Gene Signature Assessment

Stephen Blatti

November 20, 2017

### The 70-gene diagnostic signature for breast cancer progression risk

The genefu package has been a component of Bioconductor since 2011. Its DESCRIPTION file says it is an “R package providing various functions relevant for gene expression analysis with emphasis on breast cancer”. You can use this package to get acquainted with aspects of the mammaprint diagnostic test that Rafa mentioned in the lecture. Install the genefu package with biocLite.

# library(BiocInstaller)  
# biocLite("genefu")

A data.frame with information on the 70 gene signature used in the mammaprint algorithm is in the sig.gene70 data.frame. You can have a look at this:

library(genefu)

## Loading required package: survcomp

## Loading required package: survival

## Loading required package: prodlim

## Warning: package 'prodlim' was built under R version 3.4.2

## Loading required package: mclust

## Warning: package 'mclust' was built under R version 3.4.2

## Package 'mclust' version 5.3

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

## Loading required package: limma

## Warning: package 'limma' was built under R version 3.4.2

## Loading required package: biomaRt

## Loading required package: iC10

## Warning: package 'iC10' was built under R version 3.4.2

## Loading required package: pamr

## Warning: package 'pamr' was built under R version 3.4.2

## Loading required package: cluster

## Loading required package: iC10TrainingData

## Loading required package: AIMS

## Loading required package: e1071

## Warning: package 'e1071' was built under R version 3.4.2

## 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, cbind, colMeans,  
## colnames, colSums, do.call, duplicated, eval, evalq, Filter,  
## Find, get, grep, grepl, intersect, is.unsorted, lapply,  
## lengths, Map, mapply, match, mget, order, paste, pmax,  
## pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,  
## rowMeans, rownames, rowSums, sapply, setdiff, sort, table,  
## tapply, union, unique, unsplit, which, 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

You can see from this that there are 70 records in the data frame, and that there are diverse ways of describing the “genes” in the signature. How many components of the signature have a missing value for the associated NCBI gene symbol? (Remember to use is.na, never == NA.)

sum(is.na(sig.gene70$NCBI.gene.symbol))

## [1] 14

### Kinases in the 70 gene signature

Kinases are important for cell-cell communications; see the Wikipedia entry on Kinase for some background. You can use grep on the Description field of the sig.gene70 data.frame to search for substrings of long gene names. How many of the members of the 70-gene signature are genes coding for kinases?

head(sig.gene70)[,6:9]

## HUGO.gene.symbol Cytoband  
## NM\_003748 ALDH4A1 1p36  
## NM\_003862 FGF18 5q34  
## Contig32125\_RC <NA> <NA>  
## U82987 BBC3 19q13.3-q13.4  
## AB037863 <NA> <NA>  
## NM\_020974 SCUBE2 11p15.3  
## Alternative.symbols  
## NM\_003748 ALDH4|P5CD|P5CDh|P5CDhL|P5CDhS  
## NM\_003862 FGF-18|ZFGF5  
## Contig32125\_RC <NA>  
## U82987 JFY1|PUMA|PUMA/JFY1  
## AB037863 <NA>  
## NM\_020974 CEGP1|Cegb1|Cegf1|FLJ16792|FLJ35234|MGC133057  
## Description  
## NM\_003748 aldehyde dehydrogenase 4 family, member A1  
## NM\_003862 fibroblast growth factor 18  
## Contig32125\_RC <NA>  
## U82987 BCL2 binding component 3  
## AB037863 <NA>  
## NM\_020974 signal peptide, CUB domain, EGF-like 2

index <- grep("kinase",sig.gene70$Description)  
sig.gene70$Description[index]

## [1] "serine/threonine kinase 32B"   
## [2] "deoxycytidine kinase"   
## [3] "maternal embryonic leucine zipper kinase"   
## [4] "CDC42 binding protein kinase alpha (DMPK-like)"