Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 333 (2005) 1309–1314
BBRC
www.elsevier.com/locate/ybbrc

Allelic dosage analysis with genotyping microarrays

Shumpei Ishikawa [a,b], Daisuke Komura [c], Shingo Tsuji [a], Kunihiro Nishimura [d], Shogo Yamamoto [a], Binaya Panda [e], Jing Huang [e], Masashi Fukayama [b], Keith W. Jones [e], Hiroyuki Aburatani [a,f,*]

[a] *Genome Science Division, Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan*
[b] *Department of Human Pathology, Faculty of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan*
[c] *Dependable and High-Performance Computing, Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan*
[d] *Intelligent and Cooperative System, Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan*
[e] *Affymetrix, Inc., 3380 Central Expressway, Santa Clara, CA 95051, USA*
[f] *CREST, Japan Science and Technology Agency, 4-1-8 Honchou Kawaguchi, Saitama, Japan*

Received 22 May 2005
Available online 20 June 2005

Abstract

Genomic alternations, including dosage and allelic imbalance, constitute a major basis of neoplastic and other genetic disorders. Using oligonucleotide genotyping microarrays, here we report the development and usage of an algorithm, called genome imbalance map (GIM) algorithm, for allelic as well as total gene dosage analysis. Using the GIM algorithm, global genome imbalance status at over 100,000 loci was simultaneously analyzed with unprecedented accuracy and allelic discrimination.

Keywords: DNA microarray; SNP typing; Array CGH; Oligonucleotide microarray

DNA array technology has been widely used recently to analyze gene dosage, mainly in cancer [1–4]. Notable among them are advances in array CGH (comparative genomic hybridization) technology with tiling resolution of about 30,000 BAC (bacterial artificial chromosome) [1] and oligonucleotide microarrays [4–7]. Although these techniques have added valuable contributions in analyzing copy number variation in cancer cells, integrating the dosage of individual alleles will add significance in understanding the complex genomic alterations in cancer. The representational analysis of a portion of the genome using PCR amplification and subsequent interrogation using oligonucleotide arrays has proven useful in assessing copy number changes [4]. However, various experimental biases contribute to a noise component that is seen using current analytical methods. Here, we report the development of a simple yet effective algorithm which dampens the noise from the oligonucleotide microarrays for both gene and allelic dosage analysis with unprecedented accuracy.

Materials and methods

We used the whole genome sampling analysis (also called WGSA or sampling analysis) with allele-specific hybridization on arrays originally designed for SNP genotyping [8]. In this method, the genomic DNA is digested with restriction endonuclease *Xba*I or *Hin*dIII and subsequently adaptors are ligated and the fragments are amplified with adaptor-specific primers. This amplified product

* Corresponding author. Fax: +81 3 5452 5355.
E-mail address: haburata-tky@umin.ac.jp (H. Aburatani).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.06.040

represents about 10% of the whole genome which is subsequently hybridized to 25mer oligonucleotide arrays (for details please see [8]). Single or double array hybridizations can analyze over 11,000 (10K array) or 110,000 (100K set) SNPs simultaneously (for detail description on the oligonucleotide arrays, please see http://www.affymetrix.com/products/arrays/specific/10k.affx).

For both gene and allelic dosage analysis using the signal intensity ratio of the cancer to normal samples, we developed an algorithm which takes the following parameters into account: (i) *Xba*I fragment length, (ii) percentage of GC in *Xba*I fragments, (iii) local GC content (maximum GC% calculated by 100 bp sliding window within each *Xba*I fragment), (iv) percentage of GC of 25mer probe sequence, and (v) signal intensity of probes from reference sample (matched paired genome where available) (Figs. 1B, C and Supplementary Fig. 1). Parameters such as *Xba*I fragment length, percentage of GC in *Xba*I fragments, and local GC content could directly affect the PCR performance with slight variation in experimental conditions. Percentage of GC of 25mer probe sequences could account for hybridization and wash bias in individual experiments. There could be considerable difference in signal intensity from different probes in normal reference sample (Supplementary Fig. 1c), bias which might have been generated in scanning steps. Fig. 1 explains the importance of these parameters in reducing noise, resulting in a uniform increase of chromosome 21 region signals from hybridized arrays in chr21 trisomy sample (Fig. 1D). Our algorithm aimed to remove these biases (e_{bias} in formula below) from the signal intensity ratio. In our algorithm, the experimentally observed signal intensity ratio (of a particular SNP locus) S_{observed} was modeled as:

$$\log S_{\text{observed}} = \log S_{\text{adjusted}} + e_{\text{bias}},$$

$$\log S_{\text{adjusted}} = \log S_{\text{truecopy}} + \varepsilon,$$

$$e_{\text{bias}} = \sum_{i=1}^{4} \sum_{j=1}^{n_i} b_{i,j} x_i^j,$$

where S_{adjusted} is the adjusted ratio which we aim to calculate, e_{bias} is the bias derived from x_1 to x_4, S_{truecopy} is the true genome copy number ratio, x_1 is the length of *Xba*I fragment (bp), x_2 is the percentage GC of *Xba*I fragment, x_3 is the percentage GC of the probe, x_4 is the geometric mean of signal intensity of PM probes from reference sample ($\log_{10}$ scaled), and ε is the observation error which we assume obeys a zero mean normal distribution.

All the sequence information used here are derived from the July 2003 UCSC genome build (Build 34) (for detail, please refer to http://genome.ucsc.edu/). For polymorphic loci, the mean of percentage of

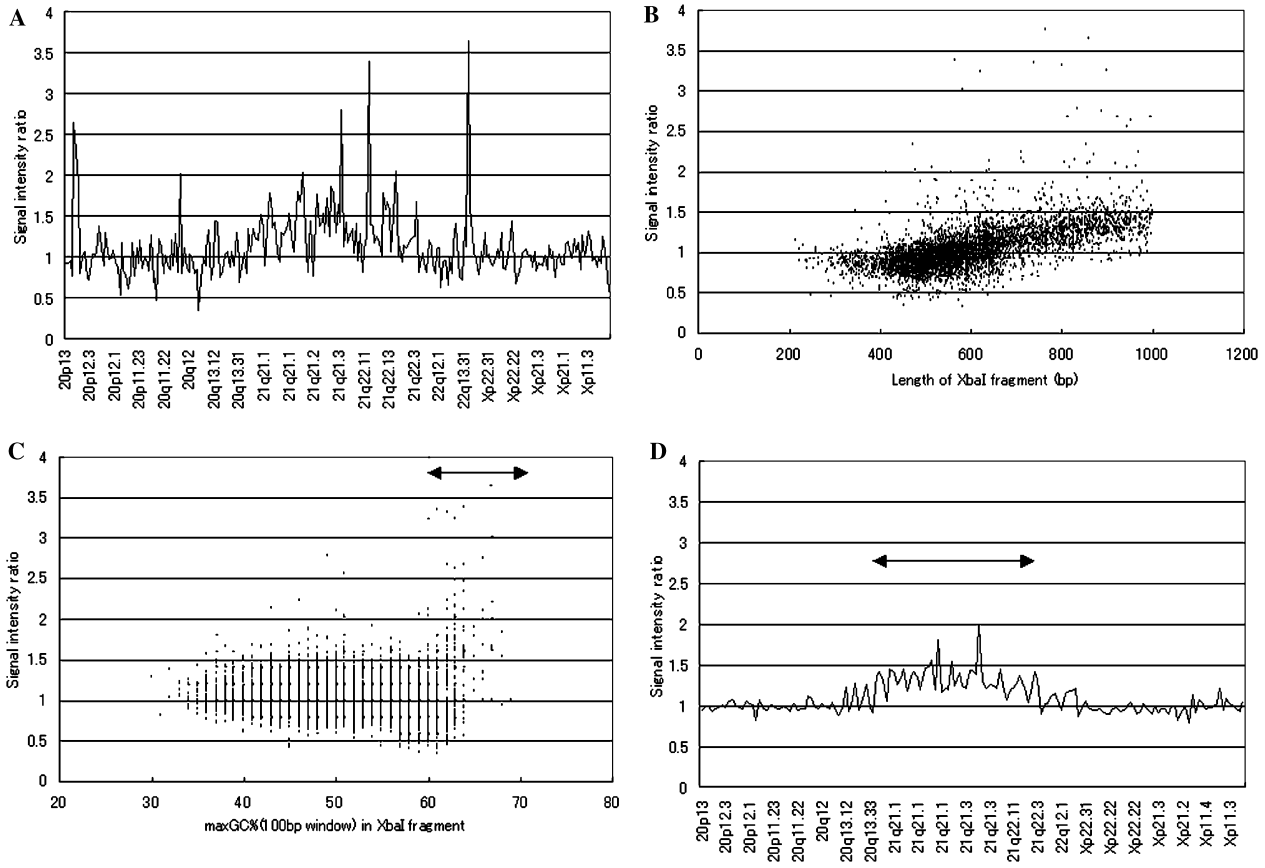

Fig. 1. Developing GIM (genome imbalance map) algorithm. (A) Raw signal intensity ratios of Down syndrome to normal diploid reference sample. Probes of mapping 10K arrays are aligned according to their physical position. Only probe sets around chr21 are shown. (B) Bias of signal intensity ratio as a function of *Xba*I fragment size. This panel shows that longer fragments are relatively well-amplified in Down syndrome sample. This can be adjusted by a polynomial function. (C) Bias of signal intensity ratio as a function of local GC contents of *Xba*I fragments (GC% is calculated using a 100 bp sliding window). In regions of higher local GC content (bi-directional arrow), an aberrantly high-signal intensity ratio appears in discontinuous fashion, suggesting that some *Xba*I fragments are not efficiently amplified in the control sample. Such regions were omitted from further analysis. (D) After modification, ratio fluctuations decrease significantly and clear increase of the signal in the chr21 regions appears (bi-directional arrow). The median and SD of signal intensity ratio are 1.30 ± 0.39 in chr21 and 1.00 ± 0.28 in other chromosomes before adjustment and 1.31 ± 0.18 and 0.99 ± 0.09, respectively, after adjustment.