Hum Genet (2002) 110 : 395–401
DOI 10.1007/s00439-002-0722-6

ORIGINAL INVESTIGATION

Jonathan D. Gruber · Peter B. Colligan
Johanna K. Wolford

Estimation of single nucleotide polymorphism allele frequency in DNA pools by using Pyrosequencing

Received: 1 February 2002 / Accepted: 4 March 2002 / Published online: 9 April 2002

Abstract Positional cloning of genes underlying complex diseases, such as type 2 diabetes mellitus (T2DM), typically follows a two-tiered process in which a chromosomal region is first identified by genome-wide linkage scanning, followed by association analyses using densely spaced single nucleotide polymorphic markers to identify the causal variant(s). The success of genome-wide single nucleotide polymorphism (SNP) detection has resulted in a vast number of potential markers available for use in the construction of such dense SNP maps. However, the cost of genotyping large numbers of SNPs in appropriately sized samples is nearly prohibitive. We have explored pooled DNA genotyping as a means of identifying differences in allele frequency between pools of individuals with T2DM and unaffected controls by using Pyrosequencing technology. We found that allele frequencies in pooled DNA were strongly correlated with those in individuals (r=0.99, *P*<0.0001) across a wide range of allele frequencies (0.02–0.50). We further investigated the sensitivity of this method to detect allele frequency differences between contrived pools, also over a wide range of allele frequencies. We found that Pyrosequencing was able to detect an allele frequency difference of less than 2% between pools, indicating that this method may be sensitive enough for use in association studies involving complex diseases where a small difference in allele frequency between cases and controls is expected.

J.D. Gruber · P.B. Colligan · J.K. Wolford (✉)
Clinical Diabetes and Nutrition Section,
Phoenix Epidemiology and Clinical Research Branch,
National Institute of Diabetes and Digestive and Kidney Diseases,
National Institutes of Health,
4212 North 16th Street, Phoenix, AZ 85016, USA
e-mail: jwolford@exchange.nuh.gov,
Tel.: +1-602-2005341, Fax: +1-602-2005335

Introduction

Linkage disequilibrium mapping is a potentially powerful approach for fine-scale localization of genetic variants conferring increased susceptibility to complex disease. The recent identification of putative susceptibility genes for type 2 diabetes mellitus (T2DM; Horikawa et al. 2001) and Crohn disease (Hugot et al. 2001; Ogura et al. 2001) indicates that this approach can be successful and serve as a general strategy for positional cloning. The recent successes of the Human Genome project, accompanied by the identification of vast numbers of single nucleotide polymorphisms (SNPs), should facilitate positional cloning efforts.

To maximize detection of linkage disequilibrium between diallelic markers and potential disease-susceptibility alleles, a significant number of densely spaced SNPs should be genotyped throughout the region of interest. However, genotyping a large number of SNPs is costly and time-consuming and is currently not economically feasible in the moderate to large sample sizes required for analyses of complex disease. One approach to circumvent the high cost of SNP genotyping in a large sample is first to screen strategically selected SNPs in pools corresponding to affected and unaffected samples and to determine whether a significant allele frequency difference exists between the two groups. This strategy has been employed using a number of technological platforms including mass spectrometry (Buetow et al. 2001;Wolford et al. 2001), kinetic polymerase chain reaction (PCR; Germer et al. 2000), denaturing high performance liquid chromatography (Hoogendoorn et al. 2000), the Invader assay (Ohnishi et al. 2001), bioluminometric assay (Zhou et al. 2001), and Pyrosequencing (PSQ; Permutt et al. 2001). In these studies, the accuracy of pooled DNA allele frequency measurements was assessed by comparison with individual DNA allele frequencies. Typically, there has been good agreement reported between pooled DNA and individual DNA measurements for most methods.

We assessed the applicability of PSQ in pooled DNA allele frequency measurements because we presently use this platform for SNP genotyping in individual samples and are confident of its sensitivity and accuracy. PSQ is a real-time sequencing method based on a four-enzyme mixture reaction that uses luciferase-luciferin light release as the detection signal for nucleotide incorporation into a target DNA strand (Ronaghi et al. 1998). A specific software program (PSQ 96 SNP Software AQ, version 1.2) is available for the evaluation of allele frequencies based on measurements of peak height, a modification of PSQ known as PSQ-AQ. Permutt et al. (2001) have recently reported a strong correlation between peak height and allele frequency by using PSQ-AQ with four different SNPs in pools of 150 diabetic and 150 non-diabetic Caucasians. Our goal here has been to expand this study of PSQ-AQ in order to estimate pool allele frequencies over a full range of potentially complicating variables, such as actual allele frequency and sequence context, that researchers should expect to encounter in real-world applications of the technology. In addition, we aimed to determine how well allele frequency differences between two pools could be detected by using PSQ-AQ. This is an important consideration, particularly in complex diseases, because allele frequency differences between case and control groups are not expected to be large.

Materials and methods

DNA quantification and pooling

DNA was obtained from Pima Indians who are participants in ongoing longitudinal studies conducted among members of the Gila River Indian Community since 1965 (Knowler et al. 1978). The study was approved by the Institutional Board of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and the Tribal Council of the Gila River Indian Community. All subjects provided written informed consent prior to participation.

Genomic DNA was prepared as described (Prochazka et al. 1993) and quantitated by fluorescence with PicoGreen reagent (Molecular Probes, Eugene, Ore.) and a GENios Fluorescence Reader (Tecan, Research Triangle Park, N.C.). Genomic DNA was diluted by hand 1:250 in TE (10 mM TRIS-HCl, 1 mM EDTA, pH 7.5) in triplicate, and a 50-μl aliquot was transferred to a Greiner Fluorotrac 2000 plate (Greiner, Dusseldorf, Germany). PicoGreen reagent was diluted 1:300 in TE, pH 7.5, and 150 μl of this was added to each diluted DNA sample and allowed to sit at room temperature for 5 min, shielded from light. Standard curves were constructed by serial dilution of a commercially quantitated Lambda DNA standard (Molecular Probes). PicoGreen-stained samples were excited at 485 nm, and emission at 520 nm was measured. The fluorescence value of the reagent blank was subtracted from the fluorescence value for each DNA sample. DNA concentrations were determined from the standard curve. Final DNA concentrations were calculated by averaging the fluorescence results from three separate dilutions.

For allele frequency calculations, we used two pools: 141 individuals with T2DM (age of onset <25 years of age) and 141 unaffected controls (no diabetes after age 45 years). Each pool contained 300 ng each DNA sample at a final concentration of 100 ng/μl. Allele frequencies were calculated from the average of four replicate measurements per pool.

Contrived "pools" were created by titrating DNA in various proportions from two samples representing both homozygous genotypes. To confirm relative DNA concentrations, the samples were fine-calibrated to each other by using an equal volume of each sample as template for PCR amplification and subsequent PSQ-AQ.

Table 1 SNP flanking sequence, PCR amplification primers, and extended primers used in PSQ reactions. All primer sequences are given in the 5'→3' direction. Each biotinylated primer is identified with a *B* preceding the oligonucleotide sequence. Bases included in the assay (i.e., 3' to the sequencing primer) are *underlined*

SNP id	Flanking SNP sequence	Forward primer	Reverse primer	Extended primer
ss1848964	AAAGTATCCC[G/A]GGAACCCTCA	B-TCATGCCCTTGTTACAAA	GATGTTCCTATTAGCCAACT	CCTATCTTGAGGGTTCC
rs1135783	AGCAGGAACA[T/C]GGGCTGTACA	GATTCTGGAAAGCATTGCGTA	B-ATGAGTGCCCACGATGTCA	TGGTGAGCAGGAACA
TSC0282391	TCAAGGAGGA[T/G]AAACAAAAAA	CAACAGAGGGAAATTAGAAGC	B-CAAGTCACTGTATTGCTCTCA	AATTGATGTCAAGGAGG
TSC0049603	TCAGTCACTG[A/T]CATTTAATCA	B-GTACAGTTGCTTCCATTTCC	TTCCCAGAGTATTTAATGGTTA	AACAGTGCAAATATGATTA
ss154882	TTCTAACAAA[T/C]GCAAAAAATA	GTCATCCACTAAACAGAAAAT	B-CAGCCATTCAATAAATCATA	AATATAGTCTTTCTAACAA
ss3592	ATTTATATTC[T/C]AGTCTCACTC	GCTCAGATGCTGTTTACGG	B-CATATTTTTAGCTGGAGTTGG	GTTTACGGATATTTATATTC
TSC0337080	TCAGAAAGAT[A/G]CAGTTTAGAA	CAGGACATGAGGAGAGTTGTT	B-GAATAAATGAATGAGTGATTGAGTG	CATAGAGTCAGAAAGAT
TSC0078338	AGGTGATTTT[C/T]TTCCCCATTT	GGCTTTATTTTCAGGAATTA	B-TCTTATCCCTCACACTCTGG	CCAATCAAGGTGATTTT
1803089a	CAGGGCAAGG[G/A]GTTCAGGGGC	CTGATGGAGAAGGTCAAGAG	B-CCTGGGTTCAGACTTCTGT	AGTCTCAGGGCAAGG
rs491061	AGGAGTGTAA[A/G]AGATATTTGA	B-ATGCCATTGAAAAGAGAATTA	TTTCCAATTACACTTAGTTTAGG	TGCTATATCTCAAATARC
TSC0049924	TGAAAAACTT[C/T]TAGAGCCCAA	GGGTGTGGAAACTGCTATT	B-CCTAGCTTCCCTTTTAGGTC	CTATTTTGTTGAAAAACTT
ss2034040	TATTTAAAAC[A/G]GATACTCAAT	B-AATGTGGCTTATGTGACTGAG	GTTGTTCCAATTATGCTATAAACT	AACAACTGAATTGAGTA
TSC0065808	TAGTGCAATT[T/C]CAGTAAAATT	TGTCACAGGGTACGGTCAAT	B-ATAAATCTGTCCATTTCTACAAG	CAAGAGATTTAGTGCAAT
ss1259272	TTAATCTGTC[C/T]TCCCCTCTCC	CAGGCCGTTCACAAACT	B-TTCCCTCAACCTAGGAGAG	TCGAATATTTCTTTAATC
ss2443968	AAATCAATAT[T/C]CCCTTTCAGA	B-ACAGCAGTTCCAATTCCGTA	CCACTTTGCCATTGCTATTT	AGAGTAGTATTTCTGAAAGG

After calibration of the two homozygotes, each pool was constructed with a final DNA concentration of 100 ng/µl. In total, for each SNP, there were 86 such pools with an allele frequency range of 0.95–0.10, arrayed on a 96-well plate so that the allele frequency in each pool differed from adjacent pools by 0.01. Four replicates of each homozygote were included as standards, and each assay was performed in triplicate.

PCR amplification and PSQ

Fifteen SNPs were selected from markers previously typed in individual DNA samples according to minor allele frequency (0.02–0.5) and flanking sequence context of the variant allele. Most of these SNPs were obtained from a public database (http://www.ncbi.nlm.nih.gov/SNP) and are designated by Reference cluster ID (rs#), NCBI assay ID (ss#), or TSC identifier (TSC#), with the exception of SNP 1803089a, a variant that was discovered in our laboratory during validation of nearby SNP rs1803089. Information on flanking SNP sequence, PCR amplification primers, and PSQ primers is shown in Table 1.

DNA amplification of pools and individual standards used 100 ng genomic DNA in a final reaction volume between 15–50 µl. For each assay, PCR conditions (specifically, primer concentration and total volume) were considered as being optimized when amplification was robust enough to produce non-SNP single-base peaks with an PSQ signal greater than 15 RLU (relative luminescence units) but not so strong as to cause peaks to widen and yield inaccurate raw peak height data. For optimization, amplification reactions were performed with the Expand Long Template PCR System (Roche Molecular Biochemicals; Mannheim, Germany) containing 1× PCR buffer 2, (in 2.25 mM $MgCl_2$), 350 µM dNTPs, 10–20 µM (assay-dependent) forward and reverse primers, 0.02 U/µl enzyme mix, and an equivalent volume of *Taq* Start Antibody (Clontech Laboratories, Palo Alto, Calif.). For genotyping, amplification conditions were as follows: 96°C for 1 min, 45 cycles of 96°C for 20 s, 57°C for 30 s, and 68°C for 30 s, and a final extension at 68°C for 7 min. Amplification of individual DNA samples was performed in a 15-µl final reaction volume containing standard 1× PCR buffer, 1.5 mM $MgCl_2$, 200 µM dNTPs, 20 µM forward and reverse primers, and 0.3 U AmpliTaq Gold (Perkin Elmer, Foster City, Calif.). Amplification conditions were as follows: 96°C for 7 min, 45 cycles of 96°C for 20 s, 57°C for 30 s, and 72°C for 45 s, and a final extension at 72°C for 5 min. In individual DNA genotyping, genotyping reproducibility was evaluated in 30 duplicate samples typed blindly for each marker. All markers had an agreement rate greater than 97%.

PSQ reactions were performed according to the manufacturer's instructions with minor modifications (Pyrosequencing, Uppsala, Sweden). If a 50-µl DNA amplification reaction was required, the post-PCR product was halved and treated as separate 25-µl samples for incubation with beads and then recombined before the NaOH denaturing step. Briefly, 7 µl (10 µg/µl) Streptavidin-coated Dynabeads (Dynal, Oslo, Norway) was added to each PCR product and then 2× binding-washing buffer (10 mM Trizma base, 2 M NaCl, 1 mM EDTA, 0.1% Tween 20, pH 7.6), equivalent to the combined volume of beads and PCR product, was added to each well, and the plates were sealed and shaken at 65°C for 30 min. Solid-phase (bound to beads) samples were transferred consecutively to PSQ plates containing 0.5 M NaOH (for 1 min), 1× annealing buffer (200 mM Trizma Base, 50 mM magnesium acetate, pH 7.6), and 1× annealing buffer containing 15 pmol sequencing primer. Following the last step, samples were heated at 80°C for 6 min and then cooled for approximately 15 min prior to PSQ analysis. PSQ was performed at room temperature on an automated PSQ 96 instrument (Pyrosequencing) according to the manufacturer's instructions. For each assay, a template-only control, a sequencing primer-only control, a biotinylated primer-only control, and a combination of sequencing and biotinylated primers were evaluated to verify that each had negligible effect on baseline signal.

Conversion of raw data to allele frequency

For each SNP assay, we converted the raw peak height data obtained by using the PSQ-AQ software for each pooled sample to allele frequency by first amplifying and genotyping several samples of each available genotype (e.g., AA, AG, GG) by using the pooled DNA amplification protocol. The height of peaks affected by the genotype of each sample was then calculated by using the PSQ-AQ software. The ratio of one allele peak to the sum of both allele peaks was plotted against the sample frequency of that allele. For example, in a G/A polymorphism, the frequency for the G allele would be 1, 0.5, and 0 for a GG, GA, and AA genotype, respectively. The equation for the linear regression best-fit line generated is: [relative raw peak height]=m*[actual allele frequency]+b, which can then be used to convert data from the DNA pools into allele frequencies. We found that multiple replicates of a single sample and individual measurement of multiple samples of the same genotype produced similar measures of standard deviation (data not shown); thus, the use of either or both approaches as standards is acceptable.

Results

We first compared allele frequency estimations derived from pooled DNA to the actual allele frequencies measured in individual samples corresponding to each pool (Fig. 1). The two pools consisted of either 141 diabetic cases or 141 unaffected controls. In a comparison of 30 observations, corresponding to 15 SNPs measured in two pools, we found that allele frequencies measured in pools were strongly correlated with allele frequencies obtained from individual genotypings of the DNA samples comprising each pool (r=0.99, *P*<0.0001).

We next measured the difference in allele frequency between case and control pools to determine whether PSQ-AQ could reliably detect allele frequency differences between groups. In the same set of 15 SNPs, we found the allele frequency difference between case and control pools to be

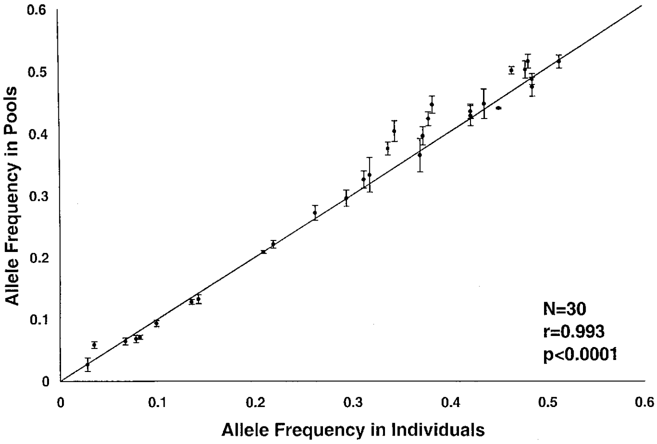


Fig. 1 Allele frequency measurements in pools versus individuals. Pools were each composed of 141 samples, and pool allele frequencies represent the average of four separate PCRs. The allele frequency as measured in individuals is shown on the *x-axis*, and the allele frequency derived from pooled samples is shown on the *y-axis*. There was a total of 30 observations, corresponding to 15 SNPs assayed in two pools. *Line* Line of identity, *y*=*x*

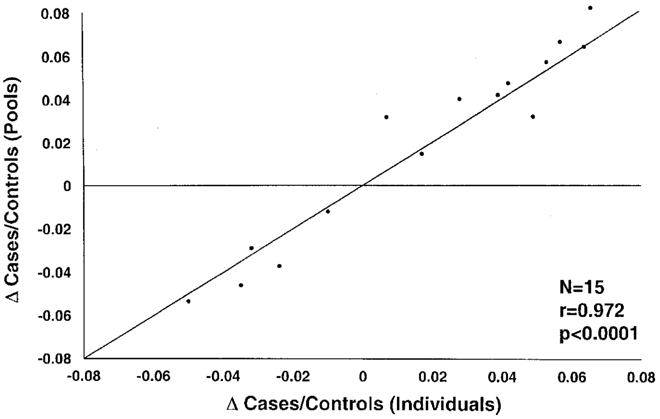
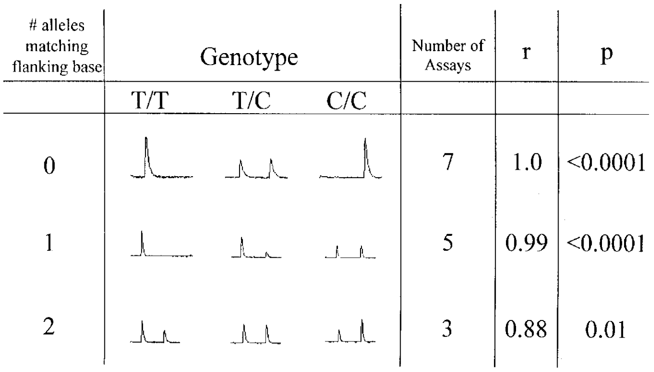


Fig. 2 Allele frequency differences between cases and controls as measured in pools and individuals. The frequency of the minor allele from the control group was subtracted from the corresponding frequency of the case group, for either individuals (*x-axis*) or pools (*y-axis*). Results are derived from 15 SNP assays. *Line* Line of identity, *y*=*x*



# alleles matching flanking base	Genotype			Number of Assays	r	p
	T/T	T/C	C/C			
0				7	1.0	<0.0001
1				5	0.99	<0.0001
2				3	0.88	0.01

Fig. 4 Effect of flanking bases on a Pyrosequencing assay. When a neighboring base is identical to an allele, it contributes to the relative height of the allele peak. *Column 1* Number of alleles matching the flanking bases, *Column 2* pyrograms representing each possible genotype according to the Column 1 class. The number of assays corresponds to how many of our 15 SNPs fit each possible category; the correlations and significance, based on two observations (pools) per assay, are shown in the final two columns

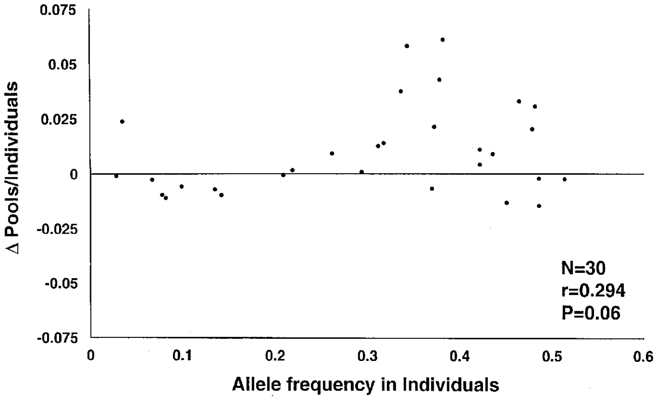


Fig. 3 The relationship between pool error and frequency of the minor allele. To determine whether pool error was related to allele frequency, we subtracted the allele frequency obtained in individuals from the frequency measured in the corresponding pool (*y-axis*) and then plotted this value against the allele frequency measured in individuals (*x-axis*), which we assumed to be the true representation

well correlated (r=0.97, *P*<0.0001) with the actual difference between cases and controls as measured in individuals (Fig. 2).

To determine whether the magnitude of error in frequency estimates of pools was influenced by the frequency of the minor allele, we compared the difference in allele frequency between pools and individuals to the allele frequency derived from individual genotypes (Fig. 3). We found no evidence for a systematic relationship between allele frequency and the extent of deviation of our measurement from the actual allele frequency, suggesting that accuracy is not differentially affected by the frequency of the minor allele. Although it appears that there is greater variability between allele frequencies measured in pools and individuals when the minor allele frequency becomes greater than 0.3, we have found that that this variability is attributable to three specific SNPs. Assays with a similar allele frequency range (i.e., >0.3) show a tight correlation between pools and individuals, indicating that this variability is probably attributable to specific assay conditions, rather than underlying allele frequency. Interestingly, for these three SNPs, we found that, despite the less tight correlation in allele frequency measurements between pools and individuals, allele frequency differences were still accurately estimated. This finding suggests that inaccuracy in allele frequency estimation in pooled samples may not adversely affect the calculation of allele frequency difference between case and control pools.

Deviation from a strict 1:1 allele peak height ratio (e.g., if comparing the signals from opposite homozygotes) might increase the error associated with accurate allele frequency estimation. In the sequencing-by-synthesis method used in PSQ, nucleotides that are both identical and adjacent cause a peak height proportional to the number of nucleotides (i.e., the height of the A peak from GAAC is approximately double that of GAC). Therefore, the context of bases directly flanking a SNP must be considered. The expected contributions of neighboring identical bases to the total height of an allele peak are shown in Fig. 4. Considering the assays (underlined bases in Table 1), we classified the 15 SNPs (two observations per SNP) according to the context of flanking bases to determine whether the accurate estimation of allele frequency in pools was affected. The assays with no interference from neighboring bases (*n*=7) and those in which one allele was identical to three or fewer bases of adjacent sequence (*n*=5) remained correlated with the actual allele frequency (r=1.0 and 0.99, *P*<0.0001). For those assays in which both bases flanking the SNP were each identical to one of the SNP alleles (*n*=3), allele frequency estimated from pools was still correlated with actual allele frequency, although less significantly (r=0.88, *P*=0.01).

To determine how well PSQ-AQ could detect small differences over a range of allele frequencies, we titrated DNA

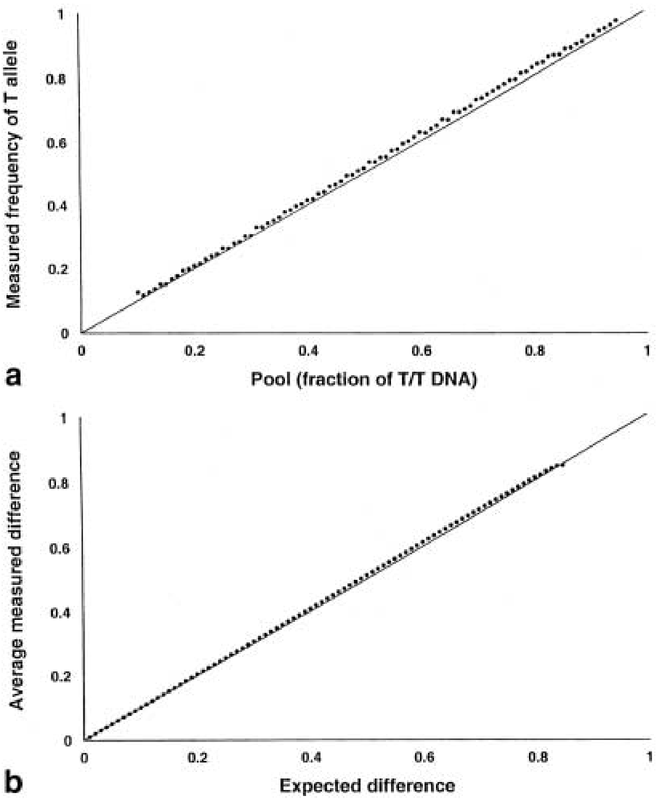


Fig. 5A, B Allele frequencies and allele frequency differences in contrived pools. **A** Allele frequency measurements; *n*=86 pools; allele frequency =[0.95, 0.94, 0.93... 0.10]) for SNP TSC0282391 (three replicates); standard deviations range from 0.0007–0.0132. **B** The averaged measured difference (*y-axis*) reflects an average of all pairwise comparisons for each level of expected difference (*x-axis*). The plot represents the average measured differences between artificial pools for SNP TSC0282391; standard deviations range from 0.0047–0.0088 for data based on more than six comparisons. **A, B** The *line* represents *y*=*x*

from two homozygous samples in various proportions representing 1% increments. For these contrived pools (*n*=86), the allele frequencies calculated from PSQ-AQ were correlated with the intended composition of the pool (Fig. 5A). We then averaged all possible pair-wise differences between these pools for each level of expected difference (Δ=[0.01, 0.02, 0.03,....0.85]), as shown in Fig. 5B. We performed this analysis with a total of four SNPs, all of which yielded similar results (data not shown).

Discussion

The object of this study was to evaluate the efficacy of PSQ-AQ in the accurate estimation of allele frequencies in pooled DNA samples. Estimation of allele frequencies in pooled DNA samples affords one way to conduct a case-control study in a cost- and time-efficient manner. We have envisioned using this approach not to obtain data to test association between a single SNP and complex disease directly, but rather to screen SNPs in pools at high density in the region of interest. Theoretically, with sufficient saturation (e.g., ~1 SNP/10 kb), a cluster of several SNPs should show significant differences between affected and unaffected pools because of linkage disequilibrium with a real susceptibility allele. Once an area has been localized, SNPs can be typed in individual samples, and the region can be scanned for likely candidate genes or thoroughly resequenced to discover novel variants. In essence, estimation of allele frequency in pools should help to narrow the critical region harboring the disease variant(s), after which direct testing of SNP-disease association and/or haplotype construction can be pursued with genotyping of individuals.

In general, allele frequency estimations in pooled samples were correlated with the expected values based on individual genotypes, suggesting that both the pool construction and PSQ-based allele quantification techniques are accurate. Whereas a number of practical considerations must be met to achieve this level of accuracy (see below), we demonstrate that allele frequency estimations in pools are accurate, independent of absolute allele frequency and the context of bases flanking the variant nucleotide. As a further confirmation of the accuracy of this method, we show that differences in allele frequency calculated from pooling data accurately reflect the differences between groups of case and control individuals.

Our studies with contrived pools show clearly that, over a large number of comparisons, PSQ is a highly accurate method for determining allele frequency differences between pools. The standard deviation of the average measured allele frequency difference remains relatively unchanged over a large range of expected differences. This suggests that larger differences between actual pool allele frequencies are more likely to be accurately detected by this method. Across all levels of biologically feasible differences, the pair-wise comparisons yielded an average difference concordant with expected values. Therefore, we can infer that large numbers of comparisons, if all reflect the same underlying allele frequency difference, can yield accurate comparison data, regardless of the magnitude of allele frequency differences. Our experiment averaged the same difference over a range of absolute allele frequencies; obviously, the actual allele frequency of case and control pools and the difference between them are fixed. However, if multiple pools with the same composition of individuals are Pyrosequenced, some errors associated with pool construction can be mitigated, similar to our comparisons of multiple contrived pools.

In an ideal PSQ assay, the relative heights of the two peaks corresponding to the two SNP alleles would directly represent the relative proportions of the alleles in the sample (e.g., a heterozygote would have two peaks of identical height, each half the size of the full peak seen with either homozygote). However, there are a number of factors that affect relative peak heights, including flanking bases that form a homopolymeric run with one or both alleles, unequal amplification of alleles, background signal, and the increased signal strength from an “A” allele because of the substitution of deoxyadenosine alpha-thio triphosphate (dATPαS) for dATP in the PSQ reaction. Although the PSQ-AQ software automatically calculates allele frequencies and raw peak heights, the best way to control for all of these potentially confounding variables is to use individual samples of known genotypes to build a standard curve that can then be utilized to convert raw peak data to allele frequencies. For this strategy to be effective, we found that the signal from the individual samples must be similar to that of the pools; both the absolute height and relative contribution of some background signals will behave in an inversely proportional fashion to the strength of the expected signal generated by the actual sequence.

The need to adjust pool data based on a standard curve of individuals adds a level of complexity to large-scale SNP survey projects insofar as a number of individuals must be genotyped or sequenced to find the appropriate standards. However, in our experience, multiple individuals have no greater variability for PSQ signal than do multiple aliquots of one individual (data not shown). Therefore, all that is required is genotyping of the SNP in a relatively small number of samples to obtain representative genotypes. If the variant is common, a small number of samples is likely to yield all three genotypes; if it is rare, the relative contribution of rare homozygotes to the pool will be small and a standard curve consisting only of the common homozygote and the heterozygote may be adequate.

In the PSQ technique, good assay design is critical for successful genotyping and allele frequency estimation. In addition to designing primers that will not impinge on amplification or PSQ, assay design also depends on the number and identity of bases to be sequenced on either side of the SNP. We found strong correlations between expected and observed allele frequencies for assays in which zero or one allele is identical to its flanking sequence, but assays in which both alleles are involved in different homopolymers have a less strong, yet still significant, correlation. In many situations, runs of identical bases can be easily avoided by designing a sequencing primer that abuts the SNP and anneals to the other bases. This strategy is especially important for SNPs in which both alleles match a flanking base; here, we show that allele frequency can be more accurately quantitated if one match is eliminated. Finally, it is important to consider that some SNPs, such as those in longer homopolymers (>4 bases), are not good candidates for PSQ, in either pools or individuals, unless the sequencing primer placement can shorten or eliminate the homopolymer.

It was beyond the scope of this study to examine the effects of pool size on allele frequency estimation in pooled DNA samples by using PSQ-AQ. However, it is certain that the use of a larger number of individuals in each pool will decrease the contribution that pipetting and DNA quantitation error will make on allele frequency estimation in pools. The manufacturer of the PSQ technique recommends a minimum number of 50 samples per pool and suggests that larger pools (n>100) are less subject to error (S. Toth, personal communication). Based on the accuracy with which we have estimated allele frequency in two pools of 141 individuals each, and considering the large sample sizes thought to be necessary to establish genetic associations in case-control designs (Risch and Teng 1998), optimal pool size will probably never fall short of the recommended minimum pool size, once factors such as population stratification, allele and locus heterogeneity, and expected mode of inheritance have been considered.

In conclusion, we have found that PSQ-AQ provides an accurate and valid approach to allele frequency estimation in pooled DNA samples. In addition, the PSQ technique is sensitive enough to detect small differences between pools across a range of disparate allele frequencies. By adjusting raw data based on the signal from individual genotypes, we show that it remains accurate under various assay conditions, such as a short homopolymeric run. For studies with large numbers of SNPs, whose goal is to narrow a region of genetic linkage by linkage disequilibrium mapping, allele quantification by PSQ in pools is an efficient approach to reduce the labor and expense involved in such experiments.

Acknowledgements We especially thank the members of the Gila River Indian Community for their continued participation in our ongoing longitudinal studies. We should also like to thank Linda Phillips for help with DNA preparation, Jon Wasson for generous sharing of data, and Dr. Clifton Bogardus for helpful discussion.

References

Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR, Braun A (2001) High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Proc Natl Acad Sci USA 98:581–584

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

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

Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, Bosque-Plata L del, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet 26:163–175

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603

Knowler WC, Bennett PH, Hamman RF, Miller M (1978) Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. Am J Epidemiol 108: 497–505

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411:603–606

Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y (2001) A high-throughput SNP typing system for genome-wide association studies. J Hum Genet 46:471–477

Permutt MA, Wasson JC, Donelan S, Skolnick G, Lin J, Suarez BK (2001) Use of DNA pools to assess allele frequencies of single nucleotide polymorphisms (SNPs) at a type 2 diabetes mellitus (T2DM) susceptibility locus. Am J Hum Genet 69: 2305s

Prochazka M, Lillioja S, Tait JF, Knowler WC, Mott DM, Spraul M, Bennett PH, Bogardus C (1993) Linkage of chromosomal markers on 4q with a putative gene determining maximal insulin action in Pima Indians. Diabetes 42:514–519

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

Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. Science 281:363–365

Wolford JK, Kobes S, Hanson RL, Bogardus C, Prochazka M (2001) Linkage disequilibrium mapping of a putative type 2 diabetes locus (1q21-q23) using pooled DNA. Am J Hum Genet 69:2000s

Zhou G, Kamahori M, Okano K, Chuan G, Harada K, Kambara H (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