Pre-process the GDSC data using BrainArray.

Paul Geeleher

August 12, 2013

Pre-process the GDSC data using the BrainArray CDF file and map the gene identifiers to gene symbols using thier annotation package. The CDF file and annotation package can be downloaded from:
http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/17.0.0/entrezg.asp

The CDF file is here:
http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/17.0.0/entrezg.download/hthgu1

The annotation file is here:
http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/17.0.0/entrezg.download/hthgu1

Once downloaded, these must be installed in the same was as any R package, for example, by using "R CMD install pacakgeName.tar.gz" from the command line.

The GDSC raw CEL files can be downloaded from ArrayExpress (accession number is E-MTAB-783). These are contained in 9 separate files (E-MTAB-783.raw.1.zip ... E-MTAB-783.raw.9.zip) which must be downloaded and unzipped into the same directory. This data is huge and should be processed on a machine with at least **32Gb** of RAM.

NB: Make sure to include the trailing backslash or forwardslash in the *celFilesLoc* variable below OR IT WILL NOT WORK.

```
> library(affy)
> celFilesLoc <- "/nas40t0/GDSC_Sanger/Expression/"
> celFileName <- paste(celFilesLoc, dir(celFilesLoc, pattern=".CEL$"), sep="")
> gdsc_raw <- ReadAffy(filenames=celFileName)
> library(affy)
> gdsc_raw@cdfName <- "HTHGU133A_Hs_ENTREZG"
> gdsc_brainarray_entrez <- rma(gdsc_raw)

Background correcting
Normalizing
Calculating Expression

> library("hthgu133ahsentrezg.db")
> x <- hthgu133ahsentrezgSYMBOL
> mapped_probes_brain <- mappedkeys(x)
> names(mapped_probes_brain) <- as.character(x[mapped_probes_brain])
> entrez2sym_brain <- as.character(x[mapped_probes_brain])
> symBrain2entrez <- entrez2sym_brain
> names(symBrain2entrez) <- names(entrez2sym_brain)
> gdsc_brainarray_syms <- exprs(gdsc_brainarray_entrez)
> rownames(gdsc_brainarray_syms) <- symBrain2entrez[rownames(exprs(gdsc_brainarray_entrez)
```

Finally, save the data.

```
> save(gdsc_brainarray_syms, file="/userhome/pgeeleher/NatMedPaper/gdsc_brainarray_syms.RD
```