Short Technical Report

Rapid SNP Allele Frequency Determination in Genomic DNA Pools by Pyrosequencing™

BioTechniques 32:1138-1142 (May 2002)

B. Neve, P. Froguel, L. Corset, E. Vaillant, V. Vatin, and P. Boutin
CNRS UPRES A 8090,
Lille, France

ABSTRACT

Individual genotyping of single nucleotide polymorphisms (SNPs) remains expensive, especially for linkage disequilibrium mapping strategies involving high-throughput SNP genotyping. On one hand, current methods may suit scientific and laboratory needs in regard to accuracy, reproducibility/robustness, and large-scale application. On the other hand, a cheaper and less time-consuming alternative to individual genotyping is the use of SNP allele frequencies determined in DNA pools. We have developed an accurate and reproducible protocol for allele frequency determination using Pyrosequencing™ technology in large genomic DNA pools (374 individuals). The measured correlation (R^2) in large DNA pools was 0.980. In the context of disease-associated SNPs studies, we compared the allele frequencies between the disease (e.g., type 2 diabetes and obesity) and control groups detected by either individual genotyping or Pyrosequencing of DNA pools. In large pools, the variation between the two methods was $1.5 \pm 0.9\%$. It may be concluded that the allele frequency determination protocol could reliably detect over 4% differences between populations. The method is economical in regard to amounts of DNA, PCR, and primer extension reagents required. Furthermore, it allows the rapid determination of allele frequency differences in case/control groups for association studies and susceptibility gene discovery in complex diseases.

INTRODUCTION

Linkage disequilibrium has been successfully used for the positional cloning of complex disease susceptibility genes [NIDDM1 (7) and Crohn's disease (8,13)]. Linkage disequilibrium mapping applied to large chromosomal regions, previously identified by linkage studies, involves the genotyping of a vast amount of single nucleotide polymorphisms (SNPs). One generally aims to genotype as many SNPs as possible in the region of interest to achieve a density of at least 1 SNP/10 kb (12). When applying candidate gene approaches, all coding and regulatory regions of the genes, which include at least 1 kb of the promoter region, exonic sequences, critical intronic sequences (splice junctions and branch-point sequences), and 3'-untranslated regions, are scanned for mutations. In a manageable chromosomal region (e.g., approximately 5 mega-bases), a combined SNP map may require genotyping of 300–500 SNPs.

The methods for SNP genotyping have expanded dramatically in the past decade (9,10). However, individual genotyping remains expensive, especially for linkage disequilibrium mapping strategies involving high-throughput SNP genotyping. Although current methods may suit scientific and laboratory needs in regard to accuracy, reproducibility/robustness, and large-scale application, a cheaper and less time-consuming alternative to individual genotyping is the use of SNP allele frequency determination in DNA pools. It has been shown that RFLP analysis by densitometry (1,3), primer extension followed by denaturing HPLC (5), real-

time PCR with allele-specific amplification (2,4), and mass spectrometry (15) can accurately determine allele frequencies in pools of DNA.

Here we present an efficient protocol for allele frequency determination using Pyrosequencing™ technology (Pyrosequencing AB, Uppsala, Sweden) in small and large DNA pools (up to 374 individuals). For several random SNPs, the reproducibility of this method was tested in DNA pools of different sizes.

MATERIALS AND METHODS

Genotyping by Pyrosequencing

SNP analyses were performed using the PSQ™ 96 system in combination with the SNP reagent kit, according to the recommendations of the manufacturer (Pyrosequencing AB). Briefly, fragments were amplified by PCR (50 cycles) with one biotinylated primer. The biotin-ssDNA was then isolated and applied to minisequencing in the PSQ 96 system (Figure 1A). To isolate the ssDNA branch, we used streptavidin-coated magnetic Dynabeads® (DynaL Biotech S.A., Compiègne, France). Further studies have proven that the streptavidin/Sephadex® isolation protocol, also provided by Pyrosequencing AB, results in equal accuracy and better sensitivity of the minisequencing reaction at a lower concentration of ssDNA. PCR primers were designed according to the manufacturer's guidelines (MWG Biotech, Courtaboeuf ON, Canada). The internal Pyrosequencing primer was designed using the online design service at www.Pyrosequencing.com. The interpretation of the results from the

minisequence reactions (Figure 1B, Pyrograms™) was provided by the PSQ 96 analysis program.

To determine allele frequencies in our populations, we pooled equal amounts of all DNA samples and amplified in triplicate SNP fragments using 5 µL DNA from these pools (approximately 80 ng DNA). Then, minisequencing was performed with an isolated ssDNA branch in the PSQ 96 system. Although similar results were obtained when we used as little as 1 µL (15.8 ± 2.9 ng) pooled DNA in the reaction, we mainly used 5 µL of the pool, which facilitated PCR optimizations. Considering that for individual genotyping approximately 50 ng DNA is used, and that 5 µL of the DNA pool represents 250 times less DNA used per individual and per experiment, our protocol remained economical. The PSQ 96 analysis program interpreted the resulting Pyrograms (Figure 1C) and automatically provided the allele frequencies [mean of % M = peak height M / (peak height M + peak height N) * 100 in triplicate, where M = mutated allele and N = normal allele].

DNA Pooling

DNA was chosen from the French Caucasian population samples previously collected (6,16), which included patients with type 2 diabetes and common obesity. The concentration of DNA samples was measured in a diluted sample (± 0.1 and 1.1 ng/µL) with PicoGreen® (Molecular Probes, Eugene, OR, USA) ($A_{ex/em} = 485/525$ nm) in comparison with a standard concentration curve. Thereafter, approximately 500 ng (approximately 50 µL) DNA were pooled to reach an equal concentration for each sample.

SNP Selection

The presence of reported SNPs in our population was confirmed by Pyrosequencing on an individual basis (8–32 individuals). In addition, this allowed us to evaluate the acquired Pyrograms and to decide if the assay was suitable for allele frequency determination in DNA pools.

In our experience, the accurate determination of allele frequency depends

more critically on the quality of the Pyrogram than does the determination of individual genotypes. Allele frequencies are deducted from the exact magnitude of the peaks of two variants (Figure 1C). Pyrosequencing AB has shown that peak height is linear to allele frequency over a range of 2%–98% (11). However, even low background signals (e.g., from false priming of the internal primer or fragment loop dimerization) may interfere with the peak height and could disturb the apparent prevalence of the variation. Thus, it is important to verify if the peak height in the individual Pyrogram is proportional to the number of bases incorporated. The intrinsic higher signals of the α -thio-dATP base used for minisequencing are automatically corrected by the PSQ 96 analysis program. Some nonspecific signals could be corrected manually if they were a stable proportion of the peak height of one incorporated base.

Acquiring an optimal Pyrogram depends highly on the design of both PCR and Pyrosequencing primers. In most cases, the loop formation of PCR fragments may be avoided by carefully designing PCR primers. For minisequencing, the oligonucleotide design program (Pyrosequencing AB) searches both forward and reverse internal Pyrosequencing primers. If the program fails to find an internal primer with a minimal acceptance score for the SNP (e.g., caused by false-priming sites), then it may be successful with other initial PCR primers for fragment amplification.

RESULTS AND DISCUSSION

Accuracy of Allele Frequencies Determined in DNA Pools

To analyze the accuracy of determined allele frequencies in DNA pools

DRUG DISCOVERY

AND GENOMIC TECHNOLOGIES

by Pyrosequencing, we compared them with allele frequencies deduced from results of individual genotyping. We chose 23 SNPs that met the selection criteria and compared their allele frequency using both methods in DNA pools of different sizes (Figure 2). The measured correlation (R^2) was 0.952 and 0.980 when calculated only with pools larger than 370 DNA samples. The overall variation between the two methods was $3.21 \pm 1.8\%$ (Table 1), and this variation seemed to be independent of the allele frequency of the SNPs (0.05–0.50; data not shown). Although the number of DNA samples in a pool did not have major effects on the variation between both methods, larger pools seem to render slightly better results than smaller pools (Table 1, compare $2.30 \pm 0.9\%$ and $2.78 \pm 2.1\%$, $n = 400$ and $3.35 \pm 0.9\%$, $n = 30$). In larger pools, the random variation in sample concentration may smooth out; however, the accuracy of individual genotyping may diminish and obscure the effect.

However, in a screen for disease-associated SNPs, the accuracy of absolute allele frequencies may be of minor im-

portance compared to the relative differences between the disease and control groups. Therefore, we compared the differences detected by both methods between the allele frequencies of the groups (Table 2). In that case, the overall variation between the two methods was $2.0 \pm 1.4\%$. Furthermore, the differences in the allele frequencies between the largest pools (Tables 1 and 2, G and H) determined by either individual genotyping or Pyrosequencing of DNA pools varied only $1.5 \pm 0.9\%$.

CONCLUSIONS

The comparison of allele frequencies determined after the genotyping of individual DNA samples versus the direct determination in DNA pools by Pyrosequencing showed that the latter is a valid method to analyze allele frequencies of SNPs. The overall correlation between the measured allele frequencies was 0.952. This is less than previously reported correlations of 0.997 (4) and 0.994/0.979 (15) measured for allele frequencies in pooled DNA sam-

Table 1. Comparison of the Two Methods to Determine Allele Frequencies

Pool (no. individuals)	No. SNPs	Differences in allele frequency (%)
A (30)	6	3.35 ± 0.9
B (44)	9	4.44 ± 3.4
C (44)	9	4.44 ± 3.6
D (66)	6	1.23 ± 0.6
E (88)	9	3.56 ± 1.4
F (96)	6	3.07 ± 1.3
G (400)	10	2.30 ± 0.9
H (400)	10	2.78 ± 2.1
Overall		3.21 ± 1.8

ples. However, these DNA pools were constructed by mixing homozygote samples for the two alleles in a different ratio. We could lose precision if the individual genotyping is not 100% accurate. Indeed, in a similar model system, a very strong correlation was shown by Pyrosequencing AB ($R^2 = 0.999$) (15). In addition, Germer et al. (4) and Ross

Figure 1. Genotyping of SNPs by Pyrosequencing (PSQ 96 system). (A) Representation of the minisequencing reaction. (B) Pyrogram interpretation for individual genotyping and (C) allele frequency determination in DNA pools by the PSQ 96 system.

Table 2. Differences between Allele Frequencies of Disease/Control Groups by Both Methods

Pool	No. SNPs	Difference of allele frequencies between groups (%)
Pool A versus F	6	1.9 ± 0.7
Pool B versus E	9	2.2 ± 1.6
Pool C versus E	9	2.0 ± 2.2
Pool D versus F	6	2.5 ± 1.7
Pool H versus G	10	1.5 ± 0.9
Overall		2.0 ± 1.4

et al. (15) studied only 1–3 SNPs in a pool of 100 and 50 DNA, respectively, while in our study, the allele frequencies of 6–10 random SNPs were compared in pools of 30–374 DNA. The replication of PCR (> triplicate) may improve the correlation but increase the costs of the allele frequency determination. Moreover, a correlation of 0.980 was calculated for pools of 370 DNA

samples. An acceptable variation of $1.5 \pm 0.9\%$ was measured when comparing the difference in allele frequency of one SNP and these large pools.

In our screen for disease-associated SNPs, the variations of allele frequency in the disease and control groups are tested for significant differences with a Chi square test. Depending on the underlying true genetic model (recessive,

dominant, or additive), the capacity of a Chi square test to detect significant differences between allele frequencies of small populations may be decreased compared to testing the genotype distribution. The smaller variation of allele frequency between the two methods observed in large DNA pools suggests that the Pyrosequencing protocol is best suited for large genomic DNA pools. In theory, the statistics with groups of 400 individuals show that a SNP with an allele frequency of 0.05 in one pool is significantly different when the prevalence in the other pool differs at least 3.4% (allele frequency of 0.08) and for an SNP with a prevalence of 0.50 at least 6.6% (allele frequency of 0.57) ($P \leq 0.01$; Chi square test). When considering that Pyrosequencing analysis with large DNA pools has a variance of $1.5 \pm 0.9\%$, it may be concluded that when SNP allele frequencies show a 4% difference or more between populations, it

DRUG DISCOVERY

AND GENOMIC TECHNOLOGIES

Figure 2. Comparison of allele frequency determination by both methods. The allele frequency of SNPs in groups A–H determined by individual genotyping was plotted against the allele frequency for these SNPs determined by Pyrosequencing in DNA pools.

is interesting for further screens.

Recently, efforts have been undertaken to add allele frequency information to public SNP databases for African American, Asian, and Caucasian populations. Nevertheless, it remains important to collect details on allele frequencies in a particular study population because several genetic differences between the populations may exist. Furthermore, comparisons of allele frequencies between groups will enable the selection of those SNPs that are associated with a specific phenotype and will be interesting to genotype on a broader scale. In addition to association studies with DNA pools of disease cases and controls, the DNA pooling of affected offspring in sibling transmission disequilibrium tests may be feasible (14).

In conclusion, the determination of SNP allele frequencies by Pyrosequencing is accurate and reproducible in large genomic DNA pools (approximately 370 DNA). The method is economical in regard to the amounts of DNA, PCR, and primer extension reagents required. Furthermore, it allows a rapid determination of allele frequency differences in disease/control groups for association studies and susceptibility gene discovery in complex diseases.

ACKNOWLEDGMENTS

This work was supported by a grant from the Genopole de Lille/Region Nord-Pas de Calais (G-6B) and by the European Union (contract GIFT QLG2-CT-1999-00546). We acknowledge the statistical advice of Christian Dina. B.N. has a post rouge of the Centre National de la Recherche Scientifique (CNRS).

REFERENCES

1. Arnheim, N., C. Strange, and H. Erlich. 1985. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. *Proc. Natl. Acad. Sci. USA* 82:6970-6974.
2. Bray, M.S., E. Boerwinkle, and P.A. Doris. 2001. High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems, and promise. *Hum. Mutat.* 17:296-304.
3. Breen, G., D. Harold, S. Ralston, D. Shaw, and D. St. Clair. 2000. Determining SNP allele frequencies in DNA pools. *BioTechniques* 28:464-470.
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. Giordano, M., M. Mellai, B. Hoogendoorn, and P. Momigliano-Richiardi. 2001. Determination of SNP allele frequencies in pooled DNAs by primer extension genotyping and denaturing high-performance liquid chromatogra-

- phy. *J. Biochem. Biophys. Methods* 47:101-110.
6. Hager, J., C. Dina, S. Francke, S. Dubois, M. Houari, V. Vatin, E. Vaillant, N. Lorentz, et al. 1998. A genome-wide scan for human obesity genes reveals a major susceptibility locus on chromosome 10. *Nat. Genet.* 20:304-308.
7. Horikawa, Y., N. Oda, N.J. Cox, X. Li, M. Orho-Melander, M. Hara, Y. Hinokio, T.H. Lindner, et al. 2000. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.* 28:163-175.
8. Hugot, J.P., M. Chamaillard, H. Zouali, S. Lesage, J.P. Cezard, J. Belaiche, S. Almer, C. Tysk, et al. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411:537-539.
9. Kristensen, V.N., D. Kelefiotis, T. Kristensen, and L. Børresen-Dale. 2001. High-throughput methods for detection of genetic variation. *BioTechniques* 30:318-326.
10. Nowotny, P., J.M. Kwon, and A.M. Goate. 2001. SNP analysis to dissect human traits. *Curr. Opin. Neurobiol.* 11:637-641.
11. Pyrosequencing AB. Estimation of SNP allele frequencies. Technical Note 103.
12. Reich, D.E., M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, et al. 2001. Linkage disequilibrium in the human genome. *Nature* 411:199-204.
13. Rioux, J.D., M.J. Daly, M.S. Silverberg, K. Lindblad, H. Steinhardt, Z. Cohen, T. Delmonte, K. Kocher, et al. 2001. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* 29:223-228.
14. Risch, N. and J. Teng. 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.
15. Ross, P., H. Hall, and L.A. Haff. 2000. Quantitative approach to single-nucleotide polymorphism analysis using MALDI-TOF mass spectrometry. *BioTechniques* 29:620-629.
16. Vionnet, N., E.-H. Hani, S. Dupont, S. Gallina, S. Francke, S. Dotte, F. De Matos, E. Durand, et al. 2000. Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2 diabetes locus on chromosome 1q21-q24. *Am. J. Hum. Genet.* 67:1470-1480.

Received 21 December 2001; accepted 15 February 2002.

Address correspondence to:

Dr. Bernadette Neve
Genetics of Multifactorial Diseases
CNRS UPRES A 8090
Institut Biologie de Lille
Institut Pasteur de Lille
1 rue Calmette 59021
Lille BP447, France
e-mail: bernadette.neve@mail-good.
pasteur-lille.fr