METHODS

Representational Oligonucleotide Microarray Analysis (ROMA) and Comparison of Binning and Change-Point Methods of Analysis: Application to Detection of del22q11.2 (DiGeorge) Syndrome

Christopher M. Stanczak,1 Zugen Chen,1 Stanley F. Nelson,1 Marc Suchard,1,5,6 Edward R.B. McCabe,1–4 and Sean McGhee2*

1Department of Human Genetics, David Geffen School of Medicine at the University of California, Los Angeles (UCLA), Los Angeles, California; 2Department of Pediatrics, David Geffen School of Medicine at UCLA; 3Mattel Children's Hospital at UCLA, Los Angeles, California; 4Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California; 5Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, California; 6Department of Biostatistics, UCLA School of Public Health, Los Angeles, California

Communicated by Paolo Fortina

DiGeorge (del22q11.2) syndrome is estimated to occur in 1:4,000 births, is the most common contiguous-gene deletion syndrome in humans, and is caused by autosomal dominant deletions in the 22q11.2 DiGeorge syndrome critical region (DGCR). Multiple microarray methods have been developed recently for analyzing such copy number changes, but data analysis and accurate deletion detection remains challenging. Clinical use of these microarray methods would have many advantages, particularly when the possibility of a chromosomal disorder cannot be determined simply on the basis of history and physical examination data alone. We investigated the use of the microarray technique, representational oligonucleotide microarray analysis (ROMA), in the detection of del22q11.2 syndrome. Genomic DNA was isolated from three well-characterized cell lines with 22q11.2 DGCR deletions and from the blood of a patient suspected of having del22q11.2 syndrome, and analyzed using both the binning and change-point model algorithms. Though the 22q11.2 deletion was easily identified with either method, change-point models provide clearer identification of deleted regions, with the potential for fewer false-positive results. For circumstances in which a clear, a priori, copy-number change hypothesis is not present, such as in many clinical samples, change-point methods of analysis may be easier to interpret. Hum Mutat 0, 1–6, 2007.

INTRODUCTION

22q11.2 Deletion (DiGeorge) syndrome (MIM# 188400) is characterized by congenital cardiovascular defects, craniofacial abnormalities, and hypoplasia of the thymus and parathyroid glands, causing T cell deficiency and hypocalcemia [Shaffer et al., 2001]. Deletion of TBX1 (MIM# 602054), resulting in abnormal development of the third and fourth pharyngeal pouches, is thought to explain the phenotype. The 3-MB DiGeorge syndrome critical region (DGCR) is deleted in approximately 90% of cases, and is the most frequent interstitial deletion in humans [Emanuel et al., 2001]. Most of the remaining cases have a smaller ~1.5-MB deletion in the region [Greenberg et al., 1988]. Approximately 25 genes are thought to lie in the 22q11.2 typically deleted DGCR. Usually the deletion occurs as a sporadic de novo deletion and autosomal dominant inheritance is observed [Wilson et al., 1991]. Low copy number repeats are present at the areas that map to the breakpoints in both the 1.5-MB and 3-MB deletions [Edelmann et al., 1999]. These low copy number repeats have >97% sequence identity, and thus predispose the region to homologous recombination and unequal crossing over.

Array comparative genomic hybridization (aCGH) techniques are microarray based copy-number analysis procedures that make use of the high density of microarrays to provide genomewide identifications of alterations in copy number [Pollack et al., 1999]. aCGH has been performed on genomic DNA from patients with DiGeorge syndrome. The techniques involved cohybridization of labeled normal and test genomic DNA to a collection of BACs or genomic PCR fragments spotted on a microscope slide. Buckley et al. [2002] constructed a 500-element array (167 cosmids, 329

Received 31 January 2007; accepted revised manuscript 4 June 2007.

*Correspondence to: Sean A. McGhee, MD, Department of Pediatrics, David Geffen School of Medicine at UCLA, 10833 LeConte Ave., Los Angeles, CA 90095-1752. E-mail: smcghee@mednet.ucla.edu

Grant sponsor: National Institutes of Health, National Institute of Child Health and Human Development; Grant numbers: R01 HD22563; U24.

DOI 10.1002/humu.20593
Published online in Wiley InterScience (www.interscience.wiley.com).
BACs, and four PCR pools) covering chromosome 22 to an average of one element every 75 kb. They reported detection of a 22q11.2 deletion in a patient with DiGeorge syndrome. Mantripragada et al. (2004) constructed an even narrower array, with 40 clones covering 6 MB around 22q11.2. With a resolution of about 25 kb, they identified 22q11.2 deletions in six patients with DiGeorge syndrome. Prescott et al. (2005) applied this technique to genomic DNA from six patients with the 22q11 deletion syndrome phenotype, but no detectable deletion by fluorescence in situ hybridization (FISH) analysis.

More recently, a new microarray-based technique for copy-number estimation, representational oligonucleotide microarray analysis (ROMA) [Lucito et al., 2003; Johansen et al., 2005; Slater et al., 2005; Huang et al., 2006; Ming et al., 2006], has been developed that makes use of oligonucleotide-based microarrays, possibly increasing the resolution of copy-number estimations, and certainly increasing the potential for standardization of arrays. We have shown that the Affymetrix GeneChip 50K Mapping Array (Affymetrix), which utilizes short 25-pb synthetic oligonucleotides as the probes on the array instead of large genomic fragments found on bacterial artificial chromosome arrays used in aCGH, offers greater resolution, reduced noise, and a simple protocol that can be automated [Stanczak et al., 2006].

Greater application of microarray copy-number analysis in DiGeorge syndrome populations could help clarify certain questions in DiGeorge syndrome, such as the role of deletions outside of the 22q11.2 region. Our purpose was to investigate whether the Affymetrix array could be utilized to detect the loss of a single copy of genomic DNA from patients with DiGeorge syndrome. We further sought to determine what method of data analysis provides the optimal interpretation of results. Ease in identifying alterations in copy number will be needed if this technology is to transfer to clinical practice.

**MATERIALS AND METHODS**

**Subjects**

**DiGeorge patient.** Several milliliters of blood were obtained from a patient with DiGeorge syndrome after informed consent approved by the University of California Los Angeles (UCLA) Institutional Review Board. The patient was a 10-month-old boy who had confirmed del(22)q11.2 syndrome by FISH analysis. He had been diagnosed as part of an evaluation for severe gastroesophageal reflux and failure to thrive, after two episodes of pneumonia and frequent otitis media. He had mildly cup-shaped ears and first degree hypospadias, but had never had documented hypocalcemia; T-cell counts were adequate. He had no history of heart disease.

**Cell lines.** DNA was also isolated from cell lines derived from two individuals with DiGeorge syndrome (Patients GM07215 and GM13325) and one with velocardiofacial syndrome (GM07939) that were obtained from Coriell Cell Repositories. All three cell lines were verified by FISH to have 22q11.2 deletions, and therefore represented appropriate known deletions on which to test our analytical approach.

Patients GM07215 and GM13325 were clinically affected with DiGeorge syndrome, and presented with an interrupted aortic arch, thymic aplasia, and atrial and ventricular septal defects. Patient GM07939 and his older brother, who died at 3 months of age, were described clinically with velocardiofacial syndrome, and presented with truncus arteriosus, hypoparathyroidism, lack of radiographic thymic shadow on chest X-ray, and facial dysmorphism. Their father had hypertelorism, micrognathia, low-set ears, relative hypoparathyroidism, and deficiency of T-lymphocytes. Patient GM07939 also showed these features, and additionally had two supernumerary digits of the upper extremity, a cleft palate, micrognathia, ventricular septal defect, and hypocalcemia.

**METHODS**

The Affymetrix Xba 50K Mapping Array (SNP chip) contains 20 perfect-match and 20 mismatch oligonucleotides at each of 58,494 SNPs across the genome (except the Y chromosome), with 328 SNPs on chromosome 22 (an average of one SNP every 150 kb), and 19 SNPs in the 3 MB 22q11.2 DOCR (an average of one SNP every 158 kb). The SNP distribution is not even across the region, as it is based on genetic, not physical, distance and limitations of the molecular method that relies on coamplification of XbaI fragments. This uneven distribution means that the resolution at which breakpoints can be determined may be limited by the available SNPs in the particular region under study. The low copy repeats in the DiGeorge critical region, where the typical breakpoints are located in del22q11.2 syndrome, do not have a high SNP resolution on the XbaI array, so although the general extent of the deletion can be determined, precise breakpoint mapping must be conducted by other methods, such as PCR or multiplex ligation-dependent probe amplification (MLPA) [Vorstman et al., 2006].

Each patient's genomic DNA was subjected to restriction digestion with XbaI (New England Biolabs, Ipswitch, MA) for 2 hr and the enzyme was deactivated by heating to 70°C for 20 minutes. Next, proprietary adaptor sequences (Affymetrix, Santa Clara, CA) containing XbaI compatible ends and primer binding sites were ligated to the digested genomic DNA fragments using T4 DNA ligase (New England Biolabs) at 16°C for 2 hr and heat inactivated at 70°C for 20 minutes. A single-primer PCR amplification was performed for 35 cycles with an annealing temperature of 59°C and an extension temperature of 72°C using AmpliTaq Gold (Applied Biosystems). The products were purified using QiAquick PCR Purification spin columns (Qiagen, Valencia, CA), quantitated by spectrophotometry, and qualitatively checked by agarose gel electrophoresis. Each sample was fragmented by DNase I (Affymetrix) at 37°C for 30 minutes, again checked by gel electrophoresis, and end-labeled using terminal deoxynucleotidyl transferase (Affymetrix) for 2 hr at 37°C. The samples were then added to hybridization buffer containing 5% DMSO (Sigma, St. Louis, MO), 5.77 mM EDTA (Ambion, Austin, TX), 0.115 mg/mL herring sperm DNA (Promega, Madison, WI), 11.5 mg/mL human Cot-1 (Invitrogen, Carlsbad, CA), 0.0115% Tween-20 (Pierce, Rockford, IL), 2.69 M tetramethyl ammonium chloride (Sigma), 0.056 M MES (Sigma), 2.5 × Denhardt's solution (Sigma), and an oligonucleotide control reagent (Affymetrix), denatured at 95°C for 10 minutes, and hybridized to the 50K SNP Chip for 18 hr at 48°C. Finally, the SNP Chips were washed and stained with streptavidin (Pierce), biotinylated goat anti-streptavidin antibody (Vector Laboratories, Burlingame, CA), and streptavidin phycoerythrin (Molecular Probes, Eugene, OR). The arrays were scanned in the UCLA Microarray Core Facility and analyzed using GCOSS and GDAS Affymetrix software and copy number was determined using the Chromosome Copy Number Analysis Tool (CNA; Affymetrix) and the Copy Number Analyzer for GeneChip (CNAO) [Nannya et al., 2005] programs.

**Analysis of Copy Number Using Circular Binary Segmentation**

We analyzed copy number using circular binary segmentation (CBS) by comparing the subject to a single control individual. The
signal intensity at each feature on the XbaI array was normalized between the subject and control by first dividing all signals by the mean signal for that chip, expressing the result as a fraction of the mean signal (fractional signal intensity). The set \( \lambda_i \) of log2 ratios were generated by comparing the patient’s raw signal intensities at marker \( i \) with those of a reference marker. Further normalization for each feature was obtained by a linear regression of the log2 ratio as a function of the average fractional signal intensity, as follows.

The mean fractional signal intensity for the \( i \)th marker, \( \lambda_i \), was estimated as the average signal intensity for that marker from the patient and reference arrays. The difference in signal intensity between patient and reference for the \( i \)th marker (\( \lambda_i \)) was plotted as a function of mean signal intensity for each marker. Ordinary least squares (OLS) regression was performed to express \( \lambda_i \) as a function of average probe intensity. The revised log2 ratio was then expressed as the residual for each feature after linear regression. We defined the revised log2 ratios, \( N_i \), as the residuals following OLS regression: \( N_i = \lambda_i - \text{OLS}(\lambda_i) \). This helped to reduce the variation due to feature specific hybridization differences.

In a similar fashion, linear regression of the log2 ratio as a function of the length of the specific probe-set fragment was performed and the log2 ratio was again revised and expressed as the residual. A third regression was performed using GC content and the log2 ratio then again revised. These regressions also helped reduce variation due to hybridization differences. There were no practical differences noted between doing sequential univariate regressions and a single multivariate regression.

The estimate of relative log2 ratio was estimated using the CBS method of Olshen and Venkatraman [2004]. After normalization, the adjusted log2 ratios \( N_i \) were then used in the CBS algorithm (DNAcopy package, www.bioconductor.org), which uses a change point model to estimate changes in copy number [Olshen and Venkatraman, 2004]. CBS computes a likelihood ratio statistic testing the null hypothesis that there is no copy-number change against the hypothesis that one or multiple changes occur at any point along the chromosome. Once the null hypothesis is rejected, the algorithm recursively identifies all changes where the likelihood ratio exceeds the critical value [Olshen and Venkatraman, 2004].

The CBS algorithm returns an R data frame containing a list of the log2 ratios throughout the genome. Most regions will have a log2 ratio of approximately zero, indicating no difference in copy number between the subject and the reference sample. The entire genome, excluding the Y chromosome, was observed for any significant changes in copy number. The CBS algorithm will occasionally return changes that involve only two or three markers, or regions in which the reported log2 ratio is only slightly different from zero. These small, two to three marker changes did not appear to be reliable, as they often had extremely high or low log2 ratios, and may represent occasional copy-number change calls due to random noise. For this reason, we manually excluded these regions from further analysis. In addition, because we were interested primarily in interstitial deletions, we made the assumption that the log2 ratio for a hemizygous deletion would be close to −1. This helped to eliminate areas identified by the segmentation algorithm in which copy number was only slightly different from baseline.

The minimum deviation from the baseline log2 ratio of zero required to consider a copy-number change was determined by comparing the baseline variance in log2 ratio with the variance in regions of known copy change, as further explained in the results, and was determined to be >0.3 or <−0.3. While a true region of copy-number change could exist with a smaller log2 ratio than this, it would likely represent heterogeneity in the cell population under analysis. We assumed for the purposes of this analysis, since we are analyzing genomic DNA from a single individual, that all cells would share the same copy-number changes. This assumes that there is no mosaicism in the individual. This is consistent with the FISH and karyotype results for these individuals, but may not necessarily be assumed in general clinical practice. Further work with mosaic individuals will be required to determine the influence of this phenomenon.

All data analysis was performed using MySQL and R (www.mysql.com; www.cran.org).

### RESULTS

The normalized SNP intensity reading for each probe set was compared to a large Affymetrix reference set of normal individuals using CNAT software and the relative genomic content was determined. A log2 ratio was calculated using a bin size of 0.5 MB along with the genotype call at each SNP (Table 1). By using the log2 of the ratio between the reference and sample, we approximate the copy number at that locus. A marked decrease in the log2 ratio to less than 1.5 and greater than 1.0 was observed for specific regions in the DGCR, located from 17–20 MB on chromosome 22, for each of the three cell lines and the DiGeorge patient (Fig. 1). We calculated the average log2 ratio for the entire chromosome 22 as well as for the DGCR deleted region and compared this to results from two normal individuals (Table 1). The chromosome 22 normalized copy number for the cell lines in the 17–20 MB deleted region was approximately 1.2, while in unaffected controls it was approximately 1.8 and 1.9, and in the patient with DiGeorge syndrome it was 1.0.

We evaluated the same set of data under multiple different algorithmic approaches to determine if other methods would make discrimination of signal from noise easier. Two algorithmic approaches were investigated. 1) CNAG [Nannya et al., 2005], uses a normalization procedure followed by binning as before to reduce noise in the sample and call copy number changes. 2) We also used our own modification of existing algorithmic approaches, consisting of a normalization procedure similar to that used in CNAG, but followed instead by CBS, a change-point method in which the likelihood of a change in copy number at each marker is calculated (Fig. 2) [Olshen and Venkatraman, 2004]. All three algorithmic approaches—CNAT, CNAG, and CBS—were

<table>
<thead>
<tr>
<th>Sample</th>
<th>Control 1</th>
<th>Control 2</th>
<th>GM07215</th>
<th>GM07939</th>
<th>GM13325</th>
<th>DiGeorge Patient 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average chromosome 22 copy number</td>
<td>2.43</td>
<td>2.43</td>
<td>2.14</td>
<td>2.27</td>
<td>2.11</td>
<td>2.34</td>
</tr>
<tr>
<td>Average chr 22:17–20 MB copy number</td>
<td>2.32</td>
<td>2.21</td>
<td>1.27</td>
<td>1.37</td>
<td>1.30</td>
<td>1.18</td>
</tr>
<tr>
<td>Chr 22 Normalized Copy Number in chr 22:17–20 MB</td>
<td>1.91</td>
<td>1.82</td>
<td>1.19</td>
<td>1.21</td>
<td>1.23</td>
<td>1.01</td>
</tr>
</tbody>
</table>

*base pairs numbered from 1 at 22p telomere.*
compared in their ability to detect the del22q11.2 deletion in our set of patients (Fig. 3).

The CNAG software [Nannya et al., 2005] improved the signal-to-noise ratio by compensating for length and GC content of PCR products and reducing noise with binning. However, it was still necessary to manually identify regions of reduced copy number, and though the deletion was clear, some ambiguity in the exact breakpoint location existed. Figure 3B shows the genome-wide log2 ratio and smoothed data for the CNAG method, with smoothed data in blue. Note that though noise is considerably reduced there remains a slope to the breakpoint areas that obscures the actual location. Also, manual identification of copy number changes would not be desirable in a clinical laboratory setting, as it would increase technician time and expertise and reduce standardization.

In contrast, the change point method used by CBS clearly and automatically identified regions of copy change (Figs. 2 and 3C). While the accuracy in larger numbers of samples will need to be determined, and the precision of breakpoint estimation evaluated, the change-point method correctly identified the del22q11.2 deletion in all of our patients and provided a clear estimate of the actual breakpoint, unlike the other methods investigated. The mean log2 ratio for the del22q11.2 deletion was $-1.01 \pm 0.07$, while the median log2 ratio overall was $-0.011$ and fifth to 95th percentiles were $-0.17$ to $0.22$. This suggested a reasonable threshold for determining a copy number change to be 0.3.

As mentioned in the Methods section, we manually inspected the CBS output to exclude suspect calls, such as those involving only two to three markers or those deviating only slightly from baseline (i.e., less than the threshold of 0.3). The data in the figures is presented without any of this manual editing. This could conceivably be automated though, for example, through the use of a hidden Markov model. With the XbaI chips, there was a median of two copy number variations (CNVs) per genome though, and automation of this process was not felt to be necessary in this setting. With higher density chips, such as the 500K arrays now available, smaller CNVs will be detected and automation may well be useful there.

DISCUSSION

Successful identification of the 22q11.2 deletion in the three Coriell cell lines and patient’s blood samples provides proof-of-principle for use of the Affymetrix 50K SNP chip to detect a single copy loss of DNA. We have previously demonstrated that it is possible to detect a deletion in Xp21 in boys with complex glycerol kinase deficiency, illustrating the ability of this technique to detect and map a complete loss of genetic material [Stanczak et al., 2006]. This investigation provides evidence that the Affymetrix 50K SNP Chips can also be used to detect a 50% loss of DNA. These data suggest that other copy number aberrations, such as gene duplications, could also be determined, as has been shown for other arrays [Wong et al., 2004; Nannya et al., 2005].

Since every array analysis method has limitations, it is advantageous to use multiple approaches. While analysis of each sample with multiple different software programs can facilitate the identification of otherwise missed aberrations or denial of false alterations, FISH or another methodology would still be necessary.
for confirmation. In addition to the method of analysis, the detectable deletion size is depends upon the density of SNPs in the region and the threshold magnitude that is employed by the algorithm. Since CNAG averages log2 ratios over one to 10 adjacent SNPs, the size of detectable deletions will vary across the genome as a function of the SNP density in the array, variations in array manufacture, and the algorithmic approach that is used [Nannya et al., 2005].
Oligonucleotide mapping arrays such as those used in this study have the advantage over FISH of providing genomewide determination of copy-number change, and in addition define SNP alleles at loci genomewide. These two features could be extremely useful in analyzing phenotypic variability in DiGeorge syndrome, as well as helping to define the role that some of the other deletions identified in association with DiGeorge syndrome may play. Work with ROMA and aCGH has demonstrated that CNVs are common in all individuals, increasing the chance that the previous alternative deletions identified in del22q11.2 syndrome may represent benign polymorphisms.

ROMA and aCGH technology, if sufficiently reliable and interpretable, would be of major benefit in the cytogenetics laboratory. However, clinical interpretation of the data is complicated by the observations that CNVs are so common [Iafrate et al., 2004; Sebat et al., 2004; Feuk et al., 2006; Wirtenberger et al., 2006]. Nevertheless, the ability to screen clinically for pathologic copy-number changes genomewide would likely improve the sensitivity of cytogenetic screening and as methods become less expensive might even permit mass screening of newborns as part of the normal newborn screening process.

del22q11.22 Syndrome is the most common interstitial deletion in the clinical setting, and our results suggest that ROMA is capable of reliably detecting these deletions. Change-point models appear to provide the most easily interpreted and standardized output of the algorithms investigated. Further work should focus on the development of change-point models to enable the greatest opportunity for ROMA to become clinically useful.

REFERENCES


