Comparative genomic hybridisation using a proximal 17p BAC/PAC array detects rearrangements responsible for four genomic disorders

C J Shaw, C A Shaw, W Yu, P Stankiewicz, L D White, A L Beaudet, J R Lupski

Background: Proximal chromosome 17p is a region rich in low copy repeats (LCRs) and prone to chromosomal rearrangements. Four genomic disorders map within the interval 17p11–p12: Charcot–Marie–Tooth disease type 1A, hereditary neuropathy with liability to pressure palsies, Smith–Magenis syndrome, and dup(17)(p11.2p11.2) syndrome. While 80–90% or more of the rearrangements resulting in each disorder are recurrent, several non-recurrent deletions or duplications of varying sizes within proximal 17p also have been characterised using fluorescence in situ hybridisation (FISH).

Methods: A BAC/PAC array based comparative genomic hybridisation (array-CGH) method was tested for its ability to detect these genomic dosage differences and map breakpoints in 25 patients with recurrent and non-recurrent rearrangements.

Results: Array-CGH detected the dosage imbalances resulting from either deletion or duplication in all the samples examined. The array-CGH approach, in combination with a dependent statistical inference method, mapped 45/46 (97.8%) of the analysed breakpoints to within one overlapping BAC/PAC clone, compared with determinations done independently by FISH. Several clones within the array that contained large LCRs did not have an adverse effect on the interpretation of the array-CGH data.

Conclusions: Array-CGH is an accurate and sensitive method for detecting genomic dosage differences and identifying rearrangement breakpoints, even in LCR-rich regions of the genome.

Abbreviations: array-CGH, microarray based comparative genomic hybridisation; CGH, comparative genomic hybridisation; FISH, fluorescence in situ hybridisation; HMM, hidden Markov model; HNPP, hereditary neuropathy with liability to pressure palsies; LCR, low copy repeat; MAD, median absolute deviation; mCGH, metaphase-CGH; PFGE, pulsed field gel electrophoresis; SMS, Smith–Magenis syndrome.
METHODS

Subjects
We analysed 25 individuals with deletions or duplications of proximal 17p using FISH and array-CGH. These included 15 patients with deletions, nine with duplications, and one with a deletion and a duplication. Control individuals (one male and one female) were unaffected parents of patients with deletions and had normal karyotypes. Peripheral blood samples from patients and family members were obtained after informed consent approved by the Baylor College of Medicine institutional review board.

FISH
FISH analysis of deletion patients was done as described before.3 Dual colour FISH analysis of duplication patients was done on metaphase and interphase preparations of human peripheral blood lymphocytes and Epstein-Barr virus transformed lymphoblasts according to a modified procedure.24

Array-CGH
A minimal tiling path of 56 BAC/PAC clones from the centromere through the CMT1A region on 17p was included in the array, along with 16 control normalisation clones from chromosomes 2, 5, 9, 10, X, and Y. The DNA from BAC and PAC clones was prepared for array spotting as described.2 Briefly, the DNA was chemically cross linked using (3-glycidoxypropyl)tri-methoxysilane (Sigma) and printed onto glass slides using an OmniGrid Accent microarrayer with Telechem Array II Chipmaker III pins (GeneMachine). Each clone was spotted in quadruplicate. Spotting was done in the Baylor College of Medicine microarray core facility. Patient DNA was isolated from peripheral blood using a Puregene kit (Gentra). The DNA was digested with DpnII restriction enzyme (New England Biolabs) and RNase A (Roche), and then purified using a QIAquick gel extraction kit (Qiagen).

Patient and sex matched control DNA (250 ng) was differentially labelled with cyanine 3-dCTP and cyanine 5-dCTP (Perkin Elmer) using a BioPrime labelling kit (Invitrogen). Each pair of patient and control DNA samples was labelled twice with the dyes reversed and hybridised to the array at 37°C for 24 hours. The microarray slides were washed at 42°C, scanned using a GenePix microarray scanner, and microarray image quantification was carried out using GLEAMS software (Nutec Sciences). Each BAC/PAC clone position was interrogated eight times; quadruplicate spottings were each examined twice with dye reversal.

Three control versus control hybridisations (two male, one female) were done and showed reproducible normalised log2(Cy3/Cy5) ratios.

Statistical analysis
Quantified array image files (.tiff) were subjected to single channel normalisation, and dye reversed array pairs were subjected to bi-chip normalisation. All analysis was done on log2 ratios. The justification for single chip normalisation in spotted arrays is well documented.25 26 The normalisation is used to remove systematic biases such as spatial and intensity artefacts, and bi-chip scaling is used to bring the dye reversed hybridisations to a common measurement scale to facilitate combining the microarray pairs for each patient. The bi-chip scaling factors for chip pairs are motivated by the MAD scaling approach.26 Briefly, the median absolute deviation about 0 of the by-clone average normalised log2 ratio is calculated for each chip in a dye reversed pair; we denote these MAD values as m for i = 1, 2. Single chip scaling factors are then determined as s = m/(m + m). Such scaling factors are well motivated statistically and they have the interpretation of drawing the dye reversed data to the y = −x line. Once scaled, the data are sign changed and averaged to form a single value for each clone for each patient. These dye reversed average data are then used to make inferences regarding the gain/loss status of each clone for each patient. The inference uses a seven state hidden Markov model (HMM), which classifies each clone for each patient into the outcomes gain, loss, and no change (described below). The HMM technique considers the data at adjacent clones as statistically dependent to form an inference at each BAC/PAC locus. The HMM results were evaluated against the previous independent FISH analyses of a subset of the patients with rearrangements.

HMM inference method
The goal of inference in array-CGH is accurate detection of chromosomal change, which avoids false positive calls. An additional goal in our experiment is precise inference for the boundaries of chromosomal lesions, to call the breakpoints. The approach we took to accomplish these goals is to incorporate the adjacency information for the printed BACs/PACs into our statistical inference by use of an HMM. HMMs are well studied inference methods for analysing dependent data, and these models have seen extensive use in biology in the field of sequence alignment. HMMs always consist of two components: a hidden sequence of unobserved states which are treated as a Markov chain, and a collection of observed emission data. An inference for the sequence of hidden states is formed using the observed emission values.

We developed a seven state HMM to call lesion boundaries and to infer the gain/loss status of each clone in each patient. The hidden states in our CGH array HMM were the gain/no change/loss status of each BAC in each patient. The emission values in our HMM were the observed normalised microarray values for each patient. The seven hidden states in our model are: initiate loss, loss, end loss, no change, initiate gain, gain, end gain. The use of initiate and end states yields better performance for calling lesion boundaries than a simpler model with fewer states. The emission distributions are all univariate Gaussian distributions determined separately for each clone, conditional on its gain/loss status in the FISH data.

We fitted and evaluated our model on 25 patients with known FISH data. To undertake fitting and to obtain inferences in an objective fashion, we took a cross validation approach. For each patient, we performed the model fit using all patients except the one under consideration. We then made inferences on the patient excluded from the fitting process. This step is done for each patient to generate the inference results.

In each fitting step, the HMM transition matrix is directly estimated from the observed transition frequencies in the remaining 24 patients. Emission distributions for each clone are determined by estimating a mean and variance value for each clone, conditional on the FISH outcome status in each patient. For clones in which no patient showed a gain/loss state, the emission distribution is estimated using the average value across all clones for which data were available. For the transition states, the emission mean was estimated to be the mean of the no change mean and the pure loss or gain mean, respectively, for the gain and loss transition states.

Inferences for each patient were made by using the Vitterbi algorithm, which is the well studied method for obtaining the highest probability path conditional on the observations and the model. A key feature is the ability of the HMM to make
correct inferences even in regions wherein data show high variance and might otherwise lead to mistaken conclusions.

RESULTS

Array-CGH validation of rearrangements previously mapped by FISH and PFGE

To test the reliability and accuracy of the array-CGH technology, we hybridised the DNA of 12 deletion patients and one duplication patient whose unusual rearrangements were previously characterised by FISH (fig 1). In addition, three patients with common SMS deletions (1780, 1949, 1957), two patients with common dup(17)(p11.2p11.2) duplications (1789, 1913), two patients with CMT1A duplications (682, 723), and one patient with a common dup(17)(p11.2p11.2) duplication and a common HNPP deletion (1006) were analysed. All eight of these patients were documented previously to have a common deletion/duplication by the presence of a rearrangement specific pulsed field gel electrophoresis (PFGE) junction fragment or FISH analysis. Those patients with common rearrangements analysed by PFGE only are assumed to have breakpoints mapping within the proximal and distal SMS-REPs (in the case of 17p11.2 deletions/duplications) or within the proximal and distal CMT1A-REPs (in the case of 17p12 deletions/duplications). Control individuals were hybridised to the array, and none of the clones was inferred as deleted or duplicated using the HMM analysis (fig 1A).

The array-CGH analysis was accurate in detecting clones displaying a gain or loss in each patient tested (figs 1–3). Importantly, both deletions (fig 1B) and duplications (fig 1C) were readily detected. Interestingly, in a rare patient with both a deletion and a duplication of 17p, both rearrangements were easily discerned (fig 1D). As expected, all three SMS patients with common deletions revealed similar patterns in the array-CGH statistical plots (fig 2).

The deletion/duplication breakpoints detected using array-CGH were consistent with those previously mapped by FISH and PFGE (fig 3). The distal deletion breakpoint of patient 993, previously unmapped by FISH, was identified by array-CGH and subsequently confirmed by FISH. This breakpoint was mapped between clones RP11-64B12 and RP11-849N15 (figs 3A, 4A). The distal deletion breakpoint of patient 357 was confirmed to extend distally beyond clone RP11-350B3 (figs 3A, 4B).

![Figure 1: Plots of array-CGH data. The 17p clones are ordered from telomere (left) to centromere (right). Individual clone mean normalised log2(Cy3/Cy5) ratios of patient to control are plotted on the Y axis and represented by dots with error bars; clones containing LCRs >20 kb are highlighted in red. A blue smoothing spline is drawn between the points. Breakpoint clones are labelled and denoted with arrows. A diagram of proximal 17p is shown at the bottom; the black circle represents the centromere, the diagonally striped boxes represent the SMS-REPs, and the dotted boxes represent the CMT1A-REPs. (A) 1950; female control, showing no dosage gain or loss. (B) 1153; large SMS deletion from CTC-3157E16 through RP11-434D2. (C) 1861; large duplication from RP11-998F8 through RP11-98L14. (D) 1006; common dup(17)(p11.2p11.2) from RP11-92B11 through RP11-434D2 and HNPP deletion from RP11-214O1 through RP11-726O12.](http://jmg.bmj.com/)
Accuracy of array-CGH breakpoint mapping

The rearrangement breakpoints were accurately predicted by array-CGH, assigning 45/46 breakpoints (97.8%) correctly to within one overlapping adjacent clone of the breakpoint identified by FISH, and 100% to within two clones (fig 3). Previous FISH analysis of some of the clones spotted on the array showed a weak signal, indicating that those clones were partially deleted. However, as it is unknown what portion of the clone must be deleted to show a loss using array-CGH, we considered either a no change or loss inference to be correct for clones shown to be partially deleted by FISH. Six of the cases analysed had one or both breakpoints located proximally or distally of the clones contained in the array (patients 357, 527, 563, 1229, 1458, and 1861). All eight of these breakpoints were correctly inferred by array-CGH to extend beyond the clones contained in the array (fig 3).

While the dosage changes and breakpoints were correctly inferred, a few clones did not perform as well as the majority when analysed independently (figs 1 and 2). RP1-836L9 did not show a gain or loss in any of the patients tested, indicating that the quality of the PAC DNA printed in the array was suboptimal for hybridisation. Likewise, in several cases, CTD-124H2, RP1-37N7, and RP11-48J14 appeared to have no dosage change, while they were shown to be deleted by FISH. Although these small inconsistencies are apparent in the plots of the statistical mean of the log2(Cy3/Cy5) fluorescence ratio (figs 1 and 2), they had no effect on the final inference (fig 3).

DISCUSSION

We have examined the capabilities of array-CGH to detect recurrent rearrangements (CMT1A duplication/HNPP deletion, dup(17)(p11.2p11.2)/SMS deletion) and map the breakpoints of unique non-recurrent rearrangements in proximal 17p, an LCR-rich genomic interval. Array-CGH allowed the detection of losses from deletion and gains from duplication. Furthermore, breakpoints for both recurrent and uniquely sized deletions and duplications were readily discerned using our inference analysis. Our study shows that array-CGH can detect dosage differences of 17p in patients compared with normal controls despite challenges introduced by complex genome architecture.

Comparison of array-CGH and FISH

While array-CGH has proved to be an accurate method for detecting genomic dosage change and mapping breakpoints, it remains unclear whether array-CGH or FISH is a more sensitive technique. Previous work has estimated the resolution of array-CGH to be as low as 40 kb when using BAC/PAC clones in the array, while the resolution of FISH is known to be as low as 1000 base pairs (bp) when cosmids and polymerase chain reaction (PCR) products are used as probes. However, finer resolution is probably feasible with array-CGH if smaller segments of DNA such as cosmids or PCR products are used to construct the array.

Array-CGH is a much more rapid and higher throughput technique than FISH, with the ability to test for hundreds or thousands of loci in a single analysis. Array-CGH is an exceptional tool for whole genome screening of dosage imbalances, some of which may remain undetected by standard disease specific FISH analysis. This technique has proved sensitive enough to detect triplications in chromosome 1p, although the sensitivity threshold of array-CGH in mosaic cell lines is yet to be thoroughly investigated. Recently, array-CGH was shown to be sufficiently sensitive to detect duplication of all clones contained on a ring chromosome 18q that was present in 75% of cells. However, the same study showed array-CGH was not sensitive enough to detect clones that were deleted in a mosaic cell line.
carrying an 18q deletion in 33% of cells. This suggests that rearrangements present in less than 75% of cells may go undetected by array-CGH, while these cases are readily identified by classical cytogenetics and FISH. In addition, array-CGH is unable to detect balanced translocations. Thus it seems that although array-CGH is revolutionising the investigation of chromosomal rearrangements, there is still a great need for classical cytogenetics and FISH studies. For FISH studies, a region of suspected abnormality must be chosen for study, while array-CGH offers the potential to test for very large numbers of loci at once. From a clinical application standpoint, the array-CGH can enable a simultaneous high resolution analysis of the entire human genome. Abnormalities identified by such a screening approach could be confirmed by a locus specific FISH test.

Methods of array-CGH analysis
Different methods of analysis are applicable to array-CGH data, depending on the information desired from the experiment. An independent analysis method considers each clone separately when assigning a gain/loss/no change state. This method is accurate for predicting small interstitial changes, as the states of the clones near the breakpoint are not dependent on the states of the adjacent overlapping clones. However, individual clones may be assigned states that are inconsistent with adjacent clones, resulting in a deletion/duplication that appears non-contiguous. Although a handful of clones may perform less well than the majority when analysed using an independent method of analysis, these small inconsistencies vanish when a dependent analysis is implemented. This analysis method considers the state (gain/loss/no change) of the adjacent, overlapping clones and the number of consistently assigned clones when assigning a state to any particular clone. Thus a clone that has an inconsistent state when analysed independently will be consistent with adjacent, overlapping clones when analysed dependently. This method works well when information regarding contiguity of the deletion/duplication is desired, but may sometimes be less accurate at predicting the breakpoints of the rearrangement. The addition of intermediate states—begin and end states—facilitates the accurate call of breakpoints. Additionally, important information regarding the physical characteristics of particular clones may be concealed when analysed dependently. A

Figure 3  Summary of array-CGH and fluorescence in situ hybridisation (FISH) data. A schematic representation of proximal chromosome 17p, and the order of clones included in the array are shown at the top. Patient numbers are given on the left. Asterisks indicate patients whose breakpoints are assumed based on the presence of a rearrangement specific junction fragment when analysed by pulsed field gel electrophoresis. The grey bars represent clones that do not have a dosage imbalance. (A) Deletion analysis. The blue bars indicate clones that are deleted based on array-CGH data. The pink bars represent clones that were deleted based on FISH analysis. Diagonal stripes indicate clones that were not used as FISH probes; dotted boxes indicate clones that showed a partial signal when used as FISH probes. (B) Duplication analysis. The purple bars indicate clones that are duplicated based on array-CGH data. The green bars represent clones that were duplicated based on FISH analysis. Patient 1006 has both a deletion and a duplication.
Figure 4  Fluorescence in situ hybridisation (FISH) mapping of breakpoints. (A) Patient 993; RP11-998F8 (red) shows two signals, while RP11-64B12 (green) shows one signal, indicating the deletion breakpoint maps between them. (B) Patient 357; RP11-350B3 shows one signal, indicating the deletion breakpoint is distal to this clone. (C) Patient 527; RP11-330B3 (green) and RP11-98L14 (red) show three signals, indicating the duplication encompasses all clones in the array, and may extend proximally or distally. (D) Patient 1458; RP11-350B3 (green) and RP11-98L14 (red) both have three hybridisation signals, showing the duplication extends beyond the clones within the array. This is a G2 cell, and thus each signal is present in duplicate. (E) Patient 563; two hybridisation signals are present for RP11-849N15 (red), while three are present for RP11-726012 (green; two of which are superimposed on the red signals and appear yellow), indicating the duplication breakpoint is between them. (F) Patient 1229; RP11-849N15 (red) gives three signals, while RP11-924A14 (green) gives two, suggesting the duplication breakpoint maps between them.

The method of analysis used in this paper merges the benefits of both independent and dependent analysis by assigning not only gain/loss/no change states to the clones, but also intermediate states. Thus transitions between a gain or loss state and a no change state (which occur at the breakpoints of the rearrangements) are more accurately predicted. In this report, 97.8% of the breakpoints identified using array-CGH matched those detected using FISH. A comparison of the average variance (across 25 cases) of the log2(Cy3/Cy5) fluorescence ratio (patient/control) for each clone showed that BACs and PACs containing LCRs did not differ from unique sequence clones in this aspect. This observation is expected, as the LCR containing and unique sequence clones perform similarly in the experiment. A general trend showed that clones within the CMT1A region had a lower average variance value than clones within the SMS region, perhaps reflecting the increased size and number of LCRs in the SMS region. Thus these data show that array-CGH is an efficient method with which to detect genomic dosage change and map rearrangement breakpoints, even in regions of the genome laden with LCRs.

**ACKNOWLEDGEMENTS**

We thank the patients for their participation. We also thank Marjorie Withers for excellent technical assistance. This study was supported in part by grants from the National Institute of Child Health and Human Development (PO1 HD39420) and the Mental Retardation Research Center (HD24064).

**Authors’ affiliations**

C J Shaw, C A Shaw, W Yu, P Stankiewicz, I D White, A L Beaudet, J R Lupski, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA.

**REFERENCES**


