Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Functional profiling and gene expression analysis of chromosomal copy number alterations

Lucía Conde[1], David Montaner[1, 2], Jordi Burguet-Castell[1], Joaquín Tárraga[1, 2], Fátima Al-Shahrour[1], and Joaquín Dopazo[1, 2*]

[1]Department of Bioinformatics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, E-46013, Spain;
[2]Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe (CIPF), Valencia, E-46013, Spain;
Joaquín Dopazo* - Email: jdopazo@cipf.es; * Corresponding author
received January 13, 2007; accepted February 11, 2007; published online April 10, 2007

Abstract:
Contrarily to the traditional view in which only one or a few key genes were supposed to be the causative factors of diseases, we discuss the importance of considering groups of functionally related genes in the study of pathologies characterised by chromosomal copy number alterations. Recent observations have reported the existence of regions in higher eukaryotic chromosomes (including humans) containing genes of related function that show a high degree of co-regulation. Copy number alterations will consequently affect to clusters of functionally related genes, which will be the final causative agents of the diseased phenotype, in many cases. Therefore, we propose that the functional profiling of the regions affected by copy number alterations must be an important aspect to take into account in the understanding of this type of pathologies. To illustrate this, we present an integrated study of DNA copy number variations, gene expression along with the functional profiling of chromosomal regions in a case of multiple myeloma.

Keywords: profile; function; gene expression; chromosomal copy number

Background:
Genomic copy number alterations such as gains or losses of chromosomal regions have been shown to be on the basis of many human pathologies. Classical approaches to characterize these genetic aberrations used comparative genomic hybridisation (CGH), in which genomic DNA was hybridised to metaphase chromosomes. **[1]** Recently, the use of different types of microarrays to directly study genomic variations in DNA copy number is becoming more and more popular. Such massive genomic approaches are known as array comparative genomic hybridisation, or Array CGH. **[2]** These new technologies along with the use of expression arrays allow for a highly accurate characterisation of the dependence of gene expression on alterations in genomic copy number. **[3]**

As in many genome-scale methodologies data analysis and, in particular, the biological interpretation of the results constitutes a well-known bottleneck. Specific problems related to the analysis of Array CGH can be circumscribed mainly to two types: appropriate mapping and visualisation of the data onto the chromosomes, and efficient copy number estimation. This last aspect has been the motivation for a number of analytical approaches recently proposed **[4]**, that can be considered the first generation of algorithms for Array CGH analysis. Obviously, copy number variations are expected to have a strong effect on gene expression. **[5, 6]** Nevertheless, the ultimate aim of studies of copy number chromosomal alterations is to understand what is the effect produced in functional terms. In the classical vision one or a few key genes are the causative factors for the this type of pathologies, and the problem consisted in identifying such genes within the region amplified or deleted. The existence of regions in the chromosomes containing coexpressing genes **[7]** which, in addition, are functionally related has recently been reported even in higher eukaryotes. **[8]** Actually, regional arrangements of genes have found to be regulated not only by copy number alterations but also by different mechanisms such as epigenetic modifications. **[9]** This reinforces the functional role of chromosomal regions including groups of functionally related genes and its possible impact on diseases such as cancer. **[10]** These observations give credence to a new vision in which chromosomal alterations can be causing effects not by altering single key genes but by acting on complete molecular sub-systems such as pathways of functionally related genes. Recently, different approaches have focused on the functional aspects of the results of microarray experiments. **[11, 12]** Nevertheless, the possible functional significance at regional level of copy number alterations has been largely ignored. Here we present a combined approach to the study of copy-number alterations, gene expression and functional profiling, exemplified in a case of multiple myeloma. **[13]**

Methodology:
Functional profiling of Array-CGH experiments under this new perspective would require of three steps: 1) detection of regions with copy number variations (the origin of the disease), 2) detection of regional alterations in gene expression (the causes of the disease) and 3) analysis of enrichment in functional terms in the detected regions (the consequences of the alteration or the functional basis of the disease). While copy number alterations can be detected by means of different methods, alterations in the levels of gene expression are not always easy to be detected using the typical methods (t-test or similar) due several factors such as small sample sizes. For this reason here we will only use plots to visualize the effect of one variable (copy number) into the other one (expression level). The third step, the

functional profiling, becomes then the most important aspect of the analysis given that it will provide a functional explanation of the molecular basis of the disease caused by copy number alterations.

Detection of copy number alterations

We have used a segmentation method which is a variant of the circular binary segmentation method **[14]**, for copy number change detection (isowindow).

The isowindow method tries to identify boundaries between regions with a significant change in the values of intensity of hybridisation of the probes by some consecutive steps. Firstly a t-test is used to determine differences between regions around all possible boundary points. Once all the candidate boundaries have been selected (a liberal p-value is used at this stage) there are sorted from small to high minimum p-values. In a second step the boundary candidates in the list with overlapping neighbourhoods are filtered to obtain a refined list of optimal non-overlapping boundary candidates. All the p-values are recalculated for the redefined neighbourhoods and a more stringent threshold is applied here. Finally, regions at both sides of each boundary candidate are again compared with a t-test. If they are not significantly different in their average hybridisation values, then they are merged as a unique region. Otherwise they define two regions with different copy number value. This is a simple and quick procedure that allows for easily changing from fine to coarse resolution by modifying the thresholds for the p-values.

We have compared isowindow to other two methods for breakpoint detection, GLAD **[15]** and circular binary segmentation (CBS) **[14]**, which are among the best performers. **[4]** In the GLAD method a likelihood function with weights determined adaptively is used to solve the copy number estimation problem locally based on data smoothed. Then, the algorithm finds, for each probe, the maximal neighbourhood in which the local constant assumption holds. Each of the constant pieces of the line define a block of probes with similar copy number among them and different copy number from that of the nearby regions. On the other hand, the CBS method selects firstly a segment of the data (a group of probes that are all consecutively arranged in the genome or in a chromosome). The copy number measures of the probes in that segment are compared to those in the reminder dataset using a t-statistic. Hence, the method can distinguish whether the segment chosen has a copy number that is higher or lower than the overall copy number in the data, assumed to be the normal reference. This scheme is iterated exhaustively for all possible segments in the dataset, spotting those that correspond to regions of altered copy number.

An approximation to the relative performances of the methods used was obtained by means of simulated data sets. Such datasets were generated by means of a piecewise constant function plus random alterations normally distributed with mean value and three different levels for the standard deviation (corresponding to noise levels 0.2, 0.5 and 1). A mean value of 0 would correspond to a normal region, without copy number alterations, while mean values lower and higher would correspond to deletions or amplifications at different degrees, respectively. Amplified and deleted regions of different sizes are randomly situated within the simulated normal chromosome and the methods have to locate them at different noise levels. The method proposed here performs at least as well as the GLAD and CBS (Table 1) while being more efficient in terms of runtimes. Isowindow shows a better performance in finding small amplicons.

	Method		
Noise level	GLAD	Isowindow	CBS
0.2	96.9	100.0	90.6
0.5	40.6	62.5	87.5
1.0	9.4	21.9	21.9

Table 1: Percentage of success in finding copy number alterations in the simulation of the four methods for copy number estimation included in ISACGH

Functional profiling of regions with copy number alterations

The final aim of a Array-CGH experiment is to find a molecular explanation for the effects of the detected copy umber alterations. The interpretation of genome-scale data is usually performed in two steps: in a first step genes of interest are selected in this case because they are located in the amplified (or lost) region detected. In a second step, the selected genes of interest are compared to the background (here the rest of genes in the chromosome) in order to find enrichment in any functional category (gene ontology, KEGG pathways, etc.) This comparison to the background is required because otherwise the significance of a proportion (even if high) cannot be determined. Different approaches have been developed to this end. **[11]** Here we will use the FatiGO+ (16) program, which uses a Fisher's exact test to determine the enrichment in different functional categories including gene ontology, KEEG pathways, Interpro functional motifs, Swissprot keywords and some regulatory elements such as transcription factor binding sites or other regulatory motifs. **[17]**

Discussion:

We have implemented all the described functionalities in a program, ISACGH (an acronym for *In Silico* Array CGH), which is used to illustrate the concept of functional profiling of CGH arrays with an example of multiple myeloma (MM), an incurable form of haematological neoplasia.

Nine MM cell lines were obtained from the DSMZ (Deuche Sammlung von Mikroorganismen und Zelkuturen GmbH, Braunschweig, Germany) and were cultured under recommended conditions. DNA and RNA were extracted using supplier's protocols. Microarray assays were performed using the CNIO OncoChip, which contains 7657 different cDNA clones of cancer related genes. **[18]** CGH experiments onto cDNA arrays and hybridisation were performed as described in **[13]** and quantified using the GenePix Pro 5.0 software (Axon Instruments Inc., Union City, CA). Cy3/Cy5 ratio values were normalized using the DNMAD tool from the GEPAS **[19, 20, 21]** and the resulting data were transformed to log2 ratios. Our purpose was to identify any possible region that contained copy number gains (amplifications), to study the expression of the genes included in that particular region and to understand the possible functional consequences of such alterations.

Using the segmentation method as implemented in the ISACGH we could detect a putative amplicon in the chromosome 18 (which remained undetected with both GLAD and CBS, because of the low density of the array, although the effect would have been the same in a high density arrays with a small amplicon) The figure shows the region (left) and the slight, although appreciable, differences in gene expression levels within the amplicon (right).

A unique feature offered by ISACGH is the possibility of obtaining a functional profile of the detected chromosomal regions. When the amplicon is analysed through the FatiGO+ program **[16, 17]** a number of GO terms arise as over-represented in the genes contained in such region. Thus, the GO terms regulation of cellular process (GO:0050794) and regulation of physiological process (GO:0050791) were significantly over-represented in the amplicon (FDR adjusted p-value= 0.0336). Genes annotated with these terms were: BCL2, MALT1, NEDD4L, MBD2, TNFRSF11A and TCF4. Some of them have annotations at more detailed levels in GO, although the number of genes was too small as to produce statistically significant results. For example BCL2 and MALT1 are annotated as negative regulation of programmed cell death (GO:0043069). These observations suggest that some processes altered, that ultimately lead to diseases, are not produced by the deregulation of one unique gene, but are the combined result of simultaneous deregulations of genes involved in a pathway or a particular biological function. In addition, these findings stress the importance of the use of functional profiling methods for the proper understanding and interpretation of the results of the genome-scale experiments. This unique feature included in ISACGH is of extreme importance since growing evidence suggests the existence of clusters of functionally related genes in the chromosomes **[8]** and the possible impact on diseases such as cancer. **[10]**

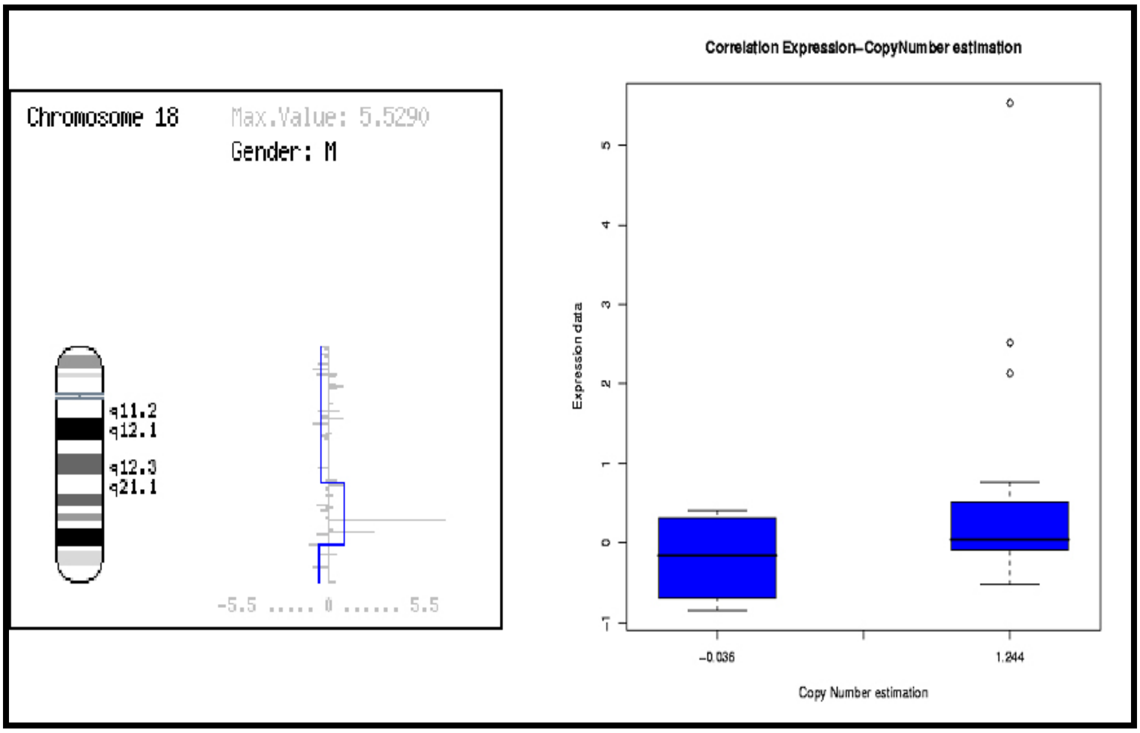

Figure 1: Detection on an amplicon in the chromosome 18 and the relationship between copy number estimation and gene expression. Left: the blue line represents the copy number estimation and the grey bars represent the individual gene expression values represented onto the same chromosomal coordinates. Right: boxplots of gene expression values for the regions with no copy number (a log-ratio of approximately 0) and the amplicon region, which is a duplication (a log-ratio of approximately 1) There is a slight increase in gene expression values in the region of the amplicon

Although ISACGH **[22]** can be used alone, it is tightly integrated in the GEPAS package. **[19, 21, 23]** GEPAS, that stands for Gene Expression Profile Analysis Suite (GEPAS), constitutes one of the most complete resources for microarray data analysis available over the web. GEPAS includes facilities for normalisations, clustering, gene selection, predictors and functional profiling. Thus, different operations (including pre-processing or normalization) can directly be performed within the same environment, without the necessity of any file reformatting step.

Conclusion:

Despite a number of applications dealing with the estimation of genomic copy number have been recently published **[4]**, there are different aspects of the analysis of Array CGH data that have been poorly addressed or even ignored. Recent evidences strongly support the existence of regional arrangements of functionally related genes **[8]**, with obvious consequences for the understanding of diseases characterised by copy number alterations, such as an important number of cancers. **[10]** This fact reduces the validity to the classical vision, in which one or a few key genes would be the causative factors of the disease, and urges to take into consideration the functional dimension in the interpretation of the effects of copy number alterations. In this new scenario, the deregulation of blocks of functionally related genes located in the chromosomal regions with copy number alterations would be behind the disease phenotype.

The methods for functional profiling have proven in many scenarios its usefulness. An obvious challenge is to increase our knowledge in different aspects of function and cooperation between genes in order to be able of applying this methods in a way that allows us to unravel new unknown functional aspects of the biology of the cell and their connections to pathologies.

Acknowledgement:

This work is supported by grants from Fundació La Caixa, NRC Canada-SEPOCT Spain, project BIO 2005-01078 from the MEC and National Institute of Bioinformatics (www.inab.org) a platform of Genoma España.

References:

[01] A. Kallioniemi, *et al., Science*, 258:818 (1992) [PMID:1359641]
[02] D. G. Albertson & D. Pinkel, *Hum Mol Genet.*, 2:R145 (2003) [PMID:12915456]
[03] E. H. Mahlamaki, *et al., Neoplasia*, 6:432 (2004) [PMID:15548351]
[04] W. R. Lai, *et al., Bioinformatics*, 21:3763 (2005) [PMID:16081473]
[05] M. Heidenblad, *et al., Oncogene*, 24:1794 (2005) [PMID:15688027]
[06] D. Pinkel & D. G. Albertson, *Nat Genet.*, 37:S11 (2005) [PMID:15920524]
[07] H. Caron, *et al., Science*, 291:1289 (2001) [PMID:11181992]
[08] L. D. Hurst, *et al., Nat Rev Genet.*, 5:299 (2004) [PMID:15131653]
[09] N. Stransky, *et al., Nat Genet.*, 38:1386 (2006) [PMID:17099711]
[10] Y. Zhou, *et al., Cancer Res.*, 63:5781 (2003) [PMID:14522899]
[11] J. Dopazo, *Omics*, 10:398 (2006) [PMID:17069516]
[12] S. Datta & S. Datta, *BMC Bioinformatics*, 7:397 (2006) [PMID:16945146]
[13] C. Largo, *et al., Haematologica*, 91:184 (2006) [PMID:16461302]
[14] A. B. Olshen, *et al., Biostatistics*, 5:557 (2004) [PMID:15475419]
[15] P. Hupe, *et al., Bioinformatics*, 20:3413 (2004) [PMID:15381628]
[16] F. Al-Shahrour, *et al., Bioinformatics*, 20:578 (2004) [PMID:14990455]
[17] F. Al-Shahrour, *et al., Nucleic Acids Res.*, 33:W460 (2005) [PMID:15980512]
[18] L. Tracey, *et al., Am J Pathol.*, 161:1825 (2002) [PMID:12414529]
[19] http://www.gepas.org
[20] J. Herrero, *et al., Nucleic Acids Res.*, 32:W485 (2004) [PMID:15215434]
[21] D. Montaner, *et al., Nucleic Acids Res.*, 34:W486 (2006) [PMID:16845056]
[22] http://isacgh.bioinfo.cipf.es
[23] J. Herrero, *et al., Nucleic Acids Res.*, 31:3461 (2003) [PMID:12824345]

Edited by Susmita Datta

Citation: Conde *et al.*, Bioinformation 1(10): 432-435 (2007)