# Determination of allele frequency in pooled DNA: comparison of three PCR-based methods

Stefan Wilkening[1], Kari Hemminki[1,2], Ranjit Kumar Thirumaran[1], Justo Lorenzo Bermejo[1], Stefan Bonn[3], Asta Försti[1,2], and Rajiv Kumar[1,2]

[1]German Cancer Research Center (DKFZ), Heidelberg, Germany, [2]Karolinska Institute, Huddinge, Sweden, and [3]Max Planck Institute for Medical Research, Heidelberg, Germany

*BioTechniques 39:853-858 (December 2005)*
doi 10.2144/000112027

*Determination of allele frequency in pooled DNA samples is a powerful and efficient tool for large-scale association studies. In this study, we tested and compared three PCR-based methods for accuracy, reproducibility, cost, and convenience. The methods compared were: (*i*) real-time PCR with allele-specific primers, (*ii*) real-time PCR with allele-specific TaqMan® probes, and (*iii*) quantitative sequencing. Allele frequencies of three single nucleotide polymorphisms in three different genes were estimated from pooled DNA. The pools were made of genomic DNA samples from 96 cases with basal cell carcinoma of the skin and 96 healthy controls with known genotypes. In this study, the allele frequency estimation made by real-time PCR with allele-specific primers had the smallest median deviation (MD) from the real allele frequency with 1.12% (absolute percentage points) and was also the cheapest method. However, this method required the most time for optimization and showed the highest variation between replicates (SD = 6.47%). Quantitative sequencing, the simplest method, was found to have intermediate accuracies (MD = 1.44%, SD = 4.2%). Real-time PCR with TaqMan probes, a convenient but very expensive method, had an MD of 1.47% and the lowest variation between replicates (SD = 3.18%).*

## INTRODUCTION

Most of the genetic contribution to complex diseases is thought to be conferred by multiple genes, each with small effects (1–3). To find these effects, large sample sizes are required. An efficient way for reducing the costs, labor time, and DNA consumption of such studies is to combine the DNA samples into pools and to determine the allele frequency in these pools (4,5). Single nucleotide polymorphisms (SNPs) are the most common type of polymorphism in the human genome, and they are relatively easy to genotype. Therefore, they are widely used as markers in association studies. In principle, any method that is able to distinguish between SNP genotypes AA, AB, and BB in a single individual can also be used to estimate the ratio between allele A and allele B in pooled DNA. Most of these methods are PCR-based. PCR is either used to initially amplify the sequence that contains the polymorphism before analyzing it or as a method to directly distinguish between variants by using allele-specific primers or probes (5).

In this study, we compared three PCR-based methods for the estimation of SNP frequencies in pooled DNA. Real-time PCR, with its high accuracy to quantify a specific DNA fragment from a broad range of starting concentrations, seemed to us a suitable method for this approach. The use of real-time PCR with allele-specific primers is the first of the tested methods. Its use for the determination of SNP allele frequencies in pooled DNA was first described by Germer et al. (6), followed by further studies that successfully applied this method (7–10). The second tested method, which is widely used for individual genotyping, is TaqMan® PCR, which discriminates between alleles by using allele-specific probes (11). This assay has previously been reported to be used for pooled samples in an end-point measurement (12) and under real-time PCR conditions (13). The third method tested in this study is quantitative sequencing, which to our knowledge has not been previously used to estimate SNP frequencies in a pooled case-control study. However, direct sequencing has previously been used to estimate the mutation frequency in pooled cDNA (14) or to detect mutations in pooled DNA (15).

## MATERIALS AND METHODS

### DNA Pool Construction

Two pools were set up by pooling 96 DNA samples from the blood of basal cell carcinoma patients and 96 samples from a healthy control group, respectively. Sample collection was approved by local ethical boards. Genomic DNA was isolated from blood samples using a QIAamp® DNA Blood Midi Kit (Qiagen, Valencia, CA, USA). DNA concentrations were measured using the PicoGreen® double-stranded DNA (dsDNA) Quantification Reagent (Invitrogen, Carlsbad, CA, USA) and the GENios® Microplate Reader (Tecan Systems, San Jose, CA, USA). Twenty nanograms of each sample were added to the pools, and the pool volumes were adjusted with water to 150 µL. To verify equal DNA concentrations, PicoGreen measurement was repeated with the pools, and minor adjustments were made. Standard real-time PCR was done with both pools in triplicate with SYBR® Green I (Invitrogen) as the fluorescent dye to confirm that both pools perform identically under PCR conditions.

### Individual SNP Genotyping

The following three SNPs were genotyped individually from 192 DNA samples: rs2066827 (TG), rs861539 (CT), and rs25487 (GA). Genotyping was done with customized TaqMan genotyping assays (Applied Biosystems, Foster City, CA, USA). For TaqMan PCR, 5 ng of genomic DNA were analyzed in a total volume of 5 µL with an ABI PRISM® 7900 Sequence Detection System (Applied Biosystems). To verify the TaqMan results, 10% of the samples were sequenced in an automated ABI PRISM 3100 Genetic Analyzer (Applied Biosystems). For the primer and probe sequences, see the Supplementary Material available online at www.BioTechniques.com. The SNPs were chosen because of their location in three genes relevant in skin carcinogenesis: *CDKN1B* (also known as *P27*), *XRCC3*, and *XRCC1*.

## Real-Time PCR with Allele-Specific Primers

The 3′ end of either the forward or the reverse primer was located directly over the SNP. To increase the PCR specificity, an extra mismatch was placed on the third or fourth base from the 3′ end of the primer. Primers were designed using the online program Primer3 (frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). See the Supplementary Material for the primer and probe sequences. Real-time PCR was performed with 10 ng DNA as a template, in a total volume of 10 µL, containing 5% dimethyl sulfoxide (DMSO), 2 mM $MgCl_2$, 30 mM KCl, 2.5% glycerol, 0.2 mM dNTPs, 0.2× SYBR Green I, 0.5× ROX reference dye (Invitrogen), 0.3 U AmpliTaq® Stoffel Fragment DNA Polymerase plus 1× buffer (Applied Biosystems), and 0.15 µM primers. For convenience, all PCR components except DNA, water, polymerase, and primers were stored as a premix at -20°C. Allele specificity and annealing temperatures were determined with a set of samples known to be homozygous for one or the other allele, respectively. For analysis, both pools were run in quadruplicate together with 10-fold dilutions of two different heterozygous samples. Each reaction was carried out separately with one of the two allele-specific primers. The allele frequencies of the pools were calculated according to the formula (6):

$$\text{frequency of the allele-A} = 1/(E^{\Delta C_t} + 1),$$

where $\Delta C_t = (A_{sample} - B_{sample}) - (A_{heterozygote} - B_{heterozygote})$. “A” and “B” stand for the cycle threshold number of the allele-specific amplification curves. “E” is the PCR efficiency, which can be deduced by the slope of the standard curve according to the equation (16):

$$E = 10^{[-1/slope]}$$

## Real-Time PCR with Allele-Specific TaqMan Probes

PCR was performed with 5 ng of DNA in a total volume of 10 µL using the same primers and probes that were used for individual genotyping. The pooled DNA was analyzed in quadruplicate. Additionally, homozygous samples for the two alleles were mixed in 9 different ratios (1:9, 2:8, … 9:1) and analyzed. These ratios were plotted as the logarithm against the cycle distance between allele-A and allele-B (Figure 1). The function of the resulting restriction graph was then used to calculate the allele frequencies in the pools. To deduce PCR efficiencies, 10-fold dilutions of individual samples from the three genotypes (AA, AB, and BB) were run in parallel.

## Quantitative Sequencing

The region around the SNP of interest was amplified by 35 cycles of PCR, taking 5 ng of DNA in a total volume of 10 µL, using the same primers previously used for the verification of individual genotyping. Sequencing reactions were performed using a BigDye™ Terminator Cycle Sequencing Kit (Applied Biosystems) in a 10 µL volume containing pretreated PCR product [30 min at 37°C and 15 min at 85°C with 0.75 µL of ExoSAP-IT™ (Amersham Biosciences, Piscataway, NJ, USA)] and sequencing primer under the following PCR conditions: 96°C for 2 min prior to 27 cycles of 96°C for 30 s, 54°C for 10 s, and 60°C for 4 min. Sequencing products were precipitated with isopropanol, washed with 70% ethanol, resuspended in 25 µL of water, and finally loaded onto an ABI PRISM 3100 Genetic Analyzer. The DNA of the two pools was analyzed in quadruplicate together with two different heterozygous samples in triplicate. Additionally, homozygous samples for the two alleles were mixed in 9 different ratios (1:9, 2:8, … 9:1). At the position of the SNP, the relative peak areas were determined from the electro-

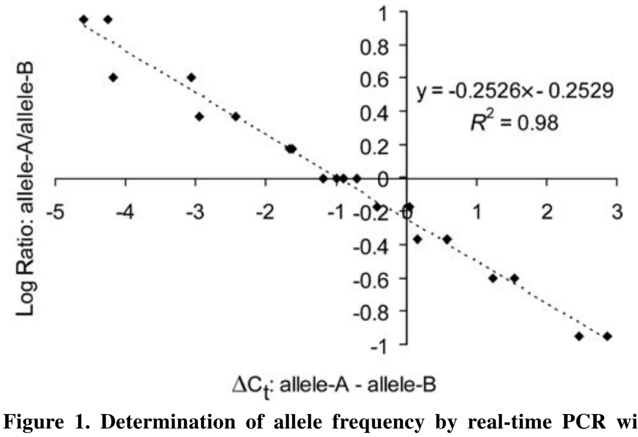

**Figure 1. Determination of allele frequency by real-time PCR with TaqMan probes.** To create a standard curve, samples homozygous for both alleles of the *CDKN1B*-SNP were mixed in 9 different ratios (1:9, 2:8, … 9:1). The ratios of allele frequencies are plotted logarithmically against the cycle distance between the amplification curves of allele-A and allele-B ($\Delta C_t$). SNP, single nucleotide polymorphism; $C_t$, cycle threshold.

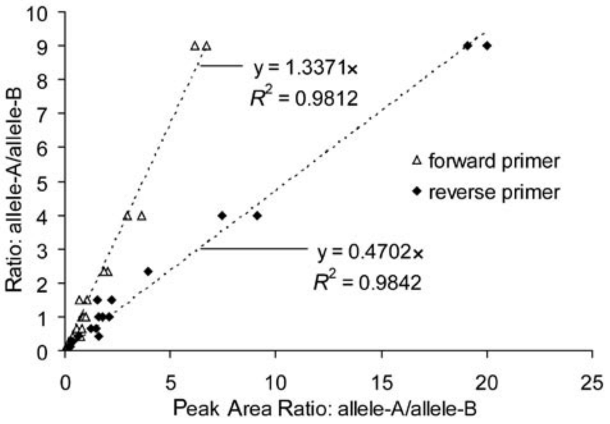

**Figure 2. Determination of allele frequency by quantitative sequencing.** To create a standard curve, samples homozygous for both alleles of the *CDKN1B*-SNP were mixed in 9 different ratios (1:9, 2:8, … 9:1). The ratios of allele frequencies are plotted against the ratios of the peak areas from allele-A and allele-B. SNP, single nucleotide polymorphism.

Table 1. Comparison of Three Methods to Detect the Allele Frequency in Pooled DNA

| Allele (Pool) | Expected | Allele-Specific Primers | TaqMan Probe | Quantitative Sequencing |
|---|---|---|---|---|
| XRCC1-A (cases) | 40.53 | 40.0 (±6.9) | 42.0 (±12.7) | 37.9 (±5.1) |
| XRCC1-A (controls) | 36.84 | 36.9 (±4.1) | 38.4 (±7.0) | 36.7 (±3.9) |
| XRCC3-T (cases) | 27.37 | 24.5 (±9.9) | 25.9 (±1.1) | 25.9 (±2.9) |
| XRCC3-T (controls) | 42.63 | 42.7 (±6.0) | 41.1 (±1.2) | 41.2 (±4.9) |
| CDKN1B-G (cases) | 20.53 | 23.7 (±9.5) | 20.2 (±2.5) | 15.2 (±4.5) |
| CDKN1B-G (controls) | 22.63 | 24.4 (±4.4) | 23.7 (±3.9) | 23.3 (±3.2) |
| $MD_{raf}$ (%)[a] | N.A. | 1.12 | 1.47 | 1.44 |
| $MD_{\Delta}$ (%)[b] | N.A. | 1.43 | 0.12 | 2.53 |
| Median SD[c] | N.A. | 6.47 | 3.18 | 4.20 |

Allele frequencies (%) were estimated for 3 SNPs in 2 pools of 96 samples. N.A., not applicable; SNP, single nucleotide polymorphism.
[a] Median deviation between the 6 expected and estimated allele frequencies.
[b] Median deviation between the 3 expected and estimated differences between cases and controls.
[c] Median of 6 standard deviations (2 pools, 3 SNPs).

pherograms using phred software (17). The ratios of the relative peak areas were plotted against the ratios of the allele frequencies. The function of the resulting regression graph (with a crossing point at the origin) was then used to calculate the allele frequency in the pools (Figure 2). Just as in the case of real-time PCR with allele-specific primers, we found that it was sufficient to use only the heterozygous samples for the correction of unequal allelic signals. This correction, known as k-correction, has been previously used in other pool studies (18–20).

### Statistical Methods

Allele frequency estimates were calculated as the median of four replicates, respectively, and standard deviations for each estimate are given as ± values. The median deviation (MD) between the expected and the observed allele frequencies was calculated for each method as the median of the 6 estimates (2 pools with 3 SNPs each). The median standard deviation was calculated for each method from 6 standard deviations (2 pools with 3 SNPs each). The median deviation between expected and observed differences between the allele frequency of cases and controls ($MD_{\Delta}$) was calculated as the median of the case-control distances of 3 SNPs for each method. All values are given as absolute percentage points. On the basis of 400 cases and 400 controls, we calculated the minimum distance between cases and controls that is required to give a significant *P* value (<0.05) in the $\chi^2$ test. To determine if the accuracies depend on the allele frequencies, allele frequencies were estimated for samples with different allele ratios (1:9, 2:8, … 9:1), and the relative errors were determined for minor allele frequencies of 10%, 20%, 30%, 40%, and 50%.

## RESULTS AND DISCUSSION

All three methods were 100% specific; that is, in samples homozygous for one allele, no signals of the other allele were detectable. Table 1 shows the expected and the observed allele frequencies for three SNPs in the two pools as well as median and standard deviations for each method. The expected frequencies were obtained from individual genotyping. The observed frequencies were estimated by three different methods.

### Real-Time PCR with Allele-Specific Primers

The most accurate estimation (MD = 1.12%) was obtained with real-time PCR using allele-specific primers, although it had the highest variation (SD = 6.47%). For the optimization of the assay, we tested primers with and without an extra mismatch. For all three SNPs, we obtained higher specificity when using primers with an extra mismatch (data not shown) as has been previously described (21,22). The alternatives for primer design are limited for this method because the 3′ end of the allele-specific primer has to be located directly on the SNP either on the plus or the minus strand of the DNA. Compared with the other methods, primer design and PCR optimization are more time-consuming and each pool has to be examined in two reactions (allele-A-specific PCR and allele-B-specific PCR). Germer et al. (6) set the PCR efficiency "E" = 2, which refers to 100% efficiency. However, we found the actual PCR efficiency (mean efficiency from allele-A-specific PCR and allele-B-specific PCR) to give more accurate results.

### Real-Time PCR with Allele-Specific TaqMan Probes

Real-time PCR with TaqMan probes showed the highest deviation from the expected allele frequencies (MD = 1.47%). However, when only the allele frequency differences between cases and controls were taken into consideration, this method gave the best estimates ($MD_{\Delta}$ = 0.12%) and it also had the best reproducibility (SD = 3.18%). The TaqMan assay has the advantage that it is a ready-to-use technique and both alleles can be analyzed in one tube. However, when both reactions take place in the same tube, there might be competing interactions. For the *XRCC1*-SNP, the calculated PCR efficiency was found to be much lower when both alleles were present in the reaction (45% efficiency) compared with a reaction with only homozygous samples (94% efficiency). In addition, we saw that amplification curves from 1 ng of one allele mixed with 9 ng of the other allele arose about 3 cycles later than 1 ng of the first allele alone. This suggests that in a mixture of both alleles, the reaction of one allele suppresses the reaction of the other allele. This can lead to a less efficient PCR and worse quantification accuracy. Furthermore, our study showed that for reliable results, it was important to run not only heterozygous samples but also different ratios of homozygous samples. This leads to a higher sample number and makes the experimental procedure more inconvenient. It will also be difficult to obtain rare homozygotes when the allele frequency of the examined SNP is low.

## Quantitative Sequencing

Although quantitative sequencing is not based on real-time PCR, it turned out to be comparably precise for the assayed SNPs. To optimize the fidelity of the presequencing PCR, we lowered the cycle number to 27 cycles (with 5 ng DNA as starting template) to keep PCR in the exponential phase. However, the accuracy was not improved (MD = 1.50%). We then took the relative peak area as a parameter for estimation. Taking the peak height instead of the relative peak area as a parameter, we obtained very similar results with a slightly worse estimation (MD = 1.66%) and the same variance. The procedure of quantitative sequencing (initial PCR plus sequencing reaction) itself takes more time compared with the other methods. However, the design of the primers is simple and flexible, and optimization of the assay is normally limited to the adaptation of the annealing temperature. Another advantage of this method is its possibility to determine various proximate SNPs at a time. For sequencing, both alleles are initially amplified in the same tube. As seen for the XRCC1 TaqMan assay, this may lead to a competition between the reactions of the two alleles and a resulting detection bias.

## Conclusions

Our comparison of the three methods showed that each of the methods had acceptable median deviations from the expected allele frequency (MD <1.5%). However, standard deviations varied between 3.2% and 6.5%. To show the impact of the standard deviation in a case-control study, we calculated the minimum difference between the allele frequencies of cases and controls that is required for a significant association of an SNP with a disease (Figure 3). This calculation was done for a given sample size of 400 cases and 400 controls. Including a standard deviation of 4.2% (median of the three methods), the minimum required distance between cases and controls increases by 4.2%. From the examined SNPs, only the allele frequencies of XRCC3 were significantly different between cases and controls. The estimated minor allele frequency for this SNP was around 25%; accordingly, the minimum required difference between cases and controls would be 11%. For this SNP, all three methods showed a difference between the estimates of cases and controls of more than 15%, and the SNP would therefore be considered to be significantly associated with the disease. However, with real-time PCR with allele-specific primers, the standard deviations were in a range where a false-positive association could not be excluded. To determine if the accuracy of the three methods depends on the allele frequency, we estimated the allele frequency in samples with known ratios of the two alleles (1:9, 2:8, … 9:1). In the range from 10% to 50% minor allele frequency, no significant differences in the accuracy could be found in any of the methods.

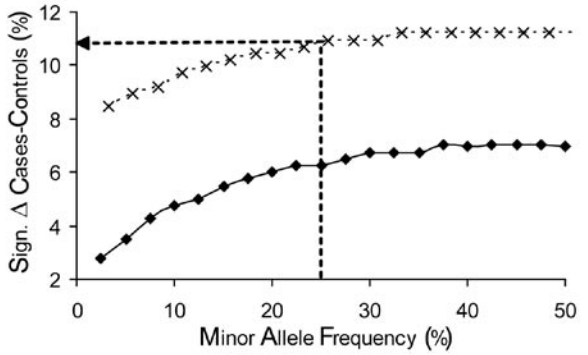

**Figure 3. Minimum significant allele frequency difference between cases and controls.** The minimum allele frequency difference between cases and controls required for a significant association ($P < 0.05$) is shown for a given sample size of 400 cases and 400 controls (continuous line). Considering a standard deviation of 4.2%, the threshold moves accordingly up and to the left (dotted line). With an estimated minor allele frequency of 25%, the required difference between cases and controls would then be 11% (dotted arrow).

The cost for the analysis of one SNP is around 350 € for the TaqMan method (primers plus probes, 300 €; master mix, 40 €; 96-well plate plus tips, 10 €), approximately 38 € for real-time PCR with allele-specific primers (3 primers, 18 €; master mix, 10 €; 96-well plate plus tips, 10 €), and approximately 117 € for quantitative sequencing (2 primers, 12 €; master mix, 10 €; 1.5× 96-well plate plus tips, 15 €; sequencing reagents, 80 €). Compared with TaqMan PCR, real-time PCR with allele-specific primers and quantitative sequencing is much cheaper and therefore might be preferred when analyzing many SNPs on the same DNA pools. However, both methods had a relatively high variation between replicates. Therefore, we would recommend performing more than four replicates using these methods. The choice of the method clearly depends on the laboratory equipment (real-time PCR, sequencer) but also on the DNA sequence around the examined SNP because some sequences may prohibit the use of one or the other method. A good initial approach might also be to analyze pools with two independent methods.

## COMPETING INTERESTS STATEMENT

*The authors declare no competing interests.*

## REFERENCES

1. **Risch, N. and K. Merikangas.** 1996. The future of genetic studies of complex human diseases. Science *273*:1516-1517.
2. **Houlston, R.S. and J. Peto.** 2004. The search for low-penetrance cancer susceptibility alleles. Oncogene *23*:6471-6476.
3. **Antoniou, A.C., P.D. Pharoah, G. McMullan, N.E. Day, B.A. Ponder, and D. Easton.** 2001. Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population-based study. Genet. Epidemiol. *21*:1-18.
4. **Norton, N., N.M. Williams, H.J. Williams, G. Spurlock, G. Kirov, D.W. Morris, B. Hoogendoorn, M.J. Owen, and M.C. O'Donovan.** 2002. Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. Hum. Genet. *110*:471-478.
5. **Sham, P., J.S. Bader, I. Craig, M. O'Donovan, and M. Owen.** 2002. DNA pooling: a tool for large-scale association studies. Nat. Rev. Genet. *3*:862-871.
6. **Germer, S., M.J. Holland, and R. Higuchi.** 2000. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. Genome Res. *10*:258-266.
7. **Shi, M., D. Caprau, J. Dagle, L. Christiansen, K. Christensen, and J.C. Murray.** 2004. Application of kinetic polymerase chain reaction and molecular beacon assays to pooled analyses and high-throughput genotyping for candidate genes. Birth Defects Res. A Clin. Mol. Teratol. *70*:65-74.
8. **Chen, J., S. Germer, R. Higuchi, G. Berkowitz, J. Godbold, and J.G. Wetmur.** 2002. Kinetic polymerase chain reaction on pooled DNA: a high-throughput, high-efficiency alternative in genetic epidemiological studies. Cancer Epidemiol. Biomarkers Prev. *11*:131-136.
9. **Dasgupta, R.K., P.J. Adamson, F.E. Davies, S. Rollinson, P.L. Roddam, A.J. Ashcroft, A.M. Dring, J.A. Fenton, et al.** 2003. Polymorphic variation in GSTP1 modulates outcome following therapy for multiple myeloma. Blood *102*:2345-2350.
10. **Rollinson, S., J.M. Allan, G.R. Law, P.L. Roddam, M.T. Smith, C. Skibola, A.G. Smith, M.S. Forrest, et al.** 2004. High-throughput association testing on DNA pools to identify genetic variants that confer susceptibility to acute myeloid leukemia. Cancer Epidemiol. Biomarkers Prev. *13*:795-800.
11. **Livak, K.J.** 1999. Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. Genet. Anal. *14*:143-149.
12. **Breen, G., D. Harold, S. Ralston, D. Shaw, and D. St. Clair.** 2000. Determining SNP allele frequencies in DNA pools. BioTechniques *28*:464-470.
13. **Xu, K., R.H. Lipsky, W. Mangal, E. Ferro, and D. Goldman.** 2002. Single-nucleotide polymorphism allele frequencies determined by quantitative kinetic assay of pooled DNA. Clin. Chem. *48*:1605-1608.
14. **Higuchi, M., S. Maas, F.N. Single, J. Hartner, A. Rozov, N. Burnashev, D. Feldmeyer, R. Sprengel, and P.H. Seeburg.** 2000. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature *406*:78-81.
15. **Amos, C.I., M.L. Frazier, and W. Wang.** 2000. DNA pooling in mutation detection with reference to sequence analysis. Am. J. Hum. Genet. *66*:1689-1692.
16. **Rasmussen, R.** 2001. Quantification on the LightCycler, p. 21-34. *In* K. Nakagawara (Ed.), Rapid Cycle Real-Time PCR, Methods and Applications. Springer Press, Heidelberg.
17. **Ewing, B., L. Hillier, M.C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*:175-185.
18. **Hoogendoorn, B., N. Norton, G. Kirov, N. Williams, M.L. Hamshere, G. Spurlock, J. Austin, M.K. Stephens, et al.** 2000. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. Hum. Genet. *107*:488-493.
19. **Le Hellard, S., S.J. Ballereau, P.M. Visscher, H.S. Torrance, J. Pinson, S.W. Morris, M.L. Thomson, C.A. Semple, et al.** 2002. SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. Nucleic Acids Res. *30*:e74.
20. **Simpson, C.L., J. Knight, L.M. Butcher, V.K. Hansen, E. Meaburn, L.C. Schalkwyk, I.W. Craig, J.F. Powell, et al.** 2005. A central resource for accurate allele frequency estimation from pooled DNA genotyped on DNA microarrays. Nucleic Acids Res. *33*:e25.
21. **Ishikawa, Y., K. Tokunaga, K. Kashiwase, T. Akaza, K. Tadokoro, and T. Juji.** 1995. Sequence-based typing of HLA-A2 alleles using a primer with an extra base mismatch. Hum. Immunol. *42*:315-318.
22. **Zhou, G., M. Kamahori, K. Okano, G. Chuan, K. Harada, and H. Kambara.** 2001. Quantitative detection of single nucleotide polymorphisms for a pooled sample by a bioluminometric assay coupled with modified primer extension reactions (BAMPER). Nucleic Acids Res. *29*:E93.

Received 21 June 2005; accepted 2 August 2005.

*Address correspondence to Stefan Wilkening, German Cancer Research Center (DKFZ), Molecular Genetic Epidemiology, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. e-mail: stefan_wilkening@web.de*