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

> Understanding the heterogeneous genotypes and phenotypes of prostate cancer is fundamental to improving the way we treat this disease. As yet, there are no validated descriptions of prostate cancer subgroups derived from integrated genomics linked with clinical outcome. In a study of 482 tumour, benign and germline samples from 259 men with primary prostate cancer, we used integrative analysis of copy number alterations (CNA) and array transcriptomics to identify genomic loci that affect expression levels of mRNA in an expression quantitative trait loci (eQTL) approach, to stratify patients into subgroups that we then associated with future clinical behavior, and compared with either CNA or transcriptomics alone.

In this document, I describe how the GEO data entry for the Cambridge cohort (training set) dataset 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. Check if the series matrix has been downloaded already, and download if not.

```
>   url <- "ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE70nnn/GSE70768/matrix/"
>   destfile <-"GSE70768_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. This involves removing the prefixes GEO gives to each variable and setting `NA` values correcly. We also check for some columns that the values are within an acceptable set and exclude any weird values. This was required as we saw that one sample seemed to have psa measurements in what would normally be the iCluster group column.

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

NOTE: Would be neater to use dplyr and tidyr for this in future

```
> pd <- pData(geoData)
> Group <- gsub("sample type: ", "", pd$characteristics_ch1)
> Group[grep("CRPC",pd$title)] <- "CRPC"
> pd2 <- data.frame("geo_accession" = pd$geo_accession, Sample = pd$description,
```

```
+       Sample_Group = Group,Gleason=gsub("tumour gleason: ","",pd$characteristics_ch1.1),
+       iCluster = gsub("(derived data) iclusterplus group: ","",pd$characteristics_ch1.3,fi
+       ECE=gsub("extra-capsular extension (ece): ","",pd$characteristics_ch1.4,fixed=TRUE),
+       PSM = gsub("positive surgical margins (psm): ","",pd$characteristics_ch1.5,fixed=TRU
+       BCR = gsub("biochemical relapse (bcr): ","",pd$characteristics_ch1.6,fixed=TRUE),
+       TotalTime = gsub("time to bcr (months): ","",pd$characteristics_ch1.7,fixed=TRUE),
+       ERG = gsub("tmprss2: ERG gene fusion status: ","",pd$characteristics_ch1.8,fixed=TRU
+       Age = gsub("age at diag: ","",pd$characteristics_ch1.9,fixed=TRUE),
+       PSA = gsub("psa at diag: ","",pd$characteristics_ch1.10,fixed=TRUE),
+       ClinicalStage = gsub("clinical stage: ","",pd$characteristics_ch1.11,fixed=TRUE),
+       PathStage = gsub("clinical stage: ","",pd$characteristics_ch1.12,fixed=TRUE),
+       FollowUpTime =gsub("total follow up (months): ","",pd$characteristics_ch1.13,fixed=T
+                 )
> pd2$iCluster <- gsub("N/A", NA, pd2$iCluster)
> pd2$iCluster[which(pd2$iCluster == "")] <- NA
> weirdValue <- setdiff(pd2$iCluster,
+         c("clust1","clust2","clust3","clust4","clust5",NA))
> if(length(weirdValue) > 0) pd2$iCluster[pd2$iCluster %in% weirdValue] <- NA
> pd2$Gleason <- gsub("N/A", NA, pd2$Gleason)
> weirdValue <- setdiff(pd2$Gleason,
+       c("10=5+5","6=3+3","7=3+4","7=4+3","8=3+5","9=5+4",NA))
> if(length(weirdValue) > 0) pd2$Gleason[pd2$Gleason %in% weirdValue] <- NA
> pd2$Gleason <- factor(pd2$Gleason,
+     levels = c("6=3+3","7=3+4","7=4+3","8=3+5","9=5+4","10=5+5"))
> pd2$ECE <- gsub("unknown",NA,pd2$ECE)
> pd2$ECE[which(pd2$ECE == "")] <- NA
> weirdValue <- setdiff(pd2$ECE, c("N","Y",NA))
> if(length(weirdValue) > 0) pd2$ECE[pd2$ECE %in% weirdValue] <- NA
> pd2$PSM <- gsub("unknown",NA,pd2$PSM)
> pd2$PSM[which(pd2$PSM == "")] <- NA
> weirdValue <- setdiff(pd2$PSM, c("N","Y",NA))
> if(length(weirdValue) > 0) pd2$PSM[pd2$PSM %in% weirdValue] <- NA
> pd2$BCR <- gsub("N/A", NA, pd2$BCR)
> pd2$BCR[which(pd2$BCR == "")] <- NA
> weirdValue <- setdiff(pd2$BCR, c("N","Y",NA))
> if(length(weirdValue) > 0) pd2$BCR[pd2$BCR %in% weirdValue] <- NA
> pd2$TotalTime <- gsub("N/A", NA, pd2$TotalTime)
> pd2$TotalTime[which(pd2$TotalTime == "")] <- NA
> pd2$FollowUpTime[which(pd2$FollowUpTime=="")] <- NA
> rownames(pd2) <- pd2$geo_accession
```

Finally, we replace the pheno data of the object and save; using compression

```
> pData(geoData) <- pd2
> camcap <- geoData
> save(camcap, file="data/camcap.rda",compress="xz")
```