Published online 18 September 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 17* ***e116***
doi:10.1093/nar/gkl601

AccuTyping: new algorithms for automated analysis of data from high-throughput genotyping with oligonucleotide microarrays

Guohong Hu, Hui-Yun Wang, Danielle M. Greenawalt, Marco A. Azaro, Minjie Luo, Irina V. Tereshchenko, Xiangfeng Cui, Qifeng Yang, Richeng Gao, Li Shen and Honghua Li*

Department of Molecular Genetics, Microbiology and Immunology, University of Medicine and Dentistry of New Jersey-Robert Wood Johnson Medical School, SRB 110, 661 Hoes Lane, Piscataway, NJ 08854, USA

Received March 9, 2006; Revised July 11, 2006; Accepted July 31, 2006

ABSTRACT

Microarray-based analysis of single nucleotide polymorphisms (SNPs) has many applications in large-scale genetic studies. To minimize the influence of experimental variation, microarray data usually need to be processed in different aspects including background subtraction, normalization and low-signal filtering before genotype determination. Although many algorithms are sophisticated for these purposes, biases are still present. In the present paper, new algorithms for SNP microarray data analysis and the software, AccuTyping, developed based on these algorithms are described. The algorithms take advantage of a large number of SNPs included in each assay, and the fact that the top and bottom 20% of SNPs can be safely treated as homozygous after sorting based on their ratios between the signal intensities. These SNPs are then used as controls for color channel normalization and background subtraction. Genotype calls are made based on the logarithms of signal intensity ratios using two cutoff values, which were determined after training the program with a dataset of ~160 000 genotypes and validated by non-microarray methods. AccuTyping was used to determine >300 000 genotypes of DNA and sperm samples. The accuracy was shown to be >99%. AccuTyping can be downloaded from http://www2.umdnj.edu/lilabweb/publications/AccuTyping.html.

INTRODUCTION

Microarray is a powerful technology for detecting and resolving a large number of nucleic acids simultaneously. cDNA microarrays (1–6) for large-scale analysis of gene expression and DNA copy number changes have been used extensively. Computer programs for all steps involved in analyzing cDNA array data have been developed. Microarrays used for genotyping are receiving more and more attention, especially after the discovery of millions of single nucleotide polymorphisms (SNPs). To meet the strong demand in high-throughput SNP genotyping, we (7) and several other groups (8–13) have developed high-throughput multiplex genotyping systems, which have been used in many studies (14–17). With these systems, a large number of SNP-containing sequences can be amplified in one or a few tubes followed by the analysis with oligonucleotide microarrays; and thousands of SNPs in a large number of samples can be genotyped in a highly efficient and affordable way. However, the immense amount of data generated from even a single microarray precludes manual processing. Automation of data analysis is an essential prerequisite for routine genotyping with microarrays.

Experimentally, data from oligonucleotide microarrays are obtained by hybridizing sample sequences to corresponding probes arrayed on solid supports. Detection of specific sequences is accomplished by either labeling the sample sequences with fluorescent dyes before hybridization or labeling the probes after hybridization. The fluorescent intensities on the probes are determined by digitizing the images of arrayed spots after scanning.

When data are obtained in good quality, the accuracy of genotyping results is usually affected by two factors, background signal and color channel bias. Normally, signal from each array spot consists of signal from specific labeling that is predominant and non-specific signal as a small portion. The amount of non-specific signal may vary depending on experimental performance. To ensure a high degree of genotyping accuracy, it is necessary to separate the non-specific noise from the specific signal.

When more than one fluorescent dye is used to label sequences of different natures, the signal intensities from different dyes could differ even if the same amounts of

*To whom correspondence should be addressed. Tel: +1 732 235 7330; Fax: +1 732 235 5223; Email:holi@umdnj.edu

sequences are present in the sample. Variation may be caused by the differences in fluorescent emission and scanning efficiency of the dyes. When an array is scanned, the 'gains' used for different fluorescent channels may vary, depending on the users' experience and the scanner performance. These factors may result in a global difference between signal intensities of the fluorescent dyes. Therefore, the microarray data from different color channels need to be normalized so that the intensities of different colors can be compared. In the case of SNPs, the two allelic sequences of a heterozygous SNP may not necessarily incorporate equal amounts of fluorescence. Therefore, a highly accurate normalization method is required to separate such a bias from the difference caused by the amounts of DNA.

Several methods for channel normalization and background estimation have been reported. One of the commonly used methods for the normalization of RNA expression data is to correct the systematic bias by using the channel signal means of all spots assuming that the average gene expression levels in the genome have little changes (18–21). However, in an SNP analysis, the number of allelic molecules labeled in one color may not be equal to those in the other color. In this case, the channel signal means would be biased. Intensity-dependent normalization methods have also been used, such as Lowess smoothing method (18,20). When the logarithms of intensity ratios [Ln(R)] are plotted against the logarithms of intensity products [Ln(I)], the Lowess method detects the systematic bias in the plot by carrying out a local weighted linear regression. The bias is corrected when the Ln(R) value of each spot is subtracted by its fitted value. This method is not ideal for analyzing SNP genotyping data either, because the amounts of signals from unequal amounts of allelic sequences present in the sample may also contribute to the local bias in an Ln(R) versus Ln(I) plot, which could be incorrectly removed by local bias subtraction.

Methods for background estimation have also been reported previously. One method for background signal estimation uses the fluorescence signal within the surrounding area of a spot as a representation of its background (22,23). This method only takes inter-spot background into consideration, while intra-spot background is the real non-specific signal contributing to the spot signal intensities. Another method involves using microarray spots containing only the printing buffer as controls. Obviously, signals from these controls may not reflect non-specific hybridization. Although printing buffer with non-probe oligonucleotides can be used as controls, these controls are always limited by the number of spots allowable on the array and may also not be representative of non-specific signal from real probes. Affymetrix genotyping microarrays use probes containing one or two mismatches as controls for each SNP (12). Hardenbol *et al*. (11) used four probes for each biallelic SNP among which the two non-allelic probes served as background controls. Di *et al*. (24) developed dynamic model-based algorithms for genotyping. The authors calculated the likelihood of the possible allelic states using signals from 56 perfect match and mismatch probes for each SNP. Mismatch probes in this method provide marker-specific assessment of background. However, these probes can only be used to determine the ratio between perfect match and mismatch probes, and not to define the background signal since in a microarray analysis, usually the mismatches between the non-specific sequences and the probes are more than one or two bases. In addition, the cost for extra probes is another concern.

We have developed new algorithms for microarray-based SNP analysis. Our algorithms take advantage of the large number of SNPs analyzed in each high-throughput assay and use signal intensities of SNPs that are very likely to be homozygous as both negative and positive controls. Therefore, these controls reflect the scenario in real samples.

MATERIALS AND METHODS

SNP microarray system

Our newly developed genotyping system can be used to genotype >1000 SNPs in a single assay (7). With this system, SNP-containing sequences are first amplified in a single tube to a detectable amount by multiplex PCR, and then used as templates to generate single-stranded DNA (ssDNA), which is then hybridized to the probes on a microarray. The probes are designed in such a way that their 3′ ends of the probes are next to the polymorphic sites in the hybridizing ssDNA. In this way, the probes can be labeled with the commonly used single-base extension method (25–27) during which single dideoxyribonucleotides (ddNTPs) are added to the probe in an allelic-specific way depending on the hybridizing allelic sequence(s). When the corresponding ddNTPs are labeled with different fluorescent chromophores (cyanine dyes, either Cy3 or C5, in our system), the allelic state of the SNPs can be determined by analyzing the amount of incorporated fluorescence. Data used in the present study were obtained with this two-color system. However, the algorithms described below could also be used for four-color systems. In addition, the genotypes of SNPs may be determined independently by using the two DNA strands as templates separately so that results from such dual-probe analysis could be compared to ensure a high degree of accuracy.

Genotyping algorithms

Based on the fluorescence intensities of the spots on microarray, genotypes of SNPs are determined in four steps: channel normalization, low-signal filtering, background subtraction and genotype determination with predicted homozygous SNPs as controls.

Using homozygous SNPs as controls. When an SNP is in a homozygous state, its probe should predominantly incorporate one color, which is designated as the signal color while the other color should be considered as the background color. The use of the signal and background colors for a large number of homozygous SNPs as controls for channel normalization and background subtraction should be a reliable method. This is because when the numbers of SNPs are large, the means of signal color intensities of the two groups of homozygous SNPs should be very close. Difference between the two means reflects channel bias. The background color for each group of homozygous SNPs would represent the sum of noises contributed by all experimental factors including the non-specific hybridization.

According to binomial distribution, the 95% confidence interval for the fraction of heterozygous SNPs in 1000 loci for a given individual would be 0.5 ± 0.031 when all SNPs are at the maximal heterozygosity of 0.5. Practically, the fraction of heterozygous SNPs should be smaller because no individual has all pre-selected SNPs at the maximal heterozygosity. Therefore, when SNPs are sorted based on their intensity ratios of the two colors, it would be safe to assume that 40% of SNPs with the highest or lowest ratios are homozygous (or 20% at each end). To validate this hypothesis experimentally, 469 SNPs at the extreme 20% of each end of a panel of 1172 SNPs in a sample were re-genotyped by the methods of restriction fragment length polymorphism (RFLP) (28,29) and direct sequencing. All these SNPs were shown to be homozygous.

Data normalization. After digitizing the image, SNPs on the microarray are first sorted based on their ratios between the two color intensities. The two 20% groups of SNPs with the highest and lowest ratios are treated as homozygous and used as controls. Signals for all spots are then normalized using the following equation:

$$S'_{ij} = S_{ij} \times \sqrt{\frac{\overline{N_{j'}}}{\overline{N_j}}}, \quad \mathbf{1}$$

where S_{ij} and S'_{ij} are the original and normalized signal intensities of spot i ($i = 1, 2, \ldots, n$ spots on an array) in channel j (g, green; or r, red) on the microarray. $\bar{N}_j$ and $\bar{N}_{j'}$ are the means of the signal intensities in the red and green channels, respectively. If j' is r and then j is g, and vice versa.

Low-signal filtering. After normalization, microarray spots with low intensities for both colors are usually removed from further analysis (12,30–33). In our program, the background intensities for the SNPs in each 20% homozygous group are first trimmed by removing the extreme values that are more than 1 SD from the mean. This step removes the sporadic outliners in the control dataset. The mean, $\bar{B}_j$, and standard deviation, σ_j ($j = r$ or g), of the homozygous subsets for channel j are then recalculated for each trimmed set and is used to filter out the low-signal spots. For any spot i, if the normalized signal intensities, S'_{ij}, in both channels match the following condition:

$$S'_{ij} \leqslant \bar{B}_j + n\sigma_i, \quad \mathbf{2}$$

where n is a user-defined value (that usually is 2), no genotype call will be made and the corresponding spot is defined as undetected with a flag of 'L' for 'low-signal'.

Background subtraction. After filtering out the low-signal spots, the 'true' signals in the remaining spots, S''_{ij}, are computed by subtracting the background mean from the normalized values. To avoid negative values for calculation of Ln(R), any negative S''_{ij} are replaced by a small value, 1 (34). To be reasonable, all S''_{ij} smaller than 1 are also set to 1.

Genotyping. A straightforward log-ratio cutoff method is employed to determine the genotype calls. Two cutoff values of Ln(R), L_r and L_g were determined using a training dataset consisting of ~160 000 genotypes obtained from the dual-probe analysis. During the training process, cutoff values from 1 to 2.4 of the Ln(R) with a 0.1 increment was tested to find the values that give the best concordance rate by comparing the two datasets obtained from independent dual-probe assays. The best concordant rate was obtained when $L_r = -L_g = 1.5$ were used. As shown in Figure 1, the two parallel lines, $y_1 = L_r$ and $y_2 = L_g$, in the scatter plot of Ln(R)'s, separate the data points into three groups. SNPs with Ln(R)'s bracketed by the two lines are classified as heterozygous, and those with Ln(R)'s outside of the bracketed range are classified as homozygous.

RESULTS

The above data preprocessing steps improve the data quality in three aspects:

(i) Centering datasets with respect to the line $y = 0$ through normalization. Although channel bias is not always visually obvious, a significant portion of datasets are biased to a certain extent and need to be normalized to ensure a high degree of accuracy. Although not often, results from a small portion of microarrays may be highly biased. In these cases, the effect of channel normalization becomes more visually obvious on the scatter plots;

(ii) Eliminating spots with signals that are not significantly higher than background. As shown in Table 1, this step eliminates 1.46 and 1.15% (low-signal rate = 1 − detection rate) of spots in the assays with 'AG' and 'CT' probes, respectively; and

(iii) Separating the three groups of spots representing SNPs in the different allelic states further away through background subtraction. The original dataset falls between the lines $y = 5$ and $y = -6$. After this step, most data points for the homozygous SNPs are out of this range. Such an effect is especially obvious for the spots with signals that are low but significantly higher than the background. As shown on the left side of the lower scatter plot in Figure 1, spots with relatively low signals are much better separated compared with the unprocessed plot in the upper panel.

Spots falling between different genotype clusters on the scatter plots are one of the major error sources. AccuTyping has a 'twilight zone' function. The twilight zones are centered by the cutoff lines. Users may adjust the width of the twilight zones based on the data quality to further insure the genotyping accuracy.

A computer software, AccuTyping, was developed based on the algorithms described above. AccuTyping takes inputs of the two color intensities digitized from scanned microarray images with one of the two popular software packages, GenePix (Axon Instrument, Union City, CA) or ImaGene (Biodiscovery, Inc., El Segundo, CA). The program may process either single datasets separately or multiple datasets in a batch. Figure 2 shows the Windows interface of the program.

When the probes are spotted in duplicate on the array, the average signal intensities are used. Occasionally, the average may not necessarily be an appropriate estimate of true signal values. For example, the average of two duplicated spots with

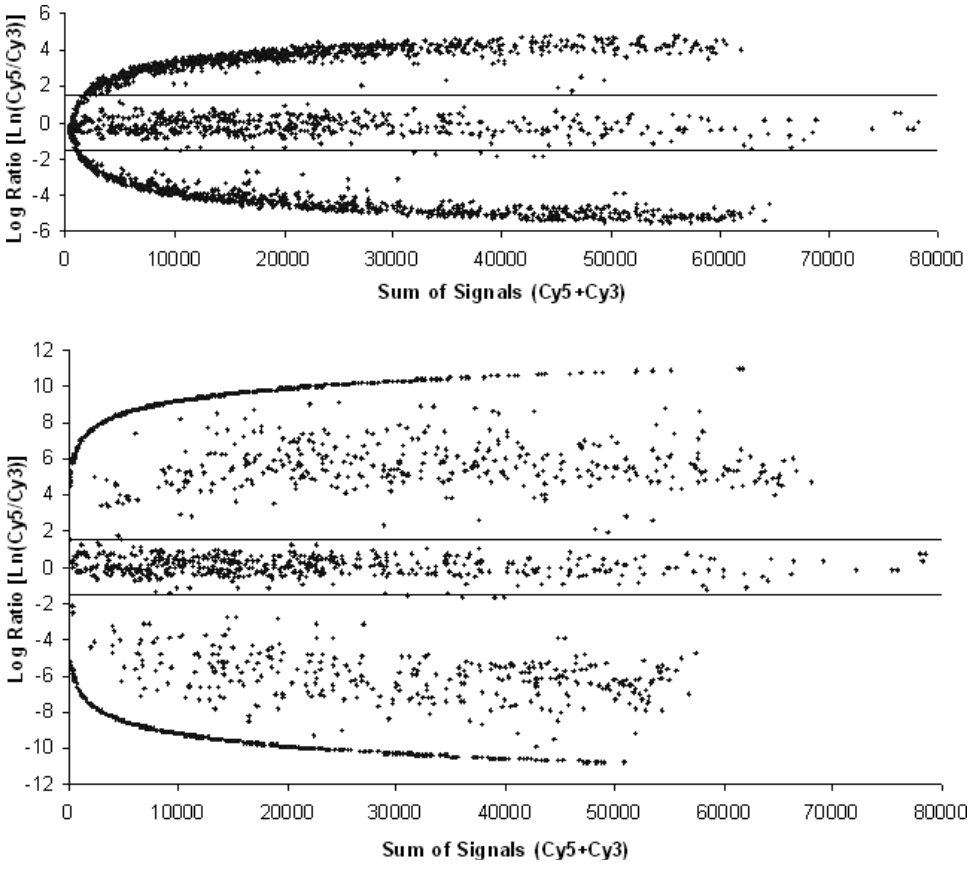

Figure 1. Scatter plots of data from a microarray for 1172 SNPs. The signal intensity log ratios, Ln(R)'s, are plotted against the signal sums of the two color intensities (R + G, Cy3 + Cy5). Upper panel, plot using the original data. Lower panel, plot after data processing. Note, spots with both color intensities smaller than the low-signal filtering values (Equation 2) were eliminated by the program and are not plotted in the lower panel. The two lines, $y = \pm 1.5$, encompass the heterozygous cluster.

predominant signals in opposite colors could give a heterozygous call. The program detects such cases and excludes them in the final genotyping step. Some spots could show apparent abnormality in the microarray images because of mechanical failure during experiment. AccuTyping excludes these spots if the user flags them. After channel normalization, low-signal filtering and background subtraction, genotype calls are made. The program also checks reproducibility between repeated assays and the accuracy by comparing the genotype calls from independent genotyping methods.

AccuTyping outputs comma separated value (CSV) files containing genotype calls, the original and processed data as well as the numerical values for the parameters used for normalization, background subtraction and low-signal filtering. The statistics of detection, concordance and accuracy, if available, are also appended in the output files. For a batch run, output results are written into a separate file for each dataset. All genotypes are pooled into an additional file to facilitate further analysis. The program is available for download from our website http://www2.umdnj.edu/lilabweb/Publications/AccuTyping.html.

AccuTyping has been used routinely for SNP genotyping with human genomic DNA and single-sperm samples (7,17) in our laboratory. Table 1 shows part of the genotyping results from typing 1172 SNPs in 24 human genomic DNA samples. The program indicates a detection rate of 98.54% with probes that incorporate fluorescently labeled nucleotides ddA or ddG (AG probes) and 98.85% with probes in the other direction, which incorporate fluorescently labeled ddC and ddT (CT probes). An average detection rate with dual probes was 97.58%, which is 1.5 and 1.8% different from the rates for single probes, respectively.

Since genotyping with probes in two different directions is carried out in independent experiments, results generated in this way can validate each other. After comparing the genotyping results obtained with probes in different directions, a small portion of the genotypes may be found to be inconsistent. As shown in Table 1, the average non-concordant rate is 2.94% with an SD of 0.36%. However, being non-concordant does not mean incorrect. Since the system involves only two colors, it is reasonable to assume that if only the probes in one direction were used, half of the non-concordant genotypes should be correct. This hypothesis was verified in our recent publication (7). Based on this method, accuracy based on the dual-probe approach is calculated and listed in Table 1. It should be pointed out that the non-concordant rates calculated in this way are used to estimate the error rates for the results obtained with single probes. Practically, non-concordant genotypes detected with dual probes will be discarded. In this case, the error rate among the concordant genotypes should be ε^2, where ε is the error rate for single probes, and the issue would be how accurate the remaining

Table 1. Summary of genotyping results of 1172 SNPs in 24 human genomic DNA samples

Sample No.	SNPs	Detectable in each direction								Detectable in Both Directions							Detectable in one directon							
	Total	AG				CT				Total		Concordant				Accuracy(CT)	Non-concordant		AG		CT		Undetectable	
		No.	%	Heterozygous				Heterozygous						Heterozygous										
				No.	%	No.	%	No.	%	No.	%	No.	%	No.	%		No.	%	No.	%	No.	%	No.	%
1	1172	1161	99.06	356	30.66	1159	98.89	362	31.23	1149	98.04	1110	96.61	337	30.38	99.97	39	3.39	12	1.02	10	0.85	1	0.09
2	1172	1151	98.21	317	27.54	1162	99.15	320	27.54	1144	97.61	1108	96.85	298	26.90	99.98	36	3.15	7	0.60	18	1.54	3	0.26
3	1172	1150	98.12	315	27.39	1160	98.98	318	27.41	1140	97.27	1102	96.67	298	27.04	99.97	38	3.33	10	0.85	20	1.71	2	0.17
4	1172	1151	98.21	323	28.06	1169	99.74	325	27.80	1150	98.12	1118	97.22	308	27.55	99.98	32	2.78	1	0.09	19	1.62	2	0.17
5	1172	1154	98.46	330	28.60	1165	99.40	326	27.98	1148	97.95	1111	96.78	309	27.81	99.97	37	3.22	6	0.51	17	1.45	1	0.09
6	1172	1161	99.06	388	33.42	1168	99.66	386	33.05	1158	98.81	1130	97.58	372	32.92	99.99	28	2.42	3	0.26	10	0.85	1	0.09
7	1172	1158	98.81	266	22.97	1149	98.04	255	22.19	1136	96.93	1108	97.54	245	22.11	99.98	28	2.46	22	1.88	13	1.11	1	0.09
8	1172	1162	99.15	274	23.58	1153	98.38	259	22.46	1145	97.70	1112	97.12	249	22.39	99.98	33	2.88	17	1.45	8	0.68	2	0.17
9	1172	1157	98.72	269	23.25	1150	98.12	266	23.13	1138	97.10	1099	96.57	249	22.66	99.97	39	3.43	19	1.62	12	1.02	3	0.26
10	1172	1158	98.81	248	21.42	1164	99.32	255	21.91	1152	98.29	1123	97.48	239	21.28	99.98	29	2.52	6	0.51	12	1.02	2	0.17
11	1172	1158	98.81	281	24.27	1151	98.21	271	23.54	1139	97.18	1109	97.37	257	23.17	99.98	30	2.63	19	1.62	12	1.02	2	0.17
12	1172	1158	98.81	265	22.88	1160	98.98	263	22.67	1147	97.87	1117	97.38	252	22.56	99.98	30	2.62	11	0.94	13	1.11	1	0.09
13	1172	1151	98.21	325	28.24	1157	98.72	318	27.48	1141	97.35	1112	97.48	304	27.34	99.98	29	2.54	10	0.85	16	1.37	5	0.43
14	1172	1145	97.70	337	29.43	1165	99.40	350	30.04	1141	97.35	1115	97.72	329	29.51	99.99	26	2.28	4	0.34	24	2.05	3	0.26
15	1172	1147	97.87	339	29.56	1160	98.98	339	29.22	1142	97.44	1110	97.20	320	28.83	99.98	32	2.80	5	0.43	18	1.54	7	0.60
16	1172	1155	98.55	330	28.57	1162	99.15	336	28.92	1147	97.87	1109	96.69	314	28.13	99.97	38	3.31	8	0.68	15	1.28	2	0.17
17	1172	1160	98.98	310	26.72	1162	99.15	318	27.37	1152	98.29	1117	96.96	296	26.50	99.98	35	3.04	8	0.68	10	0.85	2	0.17
18	1172	1161	99.06	330	28.42	1165	99.40	330	28.33	1154	98.46	1123	97.31	314	27.96	99.98	31	2.69	7	0.60	11	0.94	0	0.00
19	1172	1141	97.35	332	29.10	1157	98.72	341	29.47	1130	96.42	1095	96.90	317	28.95	99.98	35	3.10	11	0.94	27	2.30	4	0.34
20	1172	1162	99.15	330	28.40	1140	97.27	320	28.07	1132	96.59	1094	96.64	301	27.51	99.97	38	3.36	30	2.56	8	0.68	2	0.17
21	1172	1149	98.04	283	24.63	1154	98.46	289	25.04	1133	96.67	1100	97.09	268	24.36	99.98	33	2.91	16	1.37	21	1.79	2	0.17
22	1172	1159	98.89	334	28.82	1152	98.29	324	28.13	1140	97.27	1101	96.58	307	27.88	99.97	39	3.42	19	1.62	12	1.02	1	0.09
23	1172	1151	98.21	276	23.98	1164	99.32	275	23.63	1145	97.70	1109	96.86	259	23.35	99.98	36	3.14	6	0.51	19	1.62	2	0.17
24	1172	1156	98.63	320	27.68	1157	98.72	328	28.35	1143	97.53	1106	96.76	303	27.40	99.97	37	3.24	13	1.11	14	1.19	2	0.17
Average	1172	1155	98.54	312	26.98	1159	98.85	311.4	26.87	1144	97.58	1110	97.06	294	26.44	99.98	33.7	2.94	11.3	0.96	15	1.28	2.2	0.19
SD	—	5.8	0.50	33.9	2.95	6.9	0.59	36.29	3.07	7.1	0.61	8.9	0.36	34	3.01	0.50	4.1	0.36	6.96	0.59	5	0.43	1.5	0.13

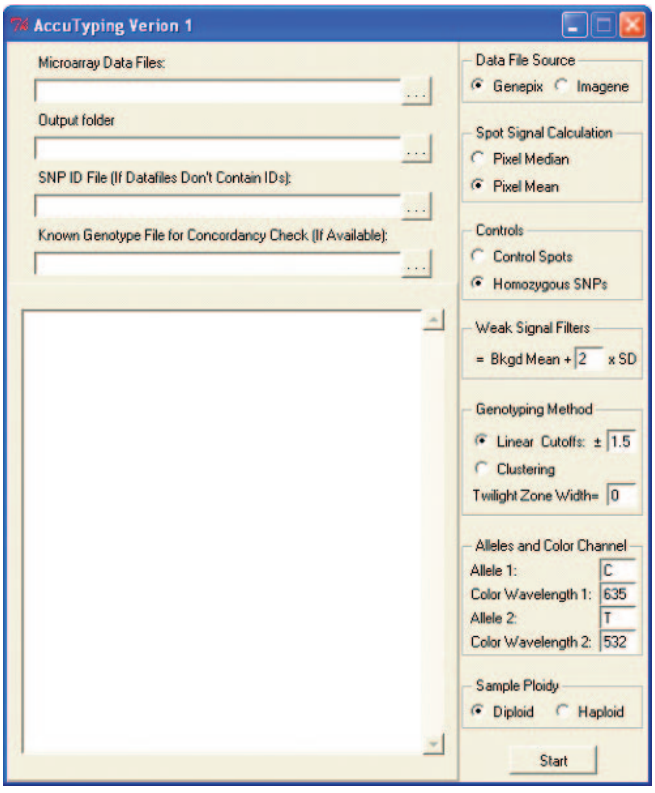

Figure 2. The graphic interface of AccuTyping.

genotype calls would be. Usually, results generated with probes in two directions can be used to address this issue. To take extra precaution, a panel of 1282 such genotypes was retyped with the RFLPs and direct sequencing methods, and all were shown to be consistent with the microarray results. These results indicate the accuracy of the dual-probe method is close or even equal to 100%.

DISCUSSION

Practically, all microarray data need to be normalized before information extraction. However, normalizations for data from SNP genotyping arrays and for those from gene expression profiling arrays are different concepts. Gene expression profiling arrays are used to learn the quantitative features of gene products in different samples. Normalization for these arrays should include (i) intra-array normalization to make intensities of different color channels comparable and (ii) inter-array normalization so that the quantities of gene products detected from different arrays can be compared. In contrast, SNP genotyping arrays are used to discriminate the allelic sequences differing by single bases. As long as these sequences can be accurately and reliably distinguished, inter-array differences may not be necessarily a major concern. By using a large number of reliably predicted homozygous SNPs as controls, we have developed very straightforward algorithms for analyzing microarray-based genotyping data. In conjunction with the dual-probe assay, our approach has achieved very high accuracy and reliability.

Algorithms (24,33) have been developed to analyze genotyping data generated by the Affymetrix SNP microarrays with very high degree of accuracy. Unfortunately, these algorithms were specifically developed for the Affymetrix system that use perfect match and mismatch probes. Such a requirement limits the application of these algorithms in three aspects: (i) most non-Affymetrix arrays do not use the mismatch probes; (ii) Affymetrix arrays contain only a small portion of known SNPs; and (iii) most genetic studies use customized specific SNP sets depending on the chromosomal regions or genes under study and many studies may not need as many SNP as those on the Affymetrix arrays. Therefore, algorithms that can be used universally for different platforms and for various SNP panels including the industrially established panels are highly desirable. Our algorithms require homozygous SNPs as controls which are abundant in any large-scale microarray analysis, and are simple and straightforward. Therefore, it can be used for this purpose.

A clustering-based algorithm for analyzing SNP genotyping data was reported by Rabbee and Speed (35) very recently. The algorithm was implemented by training the computer program with publicly available SNP genotyping data. Although the training datasets were exclusively generated by the Affymetrix SNP arrays, the algorithm may be used to analyze data generated by non-Affymetrix arrays since it does not depend on the mismatched probes. However, as pointed out by the authors, it is difficult to train the program for SNPs with alleles at a low frequency because the clustering method itself requires a certain number of samples to reach sufficient statistic confidence. When an SNP allele is at a very low frequency, the sample size required for this purpose could be very large. The same issue is true for all SNPs that do not have sufficient data available. Furthermore, when a new genotyping method is to be established, there would be not genotyping data available for training the program.

With our system, the non-concordant genotypes can be discarded to ensure a high degree of accuracy. However, an interesting issue would be how these errors are generated. Based on the dual-probe analysis, the non-concordant genotypes were found not to be randomly distributed among the samples. A small fraction ($\sim$6%) of SNPs were found to be associated with either non-concordant or undetectable signals among >30% of the samples. More detailed analyses including single-sperm analysis revealed that the majority of these SNPs were either not real SNPs (showing as heterozygous in most single-sperm samples) which could be caused by repetitive sequences or were affected by other unknown genetic variations located in the primer or probe regions. After eliminating these SNPs from further analysis, the concordant rate for the data in Table 1 would be increased from 97.06 to 98.65% with an accuracy of 99.9954%, indicating that not only is our experimental approach robust but also that our computer algorithms are highly reliable.

Other than being an appropriate method for channel normalization and background subtraction, using homozygous SNPs as controls has another advantage i.e. their large numbers. Thousands of SNPs can be analyzed by a single array, and the homozygous SNPs are in abundance. This method enhances reliability and is much more comprehensive in reflecting various factors that may contribute to data variability on the overall array as compared to using very few

controls. The cost and effort of adding extra control oligonucleotides is also eliminated. Owing to the ever-increasing capability of a small or medium sized laboratory to generate a large volume of microarray data routinely, it is expected that AccuTyping will benefit many microarray users.

ACKNOWLEDGEMENTS

We thank Dr Yong Lin and Dr Weichung Shih for their suggestion in algorithm development, and Dr Marc Ma for his valuable editorial suggestions. This research was supported by the grants R01 HG02094 from the National Human Genome Research Institute and R01 R33 CA 96309 from the National Cancer Institute, NIH, USA to H.L. Funding to pay the Open Access publication charges for this article was provided by a discretionary fund of the University of Medicine and Dentistry of New Jersey Robert Wood Johnson Medical School to H.L.

Conflict of interest statement. None declared.

REFERENCES

1. DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
2. DeRisi,J.L., Penland,L., Brown,P.O., Bittner,M.L., Meltzer,P.S., Ray,M., Chen,Y., Su,Y.A. and Trent,J.M. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet*., **14**, 457–460.
3. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a cDNA microarray. *Science*, **270**, 467–470.
4. Hegde,P., Qi,R., Abernathy,K., Gay,C., Dharap,S., Gaspard,R., Hughes,J.E., Snesrud,E., Lee,N. and Quackenbush,J. (2000) A concise guide to cDNA microarray analysis. *Biotechniques*, **29**, 548–550, 552–544, 556 passim.
5. Pollack,J.R., Perou,C.M., Alizadeh,A.A., Eisen,M.B., Pergamenschikov,A., Williams,C.F., Jeffrey,S.S., Botstein,D. and Brown,P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet*., **23**, 41–46.
6. Pollack,J.R., Sorlie,T., Perou,C.M., Rees,C.A., Jeffrey,S.S., Lonning,P.E., Tibshirani,R., Botstein,D., Borresen-Dale,A.L. and Brown,P.O. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.
7. Wang,H.Y., Luo,M., Tereshchenko,I.V., Frikker,D.M., Cui,X., Li,J.Y., Hu,G., Chu,Y., Azaro,M.A., Lin,Y. *et al*. (2005) A genotyping system capable of simultaneously analyzing >1000 single nucleotide polymorphisms in a haploid genome. *Genome Res*., **15**, 276–283.
8. O'Meara,D., Ahmadian,A., Odeberg,J. and Lundeberg,J. (2002) SNP typing by apyrase-mediated allele-specific primer extension on DNA microarrays. *Nucleic Acids Res*., **30**, e75.
9. Matsuzaki,H., Loi,H., Dong,S., Tsai,Y.Y., Fang,J., Law,J., Di,X., Liu,W.M., Yang,G., Liu,G. *et al*. (2004) Parallel genotyping of over 10 000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res*., **14**, 414–425.
10. Matsuzaki,H., Dong,S., Loi,H., Di,X., Liu,G., Hubbell,E., Law,J., Bernsten,T., Chadha,M., Hui,H. *et al*. (2004) Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nature Methods*, **1**, 109–111.
11. Hardenbol,P., Baner,J., Jain,M., Nilsson,M., Namsaraev,E.A., Karlin-Neumann,G.A., Fakhrai-Rad,H., Ronaghi,M., Willis,T.D., Landegren,U. *et al*. (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol*., **21**, 673–678.
12. Fan,J.B., Chen,X., Halushka,M.K., Berno,A., Huang,X., Ryder,T., Lipshutz,R.J., Lockhart,D.J. and Chakravarti,A. (2000) Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res*., **10**, 853–860.
13. Lindroos,K., Sigurdsson,S., Johansson,K., Ronnblom,L. and Syvanen,A.C. (2002) Multiplex SNP genotyping in pooled DNA samples by a four-colour microarray system. *Nucleic Acids Res*., **30**, e70.
14. Primdahl,H., Wikman,F.P., von der Maase,H., Zhou,X.G., Wolf,H. and Orntoft,T.F. (2002) Allelic imbalances in human bladder cancer: genome-wide detection with high-density single-nucleotide polymorphism arrays. *J. Natl Cancer. Inst*., **94**, 216–223.
15. John,S., Shephard,N., Liu,G., Zeggini,E., Cao,M., Chen,W., Vasavda,N., Mills,T., Barton,A., Hinks,A. *et al*. (2004) Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am. J. Hum. Genet*., **75**, 54–64.
16. Middleton,F.A., Pato,M.T., Gentile,K.L., Morley,C.P., Zhao,X., Eisener,A.F., Brown,A., Petryshen,T.L., Kirby,A.N., Medeiros,H. *et al*. (2004) Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. *Am. J. Hum. Genet*., **74**, 886–897.
17. Greenawalt,D.M., Cui,X., Wu,Y., Lin,Y., Wang,H.Y., Luo,M., Tereshchenko,I.V., Hu,G., Li,J.Y., Chu,Y. *et al*. (2006) Strong correlation between meiotic crossovers and haplotype structure in a 2.5-Mb region on the long arm of chromosome 21. *Genome Res*., **16**, 208–214.
18. Quackenbush,J. (2002) Microarray data normalization and transformation. *Nature Genet*., **32**, 496–501.
19. Tseng,G.C., Oh,M.K., Rohlin,L., Liao,J.C. and Wong,W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*., **29**, 2549–2557.
20. Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*., **30**, e15.
21. Fielden,M.R., Halgren,R.G., Dere,E. and Zacharewski,T.R. (2002) GP3: GenePix post-processing program for automated analysis of raw microarray data. *Bioinformatics*, **18**, 771–773.
22. Yang,M.C., Ruan,Q.G., Yang,J.J., Eckenrode,S., Wu,S., McIndoe,R.A. and She,J.X. (2001) A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol. Genomics*., **7**, 45–53.
23. Kim,J.H., Shin,D.M. and Lee,Y.S. (2002) Effect of local background intensities in the normalization of cDNA microarray data with a skewed expression profiles. *Exp. Mol. Med*., **34**, 224–232.
24. Di,X., Matsuzaki,H., Webster,T.A., Hubbell,E., Liu,G., Dong,S., Bartell,D., Huang,J., Chiles,R., Yang,G. *et al*. (2005) Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics*.
25. Pastinen,T., Raitio,M., Lindroos,K., Tainola,P., Peltonen,L. and Syvanen,A.C. (2000) A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res*., **10**, 1031–1042.
26. Pastinen,T., Kurg,A., Metspalu,A., Peltonen,L. and Syväen,A.C. (1997) Minisequencing: a specific tool for DNA analysis and diagnositics on oligonucleotide arrays. *Genome Res*., **7**, 606–614.
27. Shumaker,J.M., Metspalu,A. and Caskey,C.T. (1996) Mutation detection by solid phase primer extension. *Hum. Mutat*., **7**, 346–354.
28. Botstein,D., White,R.L., Skolnick,M. and Davis,R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet*., **32**, 314–331.
29. Kan,Y.W. and Dozy,A.M. (1978) Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc. Natl Acad. Sci. USA*, **75**, 5631–5635.
30. Cutler,D.J., Zwick,M.E., Carrasquillo,M.M., Yohn,C.T., Tobin,K.P., Kashuk,C., Mathews,D.J., Shah,N.A., Eichler,E.E., Warrington,J.A. *et al*. (2001) High-throughput variation detection and genotyping using microarrays. *Genome Res*., **11**, 1913–1925.
31. Gunderson,K.L., Steemers,F.J., Lee,G., Mendoza,L.G. and Chee,M.S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genet*., **37**, 549–554.

32. Kennedy,G.C., Matsuzaki,H., Dong,S., Liu,W.M., Huang,J., Liu,G., Su,X., Cao,M., Chen,W., Zhang,J. *et al*. (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol*., **21**, 1233–1237.
33. Liu,W.M., Di,X., Yang,G., Matsuzaki,H., Huang,J., Mei,R., Ryder,T.B., Webster,T.A., Dong,S., Liu,G. *et al*. (2003) Algorithms for large-scale genotyping microarrays. *Bioinformatics*, **19**, 2397–2403.
34. Amaratunga,D. and Cabrera,J. (2004) Processing the scanned image. In *Exploration and analysis of DNA microarray and protein array data*. Jon Wiley & Sons, Inc., Hoboken, NJ, pp. 149–53.
35. Rabbee,N. and Speed,T.P. (2006) A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics*, **22**, 7–12.