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

Stefan Wilkening1, Kari Hemminksi1,2, Ranjit Kumar Thirumaran1, Justo Lorenzo Bermejo1, Stefan Bonn3, Asta Försö1,2, and Rajiv Kumar1,2

1German Cancer Research Center (DKFZ), Heidelberg, Germany, 2Karolinska Institute, Huddinge, Sweden, and 3Max 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 PTC® 7900 Sequence Detection System (Applied Biosystems). To verify the TaqMan results, 10% of the samples were
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):

\[
\text{frequency of the allele-A} = \frac{1}{(E + 1)},
\]

where \( \Delta C_t = (A_{\text{sample}} - B_{\text{sample}}) - (A_{\text{heterozygote}} - B_{\text{heterozygote}}) \) 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):

\[
E = 10^{(-1/\text{slope})}
\]

Real-Time PCR with Allele-Specific TaqMan Probes

PCR was performed with 5 ng of DNA in a total volume of 10 \( \mu \text{L} \) 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 \( \mu \text{L} \), 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 \( \mu \text{L} \) volume containing pretreated PCR product [30 min at 37°C and 15 min at 85°C with 0.75 \( \mu \text{L} \) 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 \( \mu \text{L} \) 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-
Table 1. Comparison of Three Methods to Detect the Allele Frequency in Pooled DNA

<table>
<thead>
<tr>
<th>Allele (Pool)</th>
<th>Expected</th>
<th>Allele-Specific Primers</th>
<th>TaqMan Probe</th>
<th>Quantitative Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>XRCC1-A (cases)</td>
<td>40.53</td>
<td>40.0 (±6.9)</td>
<td>42.0 (±12.7)</td>
<td>37.9 (±5.1)</td>
</tr>
<tr>
<td>XRCC1-A (controls)</td>
<td>36.84</td>
<td>36.9 (±4.1)</td>
<td>38.4 (±7.0)</td>
<td>36.7 (±3.9)</td>
</tr>
<tr>
<td>XRCC3-T (cases)</td>
<td>27.37</td>
<td>24.5 (±9.9)</td>
<td>25.9 (±1.1)</td>
<td>25.9 (±2.9)</td>
</tr>
<tr>
<td>XRCC3-T (controls)</td>
<td>42.63</td>
<td>42.7 (±6.0)</td>
<td>41.1 (±1.2)</td>
<td>41.2 (±4.9)</td>
</tr>
<tr>
<td>CDKN1B-G (cases)</td>
<td>20.53</td>
<td>23.7 (±9.5)</td>
<td>20.2 (±2.5)</td>
<td>15.2 (±4.5)</td>
</tr>
<tr>
<td>CDKN1B-G (controls)</td>
<td>22.63</td>
<td>24.4 (±4.4)</td>
<td>23.7 (±3.9)</td>
<td>23.3 (±3.2)</td>
</tr>
</tbody>
</table>

| MD_{trf} (%) | N.A. | 1.12 | 1.47 | 1.44 |
| MD_{t} (%) | N.A. | 1.43 | 0.12 | 2.53 |
| Median Std | 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.

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_{t} = 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
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 3.2%) as a parameter, we estimated that for a SNP with a disease (Figure 3), the required difference between cases and controls would be 11% (dotted arrow).

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).
REFERENCES


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

To purchase reprints of this article, contact Reprints@BioTechniques.com