BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 7 2008, pages 889–896
doi:10.1093/bioinformatics/btn034

Genome analysis

Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes

Hyunju Lee[1,4,†], Sek Won Kong[2,3,†] and Peter J. Park[1,2,*]

[1]Harvard-Partners Center for Genetics and Genomics, [2]Informatics Program, Children's Hospital, [3]Department of Cardiology, Children's Hospital, Boston, MA, USA and [4]Department of Information and Communication, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea

Received on September 20, 2007; revised on December 29, 2008; accepted on January 22, 2008
Advance Access publication February 8, 2008
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: DNA copy number aberrations (CNAs) and gene expression (GE) changes provide valuable information for studying chromosomal instability and its consequences in cancer. While it is clear that the structural aberrations and the transcript levels are intertwined, their relationship is more complex and subtle than initially suspected. Most studies so far have focused on how a CNA affects the expression levels of those genes contained within that CNA.

Results: To better understand the impact of CNAs on expression, we investigated the correlation of each CNA to all other genes in the genome. The correlations are computed over multiple patients that have both expression and copy number measurements in brain, bladder and breast cancer data sets. We find that a CNA has a direct impact on the gene amplified or deleted, but it also has a broad, indirect impact elsewhere. To identify a set of CNAs that is coordinately associated with the expression changes of a set of genes, we used a biclustering algorithm on the correlation matrix. For each of the three cancer types examined, the aberrations in several loci are associated with cancer-type specific biological pathways that have been described in the literature: CNAs of chromosome (chr) 7p13 were significantly correlated with epidermal growth factor receptor signaling pathway in glioblastoma multiforme, chr 13q with NF-kappaB cascades in bladder cancer, and chr 11p with Reck pathway in breast cancer. In all three data sets, gene sets related to cell cycle/division such as M phase, DNA replication and cell division were also associated with CNAs. Our results suggest that CNAs are both directly and indirectly correlated with changes in expression and that it is beneficial to examine the indirect effects of CNAs.

Availability: The code is available upon request.

Contact: peter_park@harvard.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.
†The authors wish to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

1. INTRODUCTION

Nearly all cancers are caused by abnormalities in the DNA (Vogelstein and Kinzler, 2004). Structural changes of chromosomal regions such as aneuploidies, translocations, copy number aberrations (CNAs) and point mutations have been observed in various tumors (Lengauer *et al*., 1998). Among these, CNAs represent both amplifications and deletions of chromosomes, often ranging from 0.5 to 10 Mb in size. CNAs of oncogenes and tumor suppressor genes have been reported as causatively related with initiation, development and progression of cancer (Albertson *et al*., 2000; Pinkel and Albertson, 2005). With the maturation of microarray technology, CNAs studies using high-resolution array comparative genomic hybridizations (aCGH) have been performed in many types of cancer, including brain, prostate, colon, pancreatic and lung cancers (Chaudhary and Schmidt, 2006; Liu *et al*., 2006; Phillips *et al*., 2006; Pole *et al*., 2006; Tonon *et al*., 2005). These genome-wide chromosome copy number data have accelerated cancer research by allowing identification of new candidate cancer loci, classification of cancer subtypes and discovery of molecular mechanisms of cancers. In addition, meta-analyses of published aCGH datasets have revealed a relationship between the CNA pattern and cancer cell lineages (Jong *et al*., 2007; Myllykangas *et al*., 2007).

While CNAs are structural changes, measuring the level of transcripts provides additional information on whether those changes have functional consequences. Genome-wide profiling of gene expression (GE) has already shown promising possibilities in classification of cancer, prediction of treatment responses and discovery of correlated events in the clinical data such as metastasis (Bild *et al*., 2006). So far, several groups performed systematic studies to check whether CNAs are directly associated with transcriptional changes of the genes contained in those CNAs (Chaudhary and Schmidt, 2006; Jarvinen *et al*., 2006; Phillips *et al*., 2006; Pollack *et al*., 2002; Stranger *et al*., 2007). Hyman *et al*. (2002) analyzed a set of aCGH and GE profiles from the same 14 breast cancer cell lines hybridized on cDNA microarrays. They calculated the mean difference in gene expression between samples with and without amplifications divided by standard deviations for each gene and

compared with those from random permutations for estimating statistical significance. They reported that 44% of the highly amplified genes (>2.5 in copy number ratio) were up-regulated and that the percentage decreased with a lower level of amplification. Using the same statistical method, Jarvinen *et al.* (2006) analyzed CNAs and GEs from laryngeal squamous cell carcinoma cell line and found that 39% of amplified regions were up-regulated and 14% of deleted regions were down-regulated. These percentages decrease in the primary tumors: only 18% of amplified regions are up-regulated and there were no changes in the deleted regions. Chaudhary and Schmidt (2006) stimulated the prostate cancer cell line DU145 with serum and found that a large proportion of genes in deleted regions were down-regulated, but most genes in amplified regions did not show any change in GE. Although different tumor types and quantification methods can give varied estimates, these results clearly demonstrate the high impact of copy number in the transcription of those genes contained in the aberration. This direct relationship between structural changes in the DNA and gene expression has been used to identify or verify candidate cancer genes and pathways (Chin *et al.*, 2006; Hyman *et al.*, 2002; Ruano *et al.*, 2006; Soroceanu *et al.*, 2007; Sweet-Cordero *et al.*, 2006; Wolf *et al.*, 2004; Yao *et al.*, 2006).

These studies suggest that the relationship between CNAs and GEs is not simple and that the positive correlations are often but not always observed. The interaction between the two is further complicated by distant interactions in which a CNA can impact the expression of genes located elsewhere. For instance, Soroceanu *et al.* (2007) observed in glioblastoma that the DNA loss in PTEN, a known oncogene located in chr 10, is highly correlated with over-expression of IGF or EGFR, both of which are located away from chr 10. In the following, we call the relationships between CNAs and GE in the same location as a *direct* interaction and those in the different locations as an *indirect* one.

In the current study, we investigate both the direct and indirect relationships between structural changes measured by aCGH and functional changes measured by expression arrays, by analyzing three data sets in which both the copy number and expression were available. For this type of integration, there are several difficulties to overcome. The first is that the choice of data sets is limited. While both aCGH and expression data sets are plentiful, paired data sets with both DNA and RNA data on the same set of patients are scarce. It is possible to infer relationships from unpaired data sets, but that process is prone to false positives. The second issue is that the probes in the two platforms generally vary greatly, both in array type and in resolution. The newer aCGH arrays have oligonucleotide probes with much higher resolution, but the arrays in the data sets we use are two channel arrays using Bacterial Artificial Chromosomes (BACs) and thus have a low resolution, on the order of 1 MB. The platforms for expression data, on the other hand, are generally oligonucleotide arrays with higher resolution. Reconciling between the two requires resolving the many-to-one or one-to-many mappings in each chromosomal segment and may require judicious averaging of the probe values in the higher resolution platform. The third difficulty is that many genes are co-expressed and that CNAs occur simultaneously in multiple locations (Chin *et al.*, 2006). This limits the precision in locating the interacting partners. In the proposed approach, we thus deduce a set of modules, each module containing a group of co-expressed genes and a group of co-occurring CNAs. These two groups are highly correlated and provide sufficient information for pathway analysis. The relationships inferred by these modules involve distant loci and are thus fundamentally different from those derived in previous studies. Below, the proposed approach is described in detail and is applied to three data sets containing glioblastoma multiforme (GBM), bladder and breast cancer samples. In all cases, we observe that cell-cycle related pathways are enriched. More importantly, we identify several statistically significant CNAs that are associated with disease-specific pathways in each case.

2. METHODS

2.1. Data sets

The method is illustrated in Figure 1. We collected and reanalyzed three paired data sets: 34 GBM samples (Nigro *et al.*, 2005), 57 bladder tumor samples (Stransky *et al.*, 2006) and 89 breast tumor samples (Chin *et al.*, 2006). Each sample consists of a BAC array for measuring copy number and an Affymetrix GeneChip for measuring expression. Each of three datasets contained about 2400 BAC probes at an approximately megabase interval. For copy number changes, log ratio to normal samples were used as described in the original publications. Gene expression index was recalculated with the raw data (CEL files) using the GCRMA algorithm (Hubbell *et al.*, 2002). When multiple probe sets were mapped to the same RefSeq ID, we calculated the geometric mean after excluding the probe sets (with _x_at suffix) that do not map uniquely to the genome. Log-transformed values were used for further evaluation and statistical procedures.

2.2. Measuring association between CNA and expression

To investigate the association between CNAs and gene expression changes, we used the Pearson Correlation Coefficient (PCC). We first selected BAC array probes. Since many probes did not show any aberrations and thus are no longer of interest, we selected a subset of BAC probes for further analysis using the following criterion: CNAs with probes among the top 12.5% of the amplifications or the bottom 12.5% of the deletions for at least twenty percent of samples. Using PCC, we computed the association between all pairs of selected BACs from aCGH and RefSeq ID from gene expression data. The results were stored in the Correlation Matrix, as illustrated in Figure 1C. We defined the association as direct when the BAC probes and RefSeq genes were located on the same cytoband, and all other significant associations were defined as indirect. We note that while a segmentation algorithm is generally used to process aCGH data (Lai *et al.*, 2005), it results in a loss of sensitivity in this analysis, as the spatial averaging fails to take advantage of the full range of the observed log-ratios for a given probe. This is particularly true for the BAC data sets we consider here.

2.3 Biclustering for identification of modules

Because the occurrences of many CNAs are highly correlated, it is difficult to accurately distinguish among them. The same is true for expression profiles. Thus, rather than trying to relate a particular CNA with the expression of a particular gene, we search for *a set of CNAs* and *a set of expression profiles* that are highly correlated, using a biclustering approach.

Biclustering has been popular in expression profiles studies, as it attempts to find a subset of genes having similar expression patterns

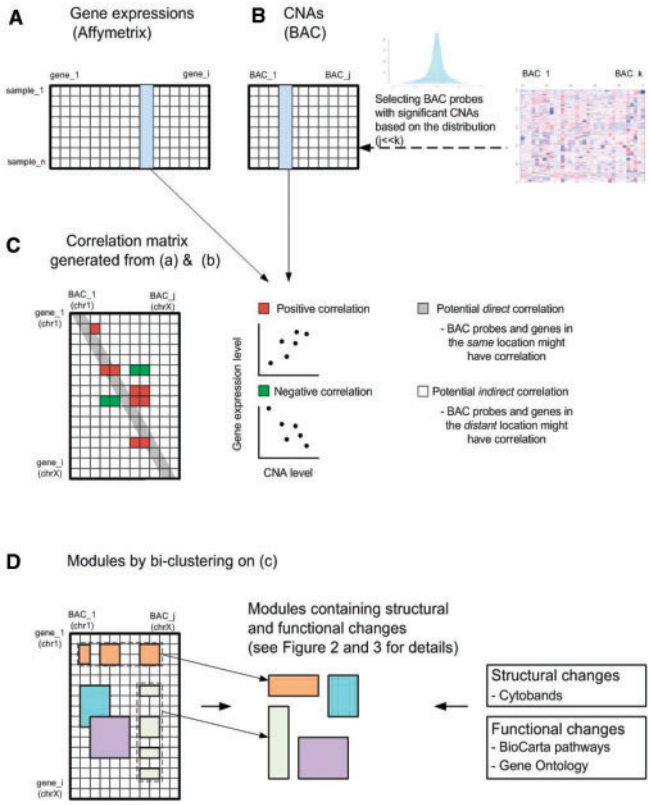

Fig. 1. A schematic of our approach. (**A**) A gene expression data set and (**B**) its paired CNA data set are collected. For CNAs, we choose BAC probes that show amplification or deletion in a given fraction of patients. (**C**) For every pair of genes and the selected BAC probes, the Pearson correlation coefficient is calculated and the correlation matrix is generated. (**D**) A biclustering algorithm is applied to the correlation matrix to obtain modules containing highly correlated genes and BAC probes. Each module is tested for enrichment of gene sets including those from Gene Ontology, Biocarta pathways and cytobands.

under a group of conditions. Such an entity is often called a module. For a comparison of various biclustering algorithms, see Prelic *et al.* (2006). In this study, a biclustering algorithm called SAMBA (Tanay *et al.*, 2004) was used to identify associated CNAs and gene expression changes (Fig. 1D). The statistical significance of generated modules was calculated by a method based on the framework developed by Tanay *et al.* (2004) and those modules with *P*-values smaller than 0.0001 were selected for further analysis (see Supplementary Material). This approach allows multiple appearances of genes and of conditions in several modules, reflecting a biological principle that genes can have multiple functions (Cheng and Church, 2000; Dudley *et al.*, 2005). The biclustering approach is appropriate for the present study, as both CNAs and genes with expression changes may participate in multiple pathways and loci distributed across different chromosomes may be related to the same biological pathway.

2.4. Enriched pathways in CNV-GE modules

To determine functional relevance of the modules identified, we tested whether the genes from expression data contained in a module were enriched for specific biological functions or signaling pathways (Fig. 1D). We collected the gene sets of biologically related functions from Gene Ontology (GO) using the annotation package in Bioconductor (http://www.bioconductor.org). Biological process GO terms with sizes between 5 and 250 were used to exclude too specific or too general ones. Additional gene sets were downloaded from the Molecular Signature Database (MSigDB) at the Broad Institute (http://www.broad.mit.edu/gsea/msigdb). We used three categories of gene sets from MSigDB: (C1) Cytobands, (C2) Manually curated pathways including BioCarta and (C3) Motif gene sets.

For each module, we calculated the hypergeometric statistics and the associated *P*-values to find enriched gene sets. To address the multiple comparison issue with respect to the large number of gene sets tested at the same time, we calculated the estimated false discovery rate using the *q*-values and <0.01 was used for enrichment threshold (Storey and Tibshirani, 2003).

To find out statistically significant structural components, we calculated a hypergeometric statistic for the enrichment of cytobands of BACs in a given module. To map BAC probes to cytobands, we used cytoband information downloaded from the UCSC golden path database (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/cytoBand.txt.gz).

3. RESULTS

3.1. Associations between CNAs and GEs

The first question was whether there are in fact many cases of strong association between distant loci. Isolated cases of such relationships have been observed, but there was no quantification of such effect previously.

PCC between all pairs of selected BACs from aCGH and RefSeq ID from gene expression data showed that the large proportion of significant associations was from different cytobands. When we controlled the significant associations as the top 1% of total number of associations, 2% (10 out of 515) of pairs in the same cytobands and 1% (4386 out of 439 151) of pairs in the different cytobands were significantly associated. The numbers were similarly high in bladder and breast cancer data sets.

These numbers clearly suggest that there are highly correlated distant loci and that studying the impact of CNAs only on the expression of those genes contained in the CNAs is not sufficient.

3.2. Modules from the biclustering method

Given our threshold for statistical significance, biclustering of CNA-GE correlation matrix generated 247, 339 and 506 modules for the GBM, bladder and breast cancer datasets, respectively. This was based on a less strict overlapping criterion between modules (Overlap factor 0.1 is used in [0,1] scale where 1 indicates non-overlap). Each module consisted of selected BAC probes and RefSeq genes that were highly correlated with each other. To select modules containing both structural and functional changes among them, we performed hypergeometric tests of the BAC probes for possible enrichment in a cytoband, and genes for gene sets as described in Methods.

3.2.1 Signaling pathway gene sets When we applied the BioCarta and the manually curated gene sets from the MSigDB C2 category, 20, 18, and 18 modules were significant in GBM, bladder and breast cancer, respectively. These modules

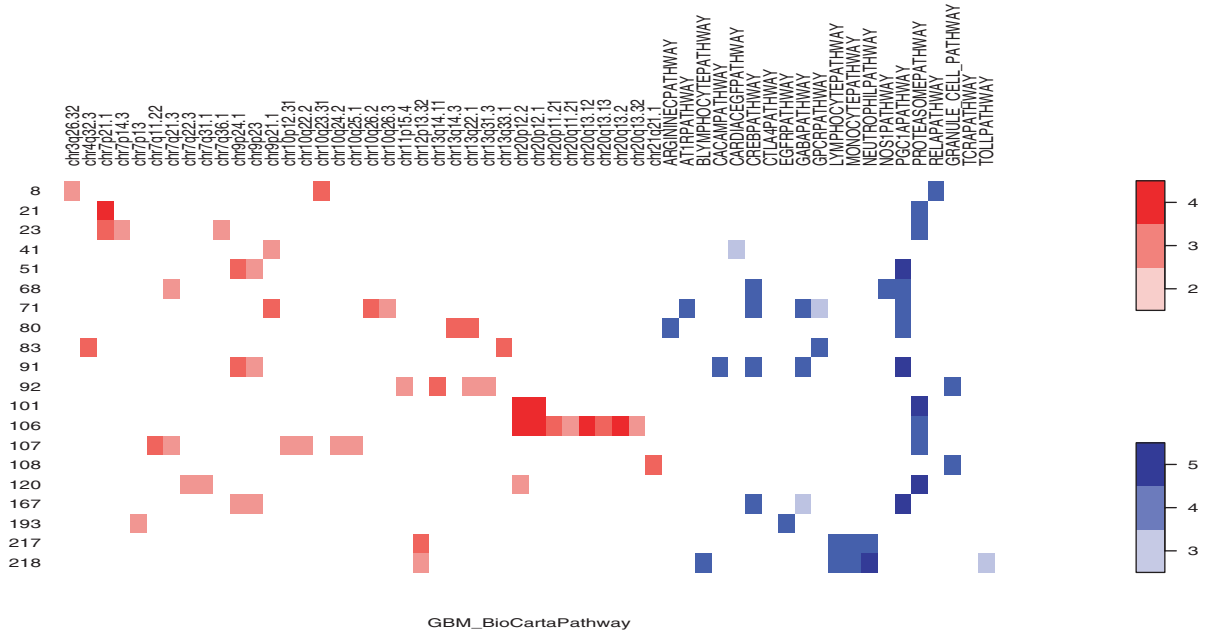

Fig. 2. Structural and functional changes observed in the GBM data set. The number on the *y*-axis is a module identifier and the names on the *x*-axis represent the enriched cytobands and pathways in the modules. The shading of the red color corresponds to the −log10 (*P*-value) from a hypergeometric test for the enrichment of BACs from a module in a cytoband. For clarity, −log10 (*P*-value)<2 is colored as white. Similarly, blue indicates the −log10 (*P*-value) from the enrichment test of BioCarta pathway in a given module. −log10 (*P*-value)<3 is colored as white.

identified signaling pathways that were highly correlated with the CNAs of distant locations (Fig. 2, Supplementary Figs S1 and S2).

In the GBM data set (Nigro *et al.*, 2005), the authors examined the paired data from 34 patients. They found that patient survival was significantly correlated with both CNAs and GE, and noticed that the aberration in a locus could be associated with the changes in expression on a different locus. For instance, they observed that CH3L1/YKL-40, a gene located on chr 1, had a strong correlation with CNAs of chr 10. Here we systematically investigated both direct and indirect association between CNAs and GE. Figure 2 shows 20 enriched categories from the MSigDB C2 gene sets for the GBM data. These categories were enriched in one or more modules deemed significant. The PGC1A pathway and the proteosome pathway, for instance, were found multiple times in different modules.

We queried all the genes from the enriched pathways in the PubMed database to check whether our findings have been reported previously. Official gene symbol and the name of specific cancer were used as keywords. We found the supporting evidence for a number of genes from the enriched modules. The result and the supporting references are summarized in Table 1. Here, we describe two examples from Table 1. Module 193 of the GBM dataset was significantly enriched for the epidermal growth factor receptor (EGFR) related pathway (uncorrected Fisher's exact *P*-value = 4.3E-05 and corresponding *q*-value <0.01), shown in Figure 3. Among the genes in this pathway, EGF, GAB3 and GRB7 were highly correlated with the CNAs of chr 7p13. This result is very interesting, for EGF itself is located on chr 4q25. It has been reported that EGFR in 7p12 is amplified in 30–50% of human GBM (Ruano *et al.*, 2006). Gefitinib, the EGFR kinase inhibitor, has been tried for the treatment of recurrent malignant glioma in selected cases (Mellinghoff *et al.*, 2005). The effectiveness of this treatment is still in debate; however, multiple lines of evidence showed that this pathway is altered in several types of cancer including GBM.

In another example, the calcium/calmodulin related pathway that includes CAMK2A, CAMK2B, CAMK2G, CAMKK2 and CALM3 was enriched in module 91. Calcineurin, a calmodulin binding protein, has been known as a brain tumor specific neuronal marker (Goto *et al.*, 1986). The potential implication of calmodulin-dependent phosphodiesterase in GBM is reviewed in Das and Sharma (2005).

In the second dataset, Stransky *et al.* (2006) identified the chromosomal region in bladder cancer samples where CNAs are partly responsible for the changes in gene expression. They discovered that several genes in a selected amplified region were regulated under the epigenetic control of H3K9 trimethylation and DNA methylation. Such regions were identified as copy number-independent regions of correlations using their Transcriptome Correlation Map, in which correlations among the expression profiles of adjacent probes are computed and stretches of probes with high correlations are selected. Copy number-dependent regions where levels of gene expression can be explained by CNAs in the same region were also identified.

In this study, we could find at least two signaling pathways where gene expression levels of genes with GO terms are correlated with CNAs in the different regions (Supplementary Fig. S4). GO:0043123 (Positive regulation of l-kappaB kinase/NF-kappaB cascade) was enriched in the three modules, where the BAC probes in chr 13q were significantly correlated with BCL10 (chr 1p22), TRAF3IP2 (chr 6q21), EDG2 (chr 9q31.3), TNFRSF1A (chr 12p13.2), LITAF (chr 16p13.13),

Table 1. Analysis of pathways for the GBM data set

Gene set name	Description	Module ID (BAC cytoband)	Genes in modules
ARGININEC	Catabolic pathways for arginine, histidine, glutamate, glutamine and proline	80 (13q14.3,13q22.1)	ALDH4A1,GLS, GLUD1,OAT
AT1R	Angiotensin II–mediated activation of JNK Pathway via Pyk2 dependent signaling	71 (9p21.1,10q26.2,10q26.3)	MEF2C,PTK2B (Lipinski *et al.*, 2005), PAK1,PRKCB1,MAPK3, MAP2K4,CALM3,ELK1
CACAM	Ca^{++}/ calmodulin-dependent protein kinase activation	91 (9p24.1,9p23)	CAMK2A,CAMK2B,CAMK2G, CAMKK2,CALM3 (Perry *et al.*, 2004)
EGF_RECEPTOR _SIGNALING		193 (7p13)	EGF,GAB1 (Kapoor *et al.*, 2004),GRB7,
GABA	Gamma-aminobutyric acid receptor life cycle	71 (10q26.2,10q26.3,9p21.1), 91 (9p24.1,9p23), 167 (9p24.1,9p23)	GABRA2 (Vlodavsky and Soustiel,2007), GABRA1,DNM1,GABRA5,NSF
GPCR	Signaling pathway from G-protein families	71 (10q26.2,10q26.3,9p21.1), 83 (13q33.1,4q32.3)	PRKAR1B,ADCY1,GNAI1 PRKAR2B,PPP3CB,PRKCB1, MAPK3,CALM3,PRKAR1B, PRKAR2B,PRKCB1 GNB1,PRKACB,GNAQ HRAS (Knobbe *et al.*, 2004),MAP2K1
NOS1	Nitric oxide signaling pathway	68 (7q21.3)	PRKAR1B (Lam-Himlin *et al.*, 2006), PRKAR2B,GRIN1,PPP3CB,PRKCB1 DLG4,CALM3
RELA	Acetylation and deacetylation of RelA in the nucleus	8 (10q23.31,3q26.32)	NFKB1 (Wu *et al.*, 2006),HDAC3, FADD (Schultze *et al.*, 2006), TNFRSF1A (Panner *et al.*, 2005)
ST_GRANULE_ CELL_SURVIVAL	Granule cell survival pathway	92 (11p15.4,13q14.11,13q22.1,13q31.3), 108(21q21.1) 108(21q21.1)	ITPKB,GNAQ,MAPK10, APC (Sunahara and Nakagawara, 2000), MAPK9,MAPK8IP1,MAPK8IP3, MAP2K4,MAPK8IP2,ASAH1, CXCL2,MAPK1
TOLL	Toll pathway	218 (12p13.32)	TLR2,CD14,LY96 FOS (Puli *et al.*, 2006),MAP2K3,IRAK1

Enriched pathways from the C2 category in MSigDB are listed. Each pathway can be enriched in more than one module. Among the genes in a given pathway, those in modules are listed.

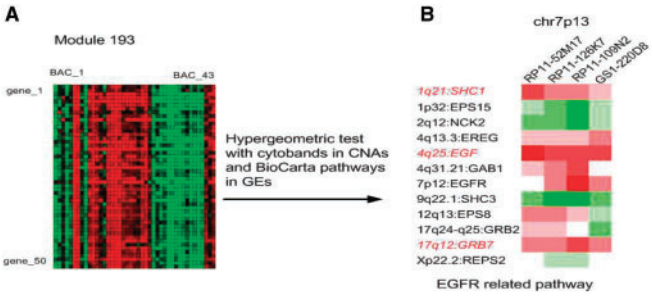

Fig. 3. (**A**) An example of a module from the GBM data set. This module contains a highly correlated set of 43 BAC probes and 50 genes. (**B**) Testing for enrichment of cytobands and pathways results in a module with four BAC probes located in chr7p13 and three genes (SHC1, EGF and GRB7) in EGFR related pathways.

TNFRSF10B (chr 8p21.3) and others. GO:0007249 (I-kappaB kinase/NF-kappaB cascade) was also enriched in a module. Interestingly, these two GO terms were closely related in terms of the number of genes shared by two, but the associated CNA loci were not same. It has been reported that NF-κB activates anti-apoptotic proteins and plays an important role in tumorigenesis and anticancer treatment (Dutta *et al.*, 2006). A complete result for the pathways is described in Supplementary Table S1.

In Chin *et al.* (2006), the authors investigated the correlations between copy number, expression and treatment responses in breast cancer. They found four regions of recurrent amplification associated with poor outcome and identified 66 genes cis-regulated by CNAs, with many genes known to be important for cancer progression. We applied our proposed method, and the result is summarized in Supplementary Table S2. The RECK pathway (Inhibition of Matrix Metalloproteinase) was significant among the MSigDB C2 gene sets. Four genes, RECK (chr 9p13.3), hRAS (chr 11p15.5), MMP2 (chr 17q12-21) and MMP14 (chr 14q11-12) were significantly correlated with chr11p15.4 and chr11p15.5. Down-regulation of RECK has been implicated in tumor angiogenesis and progression (Span *et al.*, 2003), but its role in breast cancer has not been reported yet. Our result results that RECK regulated MMPs

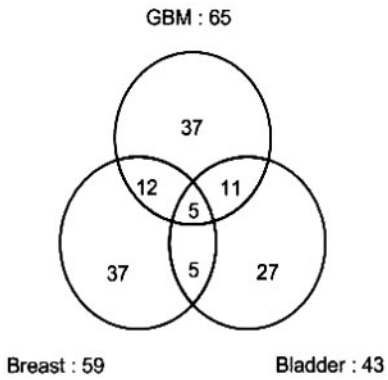

Fig. 4. Overlap among the enrichment Gene Ontology categories in the three cancer types. Among the five in the center are M phase (GO:0000279), DNA replication (GO:0006260), and cell division (GO:0051301). Cancer type-specific gene sets are described in the text.

in breast cancer should be investigated further. Epidermis development (GO:0008544) was also significantly enriched. Six genes EMP1 (chr 12p12.3), PPARD (chr 6p21.31), PLOD1 (chr 1p36), LAMC2 (chr 1q25.3), LAMB3 (chr 1q32.2) and BNC1 (chr 15q25.2) were significantly correlated with chr 3q25.33. EMP1 was reported as a novel marker of lobular breast carcinomas (Turashvili *et al*., 2007), and the loss of expression of LN5-encoding genes, LAMC2 and LAMB3, in breast cancer cell lines has also been observed (Sathyanarayana *et al*., 2003). Finally, we also found several modules that were highly enriched in immune responses, consistent with the role of immune system in developing and metastasis of breast cancer, as reviewed in de Visser *et al*. (2006).

3.2.2 Gene Ontology gene sets When we applied the Gene Ontology gene sets, 33, 29 and 52 modules were significant in GBM, bladder and breast cancer, respectively (Supplementary Figs S3, S4 and S5). In these modules, 65, 59 and 43 GO terms were enriched. The overlap among them are shown in Figure 4. Five significant GO terms were observed in common among three cancer types: M phase (GO:0000279), DNA replication (GO:0006260), locomotive behavior (GO:0007626), ATP synthesis coupled proton transport (GO:0015986) and Cell division (GO:0051301). Three of the GO terms (Cell division, M phase and DNA replication) are all tightly related to cell cycle, cell division and proliferation which are a signature of all types of cancer.

3.3 Advantage over analysis of a single data type

To show the strength of combining two data types, we also carried out pathway enrichment analysis for each type separately with the GBM data (Nigro *et al*., 2005) as an example. For GE, we used gene set enrichment analysis (GSEA) (Subramanian *et al*., 2005) to find differentially enriched gene sets between the two classes, 24 short term survivals (STS) and 10 long-term survivals (LTS), using the MSigDB C2 category of manually curated pathways, including those from Biocarta. We found that no gene sets are significant at FDR <0.25. When we decreased the significance level to the nominal *P*-value <0.01, 27 gene sets are enriched among genes up-regulated in STS and 1 gene set is enriched among genes up-regulated in LTS (data not shown). For aCGH, we first applied ISACGH (Conde *et al*., 2007) to identify the amplified and deleted regions in the chromosome by segmenting each sample. Then, we used FATIGO (Al-Shahrour *et al*., 2007) for enrichment test of Biocarta pathways between the genes in CNAs and the rest of the genes in the chromosome. When multiple-testing adjusted *P*-value was calculated for Fisher's exact test, there were no functionally enriched regions.

While it is not possible to definitively conclude that the pathways identified in the joint analysis is more functionally relevant than those from separate analysis, we have found that not many pathways are significant in single-data set analysis and that the list of pathways are significantly different. Because the relationship between the two data sets are exploited in the joint analysis, it is more likely to result in a more biologically meaningful set of pathways. We also note that the joint analysis can be carried out even when the phenotypic data (patient survival times in this case) are not available.

3.4 Higher resolution aCGH platforms

Our results above are based on the aCGH data with BAC probes, but the method obviously can be applied to the platforms of higher resolution. To illustrate this, we analyzed copy number data obtained from Affymetrix 100 K SNP arrays and expression data from Affymetrix U133 Plus 2.0 on 65 paired data sets of GBM patients (Kotliarov *et al*., 2006). Multiple probe sets were mapped into Refseq identifier and copy number estimates based on SNP probes were binned into 100 kb regions along the chromosome. A binned region was selected as amplified if the averaged log-ratio (base 2) in more than 30% of samples are greater than 2, and deleted if the log-ratio in more than 30% of the samples are less than −1.5 or that of 10% of samples are less than −2.0. Figure S6 summarizes the enriched modules. Interestingly, one of the modules was significantly enriched for cardiac epidermal growth factor pathway and, among the genes in this pathway, ADAM12, EGFR, JUN, EDN1 and PLCG1 were highly correlated with the CNAs of ch7p11.2. This shows that, while the overlap is not as strong as one would like, the two data sets from different platforms, especially for copy number estimation, commonly identify the important feature that structural changes of chr7p is correlated with functional changes of epidermal growth factor related pathways.

4 DISCUSSION

As the arrays for comparative genomic hybridization have increased in resolution, it has become possible to relate DNA copy numbers with changes in expression. Integrating these two data sets effectively is a challenging task, but is bound to result in new insights for the interplay between chromosomal instability and gene expression. A primary difficulty in this integrative analysis has been the lack of appropriate data sets. In a given cancer type, both expression data sets and copy number data sets abound, but *paired* data sets with expression and copy number from the same patients are solely lacking.

It is possible to carry out analysis even with unpaired data. For instance, Liu *et al.* (2006) identified physical clusters of genes with differential expression and then prioritized the clusters based on whether a CNA is present at the same location in a different set of GBM patients. However, the analysis becomes much more powerful when both types of data are derived from the same patient because the relationship can be inferred not just on averaged quantities but in each sample. Fortunately, there has been an increased recognition for such a design recently. In the ambitious Cancer Genome Atlas project from the National Institutes of Health (http://cancergenome.nih.gov), multiple data types including gene expression, copy number, microRNA, SNPs and DNA methylation are being generated on the same set of patients from three tumor types.

In the present work, we conducted a systematic study of how a copy number change at each location may be correlated with expression at every other location. Whereas previous studies have focused on their interaction at the same locus, we have extended this to long-range interactions. It is perhaps not surprising that there are highly correlated pairs on different loci, but the fraction of the high correlation at the same loci was extremely small (less than 1% among the most significant pairs).

One of the difficulties in an integrative study such as this is the mapping of the probes between different array platforms. Optimal probe design for expression is concentration of probes near the 3′ ends of known transcripts, whereas that for aCGH is more uniform spacing of the probes, often with higher density near oncogenes. Because of this difference in design, mapping between the two platforms involves averaging across probes in one platform to match a corresponding probe in another platform. This can result in loss of information from the higher resolution platform.

Because every pair of loci is examined, a large number of correlations is computed. Given the noise in the data, the ordering of all such pairs is not stable and interpreting each pair on the list is impractical. For a more robust analysis and clearer interpretation, we have used correlation analysis followed by biclustering to identify modules. Interpreting the modules is also not simple, however, as many of the modules have significant overlaps. Most biclustering methods have the advantage of allowing a row or a column to belong to multiple clusters, but a disadvantage is that many similar clusters can appear. Moreover, when pathway analysis is performed, each module can give multiple significant pathways. We have dealt with this problem by focusing on the pathways that appear multiple times among the clusters and examining the cluster in which the pathway has the most significant score. Our analysis of the pathways appears effective, with each data set giving tumor-type specific pathways as well as all of them giving the common cell-cycle/proliferation signatures.

ACKNOWLEDGEMENTS

This work was funded by the National Institutes of Health through R01 GM082798 and U54 LM008748 to PJP, and a systems biology infrastructure establishment grant provided by Gwangju Institute of Science and Technology to HL.

Conflict of Interest: none declared.

REFERENCES

Albertson,D.G. *et al.* (2000) Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat. Genet.*, **25**, 144–146.

Al-Shahrour,F. *et al.* (2007) FatiGO +:a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs andinteraction data with microarray experiments. *Nucl. Acids Res.*, **35**, W91–96.

Bild,A.H. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Chaudhary,J. and Schmidt,M. (2006) The impact of genomic alterations on the transcriptome: a prostate cancer cell line case study. *Chromosome Res.*, **14**, 567–586.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.

Chin,K. *et al.* (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**, 529–541.

Conde,L. *et al.* (2007) ISACGH:aweb-based environment for the analysis of Array CGH and gene expression which includes functional profiling. *Nucl. Acids Res.*, **35**, W81–85.

Das,S.B. and Sharma,R.K. (2005) Potential role of calmodulin-dependent phosphodiesterase in human brain tumor (review). *Oncol. Rep.*, **14**, 1059–1063.

de Visser,K.E. *et al.* (2006) Paradoxical roles of the immune system during cancer development. *Nat. Rev. Cancer*, **6**, 24–37.

Dudley,A.M. *et al.* (2005) A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol. Syst. Biol.*, **1**, 2005.0001.

Dutta,J. *et al.* (2006) Current insights into the regulation of programmed cell death by NF-kB. *Oncogene*, **25**, 6800–6816.

Goto,S. *et al.* (1986) Calcineurin as a neuronal marker of human brain tumors. *Brain Res.*, **371**, 237–243.

Hubbell,E. *et al.* (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.

Hyman,E. *et al.* (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.*, **62**, 6240–6245.

Jarvinen,A.-K. *et al.* (2006) Identification of target genes in laryngeal squamous cell carcinoma by high-resolution copy number and gene expression microarray analyses. *Oncogene*, **25**, 6997–7008.

Jong,K. *et al.* (2007) Cross-platform array comparative genomic hybridization meta-analysis separates hematopoietic and mesenchymal from epithelial tumors. *Oncogene*, **26**, 1499–1506.

Kapoor,G.S. *et al.* (2004) Distinct domains in the SHP-2 phosphatase differentially regulate epidermal growth factor receptor/NF-kappaB activation through Gab1 in glioblastoma cells. *Mol. Cell Biol.*, **24**, 823–836.

Knobbe,C.B. *et al.* (2004) Mutation analysis of the Ras pathway genes NRAS, HRAS, KRAS and BRAF in glioblastomas. *Acta Neuropathol. (Berl.)*, **108**, 467–470.

Kotliarov,Y. *et al.* (2006) High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. *Cancer Res.*, **66**, 9428–9436.

Lai,W.R. *et al.* (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Lam-Himlin,D. *et al.* (2006) Malignant glioma progression and nitric oxide. *Neurochem. Int.*, **49**, 764–768.

Lengauer,C. *et al.* (1998) Genetic instabilities in human Cancers. *Nature*, **396**, 643–649.

Lipinski,C.A. *et al.* (2005) The tyrosine kinase pyk2 promotes migration and invasion of glioma cells. *Neoplasia*, **7**, 435–445.

Liu,F. *et al.* (2006) A genome-wide screen reveals functional gene clusters in the cancer genome and identifies EphA2 as mitogen in glioblastoma. *Cancer Res.*, **66**, 10815–10823.

Mellinghoff,I.K. *et al.* (2005) Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. *N. Engl. J. Med.*, **353**, 2012–2024.

Myllykangas,S. *et al.* (2007) Specificity, selection and significance of gene amplifications in cancer. *Semin. Cancer Biol.*, **17**, 42–55.

Nigro,J.M. *et al.* (2005) Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. *Cancer Res.*, **65**, 1678–1686.

Panner,A. *et al*. (2005) mTOR controls FLIPS translation and TRAIL sensitivity in glioblastoma multiforme cells. *Mol. Cell Biol*., **25**, 8809–8823.

Perry,C. *et al*. (2004) CREB regulates AChE-R-induced proliferation of human glioblastoma cells. *Neoplasia*, **6**, 279–286.

Phillips,H.S. *et al*. (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*, **9**, 157–173.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet*., **37 (Suppl)**, 11–17.

Pole,J.C.M. *et al*. (2006) High-resolution analysis of chromosome rearrangements on 8p in breast, colon and pancreatic cancer reveals a complex pattern of loss, gain and translocation. *Oncogene*, **25**, 5693–5706.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA*, **99**, 12963–12968.

Prelic,A. *et al*. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Puli,S. *et al*. (2006) Signaling pathways mediating manganese-induced toxicity in human glioblastoma cells (u87). *Neurochem. Res*., **31**, 1211–1218.

Ruano,Y. *et al*. (2006) Identification of novel candidate target genes in amplicons of Glioblastoma multiforme tumors detected by expression and CGH microarray profiling. *Mol. Cancer*, **5**, 39.

Sathyanarayana,U.G. *et al*. (2003) Aberrant promoter methylation and silencing of laminin-5-encoding genes in breast carcinoma. *Clin. Cancer Res*., **9**, 6389–6394.

Schultze,K. *et al*. (2006) Troglitazone sensitizes tumor cells to TRAIL-induced apoptosis via down-regulation of FLIP and Survivin. *Apoptosis*, **11**, 1503–1512.

Soroceanu,L. *et al*. (2007) Identification of IGF2 signaling through phosphoinositide-3-kinase regulatory subunit 3 as a growth-promoting axis in glioblastoma. *Proc. Natl. Acad. Sci. USA*, **104**, 3466–3471.

Span,P.N. *et al*. (2003) Matrix metalloproteinase inhibitor reversion-inducing cysteine-rich protein with Kazal motifs: a prognostic marker for good clinical outcome in human breast carcinoma. *Cancer*, **97**, 2710–2715.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440–9445.

Stranger,B.E. *et al*. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Stransky,N. *et al*. (2006) Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet*., **38**, 1386–1396.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis:aknowledge-based approach for interpreting genome-wide expression profiles.. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Sunahara,M. and Nakagawara,A. (2000) Turcot syndrome. *Nippon Rinsho*, **58**, 1484–1489.

Sweet-Cordero,A. *et al*. (2006) Comparison of gene expression and DNA copy number changes in a murine model of lung cancer. *Genes Chromosomes Cancer*, **45**, 338–348.

Tanay,A. *et al*. (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genome-wide data. *Proc. Natl. Acad. Sci. USA*, **101**, 2981–2986.

Tonon,G. *et al*. (2005) High-resolution genomic profiles of human lung cancer. *Proc. Natl. Acad. Sci. USA*, **102**, 9625–9630.

Turashvili,G. *et al*. (2007) Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer*, **7**, 55.

Vlodavsky,E. and Soustiel,J.F. (2007) Immunohistochemical expression of peripheral benzodiazepine receptors in human astrocytomas and its correlation with grade of malignancy, proliferation, apoptosis and survival. *J. Neurooncol*., **81**, 1–7.

Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med*., **10**, 789–799.

Wolf,M. *et al*. (2004) Highresolution analysis of gene copy number alterations in human prostate cancer using CGH on cDNA microarrays: impact of copy number on gene expression. *Neoplasia*, **6**, 240–247.

Wu,M. *et al*. (2006) LRRC4, a putative tumor suppressor gene, requires a functional leucine-rich repeat cassette domain to inhibit proliferation of glioma cells in vitro by modulating the extracellular signal-regulated kinase/protein kinase B/nuclear factorkappaB pathway. *Mol. Biol. Cell*, **17**, 3534–3542.

Yao,J. *et al*. (2006) Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res*., **66**, 4065–4078.