A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays

Yasuhiro Nannya,1 Masashi Sanada,1 Kumi Nakazaki,1 Noriko Hosoya,1 Lili Wang,1 Akira Hangaishi,1 Mineo Kurokawa,1 Shigeru Chiba,1,2 Dione K. Bailey,1 Giulia C. Kennedy,1 and Seishi Ogawa1,3

1Departments of Hematology/Oncology, 2Cell Therapy/Transplantation Medicine, and 3Regeneration Medicine for Hematopoiesis, Graduate School of Medicine, University of Tokyo, Tokyo, Japan; 1Affymetrix, Inc., Santa Clara, California; and 4Core Research for Evolutional Science and Technology of Japan Science and Technology Corporation, Saitama, Japan

Abstract

We have developed a robust algorithm for copy number analysis of the human genome using high-density oligonucleotide microarrays containing 116,204 single-nucleotide polymorphisms. The advantages of this algorithm include the improvement of signal-to-noise (S/N) ratios and the use of an optimized reference. The raw S/N ratios were improved by accounting for the length and GC content of the PCR products using quadratic regressions. The use of constitutional DNA, when available, gives the lowest SD values (0.16 ± 0.03) and also enables allele-based copy number detection in cancer genomes, which can unmask otherwise concealed allelic imbalances. In the absence of constitutional DNA, optimized selection of multiple normal references with the highest S/N ratios, in combination with the data regressions, dramatically improves SD values from 0.67 ± 0.12 to 0.18 ± 0.03. These improvements allow for highly reliable comparison of data across different experimental conditions, detection of allele-based copy number changes, and more accurate estimations of the range and magnitude of copy number aberrations. This algorithm has been implemented in a software package called Copy Number Analyzer for Affymetrix GeneChip Mapping 100K arrays (CNAG). Overall, these enhancements make CNAG a useful tool for high-resolution detection of copy number alterations which can help in understanding the pathogenesis of cancers and other diseases as well as in exploring the complexities of the human genome. (Cancer Res 2005; 65(14): 6071-9)

Introduction

Genome-wide detection of copy number alterations has been drawing increasing attention in the field of cancer research as well as in the diagnosis of rare congenital disorders (1–3). In addition to studying genomic alterations in disease, the recent discovery of large-scale copy number variations in the genome of humans (4, 5) has stimulated interest in elucidating their role in the evolution of the human genome (4, 5).

The initial approach to genome-wide detection of copy number changes was comparative genomic hybridization (CGH; ref. 6). This approach first enabled the exploration of genetic alterations in cancers across the human genome at ~ 20 Mb resolution, and was later improved to less than 1 Mb resolution by replacing target metaphase spreads with a large number of discrete genomic or cDNA clones in arrays (array-based CGH; refs. 1–3). Further increase in resolution was obtained by arrays consisting of 32,433 BAC clones spanning the entire human genome, resulting in comprehensive analyses of cancer genomes at less than 100 kb resolution (7).

Another recently described approach to genome-wide copy number detection is the use of synthetic high-density oligonucleotide microarrays (8–15). These commercially available microarrays, designed to genotype 10,000, 50,000, or 100,000 single-nucleotide polymorphisms (SNP) in human genomic DNA, provide attractive alternatives to BAC-array CGH (16–19). The Affymetrix GeneChip Mapping 100K high-density oligonucleotide arrays contain 116,204 SNPs with a mean spacing of 23.6 kb, and a potential to provide the highest resolution of copy number detection currently available (19).

These arrays provide clear technical advantages including robust single-primer assay methodology, accurate and reproducible genotyping, and copy-neutral loss of heterozygosity (LOH) detection compared with array-CGH, karyotyping, and other oligonucleotide CGH arrays. Furthermore, these arrays are manufactured under stringent quality control procedures. However, there are a limited number of algorithms/software available and they suffer from certain limitations which may result in false-positive and false-negative estimations. For example, the software dChipSNP6 requires the use of a paired-normal sample, which is often unavailable, to perform the analysis (10, 20). Furthermore, dChipSNP can currently only process the Mapping 10K array and not the 50K or 100K arrays. Another alternative software/algorithm for detecting copy number changes from the Mapping 100K arrays is Chromosome Copy Number Analysis Tool (CNAT).7 CNAT uses a set of 110 normal reference individuals and thus overcomes the paired-normal requirement but does not account for experimental variation in the samples being compared (9).

Here we present an improved algorithm for copy number detection using Affymetrix GeneChip Mapping 100K arrays. This algorithm improves signal-to-noise (S/N) ratios by reducing variation in the raw signal ratios and optimizes the selection of the reference. In addition, availability of accurate genotyping information further enables LOH inference and allele-based copy number analysis. This combination of >100,000 SNP markers and

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/). Requests for reprints: Seishi Ogawa, Department of Regeneration Medicine for Hematopoiesis, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. Phone: 813-3815-5411, ext. 35609; Fax: 813-5804-6261; E-mail: sogawa-tky@umin.ac.jp.

©2005 American Association for Cancer Research.
Materials and Methods

Samples. We obtained the lung cancer cell lines CRL-5929, CRL-5872, and their immortalized lymphoblast CRL-5969, and CRL-5958, from American Type Culture Collection (Rockville, MD), and glioma cell line U251, and immature T-cell line HPB-ALL from Riken (Tsukuba, Japan). Peripheral blood samples and primary tumor specimens were collected from 96 normal volunteers and 33 patients with hematologic malignancies, and subjected to DNA extraction. Informed consent was obtained from the subjects according to the protocols approved by the Institutional Review Boards of the University of Tokyo.

Affymetrix platform. Array experiments were done according to the standard protocols for Affymetrix GeneChip Mapping 100K arrays (Affymetrix, Inc., Santa Clara, CA). Briefly, total genomic DNA was digested with a restriction enzyme (XbaI or HindIII), ligated to an appropriate adapter for each enzyme, and subjected to PCR amplification using a single primer. After digestion with DNase I, the PCR products were labeled with a biotinylated nucleotide analogue using terminal deoxynucleotidyl transferase and hybridized to the microarray. Hybridized probes were captured by streptavidin-phycocerythrin conjugates and the array was scanned and genotypes called as described (18).

Compensation of raw signal ratios and construction of the best-fit reference. Sum of signals from the 10 perfect match probes for the A allele (PA) and those for the B allele (PB) is normalized for each SNP. The mean signal intensity of all autosomal SNPs becomes the same for the two arrays being compared. Relative copy number at the ith SNP in sample 1 (Si1) and sample 2 (Si2), A1;i;i = log2(Si1/Si2), is the number of terms to be averaged and arbitrarily set to 3 to 10. For proper comparison with our system, log 2–converted values of the Genome Smoothed Analysis Copy Number (GSA_CN) were compared. FISH analyses were done as previously described (15). Probe information used in FISH analysis is supplied in Supplementary Table 2. Copy Number Analyzer for Affymetrix GeneChip mapping 100K arrays. Copy Number Analyzer for Affymetrix GeneChip Mapping 100K arrays (CNAG) ver 1.0 is the implementation of the set of algorithms described above that is written in C++ for Microsoft Windows. All the examples of copy number analysis with Affymetrix GeneChip Mapping 100K arrays presented in this article were done using CNAG. All data used in this article and CNAG can be downloaded.8

Allele-based analysis using a paired normal sample. When a paired normal sample is used for the allele-based copy number estimate, analysis is confined to those SNPs of which genotype in the reference is AB (heterozygous) and signal sums of PAs (SjA) and PBs (SjB) for the ith SNP are taken separately both in the tumor and the reference. The log 2 ratios of normalized values, $S_{iA}$ and $S_{iB}$, are similarly compensated for the different experimental conditions to calculate the corrected log 2 ratios, $A_{iA;REF}$ for PAs and $A_{iB;REF}$ for PBs, where the coefficients used for the normalization and compensation procedure are determined using all the SNPs in a region specified as diploid as is the case with analysis using references not derived from the same individual. For the purpose of allele-based copy number analysis, the corrected log 2 ratios for the heterogeneous SNPs are separated in two groups,

$$A_{iA;max} = \{ x | \max \{ c \cdot A_{iA;ref} (A), c \cdot A_{iB;ref} (B) \} \}$$

$$A_{iB;max} = \{ x | \min \{ c \cdot A_{iA;ref} (A), c \cdot A_{iB;ref} (B) \} \}$$

To obtain the allele-based copy number view, the members of each group are rearranged in the genetic order and plotted as shown in Fig. 5 (middle).

Copy number inference. Although compensation for difference in PCR conditions and optimization of the reference significantly reduce the deviation of log 2 ratios, further elimination of random noise greatly helps visualization of copy number changes along the chromosomes and efficient copy number alterations. In "local mean analysis," the sequence of the mean log 2 ratios, $A_{i;test, REF} = 1/k \sum_{i=k+1}^{i+k} A_{i;test, REF}$, is calculated, where k is the number of terms to be averaged and arbitrarily set to 3 to 10, according to the SD values after compensations of log 2 ratios and optimization of nonredundancies for diploid SNPs. In the alternative analysis using a hidden Markov model (21), the inference of copy number is more efficient and automated, in which a real state of the copy number sequence (a hidden state) along a chromosome is inferred from the observed sequence $A_{i;test, REF}$ as the state of maximum likelihood that is calculated from the state transition load and the probability of the hidden state to "emit" the observed sequence of log 2 ratios, using the Viterbi algorithm. We assumed that copy number change (state transition) is the result of a genetic recombination event between the two adjacent SNP loci, and Kosambi’s map function (1/2tanθh(2θ)) is used to transform the genomic distance, or recombination fraction between the two SNPs (θ) to state "transition probability," where θ is expressed in cM units; for simplicity, 1 cM should be 1 Mbp. The observed log 2 ratio is assumed to follow the normal distribution according to real copy number states, which gives the "emission probability." The variables of normal distribution were empirically determined from the experimental data (Supplementary Methods).

Comparison and confirmation of the results. CNAT was executed following the instructions of the suppliers. For proper comparison with our system, log 2–converted values of the Genome Smoothed Analysis Copy Number (GSA_CN) were compared. FISH analyses were done as previously described (15). Probe information used in FISH analysis is supplied in Supplementary Table 1. PCR primers were designed to amplify several adjacent fragments that are within and outside of the homozygously deleted regions in tumor samples. Primers and PCR conditions are provided in Supplementary Table 2.

8 http://www.genome.umin.jp
Results

Reduction of signal-to-noise ratio. To allow for high-fidelity copy number estimation from genotype arrays, we evaluated several variables that influence signal intensity between arrays being compared. By taking into account these variables, erroneous copy number changes can be eliminated. Initially, the raw signal intensities between normal DNA samples were compared. An arithmetic sum of the hybridization signals from all perfect match probes for each SNP was normalized so that the mean signal intensity of all autosomal SNPs becomes the same for the two arrays. Next, SDs of log 2 ratios of the normalized signals were calculated for each SNP locus and compared for

Figure 1. A, SDs of log 2 ratios for all autosomal SNPs were calculated for all combinations of array data from 96 normal individuals before (left matrix) and after (right matrix) compensations, and color coded according to the SD values. The array data are arranged in the order of experiments. Significant variations in SD values are found between different array data before compensations, whereas combinations of samples that were processed together tend to form clusters showing low SD values. After compensations, the variations in SD values among different combinations become less conspicuous. Log 2 ratios corresponding to the indicated cell before and after compensations are plotted in the chromosomal order (bottom). B, compensations of log 2 ratios for dissimilar PCR conditions between test (sample 103) and reference (sample 308) arrays. Log 2 ratios are plotted against length (right) and GC content (left) of the restriction fragment to which each SNP belongs, before compensations (top), after the first compensation for fragment length (middle), and after the second for GC content (bottom). Dependence of log 2 ratios on fragment length and GC content is serially resolved after compensations with the consequent reduction in the SD value. C, effects of compensations on SD. SD values for varying combinations of a fixed array [sample 110, blue rectangle in A (left)] with different arrays are significantly reduced after compensations. D, effects of using the best-fit multiple references on reduction in SD values. The mean SD values were calculated using an increasing number of the best-fit multiple references for 96 normal samples and plotted with SEs against the number of references averaged.
Figure 2. A, mean SD value of log 2 ratios and 25th and 75th quartiles are calculated for all combinations of 96 normal samples, before (raw log 2 ratios; lane A) and after compensations (lane B) for dissimilar PCR conditions ($N = 96C_2$). The calculations were also done after using the five best-fit references (lane C). B, genomic DNA from a glioma cell line U251 was analyzed for copy number alterations with or without compensations for different experimental conditions using different compensation algorithms. The raw log 2 ratios in the chromosomal order show widely fluctuated baselines (1; SD = 0.365), which are greatly reduced after compensation of experimental conditions and using the best-fit single (2; SD = 0.222) and multiple references (3; SD = 0.118). Regions of homozygous and hemizygous deletion are indicated by small and large arrows. The SD value is further decreased (SD = 0.114) when the second round compensation procedures are applied by specifying a region of diploid alleles (4; red bar). With this approach, diploid regions as well as other regions showing varying ploidy states are correctly assigned to their real ploidy. For proper comparison, SD values are calculated for the SNPs within the diploid range (red bar). C, SD values were also calculated and plotted for 33 tumors samples compared with a representative normal sample (sample 101) before (lane A) and after compensations using a single best-fit reference (lane B), after analysis with the average of the five best-fit references (lane C). The regression curves for compensations were determined using all autosomal SNPs (lanes A-C) or exclusively using those within a diploid region arbitrarily selected by visual inspections of the copy number output from previous steps (lane D). After a series of compensations and optimization of references, the average of log 2 ratios of every set of consecutive five (lane E) or ten (lane F) SNPs in the genetic order was taken to further reduce persisting background noises (local mean analysis; see Materials and Methods). SD values calculated using a paired normal reference with no local averages are also provided (lane G).
every combination of two arrays from 96 normal individuals (see Materials and Methods). This resulted in a considerable difference in SD values among different combinations of samples (Fig. 1A (left) and Supplementary Fig. 1). We hypothesized that different PCR kinetics among experiments could be the origin of these highly variable SD values and that adjusting for these experimental variables may allow for more accurate comparisons (Supplementary Methods). It should be noted from the block-like structures of the SD graph that the behavior of SD values is largely determined by the group of experiments done together, rather than the individual experiments.

To assess the dependence of the log 2 ratios of signals on PCR kinetics, we plotted log 2 ratios against several variables that may affect PCR reactions in different experiments, including length of PCR fragments and GC content (Fig. 1B). By compensating for the effects of these variables using two-step quadratic regressions, the log 2 ratios almost completely offset the dependencies of log 2 ratios on these variables and the SD value between the given two experiments is reduced from 0.88 to 0.40 (Fig. 1B and C). In fact, these compensation procedures dramatically improve the distribution of SD values among 96 normal samples, enabling comparisons across different samples or groups of experiments (Fig. 1A, right). SNPs located on fragments shorter than 500 bp, which account for 2.9% of total SNPs, tend to be resistant to these compensations and were precluded from the analyses altogether.

Optimization of reference selection. Although correcting for the fragment length and GC content dramatically reduces SD values, there was still considerable variation among different samples. Therefore, we explored the effects of using different reference sets. The type of reference set used largely depends on the following: (a) cases where tumor and normal DNA are available from the same patient (paired normal samples) and (b) cases where only the test or tumor sample is available. Use of a paired normal sample coupled with compensation for variability across experimental conditions generally gives lower SD values. However, the benefit of having a paired normal sample for a reference tends to be diminished when test and reference samples are processed separately (data not shown). The case when a paired normal sample for a reference is unavailable is more complicated. Although the array that shows the lowest SD value for the test sample is a candidate for the reference, we investigated choosing the average of multiple samples with the lowest SD values or best-fit samples as a reference.

To determine the effect of averaging multiple references on SD values, data from 96 normal samples were compared with the averages of varying numbers of the best-fit references, and SD values were calculated for each comparison (Fig. 1D). As the number of samples increases to five, the SD value gradually decreases to <0.15

Figure 3. Analysis of CRL-5929 using five best-fit references and the whole genome result. Red dots, raw log 2-ratio values for each SNPs. Copy number inferences from a hidden Markov model (green lines) and local mean analysis for five consecutive SNPs (blue curves) are appended. Heterozygous SNP calls are marked with green bar lines beneath the chromosomes.
for most normal samples and then reaches a plateau, suggesting that taking an average of at least five best-fit samples is sufficient to optimize the comparison (Fig. 2A).

A direct comparison of the copy number estimation using a paired normal reference and best-fit reference was also evaluated for a small cell carcinoma cell line (CRL-5929). The results using a paired normal reference show a slightly lower SD value (0.141 versus 0.177) and smaller fluctuations of baselines than with the best-fit multiple references (Supplementary Fig. 2A). However, copy number inference from the hidden Markov model (see Materials and Methods) analysis and local averaging gives an equivalent result (Supplementary Fig. 2B). This suggests that the multiple best-fit references are an excellent substitute in cases where a constitutional paired normal reference is not available.

**Validation of new algorithm using normal and tumor samples.** We evaluated the performance of this new algorithm by analyzing a variety of samples. Figure 2B shows a representative analysis of a glioma specimen, showing the dramatic reduction of baseline fluctuations or SD values when using this algorithm relative to raw log 2 ratios. The raw data (Fig. 2B-1) has an SD of 0.365, which is reduced to 0.222 after applying a single best-fit reference (Fig. 2B-2). A further improvement was seen when using multiple best-fit references, SD = 0.118 (Fig. 2B-3). Regions of homozygous and hemizygous deletion are indicated by small and large arrows, respectively. As tumor samples frequently exhibit complex chromosome abnormalities with extensive genetic imbalances (22), analysis of tumor specimens requires additional considerations in selection of references. SD values may be disproportionately inflated or overwhelmed by the effect of genetic abnormalities when all autosomal signals are included in the calculation of SD values. To circumvent this problem and to calculate proper SD values for reference selection, we adopted a two-step approach. We first made a tentative estimation of the copy number changes by using all autosomal signals and predicted the regions that are diploid by reference to other independent information (FISH, PCR, or results in the literature). In the second step, one of these diploid regions was used for normalization and for calculation of SD values to identify the best-fit references. This final step further improved the SD value, 0.114 (Fig. 2B-4).

Reductions of SD values in different tumor samples are summarized in Fig. 2C. The SD values from 33 tumor samples were calculated from raw data, as well as using the new algorithm with five best-fit references. The distribution of SD values is clearly decreased using the new algorithm. Figure 2C also illustrates the results using limited regional SNPs, local averaging of five or ten consecutive SNPs, and using a paired normal reference. After all corrections and optimizations of best-fit reference samples, a local mean procedure was applied which results in lower SD values relative to the paired normal reference sample, where no local mean was used. All of these analyses resulted in a dramatic reduction in the SD values relative to the raw data, which enables genome-wide copy number detection with a high degree of accuracy (Fig. 3).

**Sensitivity of copy number analysis using improved algorithm.** We evaluated the sensitivity of our copy number detection algorithm by calculating mean log 2 ratios for those regions having known copy number alterations. The observed

---

**Table 1. Comprehensive list of homozygous deletions and gains in tumor cell lines**

<table>
<thead>
<tr>
<th>Sample*</th>
<th>Change</th>
<th>Length (kb)</th>
<th>Cytoband</th>
<th>SNPs involved</th>
<th>Genes involved</th>
<th>Confirmation$^{a}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRL-5929</td>
<td>Deletion</td>
<td>5.5-76</td>
<td>5q13.1</td>
<td>4</td>
<td>PIK3R1</td>
<td>PCR (ID1)</td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>2,233-2,652</td>
<td>8q12.2-8q12.3</td>
<td>146</td>
<td>ASPH, GGH, others</td>
<td>FISH (RP11-234F8)</td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>1,347-1,475</td>
<td>8q24.21</td>
<td>83</td>
<td>MYC</td>
<td>FISH (RP11-145G10)</td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>1,662-1,771</td>
<td>11q14</td>
<td>113</td>
<td>ME3, SYTL2, others</td>
<td>FISH (RP11-90K17)</td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>566-627</td>
<td>12p13.33</td>
<td>23</td>
<td>CACNA1C, DCP1B, others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>825-919</td>
<td>12p13.32</td>
<td>37</td>
<td>CCND2, PARP11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>2,720-2,780</td>
<td>12p13.31</td>
<td>31</td>
<td>LAG3, GAPD, others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>693-888</td>
<td>12p12.1</td>
<td>39</td>
<td>SLC01, IAPP, others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>1,925-2,090</td>
<td>12p11</td>
<td>115</td>
<td>PTHLH, MRPS35, others</td>
<td>FISH (RP11-74F4)</td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>609-648</td>
<td>14q11.2</td>
<td>48</td>
<td>CCNBD1P1</td>
<td>FISH (RP11-98N22)</td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>1,770-1,918</td>
<td>14q11.2</td>
<td>93</td>
<td>SALL2, METTL3, others</td>
<td>FISH (RP11-70F9)</td>
</tr>
<tr>
<td></td>
<td>Deletion</td>
<td>0.9-29</td>
<td>15q23</td>
<td>3</td>
<td></td>
<td>PCR (ID2)</td>
</tr>
<tr>
<td>CRL5872</td>
<td>Gain</td>
<td>8,763-10,224</td>
<td>1q21.2-1q22</td>
<td>92</td>
<td>BCL9, IL6R, others</td>
<td>FISH (RP11-71L20)</td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>1,703-2,814</td>
<td>5p15.33</td>
<td>41</td>
<td>IRX4, NDUF56</td>
<td>FISH (RP11-20B3)</td>
</tr>
<tr>
<td></td>
<td>Gain</td>
<td>1,183-2,113</td>
<td>8p11.22-8p11.21</td>
<td>43</td>
<td>INDO, ADAM2</td>
<td>FISH (RP11-26D23)</td>
</tr>
<tr>
<td></td>
<td>Deletion</td>
<td>54-131</td>
<td>9p23</td>
<td>9</td>
<td></td>
<td>PCR (ID4)</td>
</tr>
<tr>
<td></td>
<td>Deletion</td>
<td>316-404</td>
<td>9p23</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deletion</td>
<td>391-556</td>
<td>9p22.3</td>
<td>14</td>
<td>CER1, FBM1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deletion</td>
<td>1,498-1,684</td>
<td>9p22.2</td>
<td>42</td>
<td>SIEGL2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deletion</td>
<td>1,237-1,457</td>
<td>9p21.3</td>
<td>44</td>
<td>IFNA, CDKN2A, others</td>
<td>PCR (ID5)</td>
</tr>
</tbody>
</table>

NOTE: Homozygous deletions (n = 0) and high grade gains (n > 4) were identified using a hidden Markov model.

*All samples are compared with paired normal references to distinguish between tumor-specific copy number alterations and copy number polymorphisms.

$Minimial size range is the distance of both ends of consecutive SNPs that meet the criteria. Maximal size range is the distance of the SNPs that flank the deleted or amplified regions. Based on UCSC Build 16 (hg16) assembly (http://genome.ucsc.edu/cgi-bin/hgGateway).

$The probe names used in FISH analysis and the primer set IDs used in PCR analysis are shown in parentheses. See Supplementary Tables 1 and 2 for details.
mean log 2 ratio for X chromosomes between 48 males and 48 females was $-0.49$ (SD = 0.035), which is about half of the expected value ($= -1$) for the single copy difference (Supplementary Fig. 3). Similarly, the observed mean log 2 ratio for known trisomy in CRL-5929 in chromosome 209 was 0.36, compared with the expected value of 0.585. These discrepancies are thought to result from nonspecific background hybridization, which could not be measured and subtracted before taking log 2 ratios, although the background factors can be theoretically estimated and used for performing copy number inference using a hidden Markov model (Supplementary Method). Because the SD value of log 2 ratios for baseline diploid SNPs is 0.18 ± 0.03 with best-fit references or lower with a paired normal reference, the estimated S/N ratio is ~2.0 (for trisomy) or more (for monosomy) in most cases.

Estimation of the resolution is directly proportional to the distribution of SNP markers. We determined the size of genomic alterations in CRL-5929 and CRL-5872 cell lines (Table 1). Deletions of less than 500 kb at 14q11.2 (T-cell receptor $\alpha$) and less than 393 kb at 7q34 (T-cell receptor $\gamma$) were observed in the immature T-cell line HPB-ALL; these were caused by a T-cell receptor rearrangement (23). The genomic regions comprising these deletions contain approximately 20 and 10 SNPs, respectively, from the Affymetrix GeneChip Mapping 100K arrays. As the number of SNP markers...
increases, the accuracy of size determinations increases and even smaller deletions are likely to be detected.

Finally, we compared BAC-CGH arrays, CNAT, and CNAG. BAC-arrays containing 3,151 FISH-mapped BAC/PAC clones with an average resolution of 1.0 Mb were constructed. The CRL-5929 cell line was examined using the BAC-array (Fig. 4A, top), CNAT (Fig. 4A, middle), and CNAG (Fig. 4A, bottom). The resolution on the SNP array allows for much finer mapping than BAC arrays, and additional compensations applied in CNAG provide more distinct copy number detection than CNAT. FISH analysis showed complete concordance with this method (Fig. 4A, bottom). This is more clearly seen in the immature T-cell line (HPB-ALL) on chromosome 2 (Fig. 4B). In this case, a homozygous deletion in 2p16.3 (red arrow) was detected by the 100K SNP array (Fig. 4B, right) but not detected by the BAC-array or CNAT (Fig. 4B (left), top and bottom, respectively). Lack of PCR product designed within the deleted region in HPB-ALL verified the deletion (Fig. 4B, right).

**Allele-based copy number analysis.** Another potential benefit from using a paired normal reference is the allele-based copy number analysis. In this approach, copy number estimates are done separately for each of the heterozygous SNPs (Fig. 5). The ratios can be separately defined for each of the alleles, grouped together into the larger (red) and smaller (green) ones and plotted in chromosomal order (Fig. 5; see Materials and Method). Although the analysis is confined to only heterozygous SNPs (mean heterozygosity of the Affymetrix GeneChip Mapping 100K SNPs is 0.29), it can effectively unmask regions showing copy neutral LOH due to genetic imbalances (asterisks in Fig. 5, middle), which are not detected by non-allele-based analysis (Fig. 5, top) or conventional BAC-array CGH.

**Discussion**

Recently, synthesized oligonucleotide microarrays have been used as alternatives to conventional BAC-array CGH for genome-wide copy number analysis (8–15). Large numbers of SNP markers are expected to enable ultra high-resolution copy number analysis; however, currently available algorithms do not take into account variation across different experiments, which can result in low S/N and high SD values. To address this problem, we developed a novel algorithm for Affymetrix GeneChip Mapping 100K arrays, which includes robust compensations for systematic deviations of raw signal ratios across different experimental conditions and optimizes selection of the best-fit references. Together, these features effectively reduce heterogeneity between experiments and improve the SD value of log 2 ratios from 0.67 ± 0.12 to 0.18 ± 0.03, which is considerably lower than values reported for

![Figure 5. Copy number views of allele-based (middle) and non-allele-based analyses (top and bottom) of the same tumor samples as in Fig. 3 using the local mean analysis (k = 5). In the allele-based copy number view, the larger (4max, red lines) and the smaller (4min, green lines) log 2 ratios of the two log 2 ratios for allele A and allele B signals are separately grouped. Regions having copy neutral LOH as well as monosomy can be inferred from disappearance of heterozygous SNP calls in this non-allele-based analysis (bottom). Whereas no copy number changes are observed in both chromosomes 15 and 16, chromosome 15 shows copy neutral LOH as clearly seen in the allele-based copy number view (middle). Note that the existence of copy neutral LOH in chromosome 15 is inferred from the disappearance of heterozygous SNP calls in non-allele-based analysis (bottom).](image-url)
oligonucleotide CGH-arrays (24). The SD value further decreases to 0.10 when log 2 ratios are locally averaged for consecutive five to ten SNPs (Fig. 2C). The reduction of SD values greatly contributes to high-quality copy number analysis across separately manipulated best-fit references.

Use of a paired normal reference further reduces SD values to 0.16 ± 0.03 (P < 0.05 for best-fit references), as comparison of signals is more accurately done between identical SNP loci. Use of a paired normal reference also enables allele-based analysis, which can unveil regions of copy neutral LOH in cancer cells. However, allele-based analysis only uses heterozygous SNPs, thus reducing the overall resolution by ~20% compared with the non-allele-based analysis. Recently, large regions with successive homozygous SNPs are reportedly commonly seen in leukemia samples (25).

Figure 2C illustrates that when a paired normal reference is unavailable, a best-fit reference can be used, resulting in SD reduction to 0.18 ± 0.03, which is only slightly higher than those obtained using a paired normal reference. In addition, the existence of copy neutral LOH can be also predicted from an unusually long tract of homozygous SNP calls in tumor samples (Fig. 5, bottom). Thus, when available, a paired normal reference is the ideal reference, but the average of the best-fit references is an excellent alternative that works satisfactorily in most cases, and at a lower cost.

The advantage of the copy number analysis using GeneChip Mapping 100K arrays over conventional BAC-array CGH lies in its extremely high resolution and availability of genotyp information, enabling high-density LOH analysis. Although the number of BAC clones on a single array can vary from ~3,000 to ~32,000, there exists a clear limitation to their resolution because small deletions or gains relative to the size of the average BAC clones can easily escape detection in CGH analysis (Fig. 4B). Additional limitations include (a) requirement of large amounts of BAC genomic DNA for generating spotted arrays, a process which is difficult to quality control, and (b) lack of additional genotyping information indispensable for LOH detections. All of these challenges are overcome by Affymetrix GeneChip Mapping 100K arrays. These arrays are manufactured under stringent quality control procedures; the assay requires only 250 ng of starting genomic DNA per array and provides genotyping information at >99.5% accuracy. Together, these features comprise a comprehensive approach for copy number assessment at a resolution currently not achievable by other means.

This system also has limitations. For instance, polymorphism of primer locus or restriction site would affect the signal intensity that leads to misjudgment of copy numbers. Although we can exclude these artifacts when these changes strides over different restriction fragments, validations using other methods are required to validate changes that are confined in a single fragment. It is sometimes difficult to distinguish deletions from large-scale copy number polymorphisms when best-fit references are used. In analyzing disease samples mixed with normal tissues, the amplitude of copy number changes diminishes according to the rate of normal fraction contamination. Whereas extensive SNP selection of the Mapping 100K arrays using >300 individuals likely resulted in the exclusion of SNPs on fragments with restriction site polymorphisms (via Hardy-Weinberg disequilibrium), we cannot exclude the possibility that rare mutations in restriction enzyme sites might affect amplification of affected fragments. Even if this did occur, it is highly unlikely to result in an erroneous copy number estimation because of the low likelihood that the mutation would affect more than one or two adjacent genomic fragments.

In conclusion, our improved copy number detection algorithm, combined with Affymetrix GeneChip Mapping 100K arrays, provides a powerful tool for high-resolution analysis of copy number alterations or variations across the human genome.

Acknowledgments


Grant support: Grants-in-Aid from Core Research for Evolutional Science and Technology of Japan Science and Technology Corporation; Research on Human Genome, Tissue Engineering, Ministry of Health, Labour, and Welfare, and Japan Health Science Foundation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays

Yasuhito Nannya, Masashi Sanada, Kumi Nakazaki, et al.