Bioinformatics Consortium of Taiwan

http://bct.binfo.org.tw

National Core Facility Program for Biotechnology- Bioinformatics Consortium of Taiwan

CGcgh: A tool for molecular karyotyping using DNA microarray-based comparative genomic hybridization (array-CGH)

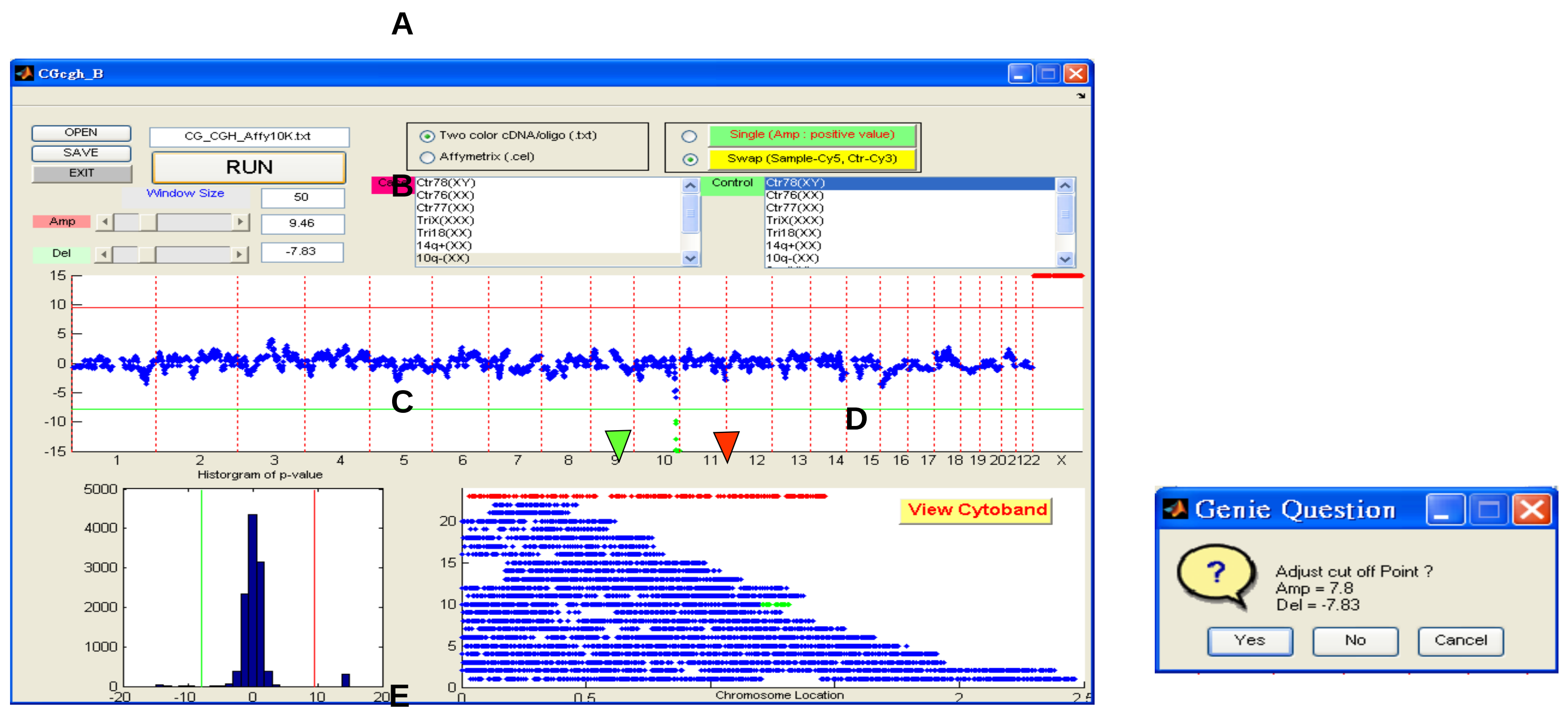
Abstract

Microarray-based comparative genomic hybridization (array-CGH) is a technique by which variations in copy numbers between two genomes can be analyzed using DNA microarrays. Array CGH has been used to survey chromosomal amplifications and deletions in fetal aneuploidies or cancer tissues. Herein we report a user-friendly, MATLAB-based, array CGH analyzing program, Chang Gung comparative genomic hybridization (CGcgh), as a standalone PC version. The analyzed chromosomal data are displayed in a graphic interface, and CGcgh allows users to launch a corresponding G-banding ideogram. The abnormal DNA copy numbers (gains and losses) can be identified automatically using a user defined window size (default value is 50 probes) and sequential student t-tests with sliding windows along with chromosomes. CGcgh has been tested in multiple karyotype-confirmed human samples, including five published cases and trisomies 13, 18, 21 and X from our laboratories, and 18 cases of which microarray data are available publicly. CGcgh can be used to detect the copy number changes in small genomic regions, which are commonly encountered by clinical geneticists. CGcgh works well for the data from cDNA microarray, spotted oligonucleotide microarrays, and Affymetrix Human Mapping Arrays (10K, 100K, 500K Array Sets). The program can be freely downloaded from http://www.mcu.edu.tw/department/biotec/en%5Fpage/CGcgh/.

Introduction

Chromosomal abnormalities are the most frequent causes of abnormal embryonic development and spontaneous abortion[1]. Comparative genomic hybridization (CGH) is a molecular cytogenetic method for genome-wide scanning of differences in DNA sequence copy number[2]. The main limitation of conventional CGH using metaphase chromosomes spread on slides, however, is its low resolution at about 10 to 20 Mb[3]. To remedy the shortcoming of limited resolution, an array CGH, was developed for performing CGH using mapped genomic clones spotted on slides. From then on, DNA microarray–based CGH, also known as genomic microarrays[4], has been widely used for the genome-wide screening of genomic imbalances in tumor cells[5]. Bacterial artificial chromosome (BAC) clones, cDNA clones and oligonucleotides have been used for array CGH. The latter includes Affymetrix Human Mapping Arrays (www.affymetrix.com) and Agilent Oligonucleotide Array-based CGH (www.agilent.com).

Many cancers are notorious on its high number of aberrant chromosomal copies, chromosomal deletions and amplifications[6]. Various academically available software programs for analyzing array CGH have been recently compared on their efficacy of assessing chromosomal abnormality of cancers such as ChARMView[7], CGH-Explorer[8], and CGH-plotter[9]. As a tertiary referral center for identification of abnormal fetuses, our interest is on the use of array-CGH for detecting fetal aneuploidies. The use of these available programs for array CGH analysis, however, often required user's manual parameter settings in order to detect single-copy changes of genomic regions accurately on several G-banding confirmed abnormally karyotyped tissues. Herein, we report the development of a user-friendly software tool, named Chang Gung CGH (CGcgh), for analyzing array CGH on both two-colored spotted microarrays and Affymetrix Human Mapping Arrays.



Adjust cut off Point ?
Amp = 7.8
Del = -7.83

Figure 1. Analysis with CGcgh on an abnormal 46XX, 10q- (q26→qter) DNA using Affymetrix 10K array
The main menu of CGcgh. (B) The log of P-value to the base 10 (Y-axis) among each chromosome (X-axis). (C) Distribution histogram of the log P values. (D) Chromosomal display of the DNA amplification (red dots) and deletion (green dots). (E) When the "RUN" button is clicked, a K-means dialog appears to ask the user to adjust the values of the threshold of log P value of amplification and deletion.

Method

CGcgh was implemented in MATLAB Version 7.4 with Microsoft Windows XP Version 5.1 and using a Graphical User Interfaces (GUIs) Design Environment tools.

Loading and pre-processing of two-colored .txt data: CGcgh was used to analyze genomic DNA signal data obtained from dual-color cDNA microarrays or Agilent oligonucleotide microarrays with data in tab-delimited text format. For the two-colored spotted microarray data, the "log ratio" values of Cy5 and Cy3 intensity were used, and we filtered the outlier data with log ratio values greater than 3 standard deviations.

Loading and pre-processing of Affymetrix .cel data: The .cel files of Affymetrix 10K, 10K 2.0, 50K Hind 240, 50K Xba 240, 250K Nsp, and 250K Sty arrays can be directly imported into CGcgh program for analysis. Four normalization methods ("APTools", "PLIER", "RMA" , "MAS5.0") were supported the program (http://www.affymetrix.com/support/developer/powertools/index.affx). After normalization, the intensity of each SNP probe set was then logarithmically transformed into the "log ratio" value for the selected "CASE" and "CONTROL".

Sliding t-test: On each chromosome, we scanned the genomic regions with a user defined window size (default value is 50 probes), and sequential student t-tests were performed to determine whether each scanning genomic region and the whole genome (except chromosome X) had the same mean values. We slid the window of genomic region along all chromosomes (X axis in **Fig. 1B**) and plotted the base 10 logarithm of P-value (Y axis in **Fig. 1B**).

K-means clustering: The histogram of log P values in the CGcgh program with an abnormal DNA sample (**Fig. 1C**) were divided into 2 or 3 groups. The major group clustered around zero, and the other two groups around 10~15 or –10~-15, each reflecting amplification and deletion, respectively. In the example pair of 46XX, 10q- versus 46XY, the 10q deletion and XX amplification region clustered into about 10~15 and –10~-15, respectively (**Fig. 1C**). We used a K-means clustering method to partition the training groups with known karyotypes to obtain the best threshold of log P values for detecting aberrations region (**Fig. 1E**). In all of our trained cases, the absolute log P values greater than 5 for cDNA microarray and greater than 6 for Affymetrix Human Mapping Arrays, indicated a significant DNA copy changes, either amplification or deletion.

CGcgh program also identifies the locations of significant copy number changes, such as amplicons and deletions (**Fig. 1D**). The position of the amplification /deletion region and the log P value can be exported as a tab-delimited text file (**Fig. 1A**). CGcgh program can also display the G-banding ideogram of each chromosome at an 850-band resolution (refer to *Homo sapiens* Genome, Build 35 version 1, NCBI) (**Fig. 2**).

Results

Comparison of CGcgh with CGH-Explorer, CGH-Plotter and ChARM

All of the parameters of the CGcgh were trained with more than 30 DNA samples from Genomic Medicine Research Core Laboratory (GMCRL), Chang Gung Memorial Hospital. **Table 1** summarizes the comparison of 15 specimens between using CGcgh, CGH-Explorer, CGH-Plotter and ChARM View. The array CGH result of every sample was also compatible with the corresponding chromosomal diagnosis (**Table 1**).

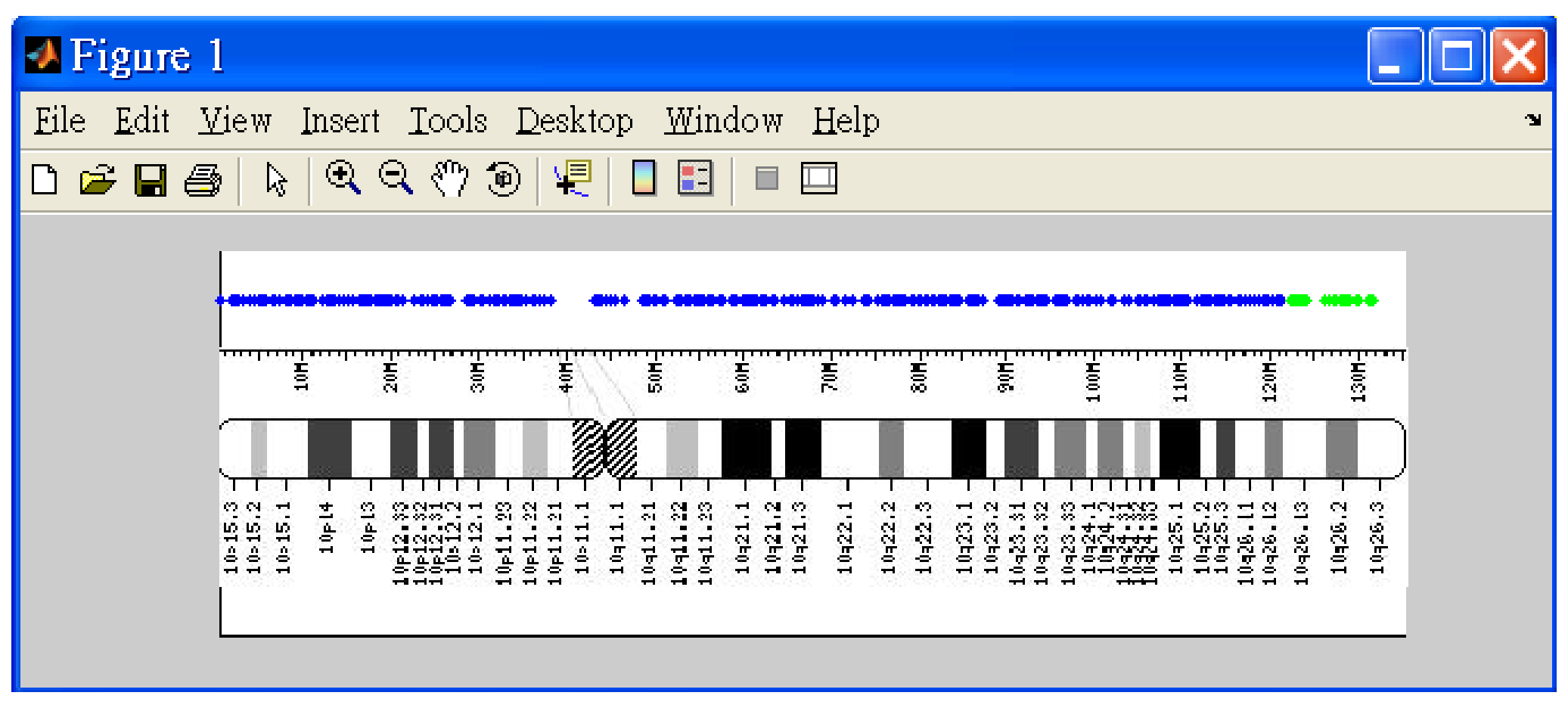


Figure 2. Ideogram of the G-banding pattern.
Selected chromosome is displayed as the ideogram of the G-banding pattern at 850-band resolution.

Analysis of .cel files derived from Affymetrix 500K Array Set

Using CGcgh, we reanalyzed the publicly available data of Affymetrix 500K Human Mapping Arrays of autism spectrum disorder derived from the GEO datasets [12]. We interrogated CGcgh program with 18 mutant DNA samples that were analyzed with the Affymetrix® Genotyping Console™ Software and confirmed with real-time quantitative PCR or fluorescence in situ hybridization by Marshall et al[13]. As showed in **Table 2**, CGcgh software specifically detected DNA fragment with 50 probes window size when the size of amplified fragment was greater than 1.43 Mb and the size of deleted fragment was greater than 1.41 MB. However, CGcgh could not detect the amplified fragment when its size was less than 0.67 Mb and the deleted fragment at the size less than 0.35 Mb.

Discussion

CGcgh, a program with a user-friendly graphic interface, is developed for detecting the single copy change in small genomic regions that are commonly encountered by clinical geneticists[10,11,14,15,16]. CGcgh supports the data generated from cDNA, spotted oligonucleotide microarrays, and Affymetrix Human Mapping Arrays. It uses a sliding t-test and K-means algorithm for analyzing array CGH without the need to install additional MATLAB toolbox.

***Table* 1. Comparison of the performance between CGcgh and three available array CGH programs**

Samples	Genomic DNA	ChipType	CGcgh	CGH-Explorer ver 2.3		CGH_Plotter		ChARM View
				Default	Optimal	Default	Optimal	Default /Optimal
				Automatic penalty Off	Automatic penalty On	Window size = 5; DC = 6	Window size = 25; DC = 2	Sign p-value= 0.01 Mean p-value= 0.01
1	47,XXX	Affymetrix 10K	XXX	O	O	+	O	NA
2	47, XY, 18+	Affymetrix 10K	XY, 18+	O	O	-	O	NA
3*	46, XX, 14q+	Affymetrix 10K	XX, 14q+ (q32->qter)	+	O	-	O	O
4	46, XX, 10q	Affymetrix 10K	XX, 10q- (q26->qter)	+	O	-	O	O
5*	46, XX, 6q-	Affymetrix 10K	XX, 6q- (6q22.1->q23.2)	+	O	+	O	O
6*	46, XY, 9q-	Affymetrix 10K	XY, 9q- (9q22.32->q31.3)	+	O	-	-	O
7*	46, XY, 4p-	Affymetrix 10K	XY, 4p- (p15.3->pter)	+	O	-	O	O
8	47, XX, 21+	GMRCL 7K	XX, 21+	+	O	-	O	NA
9	47, XX, 18+	GMRCL 7K	XX, 18+	+	O	-	O	NA
10	47, XY, 13+	GMRCL 7K	XY, 13+	+	O	-	O	NA
11*	46, XY, 12q+	GMRCL 7K	XY, 12q+ (q21.2->qter)	+	O	-	O	-
12	46, XY, 10q-	GMRCL 7K	XY, 10q- (q22->qter)	+	O	-	O	-
13	46, XX, 10q-	GMRCL 15K	XX, 10q- (q26->qter)	-	-	-	+	O
14	46, XX, 14p+	GMRCL 15K	XY, 14p+ (p22->pter)	-	O[a]	-	O	O
15*	46, XY, 4p-	GMRCL 15K	XY, 4p- (p15.3->pter)	-	O[a]	-	O	+

O: array CGH results were identical with G-banding results; + : With some false positive results; -: No result. DC: Constant for computing; O[a]: results obtained only with additional parameter setting in "Least allowed deviation" = 0.1 (Default = 0). NA: CharmView program are not applicable for whole chromosomal amplification analyses.
Referred papers of published cases: 3 [14], 5* [10], 6*[17], 7* [11], 11* [16], 15* [11]

***Table 2 Reanalysis of a data set of Affymetrix Human Mapping 500K Arrays*[a]**

FamID[a]	Sex	Chromosome	Size (Kb)	Loss/gain	CGcgh
SK0152-003	M	3p25.1-p24.3	**1,409**	loss	○
SK0205-004	F	5p15.33-p15.2	13,800	loss	○
SK0083-003	M	7q31.1-q31.31	11,023	loss	○
SK0131-003	F	7q31.1-q32.2	15,486	loss	○
SK0243-003	M	15q23-q24.2	4,289	loss	○
SK0245-005	M	15q11.2-q13.3	11,871	gain	○
SK0218-003	F	18q21.32-q23	20,358	loss	○
NA0097-000	F	Xp22.33-p22.31	5,825	loss	○
NA0002-000	M	7q36.2	66	loss	✕
MM0278-003	M	12q24.21-q24.33	18,218	gain	○
NA0067-000	M	16q24.3	265	loss	✕
MM0088-003	F	16p11.2	675	loss	✕
SK0019-004	M	16p11.2	675	loss	✕
SK0244-003	M	21q22.3	353	gain	✕
MM0109-003	F	20q13.33	**1,427**	gain	○
SK0119-003k	M	22q11.21	2,771	loss	○
SK0297-003	M	22q11.21	4,281	gain	○
SK0306-004	F	Xp11.23-p11.22	4,643	gain	○

[a] Details are in Reference #13. Symbols: ○ detectable, ✕ undetectable.

Welcome to contact with BCT
Email: bct@ym.edu.tw, Tel: 02-28267359
http://bct.binfo.org.tw

NCFPB
生技類核心設施平台維運計畫
National Core Facility Program for Biotechnology