Creating Data Sets for *pathGSARF* Analysis

Andy J. Minn, M.D., Ph.D.

Abramson Family Cancer Research Institute, University of Pennsylvania

Email contact: andyminn@exchange.upenn.edu

March 12, 2011

1 Introduction

There are several microarray expression data sets used for the RKIP metastasis signaling pathway analysis using the *pathGSARF* package. All data is on the Affymetrix platform except for the BrCa295 data sets. The BrCa295 data set is used in a clinical utility study that is part of a second manuscript. For the Affymetrix data, .CEL files are processed using the RMA method. In order to facilitate reproducibility, the method and recommendations of Kostka and Spang are used in order to provide documentation by value [1]. Three functions are used for RMA processing after an AffyBatch object is created. We provide a way to ensure that the processed data matches the original data used in our analysis.

The .CEL files for the BrCa443 data were not available at the time our analysis was initiated; therefore, we used MAS5 processed data from our previous study [2]. Results using this data are what is described in the Yun et al. manuscript. After the .CEL files recently became available, we used the processing methods described here to re-run our analysis with the BrCa443 data. Similar results were obtained and the user can optionally use either data set as described in the accompanying Sweave document. To process the BrCa443 from the raw data, the same methods for the BrCa871 data set are used.

2 GEO Accessions, Loading Libraries, and Meta-data

The .CEL file are retrieved from the Gene Expression Omnibus (GEO) website. The accession numbers for each data set are itemized below. Extract the files into a single folder for each data set and specify the path. There might be more .CEL files than samples in each data set, e.g. BrCa871, due to duplicate samples. This is addressed by matching filenames with the row names from a data frame provided for each data set. Note that compressed files are used, so the ”.gz” extension is kept.

- **BrCa871**: GEO accession numbers: GSE1456, GSE2990, GSE3494, GSE7390, GSE11121
- **BrCa443**: GEO accession number: GSE2034, GSE2603, GSE5327. The original processed data used in the Yun et al. manuscript is provided in the *pathGSARF* package.
- **BrCa295**: This is a non-Affymetrix platform and is provided in the *pathGSARF* package.

```r
> library(pathGSARF, lib.loc = "/Users/minna/R develop")
> library(hgu133a.db)
> data("Phenotype data")
> data("scl rma and u133a pmindex")
```

3 BrCa871 ExpressionSet

After processing the BrCa871 data set, the reference RMA scale is compared to the new one created (Figure 1). The two should match. It should be noted that the memory requirements for AffyBatch files can be considerable. If memory is an issue, the `justRMA()` function from the *affy* package can be used; however, the results of RMA processing cannot be verified against the reference RMA scale. An alternative method
to process the .CEL files that should be less memory intensive is to modify the method described for the BrCa344 data set so that each .CEL file is read one-at-a-time.

```r
> setwd("/Volumes/WD 300G/To process BrCa871")
> fnames <- list.files()
> mtc <- match(rownames(pdata.871), fnames)
> phenodata <- new("AnnotatedDataFrame", pdata.871)
> all(rownames(pData(phenodata)) == fnames[mtc])
[1] TRUE

> edat <- ReadAffy(filenames = fnames[mtc], phenoData = phenodata)
> edat <- bg.correct(edat, "rma")
> gc()

> edat <- quantileNormAffyBatch(edat)
> gc()

> scl.rma.new <- preproc(description(edat))$val

> plot(scl.rma$mqnts, scl.rma.new$mqnts, xlab = "Reference RMA scale",
+ ylab = "New RMA scale")

> symbol <- mget(featureNames(edat), hgu133aSYMBOL)
> genename <- mget(featureNames(edat), hgu133aGENENAME)
> locuslink <- mget(featureNames(edat), hgu133aENTREZID)
> title <- "RMA normalized BrCa871"
> fData <- data.frame(unlist(symbol), unlist(genename), unlist(locuslink))
> rownames(fData) <- featureNames(edat)
> colnames(fData) <- c("Symbol", "Gene Name", "Entrez ID")
> phenoDescrip <- c("cohort", "time to relapse", "relapse event",
+ "time to distant metastasis", "distant metastasis event",
+ "time to disease-specific death", "disease-specific survival event")
> metadata <- data.frame(labelDescription = phenoDescrip, row.names = colnames(pdata.871))
> metadata.feat <- data.frame(labeDescription = colnames(fData),
+ row.names = colnames(fData))
> phenoData <- new("AnnotatedDataFrame", data = pdata.871, varMetadata = metadata)
> featData <- new("AnnotatedDataFrame", data = fData, varMetadata = metadata.feat)
> experimentData <- new("MIAME", name = "Andy Minn", lab = lab,
+ contact = "none", title = title, url = "none")
> annotation <- "hgu133a.db"
> brca871rmaSet <- new("ExpressionSet", exprs = exprs(edat), phenoData = phenoData,
+ featureData = featData, experimentData = experimentData,
+ annotation = annotation)
> gc()
```

```r
called (Mb) gc trigger (Mb) max used (Mb)
Ncells 1981515 105.9 3493455 186.6 3493455 186.6
Vcells 43913682 335.1 1611397216 12294.0 2500646023 19078.5
```
Figure 1: Plot of reference RMA scale to new RMA scale for BrCa871.
4 BrCa443 ExpressionSet

```r
> data("BrCa443 scl rma")
> setwd("/Volumes/WD 300G/To process BrCa443")
> fnames <- list.files()
> mtch <- match(rownames(pdata.emcmskcc), fnames)
> phenodata <- new("AnnotatedDataFrame", pdata.emcmskcc)
> all(rownames(pData(phenodata)) == fnames[mtch])
[1] TRUE

> edat <- ReadAffy(filenames = fnames[mtch], phenoData = phenodata)
> edat <- bg.correct(edat, "rma")
> gc()

used (Mb) gc trigger (Mb) max used (Mb)
Ncells 2247366 120.1 3708127 198.1 3708127 198.1
Vcells 251191115 1916.5 1137133074 8675.7 2500646023 19078.5

> edat <- quantileNormAffyBatch(edat)
> gc()

used (Mb) gc trigger (Mb) max used (Mb)
Ncells 2247366 120.1 3708127 198.1 3708127 198.1
Vcells 251191115 1916.5 1137133074 8675.7 2500646023 19078.5

> edat <- summarizeAffyBatch(edat)
> scl.rma.new <- preproc(description(edat))$val

> plot(scl.rma443$mqnts, scl.rma.new$mqnts, xlab = "Reference RMA scale",
+ ylab = "New RMA scale")

> symbol <- mget(featureNames(edat), hgu133aSYMBOL)
> genename <- mget(featureNames(edat), hgu133aGENENAME)
> locuslink <- mget(featureNames(edat), hgu133aENTREZID)
> title <- "RMA normalized BrCa443"
> lab <- "Abramson Family Cancer Research Institute"
> fData <- data.frame(unlist(symbol), unlist(genename), unlist(locuslink))
> rownames(fData) <- featureNames(edat)
> colnames(fData) <- c("Symbol", "Gene Name", "Entrez ID")
> phenoDescrip <- c("identifier", "cohort", "time to metastasis",
+ "metastasis event")
> metadata <- data.frame(labelDescription = phenoDescrip, row.names = colnames(pdata.emcmskcc))
> phenoData <- new("AnnotatedDataFrame", data = pdata.emcmskcc, + varMetadata = metadata)
> featData <- new("AnnotatedDataFrame", data = fData, varMetadata = metadata.feat)
> experimentData <- new("MIAME", name = "Andy Minn", lab = lab,
+ contact = "none", title = title, url = "none")
> annotation <- "hgu133a.db"
> EMCMSKCCrmaSet <- new("ExpressionSet", exprs = exprs(edat), phenoData = phenoData, + featureData = featData, experimentData = experimentData, + annotation = annotation)
> gc()

used (Mb) gc trigger (Mb) max used (Mb)
Ncells 2003619 107.1 3708127 198.1 3708127 198.1
Vcells 45744011 349.0 909706459 6940.6 2500646023 19078.5
```
Figure 2: Plot of reference RMA scale to new RMA scale for BrCa443.
References
