Mining published lists of cancer related microarray experiments: Identification of a gene signature having a critical role in cell-cycle control

Finocchiaro G.(1,2), Mancuso F.(1,2), Muller H.(1,2)

(1) European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy (2) IFOM Foundation, Via Adamello 16, 20139 Milan, Italy

Motivation

The increasing availability and maturity of DNA microarray technology has led to an explosion of cancer related studies. In order to evaluate, integrate, and cross-validate the accumulating amount of publicly available microarray data, new methods are needed. Statistical methods have been applied to raw microarray data to identify intersections of multiple gene expression signatures from a diverse collection of data sets [1, 2]. However, this analysis is very stringent and time consuming [3]. Here we propose a fast method to detect common patterns in gene expression data by mining published lists of microarray experiments

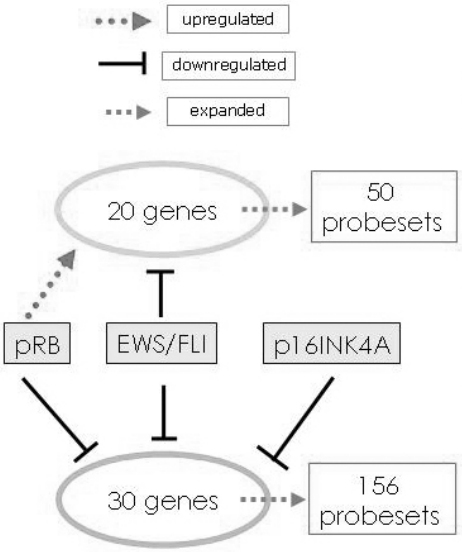
Methods

Published list of microarray experiments From the Affymetrix database of scientific publications [4], we selected 155 papers concerning both expression profiling of cancer specimens and studies of cancer related genes. From each paper, we extracted lists of genes indicated by the authors as statistically significant. We obtained 708 lists that have been annotated using a validated procedure [5]. In total, we annotated 7225 non-redundant Locuslink IDs. Microarray data We considered microarray experiments on cell lines that conditionally express a constitutively active phosphorylation site mutant of pRB or p16INK4A [6]. pRB and p16INK4A block cell cycle progression when over expressed and both proteins act as tumor suppressors [7]. pRB and p16INK4A experiments were analyzed separately using GenePicker [8]. Extraction of lists with a significantly overlapping signature Since different lists originate from different platforms, we calculated for each list the population of represented Locuslink IDs that was then used to find common genes between lists. Significance of overlapping gene content between lists was judged using the hypergeometric distribution. Seed clustering method For gene lists overlapping significantly with the lists of pRB and/or p16 regulated genes, the raw data were downloaded and used for further analysis. We used the overlapping set of genes as a seed to find genes having a similar behavior in the raw data by calculating the centroid using the median value for the overlapping genes. The Pearson's correlation coefficient ($r > 0.9$) was used to identify genes having a trend similar to the centroid. Gene Ontology classification To evaluate the enrichment of Gene Ontologies in the signatures extracted, we used a modified version of GeneMerge [9].

Results

Using GenePicker's genetic algorithm [8], we extracted 145 genes regulated by p16 (126 down, 19 up) and 328 genes regulated by pRB (200 down, 128 up). Both sets of genes were mapped on the published lists using the hypergeometric distribution (p-value < 0.001). As expected, we extracted experiments related to pRB and p16 pathways, e.g. a study on mechanisms of transcriptional regulation by pRB-E2F [10] (p-values: p16INK4A 0.00008, pRB 0.00074).

Interestingly, a strong overlap with an experiment concerning the p53 dependent growth arrest induced by Ewing's sarcoma oncoprotein EWS/FLI [11] was detected. The SOM cluster of down regulated probes was particularly enriched. In this cluster we identified: 30 genes down regulated also by pRB and p16INK4A; 20 genes up regulated by pRB (see Figure). These signatures were expanded to include genes not detected in the original publications and genes not annotated with a Locuslink ID. We obtained 156 probesets down regulated by pRB, p16 and EWS/FLI and 50 probesets up regulated by pRB and down regulated by EWS/FLI. As revealed by Gene Ontology analysis, both signatures are enriched in genes involved in cell cycle control. Down regulated genes have a critical role in G1/S phase and pRB up regulated genes have a major role in cytokinesis and mitosis. The significance of the identified overlap of pRB, p16, and pEWS-FLI regulated genes is underlined by the fact that patients with a mutated allele of the pRB gene after having been cured from bilateral retinoblastoma develop, among other cancers, Ewing sarcoma [12].



GO biological process definition	Sample Enrichment
cytokinesis	24.10
mitosis	29.25
mitotic sister chromatid segregation	71.31
cell cycle	9.27
regulation of cyclin dependent protein kinase activity	20.37
regulation of mitosis	34.56
chromosome organization and biogenesis (sensu Eukaryota)	10.76

GO biological process definition	Sample Enrichment
DNA replication	19.37
DNA replication initiation	27.73
DNA repair	12.10
mismatch repair	12.68
DNA replication, synthesis of RNA primer	42.214
cell cycle checkpoint	15.85
DNA replication checkpoint	25.37
regulation of DNA replication	25.37

Contact email: giacomo.finocchiaro@ifom-ieo-campus.it

Supplementary Information: References 1. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 2004, 101(25):9309-9314. 2. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res 2002, 62(15):4427-4433. 3. Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M: Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet 2003, 19(10):570-577. 4. http://www.affymetrix.com/community/publications/index.affx 5. Guffanti A, Finocchiaro G, Reid JF, Luzi L, Alcalay M, Confalonieri S, Lassandro L, Muller H: Automated DNA chip annotation tables at IFOM: the importance of synchronisation and cross-referencing of sequence databases. Appl Bioinformatics 2003, 2(4):245-249. 6. Vernell R, Helin K, Muller H: Identification of target genes of the p16INK4A-pRB-E2F pathway. J Biol Chem 2003, 278(46):46124-46137. 7. Sherr CJ: Cancer cell cycles. Science 1996, 274(5293):1672-1677. 8. Finocchiaro G, Parise P, Minardi SP, Alcalay M, Muller H: GenePicker: replicate analysis of Affymetrix gene expression microarrays. Bioinformatics 2004, 20(18):3670-3672. 9. Castillo-Davis CI, Hartl DL: GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 2003, 19(7):891-892. 10. Young AP, Nagarajan R, Longmore GD: Mechanisms of

transcriptional regulation by Rb-E2F segregate by biological pathway. Oncogene 2003, 22(46):7209-7217. 11. Lessnick SL, Dacwag CS, Golub TR: The Ewing's sarcoma oncoprotein EWS/FLI induces a p53-dependent growth arrest in primary human fibroblasts. Cancer Cell 2002, 1(4):393-401. 12. Ceha HM, Balm AJ, de Jong D, van 't Veer LJ: Multiple malignancies in a patient with bilateral retinoblastoma. J Laryngol Otol 1998, 112(2):189-192.