Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Usefulness of direct sequencing of pooled DNA for SNP identification and allele-frequency determination compatible with a common disease/common variant hypothesis

M.K. Kim[1,2], T.S. Nam[1], K.H. Choi[1], S.Y. Jang[2], Y.O. Kim[3] and M.C. Lee[4]

[1]Department of Neurology, Chonnam National University Medical School, Gwangju, Korea
[2]The Brain Korea 21 Project, Center for Biomedical Human Resources at Chonnam National University, Gwangju, Korea
[3]Department of Pediatrics, Chonnam National University Medical School, Gwangju, Korea
[4]Department of Pathology, Chonnam National University Medical School, Gwangju, Korea

Corresponding author: M.K. Kim
E-mail: mkkim@jnu.ac.kr

Genet. Mol. Res. 9 (2): 772-779 (2010)
Received January 10, 2010
Accepted February 4, 2010
Published April 27, 2010
DOI 10.4238/vol9-2gmr761

ABSTRACT. We examined the efficiency of direct sequencing of pooled DNA for developing common single nucleotide polymorphisms (SNPs) and its accuracy for estimating allele frequencies. A pool of 200 control DNAs was established and was used for developing SNPs and estimating minor allele frequencies (MAF). The sensitivity of the pooled DNA method for successfully detecting an SNP with an MAF >0.01 listed in the database was approximately 0.7; it was particularly efficient for detecting SNPs with MAF >0.1, which is compatible with the common disease/common variant hypothesis. The mean difference between the estimated and the observed MAFs was 0.03 ± 0.023. The pooled DNA method identified four additional SNPs, for which the allele frequency

 ©FUNPEC-RP www.funpecrp.com.br

information was not available in the database. The pooled DNA method is a cost- and time-effective tool for both qualifying and quantifying SNPs with considerable accuracy, and it can be particularly useful for dissecting the common disease/common variant hypothesis; this represents a best-case scenario for large-scale association mapping.

Key words: DNA; Pool; Single nucleotide polymorphism; Allele frequency

INTRODUCTION

Associating genetic variations with heritable phenotypes is of key interest in genetics research. Single nucleotide polymorphisms (SNPs) are the most abundant source of genetic variation in the human genome and are expected to facilitate large-scale genetic association studies for detecting alleles that confer small increments in susceptibility to complex phenotypes (Roses, 2000). The common disease/common variant (CD/CV) hypothesis predicts that the genetic risk for common diseases will often be due to disease-predisposing alleles with relatively high frequencies (Chakravarti, 1999). Although there is currently not enough empirical evidence to either prove or disprove the CD/CV hypothesis, some genetic association studies have shown results that lend support to the CD/CV hypothesis (Corder et al., 1993; Bertina et al., 1994; Altshuler et al., 2000; Reich and Lander, 2001). Therefore, information about allelic architecture that refers to the number and frequency of the alleles at a given disease locus is an important prerequisite for association studies to dissect genetic contributions to common complex genetic diseases.

The dbSNP database recently released "NCBI dbSNP Build 130" that contains about 18 million reference SNPs in the human genome, including about 800,000 SNPs in which allele frequency information is known for various ethnic populations, such as Caucasians, Africans, and Asians, including Chinese and Japanese (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). The Korean Haplotype Map (HapMap) Project also developed a Korean SNP database to provide Korean researchers with useful data-mining information for genetic association studies on common disease (Kim et al., 2008); however, since the Korean database is incomplete, many investigators in Korea depend on allele frequency data for Chinese or Japanese populations in selecting candidate SNPs for a large-scale association study. The Chinese or Japanese data have been presumed to be similar to the Korean data, but it is not clear if the data can be used exclusively in genetic association studies for Koreans without any problem. Indeed, a previous study has revealed that there is a considerable difference in allele frequency of SNPs at a locus among different ethnic groups (Kim et al., 2006), which suggests the need for SNP data that are specific to an ethnic group for whom genetic studies will be performed.

Although a variety of techniques are currently available to qualify and quantify SNPs in the human genome (Suh and Vijg, 2005), direct DNA sequencing is the most reliable method, especially in qualifying SNPs. However, direct DNA sequencing is incompatible with cost-effective, large-scale genetic studies. It is known that pooling an equal amount of DNA from individual samples and measuring the relative abundance of alleles in the pool is an efficient strategy for estimating allele frequencies in many samples because it drastically reduces the cost of the analysis and the amount of DNA consumed (Germer et al., 2000; Bang-Ce et al., 2004; Fakhrai-Rad et al., 2004). Accordingly, the combination of

 ©FUNPEC-RP www.funpecrp.com.br

direct DNA sequencing with pooled DNA methods would be a method compatible with a cost-effective, large-scale genetic study.

In this study, we aimed to elucidate both the efficiency of direct sequencing of pooled DNA in developing common SNPs and the accuracy in estimating allele frequencies in a Korean population.

MATERIAL AND METHODS

Subjects

Two hundred healthy Korean volunteers from Gwangju, Korea, who were unrelated and between 22 and 34 years of age, were enrolled in the present study. The study was approved by the Institutional Review Board of Chonnam National University Hospital, and informed consent was obtained from all participants.

DNA pooling

Genomic DNA was extracted from peripheral blood lymphocytes using a standard protocol. To screen common SNPs with a relatively high minor allele frequency (MAF) in the candidate genes in a timely and cost-effective manner, a pool of DNA was established from the healthy control group using the protocol recommended by the manufacturer (PicoGreen dsDNA Quantitation Reagent; Molecular Probes, USA). In brief, double-stranded DNAs were stained with an ultrasensitive fluorescent dye (PicoGreen dsDNA Quantitation Reagent), and the total amount of fluorescent-stained DNA in each individual was measured using a fluorometer (*Pico*fluor™ Handheld Fluorometer). The quantification was performed in duplicates for each sample. An equal amount of DNA from 200 subjects was then mixed into a pool.

SNP development and estimation of the relative allele frequencies

Eighteen candidate genes with a potential influence on the anti-epileptic drug response were tested in the present study (Table 1). In order to develop SNPs in the coding region (cSNP) of each gene, all coding regions, including the exon-intron boundary sequence of the candidate genes, were amplified using the pooled DNA from a control group by polymerase chain reaction (PCR). Approximately 350 forward and reverse primer sets (data available on request) were designed to generate PCR products ranging from 290 to 420 bp in length based on GenBank sequences. PCR assays were carried out using 1.25 U Ampli*Taq* Polymerase Gold (Applied Biosystems, USA), 100 ng pooled DNA, 2.0-2.5 mM $MgCl_2$, and 10 pM primer in a volume of 20 μL. The amplification conditions were as follows: an initial denaturation at 95°C for 5 min, followed by 35 amplification cycles (denaturation at 95°C for 30 s, annealing at various temperatures ranging from 57° to 63°C according to the primer sets for 30 s, and extension at 72°C for 1 min), and a final extension at 72°C for 7 min. The PCR products were electrophoresed on a 1.2% agarose gel, and the amplified genomic DNA fragments were extracted from the gel and purified using a QIAquick® gel extraction kit (Quiagen, Germany) according to manufacturer instructions. Direct sequencing of both strands was performed using BigDye terminator kits (PE Biosystems, USA), and each electropherogram was visually analyzed using Chromas 2.13 (Technelysium Pty Ltd., Queensland, Australia).

The relative allele frequencies for some SNPs determined in this study were estimated using the comparative method proposed by Kwok et al. (1994) and described elsewhere as follows (Jang et al., 2009): allele frequency in pooled DNA = {[reference peak height (individual) / reference peak height (pool)] / [heterozygote peak height (individual) / heterozygote peak height (pool)]} x 0.5 (Figure 1). In order to identify individual heterozygotes for the determined SNPs, 10 random DNA samples consisting of the pooled DNA were genotyped using the same PCR conditions as those used for the pooled DNA.

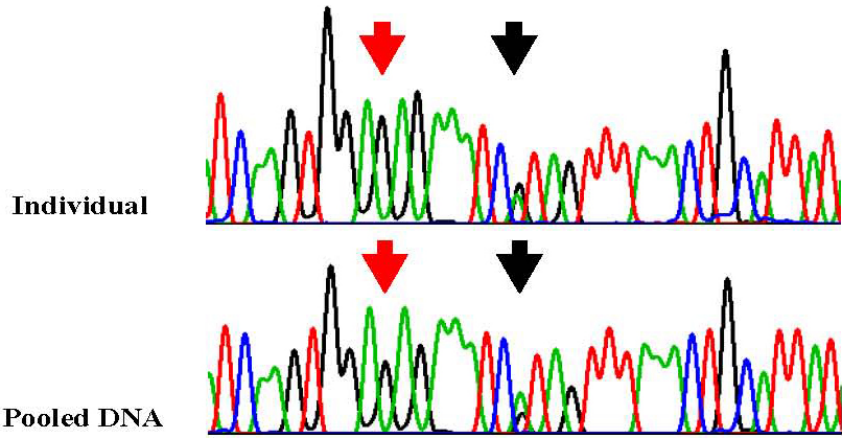

Figure 1. A comparative analysis for estimating relative allele frequencies in a pool of DNA. Allele frequency in pooled DNA = {[reference peak height (individual) / reference peak height (pool)] / [heterozygote peak height (individual) / heterozygote peak height (pool)]} x 0.5. Black arrows indicate heterozygote peaks and red arrows indicate reference peaks.

Efficiency in developing SNPs with pooled DNA

For each candidate gene, the number of SNPs developed in the present study was compared with that released in the dbSNP database. Only validated reference SNPs (RefSNPs) in which the MAF had been determined to be >0.01 for Chinese (HapMap-HCB) or Japanese (HapMap-JPT) populations in the International HapMap Project (http://www.hapmap.org/) were used as gold standards when calculating the sensitivity of the pooled DNA method in detecting the exact SNPs as they were in the database. The sensitivity was calculated as the percentage of RefSNPs in the dbSNP database correctly identified by the pooled DNA method.

Accuracy in estimating allele frequency with the pooled DNA method

For 14 randomly selected RefSNPs in which the allele frequency was estimated with the comparative method described above, direct sequencing of 200 individual DNAs in the pooled DNA was performed to determine the observed allele frequency. The difference between an estimated MAF and an observed MAF (Δ_{EO}) was measured for each RefSNP.

RESULTS

For 14 of 18 candidate genes selected in the present study, a total of 27 RefcSNPs in the HapMap-HCB and HapMap-JPT database were compatible with the condition of this study, cSNP with an MAF >0.01 (Table 1). Of the 27 RefcSNPs with an MAF >0.01 listed in the databases, 19 were reproduced by the pooled DNA method using visual analysis; the sensitivity of the pooled DNA method in successfully detecting a RefSNP with an MAF >0.01 listed in the database was approximately 0.7. Of the other 4 genes (*CHRNB2*, *VGLUT1*, *VGLUT2*, and *VGLUT3*) selected in the present study, 3 genes (except *VGLUT3)* had no compatible RefcSNP in the databases, as it was in the present study (Table 1).

Table 1. Characteristics of reference single nucleotide polymorphisms (RefSNPs).

Gene name (Ch. #)	Accession #	Database	This study	Allele	Protein residue
BCRP (4)	NM_004827	rs2231137	Yes	G/A	Val-Met
		rs2231142	Yes	C/A	Gln-Lys
MRP2 (10)	NM_000392	rs2273697	No	G/A	Val-Ile
		rs8187692	No	G/T	Arg-Leu
		rs3740066	Yes	C/T	Ile-Ile
MDR1 (7)	NM_000927	rs1128503	Yes	T/C	Gly-Gly
		rs2032582	Yes	T/G/A	Ser-Ala-Thr
		rs1045642	Yes	T/C	Ile-Ile
SCN1A (2)	NM_006920	rs7580482	No	C/T	Val-Val
		rs6432860	Yes	C/T	Val-Val
		rs2298771	Yes	G/A	Ala-Thr
SCN1B (19)	NM_001037	rs3746255	No	G/A	Gly-Gly
	NM_199037	No	rs55742440	T/C	Leu-Pro
KCNQ2 (20)	NM_004518	rs2297385	Yes	T/C	Phe-Phe
		rs1801471	No	T/A	Pro-Pro
		No	rs1801475	A/G	Asn-Thr
KCNQ3 (8)	NM_004519	rs2303995	Yes	T/C	Glu-Gly
		rs17575754	No	G/C	Leu-Leu
CLCN2 (3)	NM_004366	rs2228291	Yes	T/C	Ile-Ile
		rs9820367	No	G/C	Thr-Ser
GABRG2 (5)	NM_198904	rs11135176	Yes	C/T	Asn-Asn
		rs211037	Yes	C/T	Asn-Asn
GABRA1 (5)	NM_000806	rs35166395	Yes	C/T	Gly-Gly
EAAT1 (5)	NM_004172	rs2032892	No	G/C	Glu-Asp
EAAT2 (11)	NM_004171	rs752949	Yes	C/T	Pro-Pro
		rs1042113	Yes	A/G	Val-Val
EAAT3 (9)	NM_004170	rs2228622	Yes	G/A	Thr-Thr
		rs301430	Yes	T/C	Thr-Thr
CHRNA4 (20)	NM_000744	rs1044396	Yes	C/T	Ser-Ser
		No	rs1044397	G/A	Ala-Ala
CHRNB2 (1)	NM_000748	No	No	-	-
VGLUT1 (19)	NM_020309	No	No	-	-
VGLUT2 (11)	NM_020346	No	No	-	-
VGLUT3 (12)	NM_139319	No	rs11110359	G/A	Thr-Thr

Ch. # = chromosome number.

Of the 8 RefcSNPs that were not detected by the pooled DNA method, 5 had an average MAF <0.05, while 3 others had an average MAF >0.05 (range: 0.08-0.17), according to the HapMap-HCB and HapMap-JPT data (Table 1). For the 3 RefcSNPs that could not be detected by the pooled DNA method, even though they had a relatively higher MAF for Chinese and Japanese populations, individual genotyping was performed for 120 Korean chromosomes to determine the actual MAF for Koreans; the observed MAFs were 0.08, 0.1, and 0.09

for rs2273697, rs7580482, and rs9820367, respectively. The lowest MAF of a RefcSNP that could be detected by the pooled DNA method was 0.09 for rs2298771.

The pooled DNA method identified 4 additional RefcSNPs (rs1801475, rs1044397, rs55742440, and rs11110359), in which the allele frequency information was not released in the HapMap-HCB and HapMap-JPT database. The observed MAFs of the 4 RefcSNPs in 200 individual Korean DNAs were 0.35, 0.40, 0.26, and 0.15.

An estimation of MAFs in the pooled DNA was performed using the comparative method for 14 RefcSNPs that were randomly selected from the 27 RefcSNPs tested in the present study. The estimated MAF of each RefcSNP was compared with the observed MAF in 200 individual Korean DNAs that made up the DNA pool. The mean difference between the estimated and the observed MAFs was 0.03 ± 0.023, with a range of 0-0.08 (Table 2).

Table 2. Comparison of minor allele frequency.

dbSNP rs #	KRG data		MAF difference
	Estimated (E)	Observed (O)	$\Delta_{KRG(E-O)}$
rs2231137	0.27	0.25	0.02
rs3740066	0.27	0.28	0.01
rs1128503	0.41	0.41	0.00
rs1045642	0.30	0.35	0.05
rs2297385	0.47	0.39	0.08
rs2228291	0.48	0.44	0.04
rs11135176	0.23	0.25	0.02
rs211037	0.41	0.42	0.01
rs35166395	0.27	0.34	0.07
rs752949	0.31	0.32	0.01
rs1042113	0.28	0.31	0.03
rs2228622	0.24	0.28	0.04
rs301430	0.35	0.38	0.03
rs1044396	0.25	0.26	0.01
Mean ± SD			0.03 ± 0.023

KRG = data from the present study for 200 Koreans in Gwangju; MAF = minor allele frequency; $\Delta_{KRG(E-O)}$ = the difference between estimated and observed MAF in KRG.

DISCUSSION

Despite recent remarkable advances in the field of genetic technology, by which a number of genetic diseases (mainly Mendelian diseases) have been dissected (Stenson et al., 2003), little is known about the determinants of common complex genetic diseases, such as epilepsy. It seems that no single factor is either necessary or sufficient for a complex disease, which is the challenge of complex disease mapping because disease susceptibility due to a disease gene can be too small to be detected.

Genetic association analysis has long been considered an effective tool for dissecting complex diseases because of its greater statistical power to detect genes with a modest effect (Long and Langley, 1999). In the analysis, the frequency of alleles at polymorphic loci or, for the reason of both cost and statistical power, of haplotype tagging SNPs (htSNPs) of candidate genes is compared between an affected and a control population with a considerable sample size for identifying genes involved in complex phenotypes (Risch and Merikangas, 1996; Goldstein et al., 2003). In this context, the prerequisites for success in unraveling the genetic components of complex disease phenotypes should include both the detailed informa-

tion about allelic architectures or haplotype structures of human disease genes and a cost- and time-effective genotyping method.

A high-quality reference sequence of the human genome is now available online with the completion of the human genome project (International Human Genome Sequencing Consortium; Anonymous, 2004), and the ongoing International Haplotype Map (HapMap) project is now providing comprehensive information about haplotype pattern and htSNPs of the human genome for several populations (The International HapMap Consortium; Anonymous, 2003; Belmont and Gibbs, 2004). However, there is ample evidence that ethnic, geographic, and other characteristics of the study population greatly influence the frequencies of the different haplotype blocks (Judson et al., 2002; Salisbury et al., 2003; Schneider et al., 2003), which suggests that htSNPs identified in one population may not necessarily perform well in another and that exclusive use of SNP databases for associating genotype with phenotype would not be successful (Suh and Vijg, 2005). The Korean HapMap Project initially worked in 2003 to establish a Korean haplotype map and yield a considerable amount of Korean-specific data until now, but it covers only a small fraction of the whole human genome where the researchers' interests are concentrated (http://www.khapmap.org/). Such an imperfection in the public databases prompts the researcher to seek a cost-effective SNP discovery tool for a large-scale population-based genetic association study.

The direct sequencing of pooled DNA used in the present study was shown to be a cost- and time-effective tool in both qualifying and quantifying SNPs that had been listed in SNP databases. Comparing with the HapMap-HCB and the HapMap-JPT data, the sensitivity of the pooled DNA method in detecting RefSNP with an MAF >0.01 was about 0.7, and it was perfect in detecting RefSNP with an MAF >0.1, which is compatible with the CD/CV hypothesis. In addition, for the randomly selected 14 RefSNPs, the comparative method using pooled DNA was highly reliable in accurately estimating the MAF of the SNPs; the mean MAF difference between the estimated and the observed frequencies was only 0.03. In contrast to other prescreening methods for SNP identification, such as single-strand conformation polymorphism and denaturing high-performance liquid chromatography (Suh and Vijg, 2005), the direct sequencing of pooled DNA has advantages in its popular technique, as well as in identifying the exact position and characters of each novel SNP and estimating its allele frequency at a time without subsequent additional procedures.

In conclusion, the direct sequencing of pooled DNA is a cost- and time-effective tool in both qualifying and quantifying SNPs with considerable accuracy and can be particularly useful in dissecting the CD/CV hypothesis that represents the best-case scenario for large-scale association mapping.

ACKNOWLEDGMENTS

Research supported by the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (grant #A080307).

REFERENCES

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, et al. (2000). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* 26: 76-80.

Anonymous (2003). The International HapMap Project. *Nature* 426: 789-796.
Anonymous (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431: 931-945.
Bang-Ce Y, Peng Z, Bincheng Y and Songyang L (2004). Estimation of relative allele frequencies of single-nucleotide polymorphisms in different populations by microarray hybridization of pooled DNA. *Anal. Biochem.* 333: 72-78.
Belmont JW and Gibbs RA (2004). Genome-wide linkage disequilibrium and haplotype maps. *Am. J. Pharmacogenomics* 4: 253-262.
Bertina RM, Koeleman BP, Koster T, Rosendaal FR, et al. (1994). Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* 369: 64-67.
Chakravarti A (1999). Population genetics-making sense out of sequence. *Nat. Genet.* 21: 56-60.
Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, et al. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261: 921-923.
Fakhrai-Rad H, Zheng J, Willis TD, Wong K, et al. (2004). SNP discovery in pooled samples with mismatch repair detection. *Genome Res.* 14: 1404-1412.
Germer S, Holland MJ and Higuchi R (2000). High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res.* 10: 258-266.
Goldstein DB, Ahmadi KR, Weale ME and Wood NW (2003). Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.* 19: 615-622.
Jang SY, Kim MK, Lee KR, Park MS, et al. (2009). Gene-to-gene interaction between sodium channel-related genes in determining the risk of antiepileptic drug resistance. *J. Korean Med. Sci.* 24: 62-68.
Judson R, Salisbury B, Schneider J, Windemuth A, et al. (2002). How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics* 3: 379-391.
Kim YO, Kim MK, Woo YJ, Lee MC, et al. (2006). Single nucleotide polymorphisms in the multidrug resistance 1 gene in Korean epileptics. *Seizure* 15: 67-72.
Kim YU, Kim SH, Jin H, Park YK, et al. (2008). The Korean HapMap Project Website. *Genomics Inform.* 6: 91-94.
Kwok PY, Carlson C, Yager TD, Ankener W, et al. (1994). Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* 23: 138-144.
Long AD and Langley CH (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* 9: 720-731.
Reich DE and Lander ES (2001). On the allelic spectrum of human disease. *Trends Genet.* 17: 502-510.
Risch N and Merikangas K (1996). The future of genetic studies of complex human diseases. *Science* 273: 1516-1517.
Roses AD (2000). Pharmacogenetics and the practice of medicine. *Nature* 405: 857-865.
Salisbury BA, Pungliya M, Choi JY, Jiang R, et al. (2003). SNP and haplotype variation in the human genome. *Mutat. Res.* 526: 53-61.
Schneider JA, Pungliya MS, Choi JY, Jiang R, et al. (2003). DNA variability of human genes. *Mech. Ageing Dev.* 124: 17-25.
Stenson PD, Ball EV, Mort M, Phillips AD, et al. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* 21: 577-581.
Suh Y and Vijg J (2005). SNP discovery in associating genetic variation with human disease phenotypes. *Mutat. Res.* 573: 41-53.