Automation of analysis of high throughput RNAi data using R and cellHTS.

Abstract: RNA interference (RNAi) screening is fast becoming the technology of choice for understanding gene functions and in drug discovery. The process involves targeting a single gene with a specific siRNA (small interfering RNA) and studying the effects of silencing the gene. These experiments are useful in defining the unknown gene function and identifying targets for drugs in specific cells. Since these experiments use high throughput technologies studying thousands of genes simultaneously the data generated from the same is also voluminous. Automation of the process is therefore much called for in an environment where many scientists are working on different siRNA libraries and multiple different cell types to answer various questions. This project aimed to automate the process of analysis of data obtained from high throughput RNAi screening and standardize the results so that investigators across the organization can compare their results. The project used R (version 2.13.1), which is a statistical analysis software freely downloaded and cellHTS which is an analysis software within the ‘Bioconductor’ package of R specific for analysis of RNAi experiments. Data was obtained from a study which used cells from five Ewings sarcoma cell lines plated on a 384 well plate and transfected with siRNAs from a library of commercial siRNA and Myeloma dataset which was looking to identify druggable targets in myeloma cells. The final outcome in the study was cell viability, measured as a luminescence value corresponding to each well (each siRNA). Statistical analysis of this study involves producing quality metrics graphs, normalization of data and identification of “hits”. The software was developed in open source R statistical tool. The software intakes the data in user provided format and converts it into a format compatible with the cellHTS package. The output includes quality metrics plots like histogram of intensities, MA plots, and correlation scatter plots between replicates. Normalization can be done using any of the options provided in the cellHTS package including robustZscore and Bscore. Further quality metrics assessing the overall quality of the assay including Z-factor and SSMD is also calculated. Finally, top ‘hits’, which identifies significant luminescence values is represented in a tabular format. This software thus streamlines the process of analyzing immense volumes of high throughput screening data for statistical analysis.

Introduction: RNA interference screening is fast becoming the technology of choice for development of new drug targets. RNA interference is a mechanism within living cells to silence genes. RNAi technology makes use of this natural mechanism to design specific RNAs targeting specific genes to study the effects of silencing the gene on cell morphology/survival. The experiment primarily involves mixing the cells of interest with a siRNA (short interfering RNA) and seeing the effects of the siRNA on the cell over a period of time. Since the experiment involves studying hundreds of genes simultaneously this experiment is called high throughput.

There is huge amount of data which needs to be analyzed fast using standardized methods. cellHTS is a software package which is used to analyze HTS RNAi data. cellHTS version 2.13.1 was used for this project. cellHTS is an open source software developed as a package within Bioconductor available at [www.bioconductor.org](http://www.bioconductor.org).

The program was written in R which is open source software for statistical analysis. Bioconductor is a subset of R developed exclusively for analysis of genomic data.

Scope of the project: Scope was to automate the analysis of HTS RNAi data and produce standardized reports.

Materials and methods: R and cellHTS2 were used for programming. R was downloaded from [www.r-project.org](http://www.r-project.org). cellHTS2 version 2.13.1 was downloaded as an R package from Bioconductor website [www.bioconductor.org](http://www.bioconductor.org).

The aim was to automate the analysis using cellHTS2 for the analysis process. cellHTS2 requires that the data be input in a very specific format (Plate configuration file, Platelist file, etc). The program was written with the purpose to create these files.

The program was designed to have the following structure:

1. Data file which stores the data in raw format
2. Main file which will call all the functions
3. Functions file with functions defining all the processes for making the files
4. Report file to store the final results

**MAIN FILE**

Init: initializes the program by setting the path and the directory.

Experiment: loads experiment specific profile

Data: takes raw data input and creates following datasets: raw, augmented and cellHTS

QC first pass: creates QC plots using raw data intensities

Normalization: normalizes data based on type of normalization method defined

QC second pass: creates QC plots using normalized data

Hit identification: based on scoring method scores the data and generates final reports

*Main file is generic and only the experiment name needs to be changed/added.*

**FUNCTIONS FILE**

Profile folder

This loads all the generic profile values like the total number of wells, number of rows and columns in a 384 well plate, normalization methods available etc.

This function also lets the program load the experiment specific profile based on the experiment name defined in the main folder using the ‘switch’ command.

Experiment specific profile

This function defines experiment specific parameters like the folder where the data is stored, the raw data file structure (profile$celllines\_inRawdata) which defines where the luminescence values are stored in the raw data file, positive and negative controls used in the experiment.

The function also creates augmented file from raw data file. The structure of the augmented file is defined using profile$augDataNULLtemplate. The next step is to fill the augmented file using raw2augment function. Plate configuration file of the cellHTS format is also created in this function. Refer the cellHTS documentation for structure of Plate configuration file. Briefly the file defines which wells in each plate has ‘sample’, ‘positive’ control, ‘negative’ control, and ‘other’. This file accepts regular expressions.

Data function

This functions calls functions from experiment specific profiles to create raw, augmented and cellHTS data files and thus returns raw, augmented and cellHTS data as a list. This function also has the switch command to call experiment specific function to create augmented files and plate configuration file (raw2augment, create\_cellHTS\_Plateconfig).

Next step is to create files in cellHTS format from augmented file. First we create the directories for each specific cell line. Next step is creating the specific data file folders from the augmented file. This creates subsets of the augmented data according to plate and replicates. Data subsetting is done based on the columns which have data for specific cell lines and creating specific folders for the same.

Next step creates cellHTS objects using cellHTS commands. Refer cellHTS documentation for details on the command. Briefly readplatelist reads the raw data, configure command configures the data using plateconfiguration file, normalizeplates normalizes the data using defined normalization method, scoreReplicates scores the data and summarizeReplicates summarizes the data based on method defined.

Generate report function

This function uses the writeReport function of cellHTS to create reports, raw, normalized and scored. This is done by cell line. This function also creates box plots of raw intensities by type of probe (sample, control) and dot plot of intensities across all plates.

Utilities function

This function checks for compatibility of normalization method and transformation of data. Refer to normalization methods used in cellHTS document (excel sheet) for all methods used and type of data transformation required.

**DATA FILE**

Data file stores the raw data as an excel sheet and also has folders created using data function in functions folder. These data folders are created by cell line and within each cell line folder is cellHTS specific data files by plate and replicate.

**REPORT FILE**

Finally the report file contains report by cell line and arranged by type of data (raw, normalized, scored). Normalized report folder further contains normalized report folders according to the type of normalization method used. Similar is the case for scored report.

Thus the process of creation of cellHTS specific files is automated. Reports are also produced automatically and can be customized to the needs of the experimental design. In future more analysis methods can be implemented using the cellHTS data format as the data can be accessed easily form cellHTS objects.