SMU Data Science Review

Volume 1 | Number 2 Article 9

2018

Predictions Generated from a Simulation Engine for Gene Expression Micro-arrays for use in Research Laboratories

Gopinath R. Mavankal
Southern Methodist University, gmavankal@smu.edu

John Blevins
Southern Methodist University, jblevins@mail.smu.edu

Dominique Edwards
Southern Methodist University, daedwards@mail.smu.edu

Monnie McGee
Southern Methodist University, mmcgee@smu.edu

Andrew Hardin
To be provided, drew_hardin@yahoo.com

Predictions Generated from a Simulation Engine for Gene Expression Microarrays for use in Research Laboratories

John Blevins[1], Dominique Edwards[1], Gopinath Mavankal[1], Andrew Hardin[2] and Monnie McGee[1]

Master of Science in Data Science, Southern Methodist University, 6425 Boaz Lane, Dallas, TX 75205
Silicon Valley Data Science, Cupertino, CA[2]
{jblevins, daedwards, gmavankal, mmcgee}@smu.edu, drew_hardin@yahoo.com

Abstract. In this paper we introduce the technical components, the biology and data science involved in the use of microarray technology in biological and clinical research. We discuss how laborious experimental protocols involved in obtaining this data used in laboratories could benefit from using simulations of the data. We discuss the approach used in the simulation engine from [7]. We use this simulation engine to generate a prediction tool in Power BI, a Microsoft, business intelligence tool for analytics and data visualization [22]. This tool could be used in any laboratory using microarrays to improve experimental design by comparing how predicted signal intensity compares to observed signal intensity. Signal intensity in microarrays is a proxy for level of gene expression in cells. We suggest further development avenues for the prediction tool.

1 Introduction

We apply the Probe Logarithmic Intensity ERror (PLIER) algorithm [5] to a series of simulations using the Simulation Engine described by Hardin A. [4] to generate a data set that predicts how strength of a signal which is a proxy for gene activity would vary with changing levels of stimuli or treatments applied to the samples. Researchers studying fundamental biology or performing clinical investigations are interested in understanding how gene activity responds to stimuli and influence each other. The data generated in this report would allow investigators using microarrays to gain insights that could be used to refine laboratory experiments they conduct, to reduce the number of trials and in general increase statistical power of the experiments by comparing predicted to observed signal intensities that serve as surrogates for gene activity. In this introduction we provide an overview of biology, technology and data analysis involved in the use of microarrays.

1.1 Detecting Gene Expression using Microarrays

Gene Expression Microarrays were first used in the mid-1990s as a tool to study the simultaneous expression of thousands of genes under an unlimited number of treatments, time courses and applications. Gene expression occurs when transcription is initiated through cellular signaling pathways (Fig 1) which result in the DNA template (genome) getting copied into multiple copies of mRNA [1]. The level of mRNA production varies under different stimuli. These mRNA copies exit the nucleus where they are produced and are bound by the ribosomal apparatus in the cytoplasm which in turn uses the mRNA as a template to translate the genetic code (nucleotides) to proteins (amino acids). Gene expression is the expression of the genetic code as protein molecules that perform functions in a cell.

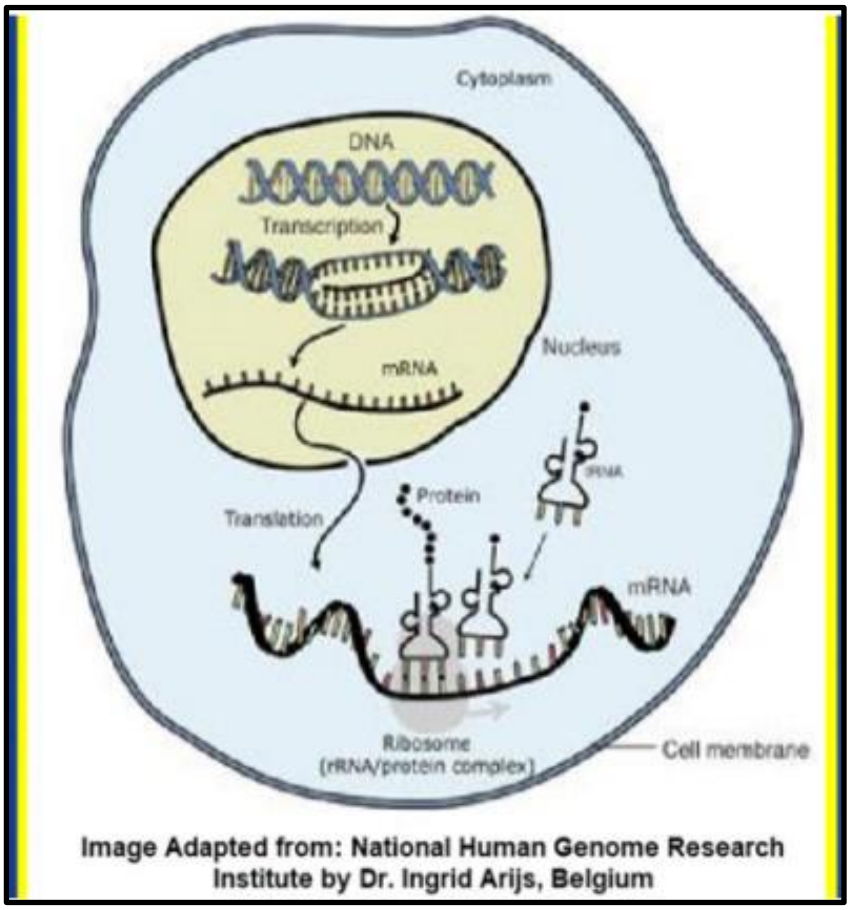


Fig. 1. Gene Expression occurs when stimuli from the environment are conveyed through signaling pathways to the interior of the nucleus where the genome resides. The figure illustrates how the complementary sequences of the DNA in the nucleus are melted or opened by

transcription factors which then results in the single copy of the protein encoding DNA region of the genome to be copied into several copies of messenger RNA (mRNA) that exit the nucleus. These mRNA are engaged by the ribosomal apparatus in the cytoplasm which utilize the mRNA as template to "translate" the genetic code into proteins. Thus, the level of protein production which is a proxy for gene expression can be measured by measuring the level of mRNA in the sample applied to a microarray plate.

Microarrays are designed to measure the level of gene expression and it uses the property of nucleotides to pair and thus bind to its complementary sequence (Fig 2). Nucleotide sequences have directionality and complementarity (lines in figure, connecting the sequences represent the pairing interaction of A to T or A to U or G to C). Although one might expect a perfectly matched sequence of probe DNA and target mRNA to bind and that of a probe DNA with a mismatch introduced in it, to not bind, experimental data shows otherwise. So, the expectation of using the mismatched probe signal as a background correction for the matched probe deposited in a different array location did not work as expected. Some mismatched probes produce a signal that is much better than other probe sequences with a perfect match. This has an implication in the way the microarray data is treated. Whereas older investigations used the mismatched and perfect match probes as proxy for signal and background, based on the empirical realization that it is not so, the work in this report treats the mismatched and matched probes as equivalent to each other and not as signal and background.

mRNA isolated or extracted from a biological sample is labelled or chemically bound to a fluorescent dye molecule. The mRNA of a specific region of hemoglobin protein is shown in Fig. 2. The unshaded sequence with uracil molecules in red represents the sequence of the mRNA. mRNA can bind to a stretch of DNA that is chemically synthesized and bound at a high concentration within a small area of a plate. The sequence of such a "probe" is shown in orange highlight in the figure. The probe sequence is in fact complementary to the gene sequence shown in blue shades in Fig. 2.

Just about 2% of the human genome codes for proteins. In the figure, the green shaded sequence is the code for the protein, hemoglobin which binds oxygen in circulating red blood cells in our body. The mRNA can be isolated from cells (hemopoietic red blood cells and not the mature red blood cells that lose the nucleus, in this instance) and a very small quantity of mRNA (nanograms) suffices for detection. The mRNA is chemically bound to fluorescent chemical. This modified mRNA can fluoresce and produce a light signal that can be detected and quantified. The mRNA (target) containing sample is injected into the chamber of the microarray plate that has more than a million probes placed in an array. This is incubated at a fixed temperature for a length of time and then washed (Fig. 4, probe array hybridization step). A fixed amount of each probe sequence is on the plate which is also referred to as a platform with identifiers such as GPL570. Platforms are designed such that all probes are sequences obtained from one organism. Thus, there is a human genome platform which is different from a platform that uses the yeast genome.

The PLIER algorithm relates the fraction of RNA in the target that was applied to the microarray plate with the fraction that gets bound to the probe positions on the microarray due to the affinity of the molecules which stems from the complementarity of nucleotides. The fluorescent-labelled mRNA that is bound to the proble would

produce a light intensity signal that is proportional to the amount of RNA that was present in the sample which was applied. This relationship between the DNA of the genome (blue shaded), the complementary sequence in the mRNA copy (not shaded) is illustrated in Fig. 2, using the gene sequence for Hemoglobin as an example.

Coding regions in a gene are often not contiguous. They get copied and spliced (joined) together in the cell. This can produce different versions of the same gene differing in length depending on the segments that are missing. These are called splice variants. In the past all probe sequences that were applied on the plates were chosen from the 3-prime end of the gene, so it would detect all splice variants. Current designs pick sequences across the coding segments and are referred to as exon arrays. Just like Moore's law, better technology in improving and increasing the amount of probe into smaller areas producing better signal quality [6] [8][9].

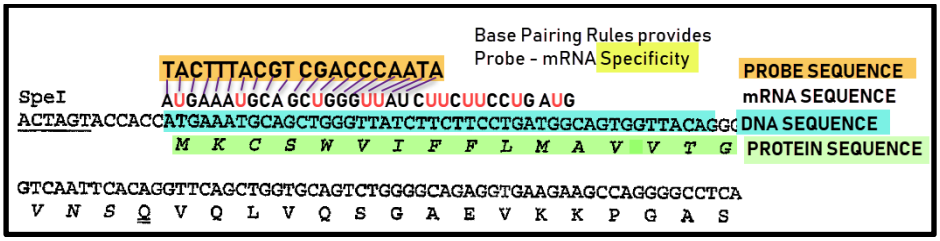


Fig. 2. The amino acid sequence of Hemoglobin protein is shown in green in alignment with the coding for it on the gene for it which is highlighted in blue. When the information is copied, it is copied into a slightly different molecule, the messenger RNA that has a complementary sequence shown without shading, with the Uracil equivalent of Thymine nucleotide molecule shown in red highlight. The probe sequence that would be designed to detect the mRNA is shown in the orange highlight. The complementarity of the nucleotide sequence over a stretch of the length of the molecule results in mRNA binding a probe that is chemically synthesized and immobilized on a small well-defined area of a micro array plate.

The probe sequences in commercially available micro-arrays are of two types, oligonucleotide arrays and cDNA arrays. Oligonucleotides could be short, 25-mer arrays (Affymetrix) or long 60-mer (Agilent) arrays. cDNA are much longer sequences. Short oligonucleotides have been shown to have the advantage of being capable of high density spotting which could improve signal but may lose the specificity of the signal. The Agilent platform using longer oligonucleotides would have more specificity but lesser density in spotting. Low levels of differential expression of several thousands of genes, which may be sufficient to modulate biological activity, poses a challenge in detecting them in studies using microarray technology. However, the accumulated microarray that result from the use of the same platform by different investigators across the globe, provides the opportunity to mine information that may have been missed in earlier analysis.

1.2 Challenges in Verifying Gene Expression using Microarrays

Using the complementarity of the sequence to detect and quantify mRNA is in fact a well-established technique much before microarray technology emerged. Assays

such as Reverse-Transcriptase Polymer Chain Reaction (RT-PCR) [11][12] are well established accurate procedures. The difference between such highly accurate RT-PCR assay and the microarray data is the difference between getting a snapshot of the amount of one gene versus getting a simultaneous snapshot of all genes. Such data also shifts the paradigm of the analysis from a reductionist approach to a non-reductionist approach, especially when the number of replicates available for such high dimensional data is very small [23]

However, if these methods were to be applied to determine the dose-response of every probe on a microarray, we are speaking of assaying over a million probes. Other challenges are discussed in [4].

Another way to verify levels of expression is to empirically examine (as opposed to theoretical examination) by using data sets derived from real data, that are referred to as plasmode data sets [13][14]. Though invaluable, "spike-in arrays", a form of plasmode data also provide a way to verify expression levels [15]. Such studies are rare, though. MicroArray Quality Control (MAQC) project launched by US Food and Drug Administration found data across experiments and platforms can be reliable and consistent.

The nature of biochemical phenomenon presents a challenge in that "while early microarray experiments focused on samples with large differences in a few genes, more recent findings stress that it is not large changes in a few genes, but rather small changes in many genes that will be important for understanding both complex diseases and the subtleties of biological processes"[16]. In addition to this is the problem of high dimensionality and low sample size of the microarray (holds true in general for all "omics") datasets. An introduction to statistical testing procedures as it relates to analysis of microarray data of tumor tissues is discussed in a book on Statistical Analysis principles [17].

1.3 Statistical Simulations offer an alternative

Statistical simulation could be used as an alternative to perform experiments to verify differential expression of genes. The pros and cons of using simulations is discussed in the technical report [4]. A general review

The report [4] discusses how the model in the dissertation [7] addresses the issues that relate to the flexibility of using simulation models to incorporate bias that could result from attempts to favor desired conclusions. This is done by using a combination of bootstrapping from real data and shaping the simulation to reflect the limits of detection at the tails of the distribution.

The simulation model presented is a location scale model [7].

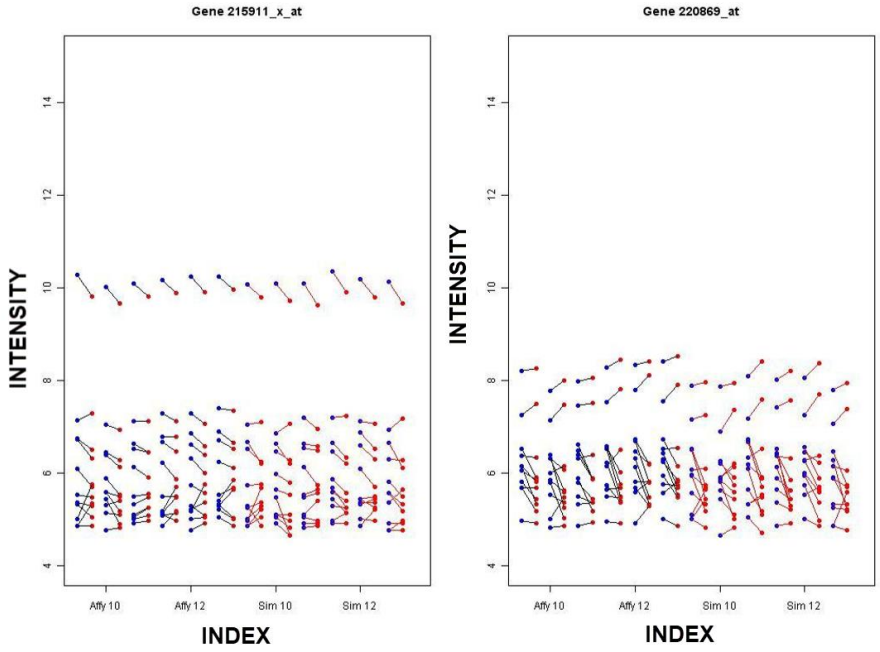


Fig. 3. Comparison of the original Affymetrix experiments and the simulated arrays for two genes from [7]. The figures show the six arrays on the original data compared to the 6 arrays on the simulated data. In this figure there are 11 probes per gene. For each array the PM and MM intensities have been plotted. The PM intensities are the values shown by the blue dots. The MM intensities are to the right of the PM intensities and are shown in red dots. A line connects the two intensities to show the relationship between the PM and MM intensities for each probe pair. The line is black for the original arrays and red for the simulated arrays

Figure 3 from [7] shows the simulation generates intensities that are like the intensities reported in experiments. The simulation engine is simulating the noise as well as the signal. Simulations reported in the literature often simulate expression levels.

While this study was performed on a platform that used DNA from Human genome, the approach would be expected to be more general and applicable to other platforms using DNA of other species.

1.4 Workflow in the laboratory that precedes Microarray data collection

Laboratories doing clinical or fundamental research involve much bench work as is shown in Fig. 4. A typical analysis workflow is described in [2]. The detection limits have fallen so low as to enable detection of mRNA from very small sized samples and even up to single cell. The limits are lowered by including an amplification step where a reverse transcriptase enzyme is used to make several copies of the isolated mRNA. This would result in complementary DNA (fig 2), which require a microarray

purposed for that by having array locations with sequences that can bind the cDNA instead of mRNA. The binding of a fluorescent label to the mRNA and the incubation conditions (probe array hybridization in Fig. 3A.) are factors that could result in lowered or improved binding and thus indirectly determine the detection of differences between control and sample or differential gene expression.

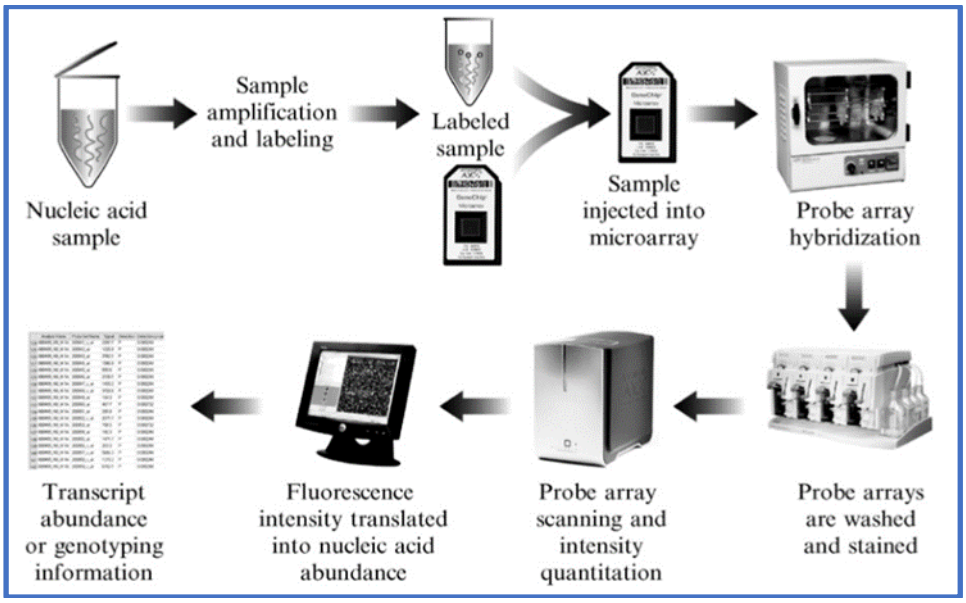


Fig. 4. Laboratory workflow that precedes data collection from [19]. The nucleic acid sample is the total RNA isolation procedure from a sample whose mRNA is labelled with a fluorescent dye. The dye-labelled mRNA is injected into the microarray cassette that in the case of the platform used in this study was a human genome platform (GPL570) with over 1.3 million probes (20-mer sequences from over 54,000 mapped genes. Mappings of probes to genes undergoes revisions and only about 600,000 of the probe positions had mappings to a gene. The mappings that have one to many were excluded by the manufacturer, but the probe sequences are made available and researchers could identify and work with them, if needed. The hybridization step on the top right panel is done at a fixed temperature and duration to allow the binding of the target mRNA to the probe on the plate. This is followed by a wash to remove all unbound target mRNA before the microarray plate is scanned for fluorescence intensity. The fluorescence intensity also referred to as the Probe intensity is a proxy for the level of gene expression or the level of mRNA template that would result in protein expression / gene expression.

2.0 Methods

We used the code in [7] to create gene level summaries of the predictions. The platform "GPL570" was selected in the GEO database [3] as it is the Affymetrix platform's newer version. This platform had 572 experiment sets (GSE), of which 49

were randomly selected. The 1442 CEL files corresponding to these experiments were used to generate simulations.

A series of simulations were created by maintaining the fold-changes for all probes to values ranging from -20 to +20, after each individual run. 17 such sets were then combined into one file with each row being merged with the gene related data corresponding to each probe position.

2.1 Pseudo code to obtain data for simulations

1. In GEO database, query and locate all records of experiments (all GSE) of a platform (GPL570).
    a. Create a sampling frame
    b. assign unique random numbers
    c. sort by assigned random numbers
    d. select first 50 experiments
2. Manually or through code, download the data.
    a. Package GEOquery, functions getGEO, getGEOSuppFiles
    b. Alternately, download CEL files from GEO to local drive
3. Create a folder that will be input to the simulation
    a. CEL files in folder

2.2 Pseudo code to generate simulation

1. Retrieve all CEL files in folder
2. For each probe position, compute normalized values as the standardized intensity measurements calculated by
    a. subtracting the mean of the probe position from all CEL files
    b. then dividing by the standard deviation of the probe position from all CEL files
    c. GPL570 has 1354896 probe positions per CEL file
3. Generate standardized files for each CEL file in a separate directory from previous step
4. Specify the parameters for the simulation run
    a. We treated all CEL files as equivalent, so the factor was assigned without regard to the CEL file's meta data identifying it as control or treatment.
    b. We specified that 20 simulations be produced in each run
    c. We specified the factor to use in the PLIER function in a file for each run as applicable to all probes that are mapped to genes.
5. Run simulation
    a. We encountered an error with one experimental set that was dropped.
    b. Simulation run produces a file with output of 20 simulations of the array which is the equivalent of 20 CEL files or 1354896 probe intensities.

2.3 Pseudo code to convert simulation output to gene level summary

1. Match each probe position to the probe pair and each probe pair to gene
    a. Package Bioconductor, library hgu133plus2probe, provides probe sets to probe position mappings
    b. Package GEOquery, function getGEO to get Simple Omnibus Format in Text (SOFT) file of platform meta-data
2. Calculate the mean and standard deviation and coefficient of variation as a percentage for each probe position from the output of each of 20 simulations for each of 16 different factors applied. Drop rows where the coefficient of variation exceeds 10%.
    a. Packages dplyr, tidyr, tidyverse
    b. Functions groupby, summarise, apply, filter
3. For each mapped gene, calculate the summarized mean and standard deviation using the probe position to gene mapping from the output of previous step. Same packages as used in previous step.
4. Generate a csv file that lists the 54K genes in the mapping by the characteristics of the gene (from platform meta data) and each of the summarized values for each of the applied factors.

2.4 R packages and tools used in this report

The soft (*Simple Omnibus Format in Text*) file of the platform GPL570 was downloaded using the function "getGEO" from the "GEOquery" package. The information from this file was written into a local file for reuse. The database of the same platform, "hgu133plus2probe" installed through the "biocLite" package was used to map the probeset to gene. This library maps only about 0.6 million of the 1.3 million probe positions in the CEL files. The other probe positions are perhaps not used by the manufacturer as each probe might be mapping to more than one gene. The sequences of all probes though are available. The probe locations are referred to as Probe interrogation positions.

Simulations were generated and saved into a file. The code used for the simulation was provided by Andrew Hardin. Parameters generated by this code is written into files which require further specification of other parameters before further use. The file "experiment_factor_affy.xlsx" specifies the factors to apply to the PLIER function. These functions are not applied on the probes that have the prefix "AFFY". The same factor was applied to all the probe sets and the simulation results were labelled to reflect the factor applied. Jupyter notebook was used to generate the simulated values that were written into files, as the code was provided in that format.

The plyr/dplyr/tidyr/tidyverse packages was used to merge the probe level simulation outputs to gene level summary statistics. Probe position refers to the individual probes and Probe name refers to the probe set. Each probe set corresponds to a gene in the platform. The output from all simulations using different fold increases were merged to one file with data at the summarized gene level for different

fold concentrations. This file was used in Power BI to generate visualizations that pertain to the filters that were applied to select the genes of interest.

Genes are annotated by the same set of key figures as are used in the GEO database and each of these key figures are represented in the output file. In Power BI these were relabeled to make them more user friendly.

2.5 Bioconductor packages and S4 classes

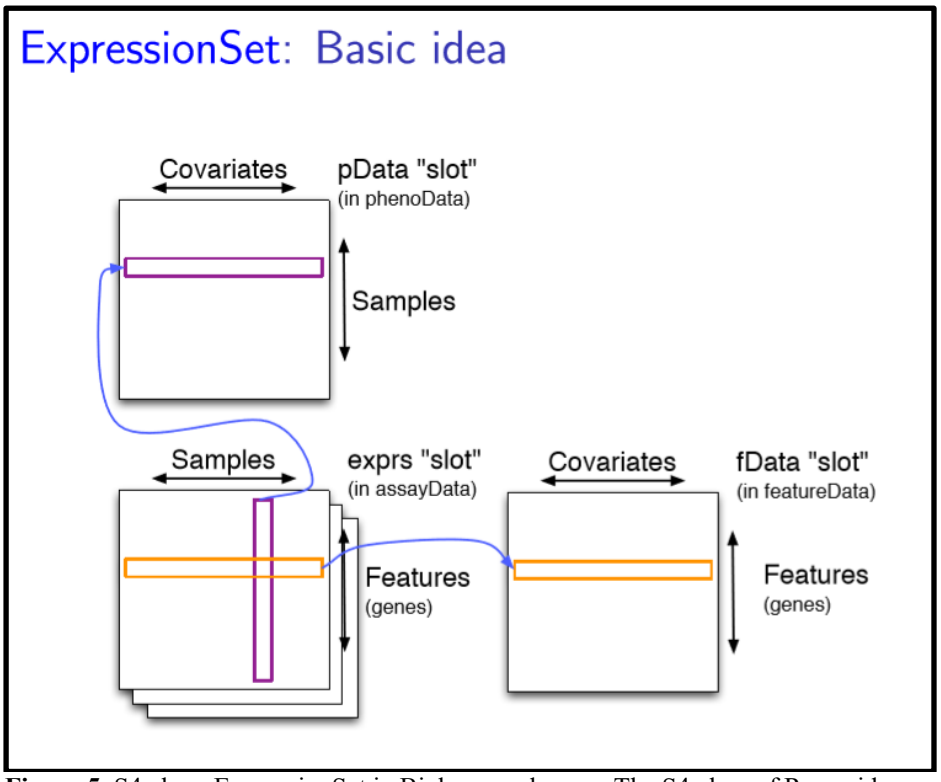


Figure 5. S4 class, ExpressionSet in Biobase package. The S4 class of R provides a container for high-throughput assay data and data on genes ("phenoData" in figure). It is a matrix data structure where the rows represent the probes in a probe set that maps to a gene, which are also referred to as features. Slot is a S4 term for "properties". The properties of the gene are different features describing the gene such as the function and location within the cell. Such data continues to get updated and thus by separating them, allows users to use previously reported experimental data with current knowledge of the gene.

S4 classes are more rigorously defined than S3 classes that are more commonly used in R. Information is organized into slots (properties). The class and slot data type must match. By enforcing type and validity, inheritance and encapsulation is enabled

which in turn allows complex projects using multiple contributors to work effectively as is the case with microarray experiments where data submission to the database must be MIAME compliant.

3 Results

The GEO database [3] maintained by National Institutes of Health, National Center for Biotechnology Information (NIH NCBI GEO) as of May 2018 has over 1.3 million samples (GSM), in 4348 dataset records (GDS) in over 97,000 experiment series (GSE) with 448653 CEL files on human genome. The site also provides an integrated access to genes of genomes. Genes can be queried using meta data on genes that are organized under several categories. The csv file we produced mirrors those fields, which were sourced from the manufacturer provided platform package in R as shown in the pseudo code in the section on Methods. A live demo of the dashboard [10] is available. A researcher using the csv file can download a free version of the tool, Power BI from Microsoft™ to generate this dashboard. Microsoft provides training online on using features such as drill down available in the tool.

3.1 Predictions from Simulations surfaced in a Power BI dashboard

The csv file output generated in this report is specific to the Affymetrix platform, GPL570, a human genome platform. See Appendix section A1 for details on this file. The features in the output include fields that are meta data on the genes in the platform and the predicted mean, standard deviation and coefficient of variation for each of 18 factors (see Fig. 6) used to predict the signal intensities.

A six-minute video explaining the use of the dashboard created in Power BI is available [10] (see footnote 1). Figs. 6 and 7 show the dashboard created in Power BI that is discussed in the video.[1]

[1] Link to Demo of BI Tool

Figure 6. Simulated output is used to create a Dashboard in Power BI™ see Appendix A2 for instructions to set it up.

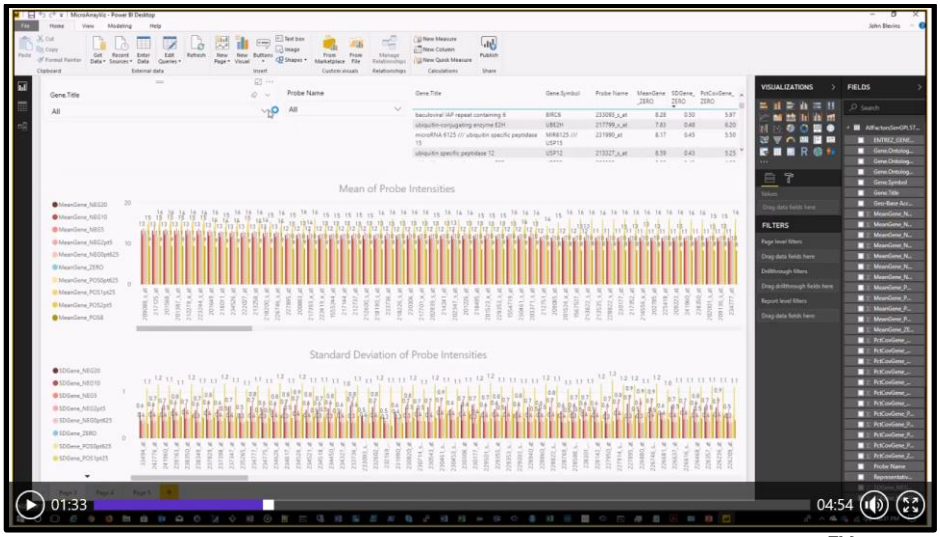


Figure 7. Simulated output is used to create a Dashboard in Power BI™ see Appendix A2 for instructions to set it up and video link on how to use it.

3.2 Use of the dashboard to explore candidate genes that might be differentially expressed

The dashboard can be created by a user as well, using the free version of Power BI available from Microsoft (see Appendix section A2 for link to download). The dashboard provides a user the ability to select genes based on criteria that range from a specific gene's ID to exploring possible candidate gene/s located using terms searched in the data about the gene.

For instance, the field (Appendix A2) "Gene.Ontology.Cellular.Component" in the file is data about the location of the protein within the cell. If the term "Golgi" (for Golgi bodies, the organelle that sorts and processes proteins) is searched in this field, it selects 3283 genes from 52033 genes in the file. If the investigator is exploring the possible candidate genes involved in metabolism within this subset, the term "metabol" entered in the field "Gene.Ontology.Biological.Process" would return 1270 hits. Furthermore, if the molecular function of candidate regulated genes of interest is a transferase, a third filter for "transferase" applied in the field "Gene.Ontology.Molecular.Function" would then yield 458 hits. As ontology relates to the basis on which inference is drawn, the investigator could restrict examination of genes known to have this function that is based only on direct assay. By adding an additional search term "inferred from direct assay" to the ontology of the molecular function, the number of hits gets reduced to 232 genes.

In this manner a researcher could narrow the listing of candidate genes to consider from 52033 rows to 232 rows,

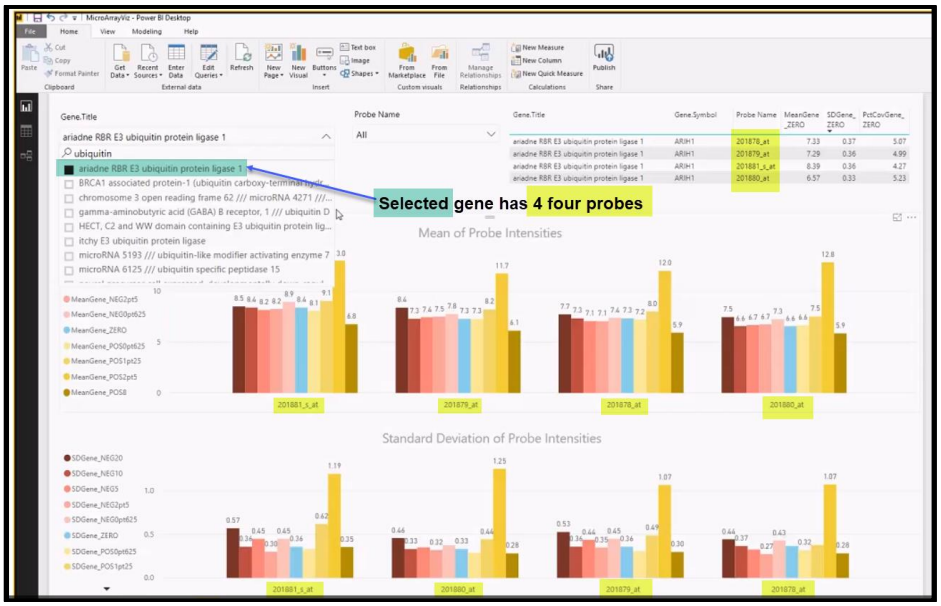


Fig. 8. The user can view the field "Gene Title" in the dashboard and select a gene, highlighted in green in the figure. On doing so, the dashboard would refresh to display the probe sets that map to this gene. The terms probe and probe sets are sometimes used interchangeably. The mapping also changes with new information, sometimes.

Probe position uniquely identifies the probe on the array and it refers to one specific sequence from the coding region of the human genome. There are two panels of data that are visible in the display. The panel on the top are the mean intensities from the simulation while the set at the bottom are the standard deviations. Sample raw data numbers that correspond to these can be seen in Appendix A2.

Figure 8 shows a visualization of one selected gene. The gene selected in Fig. 8 has 4 different probe sets mapping to it.

3.3 The value proposition of the dashboard – mine the detailed Signal Intensities at a probe position level

The dashboard shows how any selected gene/s would have their probe-set level summarized mean signal intensity change with increasing or decreasing fold concentration from the basal level of gene expression. The basal level is the intensity where the PLIER function factor 0 was applied to the data drawn from the sampled set of CEL files.

The file generated in this report does not retain the probe level data. It is possible to select a few (four) ranked set for each gene and retain the probe level prediction to compare it against the observed signal at a probe position level. Power Bi's drill down feature would enable such exploration of data.

A researcher comparing the predictions to observed signal intensities might consider parameters to change in the microarray sample preparation protocol, such as the temperature at which hybridization is carried out by considering the characteristic of the dose response and extent of variability in the predictions in the tool.

In its current form this file does not include experimental data. It is possible to include such data though it would require coding effort. Power BI features could circumvent the need for coding and enable the user to compare experimental data with the predictions. The field that could be used for this purpose, namely, the probe position in the array is available in the file.

By replacing the platform specific data used in the simulation, this method could be extended to other platforms.

4 Conclusion

The Power BI visualization can be setup by any researcher who gets the csv file output. The predictions can be refined, and additional drilldown features could be added as discussed in section 4.3 below. Such use of the tool holds the potential to realize efficiencies in conducting fewer better experiments in the clinical and fundamental research labs which could increase the adoption of the technology, considering the vast amount of data available in public databases.

4.1 Microarray technology and "Omics"

Microarray technology is essentially a query on biomolecules using a set of sequences expected to have target molecules in the sample, as discussed in section 1.3. The underlying principle of a simultaneous capture and definition of several biomolecules now extends to different technologies some of which have been around for a while.

For instance, mass spectrometry, a well-established technique is applied to identify cancer biomarkers from tissues[20]. Proteases are enzymes that act as scissors that cut specific motifs. Proteins undergo modifications after they are translated (Fig. 1). When such modifications change the function of the protein, a disease state that might result from differently modified proteins is not the result of a differently encoded gene (mutation) or a differently expressed level of expression.

The report [20] describes a technique that uses mass spectrometry (tandem MS/MS) to identify proteins from a patient with a congenital form of cataract that is the result of a different pattern of protein modification. So, the protein sequence per se is not any different between a normal and affected individual, but the congenital form of the disease results in the protein being modified in a different manner. This is one example of proteomics. In general, when analytical techniques are applied to biomolecules in a similar manner to look for patterns in other biomolecules, the term "omics" is used.

4.2 Microarray technology and "RNA seq"

RNA transcripts can be identified by a high throughput sequencing methodology that has been touted as a replacement for microarrays by some [21] while others have point to a role for both technologies [9].

Recent trend has seen a decrease in submissions of micro array data in GEO database. This does not necessarily imply that the use of micro arrays has decreased due to it being superseded by RNA seq or Nextgen sequencing [18]. There are users in industry who do not submit the data to public databases but use the array. As database submission standards have not kept pace with the technology changes, it has also resulted in fewer submissions[18].

4.3 Further work.

A back test of the prediction to the observed signals could be done by identifying the controls and treatment files of an experiment set from the meta data files. This would provide a reference measure to optimize the parameters used in this work. The parameters that could be varied to get a better fit of the predicted set of metrics would be:

1. The sample size: this report used 1442 CEL files to bootstrap the simulation. Perhaps doubling this size might improve the prediction.

2. This report selected probes with a percent coefficient of variation below 10 to improve the summarized statistics of a probe set. Perhaps a better measure might be to use the top 3, 5 or 9 ranked by percent coefficient of variation.
3. This work used 20 simulations to summarize the simulated means and standard deviation. Perhaps an increased number of simulations might yield a better fit.
4. This report used the mean statistic for the summarized values by gene. Perhaps the third quartile might be a better statistic considering the noise in the data.
5. The predictions generated in this work shows a peak that is consistent across the genes. It might be instructive to verify if it is reproducible.

References

1. The Cell: A Molecular Approach, 2nd edition, by Cooper (GM) and Sunderland (MA), Oxford University Press, 2000, (ISBN-10: 0-87893-106-6).
2. NIH, National Cancer Institute, Collaborative Bioinformatics Resource (CCBR), Bethesda, Maryland, USA, Training program document link (https://bioinformatics.cancer.gov/sites/default/files/course_material/Btep-R-microA-presentation-Jan-Feb-2015.pdf , accessed on 24 Oct 2017).
3. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan; 41(Database issue):D991-5.
4. Andrew Hardin and Monnie McGee(2015). Semi-Parametric Simulation of Affymetrix Microarrays to Obtain Realistic Output. Department of Statistical Science Technical Report Number 383. Southern Methodist University, Dallas, Texas.
5. Therneau, T.M and Ballman, K.V.(2008) "What does PLIER really do?" Cancer Informatics 6:423-431
6. Dapas, M., Kandpal, M., Bi Y, and Davaluri, R. V. Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms. Briefings in Bioinformatics, Vol18, Issue 2, pages 260 – 269, 2017, Oxford Univ. Press,
7. Hardin, A. Semi-Parametric Simulation of Affymetrix Microarrays to Obtain Realistic Output. Ph.D. Dissertation, Southern Methodist University, 2010.
8. Chen, L., Sun, F., Yang, X., Jin, Y, Shi, M., Wang, L., Shi, Y., Zhan, C., and Wang, Q. Correlation between RNA-Seq and microarrays results using TCGA data. Gene, 628 (20170, 200-204.
9. Robinson, D.G., Wang, J. Y., and Storey J. D. A nested parallel experiment demonstrates differences in intensity-dependence between RNA-seq and microarrays. Nucleic Acids Research, Volume 43, Issue 20, 16 November 2015, Pages e131. https://doi.org/10.1093/nar/gkv636 (accessed on 6 May 2018).
10. A six minute video on the tool built in this project is in the link: https://amotionbackup.blob.core.windows.net/capstone/PowerBI_Demo.mp4 accessed on7 May 2018.
11. Etienne W., Meyer M.H., Peppers, J., Meyer, R.A.J (2004) "Comparison of mRNA gene expression by RT-PCR and DNA microarray" Biotechniques 36(4): 618 - 20,622,624-6
12. Qin, L., Beyer, R.P., Hudson, F.N, et al. (2006) "Evaluation of methods for oligonucleotide array data via quantitative real-time PCR" BMC Bioinformatics 7:23
13. Gadbury GL, Xiang Q, Yang L, Barnes S, Page GP, et al. (2008) Evaluating Statistical Methods Using Plasmode Data Sets in the Age of Massive Public Databases: An Illustration

Using False Discovery Rates. PLoS Genet 4(6): e1000098. doi:10.1371/journal.pgen.1000098.

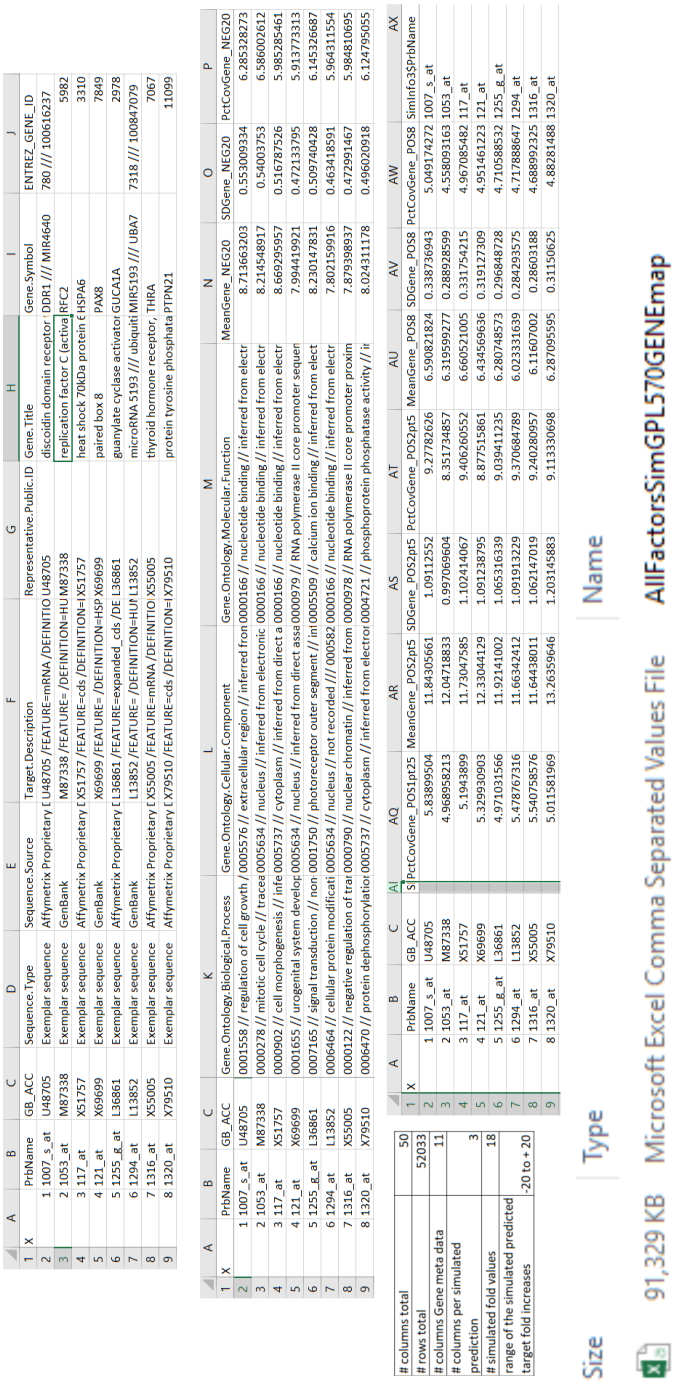
14. Guillem Rigaill, Sandrine Balzergue, Véronique Brunaud, Eddy Blondet, Andrea Rau, Odile Rogier, José Caius, Cathy Maugis-Rabusseau, Ludivine Soubigou-Taconnat, Sébastien Aubourg, Claire Lurin, Marie-Laure Martin-Magniette, Etienne Delannoy; Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis, Briefings in Bioinformatics, Volume 19, Issue 1, 1 January 2018, Pages 65–76, https://doi.org/10.1093/bib/bbw092
15. Qianqian Zhu, Jeffrey C Miecznikowski, Marc S Halfon. Preferred analysis methods for Affymetrix GeneChips. II. An expanded, balanced, wholly-defined spike-in dataset. BMC Bioinformatics, 2010, Volume 11, Number 1, Page 1.
16. Vasiliu D, Clamons S, McDonough M, Rabe B, Saha M (2015) A Regression-Based Differential Expression Detection Algorithm for Microarray Studies with Ultra-Low Sample Size. PLoS ONE 10(3): e0118198. https://doi.org/10.1371/journal.pone.0118198
17. Dunkler D., Sánchez-Cabo F., Heinze G. (2011) Statistical Analysis Principles for Omics Data. In: Mayer B. (eds) Bioinformatics for Omics Data. Methods in Molecular Biology (Methods and Protocols), vol 719. Humana Press.
18. Discussion with technical support from Affymetrix.
19. Dalma-Weishausz, D. Warrington, J. Tanimoto, E. Y., and Miyada C. G. The Affymetrix Platform, An overview, METHODS IN ENZYMOLOGY, VOL. 410 2006 Elsevier Inc.
20. Michael J. M., W. H. McDonald, A. Saraf, R. Sadygov, J. M. Clark, J. J. Tasto, K. L. Gould, D. Wolters, M. Washburn, A. Weiss, J. I. Clark, and J. R. Yates III. Shotgun identification of protein modifications from protein complexes and lens tissue. PNAS June 11, 2002. 99 (12) 7900-7905; https://doi.org/10.1073/pnas.122231399. (accessed on 10 May 2018).
21. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X (2014) Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLoS ONE 9(1): e78644. https://doi.org/10.1371/journal.pone.0078644 (accessed on 10 May 2018)
22. Whitepaper "Advanced Analytics with Power BI", Microsoft. Link to Power BI whitepapers: https://docs.microsoft.com/en-us/power-bi/whitepapers (accessed on 10 May 2018).
23. Rivas Ariel L., Leitner Gabriel, Jankowski Mark D., Hoogesteijn Almira L., Iandiorio Michelle J., Chatzipanagiotou Stylianos, Ioannidis Anastasios, Blum Shlomo E., Piccinini Renata, Antoniades Athos, Fazio Jane C., Apidianakis Yiorgos, Fair Jeanne M., Van Regenmortel Marc H. V. Nature and Consequences of Biological Reductionism for the Immunological Study of Infectious Diseases. Frontiers in Immunology, Vol 8. Pg. 612, 2017. https://doi.org/10.3389/fimmu.2017.00612 (accessed on 3 May 2018).

Appendix:

A1. Link to a blog on R and methods that was useful for work done in this report

http://biolearnr.blogspot.co.uk/ (accessed on 11 May 2018)

A2. File output of predictions from the simulations, generated in this project



	A	B	C	D	E	F	G	H	I	J
1	X	ProbName	GB_ACC	Sequence.Type	Sequence.Source	Target.Description	Representative.Public.ID	Gene.Title	Gene.Symbol	ENTREZ_GENE_ID
2	1	1007_s_at	U48705	Exemplar sequence	Affymetrix Proprietary D	U48705 /FEATURE=mRNA /DEFINITIO	U48705	discoidin domain receptor	DDR1 /// MIR4640	780 /// 100616237
3	2	1053_at	M87338	Exemplar sequence	GenBank	M87338 /FEATURE= /DEFINITION=HU	M87338	replication factor C (activa	RFC2	5982
4	3	117_at	X51757	Exemplar sequence	Affymetrix Proprietary D	X51757 /FEATURE=cds /DEFINITION=I	X51757	heat shock 70kDa protein 6	HSPA6	3310
5	4	121_at	X69699	Exemplar sequence	GenBank	X69699 /FEATURE= /DEFINITION=HSP	X69699	paired box 8	PAX8	7849
6	5	1255_g_at	L36861	Exemplar sequence	Affymetrix Proprietary D	L36861 /FEATURE=expanded_cds /DE	L36861	guanylate cyclase activator	GUCA1A	2978
7	6	1294_at	L13852	Exemplar sequence	GenBank	L13852 /FEATURE= /DEFINITION=HUI	L13852	microRNA 5193 /// ubiquiti	MIR5193 /// UBA7	7318 /// 100847079
8	7	1316_at	X55005	Exemplar sequence	Affymetrix Proprietary D	X55005 /FEATURE=mRNA /DEFINITIO	X55005	thyroid hormone receptor,	THRA	7067
9	8	1320_at	X79510	Exemplar sequence	Affymetrix Proprietary D	X79510 /FEATURE=cds /DEFINITION=I	X79510	protein tyrosine phosphata	PTPN21	11099

	A	B	C	K	L	M	N	O	P
1	X	ProbName	GB_ACC	Gene.Ontology.Biological.Process	Gene.Ontology.Cellular.Component	Gene.Ontology.Molecular.Function	MeanGene_NEG20	SDGene_NEG20	PctCovGene_NEG20
2	1	1007_s_at	U48705	0001558 // regulation of cell growth / 0005576 // extracellular region // inferred from 0000166 // nucleotide binding // inferred from electr			8.713663203	0.553009334	6.285328273
3	2	1053_at	M87338	0000278 // mitotic cell cycle // traceal 0005634 // nucleus // inferred from electronic 0000166 // nucleotide binding // inferred from electr			8.214548917	0.54003753	6.586002612
4	3	117_at	X51757	0000902 // cell morphogenesis // infe 0005737 // cytoplasm // inferred from direct a 0000166 // nucleotide binding // inferred from electr			8.669295957	0.516787526	5.985285461
5	4	121_at	X69699	0001655 // urogenital system develop 0005634 // nucleus // inferred from direct assa 0000979 // RNA polymerase II core promoter sequen			7.994419921	0.472133795	5.913773313
6	5	1255_g_at	L36861	0007165 // signal transduction // non 0001750 // photoreceptor outer segment // ini 0005509 // calcium ion binding // inferred from elect			8.230147831	0.509740428	6.145326687
7	6	1294_at	L13852	0006464 // cellular protein modificati 0005634 // nucleus // not recorded /// 000582 0000166 // nucleotide binding // inferred from electr			7.802159916	0.463418591	5.964311554
8	7	1316_at	X55005	0000122 // negative regulation of trai 0000790 // nuclear chromatin // inferred from 0000978 // RNA polymerase II core promoter proxim			7.879398937	0.472991467	5.984810695
9	8	1320_at	X79510	0006470 // protein dephosphorylatio 0005737 // cytoplasm // inferred from electroi 0004721 // phosphoprotein phosphatase activity // ir			8.024311178	0.496020918	6.124795055

	A	B	C	AQ	AR	AS	AT	AU	AV	AW	AX
1	X	ProbName	GB_ACC	PctCovGene_POS1pt25	MeanGene_POS2pt5	SDGene_POS2pt5	PctCovGene_POS2pt5	MeanGene_POS8	SDGene_POS8	PctCovGene_POS8	SimInfo3SProbName
2	1	1007_s_at	U48705	5.83899504	11.84305661	1.09112552	9.27782626	6.590821824	0.338736943	5.049174272	1007_s_at
3	2	1053_at	M87338	4.968958213	12.04718833	0.997069604	8.351734857	6.319599277	0.288928599	4.558093163	1053_at
4	3	117_at	X51757	5.1943899	11.73047585	1.102414067	9.406260552	6.660521005	0.331754215	4.967085482	117_at
5	4	121_at	X69699	5.329930903	12.33044129	1.091238795	8.877515861	6.434569636	0.319127309	4.951461223	121_at
6	5	1255_g_at	L36861	4.971031566	11.92141002	1.065316339	9.039411235	6.280748573	0.296848728	4.710588532	1255_g_at
7	6	1294_at	L13852	5.478767316	11.66342412	1.091913229	9.370684789	6.023331639	0.284293575	4.717888647	1294_at
8	7	1316_at	X55005	5.540758576	11.64438011	1.062147019	9.240280957	6.11607002	0.28603188	4.688992325	1316_at
9	8	1320_at	X79510	5.011581969	13.26359646	1.203145883	9.113330698	6.287095595	0.311506625	4.88281488	1320_at

# columns total	50
# rows total	52033
# columns Gene meta data	11
# columns per simulated prediction	3
# simulated fold values	18
range of the simulated predicted target fold increases	-20 to + 20

Size

Type

Name

91,329 KB Microsoft Excel Comma Separated Values File AllFactorsSimGPL570GENEmap

A3. Setup Dashboard in Power BI for GPL570 platform

1. Link to download desktop version of Power BI (free). (accessed on 10 May 2018).
2. Home/Get Data/ csv select file.
3. Load Data
4. Depending on the kind of data exploration, select the visualization, fields and filters displayed on the right-hand side pane of the tool.
5. Email Go Mavankal to get the csv output file to load.

The steps in A2 pertain to the Windows version of the tool, however, a free version is also available for Mac users.