

# Relations for array CGH and Gene Expression datasets

Mira Oh<sup>1</sup>, Bongjun Song<sup>2</sup>, Hyunju Lee<sup>3</sup>

## 1 Introduction

Genomic-wide technologies have been undergoing rapid and vast changes. To this end, genomic wide array comparative genomic hybridization (array CGH) detects chromosome copy number aberrations (CNAs) at a higher resolution level than conventional chromosome-based CGHs. CNAs, such as amplifications or deletions or genomic regions or entire chromosomes, which allows identification of new cancer related genes (Pinkel and Albertson, 2005).

CNAs can be identified in cancers by altering the expression of genes within a specific region as gene expression is measured using the samples as used CNAs. From subsequent copy number and expression analysis, the significant impact of widespread CNAs can be identified (Perou *et al.*, 2000; Pollack *et al.*, 2002). Therefore, the combining of CNAs and gene expression datasets have the potential to identify new cancer biomarkers.

In this paper, we applied frequency thresholds for combining multiple samples and computed correlations between CNAs and gene expressions by r-square.

## 2 Frequency threshold and correlation

Frequency of amplifications and deletions in multiple samples is used to detect aberrated regions consistent across samples. Frequency thresholds are determined by a permutation approach method. A permutation consists of a random rearrangement significance copy number changes. Genes in array CGH datasets are permuted  $B$  times. P-value of each gene is calculated by comparing permuted values and observed value. Then, false discovery rate (FDR) is used for correcting multiple hypothesis testing.

For each gene  $i$ , let  $F(i)$  be the fraction of amplifications in the observed data  $F_k(i)$  be the fraction of amplification in the  $k$ -th permuted data. Then the following value is calculated.

$$P(i) = \frac{\#\{F_k(i) \geq F(i)\}}{B}, \quad i = 1, 2, \dots, I,$$

where  $I$  is the number of total genes. Using FDR,  $P_c(i)$  is calculated to correct multiple hypothesis testing. Genes with  $P_c(i)$  less than significance level ( $\alpha$ ) are regarded as amplifications. Deletions are calculated similarly.

For each gene, correlations between CNAs and gene expression values in its neighbors within 4Mb are calculated using Pearson correlation coefficient.

## 3 Results

We used array CGH and gene expression datasets from 41 breast cancer samples (4 cell lines and 37 tumors), using cDNA microarray containing 6,095 different mapped human genes (Pollack *et al.*, 2002). Also, we used circular binary segmentation (CBS; Olshen *et al.*, 2004) method to detect CNAs in a single sample. Amplification and deletion thresholds for a single sample are used segment mean

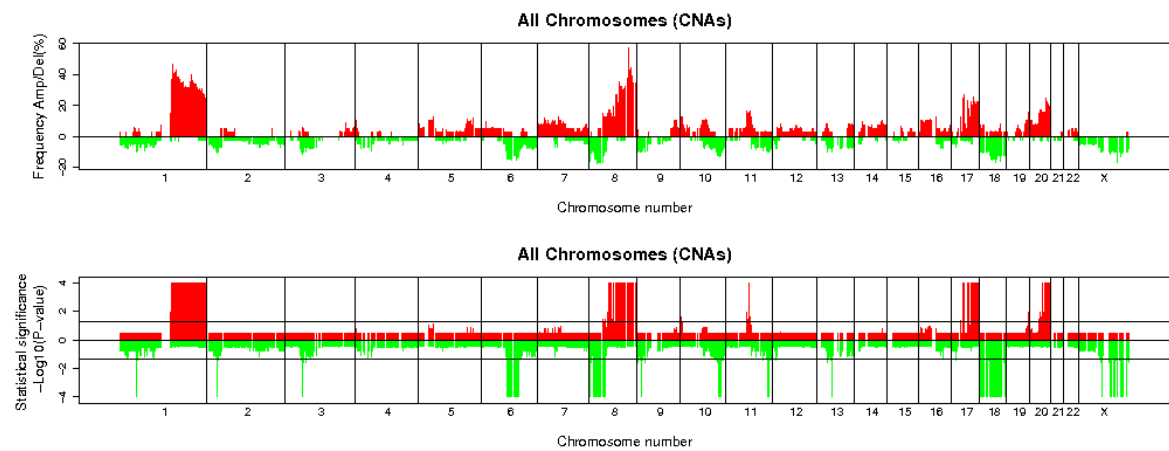
---

<sup>1</sup> Dept. of Statistics, National University of Chonnam, Republic of Korea. E-mail: omr@chonnam.ac.kr

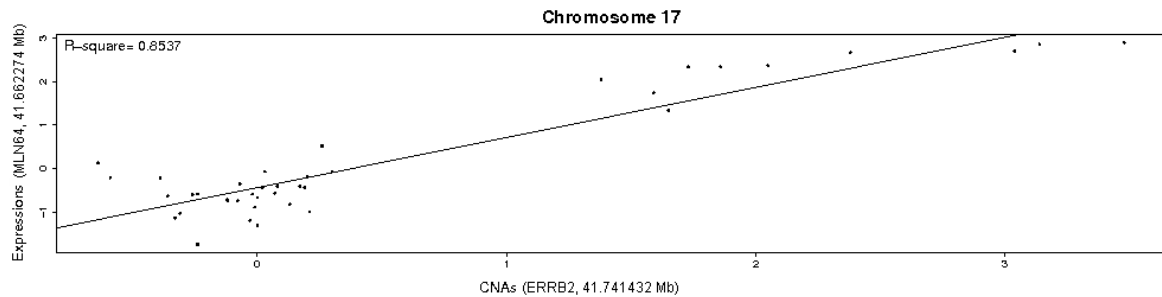
<sup>2</sup> Dept. of Information and Communications, Gwangju Institute of Science and Technology, Republic of Korea.  
E-mail: jsoju@gist.ac.kr

<sup>3</sup> Dept. of Information and Communications, Gwangju Institute of Science and Technology, Republic of Korea.  
E-mail: hyunjulee@gist.ac.kr

( $\pm 0.25$ ). Figure 1 represents CNAs in all chromosomes. Figure 2 shows correlations between CNAs and gene expressions based on r-squares.



**Figure 1: Amplifications and deletions in the breast tumors by CBS segmentation method.**



**Figure 2: Correlations between CNAs and gene expression values**

## 4 References

- [1] 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.
- [2] Perou,C.M., Sørlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen, L.A. *et al.* 2000. Molecular portraits of human breast tumours. *Nature (London)*, **406**, 747-752.
- [3] Pinkel,D. and Albertson,D.G. 2005. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, **37**, s11-s17.
- [4] Pollack,J.R., Sørlie,T., Perou,C.M., Rees,C.A., Jeffrey,S.S., Lonning,P.E., Tibshirani,R., Botstein,D., Børresen-Dale,A. 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. *Proceedings of the National Academy of Sciences*, **99**, 12963-12968.