Hum Genet (2002) 110 : 471–478
DOI 10.1007/s00439-002-0706-6

ORIGINAL INVESTIGATION

**Nadine Norton · Nigel M. Williams · Hywel J. Williams
Gillian Spurlock · George Kirov · Derek W. Morris
Bastiaan Hoogendoorn · Michael J. Owen
Michael C. O'Donovan**

Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools

Received: 29 November 2001 / Accepted: 31 January 2002 / Published online: 23 March 2002

Abstract Detecting alleles that confer small increments in susceptibility to disease will require large-scale allelic association studies of single-nucleotide polymorphisms (SNPs) in candidate, or positional candidate, genes. However, current genotyping technologies are one to two orders of magnitude too expensive to permit the analysis of thousands of SNPs in large samples. We have developed and thoroughly validated a highly accurate protocol for SNP allele frequency estimation in DNA pools based upon the SNaPshot (Applied Biosystems) chemistry adaptation of primer extension. Using this assay, we were able to estimate the difference in allele frequencies between pooled cases and controls (Δ) with a mean error of 0.01. Moreover, when we genotyped seven different SNPs in a single multiplex reaction, the results were similar, with a mean error for Δ of 0.008. The assay performed well for alleles of low frequency alleles (*f*~0.05) and was accurate even with relatively poor quality DNA template extracted from mouthwashes. Our assay conditions are generalisable, universal, robust and, therefore, for the first time, permit high-throughput association analysis at a realistic cost.

Introduction

Candidate gene association analysis is currently the only viable approach for detecting alleles that confer small increments in susceptibility to complex phenotypes (Risch and Merikangas 1996). However, to realise the potential of association approaches, assays are required for genotyping single-nucleotide polymorphisms (SNPs) that are applicable to most SNPs, require minimal optimisation, are rapid and are fully or semi-automated. Current genotyping technologies offer some of the above characteristics but are at least one to two orders of magnitude too expensive to permit the analysis of thousands of SNPs in large samples.

Previously, we (Daniels et al. 1998; Kirov et al. 2000) and others (Pacek et al. 1993; Barcellos et al. 1997; Shaw et al. 1998) have addressed the costs and labour involved in large-scale genotyping by developing methods of analysing micro-satellite allele frequencies in pooled DNA samples as originally suggested by Arnheim et al. (1985), whereas others have explored appropriate statistical methods for pooled analyses (Risch and Teng 1998). We have now developed and assessed a highly accurate protocol for SNP allele frequency estimation by using the SNaPshot (Applied Biosystems) modification of primer extension chemistry (Sokolov 1990; Kuppuswamy et al. 1991; Syvanen et al. 1993). The assay is universal (simple to multiplex) and we have used a single set of conditions to analyse more than 150 SNPs to date. Moreover, it requires no optimisation other than a simple assessment of reaction efficiency and allele-specific products are resolved by automated analysis on a capillary sequencer. As the cost per SNP is trivial (essentially the cost of the primers), the properties of this assay should permit laboratories with facilities for automated primer design, liquid handling, polymerase chain reaction (PCR) and high capacity capillary sequencers to estimate economically the allele frequencies of SNPs in genome databases and to undertake an association analysis of many thousands of candidate SNPs in large samples each year. The method should also permit large-scale candidate gene analysis in laboratories without these resources provided they have access to gel-based sequencers. However, there are a number of study settings when pooled analysis is not applicable. It is impossible to construct haplotypes, which may be crucial for the success of linkage disequilibrium mapping, although this may be offset in part by the ability to test many more markers than would be permitted by individual genotyping. Pooled analysis also makes it impos-

N. Norton · N.M. Williams · H.J. Williams · G. Spurlock
G. Kirov · D.W. Morris · B. Hoogendoorn · M.J. Owen
M.C. O'Donovan (✉)
Department of Psychological Medicine,
University of Wales College of Medicine,
Heath Park, Cardiff, CF14 4XN, UK
e-mail: odonovanmc@cardiff.ac.uk,
Tel.: +44-12920-743242, Fax: +44-12920-747839

sible to study epistasis, heterosis, and undertake *post hoc* analyses of sub-phenotypes.

Materials and methods

DNA pools

DNA pools were constructed from samples that are in routine use in our laboratory for association studies of schizophrenia, bipolar disorder and dyslexia. We determined the concentrations of all DNA samples used for pool construction by using the PicoGreen dsDNA Quantitation Reagent (Molecular Probes, Eugene, Ore.) in a Labsystems Ascent Fluoroskan (LifeSciences International, Basingstoke, UK). The quality of DNA was initially assessed by PCR amplification of microsatellite markers under standard conditions. Only DNA samples that amplified were included in DNA pools. Based on the spectrophotometer reading (at 260 nm) of the concentrated stocks, water was added to DNA samples extracted from blood to produce a target concentration of 40 ng/µl. Samples were then allowed to equilibrate at 4°C for 48 h. Each 40-ng/µl sample was then accurately quantified by the PicoGreen method, diluted to 4 ng/µl and again allowed to equilibrate at 4°C for 48 h. Each 4-ng/µl dilution was subsequently quantified by the PicoGreen method and only those at 4 ng/µl (±0.5 ng/µl) were accepted for pooling. Samples whose concentration was above this were rediluted to 4 ng/µl based upon the new readings; this process was repeated iteratively until a concentration of 4 ng/µl (±0.5 ng/µl) was obtained. Samples whose concentration was below this were rediluted from stocks through the same stages as above, with the dilution factor being dictated by the PicoGreen readings rather than spectrophotometry. Pools were then constructed by combining equal volumes of each sample. Case-control pools were constructed from patients with schizophrenia (affected) and blood-donor controls and ranged in size from 130–189 DNA samples.

Table 1 Oligonucleotide sequences for each SNP (*F*, *R* forward and reverse PCR primers, respectively, *EXT* corresponding extension primer for each SNP assay)

SNP	Sequence
DRD3 S9G, 25A→G	(F) 5’-GCTCTATCTCCAACTCTCACA (R) 5’-AAGTCTACTCACCTCCAGGTA (EXT) 5’-CTCTGGGCTATGGCATCTCTGAGCCAGCTGAGT
DRD3 –6733A→G	(F) 5’-AAGCTGGAAAAGCAGCACTC (R) 5’-CTCCTGCAGCCATTTACTGA (EXT) 5’-AGTTTGCTTTGCTTGGGTATGTCTGCT
DRD3 –4147C→T	(F) 5’-CGTCAACTTCCATGCTGCTAT (R) 5’-TAAAAAGGCAGGGGAACAGA (EXT) 5’-TCTGTAAGTCTTAATGAGGTGCTAAGGAGGAA
DRD3 –712G→C	(F) 5’-TTACATGGGAAGAATCTGGAGTGCA (R) 5’-GAGGGTGTGAGGTAGACAGATTGTG (EXT) 5’-CTAACTCTGGGACCTTATGCATATTACTTTACCTCT
DRD3 –205A→G	(F) 5’-ATCTCCTCCAGGTCAAGACTCAATT (R) 5’-CCTGTGAGGAGACAGAAAACAATAT (EXT) 5’-GAATGGGAGCTTCAAAGGGAAGGAATTAA
AC004169/30,419:G→T	(F) 5’-TGCACCCACATGCATTTCAG (R) 5’-TAGCTCACAGTGCCTGCGG (EXT) 5’-CTCCATGGGTGCACAGACGG
HTR2 A 102T→C	(F) 5’-TCTGCTACAAGTTCTGGCTT (R) 5’-CTGCAGCTTTTTCTCTAGGG (EXT) 5’-GGCTCTACAGTAATGACTTTAACTC
HTR2 A –1438A→G	(F) 5’-AACCAACTTATTTCCTACCAC (R) 5’-AAGCTGCAAGGTAGCAACAGC (EXT) 5’-TGGCTTTGGATGGAAGTGCC
PLCB2 IVS2–8G→A	(F) 5’-CTGGACTTTTTGTCCCACAT (R) 5’-TGCCCCATGGAGCTAGTA (EXT) 5’-CCCACCGGGATCCGCACCCT
PLCB2 IVS12–24C→T	(F) 5’-AACGCTGACCTTCTGTTCAT (R) 5’-AGGCTCAAATGTCCCACA (EXT) 5’-GGTGCTGAGCGGCTGAACCC
PLCB2 IVS23–39C→A	(F) 5’-AGCCCTATTTATGGGAGAAGG (R) 5’-CTCATCCCCGAGATCACC (EXT) 5’-TCAGGAAGGTGGCTTGACAG
NTS –167C→G	(F) 5’-GATACTGGGGGTTCTTTGTC (R) 5’-GAGCAACTCTTCTCCCAGAT (EXT) 5’-GCAAAGATAATGTCTGTA
AP002831/106,491:A→G	(F) 5’-TTAGCAAAGAGTCAAGCGCA (R) 5’-CACCAATACCTGTCAGTGGC (EXT) 5’-GAAAAAGAACTGTTATTGGAGTC
COMT –287C→T	(F) 5’-TAGTAACAGACTGGGCACGAA (R) 5’-GTTCAAAGGGCATTTATCATG (EXT) 5’-TGTGAGTATGGGAAGGGGAA
GRM7 674A→T	(F) 5’-ATGAACAAGGATCTCTGTGC (R) 5’-TCCAGCTTGCTCCATCTCT (EXT) 5’-GAACAAGGATCTCTGTGCTGACT

Family-based association pools were constructed from 111 probands with bipolar disorder (affected) and their 222 parents (controls). Pools of 171 cases with dyslexia (affected) and 143 controls were constructed with DNA extracted from mouthwashes. For these, the DNA was quantified by using the same protocol as described for DNA extracted from blood, except that each sample was diluted to a final concentration of 2 ng/µl (±0.5 ng/µl).

We genotyped all samples included in the pools individually for all SNPs. Most were genotyped by restriction fragment length polymorphism (RFLP) analysis; a detailed protocol for each polymorphism is available upon request. Details of the primers used in each pooled assay are shown in Table 1. Routine thermocycling conditions were used for PCR; the reaction volume was 12 µl, which comprised 24 ng pooled genomic DNA, 100 µM dNTPs and 0.5 U *Taq* DNA Polymerase (QIAGEN, Crawley, UK) in the buffer provided by the manufacturer.

Pooled genotyping

Single marker pooling

We prepared PCR fragments for primer extension by incubating 12 µl PCR product with 2 U exonuclease I (Amersham) and 1 U shrimp alkaline phosphatase (Amersham) for 45 min at 37°C on a thermocycler followed by 15 min at 80°C. Primer extension was performed by combining 1 µl treated PCR product with 5 µl SNaPshot kit, 0.15 pmol extension primer and 3 µl water. The reaction mix was incubated at 94°C for 2 min and then was subject to 25 cycles of 95°C for 5 s, 50°C for 5 s and 60°C for 5 s. To prevent unincorporated fluorescent ddNTPs obscuring the primer extension products during electrophoresis, the reactions were treated with 1 U shrimp alkaline phosphatase at 37°C for 45 min followed by 15 min at 80°C. Aliquots of 1 µl SNaPshot product and 9 µl Hi-Di

Table 2 Accuracy of estimating single marker allele frequencies in DNA pools. Data for primer extension reactions performed for each marker in a single reaction. The estimated allele frequencies and the standard error of the mean of replicate pooling measurements (*Pool*), together with the corresponding real (*Real*) allele frequency (as determined from individual genotyping) in case and control samples are presented for each SNP (*Control* and *Affected*, respectively).
n Number of samples included in each pool, *k* correction factor (ratio of peak heights for each allele of a heterozygote) for each marker, *Δ* difference in allele frequencies between controls and cases, *Error* error of pooling defined as the discrepancy between the Δ obtained by pooled and individual genotyping. The estimate of error assumes that individual genotyping is 100% accurate

Gene/SNP		Control	Affected	Δ	Error
DRD3 S9G,	*n*	184	184		
25A→G	Pool	0.665 (0.007)	0.662 (0.01)	0.003	0.007
(k=0.86)	Real	0.665	0.669	–0.004	
DRD3	*n*	184	184		
–6733A→G	Pool	0.963 (0.003)	0.978 (0.00)	–0.015	0.013
(k=0.47)	Real	0.918	0.946	–0.028	
DRD3	*n*	184	184		
–4147C→T	Pool	0.916 (0.00)	0.921 (0.002)	–0.005	0.001
(k=0.47)	Real	0.978	0.984	–0.006	
DRD3	*n*	184	184		
–712G→C	Pool	0.756 (0.004)	0.718 (0.003)	0.038	0.009
(k=0.77)	Real	0.769	0.740	0.029	
DRD3	*n*	184	184		
–205A→G	Pool	0.300 (0.001)	0.306 (0.005)	–0.006	0.014
(k=1.38)	Real	0.352	0.344	0.008	
AC004169/30,	*n*	222	111		
419: G→T	Pool	0.965 (0.001)	0.968 (0.002)	–0.003	0.004
(k=1.0)	Real	0.973	0.972	0.001	
HTR2 A	*n*	189	180		
102T→C	Pool	0.358 (0.004)	0.410 (0.005)	–0.052	0.022
(k=1.35)	Real	0.39	0.42	–0.03	
HTR2 A	*n*	189	180		
–1438A→G	Pool	0.428 (0.002)	0.443 (0.004)	–0.015	0.015
(k=0.90)	Real	0.390	0.420	–0.030	
PLCB2	*n*	171	143		
IVS2–8G→A	Pool	0.617 (0.004)	0.594 (0.005)	0.023	0.018
(k=0.97)	Real	0.646	0.641	0.005	
PLCB2	*n*	171	143		
IVS12–24C→T	Pool	0.490 (0.007)	0.486 (0.010)	0.004	0.016
(k=0.84)	Real	0.485	0.497	–0.012	
PLCB2	*n*	171	143		
IVS23–39C→A	Pool	0.283 (0.005)	0.209 (0.012)	0.074	0.010
(k=1.29)	Real	0.232	0.168	0.064	
NTS	*n*	157	160		
–167C→G	Pool	0.272 (0.003)	0.275 (0.006)	–0.003	0.006
(k=0.99)	Real	0.245	0.254	–0.009	
AP002831/106,	*n*	130	146		
491:A→G	Pool	0.200 (0.003)	0.187 (0.003)	0.013	0.001
(k=0.81)	Real	0.200	0.188	0.012	
COMT	*n*	157	160		
–287C→T	Pool	0.441 (0.009)	0.449 (0.003)	–0.008	0.006
(k=1.51)	Real	0.436	0.450	–0.014	
GRM7	*n*	130	146		
674A→T	Pool	0.464	0.533	–0.069	0.009
(k=0.71)	Real	0.481	0.541	–0.060	

formamide were combined in a 96-well 3100 optical microamp plate (Applied Biosystems), which was loaded onto a 3100 DNA sequencer (Applied Biosystems). Reactions were electrophoresed on a 36-cm capillary array at 60°C by using POP4 polymer, dye set "E" and Genescan run module "SNP36POP4_default". Electrophoresis data were processed by using Genescan Analysis version 3.7 (Applied Biosystems). Peak heights of the allele-specific extended primers were determined by using Genotyper version 2.5 (PE Biosystems) and the allele frequency in each pool was determined as the mean of four primer extension assays corrected for the degree of unequal allelic representation detected in a heterozygote as described below.

Relationship between peak height and primer concentration

To determine the relationship between primer extension signal strength and primer concentration, primer extension reactions for three different SNPs (AP002831/106,491, COMT and NT) were prepared by adding 0.031, 0.063, 0.125, 0.25 and 0.5 pmol each extension primer. Each reaction was prepared as five replicate reactions. The mean peak height for each reaction was then plotted against amount of primer added.

Multiplex primer extension

To investigate multiplex primer extension analysis, reactions were prepared in exactly the same way as described, except that equal volumes of the PCR products for each SNP were combined and then 1 μl of the mix was treated with shrimp alkaline phosphatase and exonuclease I as above. Similarly, all extension primers were combined and primer extension was performed for all SNPs simultaneously by using an aliquot of the mix. The seven extension primers used in the multiplex reaction were designed to span from 18 bp to 38 bp. To ensure that all reactions in multiplex yielded approximately equal-sized product peaks for all SNPs, the concentration of the extension primer for each reaction was adjusted based upon the peak height of a test genotype as described below.

Correction for unequal allelic detection

For a di-allelic marker, the primer extension products for each allele are not equally represented. Possible explanations include differential PCR amplification of alleles (Liu et al. 1997) and differential efficiencies of the incorporation of the ddNTPs for each allele-specific reaction (Haff and Smirnov 1997; Barnard et al. 1998). An additional explanation is the unequal emission energies of the different fluorescent dyes. In order to allow for unequal representation of alleles, the estimated allele frequencies from pools were corrected by using the mean of the ratios obtained from four analyses of a heterozygote as given by the equation: Frequency of allele $A = A/(A+kB)$ where A and B are the peak heights of the primer extension products representing alleles A and B in pools and k is the mean of four replicates of A/B ratios observed in a heterozygote (Hoogendoorn et al. 2000).

Results

DNA pooling for single markers

We tested the accuracy of pooled analysis by using 15 different SNPs in pools of cases (schizophrenia, bipolar disorder or dyslexia) and pools constructed either from unrelated controls (used in studies of schizophrenia or dyslexia) or pools constructed from the parents of the pooled bipolar cases. The data from pooled and individual genotyping are given in Table 2. Estimation of differences between cases and control pools (Δ) was extremely accurate with the mean error for Δ of 0.01 (maximum error: 0.022). Differences between cases and controls were estimated to an accuracy of less than 1.6 alleles in 100 for 13 of the 15 SNPs. Estimation of absolute allele frequencies in the pools agreed well with the results from individual genotyping with a mean error of 0.023 (maximum error: 0.063).

Only one of the 15 SNPs tested was "associated" with affected status at the conventional level of $P \leq 0.05$. This was the PLCB2 IVS23–39C→A SNP in the dyslexia sample (P=0.033). A similar result was obtained by individual genotyping (P=0.05).

DNA pooling in multiplex reactions

Figure 1 shows the average peak height and the range in measurements for five replicate reactions for three SNPs at five primer concentrations. In general, to enhance throughput, we do not wish to quantitate each PCR product. Therefore, in a multiplex reaction, each genotyping reaction will be based on differing amounts of input template. Different primer extension reactions also have different efficiencies. These factors result in the extension signals for each SNP

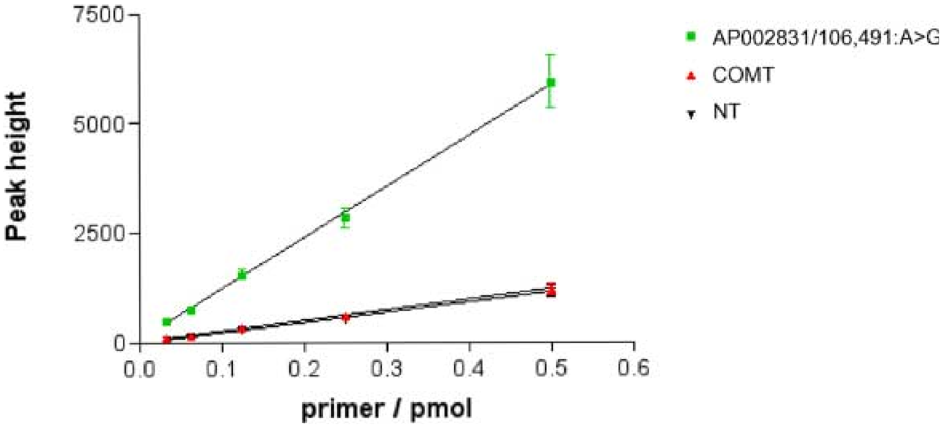

Fig. 1 To determine the relationship between primer extension signal strength and primer concentration, primer extension reactions for three different SNPs (AP002831/106,491, COMT and NT) were prepared by adding 0.031, 0.063, 0.125, 0.25 and 0.5 pmol of each extension primer. All reactions were performed as five replicates. The mean peak height for each reaction was then plotted against the amount of primer added revealing a linear relationship between the two variables

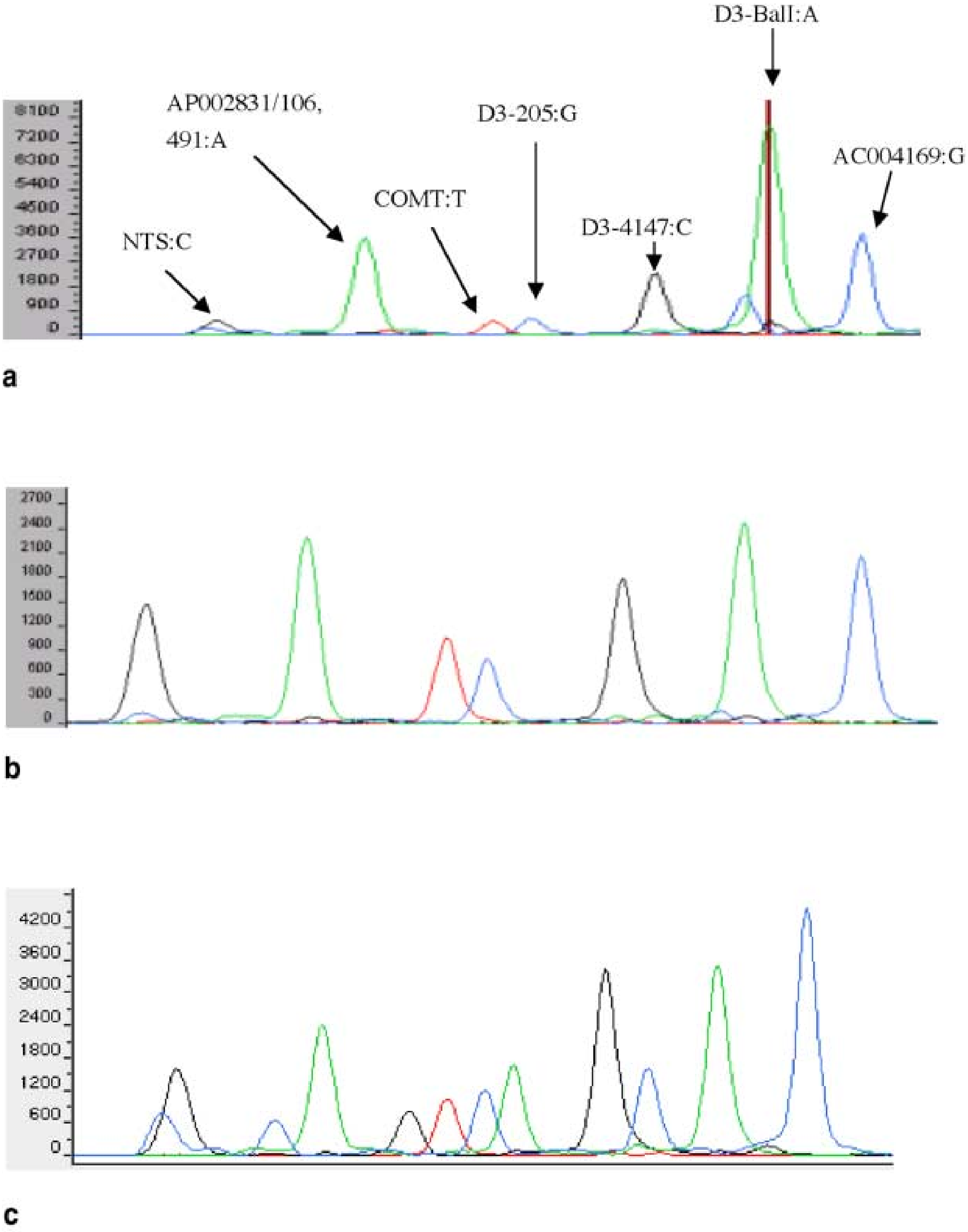

Fig. 2 **a** Seven different SNPs were amplified by using homozygous DNA samples as templates. The products were pooled and subjected to a multiplex primer extension reaction containing all seven extension primers in a single reaction. The extension primer used for each SNP was at a constant concentration of 0.5 pM. **b** As for **a**, except primer concentrations were adjusted as described. **c** As for **a**, except all seven SNPs were amplified from pooled DNA constructed from 130–222 subjects with the primer concentration being adjusted as in **b**

in a multiplex panel having different peak heights at any single set of primer extension conditions (Fig. 2a). This is a potential problem for multiplex primer extension reactions as all assays may not be simultaneously within the linear quantitative range. The simplest variable that can be adjusted is primer concentration and, as primer extension is a linear rather than exponential amplification, we predicted that the amount of product would be linearly related to the concentration of extension primer.

The data in Fig. 1 show that the relationship between primer concentration and peak height in the primer extension reaction is indeed linear. It follows that the peak height for any given primer concentration across the measured range can be predicted by a single measurement of the peak height by using a single primer concentration. This relationship was used to adjust the primer concentration to ensure that reaction strengths for each SNP in multiplex assays were approximately equal with the goal of a homozygous peak height of approximately 2000 fluorescence units.

In subsequent multiplex assays, we made this adjustment from the peak height in the single test reaction that we routinely perform to ensure that the SNP is actually polymorphic. After measuring the height of the peak of the extension product in the test reaction, the concentration of the extension primer for the required peak height is simply calculated as *Y*'/(*Y*/*X*) where, *Y*' is the required peak height, *Y* is the initial peak height and *X* is the initial

Table 3 Data for seven markers genotyped by primer extension simultaneously as a multiplex reaction. Accuracy of estimating the allele frequencies of multiple markers (*n*=7) simultaneously in DNA pools. Other details as in Table 1

Gene/SNP		Control	Affected	Δ	Error
NTS	*n*	157	160		
–167C→G	Pool (multiplex)	0.279 (0.001)	0.291 (0.001)	–0.012	0.003
	Real (individual)	0.245	0.254	–0.009	
AP002831/106,	*n*	130	146		
491:A→G	Pool (multiplex)	0.212 (0.001)	0.223 (0.002)	–0.011	0.023
	Real (individual)	0.200	0.188	0.012	
COMT	*n*	157	160		
–287C→T	Pool (multiplex)	0.469 (0.001)	0.468 (0.001)	0.001	0.015
	Real (individual)	0.436	0.45	–0.014	
DRD3	*n*	184	184		
–205A→G	Pool (multiplex)	0.340 (0.003)	0.336 (0.002)	0.004	0.004
	Real (individual)	0.352	0.344	0.008	
DRD3	*n*	184	184		
–4147C→T	Pool (multiplex)	0.980 (0.001)	0.981 (0.00)	–0.001	0.005
	Real (individual)	0.978	0.984	–0.006	
DRD3 S9G,	*n*	184	184		
25A→G	Pool (multiplex)	0.690 (0.002)	0.689 (0.002)	0.001	0.005
	Real (individual)	0.665	0.669	–0.004	
AC004169/30,	*n*	222	111		
419: G→T	Pool (multiplex)	0.973 (0.001)	0.975 (0.001)	–0.002	
	Real (individual)	0.972	0.973	–0.001	0.001

primer concentration. The effectiveness of this approach is demonstrated in Fig. 2a, b in homozygous DNA samples for seven different SNPs and in Fig. 2c in DNA pools. After adjusting the primer concentrations, we performed primer extension for each of the seven SNPs simultaneously as part of a single multiplex reaction.

Data from individual genotyping and multiplex pooling are presented in Table 3. DNA pooling in multiplex reactions gave results that were comparable to the results of single-marker pooled analysis. The mean error for Δ was 0.008 (maximum: 0.023). The error in estimating absolute allele frequency was 0.017 (maximum: 0.037).

Discussion

Previously, we (Hoogendoorn et al. 2000) and others (Germer et al. 2000) have suggested quantitative SNP allele frequency estimation in DNA pools (DNA pooling) as an interim solution to the practical problems facing association analysis until genuinely cheap and robust high-throughput individual genotyping assays become available. In order for DNA pooling to provide this solution, the pooled assay must be accurate, semi-automated, high-throughput, universal, robust, cheap and based upon widely accessible technology. A few previous studies have suggested that allele frequency estimation in pools can be fairly accurate in a diverse range of assay systems, including RFLP analysis (Breen et al. 2000), kinetic PCR (Germer et al. 2000), denaturing high-performance liquid chromatography (DHPLC)-based analysis of primer extension products (Hoogendoorn et al. 2000), fluorescent single-strand conformation polymorphism analysis (SSCP; Sasaki et al. 2001) and MALDI-TOF mass spectrometry (Ross et al. 2000). Unfortunately, some of these studies have not thoroughly tested the accuracy of the method, as they have only compared the estimated allele frequencies (determined by pooling) with the correct allele frequencies (determined by individual genotyping) for a small number of polymorphisms (Breen et al. 2000; Germer et al. 2000; Ross et al. 2000). Some methods are further limited by lack of automation, particularly RFLP analysis (Breen et al. 2000) and the requirement for time-consuming and occasionally difficult optimisation of reaction conditions, e.g. DHPLC-based analysis of primer extension products (Hoogendoorn et al. 2000). The problems of optimisation are compounded in the case of kinetic PCR by the need for very expensive primers (Germer et al. 2000). Additional problems exist with some of the more promising techniques. The SSCP method has not been validated by comparing pooled data from real complex pools against individual data obtained from individual genotyping and has the extra disadvantage of only being applicable to 73% of SNPs (Sasaki et al. 2001). Clearly, MALDI-TOF mass spectrometry (Ross et al. 2000) confers a high potential for automation but its accuracy is unknown and it requires access to equipment that is available to very few laboratories and that, given its cost, is likely to remain so for several years to come.

We believe that quantitative allele frequency measurement with fluorescent-dye terminators and primer extension provides a ready-to-use approach that meets the suggested criteria of accuracy, semi-automation, high-throughput, universality, robustness and economy, and that is based upon widely accessible technology.

In terms of accuracy, to be applicable in association studies, pooled analysis must yield a fairly accurate estimation of the absolute allele frequency and a more exact estimation of relative allele frequencies. The accuracy that we have achieved and, in particular, the astonishingly accurate estimates of relative allele frequencies are likely to be adequate for most samples. Indeed, our mean 1% error in estimating relative allele frequency differences is only around 0.5% greater than the error rate reported for individual genotyping by a first-class research laboratory (Mein et al. 2000).

Regarding automation, at present, sample loading, analysis, allele frequency estimation and statistical analysis is automated, although we visually check the electropherograms. The degree to which the rest of the process is automated will depend upon the resources of the laboratory. As we do not have a fully automated laboratory, we are uncertain as to how far one can proceed in this direction. However, as the process is similar to sequencing, laboratories geared for large-scale sequencing should be able to automate the process almost completely and consequently achieve extremely high throughput.

The assay is also universal and does not require optimisation post-PCR. To date, we have applied identical protocols to over 150 different SNPs. After PCR, all conditions except primer concentration (which can simply be determined from a single test reaction) are identical regardless of the SNP. We have also shown that the assay is applicable for case-control and haplotype-based haplotype relative risk methods (e.g. AC004169/30,419T→G, in Table 2 represents pools from parent/proband trios). Moreover, we have also shown that the method is accurate even when based on DNA from mouthwashes, as all three PLCB2 SNPs (Table 2) were genotyped in pooled DNA extracted from this source. The assay is therefore generalisable, universal and robust.

The assay, as we have presented it, involves a considerable capital outlay in a capillary sequencer. However, we have also achieved similar results by using ABI 373 and ABI 377 DNA sequencers (data not shown), which are widely available. To avoid the substantial costs of individual genotyping, we have demonstrated the methodology by estimating the accuracy of using four replicates of pools, each containing up to 222 individuals. As the purpose of replication is to allow for variance caused by measurement error; similar accuracy can be expected in larger case-control studies by combining data from a single run of four pools, each containing 225 different cases and a similar number of controls. This will yield accurate allele frequency data representing 2000 genotypes for the cost of 14 genotypes (4 pools of cases, 4 of controls, 4 of a heterozygote and one each of a single test and negative control genotype prior to pooling). The exact cost will vary between countries and even between laboratories within countries depending upon purchasing agreements. We estimate that, in our hands, with modest multiplexing (4 SNPs per run), the reagent cost per complete association study per SNP is £24 including primers. Low cost comes at a penalty. The most important may be that it is impossible to construct haplotypes and indirect association studies may suffer reduced power as a consequence. This may be offset in part by the ability to test many more markers than would be permitted by individual genotyping and is likely to be less important for direct analyses of putative candidate SNPs. Pooled analysis also makes it impossible to study epistasis, heterosis and dominance and to undertake *post hoc* analyses of sub-phenotypes. The balance of cost to benefit in these regards is unknown.

Acknowledgements The MRC (UK) and a JREI (UK) award supported this work.

References

Arnheim N, Strange C, 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 82:6970–6974

Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, Thomson G (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. Am J Hum Genet 61:734–747

Barnard R, Futo V, Pecheniuk N, Slattery M, Walsh T (1998) PCR bias toward the wild-type k-*ras* and *p53* sequences: implications for PCR detection of mutations and cancer diagnosis. Biotechniques 24:684–691

Breen G, Harold D, Ralston S, Shaw D, St Clair D (2000) Determining SNP allele frequencies in DNA pools. Biotechniques 28:464–470

Daniels J, Holmans P, Williams N, Turic D, McGuffin P, Plomin R, Owen MJ (1998) A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. Am J Hum Genet 62:1189–1197

Germer S, Holland MJ, Higuchi R (2000) High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. Genome Res 10:258–266

Haff LA, Smirnov IP (1997) Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. Genome Res 7:378–388

Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshere ML, Spurlock G, Austin J, Stephens MK, Buckland PR, Owen MJ, O'Donovan MC (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

Kirov G, Williams N, Sham P, Craddock N, Owen MJ (2000) Pooled genotyping of microsatellite markers in parent-offspring trios. Genome Res 10:105–115

Kuppuswamy MH, Hoffman JW, Kasper CK, Spitzer SG, Groce SL, Bajaj SP (1991) Single nucleotide primer extension to detect genetic diseases: experimental application to hemophilia B (factor IX) and cystic fibrosis genes. Proc Natl Acad Sci USA 88:1143–1147

Liu Q, Thorland EC, Sommer SS (1997) Inhibition of PCR amplification by a point mutation downstream of a primer. Biotechniques 22:292–296

Mein CA, Barratt BJ, Dunn MG, Siegmund T, Smith AN, Esposito L, Nutland S, Stevens HE, Wilson AJ, Phillips MS, et al (2000) Evaluation of single nucleotide polymorphism typing with Invader on PCR amplicons and its automation. Genome Res 10:330–343

Pacek P, Sajantila A, Syvanen AC (1993) Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. PCR Methods Appl 2:313–317

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Risch N, 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

Ross P, Hall L, Haff LA (2000) Quantitative approach to single-nucleotide polymorphism analysis using MALDI-TOF mass spectrometry. Biotechniques 29:620–629

Sasaki T, Tahira T, Suzuki A, Higasa K, Kukita Y, Baba S, 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

Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A (1998) Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome Res 8:111–123

Sokolov BP(1990) Primer extension technique for the detection of single nucleotide in genomic DNA. Nucleic Acids Res 18: 3671

Syvanen A, Sajantila A, Lukka M (1993) Identification of individuals by analysis of biallelic DNA markers, using PCR and solid-phase minisequencing. Am J Hum Genet 52:46–59