5-1-2002

Assessing allele frequencies of single nucleotide polymorphisms in DNA pools by pyrosequencing technology

Jon Wasson

Gary Skolnick

Latisha Love-Gregory

M. Alan Permutt

DRUG DISCOVERY
AND GENOMIC TECHNOLOGIES

Short Technical Report

Assessing Allele Frequencies of Single Nucleotide Polymorphisms in DNA Pools by Pyrosequencing™ Technology

BioTechniques 32:1144-1152 (May 2002)

Jon Wasson, Gary Skolnick, Latisha Love-Gregory, and M. Alan Permutt
Washington University School of Medicine, St. Louis, MO, USA

ABSTRACT

*Single nucleotide polymorphism (SNP) association studies searching for differences in allele frequencies between cases and controls have been widely used for genetic analysis. Individual genotyping is prohibitively expensive in large sample sizes. Pooling of samples provides the obvious advantage of higher throughput and lower cost. Here we report our results with the analysis of SNP allele frequencies in DNA pools using Pyrosequencing™ technology. For seven different SNPs, we observed a mean difference of 1.1 ± 0.6% between allele frequencies determined in two different DNA pools (*n *= 150 cases and 150 controls) compared to individually genotyped samples.*

INTRODUCTION

Linkage analysis in families with multiple affected individuals has become a standard method for discovering Mendelian disease genes (7). However, for most complex genetic diseases, a single gene is neither necessary nor sufficient for the disease etiology, and the standard positional cloning methods are inadequate to identify the genes. Linkage disequilibrium mapping in unrelated cases and controls has recently been proposed for this endeavor (5). Because the distance over which disequilibrium extends between markers and disease loci are never known, nor the degree of genetic risk contributed by any particular locus, one is tempted to genotype closely spaced markers in as many cases and controls as can be identified. Single nucleotide polymorphisms (SNPs), while biallelic, have been preferred over simple sequence repeat polymorphisms for this type of analysis, as SNPs are more abundant in the genome (9). Thus, many have sought rapid and cost-efficient methods for SNP genotyping. Pooling of DNA is a means of quickly finding regions of linkage disequilibrium with disease loci, thus requiring the accurate determination of allele frequencies in DNA pools (2). We now report that Pyrosequencing™ (Pyrosequencing AB, Uppsala, Sweden), a real-time sequencing method that employs an enzyme cascade system to monitor the release of inorganic pyrophosphate during nucleotide incorporation, is suitable for the genotyping of DNA pools. As each nucleotide is incorporated, the pyrophosphate released is quantitated by a luciferase reaction that results in a peak that is represented on a Pyrogram™ (Pyrosequencing AB). The PSQ96™ Pyrosequencer and the accompanying allele quantification software (Pyrosequencing AB) enabled us to determine rapidly and accurately the SNP allele frequencies in DNA pools containing 150 individuals. This method does not preclude larger samples.

MATERIALS AND METHODS

Patient Population

Cases (type 2 diabetes mellitus) and controls were of Ashkenazi Jewish descent as described previously (8).

DNA Isolation and Quantification

The DNA samples were isolated from whole blood using the Puregene™ kit as described (Gentra Systems, Minneapolis, MN, USA). DNA was quantified with the TKO 100 Mini-Fluorometer and Hoechst dye method as described (Hoefer Scientific Instruments, San Francisco, CA, USA). For the purposes of creating DNA pools, efforts to determine DNA concentrations accurately for each sample are critical, as errors will skew the proportion of each genotype in the pool. Spectrophotometric analysis was avoided because substances such as protein and salts may give spurious results (13). On the other hand, the larger the number of samples in the pool, the less important the individual quantifications become, as random errors in individual samples

tend to be minimized in large samples. Individual working samples were diluted with high-purity water to 10 ng/μL in sterile 1.2-mL polypropylene tubes, arranged in 96-well format racks, and stored at 4°C in a dedicated refrigerator separate from PCR products.

Construction of DNA Pools

The DNA samples were gently mixed on a rocking platform to ensure homogeneity before pipetting. Equal volumes of each sample were delivered to a sterile 55-mL polypropylene solution basin (Labcor Products, Frederick, MD, USA) using an accurately calibrated multichannel pipet. Once all the individual samples for that particular pool were dispensed, the basin containing the DNA was rocked (carefully to avoid spillage) for several minutes to do a preliminary mixing. The DNA was then pipetted into a 50-mL polypropylene tube, making every effort to recover all liquid in the basin; further mixing was done by rocking the tube for about 1 h. The pooled DNA was placed into 1-mL aliquots in sterile 1.5-mL polypropylene microtubes and stored at 4°C in the dedicated refrigerator. The tops of stored tubes were wrapped with Parafilm®. As a quality control, the uniformity of the mixing procedure was verified by genotyping replicate aliquots of the pools for several SNPs. The total volume of the pool was determined by the number of SNPs to be tested in the overall study. Each SNP assay required a minimum of 200 ng (25 ng × 8 replicates equals 20 μL of 10 ng/μL) of pooled DNA. To assay approximately 2500 SNPs required a pool total volume of 50 mL (333 μL each of the 150 individual DNAs at 10 ng/μL). The volume and concentration of the pool could be adjusted to meet the laboratory's own requirements and availability of DNA.

PCR

The reaction consisted of 2.5 μL GeneAmp™ 10× Buffer II (Applied Biosystems, Foster City, CA, USA), 2–3 μL 25 mM $MgCl_2$ solution, 0.5 μL each 20 mM dNTP (Amersham Biosciences, Piscataway, NJ, USA), 1 μL 10 pmol/μL 5′ biotin-TEG labeled, HPLC-purified primer, and 1 μL 10 pmol/μL unlabeled primer (IDT, Coralville, IA, USA), 0.25 μL (1.25 U) AmpliTaq Gold® (Applied Biosystems), 2.5 μL 10 ng/μL DNA (pool or individual), and sterile water to 25 μL total volume. Before pipetting, the DNA pool was gently vortex mixed to ensure that the solution was homogeneous. Thermal cycling was done interchangeably on a GeneAmp 9700™ (Applied Biosystems) or PTC-200™ (MJ Research, Watertown, MA, USA) using the following profile: heated lid, 95°C for 10 min, 45 cycles of 95°C for 45 s,

annealing temperature (56°C–62°C) for 45 s, 72°C for 1 min, and a final hold at 4°C. Forty-five cycles ensure that all PCR components were exhausted. PCR primers were designed with Primer 3 Software (code available at http://www.genome.wi.mit.edu/genome_software/other/primer3.html), and the predicted reaction conditions (annealing temperature and $MgCl_2$) were tested on several nonessential DNA samples. In more than 100 assays tested, the stipulated conditions for primer sets were optimal more than 95% of the time. Fragment sizes of 100–500 bp have been successfully analyzed.

PCR Plate Setup

Plates (96-well) were set up with eight replicates of the case pool in column A and eight replicates of the control pool in column B. A variable number of replicates from 3 to 10 were tested, and it was found that eight replicates most consistently resulted in a standard deviation of 2 or below. The number of replicates for each pool could be modified to meet the stringency of the application. Column C was used for quality-control samples and contained four individual DNA samples and four primer controls. The remaining columns on the plate were set up in the same way for the other SNPs to be tested. Chimney-top-style PCR plates (Phoenix Research Products, Hayward, CA, USA) were used as the deeper wells prevent spillage in downstream manipulations. The individual genotypes helped define any unusual characteristics about each SNP, such as preferential allele amplification (see Discussion section). The primer controls as described in the company protocol further defined the quality of the particular SNP assay, such as baseline noise caused by primer dimerization (Pyrosequencing AB). Four different SNPs were assayed on one plate.

Template Preparation and Pyrosequencing

The PCR product was immobilized, and single-strand isolation was performed with Dynabeads™ M-280 Strepavidin (Dynal Biotech, Oslo, Norway) as described (Pyrosequencing AB). For products over 300 bp, a 30-min hybridization at 65°C was done with 15 µL beads/sample. We utilized Hydra™ 96-well microdispensing robots (Robbins Scientific, Sunnyvale, CA, USA) to pipet the hybridized PCR products and denaturing, washing, and annealing buffers to the PSQ plates™ (Pyrosequencing AB). The sequencing reactions were performed as described (Pyrosequencing AB). No optimization was required.

Allele Quantification Software

Allele frequencies in the samples were assessed by SNP Software AQ (Pyrosequencing AB) as described, and the data were exported to a Microsoft® Excel® spreadsheet for further analysis.

Denaturing HPLC

The PCR was performed as previously described. Heteroduplex DNA was formed and analyzed as described with the Wave™ technology (Transgenomic, Omaha, NE, USA).

RESULTS AND DISCUSSION

To establish the accuracy of SNP allele frequency estimates in DNA pools, we first examined the correlation between allele frequencies and Pyrogram peak heights. To illustrate the readout, homozygous C/C, T/T, or heterozygous C/T individuals for a particular SNP are shown in Figure 1A. In this example, the C/C individual has a peak at the C nucleotide that is equal in height to the peak of the T/T individual at the T nu-

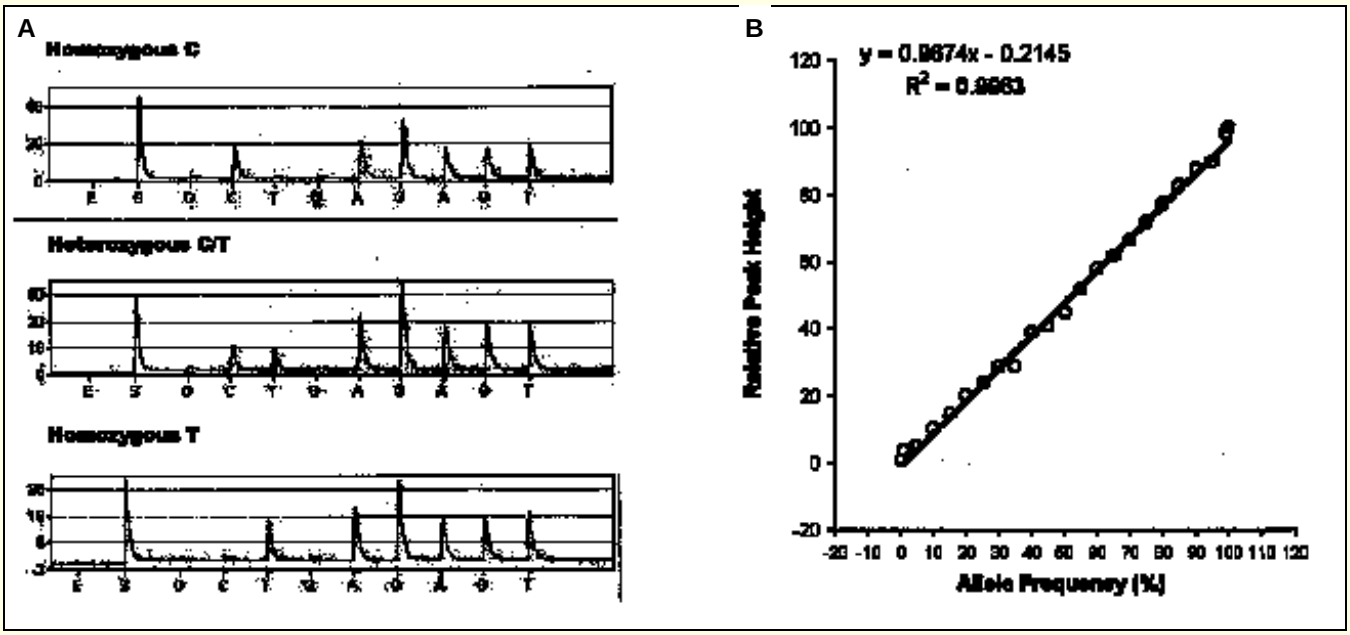

Figure 1. C/T SNP Pyrogram and allele frequencies. (A) Pyrogram for a C/T SNP sequence. (B) Regression line between allele frequencies and Pyrogram peak heights in a mixing experiment between two individuals homozygous for a particular SNP. Each point is the mean of duplicate determinations.

Table 1. Estimation of Allele Frequencies in Pooled DNA Samples Compared to Individual Genotyping

SNP		Pools	Individuals	Delta (pools versus individuals)
ID130 (C/T)[a]	Controls	17.8 ± 0.4% C (*n* = 4)	18.2% C	0.4%
	Cases	14.6 ± 0.3% C (*n* = 10)	15.4% C	0.8%
ID146 (C/T)[a]	Controls	1.4 ± 0.6% T (*n* = 4)	2.8% T	1.4%
	Cases	2.5 ± 0.2% T (*n* = 10)	1.4% T	1.1%
PKIG (C/T)	Controls	15.5 ± 0.5% T (*n* = 8)	15.0% T	0.5%
	Cases	15 ± 0.4% T (*n* = 10)	15.0% T	0.4%
EPPIN3 (G/T)	Controls	20.1 ± 0.3% G (*n* = 3)	21.7% G	1.6%
	Cases	24.3 ± 0.4% G (*n* = 10)	26.0% G	1.7%
MYBL86 (A/G)	Controls	7.4 ± 0.3% G (*n* = 8)	6.2% G	1.2%
	Cases	12.9 ± 0.5% G (*n* = 8)	11.7% G	1.2%
MYBL42 (A/T)	Controls	20.4 ± 0.3% A (*n* = 8)	22.5% A	2.1%
	Cases	27.4 ± 0.3% A (*n* = 8)	27.5% A	0.1%
HNF47 (A/G)	Controls	41.9 ± 0.7% G (*n* = 8)	39.8% G	1.1%
	Cases	36.8 ± 0.5% G (*n* = 8)	34.8% G	2.0%
			$\bar{x}$ ± SEM	1.1 ± 0.6%

Pools of DNA from 150 cases and 150 controls were constructed, and the allele frequencies were determined in replicates of the pools and compared to those determined in the same individuals.

All figures are for the minor allele.

[a]All pool and individual figures were determined with the Pyrosequencer, except for the individual determinations of these two SNPs, which were determined by denaturing HPLC.

cleotide. In contrast, the C/T individual has peak heights at both the C and T bases that are half the height of those in the homozygotes. Control nucleotides that are not part of the actual sequence, in this case Gs, are dispensed immediately before and after the nucleotides that comprise the SNP. The presence of peaks for these control bases implies that the assay has a problem and that new conditions are indicated. The E and S positions are the enzyme and substrate respectively in the reaction.

Next, a mixing experiment was performed in duplicate with the two different homozygotes (C/C and T/T) in various proportions from 1:99, 5:95, 10:90, etc. down to 90:10, 95:5, and 99:1. The SNP Software AQ then converted the peak heights to allele frequencies (Figure 1B). The R^2 statistic for the regression line relating peak height versus estimated allele frequency was 0.9963, and a test for the significance of the regression line yielded a *P* value less than 0.0001, indicating that Pyrogram peak heights accurately reflect allele frequencies. With this assurance, we constructed DNA pools and compared the allele frequencies of pooled DNA with the allele frequencies determined by genotyping the individuals that comprised the pools.

Two DNA pools were assembled, one from 150 cases and the other from 150 controls as described above, and allele frequencies were ascertained for seven different SNPs. The reported minor allele frequencies ranged from less

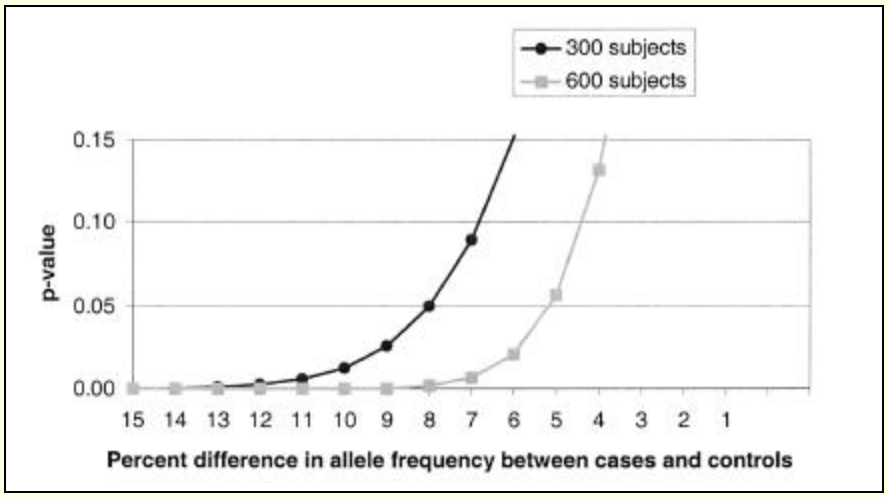

Figure 2. Significance versus percentage difference between cases and controls. Differences in allele frequencies were determined as described in the text. *P* values were calculated using the two-sample test for binomial proportions.

than 5% to 50%. SNPs that contain three or more of the same nucleotide adjacent to the polymorphic site were not selected because the light response following incorporation of several identical nucleotides is not linear (10). In our pools, minor allele frequencies for the SNPs varied from 1.4% to 41.9%. The DNA pools were assayed as few as three to as many as 10 times (Table 1), and standard errors for the replicates from 0.2% to 0.7% were observed. Two different methods of individual genotyping were employed: denaturing HPLC and PSQ96, as indicated in Table 1. The relationships between the allele frequencies estimated in pools compared to allele frequencies determined by individually genotyping are shown. The differences between pools and individual genotypes varied from 0.0% to 2.1%, with mean difference of 1.1 ± 0.6% ($\bar{x} \pm$ SEM).

To assess whether SNP allele frequencies differ between cases and controls, the measurement error, along with the number of individuals tested, must be considered. Therefore, we incorporated the error in allele frequency determination by Pyrosequencing that we observed by analysis of the seven SNPs shown in Table 1 and determined estimates of significance for $n = 300$ or $n = 600$ individuals (Figure 2). We estimated that allele frequencies that differed between cases and controls by at least 8% for $n = 300$ individuals would be required to give a difference significant at the $P < 0.05$ level, and at least 5.2% for $n = 600$ individuals. The P values for the graph were calculated using the two-sample test for binomial proportions (normal theory test) (11), including twice the measurement variance in the calculation of the z statistic. For illustrative purposes, we fixed one allele frequency at 0.30 and varied the second, incrementally adjusting it from 0.30 down to 0.15. Note that for SNPs for which the minor allele is less frequent (<0.30), a smaller difference between the allele frequencies of cases and controls may also be significant. For example, allele frequencies in cases and controls of 5% and 10% yield a significant P value ($P = 0.027$, $n_{cases} = n_{controls} = 300$), while frequencies of 25% and 30% do not ($P = 0.109$, $n_{cases} = n_{controls} = 300$), despite the fact that the difference between cases and controls is 5% in both instances. In practice, the significance of the difference between pools of cases and controls is calculated directly using the two-sample test for binomial proportions.

Below we describe some of the positive and negative aspects we have encountered in Pyrosequencing and suggest methods for maximizing results. One of the advantages of this system is its ease of use in a small laboratory setting. Because the run and analysis times are relatively short, we may conveniently share the instrument with two other laboratories. Furthermore, the instrument has required little maintenance.

After PCR, a 96-well tray can be prepared for sequencing in about 10 min using the PSQ96 Sample Prep Tool. Plates can be processed in batch fashion and stored until sequenced: one week at 4°C or for up to six months at -20°C. The instrument automates the reaction itself. The operator needs only to fill an ink jet printer-type cartridge with nucleotides, enzyme, and substrate provided in a kit and place it in the machine. The instrument computer interface is simple; for example, the sample sheet is in a 96-well format in which the data entered once can easily be copied into other cells. Pyrosequencing of 96 samples takes only 10 min. The automated analysis of genotypes takes less than 1 min. Samples with ambiguous genotypes or errors are highlighted for user editing. A timesaving feature of the SNP Software AQ is that it will calculate the minimum and maximum allele frequencies and standard deviation for multiple samples simply by pressing the control key and selecting the samples of interest. Another timesaving feature is a Web-based SNP sequencing primer design software (Pyrosequencing AB) that simplifies the primer design.

Sixty-four out of 67 SNPs tested to date have been successfully assayed, which is a 96% success rate. The three assays that failed yielded low peak heights, below 2 units. The PCR products were robust and verified by sequencing (ABI PRISM™ Big Dye Terminator Cycle Sequencing Ready Reaction Kits; Applied Biosystems). There were no problematic DNA structures (such as hairpins) near the SNPs. Different SNP sequencing primers

were also utilized. The reasons for these failures remain unresolved.

It should be noted that because of the chemistry of the Pyrosequencing reaction, incorporation of the A nucleotide results in an approximately 9% higher peak relative to peaks for other nucleotides. When estimating the allele frequencies from pools and comparing these to allele frequencies measured from the individual genotypes, an adjustment needs to be made. The adjustment factor panel of the SNP Software AQ allows the operator to correct for this effect as well as other abnormalities such as preferential allele amplification in the PCR or baseline noise in the SNP Pyrogram. On the other hand, when comparing the relative allele frequencies between pools of cases and controls, this effect will subtract out in both pools.

A limitation of measuring SNP allele frequencies in pools is that one cannot estimate Hardy-Weinberg equilibrium or construct haplotypes To obviate this limitation, once we observe allele frequencies that appear to differ between cases and controls, we genotype additional nearby SNPs and then genotype individuals for each SNP and construct haplotypes (3).

There are reports of other methods to assess allele frequencies in DNA pools such as kinetic PCR (4), MALDI-TOF mass spectrometry (12), TaqMan® (1), and direct sequencing (6). The kinetic PCR technique requires no post-PCR processing, and the correlation between measured and known allele frequencies is $R^2 = 0.997$; however, separate PCRs are necessary to assay the two SNP alleles, and the success rate of the assay is 80%. The MALDI-TOF mass spectrometry technique also has a good correlation between measured and known allele frequencies ($R^2 = 0.997$ for one SNP); however, the standard deviations of the measured peak areas are typically 10%. The TaqMan assay reports accurate results with low standard deviations; however, the analysis cost is high and assay optimizations are sometimes difficult. Direct sequencing of PCR-amplified pooled DNA yields only a rough estimate of the SNP allele frequencies, with a variance of up to 5%. Since we have not done any direct comparisons between Pyrosequencing technology and the other methods, we cannot draw any conclusions about their relative merits.

In summary, we find that Pyrosequencing technology yields estimates of allele frequencies in DNA pools to within 2% of those defined by individual genotyping. We predict that this technology will provide the opportunity to genotype many more SNPs for positional cloning and for the evaluation of candidate genes for diseases.

ACKNOWLEDGMENTS

We gratefully acknowledge Drs. Paul Goodfellow and Daniela Gerhard for their helpful comments and suggestions for the manuscript. We also acknowledge Stacey Donelan and Jiyan Ma's assistance with technical support, as well as Ingrid Borecki and Brian Suarez for statistical support. This work was supported in part by National Institutes of Health grant nos. DK16746, DK07120, and DK49583 to M.A.P.

REFERENCES

1. **Breen, G., D. Harold, S. Ralston, D.M. St. Clair, and D. St. Clair.** 2000. Determining SNP allele frequencies in DNA pools. BioTechniques *28*:464-470.
2. **Chen, A., S. Wayne, A. Bell, A. Ramesh, C.R. Srisailapathy, D.A. Scott, V.C. Sheffield, P. Van Hauwe, et al.** 1997. New gene for autosomal recessive non-syndromic hearing loss maps to either chromosome 3q or 19p. Am. J. Med. Genet. *71*:467-471.
3. **Daly, M.J., J.D. Rioux, S.E. Schaffner, T.J. Hudson, and E.S. Lander.** 2001. High-resolution haplotype structure in the human genome. Nat. Genet. *29*:229-232.
4. **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.
5. **Kruglyak, L.** 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. *22*:139-144.
6. **Marth, G., R. Yeh, M. Minton, R. Donaldson, Q. Li, S. Duan, R. Davenport, R. Miller, and P.Y. Kwok.** 2001. Single-nucleotide polymorphisms in the public domain: how useful are they? Nat. Genet. *27*:371-372.
7. **Ott, J.** 1991. Analysis of Human Genetic Linkage, 2nd ed. Johns Hopkins University Press, Baltimore, MD.
8. **Permutt, M.A., J.C. Wasson, B.K. Suarez, J. Lin, J. Thomas, J. Meyer, S. Lewitzky, J.S. Rennich, et al.** 2001. A genome scan for type 2 diabetes susceptibility loci in a genetically isolated population. Diabetes *50*:681-685.
9. **Reich, D.E. and E.S. Lander.** 2001. On the allelic spectrum of human disease. TIG *17*:502-510.
10. **Ronaghi, M.** 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. *11*:3-11.
11. **Rosner, B.** 2000. Two sample test for binomial proportions, p.356-371. *In* Fundamentals of Biostatistics, 5th ed. Duxbury Publishing.
12. **Ross, P., L. Hall, and L.A. Haff.** 2000. Quantitative approach to single-nucleotide polymorphism analysis using MALDI-TOF mass spectrometry. BioTechniques *29*:620-629.
13. **Wasson, J. and G.M. Brodeur.** 2000. Molecular analysis of gene amplification in tumors, p.10.5.1-10.5.18. *In* Current Protocols in Human Genetics, 2nd ed. John Wiley & Sons, New York.

Received 4 December 2001; accepted 26 February 2002.

Address correspondence to:
Dr. Jon Wasson
Washington University School of Medicine
Division of Endocrinology, Diabetes, and Metabolism
Campus Box 8127
660 S. Euclid Ave.
St. Louis, MO 63110, USA.
e-mail: aplab3@im.wustl.edu