Nucleic Acids Research, 2005, Vol. 33, No. 11 **3455–3464**
doi:10.1093/nar/gki643

SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data

Thomas S. Price[1], Regina Regan[2], Richard Mott[1], Åsa Hedman[1], Ben Honey[2], Rachael J. Daniels[4], Lee Smith[3], Andy Greenfield[3], Ana Tiganescu[1], Veronica Buckle[4], Nicki Ventress[4], Helena Ayyub[4], Anita Salhan[1], Susana Pedraza-Diaz[1], John Broxholme[1], Jiannis Ragoussis[1], Douglas R. Higgs[4], Jonathan Flint[1] and Samantha J. L. Knight[1,2,*]

[1]The Wellcome Trust Centre for Human Genetics, [2]Oxford Genetics Knowledge Park, Roosevelt Drive, Churchill Hospital, Headington, Oxford OX3 7BN, UK, [3]Mammalian Genetics Unit, Medical Research Council, Harwell, Didcot, OX11 0RD, UK and [4]Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DS, UK

Received April 21, 2005; Accepted May 20, 2005

ABSTRACT

Comparative genome hybridization (CGH) to DNA microarrays (array CGH) is a technique capable of detecting deletions and duplications in genomes at high resolution. However, array CGH studies of the human genome noting false negative and false positive results using large insert clones as probes have raised important concerns regarding the suitability of this approach for clinical diagnostic applications. Here, we adapt the Smith–Waterman dynamic-programming algorithm to provide a sensitive and robust analytic approach (SW-ARRAY) for detecting copy-number changes in array CGH data. In a blind series of hybridizations to arrays consisting of the entire tiling path for the terminal 2 Mb of human chromosome 16p, the method identified all monosomies between 267 and 1567 kb with a high degree of statistical significance and accurately located the boundaries of deletions in the range 267–1052 kb. The approach is unique in offering both a nonparametric segmentation procedure and a nonparametric test of significance. It is scalable and well-suited to high resolution whole genome array CGH studies that use array probes derived from large insert clones as well as PCR products and oligonucleotides.

INTRODUCTION

Cytogenetically visible segmental aneusomies have long been recognized as a common cause of human genetic disease, but less readily detectable chromosomal rearrangements can also be clinically important. For example, small genomic rearrangements may account for 14–15% of idiopathic learning disability, a common condition for which most investigations have a very low diagnostic yield (1–3). Methods to screen the genome for DNA copy-number changes are therefore likely to have important clinical applications. One such method, comparative genome hybridization (CGH) to DNA microarrays (array CGH), has attracted much attention, but reliable detection of single-copy gains or losses remains challenging.

Array CGH can reveal single-copy changes (4–6), but it is also clear that such changes can pass undetected, suggesting that the method may not be sufficiently sensitive and/or specific to be useful as a routine diagnostic tool. In a study of telomeric regions of the genome, it has been found that one false positive result would be expected for every patient analysed (7). Indeed, for some arrayed probes, 15% of the analyses would be scored as abnormal. In an array CGH study of

*To whom correspondence should be addressed at The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Churchill Hospital, Headington, Oxford OX3 7BN, UK. Tel: +44 1865 287511; Fax: +44 1865 287501; Email: sknight@well.ox.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors the 1p36 region, single copy deletions were correctly identified in all patients with 1p monosomy, but the analysis failed to identify known copy-number changes using 1p probes that map to the most terminal associated repeat region of 1p36 (8). Furthermore, a whole genome study, using probes spaced every 1 Mb across the genome, reported a 10% false positive rate and 10–15% polymorphisms (1). Thus, although whole genome array CGH has the potential to be extremely effective, there are concerns regarding the suitability of the approach in a clinical diagnostic environment where a reliable assay, providing clear, high quality results of measurable significance is required.

We set out to determine whether a method for analysing array CGH data sets could be developed that would be sufficiently robust and reliable for clinical applications. To do so, we required a series of DNA samples from individuals with previously characterized copy-number changes, the boundaries of which had been mapped accurately. We used a series of DNA samples from patients with monosomies that varied in size from 223 to 1567 kb and which had been accurately mapped onto the terminal 2 Mb of chromosome 16p. Not only has this 2 Mb region been fully sequenced, annotated and well characterized, but it is also fully represented by a tiling path of cosmid and PAC clones that have been accurately placed with respect to the sequence, thereby allowing us to generate array probes (9).

We developed a computational method that makes no distributional assumptions about the data to identify putative copy-number changes and determine their statistical significance. Since analysing probe signals independently has been shown to be error-prone (in terms of false positives and false negatives) and complicated by polymorphisms/benign variants, we adapted the Smith–Waterman algorithm (10) to identify genomic regions with copy-number changes that span multiple probes. This algorithm has been used previously to identify genomic regions with unusual properties (11). This method, which we have termed SW-ARRAY (Smith–Waterman algorithm adapted for Array CGH) represents a new approach towards optimizing array CGH analyses so that copy-number changes can be detected more accurately, thus making array CGH more suitable for a clinical diagnostic environment.

MATERIALS AND METHODS

Subjects

DNA samples from a total of 12 control subjects and 16 test subjects were used for the array CGH studies. All test subjects had well-characterized monosomies that ranged from 223 to 1567 kb and were located within the terminal 2 Mb of chromosome 16 (Table 1).

Generation of array probes

A list of the clones used in these studies is at http://www.well.ox.ac.uk/~sknight/NAR. Both the chromosome 16p tiling clones and the telomere-specific clones (used for normalization calculations) have been reported previously (9,12,13). Within the 16p tiling path, there are only five sequence gaps of <100, 600, 850, 1250 bp and 8 kb (9). DNA was extracted from the cosmid, BAC and PAC clones using a standard alkaline-lysis protocol (14). DNAs representative of these clone inserts were obtained by PCR amplification using an adaptation of the method of Fiegler *et al.* (15) described as follows. For each DNA, two 100 µl PCR reactions were performed (i) containing ~100 ng DNA, 2 mM DOP-2 primer (5′-CCGACTCGAGNNNNNNTAGGAG-3′)(MWG), 1.7× polymerization mix (85 µM each of dCTP, dGTP, dTTP and dATP), 1.7× Buffer (17mM Tris–HCl, pH 8.3 and 85 mM KCl), 4.25 mM $MgCl_2$, 6 U Amplitaq Gold (Perkin Elmer Cetus Inc) and (ii) containing the same reagents except the substitution of DOP-2 primer with DOP-3 primer (5′-CCGACTCGAGNNNNNNTTCTAG-3′)(MWG). The cycling conditions were 1× cycle of 94°C for 3 min, 9× cycles of 94°C for 1 min and 30 s, 30°C for 2 min 30 s and 0.1°C/s ramp to 72°C followed by 29 cycles of 94°C for 1 min, 62°C for 1 min 30 s and 72°C for 2 min, and finally a single cycle of 72°C for 5 min. The products were checked by agarose gel electrophoresis, the DOP2 and DOP3 products for each clone combined and the DNA precipitated in 1/10th volume 3 M sodium acetate and 2 volumes ethanol for >48 h at −70°C. The precipitated DNAs were pelleted, washed in 70% ethanol and resuspended overnight at 4°C in 21 µl dH_2O. Following a further agarose gel electrophoresis check, dimethyl sulfoxide (DMSO) was mixed with each sample to give a final arraying solution containing 50% dH_2O:50% DMSO. The samples were then transferred to 384 well microtitre plates (Amersham Biosciences) prior to arraying.

Table 1. Known map positions of monosomies in test samples

Case ID	Position of monosomy (kb from 16p)	References
1	0–223	(26)
2	33–300	(26)
3	0–300	D.R. Higgs and Dr V. Buckle, personal communication
4	0–700	(26)
5	0–775	(9,27)
6	0–900	(28)
7	0–999	(28)
8	0–1052	(29)
9	0–1100	(30)
10	0–1159	(31)
11	0–1159	(28)
12	0–1183	(28)
13	0–1400	D.R. Higgs and Dr V. Buckle, personal communication
14	0–1567	(31)
15	0–1000	(9)
16	0–1160	D.R. Higgs and Dr V. Buckle, personal communication

Array fabrication and processing

Prepared probes were spotted in quadruplicate on to CMT-GAPs slides (Corning Ltd) using a Gen III Microarray spotter (Amersham Pharmacia Biotech). The spotted microarrays were subjected to 650 × 100 µJ UV irradiation in a Stratalinker (Stratagene) and then baked for 2 hrs at 80°C. Prior to hybridization, the microarrays were immersed in boiling water for 2 min, passed through an ethanol series consisting of 70, 95 and 100% EtOH and dried by centrifugation at 160 *g* in a bench top centrifuge (Beckmann). The slides were then immersed in 3.5× SSC, 0.1% SDS, 1% BSA (Fraction V, Sigma) at 50°C for 45 min, washed by shaking in dH_2O at room temperature (RT) for 2 min followed by dehydration through an ethanol series and drying by centrifugation as above. The processed microarrays were stored in a plastic slide container at RT until ready to hybridize.

Preparation and random primed labelling of target DNAs

Target DNAs consisted of (i) control DNAs, comprising anonymous DNAs from phenotypically normal individuals of known sex, and (ii) test DNAs comprising DNAs from patients of known sex and with known 16p subtelomeric rearrangements. For each CGH experiment, control and test DNAs were digested with SauIIIA, and subsequently purified using the Wizard DNA Clean-Up System (Promega) according to manufacturer's instructions. A 5 μg sample of each purified DNA was differentially labelled with Cy3-dCTP and Cy5-dCTP (NEN) using the BioPrime Labelling Kit (Invitrogen) following The Brown Lab protocol webpage (http://cmgm.stanford.edu/pbrown/protocols/), with the exception that mixed, labelled target DNAs were made up to a final volume of 55 μl containing filtered 3.4× SSC and 0.3% SDS.

Array CGH

Each labelled target mixture was denatured at 97°C for 5 min and immediately transferred to a 37°C hotblock for 55 min to allow repetitive sequences to be blocked. The hybridization mix was then applied to a coverslip (pre-warmed at 37°C) and a pre-warmed array slide lowered on to the mix, thereby sandwiching the mix between the coverslip and the microarray spots. Hybridization was carried out in a Corning Ltd hybridization chamber submerged in a 65°C shaking waterbath for 36–48 h. Following hybridization, the coverslip was removed from the microarray slide by 10 min immersion in a Coplin Jar containing 2× SSC, 0.03% SDS at RT. The microarray slide was then passed through a series of washes on a shaking platform. The wash series was 2× SSC, 0.03% SDS for 5 min at 65°C followed by 6× 15 min washes in 0.2× SSC at RT. The slide was then dried by centrifugation at 50 *g* for 3 min at RT and stored in a light-proof container prior to image acquisition.

Image acquisition

Cy3 and Cy5 fluorescence intensity data were collected by scanning the hybridized microarrays in a PackardBell Biochip (formerly GSI Lumonics, Inc) machine using the application ScanArray (PackardBell Biochip). The data were collected at a 10 μ resolution with the laser set at 95% and the scanning rate at 100%. The Photomultiplier tube gains were typically set at 65 for Cy3 images and 55 for Cy5 images. The images were stored as TIFF files ready for analysis.

Subject data set

The data set comprised the measurements from a total of 27 array CGH experiments, of which six were generated from control versus control hybridizations (12 unrelated subjects) against the 16p arrays. Of the remaining 21 experiments, seven used test subject DNAs selected to represent a range of well-characterized 16p deletions ranging from 223 to 1567 kb in length. These subjects were investigated with prior knowledge of the deletion boundaries, to allow calibration of our analysis. During calibration experiments, the results were compared with those of single probe by probe data analyses obtained using the technique outlined by Veltman *et al*. (7). Following the calibration, 14 blind array CGH experiments were performed with anonymized samples later revealed to be made up of nine new test subjects with well-characterized 16p deletions and five of the test subjects with 16p deletions tested previously. The study of subject samples was approved by the appropriate institutional review board.

Data extraction

The fluorescence intensity of each spot (measured as the median of the pixel intensities), together with the local background (measured as the median of the pixel intensities in the region immediately surrounding the spot), were extracted from the image files using the Quantarray software and exported for further analyses (see below). Spots of poor quality were identified by eye and flagged.

Data preprocessing and normalization

The results from the 16p probes were analysed initially on a probe by probe basis using the method of Veltman *et al*. (7), with the exception that quadruplicate (rather than triplicate) spot data were available for each probe. First, the net intensities of each of the Cy3 and Cy5 channels were calculated as the raw spot intensity minus the local background. Second, control:test intensity ratios were calculated and the data edited by excluding flagged spots, excluding spots with an intensity <600 (based on the intensity values from blank spots) in the control sample and, if necessary, removing the replicate data from probes yielding SD $\geqslant$ 0.2 until SD $\leqslant$ 0.2 for the remaining replicates. The minimum number of spots accepted for analysis was two; if it was not possible to decide which two of the four spots should be included then all four spots were excluded. For the seven subjects tested with prior knowledge of the deletion boundaries, the data were normalized by multiplying all Cy3 intensities by the value needed to give a median ratio of 1:1 for 16p probes known to lie in disomic regions. For the 14 blind control versus test array CGH data sets (generated using anonymized test DNAs for which only the sex was known), the subjects tested were likely to have deletions involving the terminal 2 Mb of chromosome 16p and therefore the use of 16p probes for normalization was avoided. Instead, the data were normalized so that the median ratio value for arrayed telomere-specific probes (representing disomic regions of the genome) was 1:1. Control versus control array CGH data from arrays that included both probe sets showed that the multiplication factors required for normalization were consistent regardless of whether 16p probes or telomere probes were used for the calculations. Following normalization, the mean test:control intensity ratios were calculated for each probe by averaging the ratios of accepted spots.

Analysis of array CGH data

The data from the initial seven array CGH experiments, using samples from patients with a range of single copy deletions of chromosome 16p, were analysed in two ways, described in detail below. First, the global thresholding approach of Veltman *et al*. (7) was used. Second, the Smith–Waterman algorithm was applied, trying various formulae for the threshold parameter in order to calibrate this parameter for the analyses to follow. Subsequently, we applied our SW-ARRAY analysis using 14 'blind' test versus control array CGH data sets newly generated using anonymous test DNAs for which only the sex was known. The identities of the patients' chromosomal rearrangements were only revealed after the analyses had been performed.

Probe by probe global thresholding analysis. Data were first analysed one probe at a time, using the global thresholding method (7) to determine whether or not individual probes could detect a region of copy-number change in the seven patients with known regions of monosomy. Using the methodology of Veltman *et al*. (7), normalized data from the six control versus control hybridizations were used to calculate the overall mean of the fluorescence ratios and the SD from this mean for each arrayed 16p probe (in this way, the variation for each probe between different control hybridizations is considered). To correct for the intrinsic variability between probes, the mean intensity ratio yielded by each probe in the control versus test hybridizations was divided by the mean intensity ratio of that particular clone as calculated from the six control versus control hybridizations. This correction was also applied to the mean intensity ratios of each of the clones from the control versus control hybridizations, resulting in a ratio value of 1.0 for all clones. In this way, the corrected ratio obtained for every probe of the control versus patient hybridizations could be easily plotted and compared on the same plot with the corrected control versus control ratio (always a value of 1.0) together with the calculated SD for each clone ratio. As previously described, lower and upper threshold values of 0.8 and 1.2 were used as evidence of deletion and polysomy, respectively (7).

Multiple probe SW-ARRAY (Smith–Waterman algorithm adapted for Array CGH) analysis. The array CGH data were first corrected for the intrinsic variability between probes as described above. Next, isolated outlying probes were removed from the analysis if the logarithm of their intensity ratio fell by >2.5 rescaled MAD (= MAD/0.6745, the rescaled for normal consistency; a robust measure of SD) from the median for the other probes on the array.

The data were then analysed using an adaptation of the Smith–Waterman algorithm, a technique originally applied in bioinformatics for the local alignment of DNA and protein sequences (10), and for the identification of sequence segments with unusual properties (11). The justification for this method is that microarray CGH intensity log ratios, considered sequentially along the genome, represent a one-dimensional series of continuously distributed scores. Contiguous sequences of predominantly high values in this series may indicate polysomic regions. Conversely, sequences of predominantly low values may indicate deletions, and can be found in the same way by changing the sign of the data. The method proceeds as follows. First, a threshold value t_0 is subtracted from the log ratios, ensuring that the mean of the adjusted scores is negative. The score of a segment of consecutive probes is the sum of the corresponding adjusted log ratios. Next, high-scoring 'islands' are identified using the Smith–Waterman algorithm. A locally high-scoring segment or island is defined to be a positive-scoring segment whose score cannot be increased by shrinking or expanding the segment boundaries; the Smith–Waterman algorithm is an efficient method to identify all such islands in the data set.

More formally, let $X(p)$ be the adjusted score for the pth probe ordered along the genome. Let us define the score of the segment from p to q inclusive as

$$T(p,q) = \sum_{i=p}^{q} X(i).$$

Define $S(p)$ to be the score of the island ending at coordinate p, and $B(p)$ to be the coordinate of the beginning of the island. Then it can be shown that the following Smith–Waterman recursion will find the islands. Let $S(0) = 0$, and for $p > 0$

$$S(p) = \begin{cases} S(p-1) + X(p) & \text{if } S(p-1) + X(p) \\ 0 & \text{otherwise} \end{cases}$$

$$B(p) = \begin{cases} B(p-1) & \text{if } S(p) > 0 \\ p & \text{otherwise.} \end{cases}$$

The boundaries $\{B(p_{\max}), p_{\max}\}$ and score $S(p_{\max})$ of the overall maximum-scoring island are output by the algorithm. Note that negative-scoring loci can occur inside an island, thus allowing for occasional false positive or negative signals. In order to identify all islands, the segment corresponding to the maximum-scoring island is replaced by a sequence of zeroes and the algorithm repeated until no positive-scoring islands are detected.

The statistical significance of an island was estimated by permutation, as the proportion of times that a higher-scoring island was found in 1000 runs in which the adjusted log ratios were permuted between the probes and the highest-scoring island in the shuffled data recorded in each run. This method was based on the premise that successive scores from the permuted data approximate the null distribution of scores. When testing for a copy-number increase that does actually exist, this is in fact a conservative assumption, since the permuted data will be drawn from a mixture of the real null distribution and the distribution of scores in the polysomic region. Since the sequences analysed were only 2 Mb long, only the highest-scoring island in each sequence was tested for statistical significance. The R scripts used for pre-processing and analysing the data are available from http://www.well.ox.ac.uk/~tprice/cgh.

The optimal threshold value was found using a set of non-blind data. A sequence of 100 t_0 values between the median and (median + 0.4 × MAD) were tried, and the proportion of times that a position fell within the highest-scoring island was taken as a heuristic indicator of robustness. If this quantity takes a value near 1 at any particular position, it means that a copy-number change is indicated, and that this indication is not sensitive to the value of the threshold. Values near 0 mean that copy-number changes are not indicated, regardless of the value of the threshold. Intermediate values between 0 and 1 mean that the detection of copy-number changes is to some degree sensitive to the choice of threshold value.

RESULTS

Probe by probe global thresholding analysis of DOP-amplified 16p clone data

Control versus control hybridization for 16p probes. We first investigated the variation between probes that occurs across the 2 Mb tiling path in control versus control hybridizations (i.e. where there are no aneusomies present). Figure 1a–c shows the normalized fluorescence ratios for three normal versus normal hybridizations. The ratios vary from probe to probe, but the ratio profile for individual probes is remarkably consistent. Figure 1d shows the mean ratios and SD values obtained from these and three additional normal versus normal hybridizations. It can be seen that some probe results varied very little between hybridizations (minimum SD = 0.048) whereas others showed a larger level of variation (maximum SD = 0.202). The mean SD over all probes was 0.108.

Control versus test hybridizations for 16p probes. The case IDs of the test subjects and the known map positions of their monosomies are given in Table 1. Figure 2 shows the normalized fluorescence ratios, corrected for probe to probe variation, for two of the patients monosomic for different portions of 16p, one with a 1 Mb terminal deletion (Figure 2a) and one with a larger 1.567 Mb deletion (Figure 2b). These are compared with the normalized, corrected mean ratios (which give a reference value of 1.0) and SD values obtained from the six control hybridizations. Using this method, reduced ratios are clearly visible when compared with the control ratios for both patients, but it is also clear that not all probes lying in monosomic regions yield control versus test ratios below a threshold ratio of 0.8 (see Materials and Methods for choice of threshold ratio) and that some of these clones do not show any difference at all, falling within or close to, the SD of the corresponding mean control versus control ratios. In one case, a probe lying in

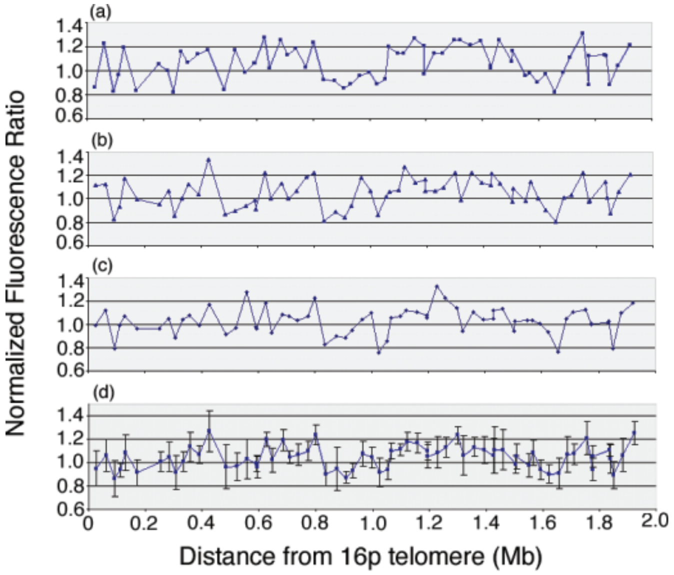

Figure 1. Normalized ratio data (blue data points) from the 16p probes for three control versus control hybridizations (**a–c**) and mean ratio and SD values (shown as black error bars) for six control versus control hybridizations (**d**). The data points are joined by lines to help visualize the similarity between hybridization profiles.

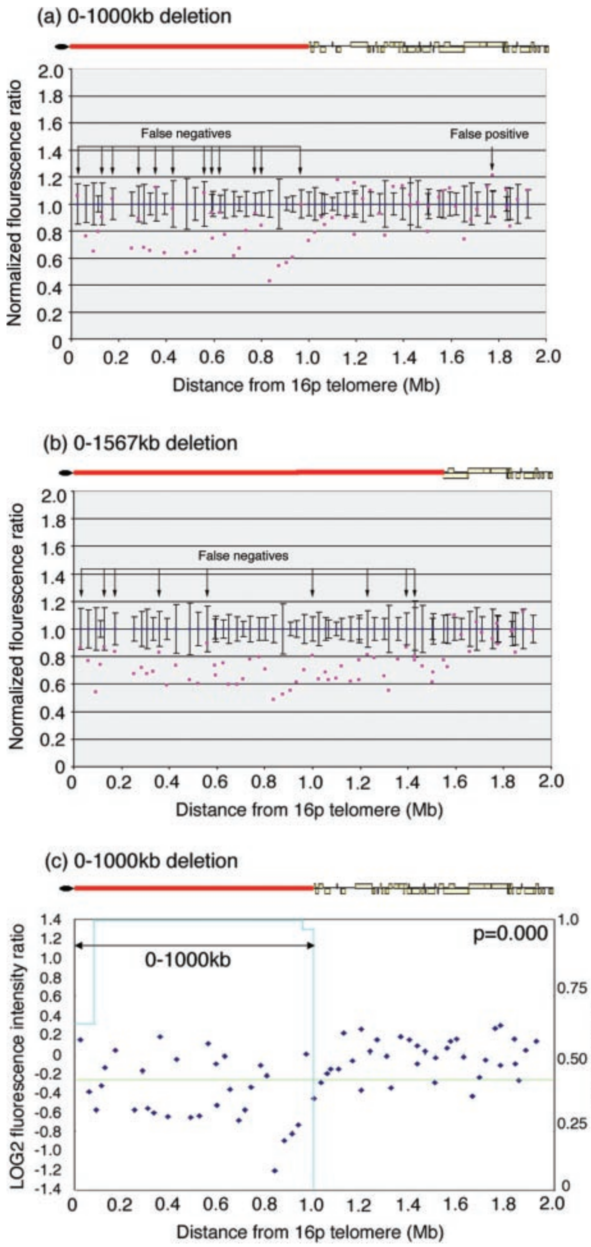

Figure 2. (**a**) and (**b**) Charts showing the mean, normalized hybridization ratios across the terminal 2 Mb of 16p for two patients, monosomic for different portions of 16p (pink data points) and in each case compared with the mean, normalized hybridization ratios and corresponding SD values (black error bars) from the control versus control hybridizations (dark blue data points). Each data point represents data for a single arrayed probe and has been corrected for probe to probe variation. The known region of monosomy for each patient is indicated by the red box in the ideogram of the terminal 2 Mb of chromosome 16p above the relevant chart. The yellow boxes on the ideograms represent known gene locations. The probes giving false negative and a false positive result are annotated and indicated by the arrows. (**c**) Example of the graphical output obtained by applying SW-ARRAY on one data set analysed initially using the probe by probe global thresholding approach [this initial analysis is shown in (a)]. Each black data point indicates the mean, normalized LOG2 fluorescence ratio for a single 16p probe. The significance value is given in the top right of the chart. The robustness values are plotted as blue lines, the threshold value as a green line. The genomic regions identified as monosomic with robustness values >0.5 are indicated by the black double-ended arrow. The known region of monosomy for each patient is indicated by the red box in the ideogram of the terminal 2 Mb of chromosome 16p above the chart and the yellow box on the ideogram is known to be disomic.

the disomic region gives a ratio above the threshold ratio of 1.2, indicating a false positive result (Figure 2a).

The results from these and the additional eight data sets are summarized in Table 2. Analysing each probe independently, 78.1% (95% CI 72.5–83.8) of the 16p probes identified the known monosomic regions correctly and 21.9% (95% CI 16.6–28.0) were false negatives using the threshold ratio of 0.8. The overall percentage of false positives across monosomic and disomic regions and including the control hybridizations was 8.9% (95% CI 7.0–11.1). The error rates in our data are broadly consistent with those reported in earlier studies.

Multiple probe analysis with SW-ARRAY

To deal with the problem of the error rates encountered when considering the probes individually, we analysed the data from array CGH experiments using SW-ARRAY. We first trained the algorithm by optimizing the choice of the threshold parameter, t_0, on a set of non-blind data (i.e. where the locations of aneusomies were known) and then validated the method on a blind data set.

Training SW-ARRAY on non-blind data. Using the known position of the rearrangements from seven initial test samples, the most useful threshold t_0 for identifying monosomy was found to be the median of the log ratios plus 0.2 times their *MAD* when rescaled for normal consistency (i.e. median + 0.2 × *MAD*). We tested thresholds over a range of plausible values to assess the sensitivity of the copy-number change identifications to different values of this parameter and found that the algorithm's performance is relatively insensitive to the threshold value. Figure 2c shows an example of the graphical output obtained by applying the algorithm to one of the data sets analysed above using probe by probe global thresholding (Figure 2a). In this example, the method correctly identified the known 0–1000 kb region of monosomy as a monosomy of 27–1055 kb. This region gave robustness values of >0.5 (predominantly 1.0) indicating insensitivity to the threshold parameter, and a permutation-based statistical significance of $P < 0.00001$. Table 3 summarizes the results obtained in this way from all seven control versus test data sets and all six control versus control data sets. For all except the smallest and the largest monosomy, the identified regions coincide closely with the true extent of the copy-number change in the subjects, indicating that the regions of monosomy are accurately located. The smallest region of monosomy (location 33–300 kb) was not detected ($P = 0.0578$) whereas for the largest monosomy (location 0–1567 kb), the location of the deletion was identified as the region from 356 to 1358 kb ($P < 0.0008$). Importantly, the range of P-values obtained from the control normal versus normal hybridizations was $P = 0.0840$–0.9688 (mean $P = 0.4321$) i.e. none gave a P-value <0.05, indicating that no false positives were found.

Validation of SW-ARRAY on blind data. Figure 3 shows the results obtained when the Smith–Waterman analytical method was applied to 14 data sets obtained from blind array CGH experiments. Following analysis, the decoded data sets were found to include five of the test samples used previously for the global thresholding and training analyses and nine new test samples. However, the blind data sets from the five test samples used previously had been newly generated and therefore were independent of the non-blind data sets used for the probe by probe global thresholding and training analyses. Of the 14 deletions, only the smallest terminal monosomy (Case 1, 0–223 kb) was not detected using the algorithm. All 13 of the remaining monosomies were identified with a high degree of statistical significance. The smallest of these was an interstitial monosomy from 33–300 kb (Case 2, accurately located as 27–356 kb, $P = 0.00086$). Of the six largest monosomies, one, Case 11 (0–1159 kb monosomy) was detected with high statistical significance and was accurately located whereas five were detected with high statistical significance and were predicted to lie in regions that nested within and covered most of the length of the known corresponding

Table 2. Summary of 16p array CGH results analysed probe by probe

Location of monosomy (kb from 16p telomere)	Correct monosomy result probe (%)	Correct disomy result probe (%)	False positive in monosomic region[a] probe (%)	False negative in monosomic region probe (%)	Overall false positive results[b] probe (%)
Test samples					
34–300	86	59	0	14	37
0–775	63	68	0	37	20
0–1000	61	82	0	39	9
0–1052	97	91	0	3	5
0–1160	86	96	0	14	2
0–1183	70	82	0	30	8
0–1567	83	100	0	17	0
Control samples					
0	NA	95	NA	NA	5
0	NA	100	NA	NA	0
0	NA	100	NA	NA	0
0	NA	98	NA	NA	2
0	NA	88	NA	NA	12
0	NA	98	NA	NA	2

[a] Copy-number increases rather than decreases.
[b] Includes false positives in disomic regions.
NA = Not applicable.

Table 3. Non-blind array CGH results obtained using SW-ARRAY

Known location of monosomy (kb from 16p telomere)	Overall significance (P-value)	Algorithm location with robustness >0.5 (kb from 16p telomere)
Test samples		
33–300	0.0578	NA
0–775	0.0042	27–775
0–1000	0.0000	27–1055
0–1052	0.0000	27–1055
0–1160	0.0000	<27–1026
0–1183	0.0000	27–1121
0–1567	0.0008	356–1358
Control samples		
0	0.9688	NA
0	0.3915	NA
0	0.0840	NA
0	0.1459	NA
0	0.8497	NA
0	0.1529	NA

NA = Not applicable.

monosomies. Importantly, for the seven cases of monosomy in the range 267–1052 kb, all were accurately located and correctly identified with a high degree of statistical significance.

DISCUSSION

We have devised a dynamic programming algorithm, SW-ARRAY, to classify regional variations in array CGH fluorescence ratios as copy-number changes. We have implemented a permutation test to assign statistical significance to these classifications and measured the robustness of the co-ordinates of the regions predicted to contain a copy-number change. Our approach can be applied to high-resolution array CGH data irrespective of the arrayed probe type. Importantly, it is scalable and so can be applied to the analysis of data from high resolution arrays such as the newly emerging whole genome large insert tiling path arrays (16) and whole genome oligonucleotide arrays (17). This method is well-suited for high resolution array CGH applications in a clinical diagnostic environment, where reliable, robust assays, providing clear, high quality results of measurable significance are required.

The method achieves optimal performance when an array comprises numerous probes that border as well as bridge a region of copy-number change. This is because SW-ARRAY works by locating the boundaries of regions of copy-number change, and is most effective for long sequences for which edge effects are minimized. When these conditions are met, we were able to detect all rearrangements. Thus, all monosomies ranging from 700 to 1052 kb in length were identified at $P < 0.0001$, and their boundaries accurately located. Of the 14 blind tests, only the 0–223 kb terminal monosomy was not detected. The smallest monosomy detected was the 33–300 kb interstitial deletion ($P = 0.00086$) and the largest monosomy detected was the 0–1567 kb deletion ($P = 0.01344$). In five blind tests from subjects with monosomies in the range 1100–1567 kb, the analysis correctly identified monosomies that were highly statistically significant, but the predicted boundaries of the regions were more sensitive to the threshold value used in the Smith–Waterman procedure (as indicated by robustness values exceeding 0.5 only for the majority rather than the whole of the region known to be deleted). Importantly, the predicted boundaries of monosomies of this size will be less sensitive to the threshold value when larger contiguous chromosomal segments are represented (as in high resolution whole genome tiling path arrays) or when there are many more data points to consider (as in high resolution oligonucleotide arrays).

Using previously published probe by probe global thresholding analysis criteria (7), we found an overall false positive rate of 8.9%. For known monosomic regions, only 78.1% of the 16p probes identified copy-number changes correctly, with 21.9% being false negatives. We also observed a wide range in the SD values when the mean control versus control hybridization ratio SD values for each probe were compared with those of all the probes on the array, suggesting that information may be lost by applying the same threshold values to all probes.

The level of resolution that may be achieved using SW-ARRAY theoretically depends on the sequence length and spatial density of the arrayed probes. However, increased resolution brings with it the increased likelihood of identifying imbalances that are due to very small regions representing phenotypically benign variants or polymorphisms. No single analytical approach will be able to distinguish between these versus pathogenic copy-number changes until such variants/polymorphisms have been well documented and annotated throughout the genome. In the meantime, it will be necessary to follow up all imbalances identified by array CGH in an attempt to determine whether they are likely to be clinically significant.

Other array CGH analytic approaches include the use of smoothing (18), 3-means clustering (19), estimating mixtures of Gaussian distributions (20,21) and hidden Markov models (HMMs) (22,23). Smoothing improves the specificity of global thresholding methods, but needs special tuning to work with different data sources. Clustering procedures, including estimating mixtures of Gaussian distributions, suffer in their performance from not taking into account the spatial dependencies in the data. HMMs have a learning algorithm, but tend to be used to analyse specific classes of data (e.g. from specific cancer cell lines) because of the algorithm's sensitivity to the topology of the HMM. Our approach, like HMM-based and Bayesian algorithms leads to a dynamic programming solution. It bears similarities to the HMM approach, but it has the advantage that it does not rely on the assumption that the data take a parametric (Gaussian) form. In contrast, the Bayesian segmentation algorithm recently put forward by Daruwala *et al.* (24) does make this assumption, though could be converted into a nonparametric algorithm under a mild assumption of bounded variance. Of all the published methods to date, only that of Olshen *et al.* (25) offers a nonparametric segmentation procedure. SW-ARRAY depends only on a single threshold parameter, and makes no other assumptions about the distribution of aneusomic segments. The results we obtained were highly robust to different choices of the threshold. Our statistical analysis is unique in offering not only a nonparametric segmentation procedure, but also a nonparametric test of significance, i.e. once a single threshold parameter has been set, the method identifies regions of copy-number change without assuming that the data follow

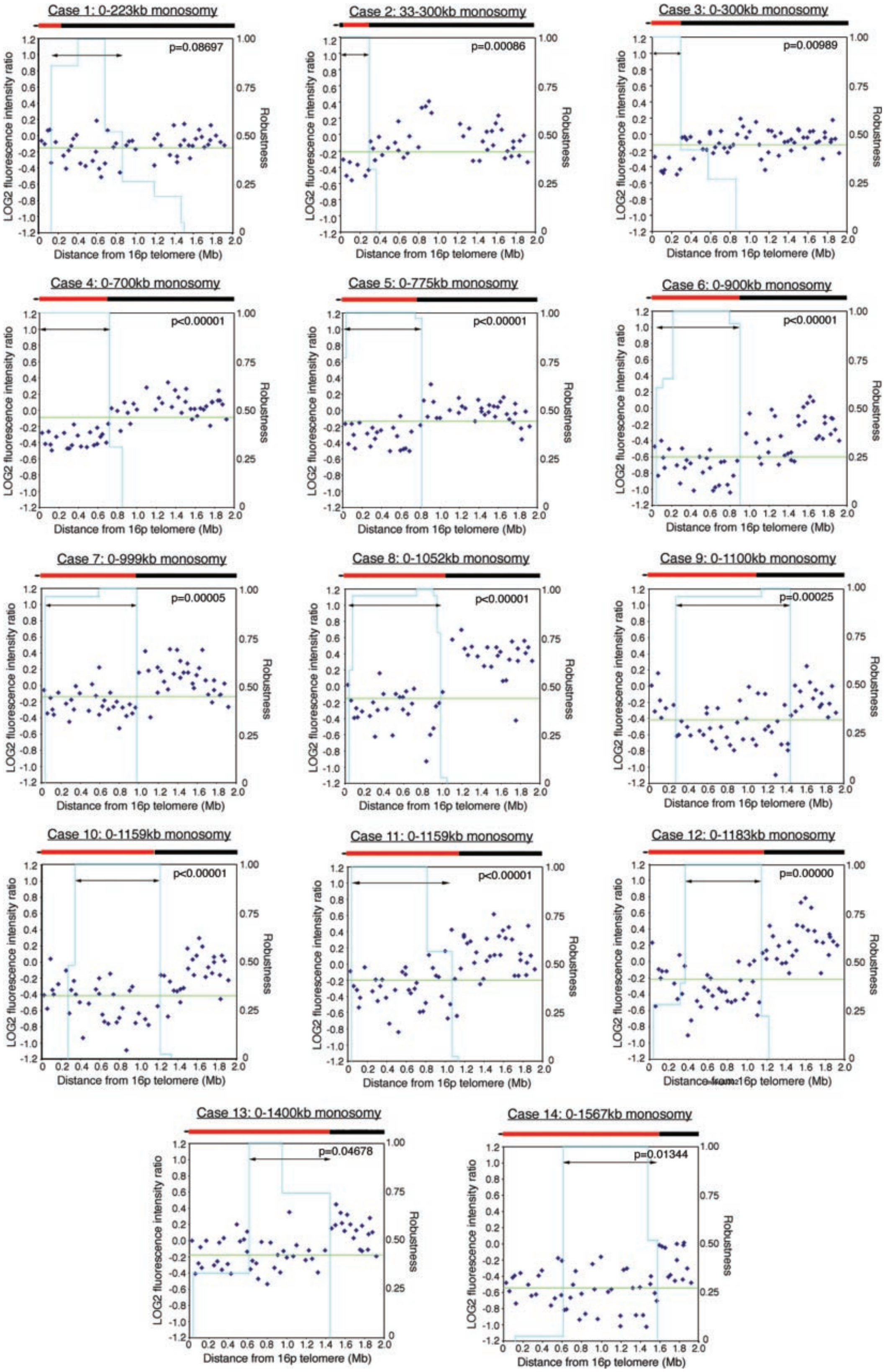

Figure 3. Results of applying SW-ARRAY on data sets with 14 blind control versus test hybridizations. Each black data point indicates the mean, normalized LOG2 fluorescence ratio for a single 16p probe. The significance values are given in the top right of each chart. The robustness values are plotted as blue lines, the threshold values as green lines. Genomic regions identified as monosomic with robustness values >0.5 are indicated by the black double-ended arrows. The known region of monosomy for each patient is indicated by the red box in the ideogram of the terminal 2 Mb of chromosome 16p above the relevant chart. The black box on the ideograms represent disomic regions.

a normal distribution and furthermore, tests the significance of these regions without making any other assumptions.

In summary, the SW-ARRAY method of data analysis represents a way to overcome the problems of low sensitivity and specificity associated with array CGH. The method can be adopted for data from any type of array probe and for all regions of the genome where there is accurate positional information across a contiguous section of chromosome. Its use makes array CGH significantly more suited to clinical diagnostic purposes.

ACKNOWLEDGEMENTS

We are very grateful to Dr Jennifer Taylor for reading the manuscript, sharing her statistical expertise and providing valued comments on the array CGH data analysis. We thank Jo Leach for performing cytogenetic analyses and Dr Niki Meston for valued discussion. We acknowledge The Health Foundation, The Wellcome Trust and The Oxford Genetics Knowledge Park for financial support. Funding to pay the Open Access publication charges for this article was provided by JISC.

Conflict of interest statement. None declared.

REFERENCES

1. Vissers,L.E., de Vries,B.B., Osoegawa,K., Janssen,I.M., Feuth,T., Choy,C.O., Straatman,H., van der Vliet,W., Huys,E.H., van Rijk,A. *et al.* (2003) Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am. J. Hum. Genet*., **73**, 1261–1270.
2. Shaw-Smith,C., Redon,R., Rickman,L., Rio,M., Willatt,L., Fiegler,H., Firth,H., Sanlaville,D., Winter,R., Colleaux,L. *et al.* (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet*., **41**, 241–248.
3. Knight,S.J.L. (2005) Subtelomeric rearrangements in unexplained mental retardation. In Fuchs,J. and Podda,M. (eds), *Encyclopedia of Medical Genomics and Proteomics*.. Marcel Dekker, Inc., New York, USA, pp. 1246–1252.
4. Solinas-Toldo,S., Lampel,S., Stilgenbauer,S., Nickolenko,J., Benner,A., Dohner,H., Cremer,T. and Lichter,P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.
5. Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C., Zhai,Y. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet*., **20**, 207–211.
6. Pollack,J.R., Perou,C.M., Alizadeh,A.A., Eisen,M.B., Pergamenschikov,A., Williams,C.F., Jeffrey,S.S., Botstein,D. and Brown,P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet*., **23**, 41–46.
7. Veltman,J.A., Schoenmakers,E.F., Eussen,B.H., Janssen,I., Merkx,G., van Cleef,B., van Ravenswaaij,C.M., Brunner,H.G., Smeets,D. and van Kessel,A.G. (2002) High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am. J. Hum. Genet*., **70**, 1269–1276.
8. Yu,W., Ballif,B.C., Kashork,C.D., Heilstedt,H.A., Howard,L.A.., Cai,W.W., White,L.D., Liu,W., Beaudet,A.L., Bejjani,B.A. *et al.* (2003) Development of a comparative genomic hybridization microarray and demonstration of its utility with 25 well-characterized 1p36 deletions. *Hum. Mol. Genet*., **12**, 2145–2152.
9. Daniels,R.J., Peden,J.F., Lloyd,C., Horsley,S.W., Clark,K., Tufarelli,C., Kearney,L., Buckle,V.J., Doggett,N.A., Flint,J. and Higgs,D.R. (2001) Sequence, structure and pathology of the fully annotated terminal 2 Mb of the short arm of human chromosome 16. *Hum. Mol. Genet*., **10**, 339–352.
10. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.
11. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
12. Knight,S.J., Horsley,S.W., Regan,R., Lawrie,N.M., Maher,E.J., Cardy,D.L., Flint,J. and Kearney,L. (1997) Development and clinical application of an innovative fluorescence *in situ* hybridization technique which detects submicroscopic rearrangements involving telomeres. *Eur. J. Hum. Genet*., **5**, 1–8.
13. Knight,S.J., Regan,R., Nicod,A., Horsley,S.W., Kearney,L., Homfray,T., Winter,R.M., Bolton,P. and Flint,J. (1999) Subtle chromosomal rearrangements in children with unexplained mental retardation. *Lancet*, **354**, 1676–1681.
14. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
15. Fiegler,H., Carr,P., Douglas,E.J., Burford,D.C., Hunt,S., Scott,C.E., Smith,J., Vetrie,D., Gorman,P., Tomlinson,I.P. and Carter,N.P. (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones [Erratum (2003) Genes Chromosomes Cancer, **37**, 223]. *Genes Chromosomes Cancer*, **36**, 361–374.
16. Ishkanian,A.S., Malloff,C.A., Watson,S.K.,, DeLeeuw,R.J., Chi,B., Coe,B.P., Snijders,A., Albertson,D.G., Pinkel,D., Marra,M.A. *et al.* (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genet*., **36**, 299–303.
17. Lucito,R., Healy,J., Alexander,J., Reiner,A., Esposito,D., Chi,M., Rodgers,L., Brady,A., Sebat,J., Troge,J. *et al.* (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res*., **13**, 2291–2305.
18. Jong,K., Marchiori,E., Meijer,G., Vaart,A.V. and Ylstra,B. (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20**, 3636–3637.
19. Autio,R., Hautaniemi,S., Kauraniemi,P., Yli-Harja,O., Astola,J., Wolf,M. and Kallioniemi,A. (2003) CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics*, **19**, 1714–1715.
20. Hodgson,G., Hager,J.H., Volik,S., Hariono,S., Wernick,M., Moore,D., Nowak,N., Albertson,D.G., Pinkel,D., Collins,C. *et al.* (2001) Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genet*., **29**, 459–464.
21. Wang,J., Meza-Zepeda,L.A., Kresse,S.H. and Myklebost,O. (2004) M-CGH: analysing microarray-based CGH experiments. *BMC Bioinformatics*, **5**, 74.
22. Fridlyand,J., Snijders,A., Pinkel,D., Albertson,D.G. and Jain,A. (2004) Application of hidden Markov models to the analysis of the array CGH data. *J. Multivar. Anal*., **90**, 132–153.
23. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M., Chi,M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
24. Daruwala,R.-S., Rudra,A., Ostrer,H., Lucito,R., Wigler,M. and Mishra,B. (2004) A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl Acad. Sci*., **46**, 16292–16297.
25. 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.
26. Horsley,S.W., Daniels,R.J., Anguita,E., Raynham,H.A., Peden,J.F., Villegas,A., Vickers,M.A., Green,S., Waye,J.S., Chui,D.H. *et al.* (2001) Monosomy for the most telomeric, gene-rich region of the short arm of human chromosome 16 causes minimal phenotypic effects. *Eur. J. Hum. Genet*., **9**, 217–225.
27. Waye,J.S., Chui,D.H.K., Higgs,D.R., Hetherington,R. and Olivieri,N.F. (1995) *De novo* deletion of the entire α-globin gene cluster in a girl with Hb H disease. *Blood*, **86**, 8a.
28. Horsley,S.W. (2000) Characterisation of chromosome 16 rearrangements in patients with alpha thalassaemia. PhD Thesis, Oxford Brookes University, Oxford, UK.
29. Raynham,H.A. (1995) The molecular basis of the ATR-16 (alpha thalassaemia/mental retardation) syndrome. DPhil, University of Oxford, Oxford, UK.

30. Rack,K.A., Harris,P.C., MacCarthy,A.B., Boone,R., Raynham,H., McKinley,M., Fitchett,M., Towe,C.M., Rudd,P., Armour,J.A. *et al.* (1993) Characterization of three *de novo* derivative chromosomes 16 by 'reverse chromosome painting' and molecular analysis. *Am. J. Hum. Genet.*, **52**, 987–997.

31. Wilkie,A.O.M., Buckle,V.J., Harris,P.C., Lamb,J., Barton,N.J., Reeders,S.T., Lindenbaum,R.H., Nicholls,R.D., Barrow,M., Bethlenfalvay,N.C. *et al.* (1990) Clinical features and molecular analysis of the α-thalassaemia/mental retardation syndromes. I. Cases due to deletions involving chromosome band 16p13.3. *Am. J. Hum. Genet.*, **46**, 1112–1126.