AMSI
AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE
National Collaboration in the Mathematical Sciences

Adding to the Galaxy of statistical methods for genomic analysis

Shian Su
Walter and Eliza Hall Institute of Medical Research

July 3, 2014

Introduction

Project Description

The goal of this project is to provide methods for medical researchers to perform their own statistical analysis on a range of common experiments, the approach taken is to create a graphical interface in the web-based platform *Galaxy* to run R commands in the back-end, and produce results very similar to what would be obtained if the data were taken to a statistician. Specifically, the type of analysis of interest is *differential expression*, the change of expression level of a gene between two experimental conditions measured by the amount of mRNA present for each gene.

RNA sequencing is a fast developing technology that allows medical researchers to quickly and accurately sequence large amounts of genetic data, proper analysis of such data provides medical researchers with useful insight into the relationship between diseases and genes. To analyse RNA sequencing data requires sophisticated statistical techniques and also the use of computers; unfortunately the statistical methods used are usually beyond what is accessible to a medical researcher, and the interaction with computers requires the knowledge of the programming language R which is equally unsuitable for inexperienced users. The current practice is for a medical researcher to bring the data they generate to a statistician for a discussion of what effects they wish to study, the statistician then suggests the statistical analysis that should be performed.

Motivation

This project is motivated by the fact that statisticians involved in biomedical data analysis spend a significant amount of their time performing quite routine procedures for various different parties. This is undesirable due to the fact that RNA sequencing (referred to RNAseq from now on) data is very new and current methods for analysing RNAseq data, especially when studying complicated interactions, is still far from mature. It is therefore of great interest to statisticians to be able to spend time researching the statistical properties of RNAseq data and developing new methods for different testing conditions, something that is hindered by their obligation to analyse increasing volumes of data from well understood experimental designs.
The creation third party tools that perform a set of unbiased procedures on raw data to produce unbiased results for the sake of academic reproducibility, it provides more reliable documentation of the statistical procedures than a custom R script written specifically for that analysis which may or may not have been preserved.

It is also important to note that there are already ways for biologists to perform the analysis described in this project, however they may be using less sophisticated methods due to such methods being more 'user friendly'. Such situations are less desirable than having medical researchers using the best statistical tools available, especially when such tools are open source, significantly cheaper than other state-of-the-art medical devices. So this project also hopefully opens up access to and consequently increases the usage of more mature statistical methods.

Background

Gene Expression

DNA holds all our genetic information, it is a double stranded molecule that can be thought of as a string of four possible letters. These letters, or bases as they are called in biology, are effectively messages that define a blueprint for a particular human. DNA does not simply determine an individual's physical traits but rather acts constantly to regulate bodily functions, therefore many disorders in living organisms can be linked to particular genes. The process by which DNA contributes to bodily functions is by first having segments of itself *transcribed* as mRNA molecules before the mRNA molecules are *translated* into proteins (strictly speaking, mRNA is translated into polypeptides that assemble into proteins).

The sections of DNA that can be used to produce useful proteins is called a *gene*

and the act of having protein produced from a gene is called *expression* of that gene. Since proteins are very diverse in structure with varying levels of stability, it is hard to isolate and measure the exact levels of proteins present, instead we measure the amount of mRNA to give a reasonable indication of expression levels.

RNA Sequencing

Newly developed technologies allow us to sequence mRNA (the actual sequencing is done on cDNA which simply copies the bases of the mRNA onto a more stable molecule), the sequences can then be mapped back to the *genome* (academic database of the whole DNA sequence and gene locations) to determine which gene the mRNA was derived from and thus infer what gene was being expressed in a particular sample of cells. RNA sequencing actually leads to a variety of interesting applications and has been instrumental in the identification of new genes, however this project focusses on the identification of changes in expression levels across two different biological samples.

Differential Gene Expression

The term differential expression simply refers to the difference in levels of expression between two genes. This is of interest to medical researchers because they can study cells in different stages of development, different states of healthy or under different treatments to identify genes with differential expression. This allows researchers to study the possible effects of particular genes and/or study the effectiveness of certain procedures in affecting the expression of particular genes. This type of study is very much central in medical research and currently many drugs combat diseases through influencing gene expression.

Hypothesis Testing

Hypothesis testing is a major branch of statistics specifically developed for the analysis of numerical observations, in particular the investigation of whether a statistically significant level of change has occurred. This requires a model of the underlying distribution of the sampled population and can precisely calculate the probability of observing what was observed given that some predetermined condition were true. In general we make the assumption that there is no difference between the samples, we then reject that hypothesis if the probability is sufficiently low for making the observations we did given the initial assumption were true.

Difficulty in Analysis

The mentioned statistical concepts are familiar to most scientists and actually the process of performing hypothesis testing is often rather straight-forward. The problem however is that RNAseq data is a relatively new form of data, in fact it follows no known distribution. The currently accepted model is to model the *counts* (number of RNAseq reads mapped to a particular gene) as Negative Binomial, however the actual counts are often over-dispersed meaning that the estimated variance is almost always lower than the true variance. This inaccuracy is acceptable when modelling the actual value of the counts, but when studying the difference between two groups of counts it is vital to accurately determine the natural variation of each group. This problem is partially addressed by the *voom* [1] procedure defined in the *limma* package for R however there is still ongoing research on more accurate estimation of variance for gene expression.

The second complication specific to such analysis is errors in multiple testing, when determining significance we usually use a 0.05 p-value, that is to say that under our null-hypothesis there is only a 0.05 probability of observing something at least as extreme as some predetermined cut-off point. But that does expose us to a 0.05 probability that even when the null-hypothesis were true, we observed something that caused us to reject it, this is known as a false-positive or false-discovery. This is usually not a problem because in most circumstances we can tolerate being wrong at a rate of 0.05, 0.01 is also a common value, some medical researchers or particle physicist calls for far more stringent p-values, CERN for example require a p-value of 3.0×10^{-7} to report a significant discovery.

The problem here is that we simultaneously test the expression of over thirty-thousand documented genes, meaning that even at a quite stringent 0.01 p-value threshold we'd be finding over 300 differentially expressed genes purely by chance. Simply lowering the p-value would lead to a greater possibility of failing to reject the null-hypothesis when it is false (these are known as false-negatives) which is sometimes equally undesirable. However this issue is addressed by one of the most cited statistical papers of all time by Benjamini and Hochberg which proposes the method of *False Discovery Rate* control [2], which effectively controls the rate of false discoveries without being overly stringent.

Analysis Procedure

In truth, all the ramblings of the previous section was not to provide particular insight but rather to highlight the extra complications in the analysis that prevent naive

applications of statistical methods. The full procedure for analysis can actually be performed by someone who isn't an expert at this particular family of data, all one would need is some introductory university level statistics and the ability to use R. The complexities of the analysis have been hidden away in the software packages RSubread [3], limma [4] and edgeR [5] for the R programming language, there is also plenty of available documentation that guides an interested user in how to use the contents of the package.

A typical analysis would start with the raw sequence reads coming out of a sequencing machine, the reads are then aligned to the genome and a table of counts is generated using functions from the RSubread package. The variance of the genes can then be estimated and linear models fit to each gene (corresponding to each row of the table), the end result is a table reporting various statistics including an adjusted p-value (adjusted for false-discovery rate) for each gene as well as few plots for visualising the data. There are a dozen specific techniques uniquely employed by the R packages, the theory for which are beyond the scope of this project, however it should be noted that the effectiveness of the methods compare very favourably to other currently used methods and limma is one of the most downloaded R packages for bioinformatics.

Even though the statistical complexity has been hidden away in a series of computer functions, to correctly call the functions and manipulate the data into a form suitable for the functions still requires moderate statistical knowledge and experience with R, neither of which should be expected from a medical researcher who has focussed all their work around understanding biology and biochemistry. It is therefore the job of statisticians to analyse the data which other researchers produce, this usually requires a significant time commitment, much of which spent on correspondence between the two parties. It would be unproductive to ask medical researchers to learn the statistics and R necessary to reliably analyse their own data but it's also quite unproductive for statisticians to use up a significant amount of their time performing routine analysis and waiting for feedback when they could be doing further research into better statistical methods, especially in an area as new and poorly understood as RNAseq data.

Graphical User Interface

The complications in using the statistical methods in R stems from the input of commands via a command line prompt. R, like almost every other programming language, is very sensitive to syntax and very minor errors can cause entire scripts to fail and are very hard to diagnose. However graphical user interfaces allow for inputs to be passed

from a more use-friendly interface to the command prompt, in effect the interface highlights the required inputs, restricts input types and potentially provides guidance for inputs. There are many available methods available for creating graphical interfaces such as javascript or Shiny in R; the web-based platform Galaxy was chosen for this particular project because it had many existing tools available for bioinformatics.

Galaxy

Web Platform

Galaxy is an open, web-based platform for data intensive biomedical research. [6–8] It hosts a variety of tools for performing many different procedures that would otherwise require users to have some knowledge of some programming language. Below is a screenshot of the main galaxy instance, the left panel shows all the available tools on that specific galaxy instance, the right panel shows the history of actions and data objects in your current Galaxy session, and the central area is where the tool interface appears.

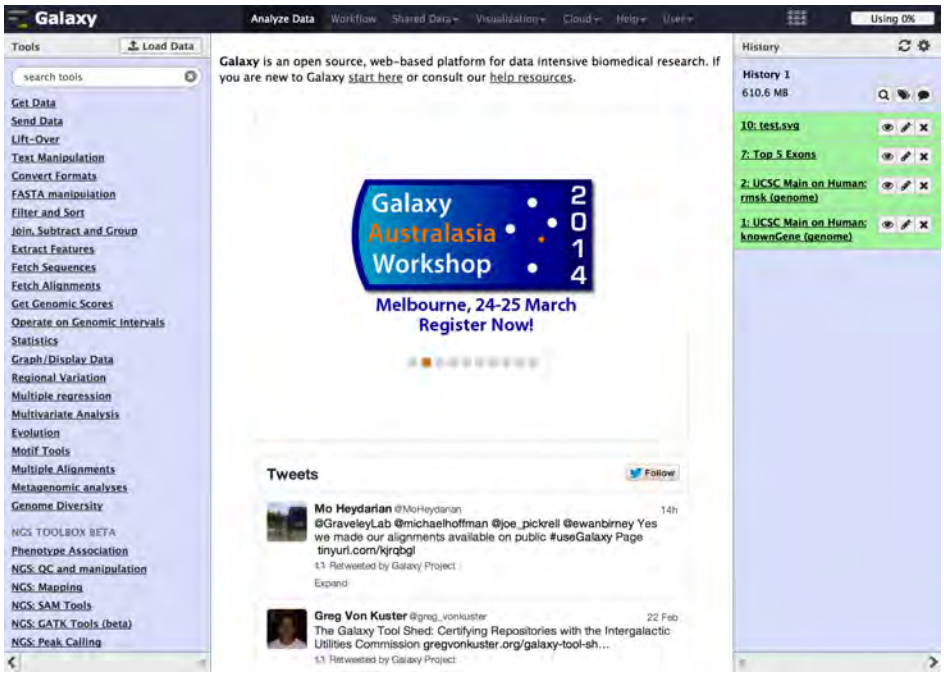


Figure 1: Screenshot of usegalaxy.org, the main public Galaxy instance.

from a more use-friendly interface to the command prompt, in effect the interface highlights the required inputs, restricts input types and potentially provides guidance for inputs. There are many available methods available for creating graphical interfaces such as javascript or Shiny in R; the web-based platform Galaxy was chosen for this particular project because it had many existing tools available for bioinformatics.

Galaxy

Web Platform

Galaxy is an open, web-based platform for data intensive biomedical research. [6–8] It hosts a variety of tools for performing many different procedures that would otherwise require users to have some knowledge of some programming language. Below is a screenshot of the main galaxy instance, the left panel shows all the available tools on that specific galaxy instance, the right panel shows the history of actions and data objects in your current Galaxy session, and the central area is where the tool interface appears.

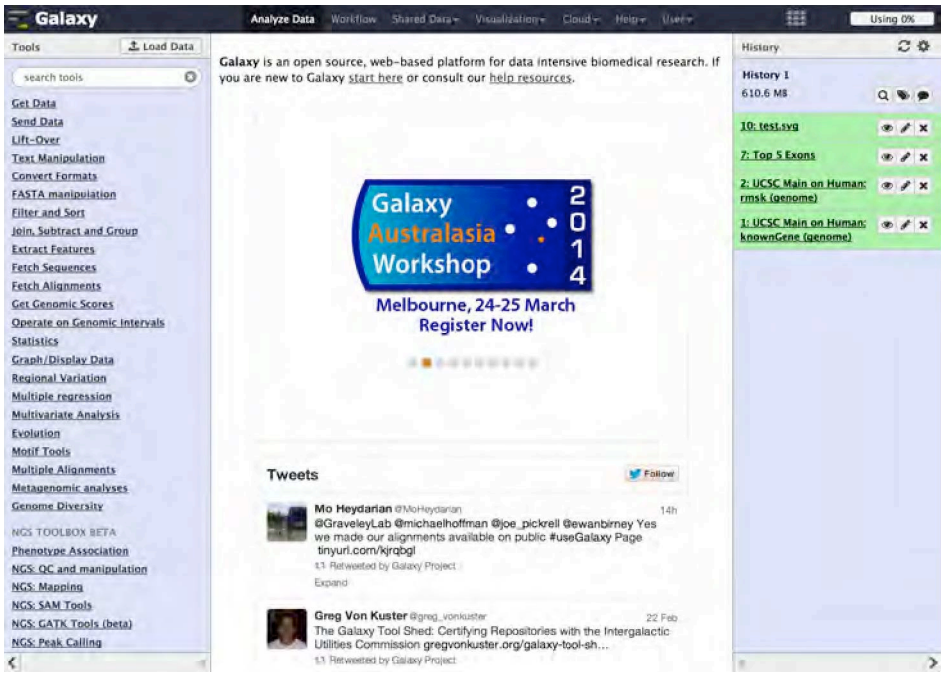


Figure 1: Screenshot of usegalaxy.org, the main public Galaxy instance.

The tools in Galaxy have their tool input interface specified in a XML file, the XML file also specified the command line prompt that is called once the tool is executed with all its inputs. The fact that Galaxy calls command line prompts means that it is able to make use of many interpreted programming languages, in this case it allowed for the use of R but most tools are written in Python or Perl which are equally powerful.

A particular benefit of such a system is that the entire Galaxy instance acts as a web service, one only has to manage the software on the server that Galaxy is hosted. All the tools can then be accessed through a web page without the need to install and maintain a range of software packages on multiple machines, something very useful for institutions where installation of new software is tightly managed.

Another extremely useful feature of Galaxy is the ability to save entire workflows, to retain both the output and all the inputs required to generate said output. Such a function is both incredibly useful for reproducing experimental results but also for performing similar types of analysis after a prolonged period of time. This was one of the key aspects of Galaxy which made it of interest in this project.

User Interface

The main purpose of Galaxy is to package complex workflows usually carried out in a script written with programming language into an interface with straightforward inputs. To make the tools more accessible, care was taken to minimise the amount of statistical terminology, names of statistical procedures were replaced with their purposes in biological analysis, and mathematical objects like design matrices for linear modelling were constructed by inferred information about the experimental groups. Annotations for each input was used whenever possible to guide the user and a more extensive description of the inputs is available on the bottom of the tool page.

Outputs

The Galaxy platform allow outputs in various file formats, many specific to the bioinformatics field. For this project the ability to output a HTML page was exploited to give more visually pleasing and insightful output compared to simple image files containing plots. As an example the *shRNAseq Tool* produces a HTML page containing various processing statistics, plots and tables. The processing statistics are there to provide users with feedback on what they input, the main purpose of this output is to help the user decide if the analysis performed was consistent with the intentions of the researcher.

Results and Discussion

The Tools

As a result of this project, four tools were produced for differential expression analysis. These are all available as open source code from https://github.com/Shians/diffexp. The tools *RSubread Mapping*, *Plot MDS* and *DiffExp* (names subject to change) are a part of a workflow that examines RNAseq data. The mapping component of this workflow is very intensive and tens of gigabytes of files are processed over at least several hours to produce the table of counts despite being many times faster than competing methods for alignment and counting.

The MDS plotting tool simply produces a multi-dimensional scaling plot that is essentially used for principal component analysis, the reason that it is an individual tool is because the MDS plot helps identify outliers which influences an input option in the DiffExp tool for sample weights. The DiffExp tool is the final step of the process, it essentially fits linear models to each row of the counts table based on the comparisons indicated by the user followed by statistical tests for significant differential expression. The result from the DiffExp tool is a plot highlighting the mean-variance trend of the log transformed count data, a plot showing the relationship between log-fold-change and average expression with statistically significant genes highlighted, and a table with the genes ranked in ascending p-value.

The shRNAseq tool performs very similar analysis on a slightly different type of experiment, the inputs can either be raw reads from sequencing machines or precomputed counts tables, it combines all the other steps into one tool because alignment for shRNA takes comparatively little time and no weights are used in the analysis.

Interface

The interface for the tools is defined in an XML file which has a style very similar to HTML files with tag structure. The essential sections are 'tool', 'inputs', 'outputs' and 'help'. The 'tool' section provides information about the tool such as version number and a unique identifier for the tool, the 'inputs' section is essentially the layout for the interface that will be presented to the user. As an example the code:

```
<inputs>
<param name="contrast" type="text" size="30"
            label="Contrasts of interest (No spaces)"
            help="Eg. Mut-WT,KD-Control"/>
</inputs>
```

Would produce the following input section:

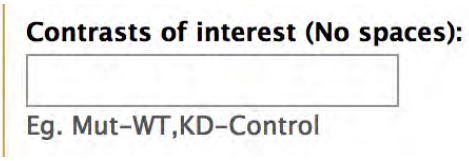


Figure 2: Contrast input section of DiffExp tool.

Note that the type of this particular input is set to 'text', a range of other types exist for different situations. The figures below illustrate a few of the input methods available as well as the ability to collapse sections of inputs based on the value of another input, a feature that is very useful for keeping the interface clean and not overwhelming.

Figure 3: Small section of DiffExp tool interface with collapsed cpm filtering options.

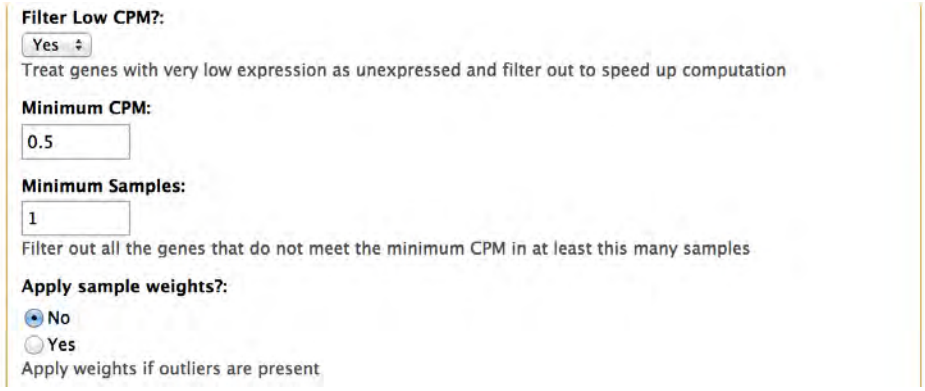


Figure 4: Small section of DiffExp tool interface with expanded cpm filtering options.

DiffExp Pipeline

The essential tool in the analysis of RNA differential expression is the DiffExp tool, the inputs required are essentially a table of counts and knowledge of the experimental groups of each sample. The table of counts can be from a variety of sources currently available for generating such information from raw reads, but the RSubread Mapping tool was created as one of the methods, it is currently limited to mouse and human genomes. The RSubread mapping tool takes the raw reads in as FastQ files and outputs the table of counts and the gene annotations.

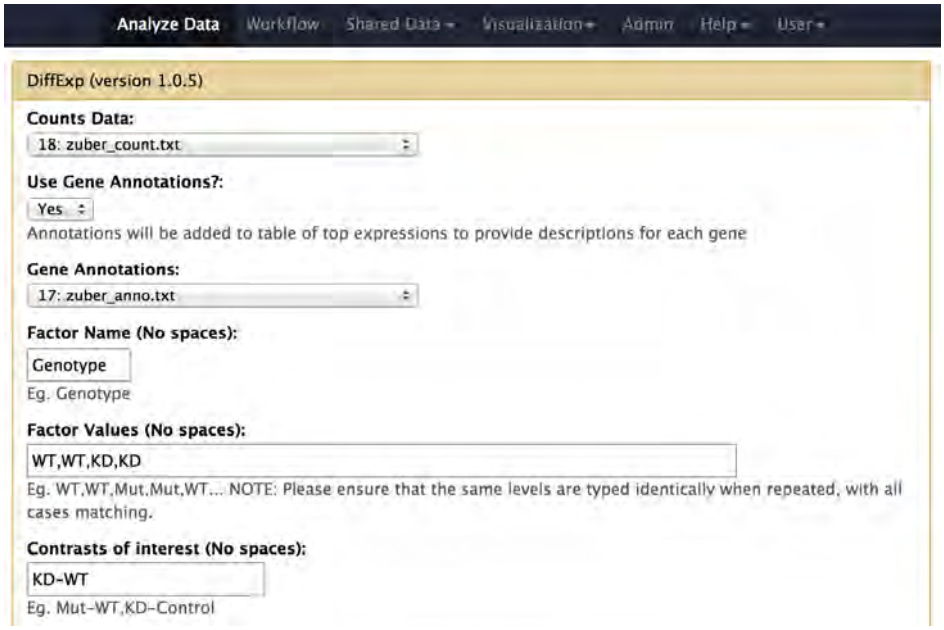
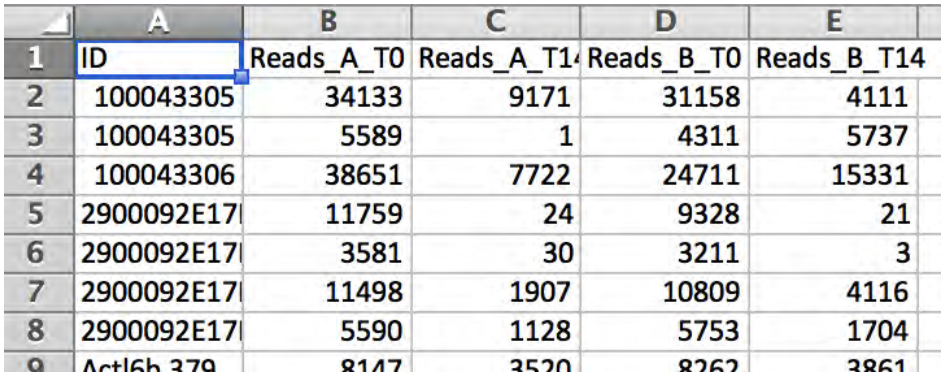


Figure 5: The required inputs for the DiffExp tool.

One may wish to attach an annotation file for the genes to make the resulting output table human-readable. The 'Counts Data' input is a simple table saved in a tab delimited format with each column representing a sample and each row representing a gene, the cells each represent the counts of some gene in some sample (The data below is actually shRNA data but the two types of data are visually identical). [9]



	A	B	C	D	E
1	ID	Reads_A_T0	Reads_A_T1	Reads_B_T0	Reads_B_T14
2	100043305	34133	9171	31158	4111
3	100043305	5589	1	4311	5737
4	100043306	38651	7722	24711	15331
5	2900092E17	11759	24	9328	21
6	2900092E17	3581	30	3211	3
7	2900092E17	11498	1907	10809	4116
8	2900092E17	5590	1128	5753	1704
9	Actl6b 379	8147	3520	8262	3861

Figure 6: Counts data from Zuber et al. (2011) in table form.

```
ID      Reads_A_T0      Reads_A_T14     Reads_B_T0      Reads_B_T14
100043305.2     34133   9171    31158   4111
100043305.4     5589    1       4311    5737
100043305.5     38651   7722    24711   15331
2900092E17Rik.377       11759   24      9328    21
2900092E17Rik.546       3581    30      3211    3
2900092E17Rik.1051      11498   1907    10809   4116
2900092E17Rik.1361      5590    1128    5753    1704
Actl6b.379      8147    3520    8262    3861
```

Figure 7: Counts data from Zuber et al. (2011) in tab delimited text form.

This is followed by some additional options that can be left on their default values for most analysis. The reason the MDS Plotting tool is not integrated with the DiffExp tool is due to the 'Apply Sample Weights?' option shown in the figure below. The MDS plot is useful for visualising the 'distance' between samples, usually referred to as principal component analysis, the presence of outliers then influences the usage of sample weights in analysis.

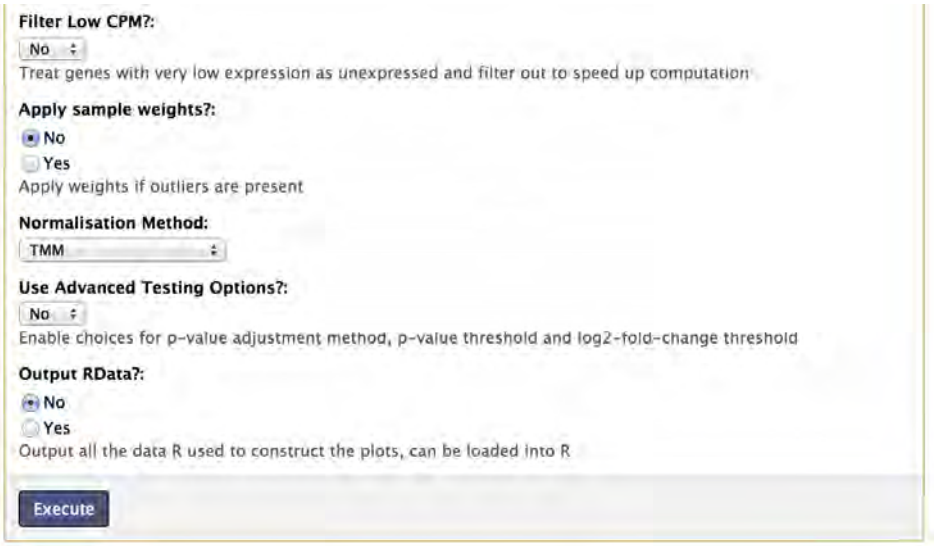


Figure 8: Additional options in DiffExp tool.

Once all inputs are set, the tool can be run and the output will be a HTML page containing the results of the analysis. The HTML output is usually used as a method of handling multiple outputs as such inputs can quickly fill up the history tab of Galaxy, however in this tool the flexibility of HTML is exploited to give more visually pleasing and informative output than plain figures. Extra information is provided in addition to the plots to give the user a better sense of their analysis, something that will hopefully be appreciated by new users and requires little extra effort as the information is readily generated by the R script. Future additions to the tool may provide links to instructions on how to interpret the plots as well as possible error checking mechanisms to warn users of potential problems with the analysis specified.

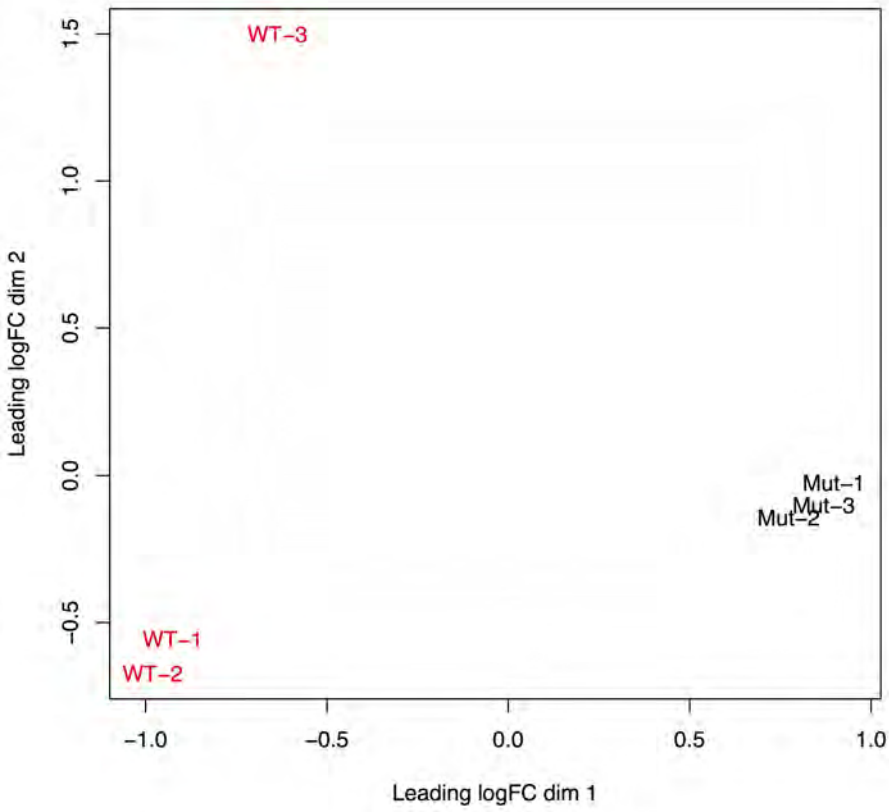


Figure 9: Output of MDS Plotting tool, shows that there is a potential outlier in this particular set of samples.

The figure above was generated using the MDS Plotter tool, it uses a variation of principle component analysis to produce a plot of the 'distance' between samples. The figure below is a mean-variance trend plot, it is a produced by *Voom* which is the primary program algorithm used in this analysis. The plot shows the trendline used to estimate the variance of each gene that was used in the hypothesis testing.

Limma Analysis Output:

All images displayed have PDF copy at the bottom of the page, these can exported in a pdf viewer to high resolution image format.

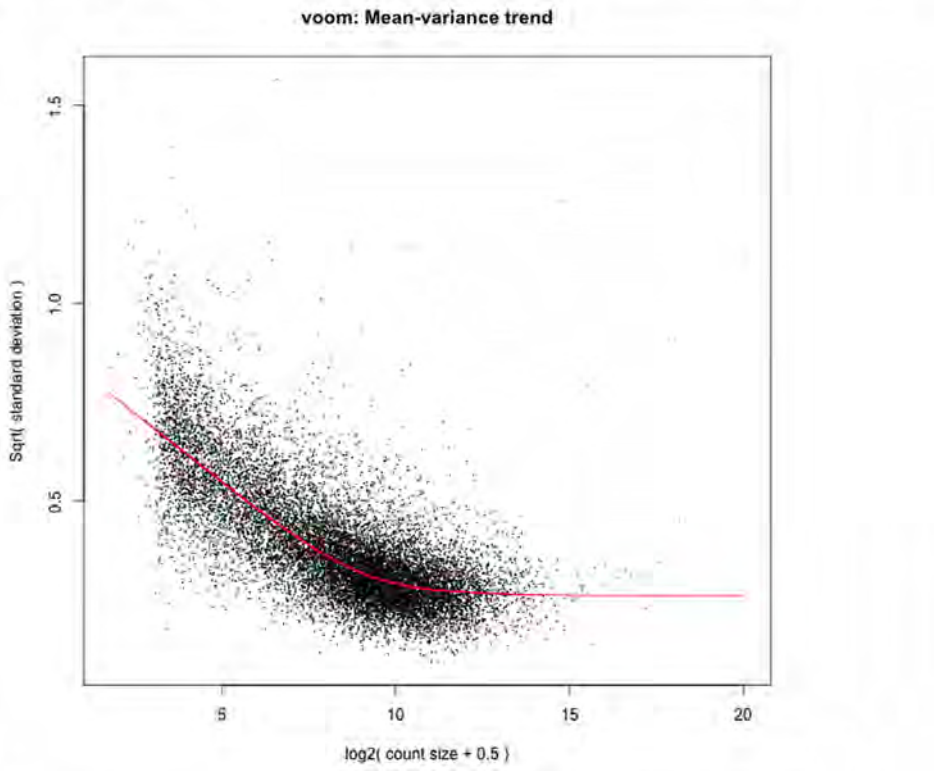


Figure 10: Part one of DiffExp HTML output.

Following the mean-variance plot is the MA plot which highlights the relationship between the degree of change measured as log-2-fold-change and the level of expression measured as log-2-counts-per-million. This plot also highlights the genes that are significant for differential expression, giving a quick visual indication of the degree of change experienced by the genes and the number of genes that are statistically significant.

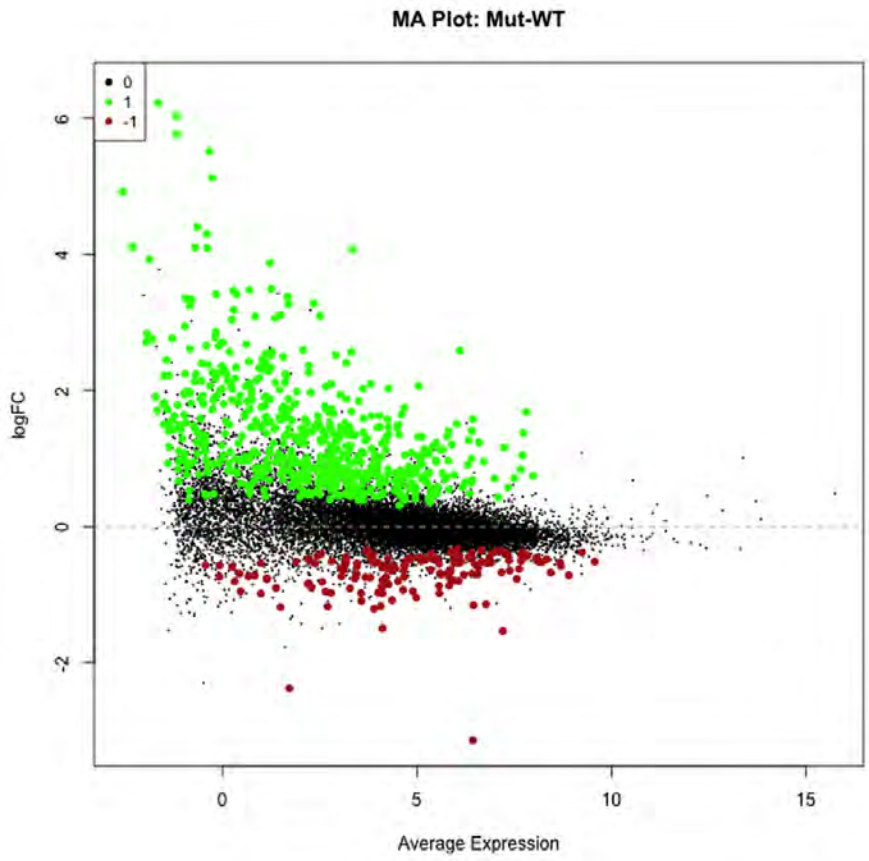


Figure 11: Part two of DiffExp HTML output.

Finally there are links to pdf versions of each graph which provides the quality necessary for use in published reports and most importantly the top expressions table. The table contains all the genes that were not filtered out and ranks them in ascending p-value, so the genes that are most likely to have been differentially expressed are at the top. This table is the main piece of information biologists study as it gives specifics on the level of expression of each gene, the amount of change experienced between the experimental conditions and direction of change. The goal in studying these table is to either associate genes with particular disorders or to determine the effectiveness of a particular treatment, being able to study all genes at once is immensely useful in many situations.

Plots:

Voom Plot (.pdf)
MA Plot(Mut-WT) (.pdf)

Tables:

Top Expressions Table(Mut-WT) (.tsv)

alt-click any of the links to download the file, or click the name of this task in the galaxy history panel and click on the floppy disk icon to download all files in a zip archive.

.tsv files are tab seperated files that can be viewed using Excel or other spreadsheet programs

Extra Information

- Genes that do not have more than 0.5 CPM in at least 3 samples are considered unexpressed and filtered out.
- TMM was the method used to normalise library sizes.
- Weights were applied to samples.
- MA-Plot highlighted genes are significant at adjusted p-value of 0.05 by the BH method, and exhibit log2-fold-change of at least 0

Summary of experimental data

	Genotype
NPC_RNA_113_C26VPACXX_ATCACG_L004.sam	WT
NPC_RNA_114_C26VPACXX_CGATGT_L004.sam	WT
NPC_RNA_11_C26VPACXX_GATCAG_L004.sam	Mut
NPC_RNA_12_C26VPACXX_TAGCTT_L004.sam	Mut
NPC_RNA_13_C26VPACXX_GGCTAC_L004.sam	Mut
NPC_RNA_14_C26VPACXX_CTTGTA_L004.sam	WT

Figure 12: Part three of DiffExp HTML output.

There is a summary of information used in this analysis, this is important as the tool itself is quite robust and will successfully produce output in many situations, it is therefore up to the user to carefully check that the specifications of the analysis are correct before they use the results generated.

Future Plans

Currently the tools are available on a locally hosted Galaxy instance at the Walter and Eliza Hall Institute, however after they are thoroughly tested and accompanying tutorials created, the tools will be released to the public Galaxy tool repository available for any Galaxy instance to install and use. The tools currently only work for single factor simple experiments, however this may be extended in the future to cover both multi-factor experiments and time-scale factor experiments. Finally due to the style of the R script, which takes arguments from the command line, it would be possible to make the tool portable into a variety of different graphical interfaces if necessary, certainly interaction with Javascript and the creation of interactive graphics is being looked into.

Final Words

This project has been successful in creating simple to use tools for any motivated biologist to perform sophisticated statistical analysis on simple experiments for differential expression. The Galaxy platform was found to be quite useful for producing an interface to run powerful R programs in a user-friendly manner, also the ability to save entire workflows was very attractive for allowing reproducibility of scientific results. The ability in Galaxy to present the output in a HTML file also provides additional presentation enhancements over what R usually provides, improving user-friendliness and making Galaxy useful beyond a simple input parser. A paper titled *'shRNA-seq data analysis with edgeR'* has also been submitted to *BMC Genomics*, the paper primarily details the usage of the edgeR package for shRNAseq analysis but suggests the Galaxy tool as a method of performing the analysis without interacting with the R programming language.

The tools created in this project are currently being used internally at the Walter and Eliza Hall institute, the shRNAseq tool has been released onto the public Galaxy tool repository and is being tested by a few bioinformaticians in Vienna. The successful introduction of these tools to medical researchers will provide much needed methods for performing their own statistical analysis, wide adoption of such methods will hopefully improve the reproducibility of scientific results. Consequently this will lighten the workload for statisticians and providing them with much needed time to work on the theory driving the statistical methods.

Additional Reading

The primary focus of this project was to provide access to sophisticated statistical methods for medical researchers in a manner that is accessible and does not require interaction with programming languages. Therefore the mathematical methods were not described in any detail, however it must be stressed that availability of the methods as well as the functional packages for applying the methods were absolutely crucial in this project.

Distribution of Counts

The previous generation of technology for performing differential analysis was using microarrays. Data generated by microarrays has the following probability density function:

$$Pr(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}.$$

Which means that microarray data is normally distributed, the most well studied probability distribution. Count data for RNAseq however follows a roughly poisson distribution, defined by its density function:

$$Pr(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k \in \mathbb{N}.$$

This is the distribution for a wide range of discrete random variables but empirical evidence suggests that count data for RNAseq is *overdispersed* meaning its variance is often higher than its mean, when they should be equal in a poisson distribution. The negative binomial distribution is currently the standard method for modelling overdispersed poisson distributions [10], it usually has the density function:

$$Pr(X = k) = \binom{k + r - 1}{k} (1 - p)^r p^k,$$

where $r > 0$, $p \in [0, 1]$ and $k \in \mathbb{N}$. However we use a slightly different parameterisation commonly used in regression for biological data where the density function is:

$$Pr(X = k) = \left(\frac{r}{r + m}\right)^r \frac{\Gamma(r + k)}{k!\Gamma(r)} \left(\frac{m}{r + m}\right)^k,$$

for $k \in \mathbb{N}$. This is the distribution used for generalised linear model fitting in shRNAseq with the edgeR package.

The DiffExp tool uses the voom function of the limma package which log-2 transforms the data such that its roughly normal, each row of the table can then have its variance estimated using the mean-variance trend shown in the output section. Then a linear model is fitted onto each of rows, the classic linear model is used, as defined by:

$$\mathbf{y} = \mathbf{X}\beta + \xi,$$

where $\mathbf{y}$ is the vector of observations, $\mathbf{X}$ is the design matrix, β is the vector of explanatory variables and ξ is a normally distributed random vector. The tests are then performed to check if one or more members of $E(\beta)$ is non-zero based on estimated variances. This is a newly proposed technique that focusses on the mean-variance relationship after modelling the transformed data as normally distributed, in contrast to all other methods which attempted to infer variance from negative binomial regression.

Acknowledgements

I would like to firstly thank my supervisor Matt Ritchie for providing me with this opportunity, his experience and guidance regarding the analytical procedures were vital in determining what was required in the interface. Additional thanks must be given to Keith Satterley and Toby Sargeant of the bioinformatics department at WEHI, their technical expertise were vital in both the development of the Galaxy tools and deployment of the Galaxy server. I also wish to express my appreciation for the assistance from Cynthia Liu and Andy Chen who helped explain elements of the statistical analysis done on the data.
It is necessary to give credit to all the developers of the RSubread, limm and edgeR packages whose amazing statistical packages I merely created a rudimentary graphical front-end for. Thanks to AMSI and CSIRO for providing funding for my work, and finally thanks to all members of the WEHI 6th floor whose willingness to discuss their work provided me with invaluable insight into medical research.

References

[1] Law, C. W., Chen, Y., Shi, W., and Smyth, G. K., 2014. *Voom! precision weights unlock linear model analysis tools for rna-seq read counts.* Genome Biology, 15(R29).

[2] Benjamini, Y. and Hochberg, Y., 1995. *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society, B57:289–300.

[3] Liao, Y., Smyth, G. K., and Shi, W., 2013. *The subread aligner: fast, accurate and scalable read mapping by seed-and-vote.* Nucleic Acids Research, 41(e108).

[4] Smyth, G. K., 2005. *Limma: linear models for microarray data.* In Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W., editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York.

[5] Robinson, M. D. and Smyth, G. K., 2010. *edger: a bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 26(139-140).

[6] Goecks, J., Nekrutenko, A., Taylor, J., and Team, T. G., 2010. *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biology, 11(R86).

[7] Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J., 2010. *Galaxy: A web-based genome analysis tool for experimentalists.* Current Protocols in Molecular Biology.

[8] Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., and Nekrutenko, A., 2005. *Galaxy: A platform for interactive large-scale genome analysis.* Genome Research, 15.

[9] Zuber, J., Shi, J., Wang, E., Rappaport, A. R., Herrmann, H., Sison, E. A., Magoon, D., Qi, J., Blatt, K., Wunderlich, M., Taylor, M. J., Johns, C., Chicas, A., Mulloy, J. C., Kogan, S. C., Brown, P., Valent, P., Bradner, J. E., Lowe, S. W., and Vakoc, C. R., 2011. *Rnai screen identifies brd4 as a therapeutic target in acute myeloid leukaemia.* Nature, 478:524–528.

Postal Address: 111 Barry Street
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 9349 4106
Web: www.amsi.org.au

[10] McCarthy, D. J., Chen, Y., and Smyth, G. K., 2012. *Differential expression analysis of multifactor rna-seq experiments with respect to biological variation.* Nucleic Acids Research, 40(10):4288–4297.

Postal Address: 111 Barry Street
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 9349 4106
Web: www.amsi.org.au

Would produce the following input section:

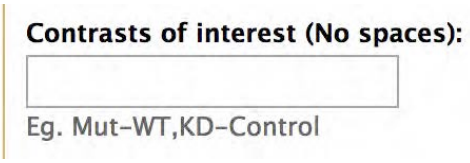


Figure 2: Contrast input section of DiffExp tool.

Note that the type of this particular input is set to 'text', a range of other types exist for different situations. The figures below illustrate a few of the input methods available as well as the ability to collapse sections of inputs based on the value of another input, a feature that is very useful for keeping the interface clean and not overwhelming.

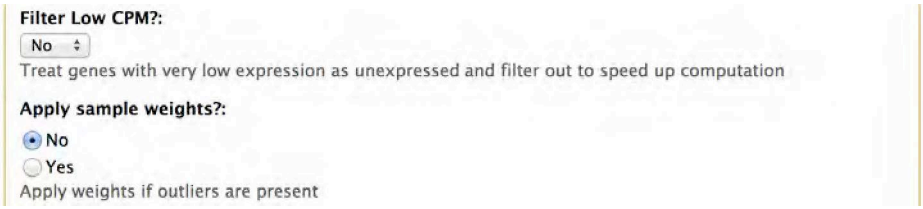


Figure 3: Small section of DiffExp tool interface with collapsed cpm filtering options.

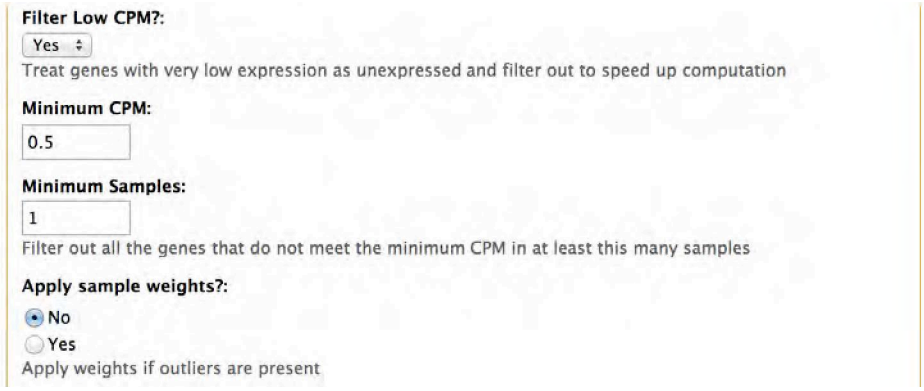


Figure 4: Small section of DiffExp tool interface with expanded cpm filtering options.

DiffExp Pipeline

The essential tool in the analysis of RNA differential expression is the DiffExp tool, the inputs required are essentially a table of counts and knowledge of the experimental groups of each sample. The table of counts can be from a variety of sources currently available for generating such information from raw reads, but the RSubread Mapping tool was created as one of the methods, it is currently limited to mouse and human genomes. The RSubread mapping tool takes the raw reads in as FastQ files and outputs the table of counts and the gene annotations.

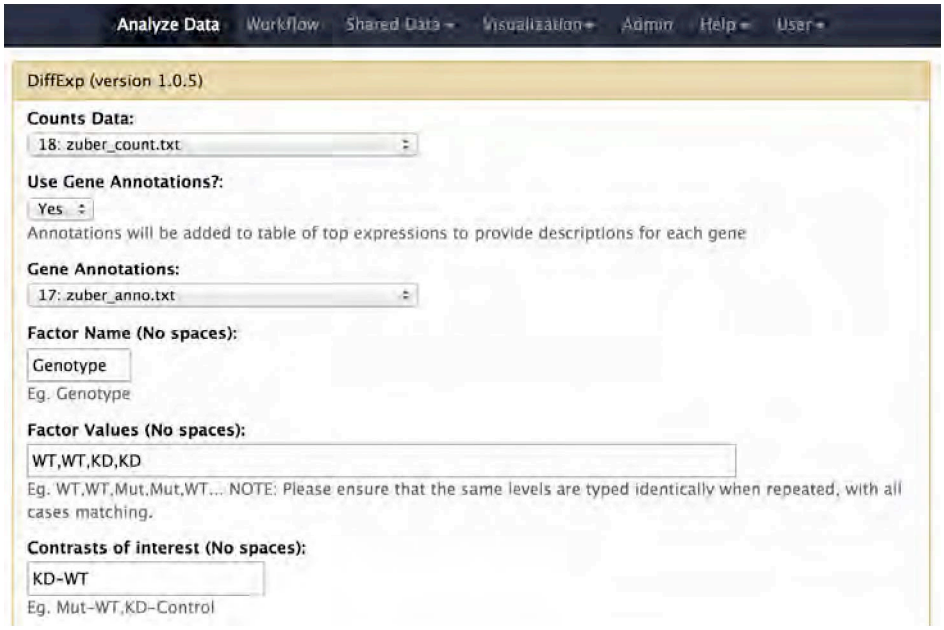
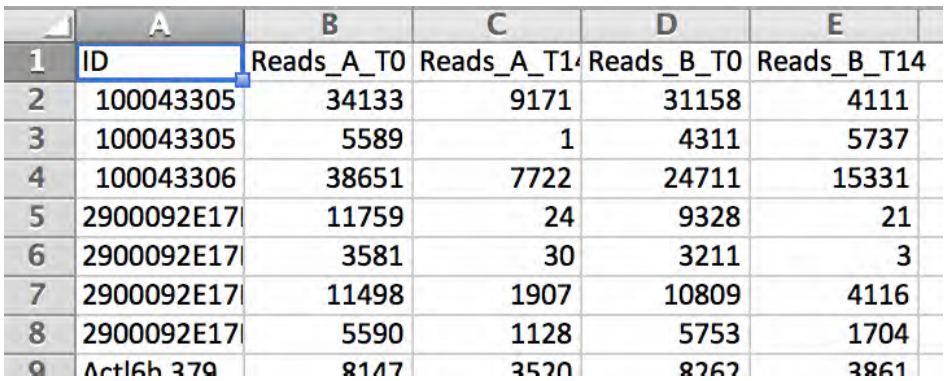


Figure 5: The required inputs for the DiffExp tool.

One may wish to attach an annotation file for the genes to make the resulting output table human-readable. The 'Counts Data' input is a simple table saved in a tab delimited format with each column representing a sample and each row representing a gene, the cells each represent the counts of some gene in some sample (The data below is actually shRNA data but the two types of data are visually identical). [9]



	A	B	C	D	E
1	ID	Reads_A_T0	Reads_A_T1	Reads_B_T0	Reads_B_T14
2	100043305	34133	9171	31158	4111
3	100043305	5589	1	4311	5737
4	100043306	38651	7722	24711	15331
5	2900092E17	11759	24	9328	21
6	2900092E17	3581	30	3211	3
7	2900092E17	11498	1907	10809	4116
8	2900092E17	5590	1128	5753	1704
9	Actl6b 379	8147	3520	8262	3861

Figure 6: Counts data from Zuber et al. (2011) in table form.

```
ID      Reads_A_T0      Reads_A_T14     Reads_B_T0      Reads_B_T14
100043305.2     34133   9171    31158   4111
100043305.4     5589    1       4311    5737
100043305.5     38651   7722    24711   15331
2900092E17Rik.377       11759   24      9328    21
2900092E17Rik.546       3581    30      3211    3
2900092E17Rik.1051      11498   1907    10809   4116
2900092E17Rik.1361      5590    1128    5753    1704
Actl6b.379      8147    3520    8262    3861
```

Figure 7: Counts data from Zuber et al. (2011) in tab delimited text form.

This is followed by some additional options that can be left on their default values for most analysis. The reason the MDS Plotting tool is not integrated with the DiffExp tool is due to the 'Apply Sample Weights?' option shown in the figure below. The MDS plot is useful for visualising the 'distance' between samples, usually referred to as principal component analysis, the presence of outliers then influences the usage of sample weights in analysis.

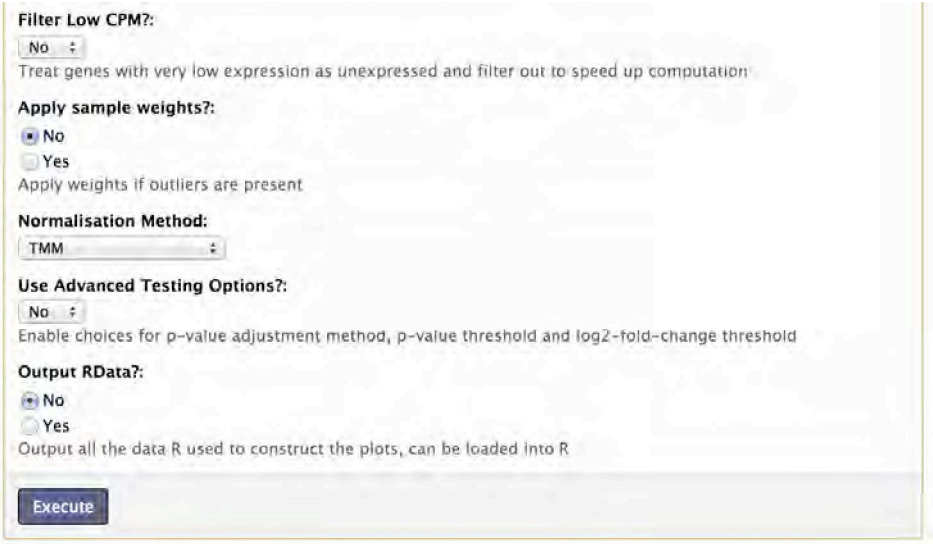


Figure 8: Additional options in DiffExp tool.

Once all inputs are set, the tool can be run and the output will be a HTML page containing the results of the analysis. The HTML output is usually used as a method of handling multiple outputs as such inputs can quickly fill up the history tab of Galaxy, however in this tool the flexibility of HTML is exploited to give more visually pleasing and informative output than plain figures. Extra information is provided in addition to the plots to give the user a better sense of their analysis, something that will hopefully be appreciated by new users and requires little extra effort as the information is readily generated by the R script. Future additions to the tool may provide links to instructions on how to interpret the plots as well as possible error checking mechanisms to warn users of potential problems with the analysis specified.

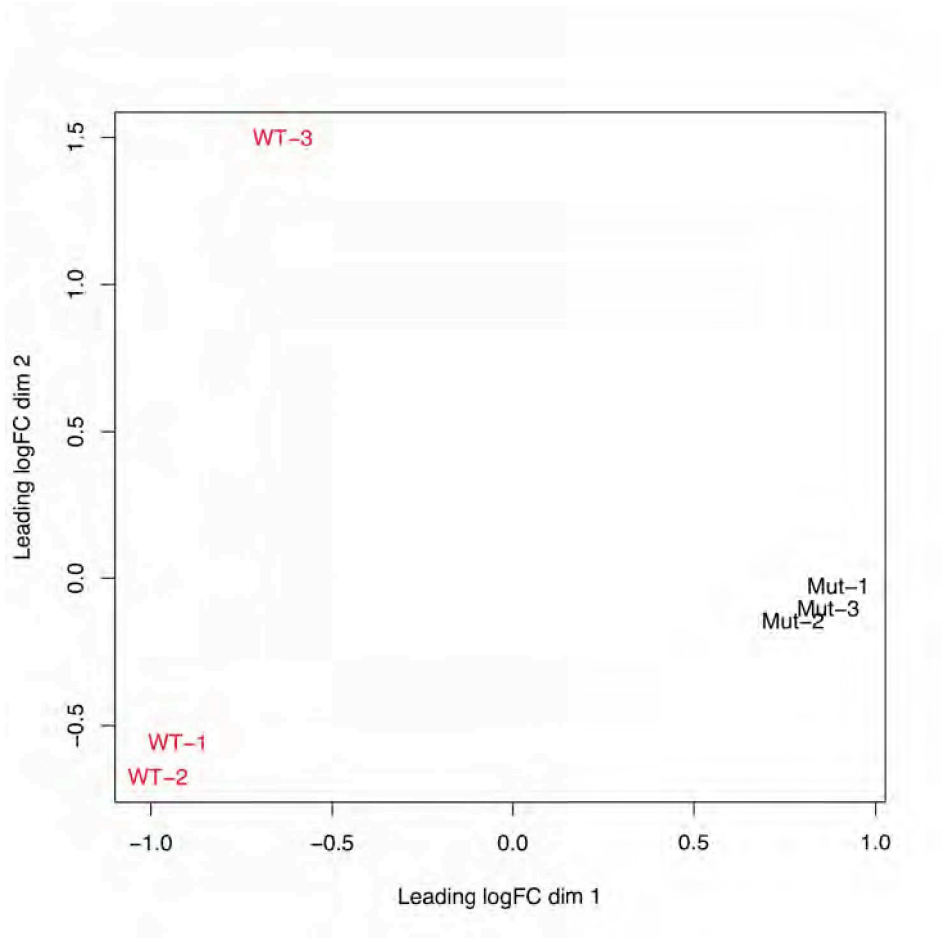


Figure 9: Output of MDS Plotting tool, shows that there is a potential outlier in this particular set of samples.

The figure above was generated using the MDS Plotter tool, it uses a variation of principle component analysis to produce a plot of the 'distance' between samples. The figure below is a mean-variance trend plot, it is a produced by *Voom* which is the primary program algorithm used in this analysis. The plot shows the trendline used to estimate the variance of each gene that was used in the hypothesis testing.

Limma Analysis Output:

All images displayed have PDF copy at the bottom of the page, these can exported in a pdf viewer to high resolution image format.

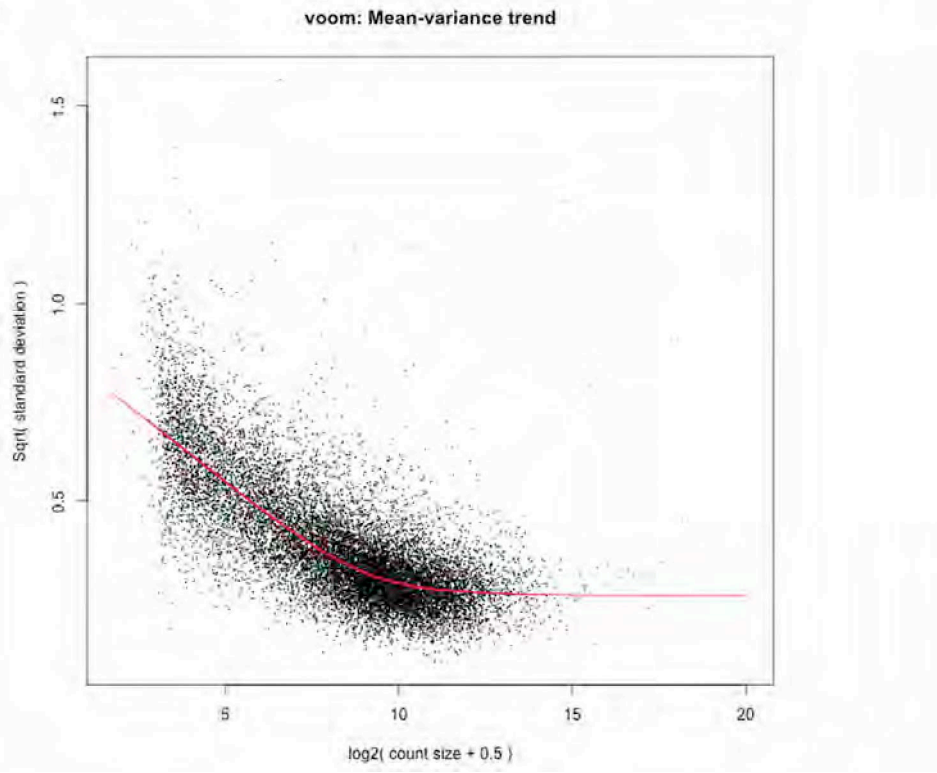


Figure 10: Part one of DiffExp HTML output.

Following the mean-variance plot is the MA plot which highlights the relationship between the degree of change measured as log-2-fold-change and the level of expression measured as log-2-counts-per-million. This plot also highlights the genes that are significant for differential expression, giving a quick visual indication of the degree of change experienced by the genes and the number of genes that are statistically significant.

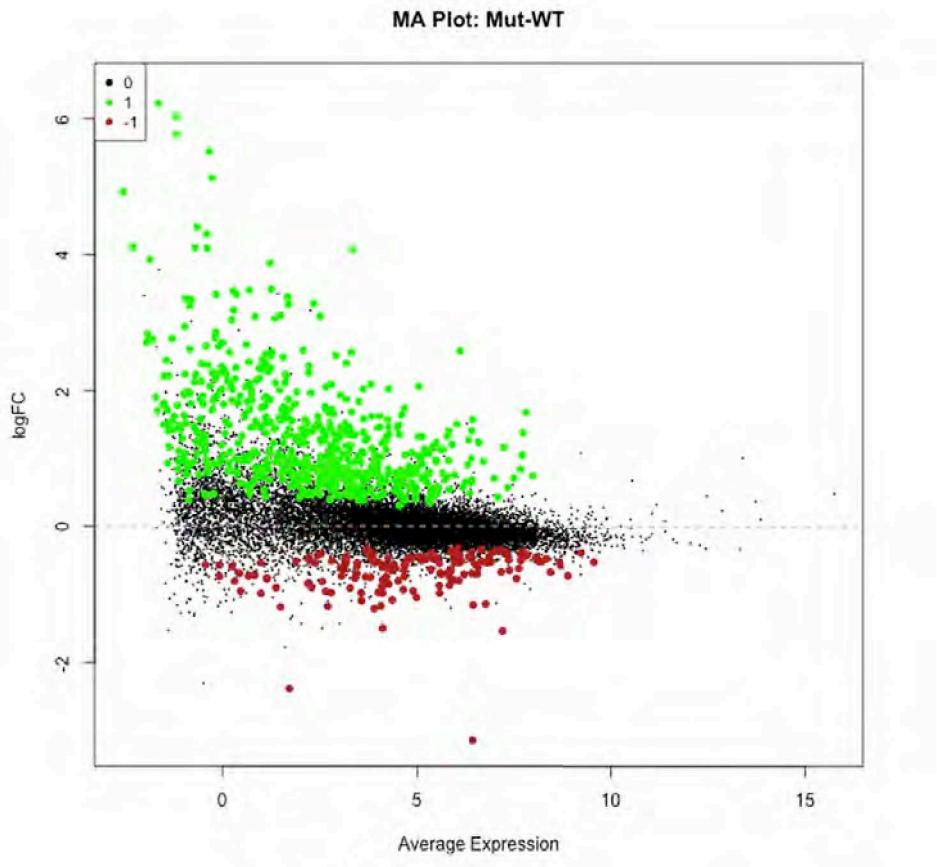


Figure 11: Part two of DiffExp HTML output.

Finally there are links to pdf versions of each graph which provides the quality necessary for use in published reports and most importantly the top expressions table. The table contains all the genes that were not filtered out and ranks them in ascending p-value, so the genes that are most likely to have been differentially expressed are at the top. This table is the main piece of information biologists study as it gives specifics on the level of expression of each gene, the amount of change experienced between the experimental conditions and direction of change. The goal in studying these table is to either associate genes with particular disorders or to determine the effectiveness of a particular treatment, being able to study all genes at once is immensely useful in many situations.

Plots:

Voom Plot (.pdf)
MA Plot(Mut-WT) (.pdf)

Tables:

Top Expressions Table(Mut-WT) (.tsv)

alt-click any of the links to download the file, or click the name of this task in the galaxy history panel and click on the floppy disk icon to download all files in a zip archive.

.tsv files are tab seperated files that can be viewed using Excel or other spreadsheet programs

Extra Information

- Genes that do not have more than 0.5 CPM in at least 3 samples are considered unexpressed and filtered out.
- TMM was the method used to normalise library sizes.
- Weights were applied to samples.
- MA-Plot highlighted genes are significant at adjusted p-value of 0.05 by the BH method, and exhibit log2-fold-change of at least 0

Summary of experimental data

	Genotype
NPC_RNA_113_C26VPACXX_ATCACG_L004.sam	WT
NPC_RNA_114_C26VPACXX_CGATGT_L004.sam	WT
NPC_RNA_11_C26VPACXX_GATCAG_L004.sam	Mut
NPC_RNA_12_C26VPACXX_TAGCTT_L004.sam	Mut
NPC_RNA_13_C26VPACXX_GGCTAC_L004.sam	Mut
NPC_RNA_14_C26VPACXX_CTTGTA_L004.sam	WT

Figure 12: Part three of DiffExp HTML output.

There is a summary of information used in this analysis, this is important as the tool itself is quite robust and will successfully produce output in many situations, it is therefore up to the user to carefully check that the specifications of the analysis are correct before they use the results generated.