Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study Ross-Adams et al. (2015) doi:10.1016/j.ebiom.2015.07.017

In this document, I describe how the GEO data entry for the Stockholm (validation) cohort of Ross-Adams et al (2015) data was processed and saved into an object for analysis in Bioconductor. First of all load the relevant libraries for grabbing and manipulaing the data

```
> library(GEOquery)
```

Now use the `getGEO` function with the correct ID. If the series matrix file has already been downloaded, don't bother to download again.

```
>   url <- "ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE70nnn/GSE70769/matrix/"
>   destfile <-"GSE70769_series_matrix.txt.gz"
>   if(!file.exists(destfile)){
+   download.file(paste(url,destfile,sep=""),destfile=destfile)
+   }
> geoData <- getGEO(filename=destfile)
```

We tidy up the data from GEO; creating a new data frame of just the clinical characteristics of interest. Make sure that only `NA` is used for missing values.

The pheno data also contains the five `iCluster` groups which were determined by integrative clustering.

```
> pd <- pData(geoData)
> pd2 <- data.frame("geo_accession" = pd$geo_accession,
+     Sample = gsub("tumour tissue_robotic radical prostatetctomy_","",pd$title),
+   Gleason=gsub("tumour gleason: ","",pd$characteristics_ch1),
+   ECE=gsub("extra capsular extension (ece): ","",pd$characteristics_ch1.2,fixed=TRUE),
+   PSM = gsub("positive surgical margins (psm): ","",pd$characteristics_ch1.3,fixed=TRUE),
+   BCR = gsub("biochemical relapse (bcr): ","",pd$characteristics_ch1.4,fixed=TRUE),
+   Time = gsub("time to bcr (months): ","",pd$characteristics_ch1.5,fixed=TRUE),
+   iCluster = gsub("derived data (iclusterplus group): ","",pd$characteristics_ch1.6,fixed=
+   PSA=gsub("psa at diag: ", "",pd$characteristics_ch1.7,fixed=TRUE),
+   ClinicalStage = gsub("clinical stage: ","", pd$characteristics_ch1.8,fixed=TRUE),
+   PathStage = gsub("pathology stage: ","",pd$characteristics_ch1.9,fixed=TRUE),
+   FollowUpTime = gsub("total follow up (months): ","",pd$characteristics_ch1.10,fixed=TRUE
> pd2$iCluster <- gsub("NA",NA,pd2$iCluster)
> pd2$Gleason <- gsub("N/A", NA, pd2$Gleason)
> pd2$Gleason <- gsub("unknown",NA,pd2$Gleason)
> pd2$ECE <- gsub("UNKNOWN",NA,pd2$ECE)
> pd2$ECE[which(pd2$ECE == "")] <- NA
> pd2$PSM <- gsub("UNKNOWN",NA,pd2$PSM)
> pd2$PSM[which(pd2$PSM == "")] <- NA
> pd2$BCR <- gsub("N/A", NA, pd2$BCR)
> pd2$BCR[which(pd2$BCR == "")] <- NA
```

```
> pd2$Time <- gsub("N/A", NA, pd2$Time)
> pd2$Time <- gsub("UNKNOWN",NA,pd2$Time)
> pd2$Time[which(pd2$Time == "")] <- NA
> pd2$FollowUpTime <- gsub("N/A", NA, pd2$FollowUpTime)
> pd2$FollowUpTime <- gsub("UNKNOWN",NA,pd2$FollowUpTime)
> pd2$FollowUpTime[which(pd2$FollowUpTime == "")] <- NA
> rownames(pd2) <- pd2$geo_accession
> pData(geoData) <- pd2
> stockholm <- geoData
> save(stockholm, file="data/stockholm.rda",compress="xz")
```