PennState

THE HUCK INSTITUTES OF THE LIFE SCIENCES

STAT 555 PROJECT REPORT

Analysis of RNA-seq data using DESeq2

Instructor
Dr. Naomi Altman

PI
Dr. Santhosh Girirajan

Submitted By
Shamitha Shetty
sbs5620@psu.edu

Date: May 02, 2016

Identification of differentially expressed genes in RNA-seq data from Drosophila knockdown models for *16p11.2* region and its functional analysis

Abstract: The *16p11.2* region has been a hotspot for various microdeletion and CNVs leading to various neurodevelopmental disorders in humans. *Drosophila melanogaster* knockdown models have been successfully used to study the effect of deletion of these orthologous genes in flies. We performed a RNA-seq analysis on six gene knockdown models in flies. Here, we suggest a workflow to perform a differential expression analysis of the RNA-seq data using the *DESeq2* package in *R*. We also perform a functional annotation study of the significant genes obtained from analysis using *DAVID*.

Introduction:

The data was obtained from RNA-seq analysis of fly models for the *16p11.2* microdeletion. This deletion is implicated in neurodevelopmental disorders of the CNV.

The *16p11.2* gene knockdown models in *Drosophila melanogaster* were constructed for studying the deletion. Ten human genes found in the *16p11.2* region have orthologous genes in *Drosophila*. Knockdown models were created for these orthologs using crosses and they were assessed for any abnormal phenotype in the progeny. RNA-seq analysis was performed for six models with an abnormal neurodevelopmental phenotype. The human genes with their orthologs in *Drosophila* are as follows:

- *C16orf53 (Pa1)*
- *CDIPT (Pis)*
- *MAPK3 (rl)*
- *DOC2A (Rph)*
- *CORO1A (coro)*
- *KCTD13 (CG10465)*

We had three biological replicates for each of these genes as well as for wild-type gene. The RNA-Seq data was recorded in a .csv file with the corresponding FlyBase Gene IDs for further analysis. A qPCR analysis of the fly heads was also performed to confirm the abnormal phenotype in these models. The objective of the analysis is to find if a mutation in any of these genes or a knockdown of the entire gene is responsible for neurodevelopmental disorders in Drosophila models and then find possible convergence in humans.

METHODS:

The overview of the basic process followed for the project is shown in Figure 1. The raw counts for the RNA-seq analysis were obtained for the biological replicates of the knockdown genes as well as the wild-type. The primary filtering, differential expression analysis and sorting the results was performed using R. Later, the significant genes were analysed for functional annotation and *Gene Ontology* using DAVID (Source: https://david.ncifcrf.gov/). The results obtained from the analysis were imported in R. An annotation of all the significant genes was performed by comparing it to the *org.Dm.eg.db* database. Further, convergence studies of the orthologous genes in humans will be performed using this data.

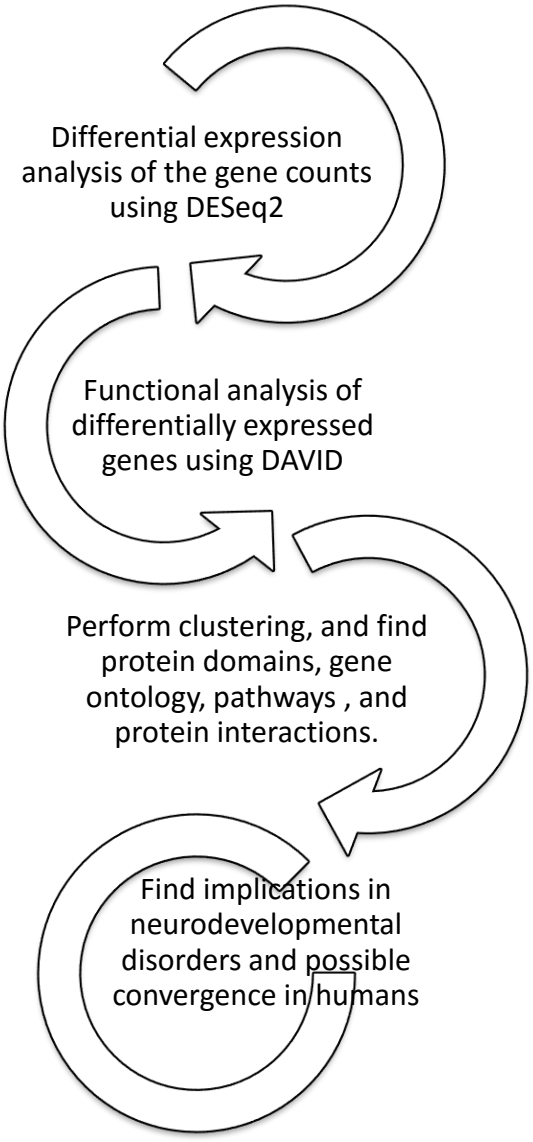

Fig. 1: An overview of the steps undertaken for the analysis

The chart of working pipeline below (Fig. 2) shows the overall working and computing flow for the analysis, right from the process of obtaining the data, the analysis performed, the functions used in R to importing the functional analysis results. A detailed R script with the significance of each chunk is provided in a .rmd format as a supplementary information file. Refer *Project_SHETTY.Rmd.*

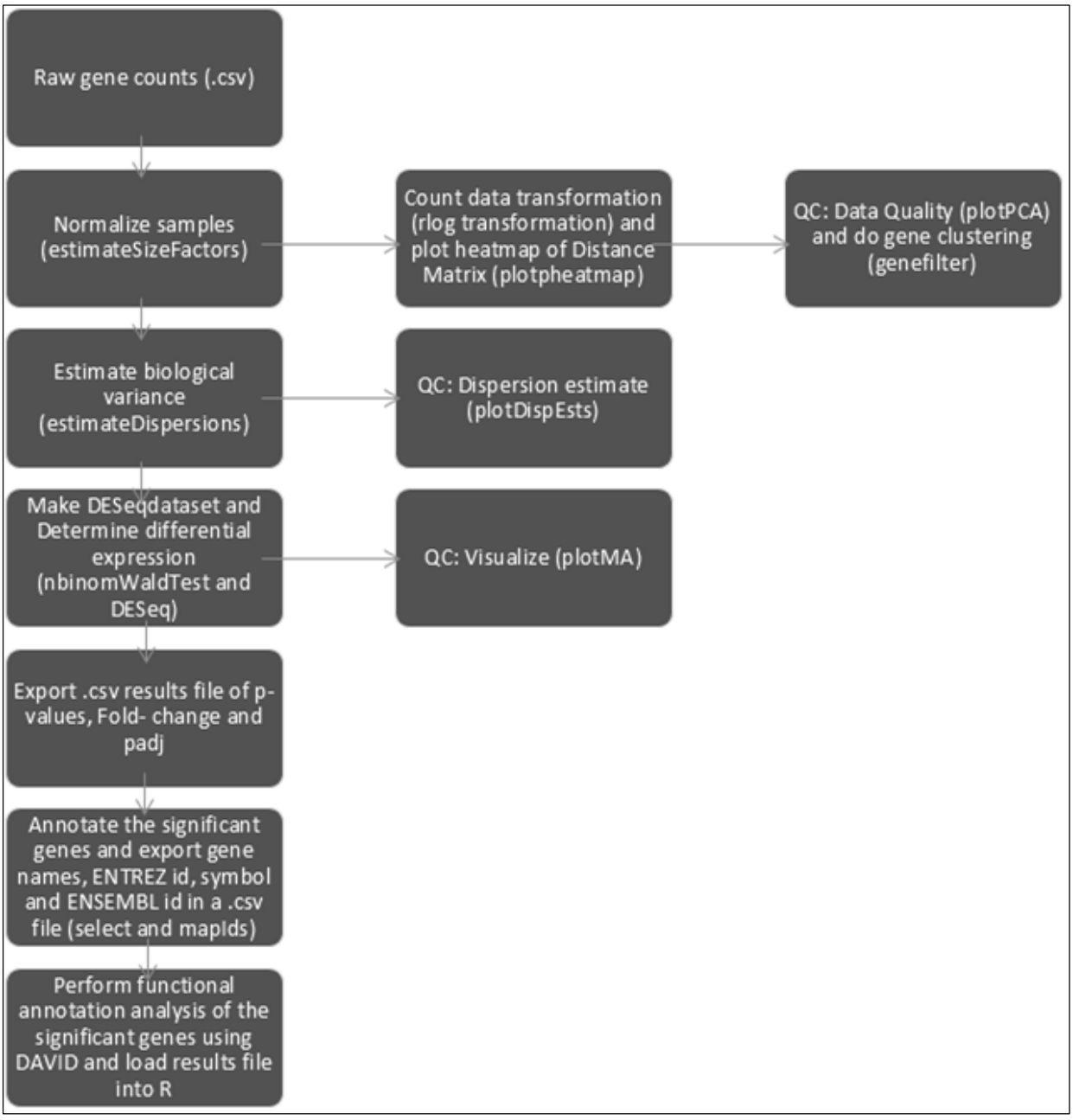

Fig.2: R pipeline for analysis

Results:

1. rlog transformation

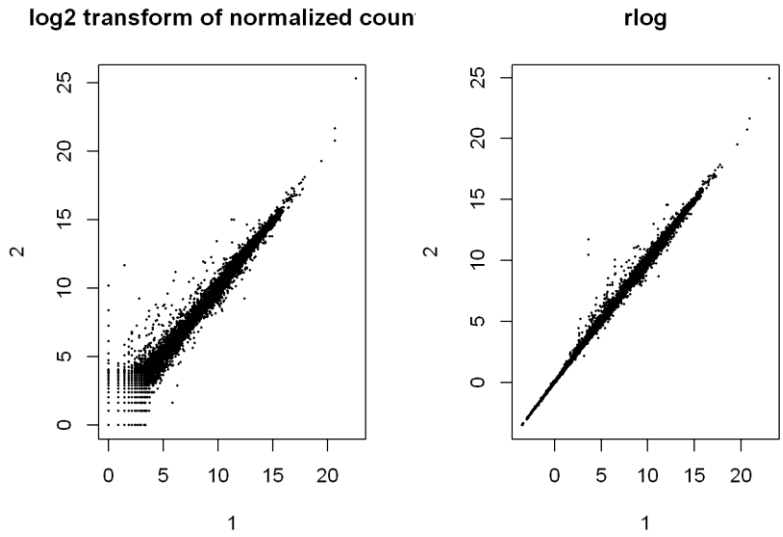

Fig. 3: Scatterplot of transformed counts from two samples. Shown are scatterplots using the log2 transform of normalized counts (left side) and using the rlog (right side).

We can see how genes with low counts (bottom left-hand corner) seem to be excessively variable on the ordinary logarithmic scale, while the rlog transform compresses differences for the low count genes for which the data provide little information about differential expression.

2. Heatmaps

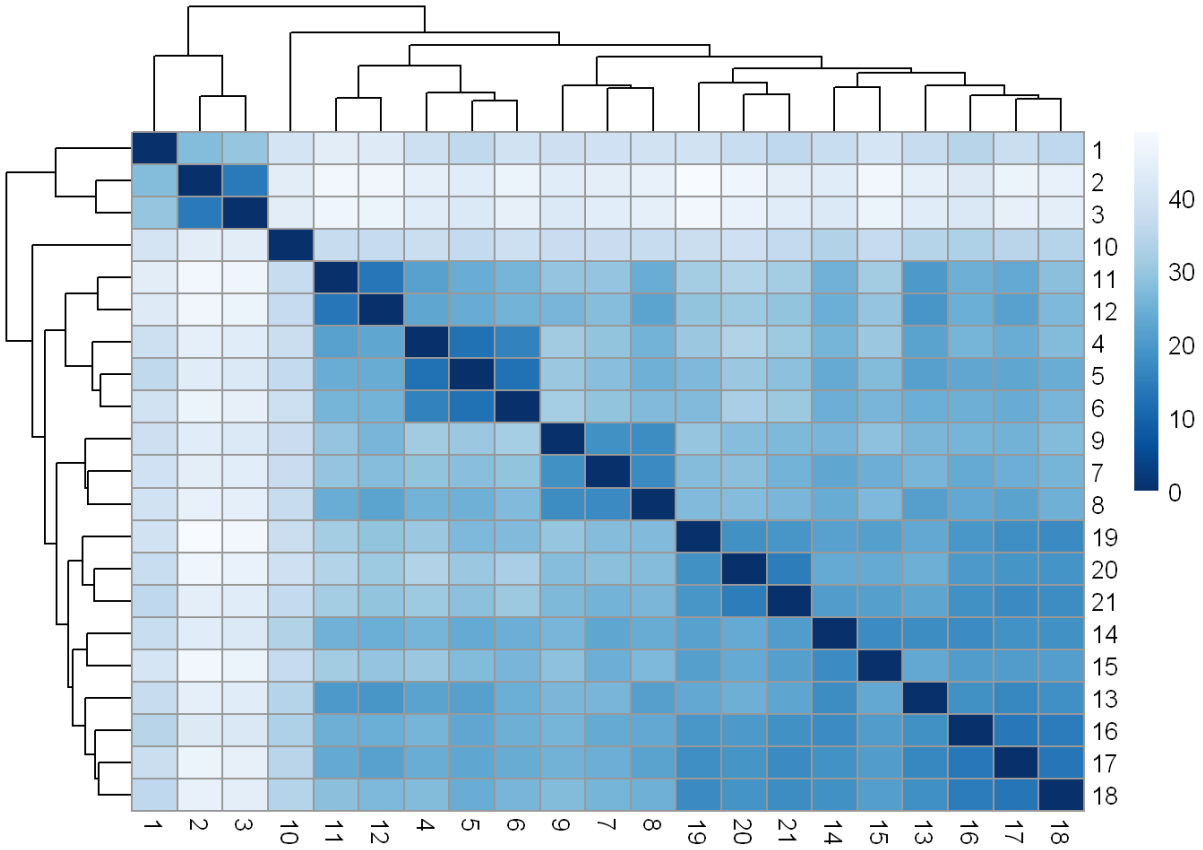

Fig.4: **Heatmap of sample-to-sample distances using the rlog-transformed values.**

This heatmap was plotted to assess overall similarity between samples and to see if it fit to the expectation from the experiment's design. The dist function was used to calculate the Euclidean distance between the samples and was used on the rlog- transformed data to ensure nearly equal contribution from all the samples.

3. MA plot

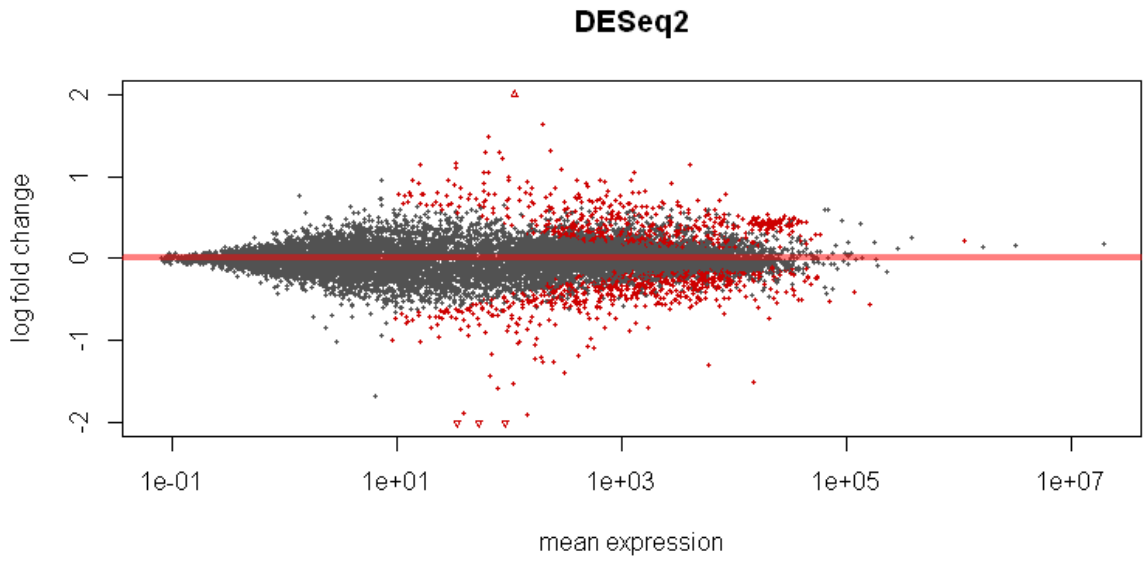

Fig. 5: MA plot of Differential expression analysis

This MA Plot was done for quality testing by plotting average of the counts normalized by size factor on the x-axis and log2 fold change on the y-axis representing each gene with a dot. Genes with an adjusted *p* value below a threshold (here 0.1, the default) are shown in red.

4. MA plot for log2FC

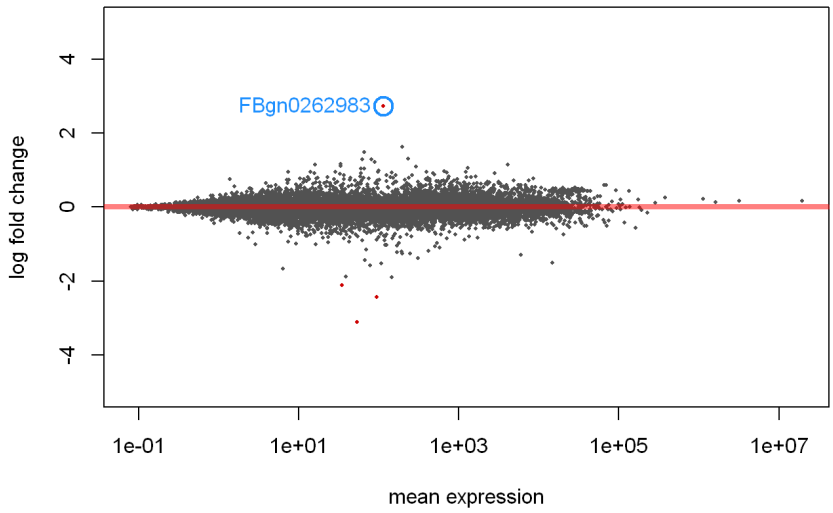

Fig. 6: **An MA-plot of a test for large log2 fold changes.**

The red points indicate genes for which the log2 fold change was significantly higher than 1 or less than -1 (treatment resulting in more than doubling or less than halving of the normalized counts). The point circled in blue indicates the gene with the lowest adjusted *p* value.

5. Histogram of pvalues

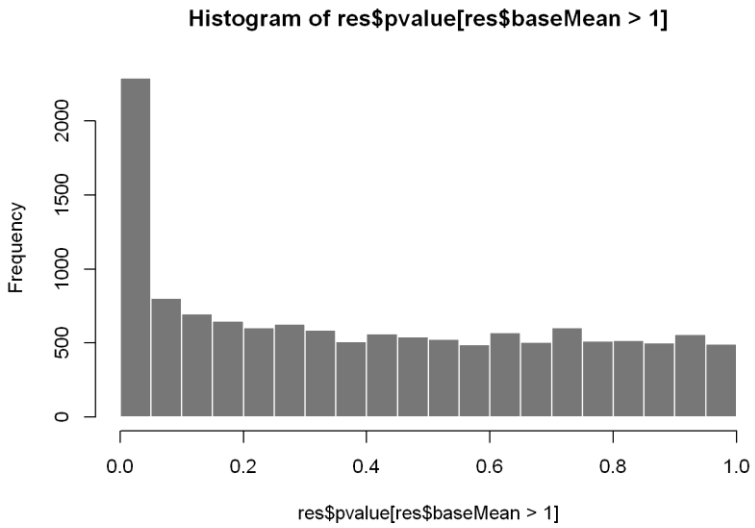

Fig. 7: Histogram of *p* values for genes with mean normalized count larger than 1.

We observe the expected shape thus showing that the analysis was successful and FDR estimation can be done.

6. Gene Clustering

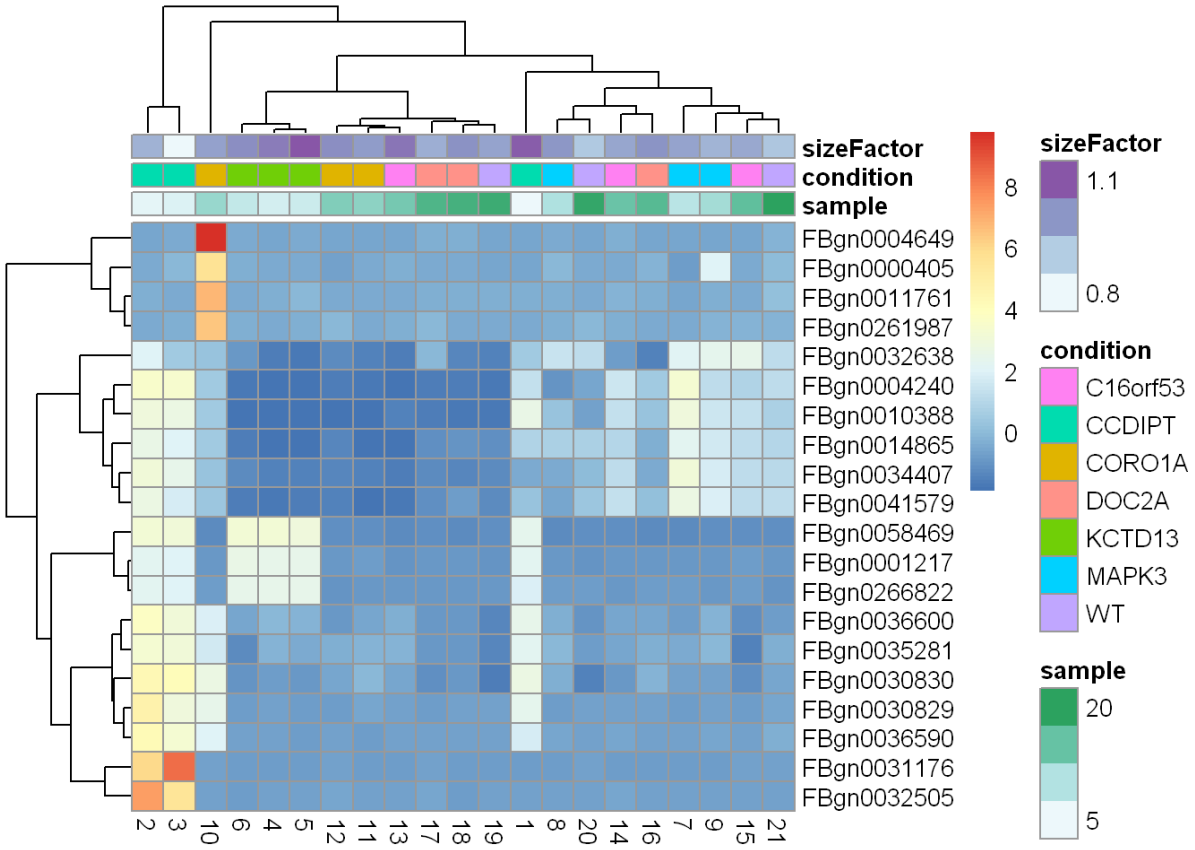

Fig. 8: Heatmap of relative rlog-transformed values across samples.
Treatment status and cell line information are shown with colored bars at the top of the heatmap and the key is given on the right of the figure. It shows the genes that covary.

7. Sample QC: Transform data to log space and visualize samples

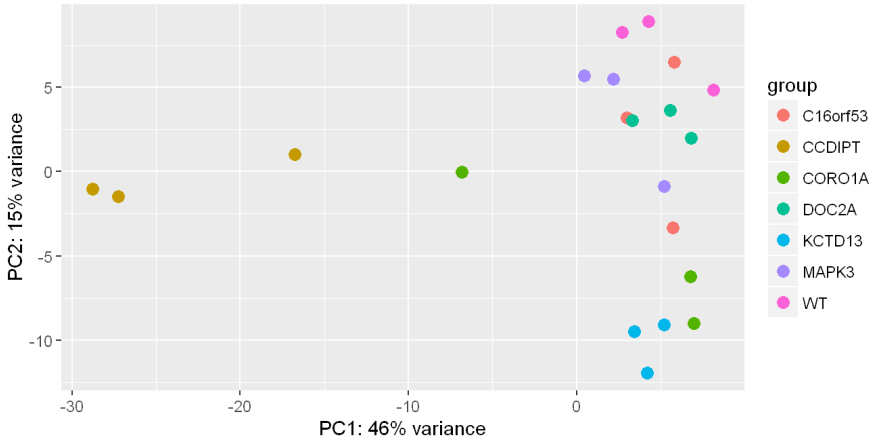

Fig. 9: PCA plot of biological variance

We can clearly see the different clusters being plotted as a result of the differential expression within these genes. CCDIPT shows most variance as compared to the rest of the samples.

8. Estimate biological variance and visualize

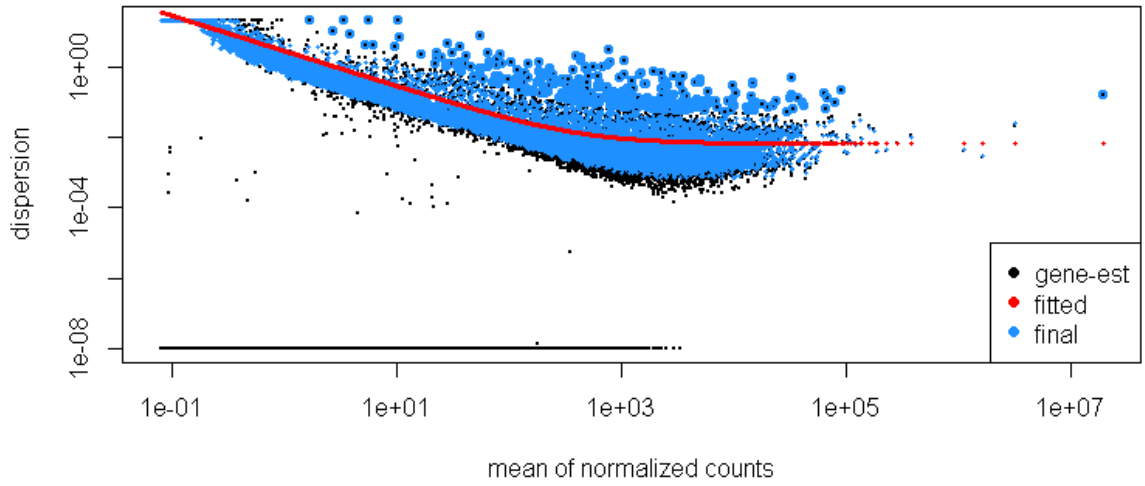

Fig. 10: Plot of mean of normalized counts vs dispersion

We can see that due to the normalization the biological variance was reduced in the sample.

9. Results of Differential expression analysis

```
> results
log2 fold change (MAP): condition WT vs C16orf53
Wald test p-value: condition WT vs C16orf53
DataFrame with 15526 rows and 6 columns
                          baseMean log2FoldChange      lfcSE        stat      pvalue      padj
                         <numeric>      <numeric>  <numeric>   <numeric>   <numeric> <numeric>
FBgn0000003           4.157232e-01   -0.028297941 0.09010026 -0.31407168 0.753466596        NA
FBgn0000008           1.373449e+03   -0.097157915 0.11940750 -0.81366679 0.415835866 0.7294320
FBgn0000014           1.005504e+00    0.009108037 0.14430764  0.06311542 0.949674596        NA
FBgn0000015           8.116016e-01    0.227695475 0.14551510  1.56475495 0.117640378        NA
FBgn0000017           1.032277e+04   -0.201885153 0.06257215 -3.22643792 0.001253414 0.0217338
...                            ...            ...        ...         ...         ...       ...
FBgn0267794           4.332910e+00    0.2601586 0.24455810   1.0637907 0.287423504        NA
FBgn0267795           2.882373e+03   -0.1862581 0.06938935  -2.6842455 0.007269371 0.0732219
__no_feature          3.179654e+06    0.1570121 0.16376547   0.9587619 0.337678721 0.6632151
__ambiguous           1.624583e+06    0.1370420 0.06300241   2.1751873 0.029616084 0.1777616
__alignment_not_unique 1.948180e+07    0.1706467 0.26103655   0.6537271 0.513287612 0.7934417
```

Fig. 11: Using the results() function of the DESeq2 analysis package

The above table containing the p-values, log2 fold change and adjusted p-values was obtained from the analysis. This was further filtered by only selecting the significant genes having padj<0.05 for GO analysis.

10. GO analysis

GOTERM_BP_FAT	neuron differentiation	RT	36	65.5	1.5E-32	12.7	1.4E-29	2.3E-29
GOTERM_BP_FAT	neuron development	RT	31	56.4	2.5E-27	12.9	1.2E-24	4.0E-24
GOTERM_BP_FAT	neuron projection morphogenesis	RT	27	49.1	7.1E-24	13.6	2.2E-21	1.1E-20
GOTERM_BP_FAT	neuron projection development	RT	27	49.1	7.8E-24	13.6	1.8E-21	1.2E-20
GOTERM_BP_FAT	cell projection organization	RT	29	52.7	7.8E-24	11.5	1.4E-21	1.2E-20
GOTERM_BP_FAT	cellular component morphogenesis	RT	32	58.2	1.9E-23	8.9	3.0E-21	3.0E-20
GOTERM_BP_FAT	cell morphogenesis involved in differentiation	RT	27	49.1	3.2E-23	12.9	4.2E-21	5.0E-20
GOTERM_BP_FAT	cell motion	RT	27	49.1	3.8E-23	12.8	4.4E-21	5.9E-20
GOTERM_BP_FAT	cell morphogenesis	RT	30	54.5	7.7E-23	9.8	7.9E-21	1.2E-19
GOTERM_BP_FAT	cell projection morphogenesis	RT	27	49.1	1.3E-22	12.2	1.2E-20	2.1E-19
GOTERM_BP_FAT	cell morphogenesis involved in neuron differentiation	RT	26	47.3	2.2E-22	13.0	1.9E-20	3.5E-19
GOTERM_BP_FAT	cell part morphogenesis	RT	27	49.1	3.2E-22	11.8	2.4E-20	4.9E-19
GOTERM_BP_FAT	axonogenesis	RT	22	40.0	1.8E-20	16.0	1.3E-18	2.9E-17
GOTERM_BP_FAT	axon guidance	RT	19	34.5	2.9E-19	20.2	1.9E-17	4.5E-16
GOTERM_BP_FAT	regulation of cell development	RT	13	23.6	4.7E-11	14.3	2.9E-9	7.3E-8
GOTERM_BP_FAT	sensory organ development	RT	19	34.5	8.2E-11	6.7	4.8E-9	1.3E-7
GOTERM_BP_FAT	regulation of nervous system development	RT	11	20.0	3.4E-10	17.6	1.8E-8	5.2E-7
GOTERM_BP_FAT	memory	RT	6	10.9	3.9E-6	24.1	7.6E-5	6.1E-3
GOTERM_BP_FAT	behavior	RT	14	25.5	4.1E-6	4.7	7.8E-5	6.4E-3
GOTERM_BP_FAT	metamorphosis	RT	13	23.6	4.1E-6	5.2	7.7E-5	6.4E-3
GOTERM_BP_FAT	regulation of cellular component size	RT	8	14.5	4.2E-6	11.7	7.7E-5	6.5E-3
GOTERM_BP_FAT	appendage morphogenesis	RT	11	20.0	4.2E-6	6.6	7.7E-5	6.6E-3
GOTERM_BP_FAT	imaginal disc-derived appendage development	RT	11	20.0	4.4E-6	6.5	7.8E-5	6.9E-3
GOTERM_BP_FAT	post-embryonic development	RT	14	25.5	4.6E-6	4.7	8.0E-5	7.2E-3
GOTERM_BP_FAT	actin cytoskeleton organization	RT	9	16.4	4.7E-6	9.1	8.1E-5	7.4E-3
GOTERM_BP_FAT	appendage development	RT	11	20.0	4.9E-6	6.5	8.2E-5	7.7E-3
GOTERM_BP_FAT	tissue morphogenesis	RT	11	20.0	5.1E-6	6.4	8.4E-5	8.0E-3
GOTERM_BP_FAT	compound eye morphogenesis	RT	11	20.0	5.7E-6	6.3	9.2E-5	8.9E-3
GOTERM_BP_FAT	actin filament bundle formation	RT	5	9.1	6.9E-6	38.0	1.1E-4	1.1E-2
GOTERM_BP_FAT	glial cell migration	RT	5	9.1	8.6E-6	36.1	1.3E-4	1.3E-2
GOTERM_BP_FAT	compound eye photoreceptor cell differentiation	RT	8	14.5	8.9E-6	10.4	1.4E-4	1.4E-2
GOTERM_BP_FAT	head segmentation	RT	5	9.1	1.1E-5	34.4	1.6E-4	1.6E-2
GOTERM_BP_FAT	eye morphogenesis	RT	11	20.0	1.1E-5	5.9	1.6E-4	1.7E-2
GOTERM_BP_FAT	eye photoreceptor cell differentiation	RT	8	14.5	1.2E-5	10.0	1.7E-4	1.9E-2
GOTERM_BP_FAT	learning or memory	RT	7	12.7	1.2E-5	13.1	1.8E-4	1.9E-2
GOTERM_BP_FAT	motor axon guidance	RT	5	9.1	1.3E-5	32.8	1.8E-4	2.0E-2
GOTERM_BP_FAT	imaginal disc-derived wing morphogenesis	RT	10	18.2	1.4E-5	6.6	1.9E-4	2.1E-2
GOTERM_BP_FAT	wing disc morphogenesis	RT	10	18.2	1.5E-5	6.6	2.0E-4	2.3E-2
GOTERM_BP_FAT	imaginal disc development	RT	13	23.6	1.5E-5	4.6	2.1E-4	2.4E-2
GOTERM_BP_FAT	morphogenesis of an epithelium	RT	10	18.2	1.8E-5	6.4	2.4E-4	2.8E-2
GOTERM_BP_FAT	actin filament organization	RT	7	12.7	1.9E-5	12.2	2.5E-4	2.9E-2
GOTERM_BP_FAT	post-embryonic appendage morphogenesis	RT	10	18.2	2.0E-5	6.3	2.6E-4	3.2E-2
GOTERM_BP_FAT	regulation of neuron differentiation	RT	6	10.9	3.0E-5	16.0	3.8E-4	4.6E-2

Fig. 12: GO analysis showing biological process involved in neurodevelopmental disorders

The GO analysis biological process (as shown in the .rmd file) using DAVID gave interesting results. The significant genes from the differential expression analysis can be seen here to be involved in a host of biological processes that contribute to neurodevelopmental health. Right from neuron development, differentiation, gliogenesis to memory these genes seem to play an important role in maintaining the health of the CNS. Any mutation in these could adversely affect the functioning of the nervous system.

11. KEGG Pathway result for GO analysis

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
KEGG_PATHWAY	Wnt signaling pathway	RT		3	5.6	4.8E-2	4.2E-1

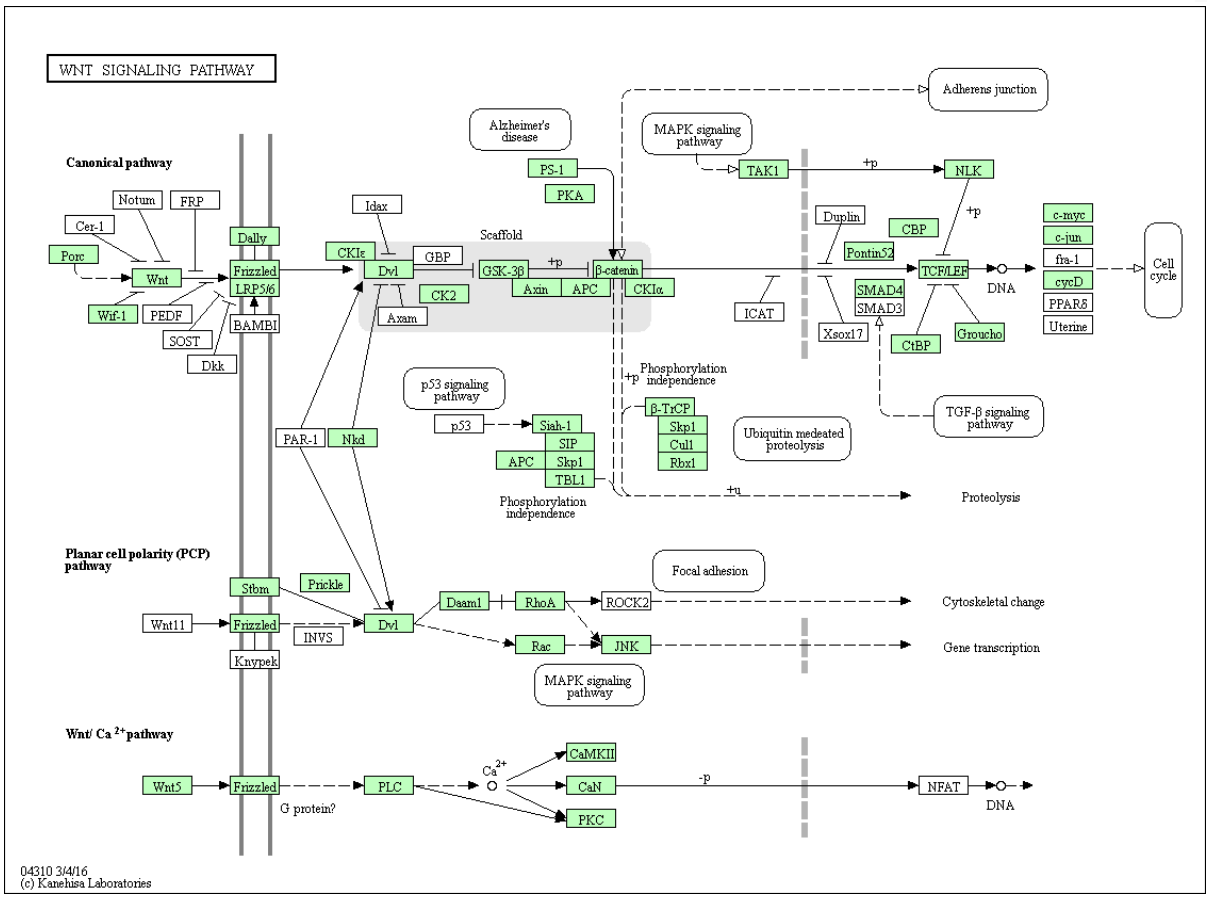

Fig. 13: The WNT Signaling pathway

The KEGG pathway from the GO analysis of the subset of the significant genes playing a role in neurodevelopmental function gave the result as 'the WNT signalling pathway'. A closer look at the pathway interactions shows that it might play a role in Alzheimer's disease.

12. Gene associations

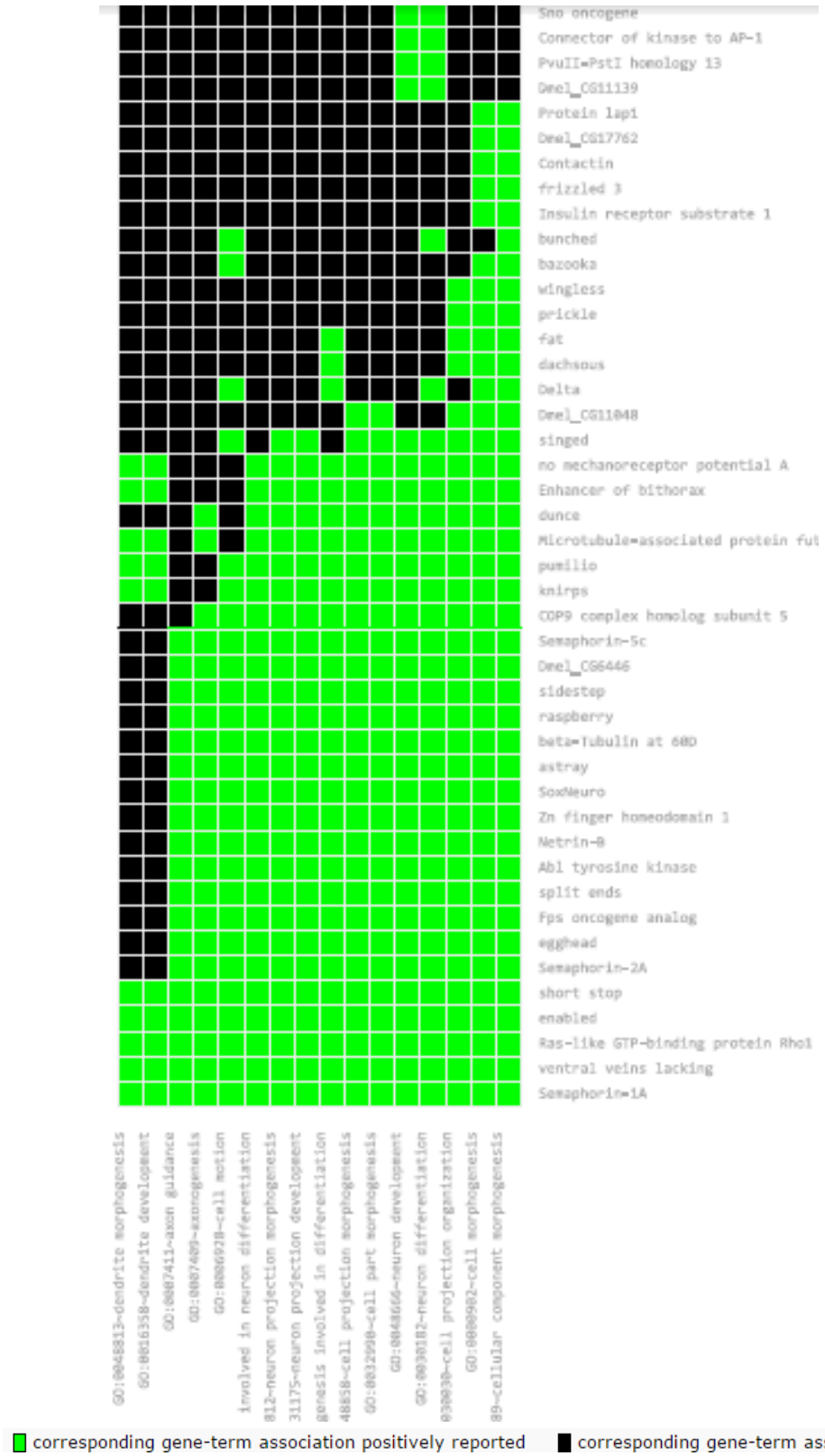

Fig. 14(a): Gene associations between the processes and significant genes

From the .rmd file we can see the various clusters that were formed according to the gene function and processes. The above cluster showing significance for neurodevelopmental disorders had a high enrichment score in the DAVID analysis. It gives a brief overview of each of the reported gene associations in the neurodevelopment processes.

13. Gene batch using kappa statistic

Similarity Score: Very High (0.75-1) High (0.5-0.75) Moderate (0.25-0.5) Low (<0.25)

#	Category	Term	Kappa
1	GOTERM_BP_FAT	axon guidance	1.00
2	GOTERM_BP_FAT	axonogenesis	0.93
3	GOTERM_BP_FAT	cell morphogenesis involved in neuron differentiation	0.84
4	GOTERM_BP_FAT	neuron projection morphogenesis	0.82
5	GOTERM_BP_FAT	neuron projection development	0.82
6	GOTERM_BP_FAT	cell projection morphogenesis	0.80
7	GOTERM_BP_FAT	cell part morphogenesis	0.80
8	GOTERM_BP_FAT	cell morphogenesis involved in differentiation	0.79
9	GOTERM_BP_FAT	neuron development	0.75
10	GOTERM_BP_FAT	cell motion	0.72
11	GOTERM_BP_FAT	cell projection organization	0.72
12	GOTERM_BP_FAT	neuron differentiation	0.68
13	GOTERM_BP_FAT	cell morphogenesis	0.64
14	GOTERM_BP_FAT	cellular component morphogenesis	0.59
15	GOTERM_BP_FAT	motor axon guidance	0.41
16	GOTERM_BP_FAT	embryonic development via the syncytial blastoderm	0.39
17	GOTERM_BP_FAT	embryonic development ending in birth or egg hatching	0.38
18	GOTERM_BP_FAT	neuron recognition	0.34
19	GOTERM_BP_FAT	cell recognition	0.34
20	INTERPRO	Plexin/semaphorin/integrin	0.34
21	INTERPRO	Semaphorin/CD100 antigen	0.34
22	SMART	Sema	0.34
23	SMART	PSI	0.34
24	GOTERM_BP_FAT	dendrite morphogenesis	0.33
25	GOTERM_BP_FAT	dendrite development	0.33

Fig. 15: Kappa statistic for the gene association

"Any given gene is associating with a set of annotation terms. If genes share similar set of those terms, they are most likely involved in similar biological mechanisms. The algorithm adopts kappa statistics to quantitatively measure the degree of the agreement how genes share the similar annotation terms. Kappa result ranges from 0 to 1. The higher the value of Kappa, the stronger the agreement. Kappa more than 0.7 typically indicates that agreement of two genes are strong. Kappa values greater than 0.9 are considered excellent."
(Source: DAVID website- Linear Search Algorithm in Gene Name Batch Viewer)

14. Annotation

```
> resannotated[1:10,]
       ENSEMBL      SYMBOL                                  GENENAME ENTREZID
1  FBgn0000003        <NA>                                      <NA>     <NA>
2  FBgn0000008           a                                       arc    43852
3  FBgn0000014        abd-A                            abdominal A    42037
4  FBgn0000015        Abd-B                            Abdominal B    47763
5  FBgn0000017         Abl                      Abl tyrosine kinase    45821
6  FBgn0000018         abo                          abnormal oocyte    44793
7  FBgn0000024         Ace                   Acetylcholine esterase    41625
8  FBgn0000028        acj6             abnormal chemosensory jump 6    47080
9  FBgn0000032      Acph-1                        Acid phosphatase 1    48445
10 FBgn0000036 nAChRalpha1 nicotinic Acetylcholine Receptor alpha1    42918
```

Fig. 16: Annotation in R of the significant genes

Finally, an annotation of all the significant genes was performed using the packages *Annotation.Dbi* and *org.Dm.eg.db.* We can see the ENSEMBL, GENENAME, SYMBOL S and

ENTREZID stored for each gene. Some genes also had missing values. These annotations can be used to perform further functional analysis and convergence studies.

Discussion:

Using the DESeq2 package it was possible to perform a very detailed differential expression analysis of the genes. The significant genes obtained from the analysis were then analysed on DAVID for functional annotation which also gave promising results. It gave a lot of information about the biological processes, pathways, clustering and gene associations between the significant genes. To confirm the results of the analysis I also ran the significant genes on the *geneontology.org* website which uses *PANTHER*.

motor neuron axon guidance								61	15	3.94	3.80	+ 4.56E-02
↳single-organism developmental process								3344	275	216.18	1.27	+ 1.25E-02
↳developmental process								3368	277	217.73	1.27	+ 1.12E-02
↳anatomical structure development								3213	264	207.71	1.27	+ 2.33E-02
↳system development								2298	208	148.56	1.40	+ 5.63E-04
↳multicellular organism development								2865	238	185.21	1.29	+ 3.89E-02
↳single-multicellular organism process								3107	258	200.85	1.28	+ 1.41E-02
PLXND1	Plexin-D1		branchiomotor neuron axon guidance		GO_Central	Homo sapiens	IBA	PANTHER:PTN001476654	plexin pthr22625		GO_REF:0000033	20150911
SLIT1	Slit homolog 1 protein		motor neuron axon guidance		UniProt	Homo sapiens	IMP		family not named pthr24373		PMID:16162649	20100122
SEMA3A	Semaphorin-3A		branchiomotor neuron axon guidance		Ensembl	Homo sapiens	IEA	Ensembl:ENSMUSP00000030714	semaphorin pthr11036		GO_REF:0000019	20160409
PLXNA3	Plexin-A3		branchiomotor neuron axon guidance		GO_Central	Homo sapiens	IBA	PANTHER:PTN001476654	plexin pthr22625		GO_REF:0000033	20150911
ALCAM	CD166 antigen		motor neuron axon guidance		Ensembl	Homo sapiens	IEA	Ensembl:ENSMUSP00000023312	cell surface glycoprotein muc18-related pthr11973		GO_REF:0000019	20160409
ERBB2	Receptor tyrosine-protein kinase erbB-2		motor neuron axon guidance		Ensembl	Homo sapiens	IEA	Ensembl:ENSMUSP00000053897	tyrosine-protein kinase receptor pthr24416		GO_REF:0000019	20160409
SEMA3F	Semaphorin-3F		branchiomotor neuron axon guidance		Ensembl	Homo sapiens	IEA	Ensembl:ENSMUSP00000141865	semaphorin pthr11036		GO_REF:0000019	20160409
NOG	Noggin		motor neuron axon guidance		Ensembl	Homo sapiens	IEA	Ensembl:ENSMUSP00000061427	bone morphogenetic protein inhibitor, noggin pthr10494		GO_REF:0000019	20160409
EGR2	E3 SUMO-protein ligase		motor neuron axon guidance		Ensembl	Homo sapiens	IEA	Ensembl:ENSMUSP00000041053	early growth response		GO_REF:0000019	20160409

Fig. 17 (a): GO analysis using *PANTHER* gave similar results and also returned a list of the orthologous genes in *Homo sapiens.*

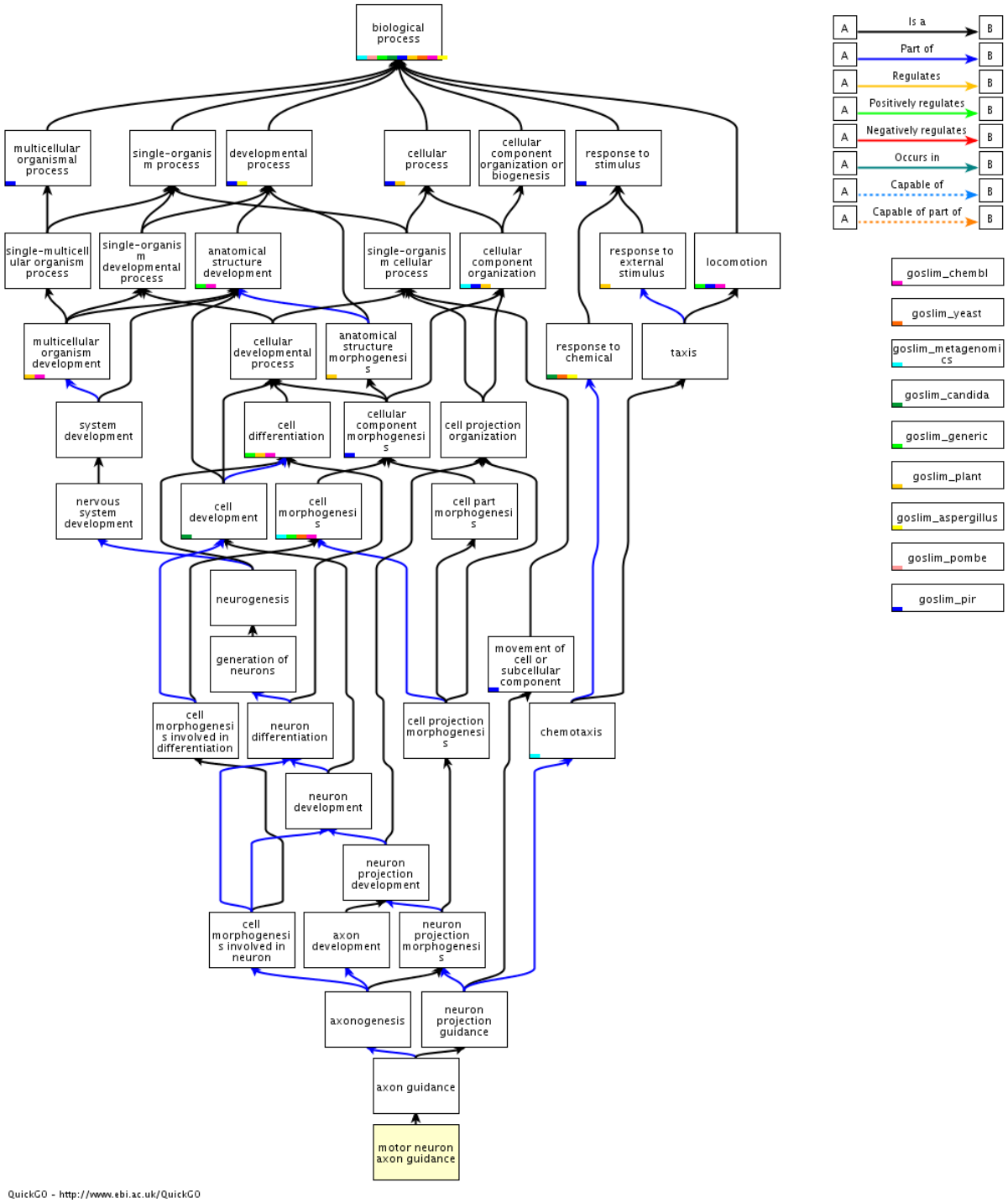

Fig. 17 (b): The graph view of the pathway and where the node 'motor neuron axon guidance' lies on it

This search gave similar results too though not exactly the same as the DAVID analysis. However, it also gave me a list of the orthologous genes that were affected in humans thus showing some kind of convergence between the two. The graph view of the process gave a detailed view of where the node lies in the pathway. This relation can be used to assess how the functioning of these genes might be affected due to mutations in any of these links.

The results obtained from the analysis showed that these genes played some role in neurodevelopment which needs to be assessed in detail further. Due to the scope of the

report only a basic analysis of the RNA-seq data was possible. A detailed analysis by reproducing the experiment and refining the suggested steps will yield more significant and definite results into the role that these genes play in causing neurodevelopmental disorders.

References:

1. Rosenfeld, J. A., Coppinger, J., Bejjani, B. A., Girirajan, S., Eichler, E. E., Shaffer, L. G., & Ballif, B. C. (2010). Speech delays and behavioral problems are the predominant features in individuals with developmental delays and 16p11.2 microdeletions and microduplications. *Journal of Neurodevelopmental Disorders*, *2*(1), 26–38. http://doi.org/10.1007/s11689-009-9037-4
2. Polyak, A., Rosenfeld, J. A., & Girirajan, S. (2015). An assessment of sex bias in neurodevelopmental disorders. *Genome Medicine*, *7*(1), 94. http://doi.org/10.1186/s13073-015-0216-5
3. Bioconductor workflows: RNA-seq workflow at the gene level
4. Princeton HTSEQ tutorials
5. DESeq analysis of RNA-seq data from turfgrass: https://github.com/mpj5142/huff_turfgrass
6. Fernandez BA, Roberts W, Chung B, Weksberg R, Meyn S, et al. Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *J Med Genet. 2009*.
7. E.K. Bijlsma, A.C.J. Gijsbers, J.H.M. Schuurs-Hoeijmakers, A. van Haeringen, D.E. Fransen van de Putte, B.-M. Anderlid, J. Lundin, P. Lapunzina, L.A. Pérez Jurado, B. Delle Chiaie. Extending the phenotype of recurrent rearrangements of 16p11.2: Deletions in mentally retarded patients without autism and in normal individuals. *European Journal of Medical Genetics, Volume 52, Issue 2, Pages 77-87*
8. PANTHER (source: pantherdb.org)
9. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc. 2009; 4(1):44–57.* doi:http://dx.doi.org/10.1038/nprot.2008.211.