Nucleic Acids Research, 2005, Vol. 33, No. 17 ***e142***
doi:10.1093/nar/gni142

PPC: an algorithm for accurate estimation of SNP allele frequencies in small equimolar pools of DNA using data from high density microarrays

Jesper Brohede[1,2], Rob Dunne[1,3], James D. McKay[4,5] and Garry N. Hannan[1,2,*]

[1]CSIRO Preventative Health National Research Flagship, Sydney, Australia, [2]CSIRO Molecular and Health Technologies, Sydney, Australia, [3]CSIRO Mathematical and Information Sciences, Sydney, Australia, [4]Menzies Research Institute, University of Tasmania, Hobart, Australia and [5]International Agency for Research on Cancer, Lyon, France

Received August 11, 2005; Revised and Accepted September 1, 2005

ABSTRACT

Robust estimation of allele frequencies in pools of DNA has the potential to reduce genotyping costs and/or increase the number of individuals contributing to a study where hundreds of thousands of genetic markers need to be genotyped in very large populations sample sets, such as genome wide association studies. In order to make accurate allele frequency estimations from pooled samples a correction for unequal allele representation must be applied. We have developed the polynomial based probe specific correction (PPC) which is a novel correction algorithm for accurate estimation of allele frequencies in data from high-density microarrays. This algorithm was validated through comparison of allele frequencies from a set of 10 individually genotyped DNA's and frequencies estimated from pools of these 10 DNAs using GeneChip 10K Mapping Xba 131 arrays. Our results demonstrate that when using the PPC to correct for allelic biases the accuracy of the allele frequency estimates increases dramatically.

INTRODUCTION

Mapping the genetic basis underlying common multifactorial diseases such as cancer through whole genome association studies has attracted much attention in recent years. The discovery and characterization of millions of single nucleotide polymorphism (SNP) markers throughout the human genome (1) and the development of genotyping technology means that genome wide association studies are becoming technically possible. Estimates hold that whole genome association studies using SNPs require genotyping of hundreds of thousands to a million markers in large groups of cases and controls (2,3). The number of cases and controls needed varies depending on disease of interest and the level of linkage disequilibrium in the study population but to achieve adequate power, sample sizes will be counted in the hundreds or thousands even in the most favorable scenarios. While genotyping cost is decreasing, the high costs associated with genotyping large numbers of individuals for this large number of markers may prove to be rate limiting for this type of study (4,5). As a means of reducing the effort and costs involved, estimating allele frequencies in pools of equimolar amounts of DNA have been explored as an alternative to individual genotyping (6,7). This strategy has successfully been used in a number of candidate gene case-control studies using both SNPs (8,9) and microsatellites (10). Many different strategies for pooling have been suggested (11) but most researchers view pooling in combination with whole genome analysis as a first screening tool to identify markers with potential interest that can be chosen for subsequent individual genotyping (4,12,13). The simplest strategy is the two-pool design where all cases are collected in the first pool and all controls are collected in a second pool but other more elaborate pooling schemes have also been suggested. For instance, creating sets of sub-pools allows stratification, not only on the basis of the disease trait but also on secondary and tertiary traits as well. This might for instance capture effects of environmental factors that are known to affect the disease in question (14).

High density microarrays capable of parallel genotyping of tens to hundreds of thousands of SNPs currently provide the strongest candidate technology for large scale genome wide genotyping (15–17). Recently, four loci associated with mild mental impairment were identified in the first pool based genome wide screen using GeneChip 10K Mapping Xba 131 arrays (13). Although these microarrays are primarily designed for individual genotyping, we and others (9, 18–21) have successfully explored the possibility of using

*To whom correspondence should be addressed. Tel. +61 2 9490 5054; Fax +61 2 9490 5010; Email: Garry.Hannan@CSIRO.au

the quantitative nature of the signal hybridization intensities from such microarrays for making allele frequency estimations in pooled DNA. To improve the accuracy of allele frequency estimates from pooled DNA, corrections have to be made to account for biases in allelic representation (22,23). This allelic representation bias is mainly caused by allele specific preferential amplification of the genomic DNA and/or differences in hybridization properties for the different probe sequences. The most common correction method is *k*-correction which uses a correction factor *k* that is empirically derived from the signal intensity pattern of heterozygote individuals (24,25). *k*-correction was recently adapted for high-density microarray data resulting in a substantial improvement in accuracy of the allele frequency estimates (20). Here, we describe a novel correction algorithm which we call the polynomial based probe specific correction (PPC) and we show that it further increases the accuracy of the allele frequency estimates compared with previously described algorithms. PPC is based on a probe pair specific hybridization profile that was empirically derived from studying unique probe responses in a reference set of 26 GeneChip 10K Mapping Xba 131 arrays. The algorithm was subsequently utilized to estimate the allele frequencies in a pool of 10 individuals that had been genotyped previously using GeneChip 10K Mapping Xba 131 arrays.

MATERIALS AND METHODS

DNA samples

The DNA samples are fully described in L. M. FitzGerald, J. Stankovich, G. Price, J. Brohede, S. Quinn, R. Thomson, D. Challis, M. Challis, C. R. Wilkinson, J. Slavin, A. Banks, K. Hazelwood, D. Mackey, G. N. Hannan, T. Dwyer, J. L. Dickinson, D. Venter and J. D. McKay (manuscript submitted). Written informed consent was obtained from all participating individuals and ethics approval was obtained from the Southern Tasmanian Human Research Ethics Committee.

Genotyping

Genotyping of individual and DNA pools were made using the GeneChip 10K Mapping Xba 131 assay according to the GeneChip Mapping Assay Manual (Affymetrix) and all reagents were supplied by the manufacturer if not stated otherwise. Briefly, 250 ng of DNA was digested by Xba I (New England Biolabs) and Xba adaptors were subsequently ligated to the ends of all fragments using T4 DNA ligase (New England Biolabs). This was used as template in a PCR amplification using AmpliTaqGold (Applied Biosystems) and a single primer complementary to the adaptor sequence. PCR products were purified from excess primer and salts by QIAquick spin-columns (QIAGEN) and a 20 µg aliquot was fragmented using DNase I. An aliquot of the fragmented DNA was separated and visualized in a 2% agarose gel in 1× TBE buffer to ensure that the bulk of the product had been properly fragmented to a size <200 bp. The fragmented samples were end-labeled with biotin using terminal deoxynucleotidyl transferase before each sample was allowed to hybridize to a GeneChip 10K Mapping Xba 131 array for 16 h.

Following hybridization the arrays were washed and stained using an Affymetrix Fluidics Station 450. Most stringent wash was 0.6× SSPE, 0.01% Tween-20 at 45°C and the samples were stained with R-phycoerythrin (Molecular Probes). Imaging of the microarrays was performed using either a GeneArray (Agilent) or a GCS3000 (Affymetrix) high-resolution scanner. Genotype calls and probe intensity data were extracted with the GeneChip DNA Analysis Software (GDAS) (Affymetrix) using default parameters.

DNA pooling

Pools were constructed from equal amounts of DNA from 10 individuals that had been genotyped previously by the GeneChip 10K Mapping Xba 131 assay. To ensure that equimolar amounts of DNA were pooled, accurate quantifications were made using PicoGreen assay (Molecular Probes) against a standard curve of λDNA.

To assess the variability in the pool construction, DNA quantification, dilution, pooling and GeneChip assay steps were performed independently three times creating pooled samples or 'true replicas': p10_rep1, p10_rep2 and p10_rep3. To capture the variation introduced by the GeneChip Mapping assay alone, technical replicas were made by independently amplifying and hybridizing the pooled DNA's of p10_rep1 two additional times (creating pooled samples or 'technical replicas': p10_rep1_tech_2 and p10_rep1_tech_3). For the purpose of evaluating the results a measure of accuracy was defined as the absolute difference between the allele frequency deduced from the individual genotyping and the corresponding estimate made using pooled DNA, averaged over all available SNPs.

Rationale for algorithm

The basis for the PPC algorithm is to correct the raw signal intensity data in a probe pair specific manner. The algorithm only utilizes the signal intensity values from the 20 probes that constitute perfect matches for every SNP that can be interrogated on a GeneChip 10K Mapping Xba 131 array. These 20 probes are divided into 10 probe pairs where each pair consists of one probe that perfectly matches the A allele (PMA) and one probe that perfectly matches the B allele (PMB). The difference between the 10 pairs is which strand (sense versus antisense) is interrogated and the position of the polymorphism relative to the centre of the 25mer that constitutes a probe (15). If considering the signal intensity from one probe pair *j* for a given SNP let:

$$x_j = \frac{A_j}{(A_j + B_j)}, \qquad \mathbf{1}$$

where A_j and B_j are the observed signal intensity values for PMA and PMB, respectively. However, the hybridization affinities of any probe pair will be unique owing to sequence specific hybridization properties. Variation in the amplification efficiency between the two different alleles and background hybridization are two additional factors that must be taken into account to make an accurate allele frequency estimate. Together this suggests that all probe pairs have distinctly different hybridization profiles which is our rationale for making a unique correction for each probe pair. In mathematical terms the hybridization profile for any given probe pair could be best described by a second-degree polynomial. In order to obtain the unique hybridization profile for the majority of

probe pairs on the GeneChip 10K Mapping Xba 131 array we capitalized on a set of 26 reference microarrays that had been used for individual genotyping. For a second-degree polynomial to be derived for a particular probe pair the minimum requirement was that at least one individual homozygous for the A allele (AA), one individual homozygous for the B allele (BB) and one heterozygote (AB) individual were present among the 26 reference microarrays. The true allele frequency (defined to be 1.0 for genotype AA, 0.5 for genotype AB and 0.0 for genotype BB) was plotted against the corresponding allele frequencies estimates made using Equation 1. A script designed in R derived the coefficients for the second-degree polynomial that describes the relationship between the true allele frequency and the estimated allele frequency. In this way the second-degree polynomial coefficients for 80 660 probe pairs (10 for each of 8066 SNPs) were successfully obtained.

Allele frequency estimates in pools of DNA

To estimate the allele frequencies from the microarrays hybridized with pooled DNA, signal intensity values from all probes were extracted by the export function of GDAS 2.0 (Affymetrix). Using only the perfect match probes, the frequency of allele A was estimated for each probe pair individually by Equation 1 followed by correction by its unique second-degree polynomial:

$$f(\mathrm{A}j) = \beta_0 + \beta_1 \times x_j + \beta_2 \times x_j^2. \quad \mathbf{2}$$

Then a median value of the 10 estimates corresponding to one particular SNP was used to represent the allele frequency for that SNP.

Table 1. Accuracy measures for allele frequency estimates using PPC

Sample	Average accuracy[a]	Largest under-estimation	Largest over-estimation	Proportion differing >10%[b]
p10_rep1	0.048 (0.061)	−0.292	0.263	10.3
p10_rep1_tech 2	0.053 (0.069)	−0.288	0.284	14.2
p10_rep1_tech 3	0.067 (0.084)	−0.364	0.400	22.4
p10_rep2	0.043 (0.056)	−0.244	0.247	7.6
p10_rep3	0.062 (0.079)	−0.629	0.445	19.3
$p10_{\text{average true replicas}}$	0.029 (0.038)	−0.194	0.206	1.8
$p10_{\text{average technical replicas}}$	0.050 (0.065)	−0.294	0.294	12.1

[a]Accuracy is measured as specified in the text. Standard deviation in brackets.
[b]Proportion of markers where the estimated allele frequency differs by more than ±10% of the deduced value.

RESULTS

Validation of PPC

Equimolar amounts of DNA from 10 individuals were pooled before being assayed on a GeneChip 10K Mapping Xba 131 array according to manufacturer's instruction. Signal intensity data was extracted and allele frequencies were estimated according to the algorithm described above. All measurements of DNA concentration, all pooling procedures and all GeneChip assays were replicated, independently, three times in order to study the variation introduced by the pooling procedure. To study the assay variation, technical replicas were made from one of the pools by repeating the GeneChip assays and hybridizations three times for one of the pools of DNA. In order to evaluate the accuracy of the allele frequency estimations from these pooling experiments all 10 DNA samples were individually genotyped. This provided a more accurate measure of the alleles that made up the pools than using population frequencies. Using only SNPs for which genotype information was available for all 10 individuals the expected allele frequencies in the pools could be deduced for 8180 SNPs. Accuracy was defined as the absolute difference between the allele frequency estimate and the allele frequency deduced from individual genotyping. An average accuracy was calculated for each replica and the results was summarized in Table 1. Computer scripts for deriving the second-degree polynomials and the subsequent allele frequency estimation were written in Perl (http://www.perl.org) or R (http://www.r-project.org<http://www.perl.org>) and are available at http://www.bioinformatics.csiro.au/publications.shtml.

True replicas versus technical replicas

For the pooled sample GeneChip replicas, the average accuracy was 0.054 and ranged from 0.043 to 0.067 (Table 1) equivalent to or better than that other reported by other published pooling studies (9,19,20). We also observed an average correlation (r^2) of 0.904 for the true replicas and 0.959 for the technical replicas (Table 2). The range of accuracy for the three true replicas (0.048, 0.043 and 0.062) was no different than that for the technical replicas (0.048, 0.053 and 0.067). However when creating a replica average, by averaging the individual SNP by SNP frequency estimates from each of the 'true' and 'technical' replicas before comparing it with the corresponding deduced allele frequency, there were different responses from the true replicas compared with the technical replicas. When averaging in this way the average accuracy

Table 2. r^2-Values from allele frequency estimates in all pools of 10 individuals

	True allele frequency[a]	p10_rep1	p10_rep1_tech2	p10_rep1_tech3	p10_rep2	p10_rep3	$p10_{\text{average true replicas}}$	$p10_{\text{average technical replicas}}$
True allele frequency[a]	1	0.966	0.960	0.923	0.948	0.929	0.978	0.962
p10_rep1		1	0.971	0.944	0.914	0.934	0.983	0.985
p10_rep1_tech2			1	0.963	0.892	0.952	0.973	0.992
p10_rep1_tech3				1	0.845	0.926	0.940	0.983
p10_rep2					1	0.865	0.949	0.895
p10_rep3						1	0.969	0.950
$p10_{\text{average true replicas}}$							1	0.978
$p10_{\text{average technical replicas}}$								1

Note that $p10_{\text{average true replicas}}$ and $p10_{\text{average technical replicas}}$ are averages of several samples rather than independent microarrays.
[a]Allele frequencies deduced from individual genotyping.

Table 3. Average accuracy in different allele frequency intervals for the average replica sample $p10_{\text{average true replicas}}$

Allele frequency interval	Average accuracy[a]	n[b]
0.0–0.1	0.019 (0.014)	125
0.1–0.2	0.029 (0.024)	345
0.2–0.3	0.032 (0.028)	631
0.3–0.4	0.029 (0.027)	744
0.4–0.5	0.027 (0.023)	793
0.5–0.6	0.028 (0.023)	805
0.6–0.7	0.028 (0.024)	779
0.7–0.8	0.030 (0.027)	747
0.8–0.9	0.032 (0.026)	501
0.9–1.0	0.026 (0.019)	235

[a]Accuracy is measured as specified in the text. Standard deviation in brackets.
[b]Number of SNPs in the interval.

Table 4. Comparisons of the accuracy using different algorithms

Reference	Average accuracy[a]	No. of SNPs	Largest under-estimation	Largest over-estimation	Proportion differing >10%[b]
This article	0.029 (0.038)	5705	−0.194	0.206	1.8
(20)	0.053 (0.060)	7059	−0.395	0.213	12.6
(19)	0.070 (0.091)	8179	−0.514	0.441	25.5
(9)	0.073 (0.092)	7633	−0.510	0.469	26.5

[a]Accuracy is measured as specified in the text. Standard deviation in brackets.
[b]Proportion of markers where the estimated allele frequency differs by more than ±10% of the deduced value.

increased to 0.029 for the true replicas and to 0.050 for the technical replicas. This increase in accuracy for the true replicas was also reflected in the other parameters shown in Table 1. For instance, the percentages of estimates that differed by >0.1 for the technical replicas 1, 2 and 3 were 10.3, 14.2 and 22.4%, respectively, while the average figure for the technical replicas was 12.1%. Corresponding percentages for the true replicas were within the same range (10.3, 7.6 and 19.3% for replicas 1, 2 and 3, respectively) but showed an exceptional improvement for the replica average, that is 1.8%. This effect was probably largely owing to the canceling out of variation in the pooling procedure (for instance pipetting inaccuracies) for the true replicas while the corresponding variation in the technical replicas was only introduced once and therefore no canceling out will occur. Table 3 shows the accuracy in relation to the deduced allele frequency for the average of the true replicas.

Comparison with previously described algorithms

To further evaluate the algorithm described here scripts in Perl and R were designed to estimate allele frequencies from the three true p10 replicas described above using previously described algorithms for identical (19,20) or very similar types of microarrays (9). All comparisons shown in Table 4 and Figure 1 are based on averaging the individual estimates of the three true p10 replicas as described above.

DISCUSSION

Estimating allele frequencies from pooled data, or allelotyping (13), has been suggested as an alternative to individual genotyping in large scale projects as a way to bring down the costs. Although pooling can be used for directly identifying disease associated markers most scientists view pooling as a first screen in order to identify markers for a subsequent targeted genotyping on an individual level. To demonstrate the utility of PPC on the estimation of allele frequencies in DNA pools, our study used pools of 10 individuals with three true replicas, resulting in a one-third study wide saving compared with individual genotyping and only a modest budget benefit. Optimal pool size for allelotyping will ultimately depend on a cost/benefit analysis specific to the individual study design, allelotyping methodology and budget restrains. Pfeiffer *et al*. (26) explores pool sizes of 3–10 while Barrat *et al*. (27) advocates an optimal pool size of 50 individuals and Le Hellard *et al*. (28) argues for pool sizes of several hundred individuals. While there appears to be clear benefits in the pool sizes described here, the effect a larger pool size has on the average accuracy when allelotyping with the PPC remains to be tested. On the basis of our results, it appeared that multiple true replicas made a significant difference to the allele frequency estimation accuracy, therefore, when considering the cost benefits of a pooling approach we suggest retaining adequate replicas for each pool regardless of size.

Differences in SNP hybridization characteristics make it difficult to make accurate allele frequency estimations in high-density microarray data. We have tackled this problem by empirically deriving a specific correction formula for each probe pair. The underlying mathematic relationship for the probe response has been accurately described with a second-degree polynomial. During the course of this project logistical regression and robust regression using Huber's M estimator (29) were also assessed as allelotyping algorithms and although they also performed fairly well in preliminary studies they were always outperformed by the second-degree polynomial which is why they were not explored further (data not shown).

The PPC is an ideal correction formula for standardized genotyping platforms like the Affymetrix GeneChip Mapping array system where large sets of reference microarrays can easily be gathered for the purpose of deriving the probe specific hybridization profiles. In addition to standardized microarrays, the Affymetrix GeneChip platform employs standardized hybridization, washing and scanning equipment which further works to keep the hybridization profile constant between experiments. Our results show that this technology platform has a very high level of consistency as reflected in the high r^2 values for the technical replicas (Table 2). Moreover, our results from averaging of replicas highlight the importance of making true replicas rather than technical replicas where only the amplification to hybridization steps are replicated. In technical replicas all biases would be re-amplified and would show up on all replica microarrays while random variation introduced in the pooling procedure would work in different directions and therefore increase the accuracy measure for true replicas. Under the conditions described here the accuracy increased from an average of 0.050 for the technical replicas to 0.029 for the true replicas. While this level of accuracy is comparable with most other technologies that have been used in conjunction with pooled DNA (Table 5), the mass-parallelism of high-density microarrays distinguish it as an ideal tool for genome wide genetic scans.

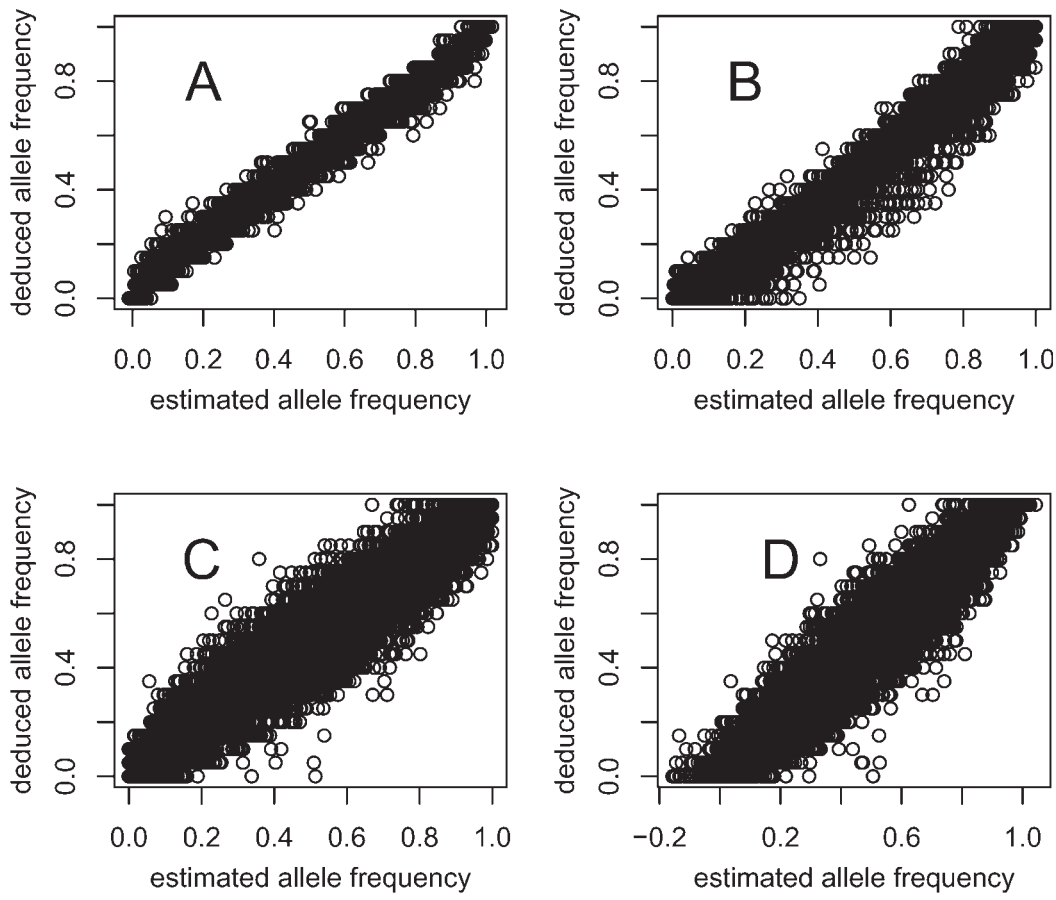

Figure 1. The figure shows the relationship between the allele frequency deduced from individual genotyping and the allele frequency estimated with (**A**) the PPC described here, (**B**) the algorithm described in (20), (**C**) the algorithm described in (19) and (**D**) the algorithm described in (9). All estimates were based on the average of three replicas as specified in the text.

When comparing the results from the previously published algorithms it's important to note the algorithm described in (9) was not developed for the GeneChip 10K Mapping Xba 131 arrays as were the other algorithms. The arrays they used in that study were similar in the way they were manufactured and the probes were also 25mers but one main difference is that there were 80 probes per SNP rather than the 40 present on the GeneChip. The 40 extra probes were an additional set of mismatch probes that would make the estimates of background signal more accurate for the chips used in (9) compared with the GeneChip 10K Mapping Xba 131 arrays used in this study. Since the performance of their algorithm was highly dependent on the subtraction of background it is probable that the estimates presented here (using the less comprehensive background estimate from our GeneChip 10K Mapping Xba 131 array data) would have been more accurate if we had used microarrays similar in design to those used in (9). Furthermore, they used their algorithm to estimate differences in allele frequency between cases and controls rather than for accurate allelotyping. While there are differences between the microarrays used in this article and the ones used in (9) we believe that they are similar enough to make an interesting comparison with the other algorithms in Table 4.

The algorithm described in (19) was based on the relative allele signal (RAS) value calculated by genotype scoring using the Modified Partitioning Around Medoids (MPAM) algorithm implemented the GDAS software. These authors stress the difference between cases and controls as the main outcome rather than accurate estimates for each DNA pool. Simpson *et al*. (20) developed this algorithm further and implemented *k*-correction. The *k*-correction uses the pattern of heterozygotes to correct the average RAS values for allelic biases. Their paper also showed the need to correct for allelic biases particularly for rare alleles. Meaburn *et al*. (21) reported that the accuracy increased from 0.077 to 0.036 when a *k*-correction was applied to GeneChip 10K Mapping Xba 131 data in a subset of 104 SNPs in a pool consisting of DNAs from 100 individuals. This was in fairly good agreement with the corresponding figures in this article (0.070 increased to 0.053 with *k*-correction). When applying the probe specific correction equations based on second-degree polynomials described in this paper the accuracy further increased to 0.029. Moreover, allelotyping using PPC occasionally resulted in estimates outside the 0–1 range which is in contrast to the algorithms that were based on the RAS values. When examining the results from $p10_{\text{average true replicas}}$ the extent of this was very low with only 16 markers (0.28%) having estimates outside the 0–1 range further supporting the high accuracy of PPC.

While the algorithms described by Butcher *et al*. (19) and Hinds *et al*. (9) have an advantage in that no prior knowledge about the probe response is required, the lower levels of accuracy might be limiting their usefulness. In contrast, both the

Table 5. Previously described accuracy measures of SNP estimates in pooled DNA using non-array based technologies

Reference	Technology	Accuracy	No. of SNPs studied	Pool size	No. of pools studied[a]
(30)	Real-time PCR	0.02	5	10	1
		0.02	3	100	1
(31)	Real-time PCR	0.003	1	56	1
		0.005	1	86	1
		0.017	1	127	1
(32)	Pyrosequencing	0.039	3	10	50
		0.026	3	20	25
		0.049	3	50	10
		0.047	3	100	5
		0.029	3	200	2
		0.067	3	479	1
(27)	Pyrosequencing	0.011	9	188	1
		0.023	9	358	1
		0.022	9	381	1
		0.021	9	739	1
(33)	Pyrosequencing	0.011	7	150	2
(28)	SnaPshot	0.015	5	96	1
(19)	SnaPshot	0.022	10	105	1
(34)	SnaPshot	0.023	15	111–220	NA
		0.017	7	130–222	NA
(28)	dHPLC	0.017	5	96	1
(35)	dHPLC	0.013	2	49–402	20
(24)	dHPLC	0.015	9	111–220	NA
(28)	MALDI-TOF	0.033	5	96	1
(36)	MALDI-TOF	0.026	8	240	1
		0.027	8	120	2
		0.027	8	60	4
(37)	PLACE-SSCP	0.017	1	78	1

The table shows an overview of results from a number of previously published papers in the DNA pooling field. The purpose was to contrast results from microarray technology with other genotyping technologies that have been used with pooled DNA. The accuracy is in some cases an average over several SNPs in multiple replicas as specified in the original references and all data has been corrected for biased allelic representation.

[a]Technical replicas of the same pool have not been included.

PPC and the *k*-correction produce highly accurate estimates but depend on prior knowledge of the behavior of the probe specific hybridization profile which limits the number of markers available for allelotyping. In particular the PPC algorithm was affected by this since a reliable second-degree polynomial could only be derived after assessing at least one individual homozygous for the reference allele, one heterozygote and one homozygote for the alternative allele. This could easily be overcome by increasing the size of databases from where the probe specific hybridization profiles was derived. This has for instance been addressed by others (20) who have designed a public database where anyone can deposit GeneChip data to constantly improve and access the *k*-correction coefficients.

A GeneChip 10K Mapping Xba 131 arrays suffices for performing linkage analysis in a family but to perform association studies in a complex population, we will need 100 000, or maybe even 500 000 to a million SNPs (3). Even given that the cost of genotyping is coming down to <1 cent per SNP, the overall cost for an association study involving hundreds of cases and controls will be very expensive. It will be a great advantage to have new algorithms developed to reduce cost while not affecting the power to locate disease genes in study populations, particularly with the recent release of the Affymetrix 500K SNPs arrays.

ACKNOWLEDGEMENTS

The authors thank Diana Brookes and Glenn Brown for technical assistance. We also wish to thank the Tasmanian Cancer council and Royal Hobart Hospital for their financial assistance to enable sample collection. Funding to pay the Open Access publication charges for this article was provided by CSIRO.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,F.S., Guyer,M.S. and Charkravarti,A. (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.
2. Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
3. Kruglyak,L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet*., **22**, 139–144.
4. Risch,N. and Teng,J. (1998) The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res*., **8**, 1273–1288.
5. Pharoah,P.D., Dunning,A.M., Ponder,B.A. and Easton,D.F. (2004) Association studies for finding cancer-susceptibility genetic variants. *Nature Rev. Cancer*, **4**, 850–860.
6. 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 the HLA class II loci. *Proc. Natl Acad. Sci. USA*, **82**, 6970–6974.
7. Michelmore,R.W., Paran,I. and Kesseli,R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl Acad. Sci. USA*, **88**, 9828–9832.
8. Butcher,L.M., Meaburn,E., Dale,P.S., Sham,P., Schalkwyk,L.C., Craig,I.W. and Plomin,R. (2005) Association analysis of mild mental impairment using DNA pooling to screen 432 brain-expressed single-nucleotide polymorphisms. *Mol. Psychiatry*, **10**, 384–392.
9. Hinds,D.A., Seymour,A.B., Durham,L.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.
10. Kirov,G., Williams,N., Sham,P., Craddock,N. and Owen,M.J. (2000) Pooled genotyping of microsatellite markers in parent-offspring trios. *Genome Res*., **10**, 105–115.
11. 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.
12. König,I.R. and Ziegler,A. (2004) Analysis of SNPs in pooled DNA: a decision theoretic model. *Genet. Epidemiol*., **26**, 31–43.
13. Butcher,L.M., Meaburn,E., Knight,J., Sham,P.C., Schalkwyk,L.C., Craig,I.W. and Plomin,R. (2005) SNPs, microarrays and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6000 children. *Hum. Mol. Genet*., **14**, 1315–1325.
14. Law,G.R., Rollinson,S., Feltbower,R., Allan,J.M., Morgan,G.J. and Roman,E. (2004) Application of DNA pooling to large studies of disease. *Stat. Med*., **23**, 3841–3850.
15. 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.
16. 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.
17. 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*, **21**, 1958–1963.

18. Uhl,G.R., Liu,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.
19. 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.
20. 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.
21. 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**, e52.
22. 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.
23. Perlin,M.W., Lancia,G. and Ng,S.K. (1995) Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. *Am. J. Hum. Genet.*, **57**, 1199–1210.
24. 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.
25. 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.
26. Pfeiffer,R.M., Rutter,J.L., Gail,M.H., Struewing,J. and Gastwirth,J.L. (2002) Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. *Genet. Epidemiol.*, **22**, 94–102.
27. Barratt,B.J., Payne,F., Rance,H.E., Nutland,S., Todd,J.A. and Clayton,D.G. (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann. Hum. Genet.*, **66**, 393–405.
28. 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.
29. Huber,P.J. (1981) *Robust Statistics*. Wiley, NY.
30. Germer,S., Holland,M.J. and Higuchi,R. (2000) High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res.*, **10**, 258–266.
31. Chen,J., Germer,S., Higuchi,R., Berkowitz,G., Godbold,J. and Wetmur,J.G. (2002) Kinetic polymerase chain reaction on pooled DNA: a high-throughput, high-efficiency alternative in genetic epidemiological studies. *Cancer Epidemiol. Biomarkers Prev.*, **11**, 131–136.
32. Lavebratt,C., Sengul,S., Jansson,M. and Schalling,M. (2004) Pyrosequencing-based SNP allele frequency estimation in DNA pools. *Hum. Mutat.*, **23**, 92–97.
33. Wasson,J., Skolnick,G., Love-Gregory,L. and Permutt,M.A. (2002) Assessing allele frequencies of single nucleotide polymorphisms in DNA pools by pyrosequencing technology. *Biotechniques*, **32**, 1144–1150.
34. Norton,N., Williams,N.M., Williams,H.J., Spurlock,G., Kirov,G., Morris,D.W., Hoogendoorn,B., Owen,M.J. and O'Donovan,M.C. (2002) Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum. Genet.*, **110**, 471–478.
35. Giordano,M., , Mellai,M., Hoogendoorn,B. and Momigliano-Richiardi,P. (2001) Determination of SNP allele frequencies in pooled DNAs by primer extension genotyping and denaturing high-performance liquid chromatography. *J. Biochem. Biophys. Methods*, **47**, 101–110.
36. Downes,K., Barratt,B.J., Akan,P., Bumpstead,S.J., Taylor,S.D., Clayton,D.G. and Deloukas,P. (2004) SNP allele frequency estimation in DNA pools and variance components analysis. *Biotechniques*, **36**, 840–845.
37. Sasaki,T., Tahira,T., Suzuki,A., Higasa,K., Kukita,Y., Baba,S. and Hayashi,K. (2001) Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. *Am. J. Hum. Genet.*, **68**, 214–218.