Web-based Tools for the Analysis of DNA Microarrays

End of Project Report

Project 5236

Authors: P. Geeleher[1,2], A. Golden[3], J. Hinde[2] and D. G. Morris[1]

[1]Teagasc, Animal Reproduction Department, Mellows Livestock Science Centre, Athenry, Co. Galway

[2]School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway.

[3]National Centre for Biomedical Engineering Science, National University of Ireland, Galway

Table of contents:

Summary

DNA microarrays are widely used for gene expression profiling. Raw data resulting from microarray experiments, however, tends to be very noisy and there are many sources of technical variation and bias. This raw data needs to be quality assessed and interactively preprocessed to minimise variation before statistical analysis in order to achieve meaningful result. Therefore microarray analysis requires a combination of visualisation and statistical tools, which vary depending on what microarray platform or experimental design is used.

Bioconductor is an existing open source software project that attempts to facilitate analysis of genomic data. It is a collection of packages for the statistical programming language R. Bioconductor is particularly useful in analyzing microarray experiments. The problem is that the R programming language's command line interface is intimidating to many users who do not have a strong background in computing. This often leads to a situation where biologists will resort to using commercial software which often uses antiquated and much less effective statistical techniques, as well as being expensively priced. This project aims to bridge this gap by providing a user friendly web-based interface to the cutting edge statistical techniques of Bioconductor.

The analysis tools that we have constructed facilitate straightforward analysis of microarray data in a web-based environment, addressing the most widely used microarray platforms and following a logical progression through an analysis pipeline that is both extensible and capable of addressing current needs.

The initial scope of this project primarily focused on analysis of Affymetrix GeneChip arrays. However the facilities for basic analysis of dual dye cDNA arrays and single dye Exiqon miRNA arrays have also been implemented and provide a solid foundation for future development.

BioconductorBuntu is a custom distribution of Ubuntu Linux that wraps the analysis tools developed by the project in an easily installable and distributable format. The server is setup by running a very straight forward installation CD. The system is best installed on a dedicated server, allowing any individuals connected to the same network to make use of the analysis tools hosted on the server.

Introduction

Eukaryotes are organisms whose cells are organized into complex structures enclosed within membranes. Most eukaryotic organisms, for example, human beings, contains billions of individual cells. Almost all of these cells contain, within each nucleus, the entire genome for that organism. This genome contains the organism's complete hereditary information in the form of deoxyribonucleic acid (DNA), that encodes a complete blueprint for all activities and structures within the organism.

In the human body, the genome consists of 23 pairs of chromosomes. One of each of this pair is inherited from the mother and the other from the father. Each chromosome is made of chains of DNA. DNA consists of two polymers made up of units called nucleotides. Each nucleotide consists of a deoxyribose sugar, a phosphate group and one of the four nitrogen bases, guanine, adenine, thymine and cytosine. These bases, which are usually represented by their first letters, G, A, T and C, are where hereditary genetic information is actually encoded. It is worth noting that one of the two strands of the DNA double helix will suffice to describe this information; this is because of complementary base pairing, whereby an A on one strand always binds to a T on the other and a C always binds to a G.

Genes are essentially segments of the DNA structure described above. Loosely speaking, a gene is a section of DNA that defines a single trait by encoding a particular pattern, about 27,000 of which exist in humans. Often though we are faced with a case where protein-coding sequences have no clear beginning or end; more technically, a gene is a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, also known as exons, transcribed regions, also known as introns and/or other functional sequence regions.

The main purpose of genes is to act as a blueprint in the creation of proteins. Proteins are made of amino acids and are responsible for the structure and activity of an organism at a cellular level. They are created as follows; starting at the 5' end (the leading end) of a gene and proceeding to the 3' end (the tail end), the information contained in the gene is transcribed into a messenger ribonucleic acid (mRNA) strand. This process is performed by an enzyme called RNA polymerase.

After transcription this mRNA molecule leaves the nucleus of the cell where it is transcribed into a protein in a process called translation. This is performed by ribosomes, which read the code carried by mRNA molecules from the cell nucleus and create proteins combining any of the 20 amino acids in the body into complex polypeptide chains. These proteins are the building blocks of the organism. This process of translating a gene into a functional product is known as gene expression.

DNA microarrays are a high throughput technology used to measure the expression levels of thousands of genes, in some cases all of the genes in a genome, simultaneously. The fundamental idea behind most microarrays is to exploit complementary base pairing (see previous section) to measure the amount of the different types of mRNA molecules in a cell, thus indirectly measuring the expression levels of the genes that are responsible for the synthesis of those particular mRNA molecules.

The spots on a microarray contain single stranded DNA oligonucleotides called probes. Each of these spots will contain DNA which is of a complementary sequence to the specific mRNA molecule that corresponds to the gene that it is targeting. An mRNA molecule which is complementary to the probe in question, should hybridise to that probe, forming a strong mRNA-DNA bond. These mRNA molecules will have previously been labelled with fluorescent dye, which means that the amount of hybridisation that has taken place can be measured by the level of fluorescence of the dye, which is examined with a scanner. This scanner then outputs a text file for each array, which contains the relevant data pertaining to that array, such as the level of fluorescence of each spot and the level of background noise. It is these text files which are subsequently computationally analysed. In theory, a spot with brighter fluorescence means that more mRNA has hybridised, which in turn infers that more mRNA was present in sample extracted from the original cell and that the gene represented by this spot is experiencing a higher level of expression.

The types of DNA microarrays most widely used today can be broadly divided into two categories, cDNA arrays and oligonucleotide arrays. The GeneChip, which is manufactured by Affymetrix, is an oligonucleotide array and is the most commonly used type of DNA microarray. They differ slightly in operation from other kinds of arrays. Each array will contain hundreds of thousands of probe spots and each of these spots will in turn contain millions of copies of an individual 25 base long DNA oligonucleotide (Figure 1)

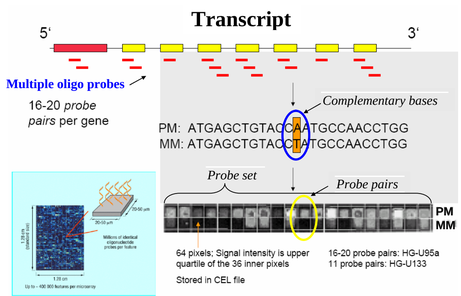

Figure 1 Affymetrix Genechip. Perfect match (PM) and mismatch (MM) probe pairs make up the set of probes representing a gene transcript. The middle base of the MM probes differs from the PM probe.

Each gene that is being targeted is represented by typically (but not necessarily always) by 11 pairs of these probes. This set of probes contains 11 perfect match (PM) probes, which are exactly complementary to the DNA sequence of a subset of 25 bases of the target gene. Each PM probe has a corresponding mismatch probe (MM), which contains the same 25 base long sequence as the PM probe, except for the fact that the middle base, or the 13th base in the chain, is substituted for the complement of the 13th base of its corresponding PM probe; so for example, a G in the 13th base of a PM probe will be replaced with a C in the MM probe. This is meant to give an estimate of non-specific binding, which occurs when mRNA that is not targeted binds to a PM probe.

cDNA microarrays differ from Affymetrix arrays in that each spot corresponds entirely to a specific gene. Sometimes duplicate spots will target the same gene, but these spots are exact copies of each other. The probes are of varying length but are generally hundreds of bases long. Instead of mRNA levels being directly measured, these arrays measure complementary DNA (cDNA), because this is more stable molecule than mRNA at these large sizes. mRNA from the original sample is reverse transcribed in a laboratory to create

an equivalent number of the more stable cDNA molecules which are then hybridised to the microarray.

These cDNA molecules are usually more than 500 bases long. Each of the probes contained on the spots on the microarary will be complementary to a cDNA molecule that represents a given gene. Thus, the measure of how much cDNA binds to its corresponding spot gives an accurate measure of the expression level of the gene in question.

Instead of expression levels of an individual sample being measured directly, two separate samples are hybridised to the same array at the one time. One of these samples is generally a control sample, while the other one is a sample of interest such as tumour tissue. Each of these samples is labelled with a particular dye; either a red fluorescent dye, Cyanine 5 or Cy5, or a green-fluorescent dye, Cyanine 3 or Cy3. When the array is read by the scanner, the differential expression level of a given gene is measured by the difference in intensity level between the red and green channel, at the spot that corresponds to the gene in question. cDNA microarrays are initially read by a scanner, which produces a TIFF image of the array. These images are then interpreted by one of a number of image analysis software packages, all of which output data in slightly different formats. This system supports analysis of data from the major platforms, including Spotfire, GenePix, BlueFuse and Agilent.

Microarrays can also be used for detection of miRNA expression levels. miRNA are short RNA molecules, generally about 22 nucleotides in length. They are encoded in genes but are not translated into proteins; instead, these molecules down regulate the expression of certain genes. They achieve this by being complementary to specific mRNA molecules created in a cell. The miRNA molecules bind to the complementary sections of these mRNAs and stop them from being converted into proteins.

Exiqon manufacture microarrays for detection of miRNA expression. The spots on these microarrays consist of Locked Nucleic Acid (LNA) probes. LNA is a modified RNA nucleotide that, because of the short length of miRNA molecules, forms a more stable bond with miRNA than standard DNA probes meaning that accuracy of measurements is increased. The miRNA molecules that are being targeted will bind to its complementary LNA probe. Other than this the processes of labelling the sample with a fluorescent dye,

hybridising the sample to the array and reading the hybridisation levels with a scanner are similar to those of other arrays.

Description of Functionality

The web based tool described in the following sections can be accessed either on http://europa.it.nuigalway.ie/cgi-bin/login.py or http://10.12.1.45/cgi-bin/login.py and these sections are best read in conjunction with the online tool.

Following connection to the server via a web browser, each user must create an account on the system via the 'Register' link, uploading their relevant details. After registration, the user may freely thereafter access the system and upload raw data. Uploaded data are assumed to be in a '.zip' archive, whose data files are unpacked and stored server-side prior to analysis. Currently, a user has three options, to upload Affymetrix R, dual dye or single dye data. At time of upload, users creating Affymetrix R experiments have the option of using Affyprobeminer's remapped CDF files instead of the default Bioconductor files. An entry, containing all relevant information is created in the database for the new experiment. As data are being uploaded and processed, progress is printed on the screen, allowing users to monitor this procedure in real time. Following data upload the next step is generally to assign the phenotypic data for the experiment. If the user is analysing Affymetrix R data, they are first presented with a screen that requires them to specify the number of factors for the experiment and the number of levels of each of these factor; as well as names for the levels and factors. Next they must assign the appropriate levels of each factor to each array. Design and contrasts matrices are created automatically from this information, allowing great flexibility in experimental design. For dual dye and single dye arrays, currently only experiments with 2 RNA sources are supported. The next stage is quality control, to ensure that data are of sufficient integrity. There are a number of options available here such as boxplots, histograms, PCA and many more diagnostic plots on normalised and unnormalised data. Several of the most commonly used normalisation options are also available. For dual and single dye data the user may also select which spot types are to be included, if a spot types file was specified at upload. There is also support for within array duplicate spots. Based on the output of quality control, it is up to the user to decide if some or all of the arrays need to be redone or removed from differential expression analysis. The 'Data Select' link allows the removal of arrays from the dataset, if it this is deemed

necessary.

The next and most significant step is differential expression analysis. Using Affymetrix R data, the user has the option of using the Limma or PUMA packages to perform this analysis. Under limma a number of preprocessing methods, such as RMA, GCRMA, MAS5 and custom methods, specifying the steps taken for normalisation, background correction, PM correction and summarisation, are available. If the user opts to use the PUMA package, data are preprocessed using PUMA's multi-mgMOS method, which implements global median scaling normalisation. The expression level information from the different arrays of each condition is then combined. Differential expression is then calculated using the Probability of Positive Log Ratio (PPLR) method. Because PUMA analysis can be a time-consuming process, even on relatively small datasets, notification of completion is emailed to users upon conclusion. For dual dye analysis the system incorporates all of normalisation methods that are available in the Limma package. There are a number of different options for normalisation between arrays, normalization within arrays, and background correction. All of these options are available at both Quality Control and Differential Expression stages. Single dye data are normalised using VSN (Huber et al., 2003) as this is the only viable option available to single dye data within Bioconductor.

Upon completion of differential expression, a ranked gene-list, containing such metrics as fold change, p-values, adjusted p-values and B-statistic is output to the viewing screen. This list contains links to gene annotation information which is retrieved using the Bioconductor package BiomaRt and is dynamically loaded using AJAX. This ranked gene-list may also be downloaded as an Excel R file. All quality control and differential expression analysis, including all files generated, are stored in the users profile for future reference. The user also has the option to download the .Rdata file containing the complete R workspace environment for a particular analysis procedure. This file can be used if they wish to continue analysis from the command line, which may be more flexible in certain situations and useful to advanced users. This also offers beginners an opportunity to study command line interactivity with the Bioconductor platform. Users may also manage their account by deleting previous experiments and analyses.

Outline of Backend Technologies

The project is entirely based on open source software which is free to redistribute, use and alter. Ubuntu Linux is the operating system upon which the server is built. Ubuntu was chosen for several reasons, one of which is that it is the most commonly used Linux distribution, meaning it will be familiar to a larger user base, as well as there being a large support base. The distribution was customised using an open source script called remastersys, which allows its users to make changes to a basic Ubuntu installation and save them as a remastered install CD. All analysis of gene expression microarray data are performed in R, specifically using Bioconductor. This distribution is available to download as “Bioconductor Buntu” from **http://bioinf.nuigalway.ie**. Pipelines facilitating the analysis of other microarray platforms are also possible to the user community by reconfiguring BioconductorBuntu accordingly. The CGI scripts which run on the server are written in Python. These scripts handle input and output to and from the xHTML, CSS and JavaScript based user interface, as well as writing to and from the database and making calls to R and Bioconductor via the RPy interface; RPy being a robust Python interface to the R Language. The Apache web server is used and the powerful open-source database MySQL is used to store and retrieve various forms of data. Exim is used as a mail transfer agent and Devcot as an IMAP server. Other technologies used by the server include, Imagemagick for image manipulation and particularly creating thumbnail versions of image files and PhpMyAdmin for web-based database administration. Considering the large data footprints associated with typical microarray analyses, BiocinductorBuntu should ideally be installed on a high end server; centralising the processing of such large amounts of data on a high specification machine, particularly when running complex analyses, will lead to greatly improved performance compared to analysis on a standard desktop machine.

Using the System

The first thing a user must do is register with the system. This is achieved by clicking the "Register" tab on the top of the page which invokes the "register.py"script and handles this process. This page contains several JavaScript functions to authenticate that the form has been filled out correctly, verifying for example that the email address supplied is valid and all required fields have been completed.

If the forum is not properly filled out, an error dialogue box will inform the user and prevent submission of the page. Once registration is complete the user may now proceed to login for the first time. The first stage in creating an experiment is to upload the raw data. At this stage the user is presented with a page that displays options to upload 3 different types of data, Affymetrix, dual dye or single dye (Figure 2)

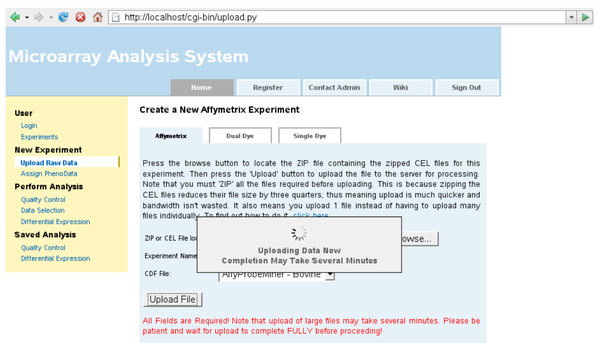

Figure 2. Screen shot showing uploading of data to Microarray analysis system

For all data types the user must package their raw data files in a zip archive. This can be accomplished with any number of freely available compression tools like 7-Zip or WinZip. Zipping the data has the double advantage of significantly compressing the data for faster upload and also means that only one file needs to be uploaded. The user must also specify an experiment name every time they upload a new dataset. There are some differences in the rest of the information supplied when uploading different types of experiments. For Affymetrix data, the user has the choice of using the default Affymetrix supplied CDF file or using the remapped CDF file from AffyProbeMiner. For Dual dye data, the user may wish to specify a “Spot Types” file. This file contains information which describes the function of certain spots on the arrays and can be used to identify, for example, control signals. Depending on the type of data being uploaded, a GenePix Array List (GAL) file may also need to be specified. This file contains information pertaining to the physical layout of the arrays used in the experiment and is needed to generate array images, as well as for certain kinds of normalisation. The information contained in a GAL file can usually be read from the actual raw data files themselves, so this field is optional. The next required field designates the image analysis program that was used to produce the files being uploaded. The options available are “Spot”, “GenePix”, “BlueFuse” and “Agilent”. Single dye experiments have all of the same upload options as dual dye (including GAL and Spot Types files), except for the fact that the names of the columns containing foreground and background intensity levels in the raw data files can also be specified.

Given the large amount of data that can be involved in microarray analysis and the often lengthy delays that are experience in processing data, one of the key usability issues of this system is that the user is never presented with a blank or static screen for any amount of time and always remains informed on what is happening on the systems back end [21]. This is particularly important where a web based system is concerned, as it is all too tempting for a user to click a browser’s stop or back button if they get even the slightest feeling something has gone wrong, which will of course halt execution of whatever is happening. This issue has been addressed in every aspect of this system where delay may be involved.

Our bovine dataset, which includes 12 arrays, weighs in at a hefty 41 megabytes, even when compressed. This upload will take several minutes, even on a quick connection.

The first screenshot on the next page shows the screens directly after the upload button has been clicked. This is what is seen while the actual data transfer is in progress. Note how the user is informed to be patient and told that the upload will take several minutes, while the spinning animation gives the all important impression that something is happening. The next screen shows what has happened after data is uploaded. For this dataset this screen takes about 30 seconds to work to completion. Crucially, progress is printed as each step is completed, so the user is not presented with a static screen until the page is fully loaded.

The next stage of almost any analysis is to assign the experiment's phenotypic data. This stage differs significantly for Affymetrix datasets, where the user has the option to specify an experimental design of up to 10 factors, with up to 10 levels of each factor. To clarify what I mean by this, consider our bovine dataset from earlier. The simplest experiments contain only one factor, our bovine experiment for example contains one factor, which we could call "Negative Energy Balance"; this factor is then said to have two levels, one of which is negative energy balance group and the other of which is a control group. An appropriate level of each factor is then assigned to each array to describe the phenotypic state of the sample. In the above example, arrays will either be designated as being from the negative energy balance group or from the control group, based on which of either level of the single factor they are assigned.

We will now consider the more complex experimental design of the Estrogen Dataset, which is one of several sample datasets that is bundled with Bioconductor. It is from an experiment on MCF7 human breast cancer cells using Affymetrix HGU95av2 arrays. The aim of the study was to identify genes which respond to estrogen and to classify these into early and late responders. This experiment is of the popular 2x2 factorial design. It contains two factors, both of which have two levels. The first factor (which we will call "Estrogen") defines two different kinds of samples, which have either estrogen absent or present. We will call the two levels of this factor "Absent" and "Present". The next factor (which we will call "time") defines the length of exposure of the samples, either 10 or 48 hours, we will call these levels "+10" and "+48".

Assignment of this phenotypic data to each of the arrays allows us to define different contrasts to assess for differential expression analysis. When the numeric values in the drop

down menus that define the amount of factors or levels is changed, more text boxes are dynamically added without reloading the page. This is achieved using JavaScript.

The next step in any analysis is to run quality control. The initial screen displays a list of available quality control options (boxplot, histogram etc.) that be selected using checkboxes. The options can be used to assess raw, preprocessed data, or both.

If working with dual dye or Affymetrix data, hovering the mouse pointer over the “Preprocessed Data” heading on the table allows the user to select from a number of different preprocessing methods. Changing this value will change how data is preprocessed for the appropriate plots.

Hovering the mouse pointer over any particular quality control option will cause a tooltip to appear that gives an outline of how that particular plot or metric can be used and how it should be interpreted. This is a useful feature for novice users.

Once the required options have been checked the user will submit this form. The next page displays everything that has been requested. Quality plots are displayed as thumbnail images, full size versions of which can be viewed by clicking the thumbnail (Fig. 3). Any metrics such as average background or scale factors are displayed in tabular form.

It is again important to note that the output of this page is piped directly to the screen as the page is generated. If every option is checked the whole page takes about a minute to print to completion for our bovine dataset, but the user is never left staring at a static or blank screen. Even the progress of any preprocessing method that is being undertaken is piped directly from R to the users screen.

Additionally, if analysing an Affymetrix experiment and the user selects the “PUMA PCA and Scree Plot” option, instead of these plots being generated there and then, the user will instead be informed that notification of completion of the plots will be emailed to them.

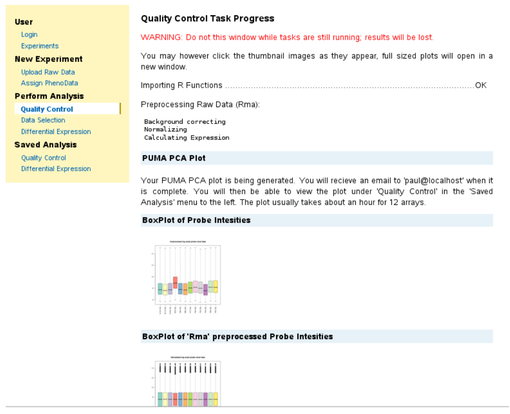

Figure 3. Screen shot of quality control output showing thumbnail images.

Due to the higher volume of data the PUMA method deals with, this plot takes a long time to generate, approximately 40 minutes for our bovine dataset. Upon completion the results of the plot can be retrieved using the “Quality Control” link under the “Saved Analysis” menu.

The data selection phase may be required following quality control. There are certain situations where a user may decide that, for quality reasons, an array is not suitable for inclusion in further analysis. The data selection page allows a user to exclude an array from subsequent analysis without having to create a new experiment.
If for example we were to decide to exclude the array “NS7.CEL” of our bovine dataset, from differential expression analysis, we can do so using this page by simply unchecking

here the array and pressing the “Update Data” button.

The next step in an analysis is to run differential expression. There are numerous different choices available at this stage and options differ significantly based on the type of dataset involved.

In the case of Affymetrix the user selects between the Limma and PUMA packages and selects a preprocessing method, they are then required to decide on which contrasts they wish to test for differential expression. This is simple for our bovine dataset, as this is an experimental design of only one factor which has only two levels; so the only contrast available is between these two levels.

But now consider the 2x2 factorial estrogen dataset described previously. In such a situation there are obviously several contrasts available, some which may be worth examining and some which may not. Two contrasts which are obviously of interest are “Present +10 VS Absent +10” and “Present +48 VS Absent +48” which will tell us which genes are calculated as being differentially expressed between the estrogen absent and estrogen present groups at 10 hours and subsequently at 48 hours. After checking the contrasts of interest and submitting the page, differential expression analysis will begin. These results can also be downloaded as a .xls format spreadsheet file, which can be viewed in an application such as Microsoft Excel or OpenOffice.

If the user opts to use the PUMA method to assess differential expression, they must once again begin by specifying the contrasts of interest; once these are submitted a message is printed notifying the user that they will receive an email upon completion of differential expression analysis. This is because differential expression analysis using PUMA takes a long time, approximately 8 hours for our bovine dataset on a machine that boasts a 2GHz Intel R CoreTM 2 Duo CPU with 4 gigabytes of RAM. This can be reduced by parallelising the differential expression process across multiple cores of a single CPU, or across multiple processing cores of multiple machines.

Differential expression analysis of dual and single dye data is similar to that of Affymetrix arrays, but because of the fact that the system currently only supports an experimental design where two levels of the same factor are compared, there is no need to specify

contrasts, as there is only one contrast available. Preprocessing options also differ significantly. Spots on dual and single dye arrays are sometimes duplicated one or more times within an array. This is to give a more reliable measure of the expression level of a gene represented by a spot. The system allows for this situation by allowing the user to specify the number of duplicate spots per array and the number of space between duplicates. Obviously this means that there must be the same number of duplicates for every spot and that the duplicates must be evenly spaced, but other than exceptional circumstances, this will always be the case. The level of correlation between these duplicate spots can then be factored into the linear model fit and subsequent differential expression analysis, hence giving a more accurate result.

Gene annotation information can be dynamically downloaded by clicking the gene names in lists of differentially expressed genes (Fig.4)

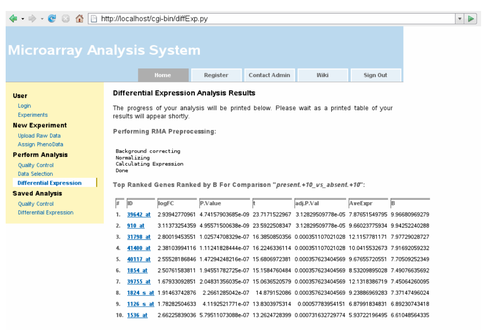

Figure 4. Screenshot showing differential expression analysis and clickable gene names.

This information can then be interpreted by a biologist. The .RData file or the spread sheet listing differentially expressed genes can be downloaded if subsequent analysis is to be

pursued using an alternative platform. The system makes a number of tools available to the user to manage data from previous experiments and analysis of experiments. The “Experiments” link in the menu on the left of the page will display a list of experiments that the user has previously uploaded and that are saved on the system. The option is available to remove any of these experiments from the system by checking them and clicking the delete key. The user may also select any previous experiments on the list for current use, which will allow them to either review previous analysis information or to perform new analysis. Clicking on any of the experiment names will bring the user to a page that shows various information about the dataset, such as the names of the files that were uploaded and data type.

Every quality control analysis is automatically saved by system and can be reviewed at a later date by the user. This is done by clicking the “Quality Control” link under the “Saved Analysis” menu. Doing this will open a page with a list of all previous quality control analysis that have been completed for this experiment. This page also allows the user to delete these previous quality control analysis if desired. By clicking on any of the analysis the user is brought to a page that describes some of the conditions under which the analysis took place, such as which arrays were selected, which preprocessing methods were used and what phenotypic data was assigned.

If during quality analysis, the user had selected a PUMA PCA and screen plot, this is where they will find those plots upon their completion. As stated previously, the user will receive an email informing them of completion, the results can then be found under this menu. This is similar to the access of previous quality analysis. The user can delete previous analysis, access information on previous analysis, and view their results in full. The user also has the option of downloading the “.RData” file, which is saved when any given analysis is completed. This file contains all of the R objects that were created during the analysis. An advanced user may download this file and load it locally in R if they wish to pursue further analysis from the command line. Clicking the “More Info” link beside the link to download the .Rdata file will provide the user with a detailed description of what the objects saved in the file are and how they were created.

If the user has specified that analysis should be performed using PUMA, similarly to quality control analysis, this is where their results will appear upon completion.

References

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol., 5, R80.

Huber W, von Heydebreck A, Sltmann H, Poustka A, Vingron M (2002) Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. Bioinformatics, 18, S96-S104.

Hubbell E, Liu WM, Mei R (2002) Robust estimators for expression analysis. Bioinformatics 18, 1585-1592.

Liu, X., Milo, M., Lawrence, N. D. and Rattray, M. (2006) Probe-level measurement error improves accuracy in detecting differential gene expression, Bioinformatics 22(17) 2107-2113.

Rafael. A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed (2003), Summaries of Affymetrix GeneChip probe level data Nucleic Acids Research 31(4) e15

Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397-420.

Wu Z, Irizarry RA (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. J Comput Biol 12(6)882-893.

Publications arising from this project:

Geeleher, P., Morris, D.G., Hinde, J. and Golden, A. (2008). Web based tools for analysis of DNA microarrays. In: Walsh Fellowship Seminar, RDS, Dublin, 12-Nov-2008, 5 pages 17620 B2

Geeleher, P., Morris, D.G., Hinde, J. and Golden, A. (2008). Bioconductorbuntu: a web based tool for microarray analysis. Bioinformatics (in press).