Normalizing Affy microarray data

All product names are given as examples only and they are not endorsed by the USDA or the University of Illinois.

INTRODUCTION

The following is an interactive demo describing a set of steps that we run to normalize Affymetrix oligonucleotide array data. We use either the RMA (Robust Multi-Array) normalization (retains probe level information; requires large amounts of RAM memory) or GCRMA (uses GC content of probes in normalization with RMA; gives one value for each probe set instead of keeping probe level information) normalization in the R packages *affy* and *gcrma*. The final data file is ready to be used as an input file for SAS. The SAS programs we run are explained on another page.
Click here for use SAS to analyze normalized Affymetrix data after RMA normalization.
Click here for use SAS to analyze normalized Affymetrix data after GCRMA normalization.

Please feel free to contact me with any questions or comments:
Steve Clough (sjclough@uiuc.edu)

DOWNLOAD DEMO SET

Click here to obtain a demo data set off .CEL files that you may use to test and learn how to normalize Affymetrix data.
The CEL file describes the intensities determined for every feature on a chip, without providing information about which probes correspond to which probe sets (such information provided by the CDF file). *Click for Affymetrix description*.

DOWNLOAD AFFY AND GCRMA IN R

To run these analyses you will need to download the FREE *affy* and *gcrma* package in R for the Affymetrix oligonucleotide array probe level data analysis, developed as part of the Bioconductor project.

The Bioconductor project website (http://www.bioconductor.org/) has links to various documents related to R/affy and R/gcrma and the statistical analyses.

R/affy. *(Click for explanations on how to download and install)*

R/gcrma. *(Click for explanations on how to download and install)*

RUNNING R/gcrma TO NORMALIZE THE DATA

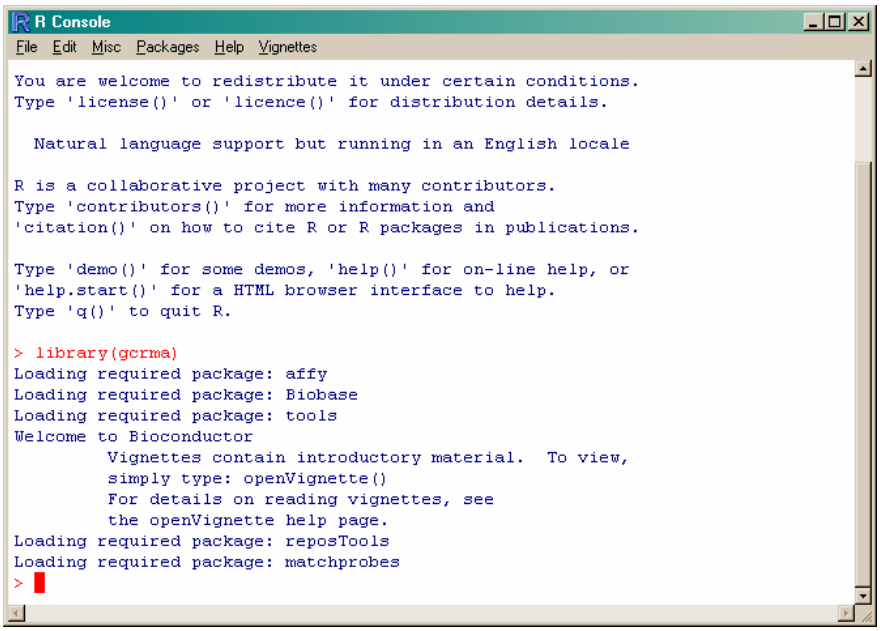
Note: the following descriptions and demo have been developed based on R version R 2.1.1.

Click here for the R/gcrma functional codes. Once you are familiar with R/gcrma this set of codes (called ".Rhistory") is all you'll need to run the normalization.

1. **PUT FILES INTO SINGLE FOLDER/DIRECTORY.**
   To run R, you need to have all the .CEL files in the same folder/directory (i.e. C:\temp\Demo\CEL_Folder).

2. **RUN GCRMA PACKAGE IN R.**
   - The first step is to load the gcrma library by opening R and simply typing:

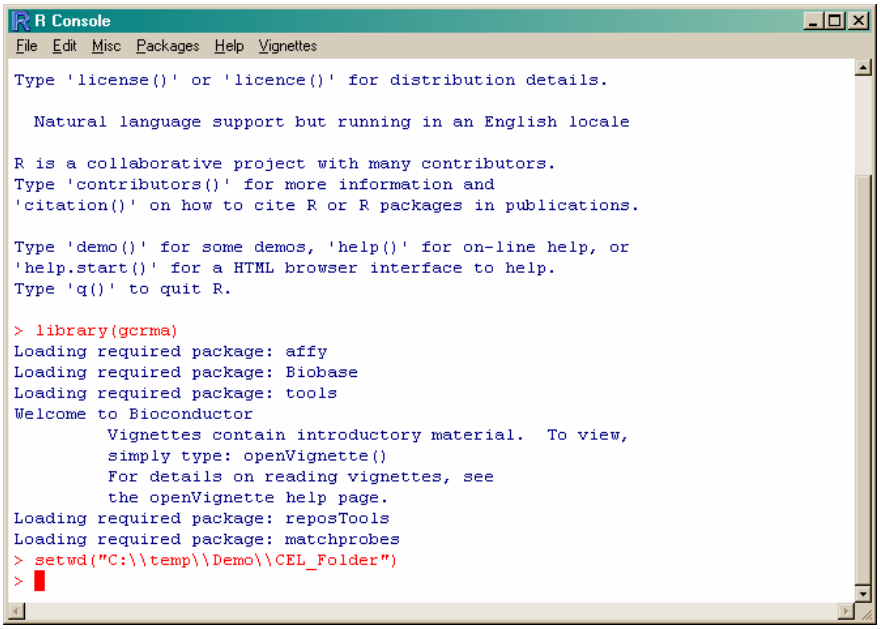
     ```
     >library(gcrma)
     ```



     Loads the required packages to run the gcrma package.

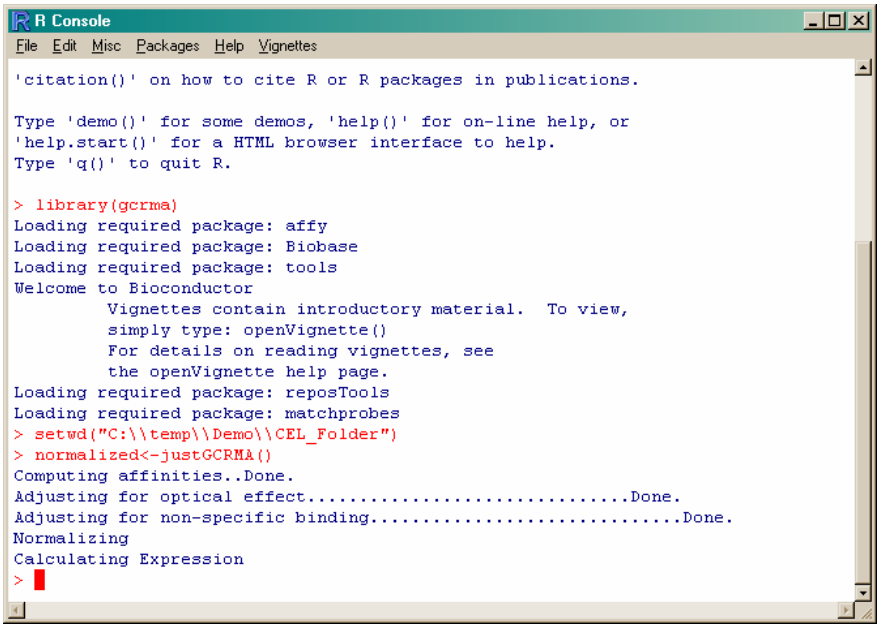
   - Set (identify) the working folder/directory where the data are located using double backslashes (i.e. C:\\temp\\Demo\\CEL_Folder)

     ```
     >setwd("C:\\temp\\Demo\\CEL_Folder")
     ```



- Normalize the data with justGCRMA() function. This function normalizes the data using the Robust Multi-Array (RMA) expression measure taking into account the GC content of the probe sequences. We prefer using justGCRMA() function because it uses less RAM memory than the standard gcrma() function. The expression measures obtained are $\log_2$ transformed.

```
>normalized=justGCRMA()
```



- To have the expression measure in an Excel readable format, save the temporary file with the normalized data as a .csv file (i.e. NormalGCRMA.csv).

```
>exprs2excel(normalized,file="NormalGCRMA.csv")
```