The Sup Wald Identification of Transition in dna copy Changes (SWITCH) package
V1.0

Victor Weigman, Andrey Shabalin, Joel Parker

1. Overview

This document introduces an R implementation of breakpoint identification in chromosomal array Comparative Genomic Hybridization (aCGH) data. The `SWITCHdna` method involves 2 main parts: 1) transition point identification and 2) segment significance testing, along with a suite of data representation functions. The transition point identification finds locations in the genome where there is a significant change in intensity within a stream of neighboring intensity levels. The first transition point is called at the location of the global maximum F statistic and then the algorithm recursively operates on either side of the transition point to find more points (passing a chosen F-value threshold). Pure intensity-based programs like this tend to yield a large number of false positive segments with no noise estimation or significance testing which includes segment size. The segment significance is based on the segment's average intensity and its size (summarized via Z-score). We used a Z score of 3 and a minimum intensity measurement of 0.09 (chosen to eliminate called "segments" of 2N chromosomes).

2. Data

2.1 Normalized intensity data.

The data used in this example is from 5 primary mouse mammary carcinomas arrayed on Agilent Whole Genome Array 244A kit. Each Agilent feature extraction file was lowess normalized according to UMD standards before Log2-ratios were formed. It is assumed that the user submits data where an experimental sample is compared to a reference pool of DNA. Unless explicitly stated, all files must be tab-delimited text. You must prepend, in the first three columns the probe, chromosome and position information, along with all intensity data. Please see Sample_aCGHdata.txt

2.2 Sample Class Descriptors

If you are planning on generating figures, a separate file is needed describing to class memberships. The txt file must be 2 columns, with headers. The first column will be your sample names. Assuming you used standard naming, with no characters in the number row, these names will match your original names in Sample_aCGHdata.txt. But if you had sample names that begin with numbers or contain non-standard characters, you will have to build this file with the sample names that R creates from your own sample names. To do this, you will take the sample column from the _SWITCH_out.txt file, remove duplicates, and take names from that file. Although Sample_cls.txt is made for Sample_aCGHdata.txt and has 1 class, you can give as many class names as you like

2.3 Annotation files

To use some functions, you will need to create annotation files. These can be for any genomic element that you are interested in: genes, cytobands, etc. The txt file, with header, must contain 4 columns: Chr, Start, Stop, Name. Typically, all chromosome

notations must be in numeric (23 for X, 24 for Y, etc) and match the notation in your original input intensity file. See ACGHannot.txt

3. Examples

3.1 Creating segments:

CNAmake recursively computes an F-test and defines segments as those probes whose f-statistic passes the threshold. In addition to defining a level of significance, the user must also decide the level of precision to which segments must be resolved. The Agilent Mouse 244A platform has, on average, one probe per 7 kb, so 35 probes are required to make segments of 250 kb (minimum segment threshold). Statistical stringency is can be adjusted through Fthresh. It is better to keep this low and let Zscore and Intensity thresholds further filter segments later

```
>source("SWITCH.r")
>CNA.ex <- CNAmake("log2itensityfile.txt",Fthresh=12,alpha=16)
```

SWITCH(data = data, Fthresh = 11, alpha = 10, a.type= 0, verbose=1)

data = file that contains CGH intensity, with first 3 columns being (Probe, Chr, Position), in genomic order

Fthresh = value for F-statistics threshold used to flag breakpoints

alpha = the %age of the chromosome a given gain/loss must be larger than to work OR 0 is min probes)

verbose = 1 will alert for each new sample processed. 0 - squelches all output to prompt

-returns-

data.frame of segments
file of segments to use for plotting functions

The output of this function is the driver of all other `SWITCH` *internal functions.*

Depending on file size, analysis time will vary, verbose =1 will report which sample is being process and the length of time for the entire procedure.

3.2 Graphics

`SWITCHdna` uses a function *plot.freq.SW* to show CNA landscape plots based on the number of samples of a class that share that CNA. The function requires a threshold for Z-score and intensity, which must be determined prior to running these functions. These values can differ slightly from dataset to dataset. Default values are given in the function call but simulations should be calculated to identify Z and intensity values from spurious

intensity data where `SWITCHdna` was run. Plot.freq.SW generates landscape plots of the frequencies of CNAs within the groups specified by the specified groups file. It also requires a Z-score and intensity cutoff. You would also need to have a separate file which contains 2 columns, one of chromosome numbers and the second of total chromosome length

```
>plot.freq.SW(data=data, grps="Sample_cls.txt", Z.score=3,
I.thresh=0.1,species=" hsChrLengths.txt", returnSummary = F)
```

creates a frequency plot for each class of interest

data = object from SWITCH
grps = file name that contains each sample and the class its associated with
Z.score = Zscore cutoff
species = a file of 2 columns, no headers, 1st column contains chromosome names (numeric), 2nd column is the length of each chromosome
returnSummary = if you want the function to return the summary table

3.3 Annotation Functions

The two annotation functions are supplied requiring the output of two prior functions. Both of which require a reference file of the genomic elements that you are interested in.

```
seg.An(sumfile=sumfile,ref=ref,pct=0.5)
```

annotates _summary.txt files for identifying genes within the region at a pre-determined percentage

sumfile = output file from plot.freq.SW or its object
ref = Reference file for containing genomic locations of region of interest, same format as with CNAgene.mat, 4 columns: Chromosome, Start, Stop, Symbol
pct = %age cutoff for annotating a segment

```
CNAgene.mat(data="filename.txt",ref="annotfile.txt",Z.score=3,I.t
hresh=0.1,return.type=1,out="filename")
```

takes direct output from CNAmat to make a M x N of gain/lost status for genes and samples

data = filename from output SWITCH
ref = reference genome file, formatted as RowID <tab> Chr <tab> Start Pos <tab> End Pos <tab> Gene Symbol
Z.score = Score to define segments that are significant
I.thresh = Score to define background level of intensity
return.type: 1 = discrete, 2 = intensity, 3 = z.score