Published online 23 August 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 15* ***e106***
doi:10.1093/nar/gkl446

A genome-wide study of preferential amplification/hybridization in microarray-based pooled DNA experiments

H.-C. Yang, Y.-J. Liang, M.-C. Huang, L.-H. Li, C.-H. Lin, J.-Y. Wu, Y.-T. Chen and C.S.J. Fann*

Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan

Received March 29, 2006; Revised May 5, 2006; Accepted June 9, 2006

ABSTRACT

Microarray-based pooled DNA methods overcome the cost bottleneck of simultaneously genotyping more than 100 000 markers for numerous study individuals. The success of such methods relies on the proper adjustment of preferential amplification/hybridization to ensure accurate and reliable allele frequency estimation. We performed a hybridization-based genome-wide single nucleotide polymorphisms (SNPs) genotyping analysis to dissect preferential amplification/hybridization. The majority of SNPs had less than 2-fold signal amplification or suppression, and the lognormal distributions adequately modeled preferential amplification/hybridization across the human genome. Comparative analyses suggested that the distributions of preferential amplification/hybridization differed among genotypes and the GC content. Patterns among different ethnic populations were similar; nevertheless, there were striking differences for a small proportion of SNPs, and a slight ethnic heterogeneity was observed. To fulfill appropriate and gratuitous adjustments, databases of preferential amplification/hybridization for African Americans, Caucasians and Asians were constructed based on the Affymetrix GeneChip Human Mapping 100 K Set. The robustness of allele frequency estimation using this database was validated by a pooled DNA experiment. This study provides a genome-wide investigation of preferential amplification/hybridization and suggests guidance for the reliable use of the database. Our results constitute an objective foundation for theoretical development of preferential amplification/hybridization and provide important information for future pooled DNA analyses.

INTRODUCTION

Large-scale international human genomic/genetic studies, such as the Human Genome Project (1), International HapMap Project (2) and ENCODE Project (3), have contributed to the further understanding of the human genome and genetic disorders. These breakthroughs were made mainly possible by the advent of mature genotyping techniques [e.g. MALDI-TOF mass spectrometry (4) and oligonucleotide microarrays (5)].

Although, high-throughput genotyping techniques are readily available, the cost is still very high for large-scale genetic studies that usually involve two high-dimension variables, i.e. large sample sizes and a large number of genetic markers. Thus, the development of pooled DNA experiment (allelotyping) technology would help reduce the cost associated with large sample sizes. Allelotyping involves mixing genomic DNAs from different study subjects to reduce the number of samples, and it is an economical alternative compared with individual genotyping experiments. Allelotyping has been broadly used in disease gene association mapping (6–11), polymorphism identification/validation (12–15), and analysis of genetic diversity (16,17). This technique has been used to type single nucleotide polymorphisms (SNPs) (18–20), short tandem repeat polymorphisms (STRPs) (21), and restriction fragment length polymorphisms (RFLPs) (22). The use of allelotyping in these methods has been comprehensively reviewed (23,24).

On the other hand, the need to reduce the cost of genotyping large numbers of SNPs has prompted the development of modern microarray-based genotyping methods (5,25). For example, the Affymetrix GeneChip Human Mapping 100 K Set provides genome-wide genotyping for each individual using only a set of two dense oligonucleotide arrays (26). This technique greatly reduces the costs of primer design and assay reagents. Integration of pooled DNA experiments and microarray-based genotyping creates a very cost-effective and high-throughput marker-typing platform for conducting large-scale genetic studies (27–32).

*To whom correspondence should be addressed at Institute of Biomedical Sciences, Academia Sinica, 128, Academia Road, Section 2 Nankang, Taipei 115, Taiwan. Tel: 886 2 27899144; Fax: 886 2 27823047; Email: csjfann@ibms.sinica.edu.tw

The success of a pooled DNA experiment mainly relies on the accurate and reliable estimation of allele frequencies of genetic markers. The estimation procedure mustly consider an adjustment for an imbalance of nucleotide reaction—referred to as 'preferential amplification' and/or 'differential hybridization'. Preferential amplification/hybridization is a function of the characteristics of different nucleotides. It is a natural phenomenon that could occur during several typing stages, such as PCR amplification, primer extension, array hybridization or signal detection (23,33), and its magnitude is quantified as the coefficient of preferential amplification/hybridization (CPA) (34,35). As the name suggests, preferential amplification/hybridization means that one allele tends to be amplified or hybridized more efficiently than another. Thus, for heterozygous individuals the fluorescence intensity of two alleles containing a SNP may differ. By definition, CPA is the ratio of average peak intensities of two alleles. A CPA > 1 indicates that the first allele tends to be amplified/hybridized more efficiently than the second allele; when CPA = 1, there is no preferential amplification/hybridization; if CPA < 1, the first allele tends to be amplified/hybridized less efficiently than the second one. This factor might have little impact on genotype calling for individual genotyping; however, it distorts the estimation of allele frequency in DNA-pooling allelotyping, where allele frequencies are estimated by calculating relative peak intensities of two alleles accumulated in a DNA pool.

In a pooled DNA study, allele frequency estimates are biased if adjustments are not made for preferential amplification/hybridization. This issue has generated much research interest (34–36). Under a feasible pooled DNA experiment, the estimation bias relates to the extent of preferential amplification/hybridization and ratio of peak intensities (RPI) of two alleles (Figure 1). For example, the positive (negative) bias for CPA = 2 and RPI = 1 (CPA = 0.5 and RPI = 1) is about 0.17 for allele frequency estimation, and the positive (negative) bias for CPA = 4 and RPI = 1 (CPA = 0.25 and RPI = 1) is about 0.30.

In many studies, additional heterozygous individuals have been collected to perform a CPA adjustment (7,35,37). Moreover, for two kinds of genotyping experiments—sequential genotyping and large-scale genotyping—one can calculate the required number of heterozygous individuals to yield a reasonably accurate and precise estimation of CPA (34). The required number of heterozygotes follows a negative binomial distribution in the former experiment and a binomial distribution for the latter. Allele frequency and RPI variability affect the required number of heterozygotes. A SNP with a low minor allele frequency and/or high RPI variability requires a large number samples to attain the necessary precision. However, additional genotyping of large numbers of heterozygous individuals increases both cost and effort. Thus, the construction of a central resource (38) for information on preferential amplification/hybridization or the use of a robust empirical estimation (36) would help diminish the cost of experimentation.

In the present study, we surveyed the distribution of CPA across the human genome and established a statistical model for CPA. We have constructed a publicly available database of CPA for different ethnic populations, and we suggest guidance for the use of CPA in DNA pooling studies. The robustness of allele frequency estimation using the database was validated by a pooled DNA experiment.

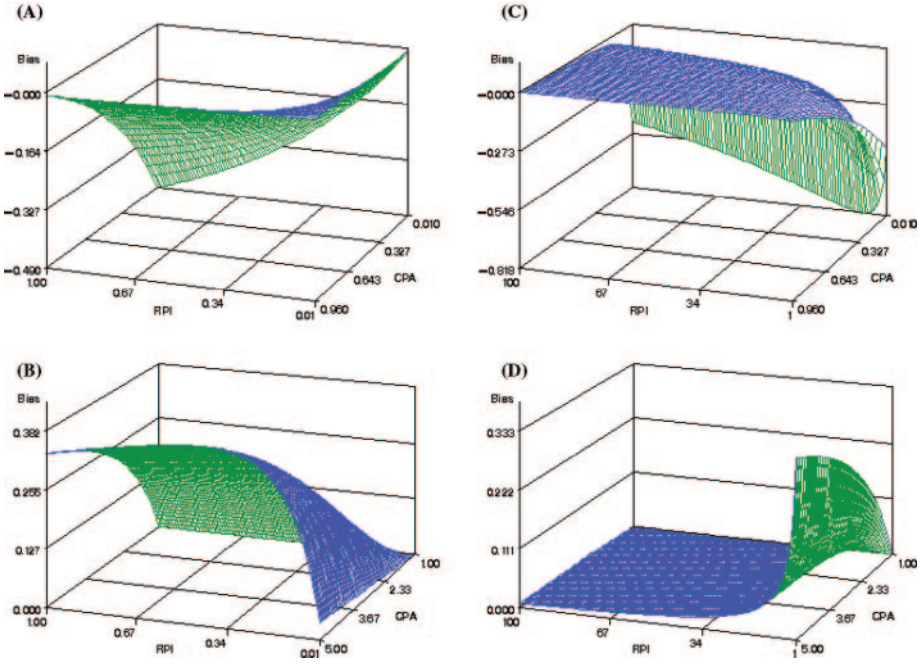

Figure 1. Relationships among CPA, estimation bias and RPI. (**A**) Both CPA and RPI are between 0.01 and 1. (**B**) CPA is between 1 and 5, and RPI is between 0.01 and 1. (**C**) CPA is between 0.01 and 1, and RPI is between 1 and 100. (**D**) CPA is between 1 and 5, and RPI is between 1 and 100.

MATERIALS AND METHODS

Study subjects

The study included 199 subjects from two panels of data. The first panel included 95 samples from the Taiwan Han Chinese Cell and Genome Bank (39). The second panel included 42 Caucasians, 42 African Americans and 20 East Asians from the Human Variation Panel (Coriell Cell Repositories). All subjects were genotyped using the Affymetrix GeneChip Human Mapping 100 K Set. Subjects in the first panel were genotyped by the National Genotyping Center (http://ngc.sinica.edu.tw) at Academia Sinica in Taiwan. Subjects in the second panel were genotyped by Affymetrix, Inc., and the data are available for download (no fee) upon request at http://www.affymetrix.com/support/technical/byproduct.affx?product=100k. Eighty-seven individuals with the best DNA quality from the first panel were selected to form a DNA pool for an allelotyping experiment.

Genotyping and genetic data

For each subject, leukocyte genomic DNA was isolated from 10 ml of blood using the Puregene genomic DNA purification kit (Gentra Systems, MN, USA). The genotyping procedure mainly followed the GeneChip Mapping Assay Protocol in the Affymetrix GeneChip Mapping 100 K Assay Manual (Affymetrix, CA). For each subject, a genotyping reaction was performed with a total of 500 ng of genomic DNA. And, 250 ng of DNA was processed by restriction enzyme digestion with XbaI and HindIII, respectively, followed by adapter ligation, PCR amplification, fragmentation, end-labeling and hybridization to microarray chips. After washing, fluorescence hybridization signals were captured using a GeneChip Scanner 3000 (Affymetrix, CA).

For each SNP genotyping, sense-strand (SS) probes and antisense-strand (AS) probes were included in seven pairs of probe quartets; each quartet contained a probe pair for each allele of a given SNP. Each probe pair contained perfect match (PM) and mismatch (MM) probe cells. Only five pairs of high-quality probe quartets (40 fluorescence signals) were selected for genotype determination. The genotype calls of SNPs were determined using the Dynamic Model (DM) algorithm (40) contained in the software GCOS version 1.2 and GDAS version 3.0 (Affymetrix). The Affymetrix GeneChip Human Mapping 100 K Set contained 116 204 SNPs with a median intermarker distance of 8.5 kb and average heterozygosity of 0.3 (26). We calculated the call rate for the genotyping of each SNP. Only data from SNPs having call rates >0.9 were included in the follow-up CPA analyses.

For pooled DNA allelotyping experiment, the DNA concentration and quality of the selected 87 subjects were determined using NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, DE). The integrity of DNA was also assessed by using gel electrophoresis with a 0.8% agarose gel. The pooling procedure was carried out by mixing equal amount of DNA at the same concentration from each individual sample. The final concentration of the pooled DNA was 50.06 ng/μl. A total of 500 ng of pooled DNA was used for allelotyping using Affymetrix GeneChip Mapping 100 K set following the genotyping procedure described above.

Statistical analyses

CPAs were estimated using three methods: arithmetic mean (35), bias-correction and geometric mean (34). Standard errors and 95% confidence intervals based on a bootstrapping procedure with 1000 replications were calculated. Appendix 1 gives details on the procedures used to estimate CPA; Appendix 2 contains information on the calculation of standard error and confidence interval. The software Pooled DNA Analyzer (PDA) (41) was used for the calculations. Based on the bias-correction CPA, genomic distributions of the estimated CPAs were investigated and model fitting was conducted based on the Shapiro–Wilks normality test (42) with Holm's multiple-test correction (43). Comparative studies of CPAs in log scale among different attribute groups (chromosomes, nucleotides, GC content and SNP location) were carried out using analysis of variance (ANOVA) analysis of regression (AOR) and analysis of covariance (ANCOVA). Correlations of CPAs among different ethnic populations were assessed by comparing Pearson correlation coefficients. The analyses were carried out using package SAS/STAT version 8 (SAS Institute, NC). To evaluate the performance of allele frequency estimation using CPAs, three types of allele frequencies were estimated by using PDA (41). Appendix 3 gives details on the allele frequency estimation. Error rates and standard errors of the unadjusted and adjusted allele frequency estimates were calculated, and Pearson correlation coefficient of the true and adjusted allele frequencies was calculated.

RESULTS

Whole-genome distribution of CPA

Figure 2 shows representative results based on the combined population containing 199 study subjects. These data present a global view of all estimated CPAs across the 22 human autosomes. For the whole genome, most of the estimated CPAs ranged from 0.5 to 2. Median, mean and standard error of the CPA estimate were 1.021, 1.077 and 0.368, respectively, and the minimum and maximum CPA were 0.238 and 4.041. The sample mode of CPAs on different chromosomes was ~1.0 (Figure 3). Extremely over-amplified or under-amplified alleles (by definition: CPA > 2 and CPA < 0.5, respectively) were observed for 5.6% of all SNPs. And the distribution of CPA was skewed to the right. These results suggest that the CPA adjustments should be applied to yield good estimates of allele frequency for all SNPs; otherwise, a serious bias may occur—especially for the SNPs with extreme CPAs (Figure 1).

The following sections offer a detailed discussion of the patterns of CPAs considering four genetic factors: chromosomes, nucleotides, GC content and SNP location. Correlations of CPAs among the three study populations are also evaluated.

Relationships between CPA and chromosomes, nucleotides, GC content and SNP location

The patterns of CPAs were quite similar across all autosomes. Means, medians and standard deviations on different chromosomes were similar to those measured for the whole genome (Figure 4A). The maximum and minimum averages of CPAs

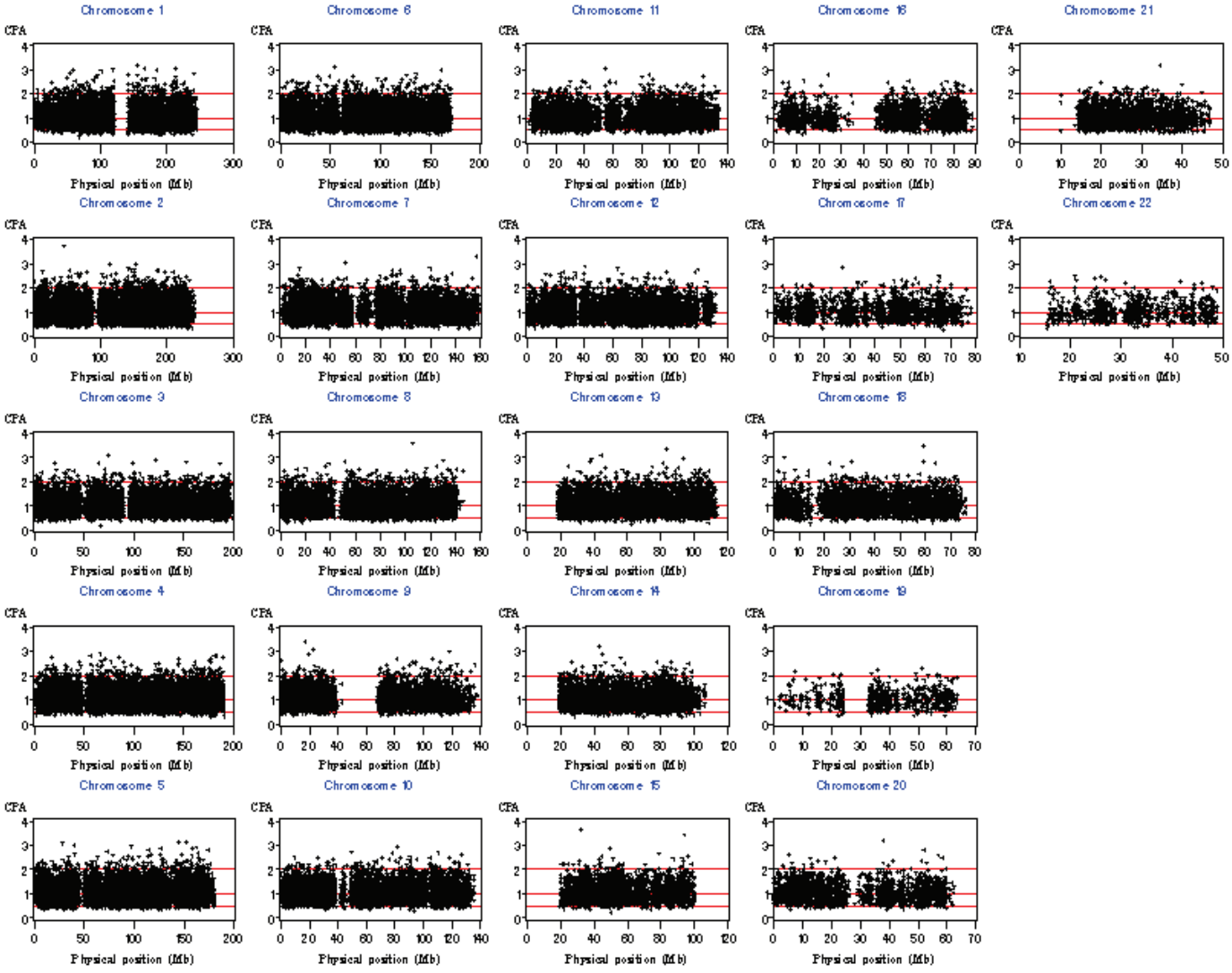

Figure 2. Scatter plots of the CPA estimates across the 22 human autosomes are given for a global view of the genome-wide CPA. The gap in each subfigure reflects the centromeric gap. In each subfigure, three red reference lines are shown for CPA = 2, 1 and 0.5 (from upper to lower).

occurred on chromosomes 18 and 19, with values 1.08 and 1.05, respectively (Table 1). One-way ANOVA showed that there was no chromosome effect on the CPA distribution ($P = 0.945$).

Preferential amplification/hybridization is mainly caused by differences in PCR amplification efficiency and hybridization and/or signal detection for alternative alleles. The degree of differential efficiency may be highly correlated with genotypes of target SNPs and percentage of GC content of probes using the hybridization-based microarray genotyping platform. Box-whisker plots for different genotypes (Figure 4B) showed that the maximum and minimum averages of CPAs were 1.272 and 0.891 for genotypes *CT* and *AG*, respectively (Table 1). One-way ANOVA revealed that the means of CPAs significantly differed among different genotypes ($P < 0.0001$). AOR showed significant correlations between CPA and the GC content of probes ($P < 0.0001$). ANCOVA was applied to further investigate the relationship between CPA and genotypes after accounting for differences in GC content. Under these conditions, there were still significant differences in CPA among genotypes ($P < 0.0001$).

We further tested signal amplification/suppression in each genotype with respect to the 4 nucleotides, yielding the following results: $C > A$ ($P < 0.0001$), $G > A$ ($P < 0.0001$), $A > T$ ($P < 0.0001$), $C > G$ ($P < 0.0001$), $C > T$ ($P < 0.0001$) and $G > T$ ($P < 0.0001$). These results suggest the relative ordering of signal amplification as $C > G > A > T$, where '>' denotes 'more efficiently amplified'. This finding provides an empirical basis for modeling preferential amplification/hybridization in different genotypes.

SNPs have different biological implications depending on their location in 3′-untranslated regions (3′-UTRs), 5′-UTRs, coding regions, introns or downstream/upstream of genes. However, one-way ANOVA showed no significant difference in CPAs between SNPs located in these different gene elements ($P = 0.4779$). The patterns of CPAs for SNPs in different locations were quite similar (Figure 4C). The mean CPA measured for all of these

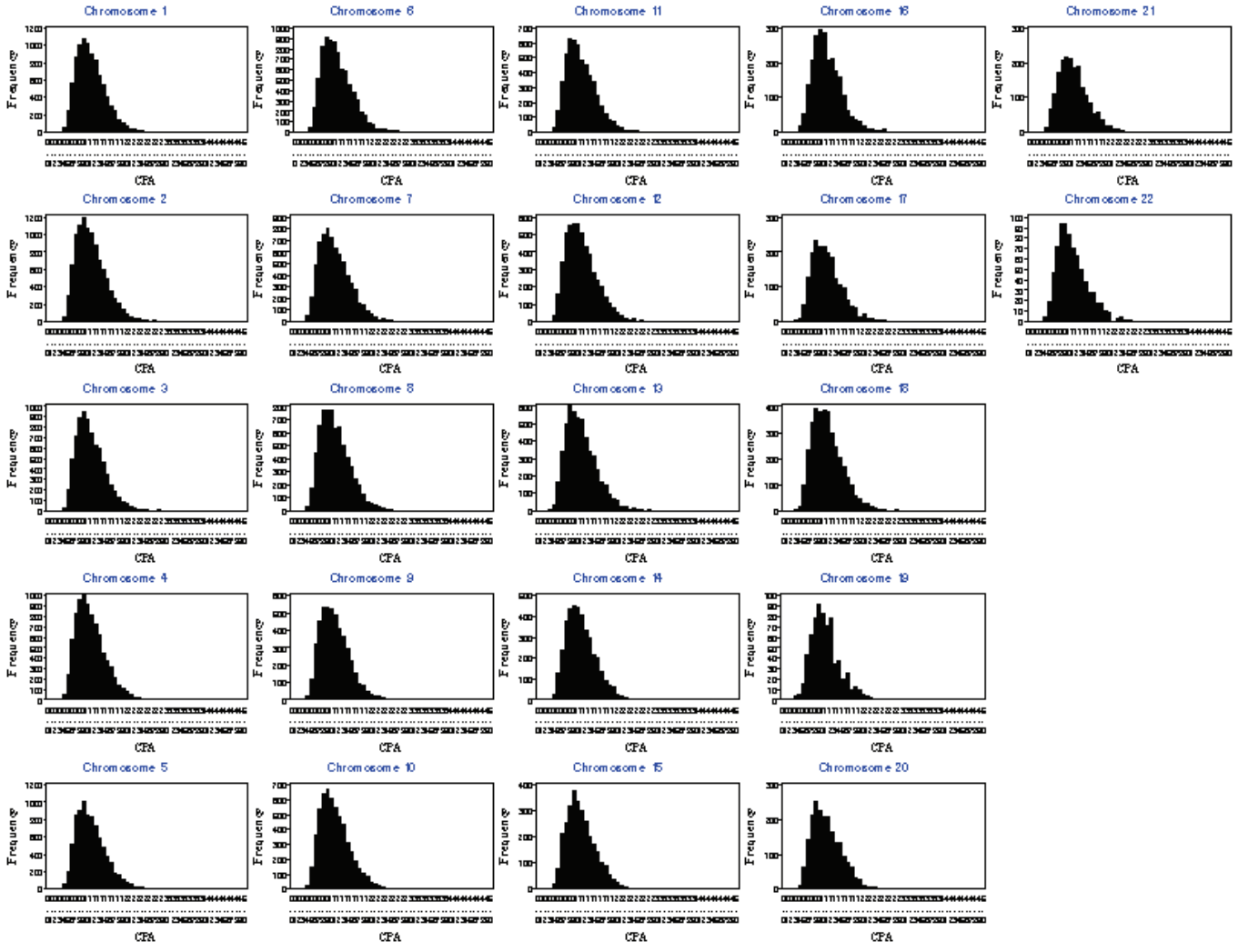

Figure 3. Histograms of CPA estimates across the 22 human autosomes show the distribution of CPA.

elements was ~1.1, and standard errors ranged from 0.34 to 0.37 (Table 1).

Effect of ethnicity on CPA

Samples in this study contained three ethnic populations, i.e. 42 African Americans, 42 Caucasians and 115 Asians (20 samples from Affymetrix and 95 Taiwanese samples). Pair-wise comparisons of CPAs among the three ethnic populations showed strong positive correlations (Figure 5); the coefficients of the Pearson correlation of CPAs between 'African American and Caucasian', 'African American and Asian' and 'Caucasian and Asian' were 0.948, 0.920 and 0.902, respectively (Pearson correlation coefficient of CPAs between the 20 Asian samples from Affymetrix and 95 Taiwanese samples was 0.996). However, there were slight discrepancies among ethnic populations and outliers. The majority of CPA differences for the pair-wise comparisons were between −1 and 1. A few striking differences between different ethnic populations were observed. For example, the maximum differences between 'African American and Caucasian', 'African American and Asian' and 'Caucasian and Asian' were 1.706, 2.119 and 2.092, respectively. The CPA-distance between African Americans and Caucasians was relatively smaller than the distance between either of these groups and Asians.

Results of ethnic-differential-CPA SNPs (i.e. SNPs with that an absolute value of a CPA difference between two ethnic populations was greater than 1) were summarized. The proportion of such SNPs across the human genome was smaller than 1.3%. The majority of these SNPs showed low minor allele frequencies (MAF) (Table 2). The proportion of the ethnic-differential-CPA SNPs within the lowest MAF bin (i.e. the MAF interval from 0 to 0.1) was ranged from 45% to 59%, but that within the greatest MAF bin (i.e. the MAF interval from 0.4 to 0.5) was <2%.

The high correlations of CPAs among different ethnic populations suggest the transferability of CPA for bias correction in most cases; however, the increased bias and reduced test power due to the observed discrepancy and a few CPA outliers suggest that an appropriate filter should be applied prior to the analysis. Therefore, we

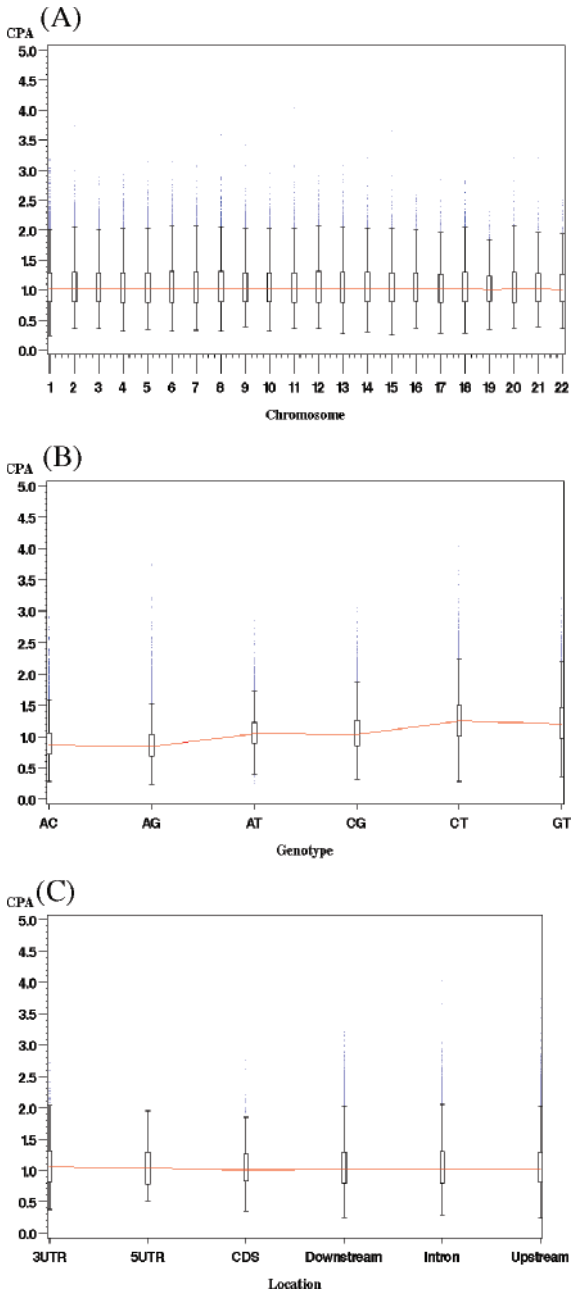

Figure 4. CPA distributions for important genetic factors. (**A**) CPA for different chromosomes. (**B**) CPA for different genotypes. (**C**) CPA for different SNP locations. Each subfigure presents box-whisker diagrams of CPAs in different categories. The red line joins the medians of CPAs across different categories. The extreme CPA values outside the 1.5 interquartile range are indicated by blue dots.

calculated both ethnic-specific and combined-population CPAs and suggest criteria for their use. A large sample size provides high reliability. The use of CPA from combined samples is suggested when CPA discrepancies among ethnic populations are small. However, a population-specific CPA should be considered to avoid misuse of CPA. The suggested criterion of SNP transferability between two populations is $|\ln(\hat{\kappa}_1) - \ln(\hat{\kappa}_2)| \leqslant (1.96 \times \sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2})$, where $\hat{\kappa}_1$ and $\hat{\kappa}_2$ are the estimated CPAs in two populations,

Table 1. Descriptive statistics of CPA with regard to different factors

	Number of SNPs	Mean of CPA	S.E. of CPA	Maximum of CPA	Minimum of CPA
Chromosome					
1	8955	1.077	0.364	3.177	0.238
2	10 094	1.081	0.371	3.738	0.357
3	7618	1.076	0.357	2.894	0.350
4	8370	1.074	0.371	2.925	0.316
5	8143	1.078	0.368	3.148	0.346
6	7843	1.079	0.374	3.143	0.318
7	6802	1.078	0.374	3.064	0.332
8	6778	1.083	0.365	3.589	0.321
9	4665	1.073	0.359	3.428	0.387
10	5534	1.078	0.365	2.951	0.313
11	5239	1.072	0.365	4.041	0.348
12	5127	1.085	0.378	2.896	0.359
13	5073	1.077	0.376	3.084	0.273
14	3913	1.077	0.366	3.208	0.301
15	2961	1.078	0.368	3.647	0.249
16	2322	1.077	0.358	2.577	0.367
17	1916	1.068	0.357	2.839	0.280
18	3487	1.083	0.370	2.832	0.284
19	669	1.049	0.339	2.310	0.344
20	2032	1.072	0.367	3.213	0.370
21	1856	1.076	0.365	3.196	0.388
22	731	1.072	0.370	2.493	0.351
Genotype					
AC	10 002	0.916	0.292	2.902	0.283
AG	37 489	0.891	0.280	3.738	0.238
AT	6761	1.077	0.275	2.848	0.249
CG	10 572	1.079	0.333	3.065	0.316
CT	38 237	1.272	0.370	4.041	0.273
GT	7067	1.239	0.374	3.208	0.365
Location					
3′-UTR	683	1.098	0.362	2.725	0.379
5′-UTR	60	1.067	0.340	1.948	0.516
Coding region	655	1.067	0.340	2.751	0.350
Intron	38 898	1.078	0.367	4.041	0.290
Downstream	32 591	1.075	0.367	3.213	0.249
Upstream	37 241	1.078	0.369	3.728	0.238

and $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the estimated standard errors of $\ln(\hat{\kappa}_1)$ and $\ln(\hat{\kappa}_2)$.

Model fitting for CPA

Statistical goodness-of-fit procedures were carried out to model CPAs. Distributions of CPAs were skewed toward the right (Figures 3 and 5). The conjecture that CPAs in a log scale follow a Gaussian distribution in each subgroup was formally tested using the Shapiro–Wilks normality test (42). More than 90% of the goodness-of-fit tests passed the normality check after Holm's multiple-test correction (43), which controls family-wise errors smaller than 1%, suggesting that the following lognormal distributions provide a good approximation to the CPA distributions:

$$f(\hat{\kappa}) = (\hat{\kappa}\hat{\sigma}\sqrt{2\pi})^{-1} \times \exp\{-[(\hat{\kappa} - \hat{\mu})/\hat{\sigma}]^2/2\}, 0 < \hat{\kappa} < \infty,$$

where the maximum likelihood estimators of mean and variance of the estimated CPA $\hat{\kappa}$ are $\exp[\hat{\mu} + (\hat{\sigma}^2/2)]$ and $[\exp\{\hat{\sigma}^2\} - 1] \cdot \exp\{2\hat{\mu} + \hat{\sigma}^2\}$, and $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and variance of $\ln(\hat{\kappa})$, respectively. This finding establishes an empirical basis for modeling preferential amplification/hybridization and provides support for theoretical development in pooled DNA analyses.

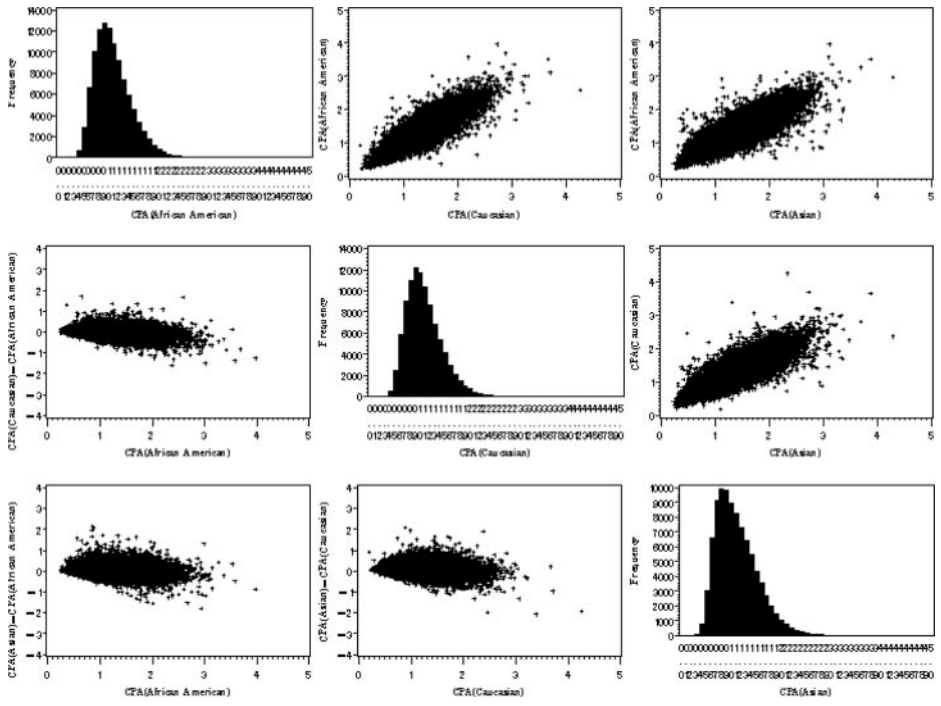

Figure 5. The CPA distribution within each population and CPA discrepancy between any two different ethnic populations. The subfigures in the diagonal show the CPA histograms in African Americans, Caucasians and Asians in order. The subfigures in the upper diagonal part are the scatter plots of CPAs between any two ethnic populations. The subfigures in the lower diagonal part show the discrepancy of CPAs between any two ethnic populations.

Table 2. The distributions of ethnic-differential-CPA SNPs between two ethnic groups

MAF	P{|CPA(African)-CPA(Caucasian)| > 1}		P{|CPA(African)-CPA(Asian)| > 1}		P{|CPA(Asian)-CPA(Caucasian)| > 1}	
	Percentage 1	Percentage 2	Percentage 1	Percentage 2	Percentage 1	Percentage 2
0.0–0.1	0.04	53.57	0.12	58.82	0.08	45.61
0.1–0.2	0.02	21.43	0.08	30.88	0.09	42.11
0.2–0.3	0.03	21.43	0.02	5.88	0.03	10.53
0.3–0.4	0.01	3.57	0.02	4.41	0.00	0.00
0.4–0.5	0.00	0.00	0.00	0.00	0.01	1.75

The distributions of ethnic-differential-CPA SNPs for African-Caucasian, African-Asian and Asian-Caucasian are shown in order. In each panel, the percentage in the first column (Percentage 1) was calculated by dividing the number of ethnic-differential-CPA SNPs by the total number of SNPs in the MAF bin. The percentage in the second column (Percentage 2) was calculated by dividing the number of ethnic-differential-CPA SNPs in a MAF bin by the total number of ethnic-differential-CPA SNPs.

The database of CPA

We constructed a CPA database containing 116 204 SNPs for use in future pooled DNA studies. The database provides three types of information: (i) SNP descriptions (chromosome number, probe set and physical position, genotype and SNP location); (ii) results from all samples (SNP call rate, allele frequency, locus heterozygosity, unadjusted and adjusted *P*-values for the test of Hardy–Weinberg Equilibrium); (iii) results from heterozygous individuals (number of heterozygous individuals used in the CPA calculation, three CPA estimates, the corresponding standard errors and 95% confidence intervals of CPA). The SNP information and annotation data are publicly available from the Affymetrix website http://www.affymetrix.com/support/technical/byproduct.affx?product= 100k. Our database is now freely accessible online at http://www.ibms.sinica.edu.tw/%7Ecsjfann/first%20flow/database.htm; the interface is shown in Figure 6. Results for genome-wide or chromosome-wide and combined-population or population-specific analyses can be obtained from the website. This database enables users to study and adjust for preferential amplification/hybridization in pooled DNA analyses.

Evaluation of allele frequency estimation in pooled samples

We carried out pooled DNA allelotyping experiments to evaluate the utility of our CPA database. Three types of allele frequencies were calculated for comparison: (i) the estimated allele frequency based on unadjusted intensity data; (ii) the estimated allele frequency based on the bias-correction CPA adjusted intensity data and (iii) the true allele frequency based on individual genotyping result. With the CPA adjustment, the bias of allele frequency estimation was obviously reduced (Figure 7 and Table 3). The error rate and standard error for the adjusted allele frequency estimates were consistently lower than that for the unadjusted estimates. Moreover, the adjusted and true allele frequencies showed a high correlation of 0.99, demonstrating the good performance

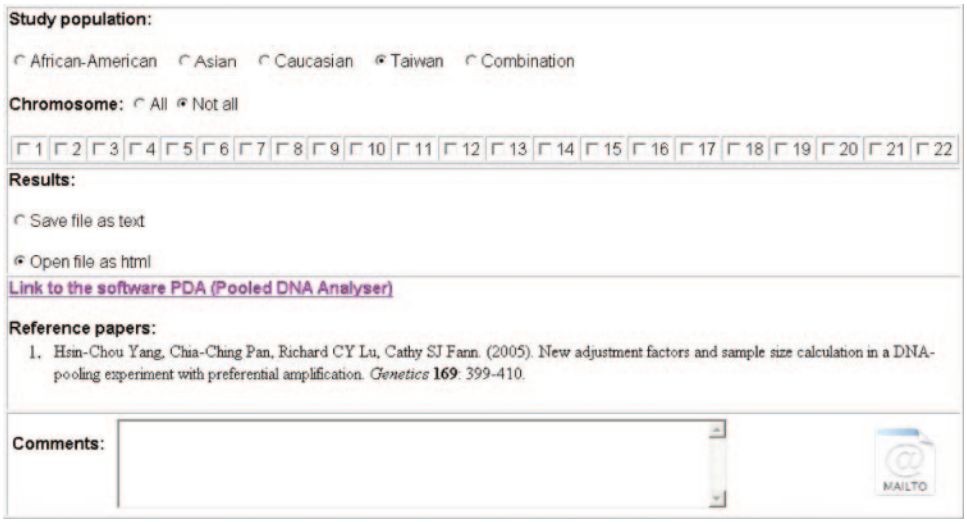

Figure 6. The interface of our web-based CPA database. The first item provides an option for outputting either population-specific or combined-population CPAs. The second option provides an option for outputting either chromosome-wide or genome-wide CPAs. Results can either be shown online or saved as an html file.

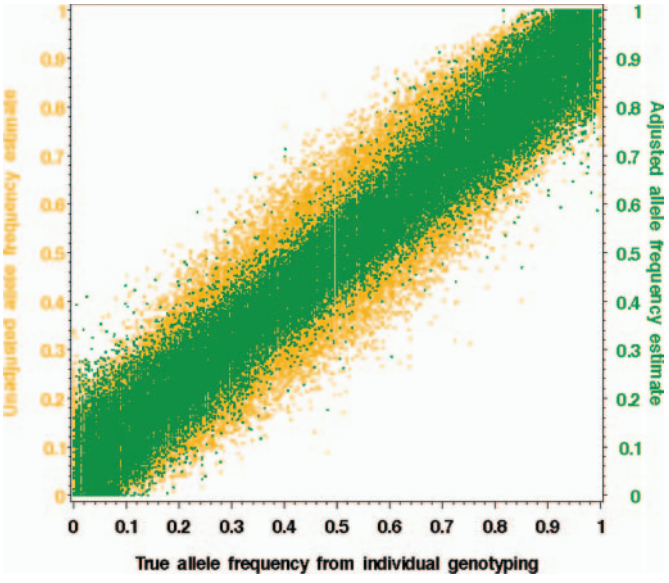

Figure 7. Scatter plots of the unadjusted and adjusted allele frequencies versus the true allele frequency. Each point denotes the coordinate of estimated allele frequency versus true allele frequency for a SNP. The yellow dots denote the results of the unadjusted allele frequency and green dots denote the results of the adjusted allele frequency.

of CPA adjustment for pooled DNA in allele frequency estimation.

CONCLUSION AND DISCUSSION

In summary, preferential amplification/hybridization plays an important role in analyses that rely on fluorescence intensity data. The adjustment for preferential amplification/hybridization has been developed to estimate allele frequency (34,35) and incorporated in test statistics for association mapping (34,44) in pooled DNA analyses. The method has been applied to microarray-based pooled DNA analysis

Table 3. Mean error rates and standard errors of the unadjusted allele frequency and adjusted frequency estimates

True allele frequency	Unadjusted estimate of allele frequency		Adjusted estimate of allele frequency	
	Mean error rate	Standard error	Mean error rate	Standard error
0.0–0.1	0.051	0.042	0.049	0.041
0.1–0.2	0.060	0.047	0.046	0.035
0.2–0.3	0.067	0.051	0.041	0.032
0.3–0.4	0.073	0.054	0.035	0.029
0.4–0.5	0.076	0.055	0.032	0.027
0.5–0.6	0.076	0.054	0.032	0.026
0.6–0.7	0.073	0.053	0.036	0.030
0.7–0.8	0.067	0.049	0.042	0.033
0.8–0.9	0.058	0.045	0.047	0.036
0.9–1.0	0.046	0.041	0.046	0.040

(29,30). Only a few papers provide systematic investigations at preferential amplification/hybridization. One study (36) has discussed preferential amplification based on 152 SNPs genotyped using the platform SnaPshot (ABI, CA). Another study (38) developed a database of adjustment index, *R*, based on results from 100 Caucasians using the GeneChip Human Mapping 10 K Array Set. Our study investigated the whole-genome behavior of preferential amplification/hybridization based on ~200 individuals using the GeneChip Human Mapping 100 K Array Set. Instead of utilizing the index *R* (38), our study focused on the index CPA because it has been developed in rich literature (34,35,44) and broadly used in practical applications (7,37,45). Our results show that CPAs are dependent on the GC content of probes and nucleotide characteristics based on ANOVA and AOR. Moreover, the results were confirmed by four-way ANCOVA, which simultaneously considers GC content, the effect of nucleotides, chromosome and SNP location. The finding that nucleotides *G* and *C* cause greater signal amplification compared with *A* and *T* is consistent with the principle that base pairing strength for *GC* (three hydrogen bonds) is higher than that for *AT* (two hydrogen bonds). Moreover, we found

that lognormal distributions properly fit CPAs and, therefore, future statistical modeling of CPA based on the distribution can be supported empirically.

We also investigated the impact of ethnicity on the estimation of preferential amplification/hybridization. This issue has never been investigated, although ethnic heterogeneity has been recognized as a critically important factor in population genetics and gene mapping studies (46,47). The CPA data transferability or data combination was only suggested upon a proper pre-selection. We constructed CPA databases based on a high-density SNP panel for specific population and combined population. To our knowledge, our study is the first to systematically discuss the whole-genome behavior of preferential amplification/hybridization with consideration of ethnicity for the CPA adjustment.

Association mapping is one of the important applications of pooled DNA analyses. Pooled DNA association tests compare differences in allele frequencies between case and control groups. A fundamental assumption of this kind of statistical test is that the two alleles relevant to a SNP are independent, i.e. that they satisfy the Hardy–Weinberg Equilibrium. Results of the association tests may be misleading if the SNPs in question violate this assumption. Therefore, in addition to marker information and CPA adjustment, the CPA database also provides the measurement of marker informativeness and verifies the Hardy–Weinberg Equilibrium. *P*-values for an exact test for the Hardy–Weinberg Equilibrium (48), with or without consideration of multiple comparisons, are provided to remind users of potential violation of this principle.

Completion of DNA-pooling association mapping relies on a well-established analytical system that includes analysis strategy and user-friendly software. Previous work established a multistage strategy consisting of adjustment of preferential amplification/hybridization, allele frequency estimation, single-point association test, multipoint association test and confirmatory association test (24). The software PDA (41) was designed to analyze data from allele-specific genotyping and array-based genotyping platforms under a multistage framework. Integration of these free resources, namely the CPA database, analytical strategy and PDA, provides a powerful strategy with which to estimate allele frequency and perform disease gene mapping in pooled DNA studies.

The costs and benefits associated with genetic typing are major concerns when assessing the feasibility of large-scale genetic studies. New-generation methods for pooled DNA analysis combine conventional DNA pooling techniques with modern microarray-based genotyping methods to meet cost-benefit requirements and achieve high-throughput and exceptional validity-reliability for large-scale genome screens. Current pooled DNA methods hardly provide detailed information for respective individuals, but such methods are quite useful when studies focus on statistical/biological inferences via allele frequency. Such methods are likely to be applied broadly in future genetic studies.

We are extending this project in the several ways. (i) We are increasing the sample size of heterozygous individuals. CPA can be estimated precisely using a modest number of heterozygous individuals (34,35); nevertheless, more heterozygous samples further increase the precision of CPA estimation. Results of this study were established on the basis of ~200 individuals, and more samples are being collected to enhance the study reliability. (ii) We are increasing the density of SNP markers. We are extending the investigation of the GeneChip Human Mapping 100 K Array Set to the Human Mapping 500 K Array Set. This new mapping panel has a median intermarker physical distance of 2.5 kb and average heterozygosity of 0.25; thus, it will provide more than twice the genetic power and SNP content relative to the 100 K Set, thereby facilitating fine-specificity positional cloning studies of complex disorders via pooled DNA analyses (online document at http://www.affymetrix.com/products/arrays/specific/500k.affx). (iii) We are extending the study populations. Although, CPAs of the majority of SNPs may be portable for different ethnic populations, the application of population-specific data to other ethnic groups runs the risk of increasing the estimation bias and reducing the testing power for some SNPs. Thus, we are collecting more samples from different groups to further expand the applicability of our database. (iv) We are enlarging the pool size. In addition to a pool size of 87, we have already carried out pooled DNA allelotyping experiments with pool sizes of 10, 30 and 50. The comparisons showed no significant differences of results among the considered pool sizes, indicating that a pool size up to 87 was still within the applicable range of this type of experiment. Therefore, we are enthusiastically conducting experiments with more samples to investigate the limitation of pool size in microarray-based pooled DNA experiments.

ACKNOWLEDGEMENTS

The authors appreciate the support from the Institute of Biomedical Sciences, Taiwan National Genotyping Center and National Clinical Core. The authors thank Mr Vincent W. Tseng for constructing the website for the CPA database. The authors also thank the two anonymous reviewers' constructive suggestions, which have largely improved the presentation of this paper. Funding to pay the Open Access-publication charges for this article was provided by Institute of Biomedical Sciences Academia Sinica Taiwan.

Conflict of interest statement. None declared.

REFERENCES

1. The International Human Genome Mapping Consortium (2001) A physical map of the human genome. *Nature*, **409**, 934–941.
2. The International HapMap Consortium (2003) The International HapMap project. *Nature*, **426**, 789–796.
3. The ENCODE Project Consortium (2004) The ENCODE (encyclopedia of DNA elements) project. *Science*, **306**, 636–640.
4. Pusch,W., Wurmbach,J.-H., Tiele,H. and Kostrzewa,M. (2002) MALDI-TOF mass spectrometry-based SNP genotyping. *Pharmacogenomics*, **3**, 537–548.
5. 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.
6. Barcellos,L.F., Klitz,W., Field,L.L., Tobias,R., Bowcock,A.M., Wilson,R., Nelson,M.P., Nagatomi,J. and Thomson,G. (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet*., **61**, 734–747.
7. Mohlke,K.L., Erdos,M.R., Scott,L.J., Fingerlin,T.E., Jackson,A.U., Silander,K., Hollstein,P., Boehnke,M. and Collins,F.S. (2002) High-throughput screening for evidence of association by using mass

spectrometry genotyping on DNA pools. *Proc. Natl Acad. Sci. USA*, **99**, 16928–16933.
8. Bansal,A., van den Boom,D., Kammerer,S., Honisch,C., Adam,G., Cantor,C.R., Kleyn,P. and Braun,A. (2002) Association testing in DNA pooling: an effective initial screen. *Proc. Natl Acad. Sci. USA*, **99**, 16871–16874.
9. Jawadi,A., Bader,J.S., Purcell,S., Cherny,S.S. and Sham,P. (2002) Family-based association tests for quantitative traits using pooled DNA. *Eur. J. Hum. Genet.*, **20**, 125–132.
10. Hinds,D.A., Seymour,A.B., Durham,K., Banerjee,P., Ballinger,D.G., Milos,P.M., Cox,D.R., Thompson,J.F. and Frazer,K.A. (2004) Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. *Hum. Genomics*, **1**, 421–434.
11. Zou,G. and Zhao,H. (2005) Family-based association tests for different family structures using pooled DNA. *Ann. Hum. Genet.*, **69**, 429–442.
12. Wolford,J.K., Blunt,D., Ballecer,C. and Prochazka,M. (2000) High-throughput SNP detection by using DNA pooling and denaturing high performance liquid chromatography (DHPLC). *Hum. Genet.*, **107**, 483–487.
13. Buetow,K.H., Edmonson,M., MacDonald,R., Clifford,R., Yip,P., Kelley,J., Little,D.P., Strausberg,R., Koester,H., Cantor,C.R. *et al.* (2001) High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc. Natl Acad. Sci. USA*, **98**, 581–584.
14. Nelson,M.R., Marnellos,G., Kammerer,S., Hoyal,C.R., Shi,M.M., Cantor,C.R. and Braun,A. (2004) Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome Res.*, **14**, 1664–1668.
15. Yang,H.C., Lin,C.H., Hung,S.I. and Fann,C.S.J. (2006) Polymorphism validation using DNA pools prior to conducting large-scale genetic studies. *Ann. Hum. Genet.*, **70**, 350–359.
16. Dubreuil,P., Rebourg,C., Merlino,M. and Charcosset,A. (1999) Evaluation of a DNA pooled-sampling strategy for estimating the RFLP diversity of maize populations. *Plant Mol. Biol. Rep.*, **17**, 123–138.
17. Hillel,J., Groenen,M.A.M., Tixier-Boichard,M., Korol,A.B., David,L., Kirzhner,V.M., Burke,T., Barre-Dirie,A., Crooijmans,R.P.M.A., Elo,K. *et al.* (2003) Biodiveristy of 52 chicken populations assessed by microsatellite typing of DNA pools. *Genet. Sel. Evol.*, **35**, 533–557.
18. Werner,M., Sych,M., Herbon,N., Illig,T., König,I.R. and Wjst,M. (2002) Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectrometry. *Hum. Mutat.*, **20**, 57–64.
19. Le Hellard,S., Ballereau,S.J., Visscher,P.M., Torrance,H.S., Pinson,J., Morris,S.W., Thomson,M.L., Semple,C.A., Muir,W.J., Blackwood,D.H. *et al.* (2002) SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Res.*, **30**, e74.
20. Barratt,B.J., Payne,F., Rance,H.E., Nutland,S., Todd,J.A. and Clayton,D.G. (2003) Identification of the sources of error in allele freqeuncy estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.*, **66**, 393–405.
21. Shaw,S.H., Carrasquillo,M.M., Kashuk,C., Puffenberger,E.G. and Chakravarti,A. (1998) Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.*, **8**, 111–123.
22. Arnheim,N., Strange,C. and Erlich,H. (1985) Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of HLA class II loci. *Proc. Natl Acad. Sci. USA*, **85**, 6970–6974.
23. Sham,P., Bader,J.S., Craig,I., O'Donovan,M. and Owen,M. (2002) DNA pooling: a tool for large-scale association studies. *Nature Rev. Genet.*, **3**, 862–871.
24. Yang,H.C. and Fann,C.S.J. (2007) Association mapping using pooled DNA. In Collins,A. (ed.), *Linkage Disequilibrium and Association Mapping*. The Humana Press, Inc., USA. In Press.
25. 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.
26. Matsuzaki,H., Dong,S., Loi,H., Di,X., Liu,G., Hubbell,E., Law,J., Berntsen,T., Chadha,T., Chadha,M. *et al.* (2004) Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nature Meth.*, **1**, 109–111.
27. Uhl,G.R., Lin,Q.R., Walther,D., Hess,J. and Naiman,D. (2001) Polysubstance abuse-vulnerability genes: genome scans for association, using 1004 subjects and 1494 single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **69**, 1290–1300.
28. Lindroos,K., Sigurdsson,S., Johansson,K., Ronnblom,L. and Syvanen,A.-C. (2002) Multiplex SNP genotyping in pooled DNA samples by a four-color microarray system. *Nucleic Acids Res.*, **30**, e70.
29. Butcher,L.M., Meaburn,E., Liu,L., Fernandes,C., Hill,L., Al-Chalabi,A., Plomin,R., Schalkwyk,L. and Craig,I.W. (2004) Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav. Genet.*, **34**, 549–555.
30. Meaburn,E., Butcher,L.M., Liu,L., Fernandes,C., Hansen,V., Al-Chalabi,A., Plomin,R., Craig,I. and Schalkwyk,L.C. (2005) Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics*, **6**, 52.
31. Meaburn,E., Butcher,L.M., Schalkwyk,L.C. and Plomin,R. (2006) Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. *Nucleic Acids Res.*, **34**, e28.
32. Macgregor,S., Visscher,P.M. and Montgomery,G. (2006) Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates. *Nucleic Acids Res.*, **34**, e55.
33. Norton,N., Williams,N.M., O'Donovan,M.C. and Owen,M.J. (2004) DNA pooling as a tool for large-scale association studies in complex traits. *Ann. Med.*, **36**, 146–152.
34. Yang,H.C., Pan,C.C., Lu,R.C.Y. and Fann,C.S.J. (2005) New adjustment factors and sample size calculation in a DNA-pooling experiment with preferential amplification. *Genetics*, **169**, 399–410.
35. Hoogendoorn,B., Norton,N., Kirov,G., Williams,N., Hamshere,M.L., Spurlock,G., Austin,J., Stephens,M.K., Buckland,P.R., Owen,M.J. *et al.* (2000) Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum. Genet.*, **107**, 488–493.
36. Moskvina,V., Norton,N., Williams,N., Holmans,P., Owen,M. and O'Donovan,M. (2005) Streamlined analysis of pooled genotype data in SNP-based association studies. *Genet. Epidemiol.*, **28**, 273–282.
37. Xu,H., Knight,J., Brookes,K., Mill,J., Sham,P., Craig,I., Taylor,E. and Asherson,P. (2005) DNA pooling analysis of 21 norepinephrine transporter gene SNPs with attention deficit hyperactivity disorder. *Am. J. Med. Genet. B*, **134**, 115–118.
38. Simpson,C.L., Knight,J., Butcher,L.M., Hansen,V.K., Meaburn,E., Schalkwyk,L.C., Craig,I.W., Powell,J.F., Sham,P.C. and Al-Chalabi,A. (2005) A central resource for accurate allele frequency estimation from pooled DNA genotyped on DNA microarrays. *Nucleic Acids Res.*, **33**, e25.
39. Pan,W.H., Fann,C.S.J., Wu,J.Y., Hung,Y.T., Ho,M.S., Tai,T.H., Chen,Y.J., Liao,C.J., Yang,M.L., Cheng,A.T.A. *et al.* (2006) Han Chinese cell and genome bank in Taiwan: purpose, design and ethical considerations. *Hum. Hered.*, **61**, 27–30.
40. Huang,J., Wei,W., Zhang,J., Liu,G., Bignell,G.R., Stratton,M.R., Futreal,P.A., Wooster,R., Jones,K.W. and Shapero,M.H. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics*, **1**, 287–299.
41. Yang,H.C., Pan,C.C., Lin,C.Y. and Fann,C.S.J. (2006) PDA: pooled DNA analyzer. *BMC Bioinformatics*, **7**, 233.
42. Shapiro,S.S. and Wilks,M.B. (1965) An analysis of variance test for normaility (complete samples). *Biometrika*, **52**, 591–611.
43. Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.
44. Visscher,P.M. and Le Hellard,S. (2003) Simple method to analyze SNP-based association studies using DNA pools. *Genet. Epidemiol.*, **24**, 291–296.
45. Johnson,M.P. and Griffiths,L.R. (2005) A genetic analysis of serotonergic biosynthetic and metabolic enzymes in migraine using a DNA pooling approach. *J. Hum. Genet.*, **50**, 607–610.
46. Akey,J.M., Eberle,M.A., Rieder,M.J., Carlson,C.S., Shriver,M.D., Nickerson,D.A. and Kruglyak,L. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLos Biol.*, **2**, 1591–1599.
47. Hinds,D.A., Stuve,L.L., Nilsen,G.B., Halperin,E., Eskin,E., Ballinger,D.G., Frazer,K.A. and Cox,D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.

48. Guo,S.W. and Thompson,E.A. (1992) Performing the exact test for Hardy–Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361–372.

APPENDIX

Appendix 1. The estimation of CPA

For each individual SNP, the intensity data contain 40 fluorescent signals, $[f(a, b, c, d), a \in A, b \in B, c \in C, d \in D]$, where $A = \{\text{SS, AS}\}$, $B = \{\text{PM,MM}\}$, $C = \{\text{Allele}_1, \text{Allele}_2\}$ and $D = \{\text{Quartet}_1, \ldots, \text{Quartet}_5\}$. Let n_h denote the number of heterozygous individuals. We introduce the CPA estimating procedure, which was formulated using the feature extraction procedure (25,30) and the bias-correction procedure (34).

In the feature extraction stage, let

$$f^*(a, \text{PM}, c, d) = \max\{f(a, \text{PM}, c, d) - \frac{1}{2}\sum_{t \in C} f(a, \text{MM}, t, d), 0\}$$

denote the adjusted PM signal of strand a for allele c in the dth quartet. The signals were calibrated by subtracting the background noise measured by intensities of mismatched cells. Based on the definition of relative allele signal (RAS) (25), the RAS of allele c was a ratio of the fluorescence signal of the allele for perfectly matched cells as follows:

$$\text{RAS}(a, \text{PM}, c, d) = \frac{f^*(a, \text{PM}, c, d)}{\sum_{t \in C} f^*(a, \text{PM}, t, d)}.$$

Medians of RAS over five quartets for the sense and antisense strands were calculated separately, and the mean of the two medians yielded the individual RAS (IRAS) of the first allele for the ith individual as follows:

$$\text{IRAS}_1(i) = \frac{1}{2}\underset{d=1,\ldots,5}{\text{median}}\{\text{RAS}(\text{SS}, \text{PM}, 1, d)\} + \frac{1}{2}\underset{d=1,\ldots,5}{\text{median}}\{\text{RAS}(\text{AS}, \text{PM}, 1, d)\}, i = 1, \ldots, n_h \text{ and}$$

$$\text{IRAS}_2(i) = 1 - \text{IRAS}_1(i).$$

The sample means of IRAS for allele 1 and allele 2 over all heterozygous individuals were

$$\overline{\text{IRAS}}_1 = n_h^{-1} \times \sum_{i=1}^{n_h} \text{IRAS}_1(i)$$

and $\overline{\text{IRAS}}_2 = n_h^{-1} \times \sum_{i=1}^{n_h} \text{IRAS}_2(i)$.

In the stage where CPA was estimated, Hoogendoorn's CPA (35) was calculated as follows:

$$\hat{\kappa}_H = n_h^{-1} \times \sum_{i=1}^{n_h}[\text{IRAS}_1(i)/\text{IRAS}_2(i)].$$

The other two CPAs, unbiased CPA and geometric CPAs (34), can be estimated as follows:

$$\hat{\kappa}_U = \hat{\kappa}_H + \frac{n_h}{n_h - 1}\left(\frac{\overline{\text{IRAS}_1}}{\overline{\text{IRAS}_2}} - \hat{\kappa}_H\right) \text{ and } \hat{\kappa}_G = \left[\prod_{i=1}^{n_h}\frac{\text{IRAS}_1(i)}{\text{IRAS}_2(i)}\right]^{1/n_h}.$$

A reference allele (i.e. allele 2) must be specified while CPA is estimated. Throughout this study, all heterozygous genotypes were rearranged as the following six genotypes: *AC*, *AG*, *AT*, *CG*, *CT* and *GT*. The second allele in each of the previous six genotypes was regarded as the reference allele.

Appendix 2. The standard error and empirical distribution of the estimated CPA

We calculated the standard error of the estimated CPA based on a bootstrapping procedure. Denote $\{[\text{IRAS}_1(i), \text{IRAS}_2(i)], i = 1, \ldots, n_h\}$ as the pairs of IRAS values of a SNP for n_h heterozygous individuals. Assume that $\{\text{IRAS}_1(i), i = 1, \ldots, n_h\}$ follows a beta distribution with a probability density function

$$f(y) = [\Gamma(\alpha + \beta)/\Gamma(\alpha)\Gamma(\beta)] \times y^{\alpha-1}(1-y)^{\beta-1}, 0 \leqslant y \leqslant 1.$$

Based on the data for the 40 fluorescent signals, we calculated the IRAS of the heterozygous individuals, $\{[\text{IRAS}_1(i), \text{IRAS}_2(i)], i = 1, \ldots, n_h\}$, as illustrated in Appendix 1. Let the sample mean and standard deviation of $\{\text{IRAS}_1(i), i = 1, \ldots, n_h\}$ be $\overline{\text{IRAS}_1}$ and S_{IRAS}. Then, the moment estimates of parameters α and β can be calculated as follows:

$$\hat{\alpha} = \{[\overline{\text{IRAS}_1}^2(1 - \overline{\text{IRAS}_1})]/S_{\text{IRAS}_1}\} - \overline{\text{IRAS}_1} \quad \textbf{1}$$

and

$$\hat{\beta} = \{[\overline{\text{IRAS}_1}(1 - \overline{\text{IRAS}_1})]/S_{\text{IRAS}_1}\} - (\hat{\alpha} + 1). \quad \textbf{2}$$

Bootstrap samples were drawn from the empirical distribution $\text{Beta}(\hat{\alpha}, \hat{\beta})$. For the bth bootstrap sample, $\{\text{IRAS}_c^{(b)}(i), i = 1, \ldots, n_h, c \in C\}$, CPA was re-estimated to obtain $\hat{\kappa}^{(b)}$. The procedure was repeated L times to obtain $\{\hat{\kappa}^{(b)}, b = 1, \ldots, L\}$. Finally, we took the sample standard deviation of $\{\hat{\kappa}^{(b)}, b = 1, \ldots, L\}$ to calculate the standard error of the CPA estimator, i.e. $S_{\hat{\kappa}} = [\sum_{b=1}^{L}(\hat{\kappa}^{(b)} - \tilde{\kappa})^2/(L-1)]^{1/2}$, where $\tilde{\kappa} = \sum_{b=1}^{L}\hat{\kappa}^{(b)}/L$.

Appendix 3. The estimation of allele frequency in a DNA pool

We calculated three types of allele frequencies for comparison, including the true, unadjusted and adjusted allele frequencies. The true allele frequency was yielded by calculating the proportion of the number of a specific allele and the total number of alleles based on individual genotyping result. The unadjusted allele frequencies were estimated by calculating the relative IRAS based on pooled allelotyping result as follows:

$$\hat{p}_1^{\text{Unadjusted}} = \text{IRAS}_1^{\text{Pool}}/(\text{IRAS}_1^{\text{Pool}} + \text{IRAS}_2^{\text{Pool}}) \text{ and } \hat{p}_2^{\text{Unadjusted}} = 1 - \hat{p}_1^{\text{Unadjusted}},$$

where $\{\text{IRAS}_c^{\text{Pool}}, c \in C\}$denoted the IRAS value of allele c in a DNA pool. The adjusted allele frequencies were estimated by incorporating the estimated CPA in the estimation of the unadjusted allele frequency as follows:

$$\hat{p}_1^{\text{Adjusted}} = \text{IRAS}_1^{\text{Pool}}/(\text{IRAS}_1^{\text{Pool}} + \hat{\kappa} \times \text{IRAS}_2^{\text{Pool}}) \text{ and } \hat{p}_2^{\text{Adjusted}} = 1 - \hat{p}_1^{\text{Adjusted}}.$$