Clonality: A Package for Clonality testing

Irina Ostrovnaya

March 30, 2012

Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center
ostrovni@mskcc.org

Contents

1 Overview

This document presents an overview of the `Clonality` package. This package can be used to test whether two tumors are clonal (metastases) or independent (double primaries) using their copy number or loss of heterozygosity (LOH) profiles. For LOH data it implements Concordant Mutations (CM) test (Begg et al., 2007) and Likelihood Ratio (LR) test (Ostrovnaya et al., 2008). For copy number profiles the package implements the methodology based on the likelihood ratio described in (Ostrovnaya et al., 2010).

2 Copy number profiles

We will show how to test independence of the copy number profiles from the same patient using breast cancer data. The BAC arrays of the pairs of lobular carcinoma in situ (LCIS) and invasive lobular carcinoma (ILC) were studied in (Hwang et al., 2004) and available at `http://waldman.ucsf.edu/Breast/Hwang.data.xls`. We will load package `gdata` in order to read the excel file.

```
> library(DNAcopy)
> library(Clonality)
> library(gdata)
```

We will read the dataset and remove rows or columns with too many NAs.

```
> data<-read.xls("http://waldman.ucsf.edu/Breast/Hwang.data.xls")
> data<-data[!is.na(data[,2]),]
> data<-data[apply(is.na(data),1,sum)<=50,]
> data<-data[,apply(is.na(data),2,sum)<=1000]

> data[1:5,1:10]

       Clone Chromosome Position LC02LCIS LC03LCIS LC04LCIS LC06LCIS LC07LCIS
2 RP11-82D16          1        2    0.139    0.121    0.184    0.164   -0.151
3 RP11-62M23          1        3    0.097   -0.009    0.160   -0.024   -0.136
4 RP11-111O5          1        4    0.062   -0.060    0.189    0.050   -0.086
5  RP11-51B4          1        5   -0.059   -0.165    0.100   -0.036   -0.291
6 RP11-60J11          1        6    0.146    0.054    0.174    0.073   -0.071
  LC12LCIS LC17LCIS
2   -0.070    0.180
3   -0.104    0.032
4    0.012   -0.013
5   -0.277   -0.134
6    0.027    0.068
```

Rows of data correspond to probes (genomic markers). The first column is probe name; the second column is the chromosome where the probe is located; the third column is probe's genomic position recorded as an index. All subsequent columns correspond to the samples and contain log-ratios.

Since there are no genomic locations in this dataset, we will download another dataset and map the genomic locations to the probes. If the genomic locations were known, we would not need this step and the column with the probe names.

```
> arrayinfo<-read.xls("http://waldman.ucsf.edu/Colon/nakao.data.xls")  #needed to extract
> data$Position<-arrayinfo$Mb[match(toupper(as.character(data[,1])),toupper(as.character(ar
> data<-data[!is.na(data$Position),]
```

Now we will remove repeated genomic locations:

```
> length(unique(paste(data$Chromosome, data$Position))) #there are repeated genomic locatio

[1] 1740

> data<-data[c(TRUE,data$Position[-1]!=data$Position[-1864]),] #discard probes with repeate
> data<-data[data$Chromosome<=22,] #getting rid of X and Y chromosomes

> dim(data)

[1] 1696   51
```

As the final step of data preparation, we have to create a CNA (copy number array) object as described DNAcopy. To save computational time, we only take the first three patients. (As a result, gain/loss frequencies used for analysis will be very imprecise and the reference distribution will have very few comparisons.)

```
> dataCNA<-CNA(as.matrix(data[,c(4:6,28:30)]),maploc=data$Position,chrom=data$Chromosome,sa
```

Our methodology allows at most one genomic change per chromosome arm, estimated by the one-step Circular Binary Segmentation (CBS) algorithm ((Venkatraman and Olshen, 2007)).

If the data had many more than 15,000 markers, most outstanding, and likely a short change would be picked up, which would not be representative of the chromosome pattern. To avoid this, one can use the following function:

```
> dataAve<- ave.adj.probes(dataCNA,2)

Total number of markers after averaging is 842
```

Here we have averaged every two consecutive marker. For this dataset, though, averaging is not necessary.

The chromosomes should be split into arms before the clonality analysis since it increases the number of independent genomic units.

```
> dataCNA$maploc<-dataCNA$maploc*1000 #transforming maploc to Kb scale
> dataCNA$chrom<- splitChromosomes(dataCNA$chrom,dataCNA$maploc)  #splits the chromosomes i

chrom
chr01p chr01q chr02p chr02q chr03p chr03q chr04p chr04q chr05p chr05q chr06p
    49     60     18     34     38     38     25    113     17     71     33
chr06q chr07p chr07q chr08p chr08q chr09p chr09q chr10p chr10q chr11p chr11q
    35     59     92     44     72     34     63     29     78     54     81
chr12p chr12q chr13q chr14q chr15q chr16p chr16q chr17p chr17q chr18p chr18q
    16     60     38     65     60     19     42     16     50     15     30
chr19p chr19q chr20p chr20q chr21q chr22q
    12     21     31     46     25     13
```

Next we have to create a vector of patient labels that matches the samples.

```
> ptlist<-substr(names(dataCNA)[-c(1,2)],1,4)
```

Finally, we can run the clonality analysis:

```
> results<-clonality.analysis(dataCNA, ptlist,   nmad = 1.25,  reference = TRUE, allpairs =

Calculating LR...
Calculating reference LR: %completed 17, 33, 50, 67, 83, 100,
```

The main information is in the output LR:

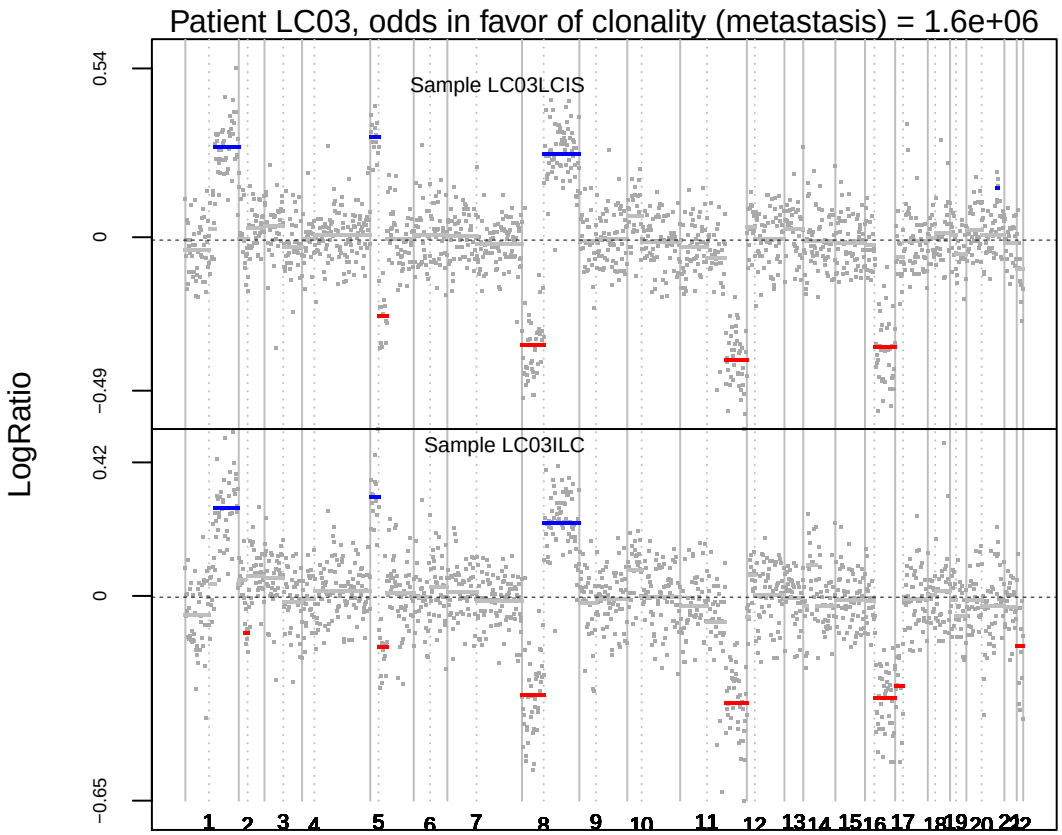
```
> results$LR
```

```
   Sample1 Sample2          LR1          LR2 GGorLL NN GL GNorLN
1 LC02LCIS LC02ILC  0.003278829 1.622146e-02      3 17  0     19
2 LC03LCIS LC03ILC 75.121794800 1.574651e+06      7 28  0      4
3 LC04LCIS LC04ILC  6.243825075 1.526120e+07      6 26  0      7
                                    IndividualComparisons LR2pvalue
1                                               chr01q 4.95 0.8333333
2                 chr01q 25.32; chr05q 17.77; chr11q 46.59 0.0000000
3 chr01p 42.48; chr01q 16.54; chr10q 22.81; chr11q 152.43 0.0000000
```

The likelihood ratios LR2 for sample LC02 is much smaller than 1, therefore these tumors are independent. Patients LC03 and LC04 have LR2 much higher than one, and we can conclude that their tumors are clonal. The reference distribution for LR2 under the hypothesis of independence is constructed by pairing tumors from different patients that are independent by default. The p-value column reflects the percentiles of a particular patient's LR2 in the reference distribution: clonal tumors would have small p-values.

We can view the genomewide plots of patient LC03 using:

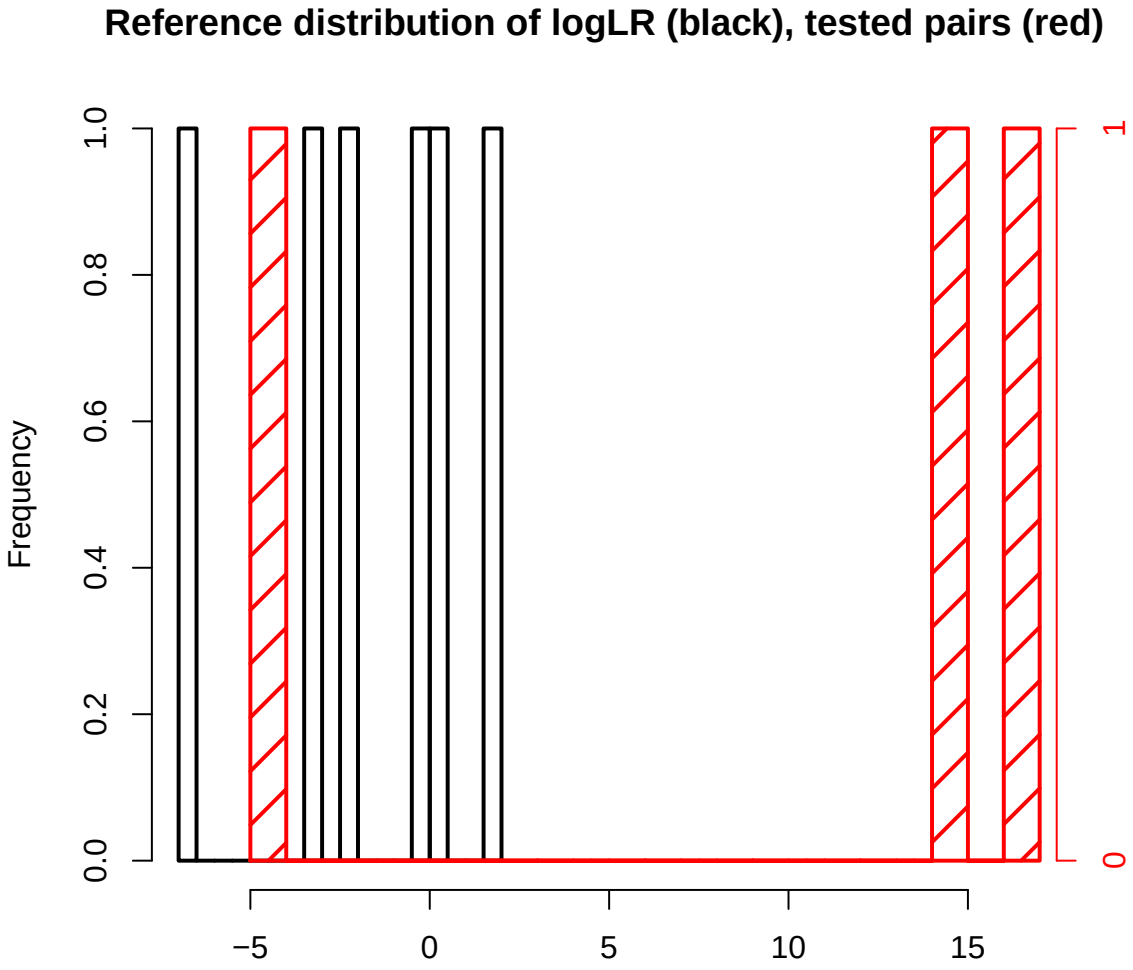
```
> genomewidePlots(results$OneStepSeg, results$ChromClass,ptlist , c("LC03LCIS", "LC03ILC"),i
```



Patterns for each chromosome would be plotted by:

The overlap between the histograms of LR2 from original pairs of tumors and the reference distribution are produced by:

```
> histogramPlot(results$LR[,4], results$refLR[,4])
```



2.1 Choice of segmentation algorithm

Note that the user can potentially specify the segmentation method to be used. Currently the default behavior of the clonality.analysis function is to use the CBS algorithm to identify the most significant change in each chromosome arm. The internal function for this purpose is "oneseg" called as oneseg(x, alpha, nperm, sbdry)

There are 4 arguments to oneseg:

x: is the finite logratio data ordered by genomic position.
alpha: the significance level used by CBS.
nperm: the number of permutations for the reference distribution.
sbdry: early stopping boundary for declaring no change (calculated from alpha and nperm).

The output of this function is a vector of 3 numbers where the first is the number of change-points detected (must be 0, 1 or 2), and the second and the third numbers are the start and end of the left segment if there is only one change-point, and of the middle segment when there are 2 change-points.

The function allows the user to specify alternative alpha and nperm for 'oneseg' as a list using the segpar argument e.g. segpar=list(alpha=0.05, nperm=1000). Since sbdry is always calculated in clonality.analysis function from alpha and nperm it is not specified.

Alternate segmentation algorithm can be used. It requires the user to create a function that takes the ordered logratio from one chromosome arm as argument "x" as in oneseg. The name of this function should not be 'oneseg' and is passed through the 'segmethod' argument and all other necessary arguments that are needed passed as a list through 'segpar' argument.

3 LOH data

The LOH data has to be combined in a matrix where first column has marker names and the following columns have LOH calls for each sample. Here we simulate a dataset with 10 pairs of tumors and 20 markers. First pair of tumor is clonal, and the rest of them are independent. If the marker is heterozygous and there is no LOH, then it is denoted by 0. LOH at maternal or paternal alleles is marked by 1 or 2.

```
> set.seed(25)
> LOHtable<-cbind(1:20,matrix(sample(c(0,1,2),20*20,replace=TRUE),20))
> LOHtable[,3]<-LOHtable[,2]
> LOHtable[1,3]<-0

> LOHtable[,1:5]

      [,1] [,2] [,3] [,4] [,5]
 [1,]    1    1    0    2    0
 [2,]    2    2    2    0    0
 [3,]    3    0    0    1    0
 [4,]    4    2    2    2    2
 [5,]    5    0    0    1    1
 [6,]    6    2    2    0    2
 [7,]    7    1    1    2    1
 [8,]    8    1    1    2    2
 [9,]    9    0    0    0    1
[10,]   10    0    0    2    0
[11,]   11    0    0    0    2
[12,]   12    1    1    2    0
[13,]   13    2    2    2    0
[14,]   14    1    1    1    2
[15,]   15    2    2    1    0
```

```
[16,]   16    0    0    0    1
[17,]   17    1    1    0    0
[18,]   18    2    2    1    2
[19,]   19    1    1    0    0
[20,]   20    2    2    0    0
```

```
> LOHclonality(LOHtable,rep(1:10,each=2),pfreq=NULL,noloh=0,loh1=1,loh2=2)
```

```
Testing clonality for patient 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,  Done
   Sample1 Sample2  a  e f g h Ntot             CMpvalue LRpvalue
1        1       1 13 13 0 1 6   20 2.20457220717234e-08        0
2        2       2  3  6 4 6 4   20    0.633257174327221        1
3        3       3  6  9 2 5 4   20     0.13247418005031    0.458
4        4       4  6  9 4 5 2   20    0.271731009940983    0.807
5        5       5  3  6 7 5 2   20    0.768026723950271        1
6        6       6  1  5 8 3 4   20    0.964059678147575    0.442
7        7       7  6 12 3 4 1   20    0.607636663320756        1
8        8       8  5 11 4 2 3   20    0.585520481546597    0.719
9        9       9  4  7 6 5 2   20    0.597049677704141    0.911
10      10      10  6 10 3 6 1   20    0.424944369195046    0.794
```

First p-value is small, indicating clonality, for both CM and LR tests. The rest of the p-values are not significant.

Markers that are not informative (e.g. homozygous) in a particular tumor should be given NA instead of a call. Such markers will be dropped from the analysis of this specific patient.

Below are the details of the session information:

```
R version 2.15.0 (2012-03-30)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] gdata_2.8.2     Clonality_1.4.0 DNAcopy_1.30.0

loaded via a namespace (and not attached):
[1] gtools_2.6.2 tools_2.15.0
```

References

Begg, C., Eng, K., and Hummer, A. (2007). Statistical tests for clonality. *Biometrics*, 63:522–530.

Hwang, E., Nyante, S., Chen, Y., Moore, D., DeVries, S., Korkola, J., Esserman, L., and Waldman, F. (2004). Clonality of lobular carcinoma in situ and synchronous invasive lobular cancer. *Cancer*, 100(12):2562–72.

Ostrovnaya, I., Olshen, A., Seshan, V., Orlow, I., Albertson, D., and Begg, C. (2010). A metastasis or a second independent cancer? evaluating the clonal origin of tumors using array copy number data. *Statistics in Medicine*, 29:1608–1621.

Ostrovnaya, I., Seshan, V., and Begg, C. (2008). Comparison of properties of tests for assessing tumor clonality. *Biometrics*, 68:1018–1022.

Venkatraman, E. and Olshen, A. (2007). A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23:657–663.