Hum Genet (2004) 114 : 464–467
DOI 10.1007/s00439-004-1087-9

ORIGINAL INVESTIGATION

**Xiaofeng Zhou · Steven W. Cole · Shen Hu**
**David T. W. Wong**

# Detection of DNA copy number abnormality by microarray expression analysis

Received: 6 November 2003 / Accepted: 12 January 2004 / Published online: 11 February 2004


**Abstract** Gene copy-number abnormalities (CNAs) are characteristic of solid tumors and are found in association with developmental abnormalities and/or mental retardation. The ultimate impact of CNAs is exerted by the altered expression of encoded genes. We have utilized high-density oligonucleotide arrays from Affymetrix to identify DNA CNAs via their impact on mRNA expression levels. In these studies, we have used three different trisomic cell lines (trisomy 9, trisomy 18, trisomy 21) as models of CNAs and have compared mRNA expression in those trisomic cells with that observed in diploid cell lines of matched tissue origin. Our data clearly show that genes from CNA chromosome regions are substantially over-represented ($P<0.000001$ by chi-square analysis) in the differentially expressed subset from comparisons of all three trisomic cell lines with normal matching cells. In addition, we have been able to detect the origin of the duplication by a statistical scan for over-expressed genes. These data show that microarray detection of differential mRNA expression can be used to identify significant DNA CNAs.

X. Zhou · S. Hu · D. T. W. Wong
Laboratory of Head and Neck Cancer Research,
Dental Research Institute, School of Dentistry,
University of California at Los Angeles, Los Angeles, Calif., USA

S. W. Cole
Division of Hematology-Oncology, Department of Medicine,
David Geffen School of Medicine,
University of California at Los Angeles, Los Angeles, Calif., USA

D. T. W. Wong
Division of Oral Biology & Medicine, School of Dentistry,
University of California at Los Angeles, Los Angeles, Calif., USA

D. T. W. Wong
Jonsson Comprehensive Cancer Center,
University of California at Los Angeles, Los Angeles, Calif., USA

D. T. W. Wong
Molecular Biology Institute,
University of California at Los Angeles, Los Angeles, Calif., USA

D. T. W. Wong (✉)
UCLA School of Dentistry,
PO Box 951668, Los Angeles, CA 90095-1668, USA
Tel.: +1-310-2063048, Fax: +1-310-8250921,
e-mail: dtww@ucla.edu

## Introduction

DNA copy-number abnormalities (CNAs; amplifications and deletions) are characteristic of solid tumors (Schwab 1999**;** Popescu and Zimonjic 1997) and are found in association with developmental abnormalities and/or mental retardation (Capone 2001). Various techniques have been developed for detecting CNAs, including comparative genomic hybridization (CGH) and loss of heterozygosity or allelic imbalance (Albertson and Pinkel 2003; Kashiwagi and Uchida 2000; Forozan et al. 1997). The ultimate impact of CNAs is exerted by the altered expression of encoded genes. We have utilized the Affymetrix high-density oligonucleotide array (HG-U133A) to identify DNA CNAs via their impact on mRNA expression levels. In these studies, we have used three different trisomic cell lines (trisomy 9, trisomy 18, trisomy 21) as models of CNAs and have compared mRNA expression in those trisomic cells with that observed in diploid cell lines of matched tissue origin.

## Materials and methods

The trisomic cells, trisomy 9 (GM03226, containing an additional truncated copy of chromosome 9, 47, XY, +del(9)(9pter→q11)), trisomy 18 (GM00734, containing an additional copy of chromosome 18, 47, XX, +18), and trisomy 21 (GM02067, containing an additional copy of chromosome 21, 47, XY, +21), and their respective age, gender, and cell-type matched normal control cells GM00302, GM04552, and GM05386 were obtained from Coriell Cell Repositories/NIGMS (http://locus.umdnj.edu/nigms/). Cells were grown under standard culture conditions (MEM Eagle-Earle BSS, 2× essential and non-essential amino acids and vitamins, with 2 mM L-glutamine), and total RNA was isolated by using a Qiagen RNeasy Kit. Labeled cRNA was synthesized and hybridized to Affymetrix U133A GeneChip high-density oligonucleotide arrays according to the standard Affymetrix protocol. Paired comparison analyses were performed for trisomy cells and their re-

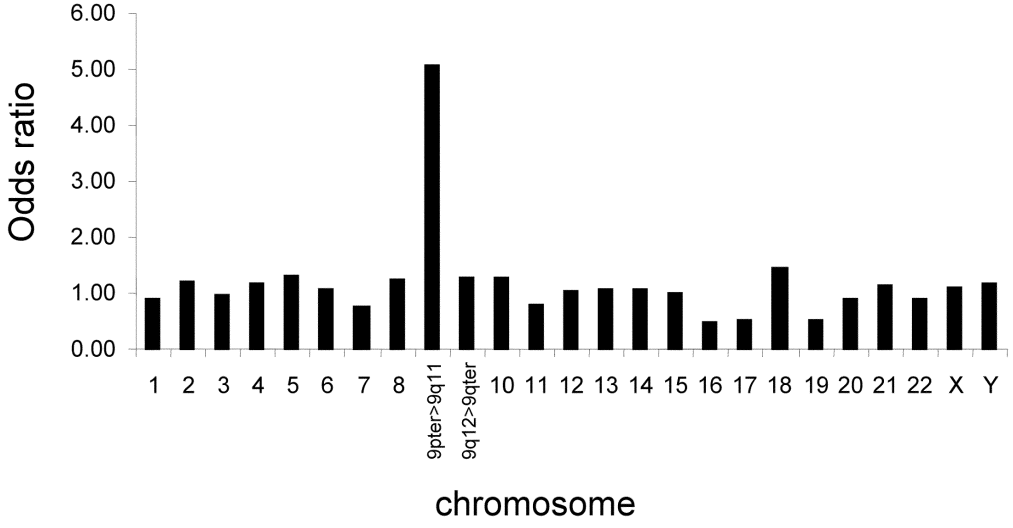


**Fig. 1** Chromosomal distribution of over-expressed transcripts for trisomy 9 cells. Concentrations of mRNA for 19,826 human transcripts were assayed by Affymetrix U133A high-density oligonucleotide array in GM03226 cells, which contained an additional truncated copy of chromosome 9 (9pter→q11). The mRNA levels in trisomy 9 cells were compared with those of normal cells from gender-, age-, ethnicity-matched donors. By using Affymetrix Microarray Suite 5.0, increased transcription was declared when the change *P*-value (critical *P*-value) was less than 0.0045. Differentially expressed transcripts were then mapped to their chromosomal location, and the relative prevalence of over-expressed transcripts from each chromosome was plotted as the odds ratio relative to the value that would be expected based on the distribution of chromosomal locations across all transcripts assayed

spective controls by using the Statistical Expression Algorithm of the Affymetrix Microarray Suite 5.0. The default settings were used for declaring over-expressed transcripts (change *P*-value <0.0045). The extent to which transcripts from a given chromosome were over-represented among the set of over-expressed genes was indicated by an odds ratio relative to the basal representation of genes from that transcript in the entire Affymetrix U133A sampling frame. Statistical significance of excess representation was evaluated by using the chi-squared test, which produced a global test statistic indicating departure from expected incidence across all chromosomes ($\chi^2$ with 23 df; Fleiss 1981).

To identify the specific chromosome showing significant CNAs, the global test statistic was decomposed into constituent values for each chromosome ($\chi^2$ with 1 df expressed as a percentage of the total $\chi^2$ value with 23 df). The trisomy 9 cell line contained a duplication of only a portion of chromosome 9 (9pter→q11), and we therefore analyzed the regions 9pter→q11 and 9q11→qter separately.

To determine whether we could identify the origin of a significant CNA from over-expression data, we fitted a simple statistical model to the data from chromosome 9 that included a parameter θ estimating the chromosomal location at which the incidence of over-expression rose from the diploid base rate of β to an elevated rate of δβ in the trisomic region. Formally, this statistical model expresses the probability of over-expression for each of *N* assayed transcripts as a function of the chromosomal location of its transcription start site and the origin of trisomy (Pr[gene *n* is overexpressed]=$\delta_{\theta n}\beta$, with *n*=1, 2,...*N* indexing the ordinal position of transcription start sites beginning with 9pter and ending at 9qter, θ indicating chromosomal location at which trisomy begins, and the subscripts θ*n* indicating the dependence of δ on both the location of the transcription start side of gene *n* and the origin of trisomy). Transcripts originating outside of the trisomic region ($n<\theta$) are over-expressed at a base rate β (i.e., $\delta_{\theta n}=1$), and transcripts originating within the trisomic region ($n>\theta$) are over-expressed at an altered rate $\delta_{\theta n}\beta$ ($\delta_{\theta n}\neq 1$). The model was fitted by maximum likelihood (binomial probability density), and the sampling distribution of θ was estimated by nonparametric bootstrapping (2000 resamplings of the *N*=733 ordered transcripts from chromosome 9 present in the Affymetrix U133A array; Efron and Tibshirani 1993).

## Results and discussion

To evaluate the effect of DNA CNAs on RNA expression in trisomy cells, differentially expressed transcripts were mapped to their respective chromosomal locations. For the trisomy 9 cell, genes located in the region in which there was an additional copy (9pter→q11) had a significantly higher prevalence in the over-expressed set than would be expected based on the prevalence of transcripts from this region in the entire set of transcripts assayed by the Affymetrix U133A array (Fig. 1). Similar results (all $P<0.00001$) were observed with trisomy 18 (odds ratio= 7.28) and 21 (odds ratio=2.20) cell lines (Table 1). These data show that it is feasible to use microarray detection of differential mRNA expression to identify significant DNA CNAs.

To determine whether we could identify the origin of a significant CNA by using over-expression data, we fitted a simple statistical model to the data from chromosome 9 of trisomy 9 cells as described above. Analysis showed that differential gene expression increased from a base rate of 6.6% to 30.3% in the vicinity of locus 202 of the 733 ordered loci on chromosome 9 (95% confidence interval=[175, 227], $\chi^2(1)=72.4$, $P<0.000001$). This corresponds to a location 38.4 Mb from chr9pter (Fig. 2). This estimate of the origin of trisomy derived for over-expression analysis agrees closely with the 9pter→q11duplication previously documented by cytogenetic methods and which would correspond to a break-point at ordered locus 208 (39.1 Mb from 9pter). These findings suggest that changes in over-expression rates can be used to localize the origin of CNAs.

**Table 1** Copy-number abnormality detected by microarray expression analysis on trisomic cells

| Cell | CNA | Baseline distribution[a] | Over-expression distribution[b] | Odds ratio[c] | Chi square[d] | *P*-value[e] | Chi square [f] fraction |
|---|---|---|---|---|---|---|---|
| GM03226 | +9pter→q11 | 0.01014 | 0.04941 | 5.07 | 181.57 | <0.000001 | 0.711 |
| GM00734 | +18 | 0.01504 | 0.1000 | 7.28 | 393.87 | <0.000001 | 0.888 |
| GM02067 | +21 | 0.01166 | 0.02529 | 2.20 | 40.39 | <0.000001 | 0.462 |

[a]The fraction of all assayed transcripts localized to the chromosomal region in *column 2*
[b]The fraction of over-expressed transcripts localized to the chromosomal region in *column 2*
[c]Odds ratio: odds of overexpression for transcripts from the chromosomal region in *column 2* relative to the odds of all transcripts originating from that region
[d]Chi square: difference between observed incidence of over-expression and incidence expected based on homogenous over-expression rates across all chromosomes
[e]*P*-value: probability of chi-square test statistic greater or equal to that observed in *column 6* by chance alone under the assumption of homogenous over-expression across chromosomes
[f]Chi square fraction: fraction of the genome-wide departure from expectation ($\chi^2$ df=23 in *column 6*) that can be attributed to the specific CNA listed in *column 2*

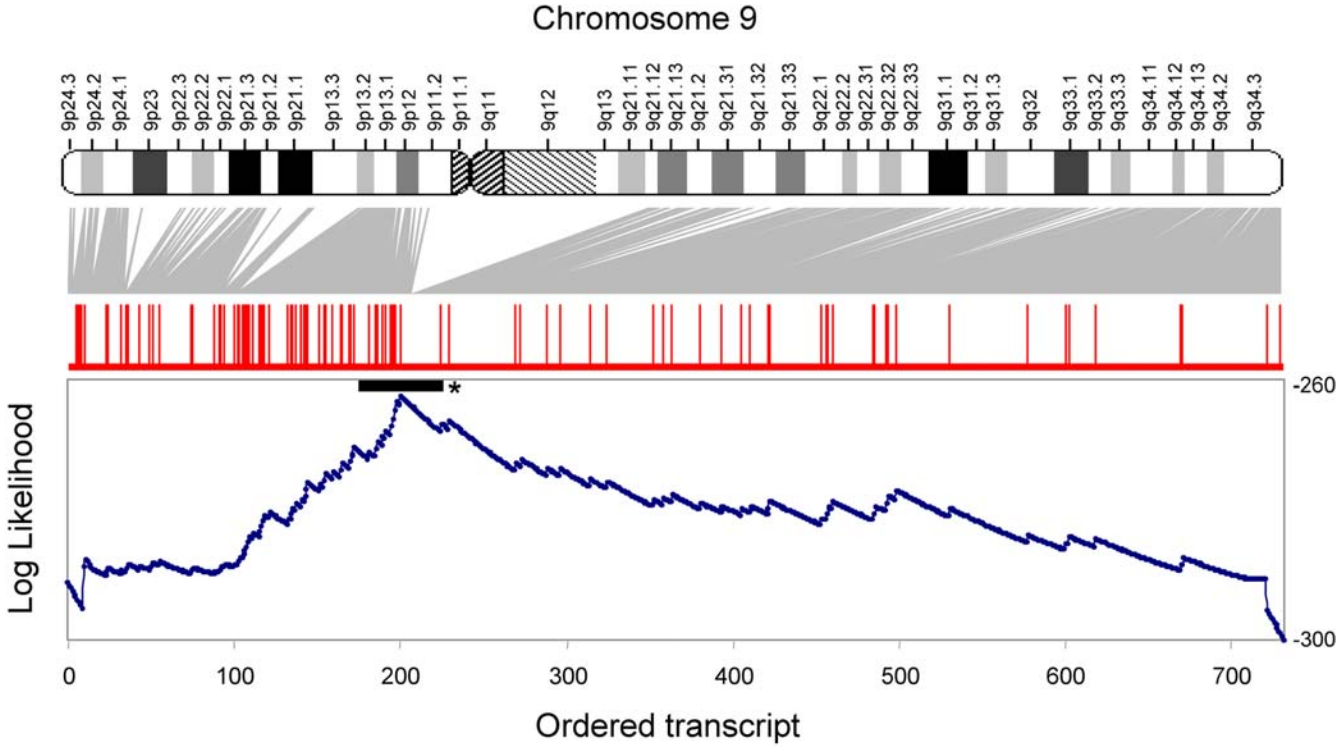


**Fig. 2** Identification of CNA origin by differential expression. Over-expressed transcripts in trisomy 9 cells were identified as described in Fig. 1 and ordered according to sequence on chromosome 9. The *bar plot* displays results of analysis ordered according to transcription start site, with *vertical bars* indicating significant over-expression. A single break-point model allowing differential density of over-expression was fitted by maximum likelihood. The log likelihood associated with breakpoints at each ordinal position on chromosome 9 was plotted *below* with the maximum likelihood value serving as the estimated origin of CNA. *Black bar* (*asterisk*) gives the 95% confidence interval for the origin of CNA over 2000 bootstrap resamplings. *Gray lines* (*middle*) map ordinal positions of each assayed transcript to the relevant chromosomal location. The 95% confidence interval captures the origin of trisomy as defined by cytogenetic analyses at ordered transcript 208, at a distance of 38.4 Mb from 9pter

In summary, our data clearly show that genes from CNA chromosome regions are substantially over-represented ($P$<0.000001 by chi-square) in the differentially expressed subset for all three trisomic cell lines. Furthermore, breakpoint analysis of trisomy 9 cells demonstrates that high resolution mapping of CNA origin can be derived from differential mRNA expression analysis. Expression-based detection of DNA CNAs may thus provide a complementary approach to the standard genomic and cytogenetic methods, such as CGH and fluorescence in situ hybridization, which directly measure the changes in genomic DNA content. The novelty of the method described here is that it uses expression data to infer the causative genomic changes. This is most useful when DNA-based data is not available, e.g., in attempts to extract genomic information from archived expressional data of clinical tumor samples. The resolution of this method is variable and depends primary on the densities of genes in particular chromosome regions but also on the probe set in the particular array platform. In principle, this technology should be equally suitable for detecting deletions. However, more

experimental and statistical studies are needed to define its sensitivity in detecting deletions.

**Acknowledgments** This work was supported in part by NIH PHS grants R21 CA94216 (to D. Wong), R21 AI49135, and R01 AI52737 (to S. Cole), and NIH training grant DE07296-07 and a CRFA fellowship (to X. Zhou). The Affymetrix U133A array hybridization and scanning were performed in the UCLA DNA microarray facility.

## References

Albertson DG, Pinkel D (2003) Genomic microarrays in human genetic disease and cancer. Hum Mol Genet (in press)

Capone GT (2001) Down syndrome: advances in molecular biology and the neurosciences. J Dev Behav Pediatr 22:40–59

Efron B Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

Fleiss JL (1981) Statistical methods for rates and proportions. Wiley, New York

Forozan F, Karhu R, Kononen J, Kallioniemi A, Kallioniemi OP (1997) Genome screening by comparative genomic hybridization. Trends Genet 13:405–409

Kashiwagi H, Uchida K (2000) Genome-wide profiling of gene amplification and deletion in cancer. Hum Cell 13:135–141

Popescu NC, Zimonjic DB (1997) Molecular cytogenetic characterization of cancer cell alterations. Cancer Genet Cytogenet 93:10–21

Schwab M (1999) Oncogene amplification in solid tumors. Semin Cancer Biol 9:319–325