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/affy TO NORMALIZE THE DATA MAINTAINING PROBE INFORMATION

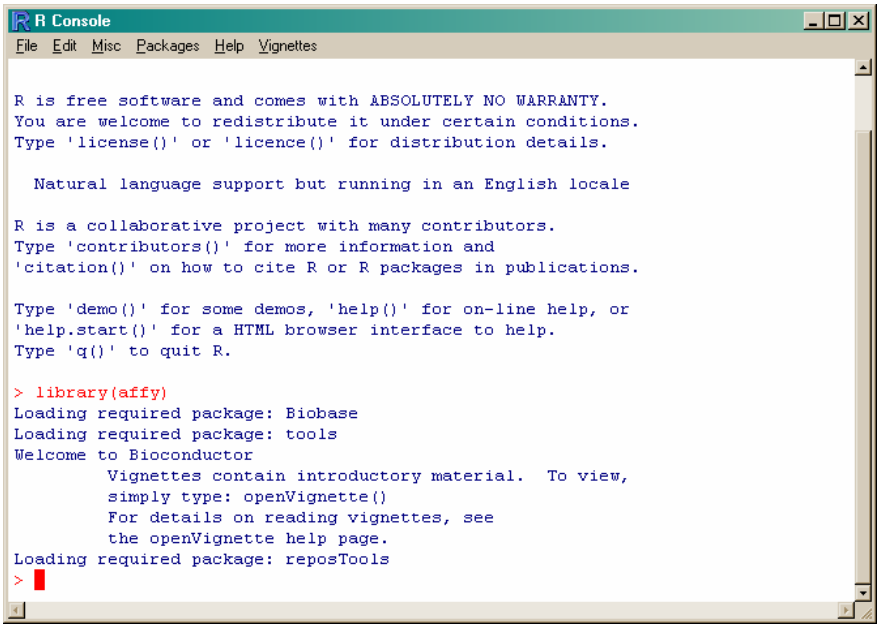
Note: the following descriptions and demo have been developed based on R version R 2.1.1.

Click here for the R/affy functional codes. Once you are familiar with R/affy 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 AFFY PACKAGE IN R.**
   - The first step is to load the affy library by opening R and simply typing:

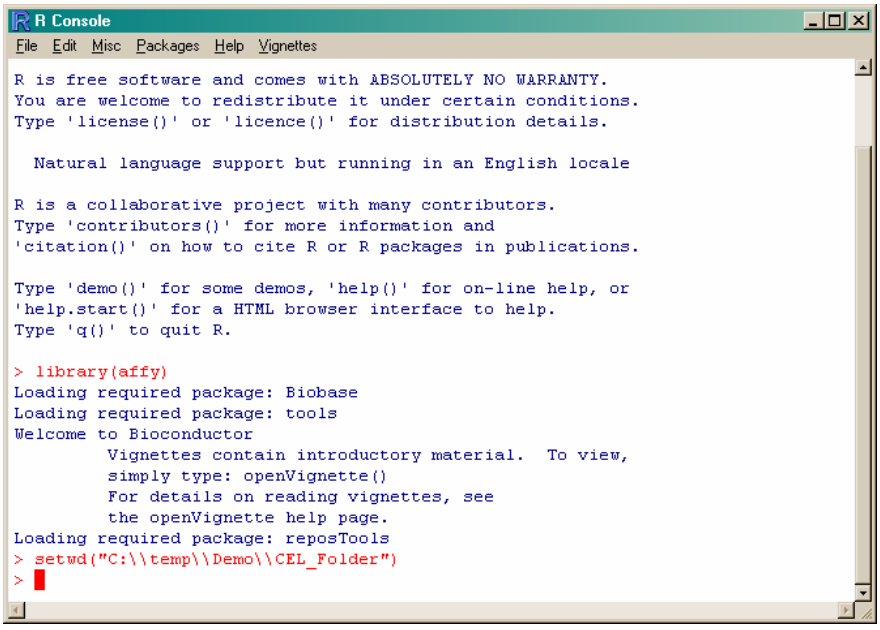
     ```
     >library(affy)
     ```



   Loads the entire required packages to run the affy package.

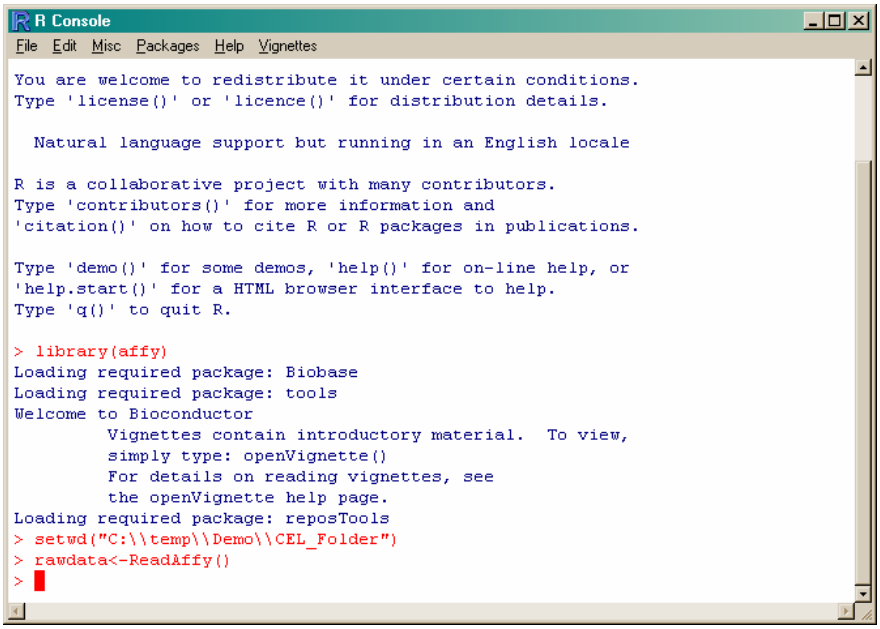
- Set (identify) the working folder/directory where the data are located using double backslashes (i.e. C:\\temp\\Demo\\CEL_Folder)

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



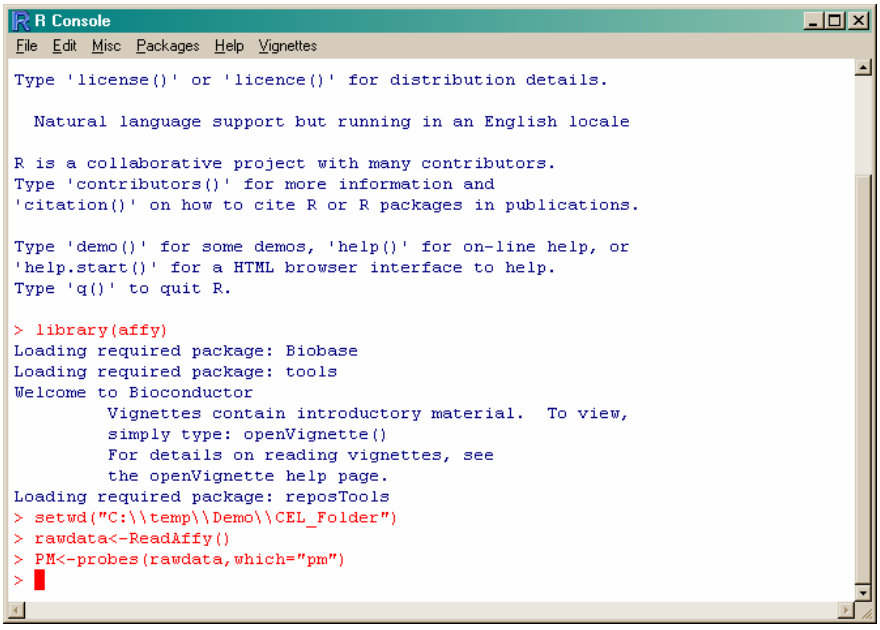
- Read the raw data into the file “rawdata”.

```
>rawdata<-ReadAffy()
```



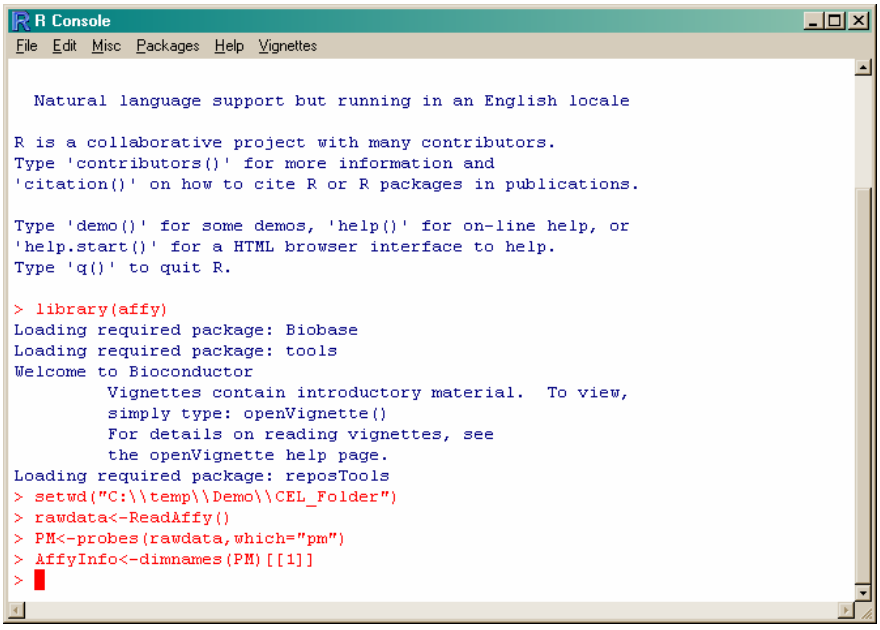
- Extract perfect match probe intensities from the “rawdata” file into a new file called “PM”.

```
>PM<-probes(rawdata,which="pm")
```



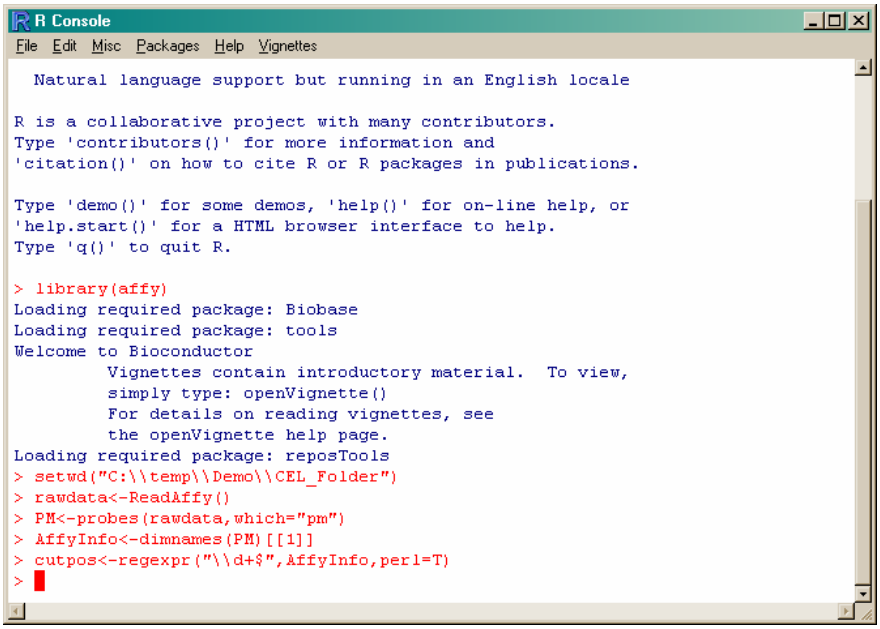
- Retrieve the gene IDs from the first column of the “PM” file and save in “AffyInfo” file.

```
>AffyInfo<-dimnames(PM)[[1]]
```



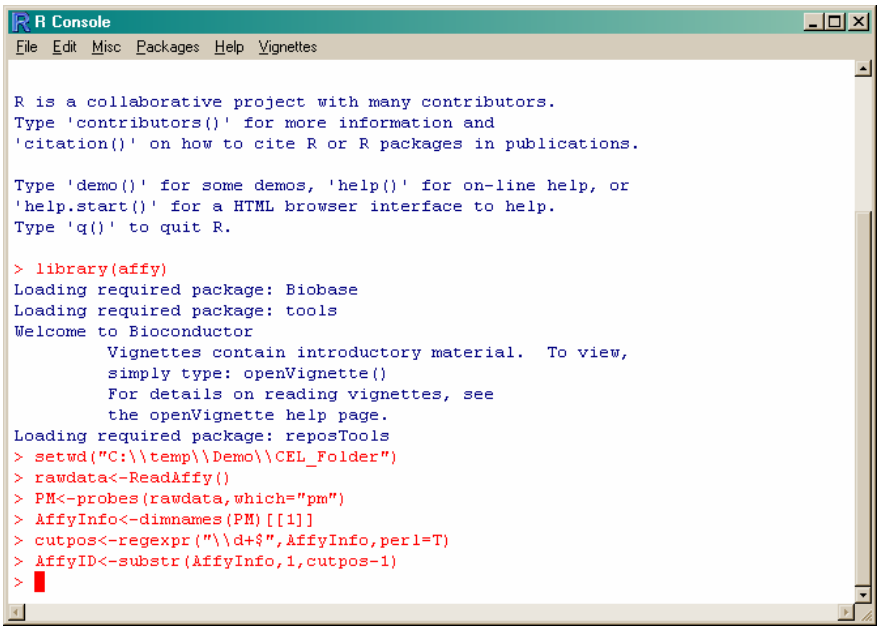
- Look for number of digits following each probe name within “AffyInfo” data set and save into “cutpos” file. Gives a -1 if no digits follow the name.

```
>cutpos<-regexpr("\\d+$",AffyInfo,perl=T)
```



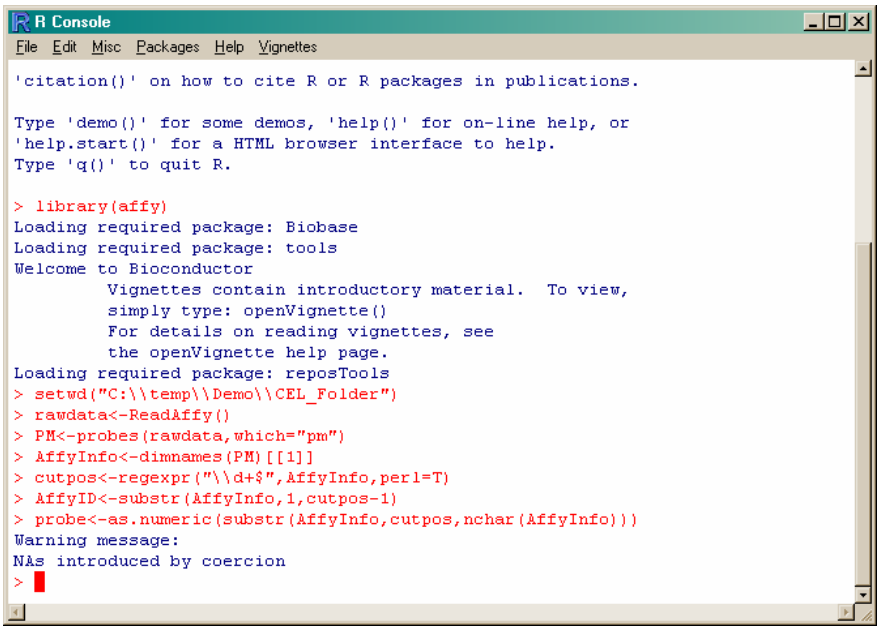
- Extract the digits following the probe names in “AffyInfo” data set and save these IDs as “AffyID” file.

```
>AffyID<-substr(AffyInfo,1,cutpos-1)
```



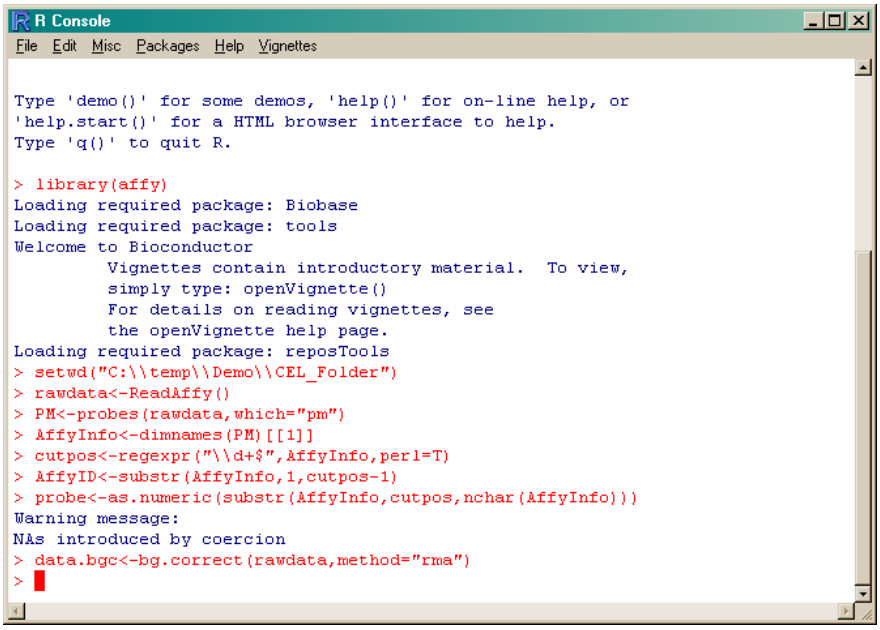
- Take numeric objects from the “AffyInfo” data set, which are the probe IDs (most are 1-11) and save as “probe” file.

```
>probe<-as.numeric(substr(AffyInfo,cutpos,nchar(AffyInfo)))
```



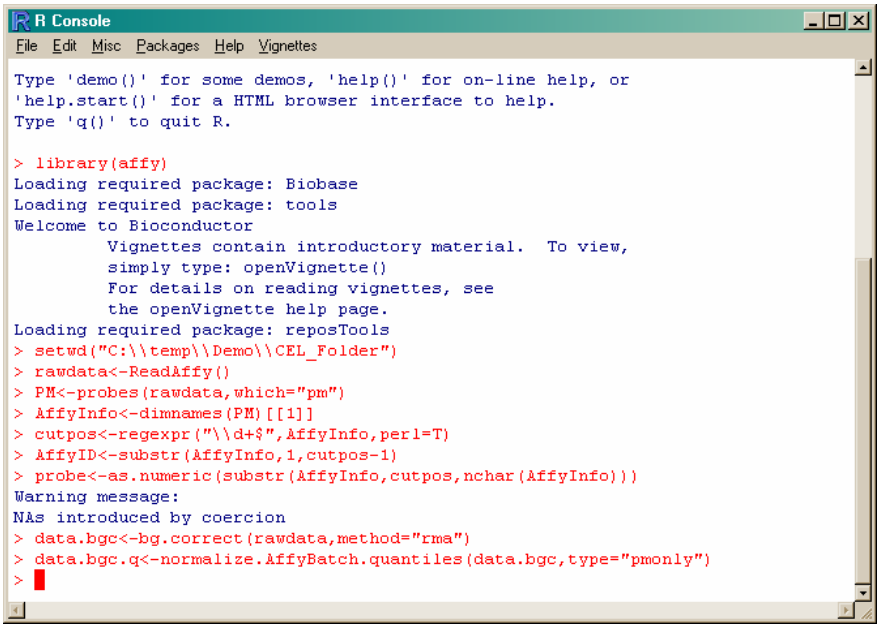
- Raw data background corrects probe intensity values with the RMA method.

```
>data.bgc<-bg.correct(rawdata,method="rma")
```



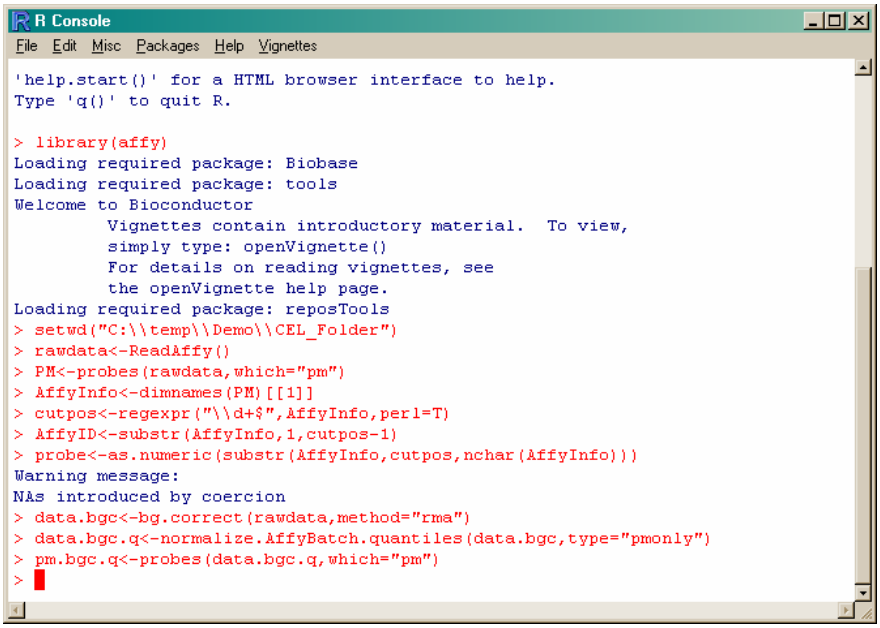
- Normalize the perfect match probe level intensities based upon quantiles method.

```
>data.bgc.q<-normalize.AffyBatch.quantiles(data.bgc,type="pmonly")
```



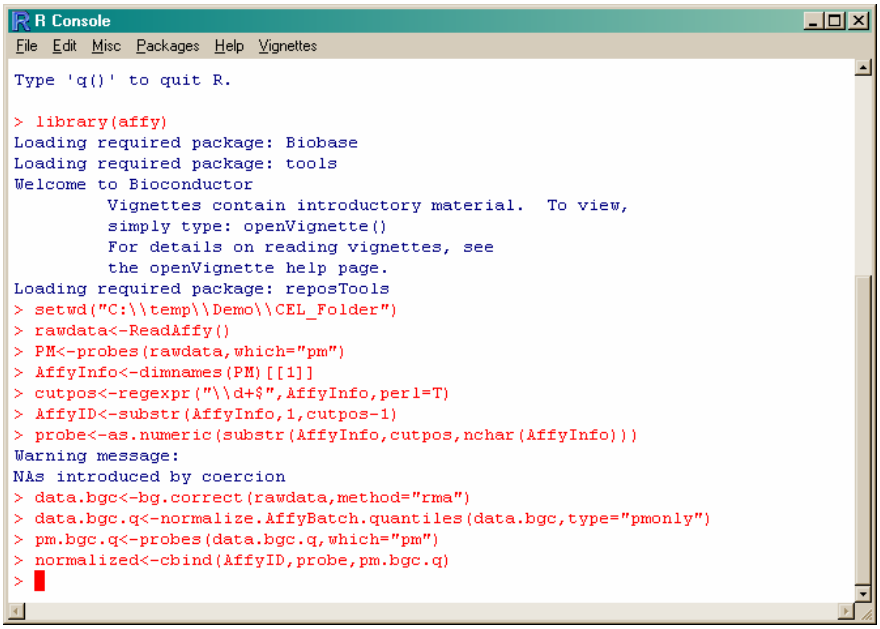
- Extract perfect match probe intensities from the “data.bgc” file into a new file called “pm.bgc.q”.

```
>pm.bgc.q<-probes(data.bgc.q,which="pm")
```



- Combind “AffyID”, “probe”, and “pm.bgc.q” file into a new file called “normalized”, which contains the normalized PM data.

```
>normalized<-cbind(AffyID,probe,pm.bgc.q)
```



- To have the expression measure in an Excel readable format, save the “normalized” file as a .csv file (i.e. NormalR.csv)

```
>write.table(normalized,file="NormalR.csv",sep=",",row.names=FALSE,
quote=FALSE)
```