Analysis of Differential Expression of Fruit Flies Using Voom

Alexander Stocking[1] & *Sivakami Thinnappan*[2]

[1,2] *Department of CS (Computer Science), Purdue University Fort Wayne, IN*

Students: stoca01@pfw.edu, thins02@pfw.edu

ABSTRACT
Differentially expressed genes are the molecular basis of phenotypic variation. DNA microarrays is the most popular approach to understand the abundance of mRNA in respect to genes. The micro arrays technique is an old, traditional way to quantify the genes. The cDNA, also known as RNA-Seq, technique emerged with high throughput. The main goal is to analyze the differentially expressed genes in female fruit flies and the link between those values and their eye sizes (small, medium, or big). Drosophila melanogaster is used as the model for development, mutation in drosophila embryos lead to drosophila embryogenesis and these embryos are developed in 2h intervals [1]. Drosophila genetic reference panel (DGRP) is generated from 205 individual flies which makes dgrp a collection of inbred lines of sample. Differential expressed genes are compared using voom (+limma) technique from the RNA-seq with the DGRP and degeneration model RPR datasets.

KEYWORDS
RNA-Seq; Gene; Differential Expression; Fruit Fly; Voom; Limma; EdgeR

INTRODUCTION

For a time, we had assumed that we could determine the physical traits of an organism through simple analysis and comparison of RNA and protein sequences. This proved difficult though as more research went into how much that DNA is expressed in the RNA that is sent out. Not all genes were equal, so it became increasingly difficult to say the difference in a single sequence is the sole cause of the physical change in the organism. So instead, in this paper we aim to explore the actual expression of those RNA sequences using a method called Voom to find if there is a link between the differential expression of a gene and its effect on eye size.

In this paper, we will be working in R and using the method voom from the package limma. Limma is a package contained in the Bioconductor group of packages. It was originally intended to work with micro arrays, but with the help of EdgeR and a few other tertiary methods, we can use this powerful tool to analyze our expression data. What it does is it transforms count data into log2 counts per million, estimates the mean variance, generates the weights for each gene, and prepares it for linear modeling. We picked it as it did quite well in the tests shown in the Soneson paper. It found a decent number of significant genes without overestimating, had a good false detection rate, and was decently fast [2].

METHODS AND PROCEDURES

Before getting started, we needed to do quite a bit of preprocessing. Both the expression set, and the eye size set have strains that the other does not. So, after filtering and removing those strains, we then needed to get a better grasp of the data. We made histograms, seen in **Figure 1,** of the mean expression values in the expression set and the means of the eye sizes in the RPR set.

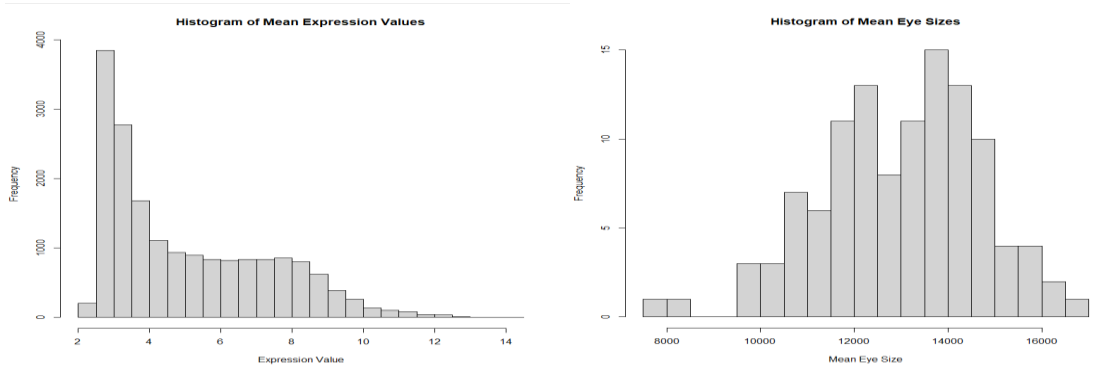

Figure 1: A histogram of the mean expression counts on the left and a histogram of the eye size averages from the cleaned RPR data set on the right.

These visualizations helped us with two things. First was that we could determine a limit for low-expressed genes that will have a negligible effect on the eye size. Since right around 3 seemed to be the most abundant value expressed in this set, we decided to limit any gene that did not have any expression value above 30 when converted to counts per million. We did this by simply grabbing the max expression from every gene and checking them against the cutoff point. We specifically avoided using the mean of the entire range of values because it could remove a differentially expressed gene if multiple low values pulled the mean down below the cutoff point. All the genes had at least some significant expression so none of them needed to be dropped. The second thing the visualizations helped us with was that it gave us an idea of what the binning for the eye sizes might look like when we create the quantiles. We can see though that there is a fairly normal distribution in the data with only a couple of outliers in the small end, so we took that into consideration when making those groupings for the eye sizes.

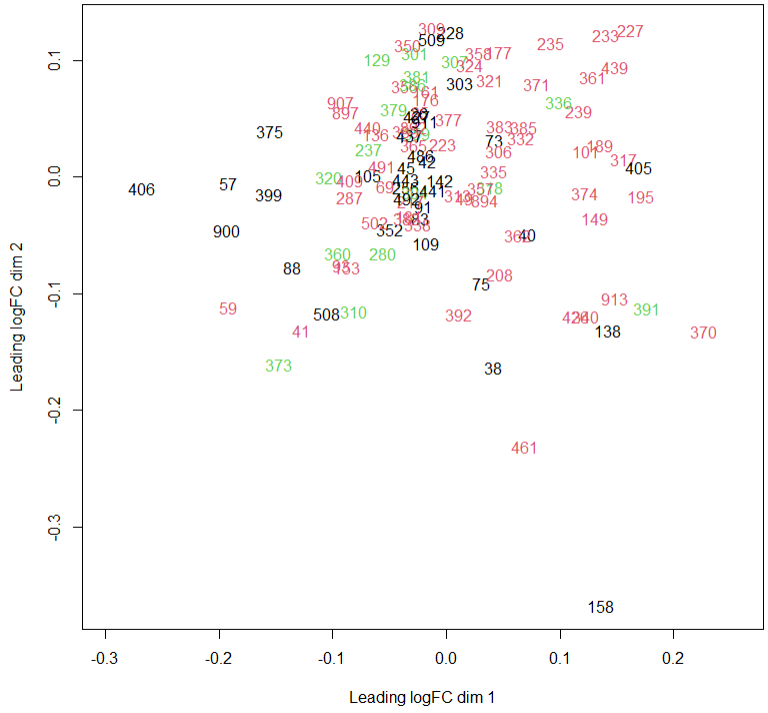

Figure 2: A multidimensional scaling plot that uses the eye sizes to group the strains considered small (black), medium (red), and large (green)

The last visualization made was a multidimensional scaling plot seen in **Figure 2**. This plot uses the eye sizes to group the strains and then plot the variance between them. So, strains that are closer together had similar expression values of their genes, while those far apart are much more different in the expression values of the genes. The small and large groups, black and green respectively, seemed to show a bit more variance than the normal sized group of red. That could be because of the amount of data points for the normal eye size versus the other two groups. So, it was not anything concrete, but it was important enough to keep in mind later down the line.

After preprocessing and visualization, we then utilized the DGEList method from EdgeR to format our expression data into an object that voom can understand. We then needed to do just a few operations onto that object like recalculating the normalization factors as well as the filtering we already discussed. We then created a model matrix using the eye size groups made from the quantiles [3]. Finally, we organized the final product into readable tables so we could easily grab the most expressive genes.

RESULTS

Using voom to create an object of analyzed data also gives us a plot of the mean variance of the data seen in **Figure 3**. While it does not tell us anything major about the results of the data, it does inform us about the noisiness of the data. For example, if we had not filtered our data and it had low-expressed genes, it would affect voom and we would see that represented as a much more chaotic red line. That line is the average expression of the genes that voom then uses to create weights for each gene, so a nice smooth line is exactly what we are looking for [4].

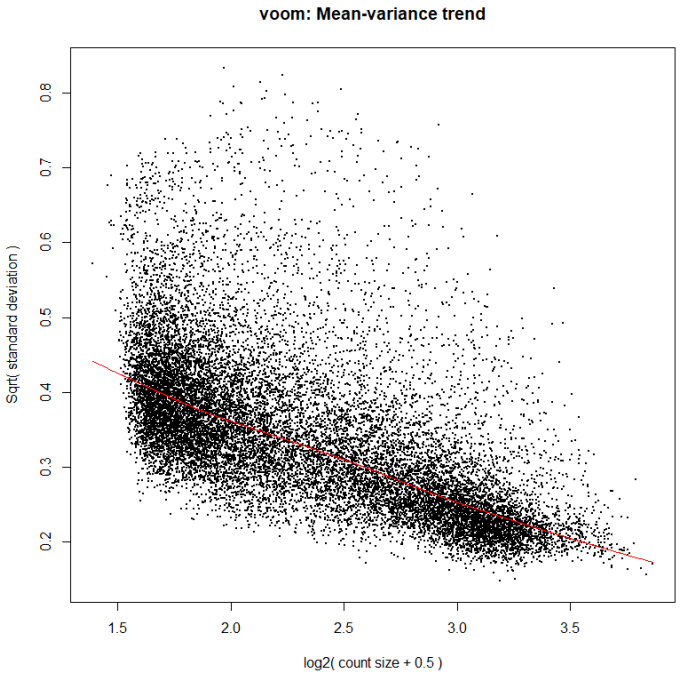

Figure 3: The mean variance plot printed out after applying voom to the expression data. The expression values are changed to counts per million reads based on the normalization factors. The red line is the averaged expression fitted to the standard deviation.

The first table we got was simply fitted using the model matrix we created earlier with an empirical Bayes method applied to it [4]. The initial results, seen in **Figure 4**, were not very promising as it seemed like the expression between groups was not vastly different. We can see the averaged expression values in each group and most of them are only different by a few hundredths. Plotting out the expression values of the top five genes against the eye size of the corresponding strain, seen in **Figure 5**, gave us a better idea of the variation in expression. The first gene, FBgn0038148, showed a lot of promise, but as we continued plotting each gene expression values, the results became less conclusive. Only one other gene, FBgn0038108, had a significant regression line when plotted out like the first gene. Not only that, but the change in expression value is so small in some of these that it was hard to say if that was different enough to even be a substantial change. These were just not good enough to confidently draw any conclusions from. So, we went about making comparisons between the groups and use those comparisons to improve voom's initial findings.

	eyeSizessmall	eyeSizesmedium	eyeSizeslarge	AveExpr	F	P.Value	adj.P.Val
FBgn0038148	7.288254	7.290331	7.300527	7.291657	3301837	1.20E-280	2.19E-276
FBgn0030034	6.639427	6.645575	6.651578	6.64481	3071875	7.30E-279	6.62E-275
FBgn0036373	6.70946	6.706334	6.711069	6.708219	2961397	5.86E-278	3.54E-274
FBgn0261599	7.20533	7.204838	7.211552	7.206481	2697281	1.19E-275	5.38E-272
FBgn0000556	7.265484	7.26813	7.268411	7.267991	2683199	1.60E-275	5.79E-272
FBgn0025286	7.141249	7.134719	7.146392	7.138796	2504483	8.04E-274	2.43E-270
FBgn0004045	7.315057	7.318762	7.319982	7.318111	2429786	4.50E-273	1.16E-269
FBgn0034968	7.141348	7.133166	7.145127	7.137901	2407292	7.63E-273	1.73E-269
FBgn0038108	6.610336	6.616335	6.619915	6.615485	2377025	1.57E-272	3.16E-269
FBgn0001324	6.696815	6.704108	6.707431	6.702318	2346139	3.29E-272	5.97E-269

Figure 4: The table data of the top 10 differentially expressed genes initially found by voom.

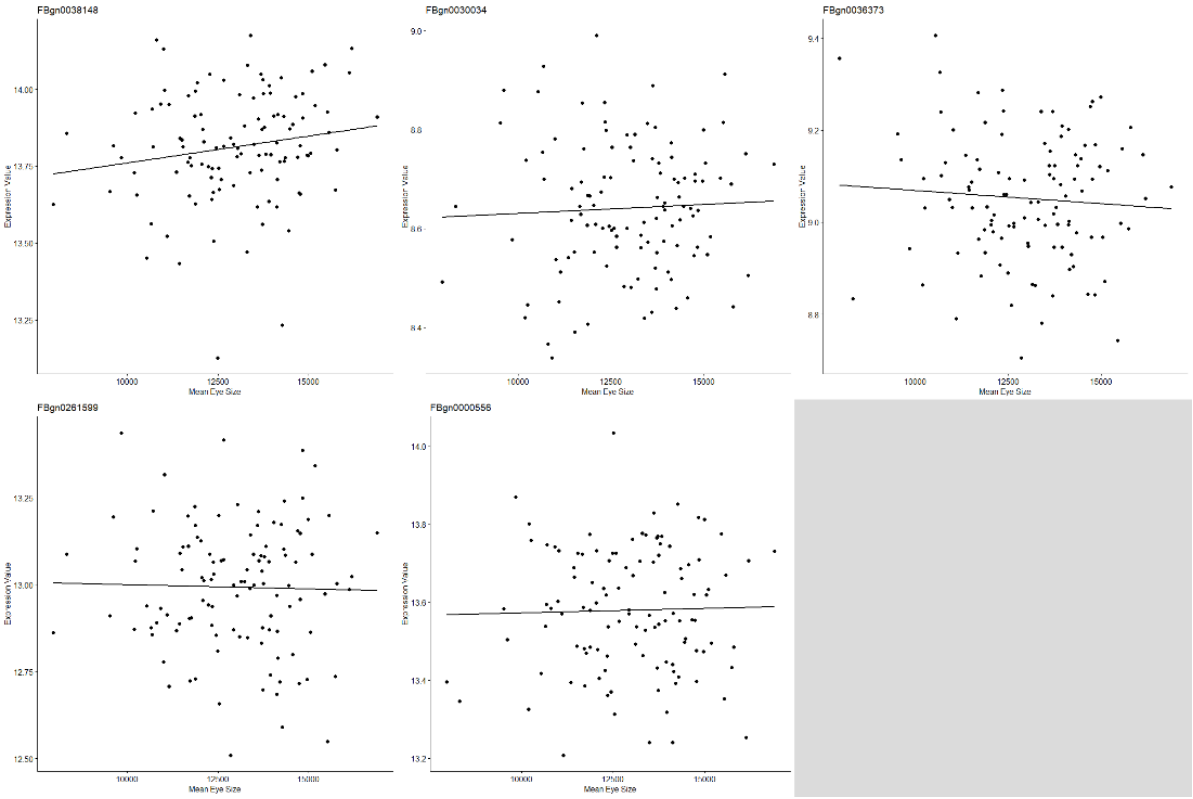

Figure 5: The plot of the top 5 differentially expressed genes initially found by voom.

Creating a simple contrast matrix, comparing each group to the others, and then using that to refit the data gave us much more promising results [4]. Not only do the plots show a clearer line of regression and therefore correlation, but the change in the expression is a lot wider as well. While there are some that we would be hesitant to say have a significant effect on eye size, these results show a lot more promise than the initial findings. The tradeoff though, was that the p-values and adjusted p-values were now much higher. The data looks more conclusive, but it holds less confidence in what it appears to be saying.

	eyeSizeslarge... eyeSizesmediu m	eyeSizesmediu m...eyeSizessm all	eyeSizeslarge... eyeSizessmall	AveExpr	F	P.Value	adj.P.Val
FBgn0035620	0.017449266	-0.17936296	-0.1619137	5.858138	12.357517	1.39E-05	0.2525425
FBgn0027584	-0.004958454	-0.36814858	-0.37310703	5.834361	10.597994	6.02E-05	0.4507004
FBgn0030594	-0.036535755	-0.14746926	-0.18400501	6.157179	10.344829	7.45E-05	0.4507004
FBgn0036993	0.032478218	-0.13520362	-0.1027254	5.959036	9.523969	1.50E-04	0.5488396
FBgn0039342	0.022616888	-0.12652274	-0.10390585	6.786504	9.51339	1.51E-04	0.5488396
FBgn0039085	-0.064496052	-0.12711583	-0.19161188	5.460847	9.033089	2.29E-04	0.5969696
FBgn0035667	0.043672269	-0.09374099	-0.05006872	6.690378	9.024299	2.30E-04	0.5969696
FBgn0054043	-0.061284733	-0.12348902	-0.18477375	5.874565	8.65139	3.18E-04	0.6563564
FBgn0036518	0.001984224	0.04503949	0.04702372	6.579366	8.508924	3.60E-04	0.6563564
FBgn0036659	-0.046976902	-0.07803003	-0.12500693	5.670124	8.380254	4.03E-04	0.6563564

Figure 6: The table data of the top 10 differentially expressed genes after comparisons were made between each group.

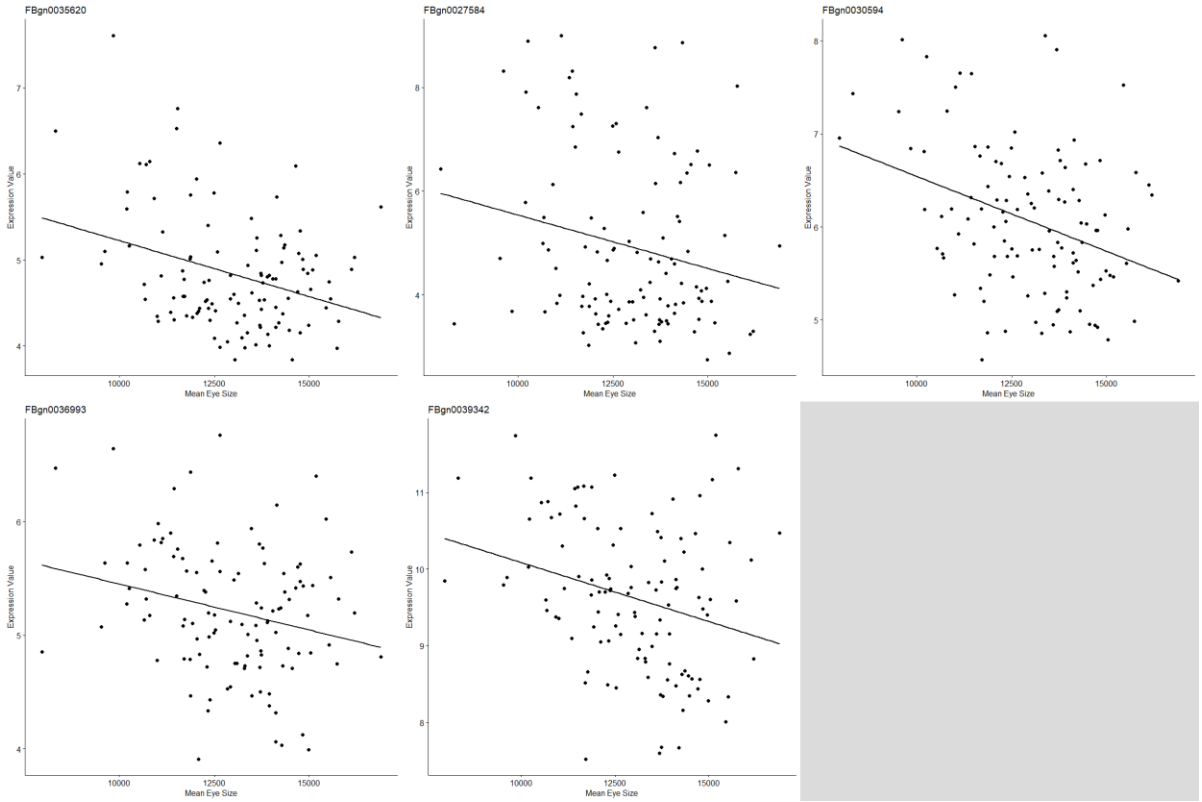

Figure 7: The plots of the top 5 differentially expressed genes found in the new comparison table.

CONCLUSION

Out of all the found genes and other data collected, FBgn0038148, FBgn0035620, and FBgn0030594 are some of the main genes we would recommend further exploring. The top gene found initially has just enough of a visible effect that we feel it deserves some analysis. There is also the other gene, FBgn0038108, that could maybe hold some interesting information. The reason why we singled those two out is because those two were the only ones to have any sort of significant correlation in their scatterplots. Again though, the change in the expression value is not as significant as we would like so these two genes could be dead ends.

Now, the contrast genes all seem to have a decent effect with their expression values, but they do not have nearly the same confidence with their p-values. Not only that, almost all the genes found in the contrast table had a negative correlation between expression and size. It is concerning that it only found one positive correlation out of the entire top 10. This means that the data is suggesting it is not the expression of a gene that controls eye size, but the lack of one that seems to be affecting it the most. While many from the contrast table would be worth the time to investigate, we recommend the first and third in that set. The second gene in that set, FBgn0027584, just has a bit too much of a spread in its scatterplot to be confident in that negative correlation.

There is a slight worry about the results we got in the sense that this is still not a conclusive answer to the question of what is causing the eye size of these flies to fluctuate. Sure, we have some genes that show a possible influence, but we cannot say that any of the correlations found are the actual cause. With the data we received, we suspect it may be a combination of genes expressing themselves in certain ways that leads to these specific eye sizes. Finding the connection between multiple genes and their expression as a group compared to the eye size of that strain will most likely be the next step. Hopefully, we have at least made a starting point for more research since this is only scratching the surface.

ACKNOWLEDGMENTS

A big thank you to Aaron Schlorke and his group who helped us understand a lot of the process. We would not have been able to get past the roadblocks we experienced working with limma without the analogous work they did.

REFERENCES

1. Nils Kolling (2015), Drosophila melanogaster development in *Quantitative genetics of gene expression during fruit fly development.*
2. Charlotte Soneson and Mauro Delorenzi: *A comparison of methods for differential expression analysis of RNA-seq data.* BMC Bioinformatics 2013 14:91.
3. Charity law, Monther Alhamdoosh, Shain su, Xueyi Dong, Luyi Tain, Gordon K. Smyth, and Matthew E. Ritche, *RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edge*R, *https://www.bioconductor.org/packages/devel/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html*.
4. University of California Davis*, RNA-Seq Workshop: Differential Expression with Limma-Voom,* https://ucdavis-bioinformatics-training.github.io/2018-June-RNA-Seq-Workshop/thursday/DE.html

ABOUT STUDENT AUTHORS

Alex Stocking is a CS graduate from Purdue University Fort Wayne who will graduate in May 2021.
Sivakami Thinnappan is a CS graduate from Purdue University Fort Wayne who will graduate in December 2022.

PRESS SUMMARY

Genes are more than just the nucleotides that they are made of, they also have an important value when analyzing how that gene might affect physical traits of the animal, expression. Expression is simply how much of that gene's code is being sent out in the body to be processed and executed. Microarrays have been one of the most popular ways of analyzing this massive data, but a new, more powerful method of analyzing known as RNA-Seq analysis makes it easier than ever. In this paper, we will use the expression and eye size data of 205 individual female fruit flies to find if there is a connection between the expression of those genes and the eye size of said fruit fly. We will show how you can use the language R and the voom method from the limma package to accomplish this. The work done in this paper will hopefully bring us one step closer to understanding the interaction between gene expression and physical traits.