Hum Genet (2000) 107 : 488–493
DOI 10.1007/s004390000397

ORIGINAL INVESTIGATION

**Bastiaan Hoogendoorn · Nadine Norton · George Kirov
Nigel Williams · Marian L. Hamshere · Gillian Spurlock
Jehannine Austin · Mark K. Stephens · Paul R. Buckland
Michael J. Owen · Michael C. O'Donovan**

Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools

Received: 26 June 2000 / Accepted: 30 August 2000 / Published online: 11 October 2000

Abstract At present, the cost of genotyping single nucleotide polymorphisms (SNPs) in large numbers of subjects poses a formidable problem for molecular genetic approaches to complex diseases. We have tested the possibility of using primer extension and denaturing high performance liquid chromatography to estimate allele frequencies of SNPs in pooled DNA samples. Our data show that this method should allow the accurate estimation of absolute allele frequencies in pooled samples of DNA and also of the difference in allele frequency between different pooled DNA samples. This technique therefore offers an efficient and cheap method for genotyping SNPs in large case-control and family-based association samples.

Introduction

Association studies allow the detection of genes with a small effect in complex disorders. However, the number of candidate polymorphisms for any given disorder is usually extremely large, as are the sample sizes required. It is clear therefore that such studies require inexpensive and non-labour intensive methods of genotyping single nucleotide polymorphisms (SNPs) in large populations. At present, various methods exist for genotyping SNPs. However, none provides sufficient economies of cost and labour to permit testing several hundred candidate gene hypotheses, much less the systematic genome scans that have been proposed (Risch and Merikangas 1996). In the future, emerging technologies such as DNA chips might permit genuinely large-scale genotyping endeavours but researchers urgently require interim solutions. Previously, we (Daniels et al. 1998; Kirov et al. 2000) and others (Barcellos et al. 1997; Shaw et al. 1998) have addressed this problem by developing methods of analysing microsatellite allele frequencies in pooled DNA samples, whereas another group has explored appropriate statistical methods for pooled analyses (Risch and Teng 1998). We have now followed up this work by developing a cheap accurate method for estimating and comparing allele frequencies of SNPs in DNA pools.

The method combines the genotyping specificity of allele-specific primer extension assay (e.g. Syvanen et al. 1993; Haff and Smirnov 1997; Syvanen 1999) with the quantitative accuracy of high performance liquid chromatography (HPLC). After the polymerase chain reacion (PCR), a primer is annealed immediately upstream from the polymorphism. In the presence of the appropriate dNTPs and ddNTPs, the primer is extended by one or more bases depending upon the allele sequence at the polymorphic site yielding two allele-specific products (Fig. 1A). After the extension reaction, the allele-specific products are separated by denaturing HPLC (DHPLC; Oefner and Underhill 1998) on a WAVE DNA Fragment Analysis System (Transgenomic, Santa Clara, Calif.) at an oven temperature of 70°C. Allele-specific extended primers are detected by UV absorbency (Fig. 1B) and the allele frequencies are calculated from the absorbencies of the peaks (in mV).

To demonstrate the utility of this approach, we report data on nine polymorphisms of interest to our group (Table 1). Our data show that this method should allow the accurate estimation of absolute allele frequencies in pooled samples of DNA and of the difference in allele frequency between different pooled DNA samples. This technique therefore offers an efficient and cheap method for genotyping SNPs in large case-control and family-based association samples.

B. Hoogendoorn · N. Norton · G. Kirov · N. Williams
M. L. Hamshere · G. Spurlock · J. Austin · M. K. Stephens
P. R. Buckland · M. J. Owen · M. C. O'Donovan (✉)
Department of Psychological Medicine,
University of Wales College of Medicine,
Heath Park, Cardiff, CF14 4XN, United Kingdom
e-mail: odonovanmc@cardiff.ac.uk,
Tel.: +44 1222 743242, Fax: +44 1222 747839

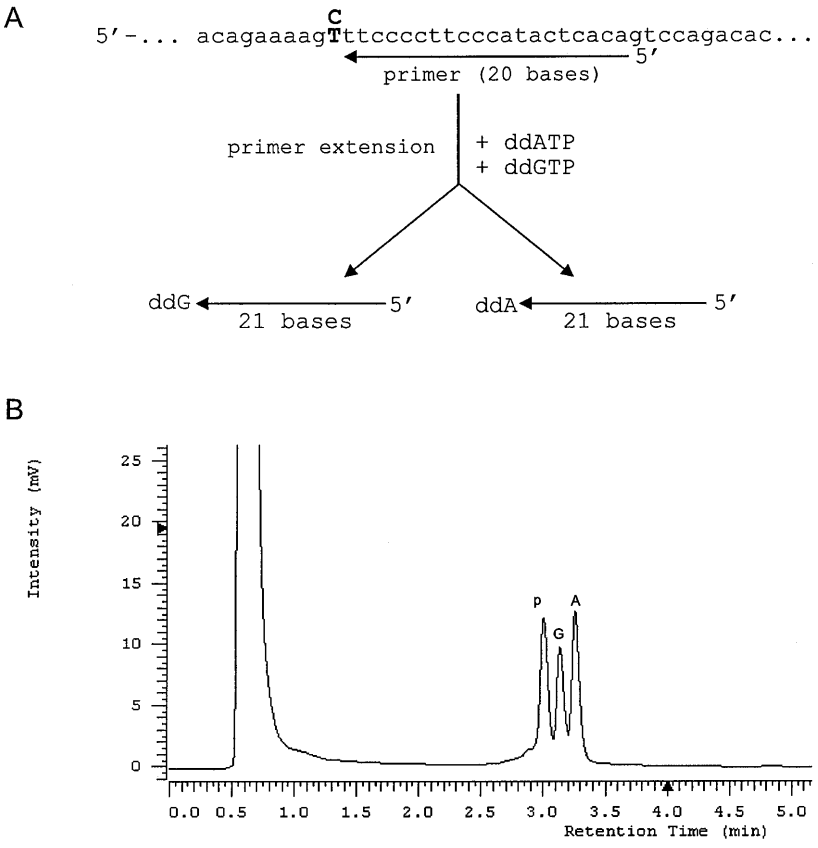

Fig. 1 **A** Schematic diagram showing the principle of primer extension. Depending upon which allele is present, the primer is extended by either ddG or ddA and can then be distinguished on the basis of sequence or size-dependent retention. **B** Chromatogram of the primer extension reaction shown in **A**, as performed on a heterozygous individual (*P* unextended primer, *G*, *A* alleles in order of elution)

Materials and methods

PCR procedure

PCR primer sequences are given in Table 2 (L and R). The four SNPs on chromosome 4p are anonymous SNPs. They are defined in Table 1 by their accession number, the position of the SNP, and the nature of the polymorphism. PCR was performed in 25 µl containing 20 pmol of each primer, 30–80 ng genomic DNA, 100 µM dNTPs and 0.5 U *Taq* DNA Polymerase (QIAGEN, Crawley, UK) in the buffer provided by the manufacturer. Amplification was performed in a PTC 225 DNA Tetrad thermal cycler (MJ Research, Genetic Research Instrumentation, Rayne, UK) under routine conditions. Detailed PCR conditions are available upon request.

Primer extension reactions

PCR fragments were prepared for primer extension by exonuclease I and shrimp alkaline phosphatase digestion as described in Hoogendoorn et al. (1999). Primer extension reactions were performed as described in Hoogendoorn et al. (1999) but the thermal cycler steps were shortened. An initial denaturation step of 2 min at 94°C was followed by 50 cycles of 94°C for 5 s, 43°C for 5 s and 60°C for 5 s. At the end of cycling, the reaction was heated to 94°C for 30 s and immediately placed on ice. The extension primers are shown in Table 2 (EXT). Nucleotide compositions and product lengths for the primer extension reactions are listed in Table 3, which also provides details about the nucleotides incorporated for each allele in the primer extension reaction (the polymorphic nucleotide is given in bold).

Construction of pools

The concentration of all DNA for pool construction was determined by using the PicoGreen dsDNA Quantitation Reagent (Molecular Probes, Eugene, Ore., USA) in a Labsystems Fluoroskan Ascent (LifeSciences International, Basingstoke, UK).

PCR product “pseudo pools” were prepared as follows. For a particular SNP, homozygous samples for each allele were amplified by PCR and the quantitated products were then mixed in different proportions to provide pools with different allele frequencies. Genomic DNA “pseudo pools” were prepared by mixing quantitated genomic DNA from two individuals, each of whom was homozygous for a different allele, in different proportions to provide pools with different allele frequencies.

True genomic DNA case-control pools were constructed from patients with schizophrenia (affecteds) and blood donor controls (for the number of subjects, see Table 1). Family-based association pools were also constructed from 111 probands with bipolar disorder I (affecteds) and their 222 parents, as detailed in a previous paper (Kirov at al. 2000). All subjects included in the pools were also individually genotyped by restriction fragment length polymorphism analysis of PCR products or by primer extension. A detailed protocol for each polymorphism is available upon request. The case-control pools were each analysed 16 times for each marker and the final allele frequency was determined as the mean of the values obtained, corrected as below, for differential representation of the alleles in a heterozygote. The family-based pools were each analysed six times and corrected as below.

HPLC analysis

HPLC was performed on a WAVE DNA Fragment Analysis System (Transgenomic) containing a DNASep column held at an oven temperature of 70°C. Primer extension products were eluted from the column by using a linear acetonitrile gradient in a 0.1 M tri-

Table 1 Allele frequency estimation by individual (*Real*) and pooled genotyping (Δ difference in allele frequencies between controls and cases, *SE* standard error of the mean). The correction factor (*k*= ratio of absorbencies for each allele in analysis of a heterozygote) for each marker is given in *column 1*

Polymorphism		Controls/parents (SE)	Affecteds (SE)	Δ
	n	157	160	
NTS (k=0.87, SE=0.02)	Pool	0.740 (0.003)	0.722 (0.003)	0.018
	Real	0.755	0.746	0.009
	n	157	160	
COMT (k=1.36, SE=0.005)	Pool	0.528 (0.006)	0.514 (0.003)	0.014
	Real	0.564	0.550	0.014
	n	189	180	
5HT2A (k=1.81, SE=0.06)	Pool	0.391 (0.016)	0.427 (0.014)	–0.036
	Real	0.390	0.420	–0.030
	n	130	146	
PRODH (k=0.87, SE=0.021)	Pool	0.792 (0.003)	0.788 (0.004)	0.004
	Real	0.800	0.812	–0.012
	n	130	146	
GRM7 (k=1.29, SE=0.052)	Pool	0.545 (0.004)	0.483 (0.004)	0.062
	Real	0.519	0.459	0.060
	n	222	111	
AC006230/62,541: A→G (k=1.41, SE=0.01)	Pool	0.843 (0.005)	0.845 (0.003)	–0.002
	Real	0.845	0.853	–0.008
	n	222	111	
AC006230/142,369: A→G (k=1.76, SE=0.03)	Pool	0.495 (0.003)	0.525 (0.002)	–0.030
	Real	0.495	0.514	–0.019
	n	222	111	
AC004169/30,419: G→T (k=2.05, SE=0.01)	Pool	0.959 (0.005)	0.959 (0.003)	0.000
	Real	0.973	0.972	0.001
	n	222	111	
AC004169/140,113: C→T (k=1.37, SE=0.04)	Pool	0.731 (0.004)	0.734 (0.004)	–0.003
	Real	0.736	0.750	–0.014

ethylamine acetate buffer (TEAA), pH 7.0, at a constant flow rate of 0.9 ml/min. The gradient was created by mixing eluents A and B. The composition of eluent A was 0.1 M TEAA (pH 7.0), 0.1 mM Na_4EDTA, and that of eluent B was 25% acetonitrile in 0.1 M TEAA (pH 7.0). Each analytical gradient was composed of 23%–36% eluent B and each run took 5.5 min. At the end of each analytical run, the column was washed by using the Transgenomic WAVEAccelerator System set to deliver 400 µl wash solution (50% acetonitrile in 0.1 M TEAA, pH 7.0). To allow reconditioning of the column, the eluent mix was kept at 18% eluent B for 60 s at the end of each wash. The eluted products were detected by using a UV detector set at 260 nm. Allele frequencies were estimated for each run from the UV absorbencies (in mV) of the allele-specific extended primers.

Correction for unequal allelic detection

In order to allow for unequal representation of alleles from known heterozygotes, estimated allele frequencies from pools were corrected by using the mean of the ratios obtained from eight analyses of a true heterozygote. The frequency *f* in the pool of allele A is then calculated as $f(a)=A/(A+kB)$ where A and B are the absorbencies of the primer extension products representing alleles A and B, respectively, and *k* is the mean of the A/B ratios observed in a heterozygote.

Polymorphisms

The polymorphisms chosen for analysis in the case-control pools were a C→G in the proneurotensin gene (NTS), an A→G in the serotonin 5HT2A receptor, an A→G in catechol O-methyltransferase (COMT), an A→G in the promoter region of proline oxidase (PRODH) and an A→T in the glutamate receptor, metabotropic 7 (GRM7). The polymorphisms analysed in the family-based association were detected in a series of bacterial artificial chromosome clones.

Results

The analysis of a heterozygote offers the perfect natural experiment for determining the ratio of the two allele-specific products corresponding to an allele frequency of 0.5. With an equal representation of alleles, this ratio is expected to have a value of 1. However, on no occasion was this observed, with the maximum deviation from expectation being observed for the AC004169/30,419: G→T polymorphism (Table 1). Possible reasons for the unequal representation of alleles are discussed later. In subsequent

Table 2 Oligonucleotide sequences

SNP	Sequence
NTS	(L) 5'-GATACTGGGGGTTCTTTGTC (R) 5'-GAGCAACTCTTCTCCCAGAT (EXT) 5'-GCAAAGATAATGTCTGTA
5HT2A	(L) 5'-AACCAACTTATTTCCTACCAC (R) 5'-AAGCTGCAAGGTAGCAACAGC (EXT) 5'-TGGCTTTGGATGGAAGTGCC
COMT	(L) 5'-TAGTAACAGACTGGGCACGAA (R) 5'-GTTCAAAGGGCATTTATCATG (EXT) 5'-TGTGAGTATGGGAAGGGGAA
PRODH	(L) 5'-TTAGCAAAGAGTCAAGCGCA (R) 5'-CACCAATACCTGTCAGTGGC (EXT) 5'-GAAAAAGAACTGTTATTGGAGTC
GRM7	(L) 5'-ATGAACAAGGATCTCTGTGC (R) 5'-TCCAGCTTGCTCCATCTCT (EXT) 5'-GAACAAGGATCTCTGTGCTGACT
AC006230/62,54: A→G	(L) 5'-AATGATATTCCAACCCAGAGG (R) 5'-ATGGTGCCATGGTTTGTCTG (EXT) 5'-ATGGTTTGTCTGTGTGTGCAT
AC004169/30,419: G→T	(L) 5'-TGCACCCACATGCATTTCAG (R) 5'-TAGCTCACAGTGCCTGCGG (EXT) 5'-CTCCATGGGTGCACAGACGG
AC004169/140,113: C→T	(L) 5'-ATCTGCTTGTTGAGCACCCT (R) 5'-CTGGCTCACTCTCCCAACTC (EXT) 5'-GGCACCTTTTCCAGGAAGCC
AC006230/142,369: A→G	(L) 5'-AACATGGCTTTAATGGAAGGG (R) 5'-GTGGTGACTTTGGGGAAAGAG (EXT) 5'-GCTGGCTGATCAGAAAAAGG

Table 3 Nucleotide composition and product length in order of elution for primer extension reactions. The final two columns give the size of the extended primer for each allele and the nucleotides incorporated into the primer for each allele in the primer extension reaction. The polymorphic nucleotide is given in *bold*

Polymorphism	Primer length (bp)	Nucleotide composition of primer extension reaction	SNP	Primer extension products per allele in basepairs. (bases added to primer)	
				Allele 1	Allele 2
NTS	18	ddCTP, ddGTP	C/G	C: 19 (**C**)	G: 19 (**G**)
5HT2A	20	ddGTP, dATP	A/G	G: 21 (**G**)	A: 22 (**A**G)
COMT	20	ddGTP, ddATP	A/G	G: 21 (**G**)	A: 21 (**A**)
PRODH	23	dA, ddT, ddG	A/G	G: 24 (**G**)	A: 25 (**AT**)
GRM7	23	ddA, dT, ddC	A/T	A: 24 (**A**)	T : 25 (**T**C)
AC006230/62,541: A→C	21	ddA, dGTP, dCTP, dTTP	A/G	A: 22 (**A**)	G: 24 (**GT**A)
AC006230/142,369A→G	20	ddA, dGTP, dCTP, dTTP	A/G	A: 21 (**A**)	G: 23 (**G**CA)
AC004169/30,419G→T	20	ddG, ddC, dTTP, dATP	G/T	G: 21 (**G**)	T: 24 (**T**AAC)
AC004169/140,113C→T	20	ddC, dGTP, dATP, dTTP	C/T	C: 21 (**C**)	T: 23 (**T**TC)

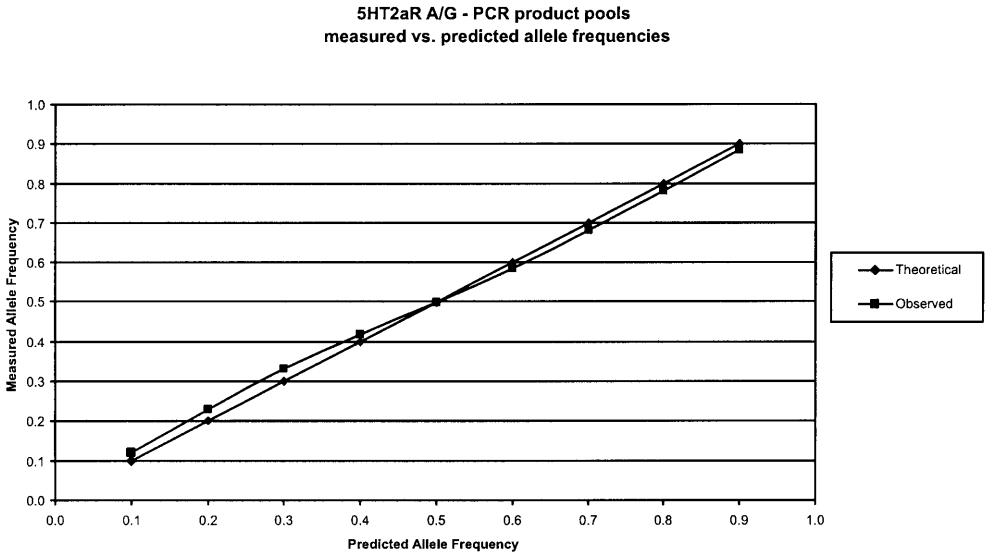

Fig. 2 Allele frequencies calculated from analysis of PCR product pools (*Observed*) for the 5HT2AR polymorphism. Each datapoint is the mean of nine replicates. Predicted allele frequencies are based upon the amount of PCR product from each homozygote used to construct the pool. The observed ratios were corrected by the ratio of the absorbencies obtained from the pool simulating an actual allele frequency of 0.5

analyses, we simply corrected the observed allele frequencies for differential representation as described above.

We then determined the quantitative relationship between the true allele frequencies and the allele frequencies measured by the assay. This was achieved by analysing PCR product "pseudo pools" constructed by pooling quantitated PCR products from two subjects who were homozygous for each of the different alleles. Pools were constructed with allele frequencies ranging from 0.1 to 0.9. Each pool was analysed nine times. The estimated allele frequencies were linearly related to the expected ratios. This is illustrated for the 5HT2A polymorphism in Fig. 2. Other polymorphisms gave similar results (data not shown).

We then tested the reproducibility of the assay across a range of allele frequencies by analysing 27 different genomic DNA "pseudo pools". These were constructed by mixing different amounts of genomic DNA from homozygotes to produce allele frequencies ranging from 0.4 to 0.6 for 5HT2A and NTS and from 0.3 to 0.7 for COMT. Each pool was analysed seven times. Across the range of allele frequencies tested, pooled analyses were highly reproducible. Thus, the coefficients of variation (standard deviation/mean) for the genomic DNA pools were 0.022 (NTS), 0.026 (5HT2A) and 0.031 (COMT).

Having established the basic properties of the assay, we tested its performance as a tool for genotyping in a series of true DNA pools. Since it appeared evident that large differences in allele frequency could be easily detected (Fig. 2), we constructed test pools that only differed slightly in their allele frequencies. The results are shown in Table 1. The estimates of absolute allele frequencies in the pools agreed well with the results from individual genotyping, with a mean error of 0.014 and a maximum error of 0.036 (COMT). More importantly for association studies, estimations of the differences in allele frequencies between pools (Δ) of cases and controls also agreed well with true differences (Table 1), with a mean error of 0.006 and a maximum error of 0.016 (PRODH).

Discussion

The objective of this research was to develop a cheap, accurate and rapid method for estimating SNP allele frequencies in association studies. First, in order to test the primer extension reaction per se, the effect of variation in PCR was removed by constructing PCR product "pseudo pools". Second, in order to assess the effect of inter-sample variation in PCR and primer extension efficiencies, genomic DNA "pseudo pools" were constructed. The data from our analyses of both sets of "pseudo pools" indicate that analysis of SNPs in pools by using primer extension and DHPLC is quantitative and highly reproducible.

In order to assess the method in a naturalistic setting, we then constructed sets of true genomic DNA pools. To be applicable in association studies, pooled analysis must yield an approximate estimate of the absolute allele frequency. We were initially concerned that this might be precluded by differential PCR amplification of alleles (Liu et al. 1997; Barnard et al. 1998), differential efficiencies of incorporation of the appropriate dNTPs/ddNTPs for each allele-specific reaction (Haff and Smirnov 1997) and minor differences in the absorbencies of the two allele-specific products, which differ slightly in size and 3' base composition. However, whereas we have observed non-equal representation of alleles in the analysis of heterozygotes, this readily yields a correction factor that allows a reasonable estimation of absolute allele frequencies in pools. It should also be noted that the estimation of differences between pools is also affected by the use of the correction factor. Although this is negligible when the difference between frequencies in pools is small or the correction factor is close to unity, we recommend that pooled analyses are corrected wherever possible.

The second critical factor in applying a pooling method to association studies is the accuracy of estimating differences in allele frequencies between pools of affecteds and controls. Our data show that when compared

with individual genotyping, primer extension and DHPLC analysis allow the accurate estimation of true allele frequency differences. Thus, from all data listed in Table 1, the mean error is 0.67%, with a range of 0% to 1.6%. This degree of accuracy is likely to be adequate in most cases and might actually rival that of individual genotyping (Mein et al. 2000) since there is less opportunity for human error.

By using the procedure described above, accurate allele frequency estimates were initially obtained for case-control sample pools based upon approximately 40 reactions (eight heterozygotes, and two pools each with 16 replicates). This represents economies of scale by a factor of 50 assuming that sample sizes of up to 1000 cases and 1000 controls are possible (Pacek et al. 1993). Further improvements in scale can then be achieved by reducing the number of replicated analyses for each pool (family-based sample) to six with no significant loss in accuracy.

Since this work was undertaken, another group have reported the analysis of SNPs in pools (Germer et al. 2000) based upon kinetic PCR and allele-specific PCR amplification, with similar results. Thus, pooled analysis of SNPs based upon two different platforms confirms that SNP analysis in DNA pools is a viable procedure for large-sample association studies. Since both methods essentially depend upon quantitative PCR, we envisage that other methods allowing the quantitative analysis of allele-specific PCR products will also be applicable to pooled SNP analysis. However, in contrast to many of these, our protocol does not require expensive reagents. Furthermore, a single set of primer extension cycling conditions is applicable to virtually all SNPs and therefore the method does not require intricate optimization steps.

We conclude therefore that primer extension and DHPLC analysis is precise and economical enough to permit high throughput screens of candidate genes in case-control and family-based association study samples. We also envisage that it should be possible to increase the throughput of the primer extension method to permit pooled analysis of tens of thousands of SNPs in realistic time frames by introducing fluorescence-based multi-channel HPLC detection systems or capillary electrophoresis systems.

Acknowledgements This work was supported by the MRC (UK) and the Wellcome Trust. G.K. is a Wellcome Trust Advanced Fellow.

References

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

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, Owen MJ, Oefner PJ, Williams N, Austin J, O'Donovan MC (1999) Genotyping single nucleotide polymorphisms by primer extension and high performance liquid chromatography. Hum Genet 104:89–93

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

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 MJ, Siegmund T, Smith AN, Esposito L, Nutland S, Stevens HE, Wilson AJ, Phillips MS, Jarvis N, Law S, Arruda M de, Todd JA (2000) Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. Genome Res 10:330–343

Oefner PJ, Underhill PA (1998) DNA mutation detection using denaturing high-performance liquid chromatography (DHPLC). In: Dracopoli NC, Haines JL, Korf BR, Moir DT, Morton CC, Seidman CE (eds) Current protocols in human genetics (suppl. 19). Wiley, New York, pp 7.10.1–7.10.12

Pacek P, Sajantila A, Syvanen A-C (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

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

Syvanen A (1999) From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. Hum Mut 13:1–10

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