SHORT REPORT

Identification of discrete chromosomal deletion by binary recursive partitioning of microarray differential expression data

X Zhou, S W Cole, N P Rao, Z Cheng, Y Li, J McBride, D T W Wong

J Med Genet 2005;**42**:416–419. doi: 10.1136/jmg.2004.025353

DNA copy number abnormalities (CNA) are characteristic of tumours, and are also found in association with congenital anomalies and mental retardation. The ultimate impact of copy number abnormalities is manifested by the altered expression of the encoded genes. We previously developed a statistical method for the detection of simple chromosomal amplification using microarray expression data. In this study, we significantly advanced those analytical techniques to allow detection of localised chromosomal deletions based on differential gene expression data. Using three cell lines with known chromosomal deletions as model system, mRNA expression in those cells was compared with that observed in diploid cell lines of matched tissue origin. Results show that genes from deleted chromosomal regions are substantially over-represented ($p<0.000001$ by χ^2) among genes identified as underexpressed in deletion cell lines relative to normal matching cells. Using a likelihood based statistical model, we were able to identify the breakpoint of the chromosomal deletion and match with the karyotype data in each cell line. In one such cell line, our analyses refined a previously identified 10p chromosomal deletion region. The deletion region was mapped to between 10p14 and 10p12, which was further confirmed by subtelomeric fluorescence in situ hybridisation. These data show that microarray differential expression data can be used to detect and map the boundaries of submicroscopic chromosomal deletions.

DNA copy number abnormalities (amplifications and deletions) are characteristic of tumours,[1 2] and are found in association with developmental abnormalities and/or mental retardation.[3] Several techniques have been developed for detecting CNA, including comparative genomic hybridisation (CGH), fluorescence in situ hybridisation (FISH), and loss of heterozygosity (LOH).[4–7] Recently, several groups have observed that chromosomal alterations can lead to regional gene expression biases in human tumours and tumour derived cell lines.[8–10] These studies suggested that a fraction of gene expression values (15–25%) are regulated in concordance with chromosomal DNA content. Statistical methods developed by our group and others have shown promising results for detecting CNA based on differential gene expression.[10 11] Crawley and colleagues used measures of gene expression bias to identify entire chromosomal arms showing aberrant expression.[10] We recently found that a maximum likelihood statistical model could be used to localise the origin of chromosomal amplification within a chromosome that had already been identified as showing global expression abnormalities.[11] In the present study, we adapt that statistical approach to detect the origin of chromosomal deletion based on gene expression data. Using three cell lines with known chromosomal deletions as model system, we compared mRNA expression in those cells with that observed in diploid cell lines of matched tissue origin.

The deletion cells del(7)(GM03240,46,XY,del(7)(q34)), del(9)(GM00870,46,XX,del(9)(p21)), del(10)(GM03047, 46,XY, and del(10)(p11.2)), generated from patients with congenital anomalies and mental retardation, and 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 (minimum essential medium Eagle-Earle BSS, 2× essential and non-essential amino acid and vitamin, with 2 mmol/l L-glutamine). Total RNA was isolated using a Qiagen RNeasy kit, and cRNA was synthesised, labelled, and fragmented, then hybridised to Affymetrix U133+ 2.0 GeneChip high density oligonucleotide arrays according to the manufacturer's standard protocol. Paired comparison analyses were performed for deletion cells and their respective controls using the statistical expression algorithm of the Affymetrix Microarray Suite 5.0 software. Default settings were used to identify underexpressed transcripts (downregulated at $p<0.002$). The extent to which transcripts from a given chromosome were over-represented among the set of underexpressed genes was indicated by an odds ratio relative to the basal representation of genes from that chromosome in the entire Affymetrix sampling frame. Statistical significance of excess representation was evaluated using the χ^2 test, which produced a global test statistic indicating departure from expected incidence across all chromosomes ($\chi^2$with 23 df).[12]

To identify the specific chromosome showing significant CNA, the global test statistic was separated into constituent values for each chromosome (χ^2 with 1 df expressed as a % of the total χ^2 value with 23 df). For deletion cell lines, the diploid and the haploid (deleted) regions were analysed separately. The differentially expressed transcripts were mapped to their respective chromosomal locations. Genes located in the region where there was a deletion (single copy of the chromosomal region), were found to have a significantly higher prevalence in the underexpressed set than would be expected based on the prevalence of transcripts from that region in the entire set of transcripts assayed by the Affymetrix array (all $p<0.00001$, with odds ratios of 3.13, 2.10, and 3.54 for del(7), del(9), del(10) cells, respectively) (table 1). These data show that it is feasible to use microarray detection of differential mRNA expression to identify DNA copy number abnormalities.

Abbreviations: CGH, comparative genomic hybridisation; CNA, copy number abnormality; FISH, fluorescence in situ hybridisation; LOH, loss of heterozygosity

Table 1 Chromosomal deletion detected by microarray expression analysis

Cell	Karyotype*	Baseline distribution†	Underexpression distribution‡	Odds ratio§	χ^2¶	p**	χ^2 fraction††
GM03240	46, XY, del(7)(q34)	0.00517	0.28500	3.13	80.82	<0.000001	0.717
GM00870	46, XX, del(9)(p21)	0.00967	0.17647	2.10	26.15	<0.000001	0.207
GM03047	46, XY, del(10)(p11.2)	0.00742	0.24042	3.54	121.69	<0.000001	0.730

*Karyotypes were provided by Coriell Cell Repositories/NIGMS. †The fraction of all assayed transcripts localised to the chromosomal deletion region in column 2; ‡fraction of underexpressed transcripts localised to the chromosomal deletion region in column 2. §Odds ratio: odds of underexpression for transcripts from the chromosomal deletion region in column 2 relative to the odds of all transcripts originating from that region.; ¶χ^2: difference between observed incidence of underexpression and incidence expected based on homogenous underexpression rates across all chromosomes.; **p value: probability of χ^2 test statistic ⩾ that observed in column 6 by chance alone under the assumption of homogenous underexpression across chromosomes; ††χ^2 fraction: fraction of the genomewide departure from expectation (χ^2 (23 df) in column 6) that can be attributed to the specific DNA listed in column 2.

To determine whether we could identify the boundaries of chromosomal deletion from underexpression data, we fitted a simple breakpoint statistical model to the data from chromosomes 7, 9, and 10 for the del(7), del(9) and del(10) cells respectively. A parameter θ was employed to indicate the chromosomal location at which the incidence of underexpression increases from the diploid base rate of β to an elevated rate of δβ in the deletion region. This statistical model expresses the probability of underexpression for each of N assayed transcripts as a function of the chromosomal location of its transcription start site and the origin of haploid DNA. (Pr(gene n is underexpressed) = $\delta\theta_n\beta$, with n = 1, 2, … N indexing the ordinal position of transcription start sites beginning with pter and ending at qter, θ indicating chromosomal location at which deletion begins, and the subscripts θn indicating the dependence of δ on both the location of the transcription start site and the origin of deletion of gene n). Transcripts originating outside of the deletion region ($n<\theta$) are underexpressed at a base rate β (that is, $\delta_{\theta n} = 1$), and transcripts originating within the deletion region ($n>\theta$) are underexpressed 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 non-parametric bootstrapping (2000 resamplings of the ordered transcripts from chromosome 7, 9, and 10 present in the Affymetrix array).[13] Analysis showed

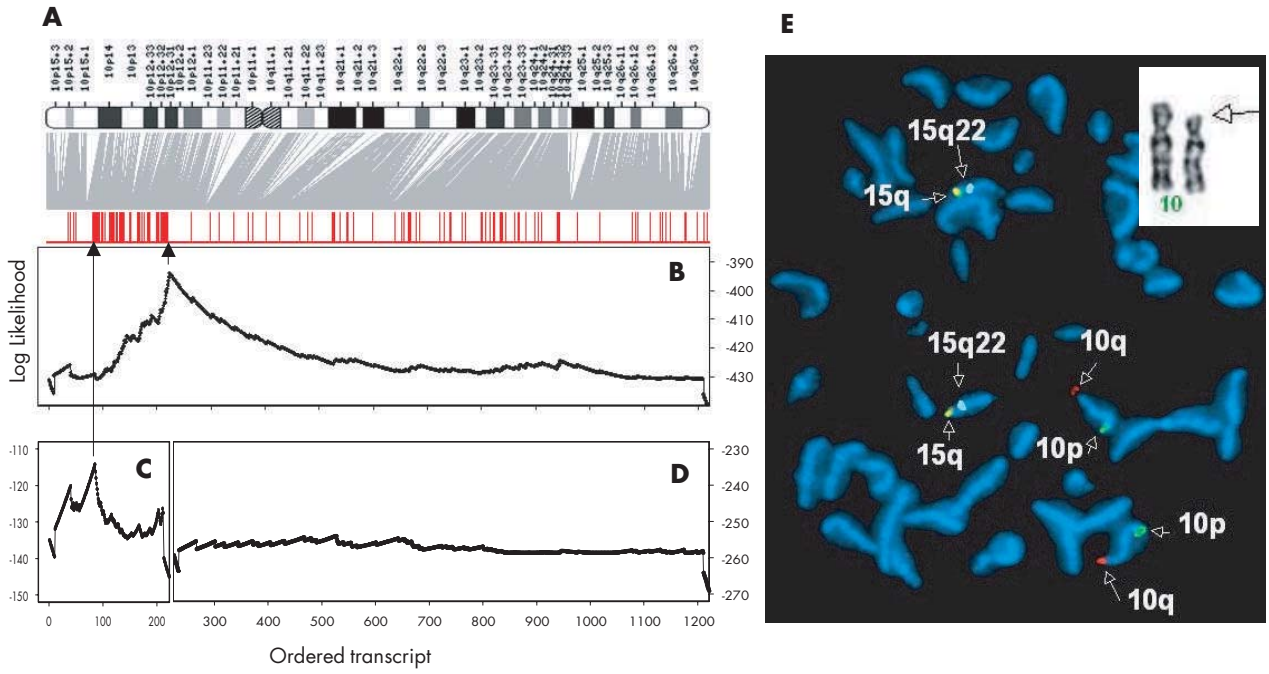

Figure 1 Mapping the boundaries of chromosomal deletion by differential expression. (A) Underexpressed transcripts in del(10) cells were identified using Affymetrix Microarray Suite 5.0, with decreased transcription declared when change in p value was <0.002. The transcripts were ordered according to sequence on chromosome 10, with red bars indicating the transcription start site of genes identified as significantly underexpressed in del(10) cells relative to a tissue matched normal control cell. (B) As detailed in the text, a single breakpoint model allowing differential density of underexpression was fitted by maximum likelihood. The log likelihood associated with breakpoints at each ordinal position on chromosome 10 was plotted (black line) with the maximum likelihood value serving as the estimated origin of CNA. Grey lines map the ordinal positions of each assayed transcript to its chromosomal location. Significant change in the prevalence of underexpressed transcripts was identified at ordered transcript 224, 28.1 Mb from 10pter, agreeing with the previously defined origin of deletion by cytogenetic analyses. (C) To determine whether deletion extended to the p terminus, transcripts 1–223 were re-scanned, and a second significant change in the prevalence of underexpressed transcripts was identified at ordered transcript 85, 12.2 Mb from 10pter. (D) No significant change in the prevalence of underexpressed transcripts was identified in the region ranging from ordered transcript 224 to 10qter. Together with the results from (B), these data indicate a single partial deletion of chromosome 10p spanning the region 10p14 to 10p12. (E) Subtelomere FISH verified results from the maximum likelihood expression based analysis by confirming that the 10p deletion was interstitial with the intact subtelomere regions. Probes used are: 10ptel006 (10pter probe, green); 10qtel24 (10qter probe, red); PML (15q22 probe, aqua) and AFMA224XHI (15qter probe, yellow). Two normal signals for both 10p and 10q subtelomeres were clearly identified. Inset: G banded chromosome 10 of del(10) cell showing the deletion of p arm of the chromosome 10.

Table 2 Identification of CNA origin by microarray differential expression analysis

Cells	Deletion region*	Breakpoint		Predicted breakpoint		Pre-point prevalence**	Post-point prevalence††	Odds ratio‡‡
		Locus†	Mb‡	Locus§	Mb¶			
GM03240	7q35→qter	1342/1493	143.0/158.2	1354 (1327 to 1406)	147.8 (142.5 to 150.2)	8.6%	35.3%	5.76
GM00870	9pter→p13	187/1185	33.2/136.3	112 (40 to 243)	19.4 (5.5 to 35.1)	26.8%	9.7%	3.41
GM03047	10pter→p12	232/1221	29.1/134.8	224 (197 to 252)	28.1 (26.6 to 32.6)	29.5%	7.2%	5.37

*Deletion regions are derived from karyotypes data that provided by Coriell Cell Repositories/NIGMS; †the expected breakpoint locus determined based on previous cytogenetic evaluation/total number of assessed loci for the specific chromosome; ‡the expected base pair location of the breakpoint/total base pairs on the specific chromosome; §point estimate (and bootstrap 95% confidence interval) of the ordered locus at which chromosomal deletion begins; ¶point estimate (and bootstrap 95% confidence interval) of the chromosomal location (in megabases from pter) at which chromosomal deletion begins; **prevalence of microarray declared decreases in gene expression for genes with transcription start sites p terminal to the estimated breakpoint; ††prevalence of microarray declared decreases in gene expression for genes with transcription start sites q terminal to the estimated breakpoint; ‡‡odds of declared underexpression in the post-breakpoint region relative to that in the pre-breakpoint region.

that, for del(10) cells, underexpressed genes increased from a base rate of 7.2% to 29.5% in the vicinity of locus 224 of the 1221 ordered loci on chromosome 10 (95% confidence interval 197 to 252, likelihood ratio χ^2 91.1, $p<0.000001$). This corresponds to a location 28.1 Mb from chr10pter (fig 1B). This estimate of the breakpoint of deletion from underexpression analysis agrees closely with the previously documented breakpoint (10p12) by cytogenetic methods, which would correspond to a breakpoint at ordered locus 241 (30.7 Mb from 10pter). Similar results were observed for the del(7) and del(9) cell lines, in which the identified breakpoints also agreed closely with karyotypes (table 2). These findings suggest that changes in underexpression rates can be used to pinpoint the boundaries of chromosomal deletions.

To determine whether the analysed chromosomes might contain novel abnormalities not previously detected, we applied the same maximum likelihood breakpoint analysis to each of the subregions defined by the results of the initial breakpoint analysis. For example, the initial analysis of chromosome 7 identified a breakpoint at ordered locus 1354 of the total 1493 chromosome 7 transcripts present in the Affymetrix sampling frame (table 2). In subsequent analyses, we scanned one fragment spanning ordered loci 1–1353 and another fragment spanning loci 1354–1493. Analyses of fragment data from chromosome 7 of del(7) cells and chromosome 9 of del(9) cells failed to suggest any further non-homogeneity in differential expression rates. However, analysis of the pter fragment of chromosome 10 from del(10) cells revealed a significant decrease in the incidence of downregulated genes in the vicinity of ordered locus 85 (out of the 223 total loci spanning 10pter-10p12) (fig 1C). The change in incidence was highly significant ($\chi^2(1) = 36.16$, $p<0.0001$), with the prevalence of downregulated genes increasing from 7.1% in the telomere–proximal region to 43.5% in the centromere–proximal region (odds ratio 10.13). These results suggested that del(10) cells retain normal diploid gene expression in the region 10pter-10p14, and that chromosomal deletion may be limited to the region 10p14-10p12. This hypothesis contradicts with the karyotype provided by the cell vendor, which indicates a complete deletion of 10pter-10p12. To resolve the contradiction, we carried out FISH as described previously, with subtelomere probes specific to 10p and 10q.[14] As shown in fig 1E, del(10) cells clearly maintain two subtelomeres on chromosome 10 (both pter and qter). Probes for chromosome 15 were used as internal control. Thus the statistical analysis of differential expression data can identify and map the boundaries of discrete chromosomal deletions.

In summary, our data clearly show that genes from deleted chromosome regions are substantially over-represented (χ^2, $p<0.000001$) in the underexpressed subset for all three deletion cell lines. Furthermore, recursive application of a statistical breakpoint analysis can generate a high resolution mapping of the bounds of localised chromosomal deletions not previously recognised. This successive decomposition of heterogeneity in differential gene expression is reminiscent of the binary recursive partitioning strategies employed in non-parametric regression[15] and could conceivably be applied to mapping other types of CNA (such as localised amplification). Expression based detection of DNA copy number abnormalities may thus provide a complementary approach to well established genomic and cytogenetic methods such as CGH and FISH, which directly measure changes in genomic DNA content. The present method is novel in using indirect functional data (transcription) to infer the underlying causative genomic changes. This approach is likely to be most useful when DNA based data are not available (for example, attempts to extract genomic information from archived expression data from clinical tumour samples), or when analysts seek to generate hypotheses about structural bases for differential gene expression in microarray data. The resolution of this method depends inherently on the density of genes in different chromosomal locations, and the specific set of genes represented on a particular microarray platform. Given the variability in these values, it is difficult to specify the resolution of the present technique in DNA base terms. However, given the magnitude of expression changes observed here, the present technique should be able to localise CNAs to contiguous regions spanning as few as 40 genes. These data show that statistical analysis of differential expression data can accurately identify the origin of CNAs in well defined model systems (see also Zhou *et al*[11]), but further experimental and statistical studies will be required to evaluate the feasibility of this approach for identifying CNAs in clinical tumour samples. However, the present results suggest that expression based analysis of chromosomal abnormalities could provide a novel means for defining pathogenic structural abnormalities in cases where DNA data are not directly available.

ACKNOWLEDGEMENTS

This work was supported in part by NIH PHS grants R21 CA97771 and R01 DE015970–01 (to D Wong), R21 AI49135 and R01 AI52737 (to S. Cole), T32 DE07296–07, K22 DE014847-01, and a TRDRP grant 13KT-0028 (to X Zhou). The Affymetrix U133+ 2.0 array hybridisation and scanning were performed in the UCLA DNA microarray facility.

Authors' affiliations

X Zhou, Y Li, J McBride, 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, CA, USA
S W Cole, Division of Hematology-Oncology, Department of Medicine, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA, USA

S W Cole, D T W Wong, Jonsson Comprehensive Cancer Center, University of California at Los Angeles, Los Angeles, CA, USA
S W Cole, D T W Wong, Molecular Biology Institute, University of California at Los Angeles, Los Angeles, CA, USA
N P Rao, Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA, USA
Z Cheng, Department of Human Genetics, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA, USA

Competing interests: none declared

Correspondence to: Dr D Wong, UCLA School of Dentistry, PO Box 951668, Los Angeles, CA 90095–1668, USA; dtww@ucla.edu

Received 15 July 2004
Revised 15 July 2004
Accepted 10 September 2004

REFERENCES

1 **Schwab M**. Oncogene amplification in solid tumors. *Semin Cancer Biol* 1999;**9**:319–25.
2 **Popescu NC**, Zimonjic DB. Molecular cytogenetic characterization of cancer cell alterations. *Cancer Genet Cytogenet* 1997;**93**:10–21.
3 **Capone GT**. Down syndrome: advances in molecular biology and the neurosciences. *J Dev Behav Pediatr* 2001;**22**:40–59.
4 **Albertson DG**, Pinkel D. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* 2003;12 Spec No 2:R145–52.
5 **Natarajan AT**, Boei JJ. Formation of chromosome aberrations: insights from FISH. *Mutat Res* 2003;**544**:299–304.
6 **Kashiwagi H**, Uchida K. Genome-wide profiling of gene amplification and deletion in cancer. *Hum Cell* 2000;**13**:135–41.
7 **Forozan F**, Karhu R, Kononen J, Kallioniemi A, Kallioniemi OP. Genome screening by comparative genomic hybridization. *Trends Genet* 1997;**13**:405–9.
8 **Phillips JL**, Hayward SW. The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. *Cancer Res* 2001;**61**:8143–9.
9 **Virtaneva K**, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle A, Krahe R. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci USA* 2001;**98**:1124–9.
10 **Crawley JJ**, Furge KA. Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. *Genome Biol* 2002;**3**:RESEARCH0075.
11 **Zhou X**, Cole SW, Hu S, Wong DT. Detection of DNA copy number abnormality by microarray expression analysis. *Hum Genet* 2004;**114**:464–7.
12 **Fleiss JL**. *Statistical methods for rates and proportions*. New York: John Wiley and Sons, 1981.
13 **Efron B**, Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman & Hall, 1993.
14 **Pettenati MJ**, Jackle B, Bobby P, Stewart W, Von Kap-Herr C, Mowrey P, Rao PN, May KM. Unexpected retention and concomitant loss of subtelomeric regions in balanced chromosome anomalies by FISH. *Am J Med Genet* 2002;**111**:48–53.
15 **Hastie T**, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference and prediction*. New York: Springer, 2001.