intCNGEan: Nonparametric testing of copy number induced differential gene expression

Wessel N. van Wieringen

Department of Epidemiology and Biostatistics, VU University Medical Center
P.O. Box 7075, 1007 MB Amsterdam, The Netherlands
&
Department of Mathematics, VU University Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
wvanwie@few.vu.nl

Contents

1 Introduction

This vignette shows the use of the `intCNGEan` package for the integrative analysis of DNA copy number and gene expression data. The following to features are discussed in detail:

- application of the methodology described in Van Wieringen and Van de Wiel (2009), and
- joint plotting of data from the two molecular levels.

These are illustrated on an example data set, which is introduced first.

2 The Pollack breast cancer data

The breast cancer data set of Pollack *et al.* (2002) is available at `http://www.pnas.org`. Pollack *et al.* (2002) used cDNA microarrays to measure copy number and gene expression of 41 primary breast tumors. As both profile types are generated on the same platform, features are automatically

matched. The pre-processing of the copy number data consists of removal of clones with more than 30% missing values, imputation of remaining missing values using the k-nearest neighbor method (Troyanskaya *et al*, 2001), mode normalization, segmentation using the CBS method of Olshen *et al.* (2004), and calling using CGHcall of Van de Wiel *et al.* (2007). The gene expression data are within-array median normalized. After pre-processing (and removal of genes on the sex chromosomes) the copy number and expression profiles consists of 5816 features.

```
> # load the full Pollack breast cancer data
> library(intCNGEan)
> data(pollackCN)
> data(pollackGE)
```

The code above loads a `cghCall` and `expressionSet` object containing annotated DNA copy number and gene expression data, respectively.

3 Matching

The first step of an integrative analysis comprises the matching of the features of both platforms. The objective of this matching step is to assign the appropriate DNA copy number to each feature on the gene expression array. Note this is not the same as to reproduce the matching produced by, say, Ensemble. In order to do the matching one needs to specify whether base end pair information of the features is available (start base pair information is assumed to be available, otherwise matching by genomic location is pointless). The `cghCall` and `expressionSet` objects for the Pollack data set contain no end base pair information, which one would specify via the `CNpbend` end `GEbpend` parameters. In addition, one needs to specify how features are matched. Currently, only three methods are implemented. Here we choose for the most conservative option `method = "overlap"`, which matches each gene from the expression array to the feature from the copy number platform with the maximum percentage of overlap. If the maximum percentage of overlap equals zero, the gene is excluded from further analyses for it could not be matched.

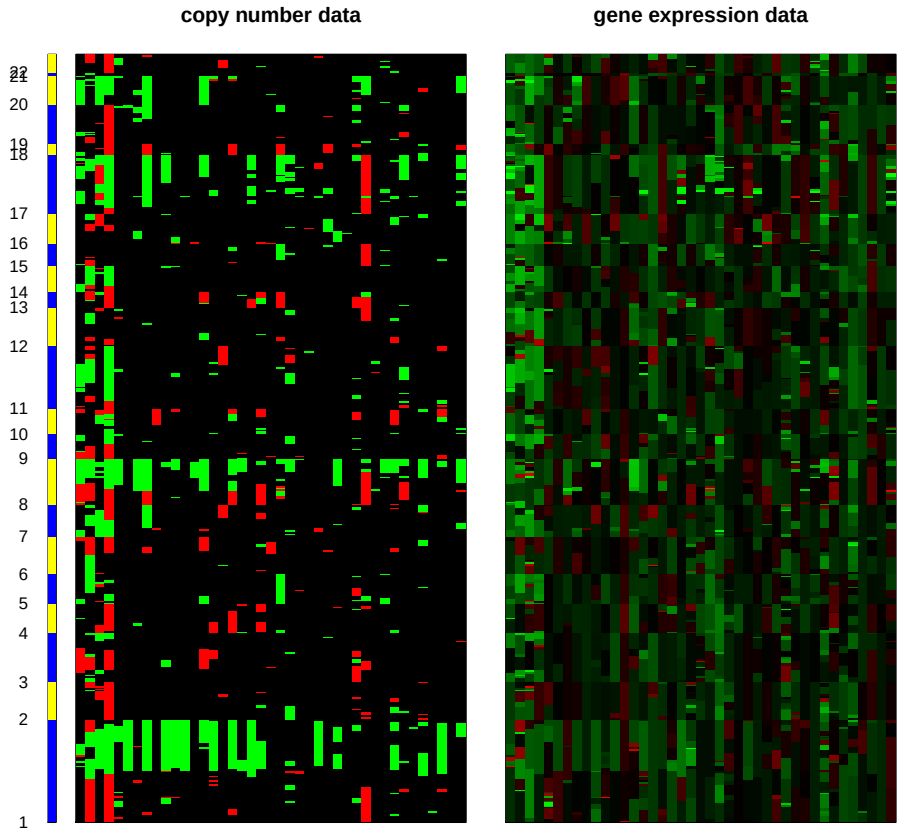
```
> # match data from the two platforms
> CNGEdataMatched <- intCNGEan.match(pollackCN, pollackGE, CNbpend = "no",
    GEbpend = "no", method = "overlap")
> pollackCN <- CNGEdataMatched$CNdata.matched
> pollackGE <- CNGEdataMatched$GEdata.matched
```

As both profile types in the Pollack data set are generated on the same platform, features are automatically matched. The matching step can thus be skipped.

4 Joint plotting

To get a overall impression of the relation between DNA copy number and gene expression data, plot the heatmaps of both molecular levels simultaneously:

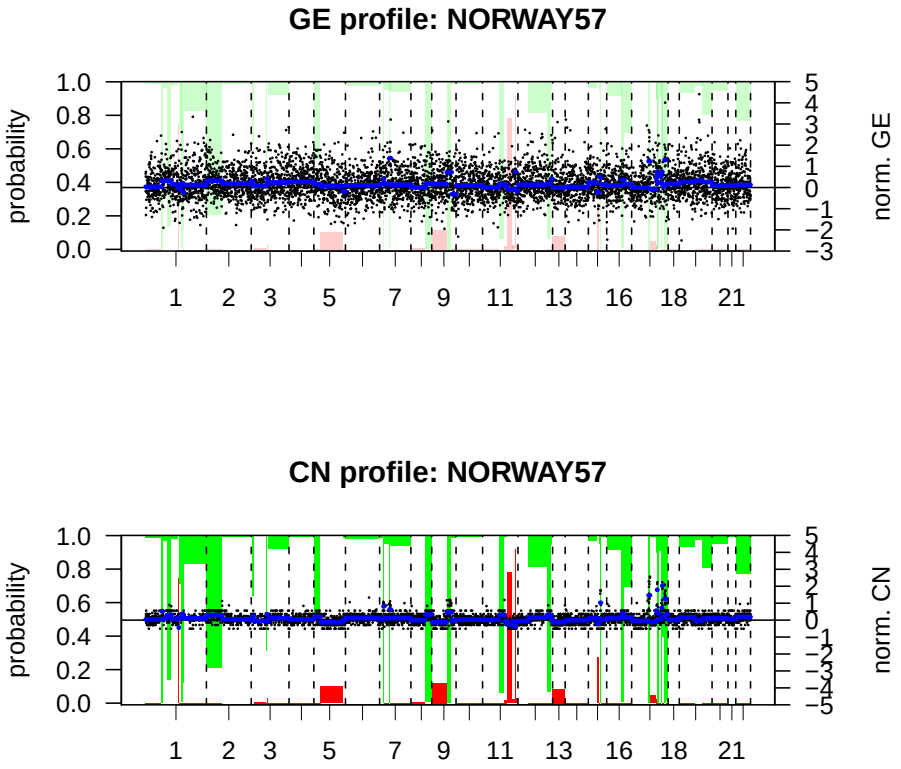
```
> intCNGEan.heatmaps(pollackCN, pollackGE, location = "mode", colorbreaks = "equiquantiles")
```



Common features in DNA copy number and gene expression data become more emphasized if, prior to simultaneous heatmap plotting, the samples both are ordered in accordance with, say, a hierarchical clustering of either of the two data sets. At all time the order of the samples should be the same for both DNA copy number and gene expression data.

Alternatively, one may be interested in the relation between DNA copy number and gene expression levels within an individual sample. This is visualized by plotting the profiles of two samples on top each other. This plotting may be limited to a particular chromosome of interest via the `chr` parameter.

```
> intCNGEan.profilesPlot(pollackCN, pollackGE, 23)
```



The color coding in the background indicate the aberration call probabilities as produced by `CGHcall` (Van de Wiel *et al.*, 2007).

5 Pre-test and tuning

The method of Van Wieringen and Van de Wiel (2009) exploits the census of cancer genes (Futreal *et al.*, 2004), which distinguishes between proto-onco and tumor-suppressor genes associated with gain and loss, respectively. This gain (or loss) of a particular genomic segment is, through the central dogma of biology, likely to result in increased (or decreased) transcription levels of the genes on the segment. Motived by Figure 1b of Beroukhim *et al.* (2010), it is assumed that, within cancer of a particular tissue, a gene cannot be a proto-onco gene as well as tumor-suppressor gene for that tissue.

Unfortunately it is unknown for every gene whether it is a proto-onco or tumor-suppressor gene. Consequently, one does not know whether to compare the gene expression between samples with a normal and gain, or between those with a loss and normal. This is decided by the array CGH data: e.g., if, for a particular gene, the call probability mass (as measured over the samples) of a gain exceeds that of a loss, the loss and normal call probability masses will be merged and the 'no-gain vs. gain' comparison is carried out for this gene.

Also prior to testing, it is recommendable to discard genes beforehand. This benefits the overall (FDR) power of the testing procedure. Exclusion of genes is done:

1. On the basis of the sum of a gene's marginal call probabilities of loss and gain. If it is smaller than `minCallProbMass`, the gene is discarded from further analysis. Effectively, this ensures identifiability of the copy number effect on expression levels.

2. On the basis of a metric, calculated from the DNA copy number data only, which aims to identify two situations for which the test is likely to have low power.

   - The first situation occurs when there is an unbalance between the expected call probabilities, as assessed *over* all samples. For instance, genes with

     $$\sum_{i=1}^{n} P(\text{sample } i \text{ has a loss at the location of gene } j) = 0.001$$

     and

     $$\sum_{i=1}^{n} P(\text{sample } i \text{ has no aberration at the location of gene } j) = 0.999$$

     have an unbalanced call probability distribution. A priori one expects that the proposed tests may not be powerful to detect a shift for such genes.
   - The second situation occurs when many samples individually (i.e. *within* sample) have a uniformly distributed call probability mass: $P(\text{sample } i \text{ has loss at the location of gene } j) = \frac{1}{2} = P(\text{sample} i \text{has an aberration at the location of gene} j) = 0.999$. This indecision on the call is propagated into the test, resulting in low power.

   The cut-off for this metric is chosen in such a way that the expected number of true discoveries is maximized.

The following command line performs the pre-testing and tuning:

```
> pollackTuned <- intCNGEan.tune(pollackCN, pollackGE, "wmw", ngenetune = 1000,
    nperm_tuning = 1000, minCallProbMass = 0.05)
```

To obtain the number of excluded genes:

```
> print(nrow(pollackGE) - nrow(pollackTuned$datafortest))
```

The `pollackTuned` object, a list, contains a vector of the genes that are passed on for testing.

The tuning excluded 664 of the 5816 genes, roughly 11%, from testing. The number of excluded genes depends among others on the DNA copy number profiles. If these are 'wild', exhibiting many aberrations all over the genome, we expect most genes to have a reasonably balanced expected (over the samples) call probability distribution. If, however, there are only few genomic regions aberrated, the contrary is expected, and more genes are expected to be excluded. The number of excluded genes also depends upon the number of genes whose expression is affected by copy number changes. This, in combination with an FDR rule, increases the probability of detecting shifts for genes with unbalanced or imprecise call probability mass.

6 Testing

We are now ready to test for DNA copy number induced differential gene expression on the set of selected genes:

```
> pollackResults <- intCNGEan.test(pollackTuned, "univariate", "wcvm",
    nperm = 10000, eff.p.val.thres = 0.10)
```

The number of significant genes, obtained through:

```
> fdrCutoff <- 0.10
> print(sum(pollackResults$adj.p < fdrCutoff))
```

equals 869 at a FDR significance level of 0.05 and 1268 at 0.10. This large number of significant genes is in line with 'major direct role' of DNA copy number alterations in the transcriptional program as claimed by Pollack *et al.* (2002), but forces us to look not only at statistical significance, but also at biological relevance. Gene prioritization (ranking) could be done by using the effect size and/or the coefficient of determination.

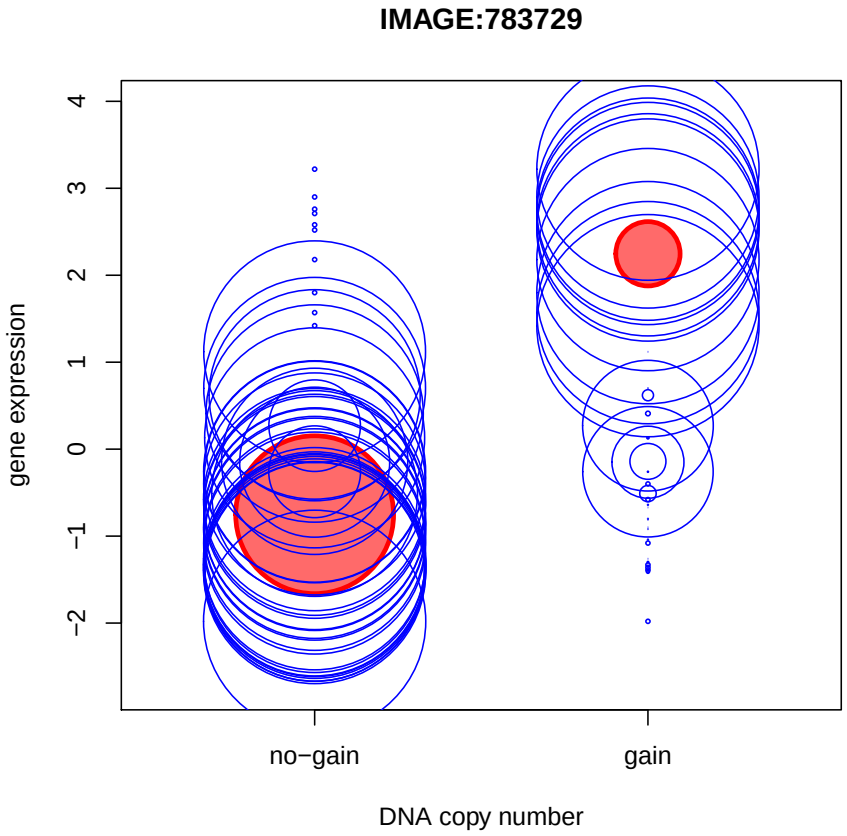
A further global view of the effect of DNA copy number on gene expression levels is provided by a histogram of the effect sizes for all selected genes, which may be obtained through:

```
> hist(pollackResults$effect.size, n=50, col="blue", border="lightblue",
    xlab="effect size", main="histogram of effect size")
```

For the Pollack data the histogram shows an effect size distribution that is clearly shifted away from zero, indicating that many genes have affected expression levels, in turn confirming the aformentioned 'major direct role'.

Among the significant genes is the ErbB2 gene. ErbB2 is a well-known cancer gene in breast cancer. It has an increased copy number in approximately 25% percent of the breast carcinoma's (Kallioniemi *et al.*, 1992), and associated over-expression. This is confirmed in the figure below where the relation between the copy number and expression data for ErbB2 is depicted. The estimated effect size of the DNA copy number on the gene expression equals 2.74 for ErbB2, and, with $R^2 = 0.69$, explains 69% of the variation in its gene expression. The multiplicity corrected p-value of the proposed tests for ErbB2 smaller than 0.0001.

```
pre-test...
tuning started...
50 of 100 resamples done, and counting...
100 of 100 resamples done, and counting...
ready: tuning done
```



7 Regional analysis

As neighboring genes share the same array CGH signature, one expects that their expression levels are affected in a similar fashion. This opens the possibility of borrow information across the genes within each region (defined as a series of adjacent clones with the same DNA copy number signature), but test for copy number induced differential expression per gene. This is done by shrinking the the test statistics within the region. In order to perform such a 'regional analysis' change the `analysis.type` parameter:

```
> pollackResults <- intCNGEan.test(pollackTuned, "regional", "wcvm",
    nperm = 10000, eff.p.val.thres = 0.10)
```

Compare this to the code of the univariate analysis.

If one wishes to assess the effect of DNA copy number on the expression levels of all genes in the region simultaneously (making an inference at the level of the region rather than that of individual genes), one may consult the `RCMgenomics`-package with implements the method of **?**. Their random coefficients model facilitates a global analysis of DNA copy number aberration associated regional co-expression at the level of the region (rather than its genes). It allows to assess a) whether there is a shared CNA effect on the expression levels of the genes within the region, and b) whether the CNA effect is identical for all genes.

References

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., McHenry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y. J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., Weir, B. A., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. A., Tabernero, J., Baselga, J., Tsao, M. S., Demichelis, F., Rubin, M. A., Janne, P. A., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. A., Loda, M., Beer, D. G., True, L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R., Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**(7283), 1899–905.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer*, **4**, 177–183.

Kallioniemi, O. P., Kallioniemi, A., Kurisu, W., Thor, A., Chen, L. C., Smith, H. S., Waldman, F. M., Pinkel, D., and Gray, J. W. (1992). ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization. *PNAS*, **89**, 5321–5325.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Borresen-Dale, A. L., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *PNAS*, **99**, 12963–12968.

Troyanskaya, H., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Van de Wiel, M. A., Kim, K. I., Vosse, S. J., Van Wieringen, W. N., Wilting, S. M. and Ylstra, B. (2007). CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, **23**, 892–894.

Van Wieringen, W. N. and Van de Wiel, M. A. (2009). Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, **65**(1), 19–29.

Van Wieringen, W. N., Berkhof, J., and Van de Wiel, M. A. (2010). A random coefficients model for regional co-expression associated with DNA copy number aberrations. *Statistical Applications in Genetics and Molecular Biology*, accepted for publication.

,