Integrative genomic profiling of human prostate cancer
Taylor et al. (2010) doi:10.1016/j.ccr.2010.05.026

Current knowledge of prostate cancer genomes is largely based on relatively small patient cohorts using single modality analysis platforms. Here we report concordant assessment of DNA copy number, mRNA and microRNA expression and focused exon resequencing in prostate tumors from 218 patients with primary or metastatic prostate cancer with a median of 5 years clinical follow-up, now made available as a public resource. Mutations in known, commonly mutated oncogenes and tumor suppressor genes such as PIK3CA, KRAS, BRAF and TP53 are present but generally rare. However, integrative analysis of mutations with copy number alterations (CNAs) and expression changes reveal alterations in the PI3K, RAS/RAF and androgen receptor (AR) pathways in nearly all metastatic samples and in a higher frequency of primary samples than previously suspected based on single-gene studies. Other new findings include evidence that the nuclear receptor coactivator NCOA2 functions as a driver oncogene in 20 percent of primaries. Tumors with the androgen-driven TMPRSS2-ERG fusion were significantly associated with a small, previously unrecognized, prostate-specific 3p14 deletion that, through mRNA expression and resequencing analysis, implicates FOXP1, RYBP and SHQ1 as candidate cooperative tumor suppressors. Comparison of transcriptome and DNA copy number data from primary tumors for prognostic impact revealed that CNAs robustly define clusters of low- and high-risk disease beyond that achieved by Gleason score. In sum, this integrative genomic analysis of a substantial cohort of tumors clarifies the role of several known cancer pathways in prostate cancer, implicates several new ones, reveals a previously unappreciated role for CNAs in prognosis and provides a blueprint for clinical development of pathway inhibitors.

In this document, I describe how the (Affymetrix) expression data data entry for the Taylor Prostate Cancer 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)
> library(org.Hs.eg.db)
```

Now use the 'getGEO' function with the correct ID

```
>   url <- "ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE21nnn/GSE21034/matrix/"
>   destfile <-"GSE21034-GPL10264_series_matrix.txt.gz"
>   if(!file.exists(destfile)){
+   download.file(paste(url,destfile,sep=""),destfile=destfile)
```

```
+   }
> geoData <- getGEO(filename=destfile)
```

The feature data only has Entrez ID and Gene Bank Accession, whereas we want to use the more-common gene name in some analyses. We use the organism-level package to do the mapping.

```
> library(org.Hs.eg.db)
> entrezid <- unlist(mget(as.character(fData(geoData)[,2]),
+           revmap(org.Hs.egREFSEQ),ifnotfound = NA))
> anno <- data.frame(fData(geoData), Entrez = entrezid)
> anno <- anno[!is.na(anno$Entrez),]
> fData(geoData) <- data.frame(anno,
+         Gene=unlist(mget(as.character(anno$Entrez),
+                          org.Hs.egSYMBOL,ifnotfound=NA)))
```

We tidy up the data from GEO; creating a new data frame of just the clinical characteristics of interest.

```
> pd <- pData(geoData)
> pd2 <- data.frame("geo_accession" = pd$geo_accession,
+   Sample = gsub("sample id: ", "", pd$characteristics_ch1),
+   Sample_Group = gsub("disease status: ","", pd$characteristics_ch1.2),
+   Gleason=gsub("biopsy_gleason_grade: ", "", pd$characteristics_ch1.4),
+   Stage = gsub("clint_stage: ", "", pd$characteristics_ch1.5),
+   Path_Stage = gsub("pathological_stage: ","",pd$characteristics_ch1.6))
> rownames(pd2) <- pd2$geo_accession
```

Extra information, particulary that relating to survival, can be found in the supplmentary table for the paper. For convenience, we save this table as a csv file in the package.

```
> fpath <- system.file("extdata", "STable1.csv",package="prostateCancerTaylor")
> extraClin <- read.csv(fpath)
> pd2 <- merge(pd2, extraClin, by.x=2,by.y=1,all.x=TRUE,sort=FALSE)
> pd2$Time <- pd2$BCR_FreeTime
> pd2$Event <- pd2$BCR_Event == "BCR_Algorithm"
> pd2$Gleason <- paste(pd2$PathGG1,pd2$PathGG2,sep="+")
> pd2$Gleason[is.na(pd2$PathGG1) | is.na(pd2$PathGG2)] <- NA
> pData(geoData) <- pd2[,c("Sample","geo_accession",
+ "Sample_Group","Gleason","Stage","Path_Stage",
+ "Type","Copy.number.Cluster","Time","Event")]
> taylor <- geoData
> save(taylor, file="data/taylor.rda",compress = "xz")
```