The algorithm used in the system is based on building and searching suffix trees.[28] Suffix trees allow for a quick implementation of many strings, in this case long sequences, and can be used to find patterns within the sequences.[33] The Mummer matching algorithm builds and searches a data structure containing suffix trees.

BLAST is a basic local alignment search tool that was developed in the 1990s. BLAST approximates sequence alignments, so that the local similarity measures of the sequences are optimum. BLAST can be used for multiple genome comparisons, and there are many sub-programs of the BLAST tool.[30] BLASTN is a form of BLAST that compares the nucleotides of sequences, in relation to their similarity. [31]The results from running BLASTN can be visualised using ACT.

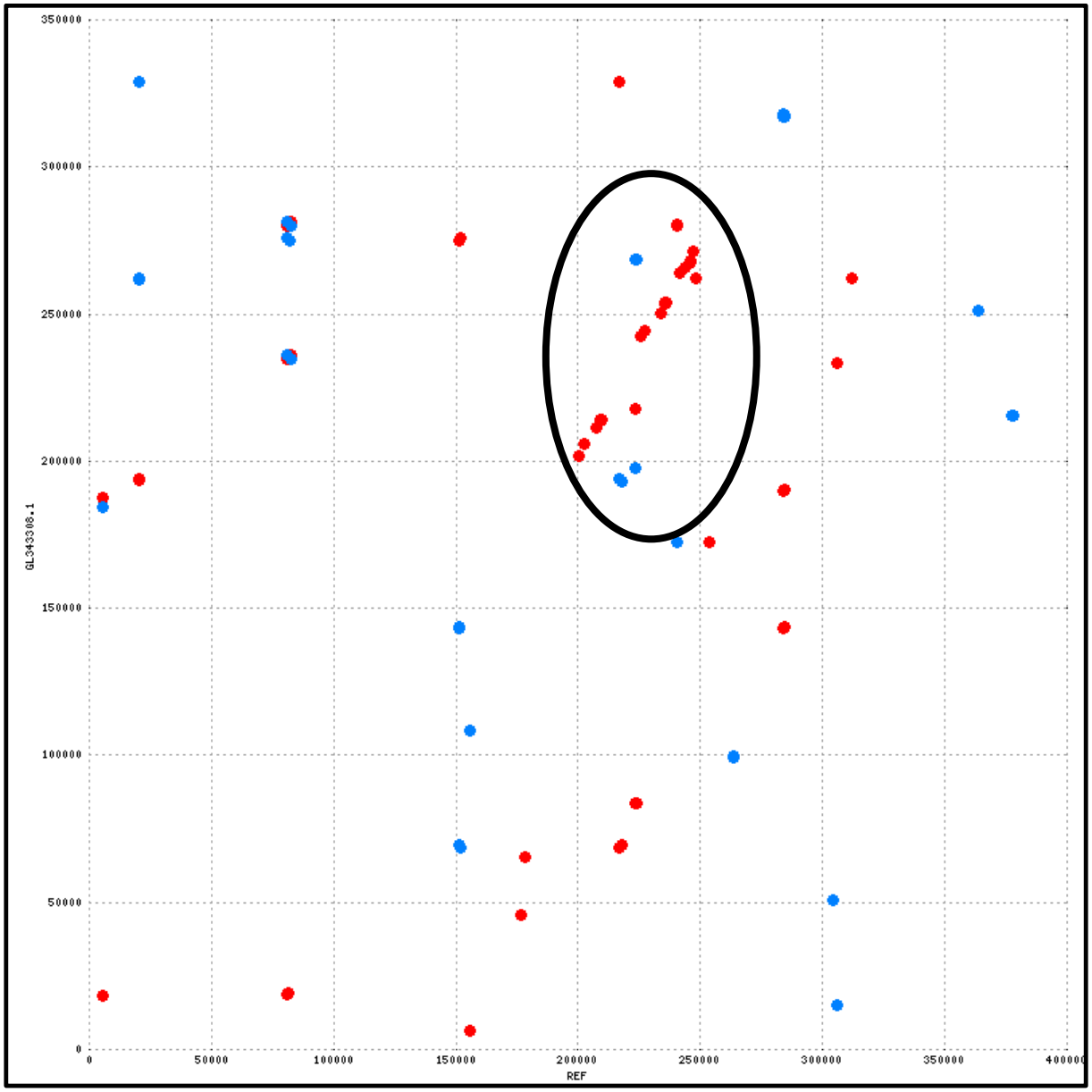
ACT (Artemis comparison tool) is a tool for displaying pairwise comparisons between two sequences. The comparison file is usually taken from a BLASTN or Mummer file. ACT was developed in 2005 by T.Carver who used the Artemis code (Rutherford et al. 2002) to create ACT. The two sequences from the input are aligned with the subject sequence above the reference sequence. Coloured bands join the sequences indicating matching regions. Red lines represent matching regions on the forward strands and blue represents matching regions on the reverse strand. The intensity of the line in proportional to the percentage identity between the two sequences. [29]

Implementation

Using ENSEMBL the genomic regions containing the PTP4A2 gene of both the Chicken and Anole Lizard were downloaded as fasta files, with around 10000 bases either side of the gene.[35] The larger genome region was chosen to give a more general comparison between the two genomes. The whole chromosome could not be compared for the species as the sequences were too long to be run with BLASTN.[18]

Mummer was implemented using the two genomic regions, commands used can be viewed in the appendix of this report. The plot created by Mummer gives a visual representation of the analysis.[28]

Plot 6: Mummer Plot

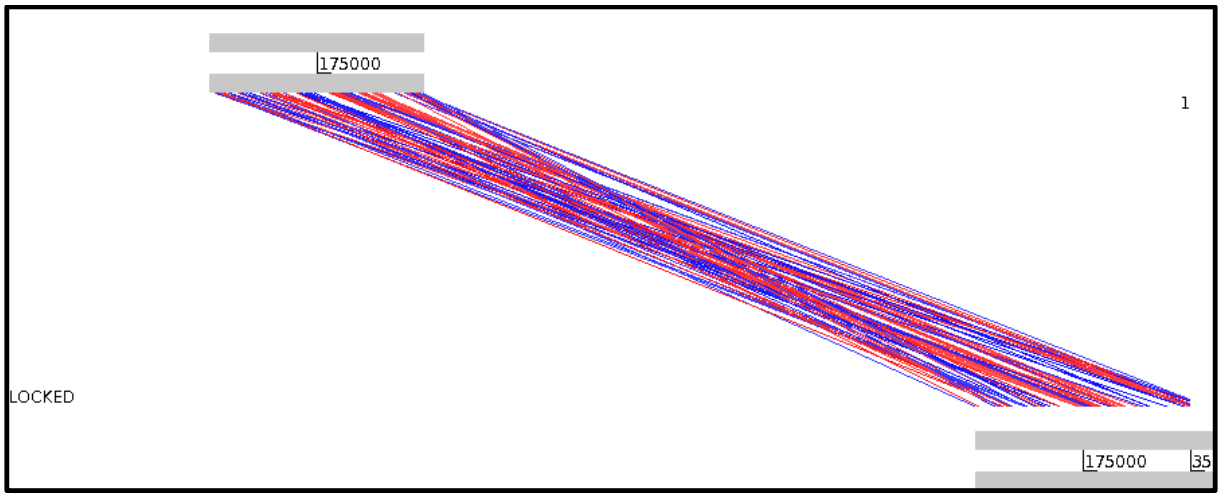


In the above plot, “REF” is the chicken genome region and the lizard genome is along the y-axis. Red dots are placed where there is a match in the genome sequences in the forward strand, and blue dots represent matches in the reverse strand. If the sequences were identical there would be two straight lines on the y=x and y=-x diagonals.[28] The

mummer plot displays a region of similarity between the two genomes in the forward direction, as circled. This could be where the PTP4A2 gene is located. In this similar region the sequences are almost identical, as the dots are almost on the y=x diagonal. There are other points of similarity between the two sequences on both the forward and reverse strand, but the sections are short and could be coincidental.

ACT was also used to display similarities between the sequences. BLASTN was run on the data first of all and then ACT was used to visualise the results.[31,29] Commands used to do this can be found in the appendix. The following image was obtained from ACT,

Image 2: ACT Visualisation



The top line represents the Anole lizard genome region and the bottom line represents the chicken genome region. Red lines represent sequence matches in the forward strand, whereas blue lines represent sequence matches in the reverse strand. The ACT display suggests that there are many similarities between the two species genome regions, however they are not identical, as was expected. There are many regions of synteny on both the forward and reverse strands.

Discussion

Both the phylogenetic tree analysis and comparative genomics techniques suggest that the chicken and Anole lizard PTP4A2 genes are very similar in some regions. The phylogenetic trees support the existence of an evolutionary relationship, which could be a common ancestor between the Anole lizard and the chicken. This is as expected as the chicken and lizard PTP4A2 genes are orthologous. The analysis indicates that the similarity between the two genes is too high to be a result of coincidence, but is an inference of a relationship. However to prove this further experimentation would have to be undertaken, and the two species' genetic evolution compared more closely.

The analysis between the chicken and Anole lizard PTP4A2 genes could easily be replicated using alternate species or alternate genes, to give a more general overview of the chicken's relationships with other species. This analysis could also be repeated to find the relationships in relation to other significantly differentially expressed genes, such as MMEL1, which was found to be the most significantly gene using the DESeq2 package. The analysis could also be restricted to finding the similarities between the chicken and other bird species, such as the duck, where we would expect higher levels of synteny, than those found between the chicken and the Anole lizard genes.

A different normalisation model could be implemented prior to applying limma, to observe the differences in the genes found to be most significant. Overall the methods shown here support the argument that bioinformatics can be messy, as the DESeq2 and limma methods gave different results due to the variance between them. However the comparisons between the genes can still be seen as good indicators of similarity, and infer evolutionary relationships between the species studied. The project has highlighted the advantages of using programming software in comparative genomics and the analysis of gene expression. Most packages available, can significantly reduce the time taken to analyse data from microarray experiments, allowing significant biological discoveries to be made more frequently. Most of the packages developed are based on prior models and techniques that have been used for many years, this suggests that many models are still relevant and acceptable to use today.

The development of these comparison and analysis techniques in the future will increase the efficiency to which gene expression levels can be calculated and phylogenetic relationships investigated, This will allow for a greater depth of evolutionary investigation, and a greater insight into how organisms are related at a molecular level. However the results of comparative genomics and phylogenetic analysis can only be used as an indicator of biological significance, and further experimentation would be needed to prove the inferred relationships.

References

1. De Azevedo WF, Leclerc S, Meijer L, Havlicek L, Strnad M, Kim SH (1997). "Inhibition of cyclin-dependent kinases by purine analogues: crystal structure of human cdk2 complexed with roscovitine". *Eur J Biochem* 243 (1-2): 518–526.
2. Martin, DMA. *BS32010 teaching server. http://ts-ug-dev.lifesci.dundee.ac.uk/BS32010/expression/data/GCOS_Cell (accessed 20th March 2014).*
3. RStudio (2012). RStudio: Integrated development environment for R (Version 0.96.122) [Computer software]. Boston, MA. Retrieved May 20, 2012.
4. Gentleman, R.; Carey, V.; Huber, W,; Irizarry, R.; Dudoit, S. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer.
5. Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. 2004. “Affy---analyssis of Affymetrix GeneChip data at the probe level”. *Bioinformatics* 20, 3(Feb. 2004), 307-315.
6. Irizarry, R. A., Hobbs, B., Collin, F., and Spees, T. P., 2003. “Exploration, normalization, and summaries of high density oligonucleotide array probe level data”. *Bioinformatics* 4, 920, 249-264.
7. Smyth, G. K.,2005. “Limma: linear models for microarray data” Springer, New York, 397-420.
8. Lonstedt I, Speed TP: Replicated micro-array data. Stat Sin 2002;12:31–46.
9. C.M. Bishop (2005). *Neural networks for pattern recognition*. Oxford University Press
10. Smyth GK, Ritchie M, Thorne N, Wettenhall J.2007. “limma:Linear Models for Microarray Data User's Guide”. *R User Guide.*

11. Zhao Z, Lee CC, Monckton DG, Yazdani A, Coolbaugh MI, Li X, Bailey J, Shen Y, Caskey CT (Sep 1996). "Characterization and genomic mapping of genes and pseudogenes of a new human protein tyrosine phosphatase". *Genomics* 35 (1): 172–81.
12. Love, M., Anders S., Huber, W.2014. “Differential analysis of count data – the DESeq2 package”.*R-user guide.*
13. DeGroot, Morris H. (1986). *Probability and Statistics* (Second Edition ed.). Addison-Wesley. 258–259.
14. Harrell, Frank E., Jr. (2001). "Section 9.3.1". *Regression modeling strategies*. New York: Springer-Verlag.
15. Benjamini, Yoav; Yekutieli, Daniel (2001). "The control of the false discovery rate in multiple testing under dependency". *Annals of Statistics* 29 (4): 1165–1188.
16. Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society, Series B* 57 (1): 289–300
17. Durinick, S., Moreau, Y., Kasprzyk, A., Davis, S., and De Moor, B.2005. “BioMart and bioconductor: a powerful link between the biological databases and microarray data analysis”. *Bioinformatics* 21, 3439-3440
18. Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, et al. 2014. “Ensembl 2014” *Nucleic Acids Research* 42 Database issue:D749-D755
19. NCBI (2014) *MMEL1 membrane metallo-endopeptidase-like 1 [Homo sapiens (human)],* Available at: *http://www.ncbi.nlm.nih.gov/gene/79258* (Accessed: 25th March 2014).
20. Anders, S.and Huber, W. 2010. “Differential expression analysis for sequence count data2. *Genome Biology* 11, 106
21. Waterhouse, a.M., Procter, J.B., Martin, D,M,A., Clamp, M. and Barton, G.J. (2009).”Jalview Version 2 – a multiple sequence alignment editor and analysis workbench”. *Bioinformatics* 25 (9), 1889-1191.
22. Martin, D.M.A., Procter, J., Waterhous, A., Shehata, S. and Barton, G.J. “Jalview 2.8: A manual and introductory tutorial”. Available at http://www.jalview.org/tutorial/TheJalviewTutorial_screen.pdf, Accesed 25th March 2014.
23. Paradis E., Claude J. & Strimmer K. 2004. “APE: analyses of phylogenetics and evolution in R language”. *Bioinformatics* 20: 289-290.
24. Schliep K.P. 2011. “phangorn: phylogenetic analysis in R”. *Bioinformatics*, 27(4) 592-593
25. S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb. 2010. “Picante: R tools for integrating phylogenies and ecology”. *Bioinformatics* 26:1463-1464.
26. Gascuel O. 1997.“BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data”. *Molecular Biology and Evolution* 14(7): 685-695.

27. The Zoological Society of London (2013) *Evolutionary distinct and globally endangered* , Available at: *http://www.edgeofexistence.org/about/edge_science.php* (Accessed: 25th March 2014).
28. Delcher, A. L., Phillippy, A., Carlton, J. and Salzberf, S. L. 2002."Fast Algorithms for Large-scale Genome Alignment and Comparision." *Nucleic Acids Research* (2002), Vol. 30, No. 11 2478-2483.
29. Carver TJ, Rutherfor KM, Berriman M, Rajandream MA, Barrell BG and Parkhill J. "ACT: the Artemis Comparison Tool". *Bioinformatics (Oxford, England)* 2005:21;16;3422-3423
30. Altschul, S., Gish, W., Webb, M., Myers, E. W. and Lipman, D.J.1990. "Basic Local Alignment Tool". *J.Mol.Biol.* 215(3), 403-410.
31. Madden, T.L., Tatusov, R.L. & Zhang, J. 1996. "Applications of network BLAST server" *Meth. Enzymol*. 266:131-141.
32. Apache web server version 2.1 (http://www.apache.org)
33. Zamir, Oren; Etzioni, Oren (1998), "Web document clustering: a feasibility demonstration", *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA: ACM, 46–54
34. Ihaka, Ross (1998). "*R : Past and Future History"*. Interface98 (Technical report). Statistics Department, The University of Auckland, Auckland, New Zealand.
35. Macfarlane, F.2014. (https://github.com/fmacfarlane/Project.git).
36. Schofield, P (2014) *Bioinformatics,* Available at:*http://www.compbio.dun-dee.ac.uk/user/pschofield/Teaching/Bioinformatics/*(Accessed: March 2014).

Appendix

Other than the commands shown below all commands/scripts used to complete the analysis for this project can be found in the repository:[35]

https://github.com/fmacfarlane/Project.git

Commands used to run mummer are as follows:

```
$ time mummer -mum -b -c chick_region.fna anole_region.fna > wga_output/mummer.mums
```

This sends the output into a file mummer.mums

```
$ head wga_output/mummer.mums
```

This allows visualisation of the results

```
$ wc wga_output/mummer.mums
```

This shows the word count of the output file.

```
$ mummerplot --png --prefix=wga_output/mummer wga_output/mummer.mums
```

This produces a Mummer plot of the relationship between the two sequences.

Commands used to run BLASTN on the data are as follows:

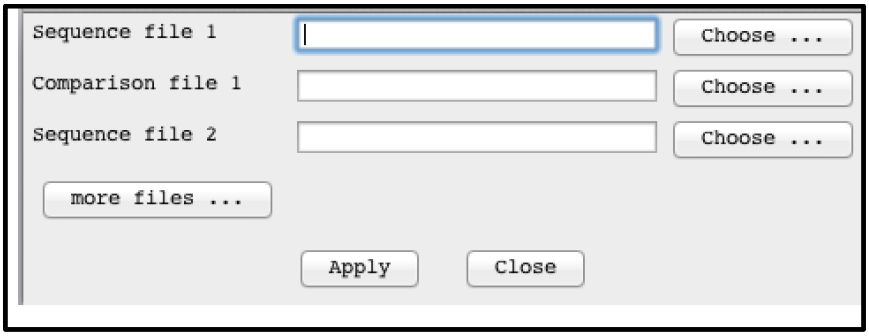
$ time blastn -query chick_region.fna -subject anole_region.fna -outfmt 6 -out wga_output/blastn.tab -task blastn

This runs BLASTN on the two genome region sequences, and prints the resulting table into the output file, “blastn.tab”.

$ act

Opens the ACT interface on Xming.

The ACT interface is shown as,



For the purposes of this project, the “anole_region.fna” was set as sequence file 1, the “chick_region.fna” was set as sequence file 2, and the “blastn.tab” file was set as he comparison file. The resulting image can be viewed in the comparative genomics section of the report.