Published online February 14, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 4 ***e28***
doi:10.1093/nar/gnj027

Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans

Emma Meaburn*, Lee M. Butcher, Leonard C. Schalkwyk and Robert Plomin

Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, De Crespigny Park, London, SE5 8AF, UK

Received October 26, 2005; Revised December 6, 2005; Accepted January 27, 2006

ABSTRACT

The identification of quantitative trait loci (QTLs) of small effect size that underlie complex traits poses a particular challenge for geneticists due to the large sample sizes and large numbers of genetic markers required for genomewide association scans. An efficient solution for screening purposes is to combine single nucleotide polymorphism (SNP) microarrays and DNA pooling (SNP-MaP), an approach that has been shown to be valid, reliable and accurate in deriving relative allele frequency estimates from pooled DNA for groups such as cases and controls for 10K SNP microarrays. However, in order to conduct a genomewide association study many more SNP markers are needed. To this end, we assessed the validity and reliability of the SNP-MaP method using Affymetrix GeneChip® Mapping 100K Array set. Interpretable results emerged for 95% of the SNPs (nearly 110 000 SNPs). We found that SNP-MaP allele frequency estimates correlated 0.939 with allele frequencies for 97 605 SNPs that were genotyped individually in an independent population; the correlation was 0.971 for 26 SNPs that were genotyped individually for the 1028 individuals used to construct the DNA pools. We conclude that extending the SNP-MaP method to the Affymetrix GeneChip® Mapping 100K Array set provides a useful screen of >100 000 SNP markers for QTL association scans.

INTRODUCTION

With the advent of highly multiplexed microarray systems for single nucleotide polymorphism (SNP) genotyping that can genotype hundreds of thousands of SNPs (1) genomewide association scans are underway, although problems remain (2,3). A major problem for the investigation of common disorders and complex traits is that large samples are needed to detect reliably the many quantitative trait loci (QTLs) of very small effect size likely to be responsible for the ubiquitous heritability of these traits (4). Because microarrays are expensive and can only be used once, employing SNP microarrays to genotype large samples is not economically viable for most researchers.

One way to screen large samples is to pool DNA for groups such as cases and controls (5). We have combined the strengths of microarrays to genotype large numbers of SNPs and DNA pooling to genotype large samples by genotyping pooled DNA on microarrays, a method we call SNP microarrays and pools (SNP-MaP). SNP-MaP allele frequency estimates for groups such as cases and controls (or low and high individuals for quantitative traits) can be compared to nominate SNPs that show the greatest allele frequency differences; these nominated SNPs can then be confirmed with individual genotyping and traditional parametric statistics. We have shown that pooled DNA can be genotyped reliably on microarrays with 10 000 SNPs (6,7) and we have used the SNP-MaP method and 10K SNP microarrays in a multistage design to identify four SNPs associated with cognitive disability and ability in a sample of 6000 children (8). Until now, our research has applied the SNP-MaP method to the Affymetrix GeneChip® Mapping 10K Array which uses a single endonuclease restriction digestion and a single primer set to amplify >10 000 SNPs distributed throughout the genome. Advances in microarray technology, coupled with increased resolution of the human genome sequence and its SNPs, have led to the recent development of the Affymetrix GeneChip® Mapping 100K Array-set (9). The 100K microarray set genotypes 116 204 SNPs, with a median and mean intermarker distance of 8.5 and 23.6 kb, respectively, with 92% of the genome within 100 kb of a SNP. Even though genomewide association studies are likely to require >100 000 SNP genetic markers (10), the 100K microarray set represents an important step towards conducting systematic genomewide associations.

*To whom correspondence should be addressed. Box Number P082, Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, De Crespigny Park, London SE5 8AF, UK. Tel: +44 20 7848 0748; Fax: +44 20 7848 0895; Email: e.meaburn@iop.kcl.ac.uk

As described in the following section, there are important differences between the 10K and 100K microarrays which warrant the present investigation which assesses whether the SNP-MaP method can be extended from the 10K microarray to the 100K microarray set.

The 100K microarray-set assay

Although the assay procedure and microarray technology for the 100K microarray set is similar to that of the 10K microarray, there are four important differences. First, the 10K microarray used just one restriction endonuclease (XbaI) and one microarray, whereas the 100K microarray set uses a second restriction endonuclease (HindIII) and a second microarray. The 100K XbaI microarray genotypes 58 960 SNPs and the HindIII microarray genotypes 57 244 SNPs. The use of a second restriction endonuclease and so many additional SNPs could affect SNP-MaP results.

Secondly, similar to the 10K microarray, enzyme digestion is followed by preferential amplification of fragments of a certain size. However, the 10K microarray uses *Taq* polymerase, which preferentially amplifies fragments in the range of 250–1000 bp, whereas the 100K microarray set uses Platinum *pfx* polymerase (Invitrogen Corporation), which preferentially amplifies larger fragments of digested DNA in the range of 250–2000 bp. This amplicon size represents ~300 Mb of sequence in comparison with ~60 Mb of sequence generated in the 10K microarray. The larger amplicon size of the 100K microarray set could reduce the number of copies generated during the extension phase of PCR and thus make 100K SNP-MaP allele frequency estimates less accurate than the 10K microarray, which could especially affect estimates based on pooled DNA.

Thirdly, the feature size of the 100K microarray set has been downsized from 18 to 8 μm^2 to enable a higher density of probes on each microarray. This might decrease the reliability and accuracy of deriving allele frequency estimates from pooled DNA, as fewer copies of oligonucleotide probes are available for DNA hybridization.

Finally, the SNPs on the 10K microarray were preferentially selected for assay performance from a starting set of 55 605 SNPs. For the 100K microarray set, SNPs were selected from a starting set of 3 031 331 SNPs, with 1 833 423 SNPs from dbSNP (11) and the SNP Consortium (12) in addition to 1 197 908 SNPs discovered by Perlegen Sciences (13). Of these, 535 564 SNPs were predicted to be in XbaI or HindIII genomic DNA sites, and following *in silico* and empirical screening, a final set of 116 204 SNPs was selected. The larger numbers of SNPs raises the possibility that some are less well characterized and may be less reliable.

Study design

These differences between the 10K and 100K microarray assays warrant an investigation of the reliability, validity and accuracy of the SNP-MaP method as applied to the Affymetrix GeneChip® mapping 100K microarray set. Using five independent DNA pools consisting of >200 individuals each, reliability was assessed by comparing the SNP-MaP allele frequency estimates for the five pools. In addition, we compared allele frequency estimates for the 7956 SNPs common to both the 10K microarray and the 100K microarray set. Validity was assessed by comparing SNP-MaP allele frequency estimates with allele frequencies obtained by individual genotyping in two ways. First, we compared our SNP-MaP estimates with individually genotyped results for SNPs on the 100K microarray set from a publicly available reference sample of 42 Caucasian individuals. Second, we individually genotyped DNA from subjects in the five DNA pools for 26 SNPs on the 100K microarray set.

We also used a 100K microarray set to genotype an individual reference genomic DNA sample provided by Affymetrix as a positive internal control to ensure all steps of the assay were performed correctly.

MATERIALS AND METHODS

Samples

Five independent pools of DNA were created from a sample of 1028 white Caucasian individuals (538 females and 490 males) randomly selected from a community-based sample of >14 000 children in the Twins Early Development Study [TEDS; see (14)], which we used in an SNP-MaP study of cognitive ability and disability with the 10K microarray (8).

DNA quantification and pool construction

DNA samples were extracted from buccal swabs (15), quantified using a spectrophotometer (260 nm) and diluted to a target concentration of 50 ng/µl. Each sample was subsequently quantified in triplicate using fluorimetry (employing PicoGreen® dsDNA quantification reagent; Cambridge Bioscience, UK) and samples that were accurately quantified (±0.5 ng/µl) were accepted for pooling. Each individual's DNA was randomly assigned to one of five DNA pools, thus providing five independent pools constructed from between 204 and 206 individuals. Each individual contributed 79.1 ng of DNA to a DNA pool. Each pool concentration ranged from 13.33 to 13.57 ng/µl.

SNP microarray allelotyping of pooled DNA

Because pooled DNA can be used only to estimate allelic frequency, not genotypic frequency, we refer to allelotyping rather than genotyping. Each of the five DNA pools was allelotyped using the GeneChip® Mapping 100K Array set in accordance with the standard protocol for individual DNA samples (see the GeneChip® Mapping 100K Assay Manual for full protocol). Each microarray was scanned using the GeneChip® scanner 3000 and GeneChip® Operating software (GCOS) v1.1.1 with patch 5. Cell intensity (.cel) files were generated and subsequently saved and transported as Cabinet (.CAB) files (using Data Transfer Tool v1.1) to a workstation that contained GCOS software v1.2. Using GCOS v1.2, new .cel files were generated and analyzed using GeneChip® DNA Analysis Software (GDAS) v3.0.

Each of the five DNA pools and a reference DNA individual provided by the manufacturer (sample number 100103) was assayed on a separate microarray set. Each sample was independently amplified before hybridization.

Generation of SNP-MaP allele frequency estimates

The 10K microarray used a Modified Partitioning Around Medoids (MPAM) mapping algorithm to analyze the cell intensity data of the labeled DNA that had hybridized to the oligonucleotide probes (16). The MPAM mapping algorithm used relative allele signals (RAS), a measure of the intensities of the perfect match (PM) probes for alleles A and B of an SNP, to make genotype calls. However, the SNP-MaP method bypasses genotype calls, using instead the average of RAS 1 (sense strand) and RAS 2 (anti-sense strand) values (RAS_{AV}) as a quantitative index of allele frequencies in pooled DNA (7).

For the 100K microarray set, a new model-based genotyping algorithm has been developed by Affymetrix, called the dynamic modeling (DM) mapping algorithm (17), which no longer generates RAS scores. Instead, we generated RAS scores manually using the RAS score algorithm (see Affymetrix® GeneChip® DNA Analysis Software users' guide for full information on the algorithm used to derive RAS values) with the raw probe intensity data exported as a .txt file. Because the calculation of RAS values is computationally intensive, we wrote a script in R that calculates RAS 1 and RAS 2 values for each SNP. These programs are freely available to download at http://sgdp.iop.kcl.ac.uk/oleo/affy.

Correction for differential florescence signals

In addition, we investigated the effect of correcting signal intensities for differential florescence—a process known as *k*-correction (18). In theory, a heterozygous individual should yield a 50:50 ratio of florescence intensities for each of the two alleles. In practice, however, this is often not the case; the presence of a single base pair change in a 25mer oligo will subtly alter the hybridization kinetics and produce unique florescence intensity. This can have important consequences when comparing allele-frequency estimates from DNA pools with those derived from individual genotyping.

k-correction requires heterozygous scores for all SNPs on the array set, and consequently would require the genotyping of hundreds of individuals, especially for rare SNPs. For this reason, we have established a central resource for the accumulation of RAS 1 and RAS 2 values from Affymetrix arrays for heterozygous individuals: http://cogent.iop.kcl.ac.uk/rcorrection.cogx (19). Using this resource, *k* can be estimated from independent heterozygotes, currently ranging from 1 to 89 individuals for the 100K microarray set, and from 1 to 40 individuals for the 10K microarray set.

The heterozygous scores for SNPs on the 100K microarray set were downloaded, and *k*-values were derived for each SNP using the following equation.

$$k = \frac{\text{Correction Value}}{1 - \text{Correction Value}}$$

where Correction Value is the average of the RAS 1 and RAS 2 values for the panel of individual heterozygotes. For individual samples, RAS 1 and RAS 2 values vary between 0 (BB homozygote) and 1 (AA homozygote). The RAS 1 and RAS 2 values for AB heterozygotes cluster (on average) around 0.5, and so *k* is 1.0 if there is no differential florescence.

k-corrected SNP-MaP allele frequency estimates for allele A ($\hat{A}$) of a SNP can be calculated as follows:

$$\hat{A} = \frac{A}{A + k(1 - A)}$$

where A is the RAS_{AV} score for allele A of a SNP.

Investigation of non-specific hybridization

The 10K MPAM mapping algorithm, in addition to calculating RAS 1 and RAS 2 scores, assessed the degree of non-specific hybridization of the sample DNA to the probes on the microarray by comparing the intensities for the mismatch probes (MM) with the intensities of the Perfect Match (CAM) probes to calculate a discrimination score, called DS_{snp}. This calculation is no longer performed with the DM mapping algorithm for the 100K microarray set, and so we calculated the DS_{snp} for each SNP, again using the freely available script in R (http://sgdp.iop.kcl.ac.uk/oleo/affy; see Affymetrix® GeneChip® DNA Analysis Software users' guide for full information on the algorithm used to derive DS_{snp}).

DS_{snp} values should range from 0 to 1 with higher scores indicating greater discrimination between PM and MM probes and less non-specific hybridization. However, as the 25mer probes only differ by 1 bp between PM and MM probes, it is not uncommon for MM intensities to fluoresce at similar and sometimes even higher intensities. In cases where MM intensities are higher than PM intensities, the DS_{snp} value will be negative. Non-specific hybridization of sample DNA to the probes on the microarray produces background noise that might especially affect allele frequency estimates for pooled DNA. For this reason, DS_{snp} values were calculated for each SNP on the microarray set and compared between the DNA pools and individual sample.

For our analysis, although DS_{snp} values were obtained for each SNP on the microarray we focus only on SNPs with a DS_{snp} value $\geqslant 0.04$.

Individual genotyping

We compared SNP-MaP allelic frequency estimates from pooled DNA allelotyped on the 100K microarray set to estimates based on individual genotyping in two ways. First, we compared our SNP-MaP estimates to individually genotyped results for a reference sample of 42 Caucasian individuals that is publicly available from the NetAffx™ Analysis Centre (http://www.affymetrix.com/analysis/index.affx), a web-based tool providing extensive annotation for each SNP on the 100K microarray set (20). The 42 individuals were obtained from the human variation panel (http://coriell.umdnj.edu/) and genotyped by Affymetrix, and are of mixed gender, self declared Caucasian, unrelated and healthy.

Second, for 26 SNPs from the microarray set, we individually genotyped DNA from the 1028 subjects who were included in the five DNA pools. Individual genotyping was performed by Kbiosciences, which uses a mixture of competitive allele specific PCR (KASPar) and TaqMan genotyping assays (http://www.kbioscience.co.uk/). We assessed the genotyping error rate for each SNP using blind duplicate samples and members of MZ twin pairs.

RESULTS

Detection rates for pooled DNA

Good hybridization signal intensities were obtained for all five DNA pools. SNPs were excluded from analysis if there was inadequate discrimination between specific versus non-specific hybridization of DNA to the probes using a DS_{snp} value <0.04. As shown in Table 1, SNP-MaP allele frequency estimates were obtained for 109 994 SNPs (94.7%) across all five DNA pools for the 100K microarray-set: 54 778 SNPs (92.9%) on the XbaI microarray and 55 216 SNPs (96.5%) on the HindIII microarray. Table 1 also lists the number of SNPs estimates successfully obtained on fewer than five DNA pools.

Detection rate and reproducibility for a reference individual

DNA from a reference individual provided by Affymetrix was assayed at the same time as the DNA pools in order to ensure that the assay was performed correctly. GDAS was used to derive genotype calls for the individual sample, with DS_{snp} value again set at <0.04. Good hybridization signal intensities were obtained, and genotype calls were obtained for 110 738 SNPs (95.3%) of which 55 186 were on the XbaI microarray and 55 552 SNPs were on the HindIII microarray. Comparing our genotypes for the reference individual to the published genotypes, the agreement was 99.4% for XbaI and 99.8% for HindIII. This confirms that the assay was performed correctly, and that no sample contamination occurred.

Reliability of SNP-MaP allele frequency estimates

Reliability: comparing SNP-MaP estimates across five independent DNA pools for the 100K microarray set. In order to assess reliability, the SNP-MaP allele frequency estimates for each of the five DNA pools were compared. It should be reiterated that each of the five DNA pools is constructed from different individuals and assayed on a separate microarray set. Thus, differences between SNP-MaP allele frequency estimates incorporate sampling variance (i.e. true allele frequency differences between the pools), pool construction error, as well as all other sources of measurement error.

Our analysis focused on the subset of SNPs on the microarray set for which *k*-values were available and where 100K SNP-MaP allele frequency estimates were obtained across all five DNA pools ($N = 97\,605$ SNPs, of which 50 254 SNPs are on the XbaI microarray and 47 351 SNPs on the HindIII microarray).

Table 1. Number of SNPs successfully estimated from the 100K microarray set across five or fewer of the independent DNA pools

Number of DNA pools	XbaI Frequency	%	HindIII Frequency	%	Combined Frequency	%
5	54 778	92.9	55 216	96.5	109 994	94.5
4	2350	4.0	1481	2.6	3831	3.3
3	978	1.6	337	0.6	1315	1.1
2	515	1.4	130	0.2	645	0.6
1	241	0.4	60	0.1	301	0.3
0	98	0.2	20	0.0	118	0.1
	58 960	100	57 244		116 204	100

As shown in Table 2, the uncorrected SNP-MaP allele frequency estimates were highly correlated across the five DNA pools, ranging from 0.960 to 0.977 (average of 0.969) across the XbaI and HindIII microarrays. As indicated by the high correlations, the SNP-Map allele frequency differences across DNA pools are small, ranging from 0.046 to 0.058 (average of 0.054). As expected for relative comparisons between DNA pools (which is the goal for case–control studies), applying *k*-correction to the same set of SNPs had little effect: the correlations across DNA pools ranged from 0.958 to 0.975 (average of 0.966) across both microarrays, and the allele frequency differences across DNA pools ranged from 0.047 to 0.06 (average of 0.056). The correlation between the uncorrected and *k*-corrected allele frequency estimates averaged across the five DNA pools was 0.980.

Presenting these differences as standard deviations (SDs) is illuminating in relation to issues of power for the SNP-MaP method. SDs for SNP-MaP allele frequency estimates across the five DNA pools for both the XbaI and HindIII microarrays are both similar and small (0.044 and 0.041, respectively). Figure 1 illustrates the distribution of SDs across the 100K microarray set for 109 994 SNPs for which data were available for all five DNA pools.

Using the average SD of 0.043, we modeled power in STATA with the 'sampsi' command. Five independent case pools and five independent control pools would yield 80% power ($P = 0.05$, two-tailed) to detect SNP-MaP allele frequency differences of 0.075 between groups (e.g. 0.500 versus 0.575) and 99% power to detect differences of 0.115.

Reliability: comparing SNP-MaP estimates for the 10K and 100K microarray sets. As an additional check on the reliability of the Affymetrix GeneChip® Mapping 100K set, the same five DNA pools were also assayed on the previously validated Affymetrix GeneChip® Mapping 10K Array Xba I 131 array. SNPs (7956) were assayed both on the 10K XbaI microarray and the 100K XbaI microarrays; for 6597 of these SNPs, SNP-MaP allele frequency estimates were available for both the 10K and 100K microarrays for all five DNA pools. Again, because we wished to examine the effect of *k*-correction, the analysis focuses on 5708 SNPs for which both 10K *k*-values and 100K *k*-values were

Table 2. Correlations for uncorrected and *k*-corrected SNP-MaP allele frequency estimates between five independent DNA pools

	Pool 1	Pool 2	Pool 3	Pool 4
Uncorrected				
Pool 1				
Pool 2	0.970 (0.053)			
Pool 3	0.973 (0.050)	0.977 (0.046)		
Pool 4	0.966 (0.056)	0.970 (0.053)	0.973 (0.050)	
Pool 5	0.960 (0.055)	0.966 (0.055)	0.970 (0.052)	0.962 (0.058)
k-Corrected				
Pool 1				
Pool 2	0.967 (0.055)			
Pool 3	0.970 (0.052)	0.975 (0.047)		
Pool 4	0.962 (0.058)	0.967 (0.055)	0.970 (0.051)	
Pool 5	0.963 (0.057)	0.962 (0.057)	0.966 (0.054)	0.958 (0.060)

Values in parentheses are the mean allele frequency differences. ($N = 97\,605$ SNPs).

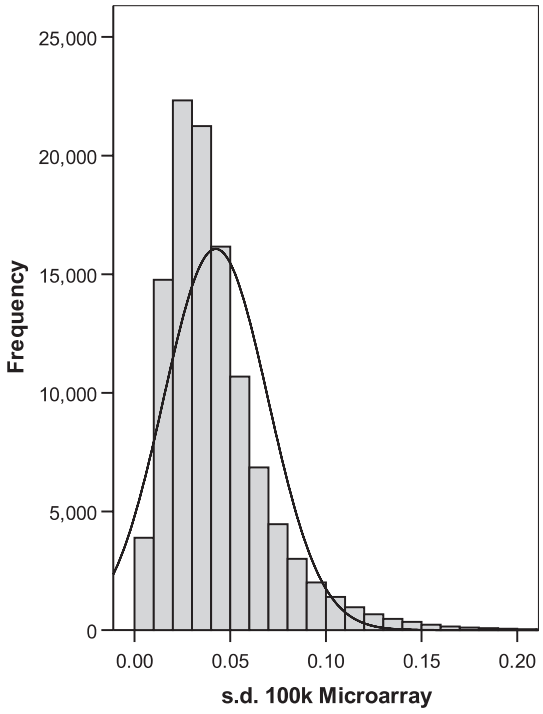

Figure 1. Histogram of SDs for 109 994 SNPs.

available. The uncorrected SNP-MaP allele frequency estimates for the 5708 SNPs were compared across the 10K SNP-MaP and 100K SNP-MaP assays. Similar to the results for the 100K comparisons across five DNA pools, the correlations between the 10K and 100K allelic frequency estimates were high, ranging from 0.911 to 0.967 with an average correlation of 0.940. The mean differences are also small, ranging from 0.055 to 0.090 (average of 0.073). Again, applying *k*-correction to the same set of SNPs had little effect; the correlations ranged from 0.909 to 0.959 (average of 0.934), and the mean differences ranged from 0.054 to 0.081 (average of 0.069).

Validity: comparison of SNP-MaP allele frequencies to individual genotyping

Validity comparisons with the NetAffx™ dataset. Focusing on the subset of SNPs that yielded an allele frequency estimate across all five DNA pools and also had *k*-values ($N = 96\,605$ SNPs), the SNP-MaP allele frequency estimates for the five DNA pools were correlated with population allele frequency estimates from NetAffx™.

As shown in Table 3, the uncorrected SNP-MaP allele frequency estimates (RAS_{AV} scores) correlate strongly with NetAffx™ allele frequency estimates, ranging from 0.933 to 0.943 across the five DNA pools with an average of 0.939, indicating that RAS_{AV} values for the 100K microarray set provide a valid measure of allele frequency in pooled DNA. The average difference in allele frequency estimates was 0.081. The scatterplot shown in Figure 2a indicates that despite the high correlation and the low average allele frequency differences, some of the allele frequency differences are substantial.

Table 3. Correlations for uncorrected and *k*-corrected SNP-MaP and NetAffx™ allele frequency estimates for five independent DNA pools

Array N SNPs	Combined 97 605	XbaI 50 254	HindIII 47 351
Uncorrected SNP-MaP estimates			
Pool 1	0.940 (0.081)	0.929 (0.086)	0.950 (0.076)
Pool 2	0.940 (0.081)	0.931 (0.084)	0.949 (0.077)
Pool 3	0.943 (0.079)	0.936 (0.081)	0.949 (0.077)
Pool 4	0.933 (0.085)	0.925 (0.087)	0.941 (0.082)
Pool 5	0.939 (0.080)	0.936 (0.081)	0.942 (0.079)
Average	0.939 (0.081)	0.931 (0.084)	0.946 (0.078)
k-Corrected SNP-MaP estimates			
Pool 1	0.960 (0.065)	0.953 (0.068)	0.967 (0.062)
Pool 2	0.960 (0.066)	0.957 (0.066)	0.964 (0.066)
Pool 3	0.964 (0.063)	0.960 (0.064)	0.967 (0.062)
Pool 4	0.953 (0.070)	0.948 (0.072)	0.958 (0.069)
Pool 5	0.957 (0.066)	0.962 (0.063)	0.953 (0.070)
Average	0.959 (0.066)	0.956 (0.067)	0.962 (0.066)

Values in parentheses are the mean allele frequency differences.

k-correction of SNP-MaP allele frequency estimates increased the average correlation between SNP-MaP estimates of allele frequencies and NetAffx™ allele frequency estimates from 0.939 to 0.959. The mean difference between *k*-corrected microarray estimates and NetAffx™ was attenuated from 0.081 to 0.066. The scatterplot shown in Figure 2b (as compared with Figure 2a) illustrates the higher correlation and smaller differences for *k*-corrected scores.

Validity comparisons with individual genotyping for the same individuals. We would expect that correspondence between allele frequency estimates from pooled DNA and from individual genotyping would be greater when both sets of estimates are based on the same individuals. For 26 SNPs on the 100K microarray set, individuals in the five DNA pools were individually genotyped in order to obtain allele frequency estimates based on the same sample, rather than the independent NetAffx™ population (20).

As shown in Table 4, the uncorrected SNP-MaP allele frequency estimates (RAS_{AV} scores) correlate well with absolute allele frequency estimates as determined by individual genotyping, with an average correlation of 0.903 and a mean difference of 0.065 (ranging from 0.007 to 0.155). As expected, *k*-correction of the pooled allele frequency estimates increased the correlation to 0.971, and decreased the mean allele frequency difference to 0.036 (ranging from 0.008 to 0.077).

However, it should be noted that using just two independent DNA pool comparisons might result in an unacceptable level of false positive results in a case–control association study. For example, looking at the 26 SNPs in Table 4 we observe SNP-MaP allele frequency differences between independent pools (e.g. pool 1 versus pool 2) ranging from 6 to 27%. Of these, 68.08% are true positive results (i.e. the difference between the two DNA pools at the individual genotyping level is >5%). Therefore, in order to minimize false positive results in a case–control association study using this approach we recommend the use of multiple DNA pools of independent samples (and parametric test statistics) and an independent replication stage. Importantly, we observed a very small number of false negative results (3.84%).

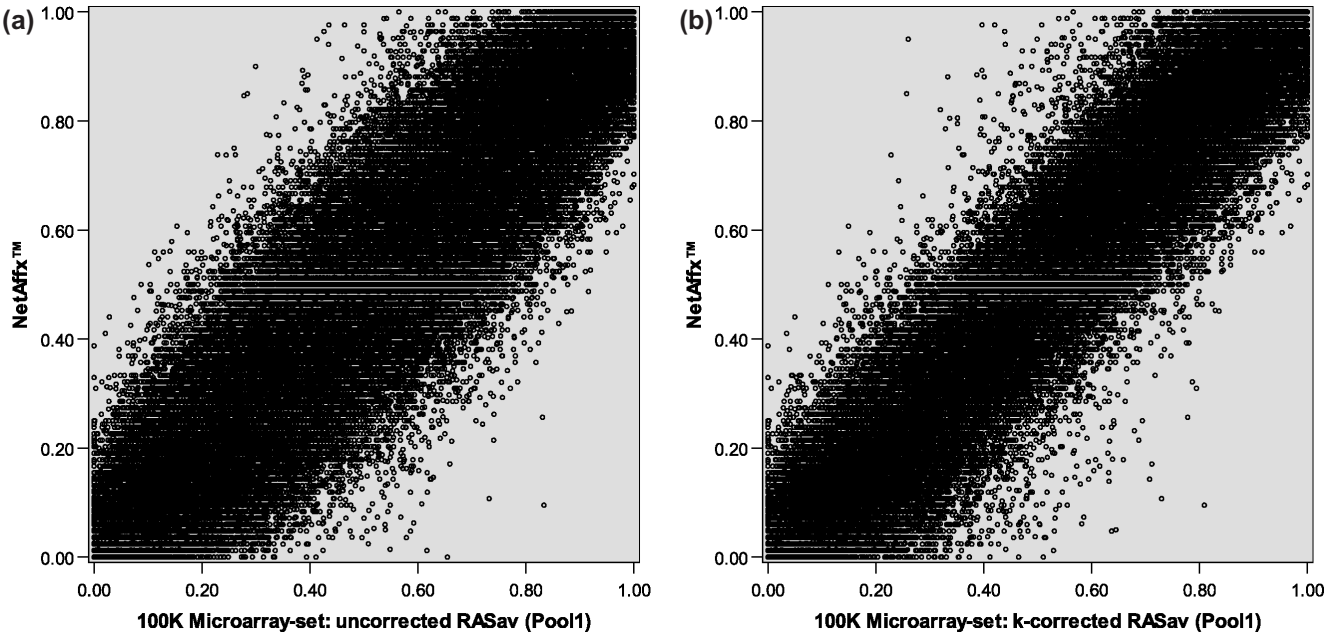

Figure 2. Scatter plot of (**a**) uncorrected and (**b**) *k*-corrected SNP-MaP allele frequency estimates from pooled DNA versus NetAffx™ allele frequency estimates from individual genotyping. The scatterplot shown is for DNA pool 1 for 97 605 SNPs.

Table 4. Summary of SNP-MaP allele frequency estimates versus individual genotyping estimates for 26 SNPs

DbSNP ID	Uncorrected SNP-MaP allele frequency estimates					*k*-Value	Individual genotyping (IG) allele frequency estimates					Average difference between SNP-MaP and IG	
	Pool 1	Pool 2	Pool 3	Pool 4	Pool 5		Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Uncorrected	*k*-Corrected
rs1002666	0.34 (0.31)	0.20 (0.18)	0.21 (0.18)	0.24 (0.22)	0.47 (0.44)	1.16	0.23	0.22	0.23	0.27	0.27	0.08	0.08
rs10493112	0.57 (0.53)	0.59 (0.55)	0.57 (0.53)	0.57 (0.53)	0.59 (0.55)	1.19	0.55	0.53	0.55	0.55	0.53	0.04	0.02
rs1343726	0.31 (0.32)	0.29 (0.29)	0.30 (0.31)	0.30 (0.31)	0.37 (0.38)	0.97	0.31	0.33	0.28	0.29	0.32	0.03	0.03
rs1386468	0.31 (0.32)	0.32 (0.32)	0.31 (0.31)	0.27 (0.27)	0.23 (0.23)	0.97	0.32	0.34	0.29	0.32	0.28	0.03	0.03
rs1480952	0.33 (0.3)	0.35 (0.32)	0.36 (0.32)	0.29 (0.26)	0.31 (0.28)	1.17	0.29	0.29	0.28	0.28	0.27	0.05	0.02
rs2050632	0.60 (0.58)	0.60 (0.58)	0.65 (0.62)	0.65 (0.63)	0.67 (0.65)	1.10	0.57	0.60	0.64	0.58	0.60	0.03	0.03
rs2254209	0.23 (0.23)	0.18 (0.18)	0.18 (0.18)	0.21 (0.21)	0.2 (0.2)	0.97	0.22	0.18	0.19	0.21	0.20	0.01	0.01
rs2292734	0.48 (0.41)	0.58 (0.51)	0.55 (0.48)	0.53 (0.46)	0.46 (0.39)	1.31	0.37	0.41	0.46	0.40	0.37	0.12	0.05
rs2382591	0.18 (0.16)	0.16 (0.14)	0.18 (0.16)	0.20 (0.18)	0.18 (0.17)	1.11	0.17	0.15	0.14	0.13	0.16	0.03	0.02
rs2409411	0.73 (0.7)	0.77 (0.74)	0.78 (0.75)	0.72 (0.69)	0.62 (0.59)	1.17	0.67	0.66	0.68	0.67	0.61	0.07	0.04
rs2593963	0.41 (0.39)	0.48 (0.46)	0.50 (0.48)	0.54 (0.52)	0.54 (0.52)	1.09	0.42	0.39	0.42	0.42	0.43	0.08	0.07
rs2832886	0.57 (0.6)	0.47 (0.5)	0.48 (0.51)	0.54 (0.58)	0.53 (0.56)	0.87	0.60	0.59	0.54	0.53	0.56	0.05	0.03
rs2834036	0.69 (0.66)	0.67 (0.65)	0.62 (0.59)	0.67 (0.64)	0.67 (0.65)	1.12	0.66	0.66	0.61	0.63	0.59	0.03	0.02
rs3811021	0.72 (0.85)	0.70 (0.84)	0.65 (0.81)	0.61 (0.78)	0.58 (0.75)	0.45	0.81	0.81	0.80	0.79	0.81	0.15	0.03
rs3935801	0.34 (0.36)	0.30 (0.31)	0.29 (0.3)	0.37 (0.39)	0.45 (0.47)	0.92	0.40	0.36	0.34	0.33	0.41	0.05	0.05
rs4128492	0.74 (0.8)	0.58 (0.67)	0.64 (0.72)	0.65 (0.73)	0.59 (0.68)	0.69	0.77	0.77	0.75	0.76	0.76	0.12	0.06
rs4509467	0.55 (0.56)	0.72 (0.73)	0.72 (0.73)	0.64 (0.66)	0.65 (0.66)	0.95	0.67	0.64	0.69	0.69	0.63	0.06	0.06
rs4754752	0.79 (0.78)	0.80 (0.79)	0.83 (0.82)	0.81 (0.8)	0.82 (0.81)	1.05	0.80	0.79	0.79	0.77	0.79	0.03	0.02
rs6691482	0.45 (0.4)	0.49 (0.44)	0.52 (0.47)	0.51 (0.46)	0.51 (0.46)	1.21	0.41	0.41	0.43	0.46	0.45	0.06	0.02
rs6763768	0.42 (0.34)	0.60 (0.52)	0.45 (0.36)	0.55 (0.46)	0.48 (0.39)	1.43	0.34	0.41	0.34	0.39	0.37	0.13	0.05
rs7197569	0.57 (0.54)	0.65 (0.63)	0.55 (0.53)	0.58 (0.55)	0.58 (0.56)	1.09	0.53	0.61	0.56	0.57	0.57	0.02	0.02
rs725272	0.64 (0.58)	0.62 (0.55)	0.67 (0.61)	0.68 (0.62)	0.70 (0.64)	1.30	0.58	0.58	0.57	0.62	0.60	0.07	0.02
rs726523	0.31 (0.29)	0.23 (0.21)	0.28 (0.26)	0.28 (0.26)	0.32 (0.29)	1.13	0.27	0.19	0.24	0.26	0.22	0.05	0.03
rs758240	0.64 (0.74)	0.57 (0.67)	0.40 (0.51)	0.43 (0.54)	0.63 (0.73)	0.64	0.67	0.64	0.60	0.61	0.65	0.10	0.07
rs9301694	0.20 (0.34)	0.21 (0.36)	0.08 (0.16)	0.19 (0.32)	0.16 (0.28)	0.48	0.32	0.34	0.31	0.33	0.31	0.16	0.05
rs991684	0.26 (0.35)	0.25 (0.33)	0.25 (0.34)	0.28 (0.37)	0.20 (0.28)	0.67	0.31	0.30	0.29	0.32	0.27	0.05	0.04
Average												**0.065**	**0.036**

Values in parentheses show *k*-corrected SNP-MaP allele frequency estimates.

Non-specific hybridization

Non-specific hybridization for SNPs on the 100K microarray set. For SNPs with DS_{snp} values ≥ 0.04, the distribution of DS_{snp} values for the XbaI and HindIII microarrays were compared for each of the five DNA pools. Average DS_{snp} values were similar for XbaI and HindIII: 0.408 and 0.418, respectively. The distribution of DS_{snp} values obtained for the reference individual is also similar (average DS_{snp} value of 0.37, for XbaI and 0.38 for HindIII). These results

indicate that DNA pools are not inherently noisier than individual samples.

DISCUSSION

Despite reasons to expect that SNP-MaP estimates for the 100K microarray set might not be as reliable and valid as the 10K microarray, the present results for the 100K microarray set are as promising as our previous results for the 10K microarray (7). For the 100K microarray set, allelotyping of SNPs for pooled DNA was as successful as individual genotyping—95% of SNPs yielded interpretable results ($DS_{snp} \geqslant 0.04$), which means that results for nearly 110 000 SNPs can be expected for the 100K microarray set. Concerning reliability, the average correlation among the five subpools in the present study using the 100K microarray set was 0.969; our previous work on the 10K yielded an average correlation of 0.955. Concerning validity, the average correlation for the 100K microarray set was 0.939 between our SNP-MaP allele frequency estimates using pooled DNA allelotyped and individual genotyping results from the NetAffx standardization sample; for the 10K microarray, the average correlation was 0.904.

As expected, *k*-correction made no difference for reliability which involves relative comparisons between comparing allele frequency estimates for different groups; it should be emphasized that relative comparisons between groups such as case versus control groups is the purpose of SNP-MaP. For validity comparisons between SNP-MaP allele frequency estimates from pooled DNA and estimates based on individual genotyping, *k*-correction improved the correlations.

Despite the high reliability of SNP-MaP estimates, the average difference in allele frequency estimates from pooled DNA is 0.036. In other words, SNP-MaP can only detect allele frequency differences >0.036 between two DNA pools. In order to increase the sensitivity to detect allele frequency differences between groups, multiple DNA pools of independent subsamples from each group are recommended. The use of multiple independent DNA pools also permits the use of parametric statistics because it assesses sampling variation. With five independent subpools as in the present experiment, the SD is 0.041, which implies that allele frequency differences of 0.075 between groups (e.g. allele frequencies of 0.500 for cases and 0.575 for controls) can be detected with 80% power ($P = 0.05$, two-tailed). Doubling the number of replicate DNA pools from 5 to 10 pools does not alter the SD but will alter the SEM by a function of the square root of the number of replicates. That is, with five replicates we observed a mean SEM of 0.19, whilst 10 replicates should yield an SEM of ~0.013, which would yield 80% power to detect differences of 0.053 and 99% power to detect differences of 0.082. Although doubling the cost of an experiment seems a considerable price to pay for these small gains in power, we advocate the use of 10 replicates in order to maximize power to detect QTLs of small effect size.

We also recommend that individual genotyping be used to confirm SNP-MaP screening. Because SNP-MaP estimates of allele frequency involve errors of estimation due to pooling DNA, group differences in allele frequency estimates will be reduced when SNPs nominated by SNP-MaP are individually genotyped. For this reason, it is unlikely that allele frequency differences between groups as small as 0.05 can be detected reliably—a more reasonable target is SNP-MaP differences >0.10. Power to detect allele frequency differences at the confirmation stage of individual genotyping depends directly on sample size.

Although power is the crucial issue in detecting QTL associations of small effect size, the issue of the balance between false positive and false negative results becomes especially important when so many tests are conducted. For example, using a nominal *P*-value of 0.05, 5000 statistically significant results are expected by chance alone; winnowing the true results from the false positives will be difficult to resolve statistically. Although the obvious statistical solution is to increase the *P*-value to protect against false positive results due to multiple testing (21), a multistage approach could provide a better balance between false positive and false negative findings (3). In the end, the solution to this conundrum will be empirical rather than statistical: independent replication.

It is generally agreed that >100 000 SNPs are needed for genomewide association scans. Because the SNP-MaP approach works equally well for the 100K microarray set as for the 10K microarray, we anticipate that the approach will also work for the 500K microarray set which is now available. It should be mentioned that the SNP-MaP approach is also likely to work for any other SNP microarrays such as gene-based microarrays, or microarrays with functional SNPs that would permit more powerful direct association analyses rather than indirect association analyses that rely on linkage disequilibrium between SNPs and QTLs.

Limitations of the SNP-MaP approach include the additional error that comes from estimating an average allele frequency from pooled DNA rather than from each individual. Accuracy would of course be better if each individual's DNA were genotyped on separate microarrays, but the expense would prohibit most researchers from studying the very large samples needed to detect QTLs of very small effect size. For example, assuming a cost of £500 per 100K microarray set, it would cost £500 000 to genotype a sample of 1000 individuals on separate microarrays. In comparison, a SNP-MaP case–control study using 10 independent case pools and 10 independent control pools with a replication design of an additional 10 case and 10 control pools would cost £30 000, including the cost of DNA pool construction. The cost of confirmation with individual genotyping will then depend largely upon how many statistically significant SNPs are selected. Assuming a cost of £0.05 per genotype, even if 4700 SNPs (far more SNPs than would reasonably be followed up) were individually genotyped the total SNP-MaP study cost would be half that of using separate microarrays—£250 000.

Our results indicate that SNP-MaP approach yields substantial reliability and validity to screen for the largest allele frequency differences between case and control groups. A greater limitation is that pooled DNA can only be used to estimate allele frequencies rather than genotypic frequencies, which means that haplotypes cannot be investigated at the SNP-MaP screening stage, although haplotypes could be incorporated into individual genotyping strategies at the confirmation stage. These costs are offset by the tremendous benefits of

screening many thousands of SNPs using the very large samples needed to detect QTLs of small effect size.

ACKNOWLEDGEMENTS

This work was supported by UK Medical Research Council grant G0500079 and Wellcome grant GR75492MA. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Syvanen,A.C. (2005) Toward genome-wide SNP genotyping. *Nature Genetics*, **37**, S5–S10.
2. Carlson,C.S., Eberle,M.A., Kruglyak,L. and Nickerson,D.A. (2004) Mapping complex disease loci in whole-genome association studies. *Nature*, **429**, 446–452.
3. Thomas,D.C., Haile,R.W. and Duggan,D. (2005) Recent developments in genomewide association scans: a workshop summary and review. *Am. J. Hum. Genet.*, **77**, 337–345.
4. Zondervan,K.T. and Cardon,L.R. (2004) The complex interplay among factors that influence allelic association. *Nature Rev. Genet.*, **5**, 89–100.
5. Sham,P.C., 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.
6. Butcher,L.M., Meaburn,E., Liu,L., 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.
7. Meaburn,E., Butcher,L.M., Liu,L., Fernandes,C., Hansen,V., Al-Chalabi,A., Plomin,R., Craig,I.W. and Schalkwyk,L. (2005) Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics*, **6**, 52.
8. 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 6,000 children. *Hum. Mol. Genet.*, **14**, 1315–1325.
9. Matsuzaki,H., Dong,S., Loi,H., Di,X., Liu,G., Hubbell,E., Law,J., Berntsten,T., Chadha,M., Hui,H. *et al.* (2004) Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nature Methods*, **1**, 109–111.
10. Ke,X., Hunt,S., Tapper,W., Lawrence,R., Stavrides,G., Ghori,J., Whittaker,P., Collins,A., Morris,A.P., Bentley,D. *et al.* (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.*, **13**, 577–588.
11. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
12. Holden,A.L. (2002) The SNP consortium: summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques* Suppl. **22–24**, 26.
13. Patil,N., Berno,A.J., Hinds,D.A., Barrett,W.A., Doshi,J.M., Hacker,C.R., Kautzer,C.R., Lee,D.H., Marjoribanks,C., McDonough,D.P. *et al.* (2005) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.
14. Trouton,A., Spinath,F.M. and Plomin,R. (2002) Twins Early Development Study (TEDS): a multivariate, longitudinal genetic investigation of language, cognition and behaviour problems in childhood. *Twin Res.*, **5**, 444–448.
15. Freeman,B., Smith,N., Curtis,C., Huckett,L., Mill,J. and Craig,I. (2003) DNA from buccal swabs recruited by mail: evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping. *Behav. Genet.*, **33**, 67–72.
16. 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.
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 100 K SNPs on oligonucleotide microarrays. *Bioinformatics*, **21**, 1958–1963.
18. 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.
19. 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.
20. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.
21. Freimer,N.B. and Sabatti,C. (2005) Guidelines for association studies in human molecular genetics. *Hum. Mol. Genet.*, **14**, 2481–2483.