Section 1: What we measure and why

Mammaprint Gene Signature

- Exploring genes used in the Mammaprint gene signature - assess risk of breast cancer
- Diagnostic signature using gene expression levels of 70 genes
- Information about the 70 gene signature used in the Mammaprint algorithm

```
library(genefu)
```

```
## Loading required package: survcomp

## Loading required package: survival

## Loading required package: prodlim

## Loading required package: mclust

## Package 'mclust' version 5.4.7
## Type 'citation("mclust")' for citing this R package in publications.

## Loading required package: limma

## Loading required package: biomaRt

## Loading required package: iC10

## Loading required package: pamr

## Loading required package: cluster

## Loading required package: impute

## Loading required package: iC10TrainingData

## Loading required package: AIMS

## Loading required package: e1071

## Loading required package: Biobase

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following object is masked from 'package:limma':
##
##     plotMA

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
data(sig.gene70)
dim(sig.gene70)
```

```
## [1] 70  9
```

```
head(sig.gene70)[,1:6]
```

```
##                          probe correlation average.good.prognosis.profile
## NM_003748           NM_003748   -0.420671                      0.12350000
## NM_003862           NM_003862   -0.410964                      0.05159091
## Contig32125_RC Contig32125_RC   -0.409054                      0.05409091
## U82987                 U82987   -0.407002                      0.06150000
## AB037863             AB037863   -0.402335                      0.06334091
## NM_020974           NM_020974   -0.399987                     -0.06231818
##                EntrezGene.ID NCBI.gene.symbol HUGO.gene.symbol
## NM_003748               8659          ALDH4A1          ALDH4A1
## NM_003862               8817            FGF18            FGF18
## Contig32125_RC            NA             <NA>             <NA>
## U82987                 27113             BBC3             BBC3
## AB037863                  NA             <NA>             <NA>
## NM_020974              57758           SCUBE2           SCUBE2
```

Assessment: Phenotypes

- COPDSexualDimorphism.data package - phenotypes (cols) individuals (rows)
- Data to assess incidence of COPD and emphysema by gender and smoking status
- The pkyrs variable in the expr.meta data.frame represents pack years smoked. Other variables include gender and diagmaj (disease status). These variables correspond to phenotypes.

```
library(COPDSexualDimorphism.data)
data(lgrc.expr.meta)
```

Assessment: Chromosomes and SNPs

- GWAS (Genome-wide association studies)
- Comparing individuals with disease vs. controls using SNP chips or DNA sequencing.
- SNPs with association are investigated for disruption of gene regulation or function
- Bioconductor *gwascat* package

```
library(gwascat)
```

```
## gwascat loaded.  Use makeCurrentGwascat() to extract current image.
```

```
##  from EBI.  The data folder of this package has some legacy extracts.
```

```
data(ebicat_2020_04_30)
ebicat_2020_04_30
```

```
## gwasloc instance with 50000 records and 38 attributes per record.
## Extracted:  2020-04-30 23:24:51
## metadata()$badpos includes records for which no unique locus was given.
## Genome:  GRCh38
## Excerpt:
## GRanges object with 5 ranges and 3 metadata columns:
##       seqnames    ranges strand |   DISEASE/TRAIT        SNPS   P-VALUE
##          <Rle> <IRanges>  <Rle> |     <character> <character> <numeric>
##   [1]       10  58153390      * | Crohn's disease   rs1819658     9e-17
##   [2]        1 206766559      * | Crohn's disease   rs3024505     2e-14
##   [3]       13  42478744      * | Crohn's disease   rs2062305     5e-10
##   [4]       19   1124836      * | Crohn's disease    rs740495     8e-12
##   [5]       12  40398498      * | Crohn's disease  rs11564258     6e-21
##   -------
##   seqinfo: 24 sequences from GRCh38 genome
```