BMB
Reports

BMB Rep. 2013; 46(6): 305-309
www.bmbreports.org

Inference of kinship coefficients from Korean SNP genotyping data

Seong-Jin Park[#], *Jin Ok Yang*[#], *Sang Cheol Kim, Jekeun Kwon, Sanghyuk Lee & Byungwook Lee*[*]
Korean BioInformation Center (KOBIC), KRIBB, Daejeon 305-333, Korea

The determination of relatedness between individuals in a family is crucial in analysis of common complex diseases. We present a method to infer close inter-familial relationships based on SNP genotyping data and provide the relationship coefficient of kinship in Korean families. We obtained blood samples from 43 Korean individuals in two families. SNP data was obtained using the Affymetrix Genome-wide Human SNP array 6.0 and the Illumina Human 1M-Duo chip. To measure the kinship coefficient with the SNP genotyping data, we considered all possible pairs of individuals in each family. The genetic distance between two individuals in a pair was determined using the allele sharing distance method. The results show that genetic distance is proportional to the kinship coefficient and that a close degree of kinship can be confirmed with SNP genotyping data. This study represents the first attempt to identify the genetic distance between very closely related individuals. [BMB Reports 2013; 46(6): 305-309]

INTRODUCTION

Human genetic variations, referring to all of the genetic characteristics observed within the human genome, are responsible for human diversities (1). Among these genetic variations, single nucleotide polymorphisms (SNPs) are the most common genetic variations between individuals. SNPs are present at a frequency of approximately 1 in every 1,000 bases in the human genome (2). Variations in the human genome can affect how humans develop diseases and respond to pathogens and drugs. SNPs are also thought to be a key to realizing the concept of personalized medicine (3).

Current genotyping technologies allow the analysis of a set of a million SNPs spread across the whole genome for a few hundred dollars per person (4). Thus, large-scale studies involving hundreds of thousands of SNPs and thousands of individuals are feasible. Genome-wide association studies (GWAS) have been widely used to identify common variants that contribute to variation in complex human phenotypes and diseases (5).

Pedigree integrity is important in population-based data with unknown family structures (6). Furthermore, family-based linkage analysis has been tremendously successful in identifying genes underlying diseases with Mendelian inheritance patterns (7). The high-throughput genotyping presents opportunities for pedigree error detection using millions of SNPs and for the identification of the degree of relatedness between a pair of individuals. Knowing the relationships between family members, as in a pedigree, can help produce a more accurate estimate of each individual's haplotypes, i.e., sequences of alleles on a chromosome (8).

The determination of relatedness between individuals in a family pedigree is particularly important for analysis of common complex diseases (9). Inferring the distance of the blood relationship between close or distant relatives is the first key step in various disease-gene association studies. Here, we defined 'kinship coefficient' as the level of the relationship between two persons related by blood, such as parent to child, one sibling to another, grandparent to grandchild or uncle to nephew, first cousins, etc. Previous studies have been mainly focused on the analysis of inter-generational relationships and pedigree comparisons (8, 10-12). However, few attempts have been made to determine the genetic distances of very closely related individuals based on SNP genotyping data.

In this study, we present a method to infer close inter-familial relationships from SNP genotyping data and ranking the kinship with a kinship coefficient of 1^{st}-6^{th} degree in Korean families. The kinship coefficient can be used to verify relationships, reconstruct pedigrees, detect pedigree errors, analyze forensic DNA data, and to identify unknown relationships among family members. In this article, we calculated allele sharing distances (ASD) (13) from SNP genotyping data from 43 individuals of two different Korean families, measured the uncertainty from the calculated ASD scores and translated these scores into kinship coefficients to identify the degree of kinship among Koreans. The kinship coefficient can be used to verify relationships, to reconstruct pedigrees, to detect pedigree errors, to analyze forensic DNA data, and to identify unknown relationships between family members.

[*]Corresponding author. Tel: +82-42-879-8531; Fax: +82-42-879-8519; E-mail: bulee@kribb.re.kr
[#]These authors contributed equally to this work.

http://dx.doi.org/10.5483/BMBRep.2013.46.6.177

Received 27 August 2012, Revised 23 November 2012, Accepted 14 December 2012

Keywords: Allele sharing distance, Kinship coefficient, Korean family, Reference standard, SNP

ISSN: 1976-670X (electronic edition)

RESULTS AND DISCUSSION

We evaluated SNP markers located in four genomic regions, CoRS, and ADME functions from the two microarray platforms in 43 individuals from two Korean families (Table 1). We considered 298 pairs (42 pairs in C_family and 256 pairs in K_family) of individuals from the same family and classified their degree of kinship from 1st to 6th degree relatives (Table 2). The ASD values of the pairs were calculated using equation (eq. 1). We constructed an automatic pipeline for identifying ASDs from SNP genotyping data.

We obtained the genetic distance scores for 1st-6th degree relatives based on data from the Affymetrix and Illumina platforms, respectively. In the two platforms, we found that genetic distance is proportional to degree of kinship. Fig. 1 shows the genetic distance scores for 1st-6th degree relatives based on data from the Affymetrix platform. It also shows that genetic distance can be measured for all degrees of kinship, but the difference in the genetic distance becomes smaller as the degree of kinship is higher. These results also show that a genetic distance between 2nd degree relatives, such as sibling/sibling, is closer than 1st degree relatives, such as parent-child. Theoretically, the parent/child shares and sibling/sibling both share 50% of their genetic material. However, the degree of genetic relatedness for siblings is not necessarily 50%, unlike the absolute 50% ratio between parents and children. The actual ratio between siblings can vary from 100% at one extreme (identical twins) to an exceedingly unlikely 0%. We used two sample *t*-test to test whether the genetic distances between 1st and 2nd degree relatives are statistically different. We found that their values were significantly different (Table 3). If both father and mother are from a genetically isolated community, their child will likely have many more similar genes. In that case, siblings will have more than 50% similarity, but the similarity depends how similar their mother and father are.

Genetic distances were calculated for non-blood related individuals to validate the distance of degree of kinship. We measured the genetic distance between C_family members and K_family members, who do not share common ancestors in their genealogical histories. We found that the distance between non-blood related individuals was higher than that for 6th degree relatives.

We assigned the genetic distance values in this study as a standard of the 1st-6th degree and officially registered these distance values as reference standards for 1st-6th degrees of Korean kinship in the National Center for Standard References Data (NCSRD). Using this data, a degree of kinship between two unknown individuals in a Korean family can be inferred or corrected by simply comparing the genetic distance values of registered reference standards.

We have constructed a FTP site (ftp://ftp.kobic.re.kr/pub/genomesrd) for these genomic reference standards. The users can download the SNP genotyping data for the 43 analyzed Korean individuals and the genetic distance scores for the 1st-6th degrees of Korean kinship. The user can also download the genetic distance values from the homepage of the NCSRD (http://www.srd.re.kr), written in Korean.

We have proposed a method to infer close relationships between two individuals using high-density genotyping data and developed the reference standards of degree of kinship in Korean families. Our approach is the first attempt to calculate the genetic distances of very closely (blood) related individuals. Our approach, based on the allele sharing between any pair of individuals, was sufficient to classify relative pairs as parent-offspring pairs, twins, full siblings, or 2nd, 3rd, 4th, 5th, or 6th degree relatives.

Our method can be performed rapidly for a single pedigree or pair of individuals, and will be useful for a wide range of applications, including forensic DNA analysis (assuming that current forensics technology transitions to high-density SNP genotyping) and relative testing and correcting. Knowing the exact degree of relatedness is also important in determining the relationships between long-separated family and when performing organ transplantation. Our results will be further applied toward automated pedigree reconstruction and association mapping in the absence of a pre-specified pedigree or in the presence of unknown genletic relatedness in the sample. However, our study has an important limitation in that the ref-

Table 2. The number of relationships in the C_family and K_family

Degree of relationship	C_family	K_family	Total
1	12	32	44
2	11	26	37
3	7	53	60
4	12	65	77
5	0	60	60
6	0	20	20
Total	42	256	298

Table 1. Selected SNP markers from Affymetrix and Illumina platforms

Platform	CDS	UTR	Intron	Intergenic	ADME[a]	CoRS[b]
Affymetrix	8,611	7,348	327,494	559,793	13,648	306,149
Illumina	33,437	24,969	430,956	646,093	12,404	306,149

[a]absorption, distribution, metabolism, and excretion, which describe the disposition of a pharmaceutical compound within an organism. [b]Common SNP rs unmber present in both the Affymetrix and Illumina microarray chips.

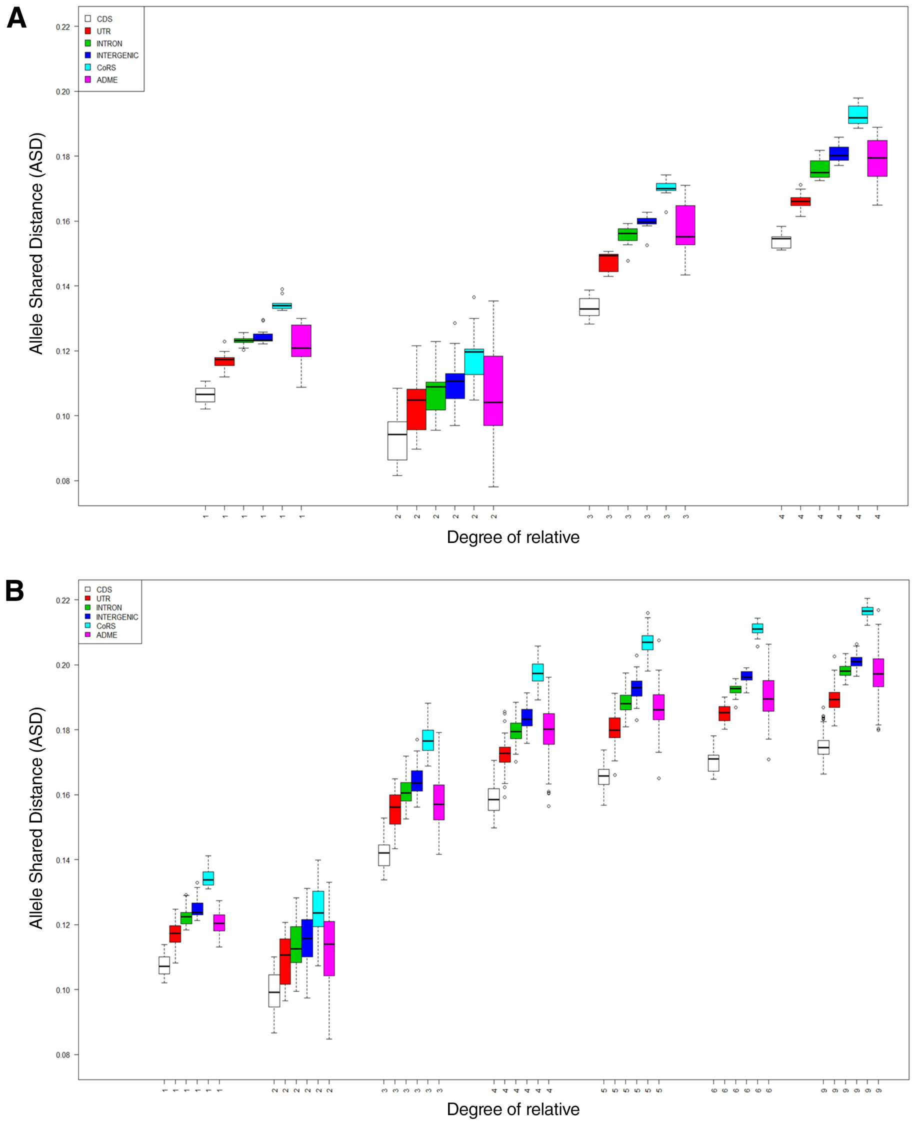

Fig. 1. Genetic distances of 1st-6th (or 4th) degree relatives and unrelated individuals calculated using the ASD algorithm in Affymetrix platform: (A) Choi family, (B) Kang family. The 9th degree relative in X axis of Kang family represents unrelated individuals. CDS and UTR represent 'coding sequence' and 'untranslated region', respectively. CoRS means 'common SNP rs unmber', which represents in both the Affymetrix and Illumina microarray chips. ADME is an acronym in pharmacokinetics and pharmacology for absorption, distribution, metabolism, and excretion, and describes the disposition of a pharmaceutical compound within an organism.

Table 3. P values of two sample *t*-test between 1st and 2nd degree of kinships in Affymetrix and Illumina platforms

Platform	CDS	UTR	Intron	Intergenic	ADME	CoRS
Affymetrix	2.33E-09	6.11E-08	2.64E-10	5.45E-09	5.68E-10	1.81E-04
Illumina	3.65E-10	1.26E-09	1.87E-11	1.28E-10	3.57E-11	9.42E-04

erence standard in the results is valid only for Korean families.

MATERIALS AND METHODS

DNA extraction and genotyping

Blood samples were obtained from 43 Korean individuals from two families, the Choi (C_family) and Kang (K_family) families. The C_family and K_family consisted of 19 and 24 individuals, respectively, whose relatedness was known to 6 degrees. Genomic DNA (gDNA) was extracted from peripheral blood leukocytes using the QIAamp DNA Stool Kit (Qiagen) according to the manufacturer's instructions. This study was approved by the Institutional Review Board (IRB) of the Faculty of Medicine at The Catholic University of Korea. Informed consent was obtained from all participants.

SNP data was obtained using the Affymetrix Genome-wide Human SNP array 6.0 and Illumina Human 1M-Duo chips as recommended by the manufacturers. We obtained 934,969 (Affymetrix) and 1,199,187 (Illumina) SNP calls with a sample call rate of ≥99%, reproducibility of ≥99.9%, and Mendelian inconsistence of ≤0.1%. We filtered SNP markers with a minor allele frequency ≥0.05% and Hardy-Weinberg equilibrium (HWE) P $\leq 10^{-6}$. The filtered SNPs were used for subsequent genetic distance analysis.

Selection of SNP markers

From the filtered Affymetrix and Illumina SNP markers, we selected SNP loci using three different criteria. First, we selected SNP markers that were located in four chromosomal regions (CDS, UTR, intron, and intergenic) based on a SNP annotation file (version 128) from the UCSC genome browser (14) and annotation data from Affymetrix and Illumina. Second, we extracted common SNP rs numbers (CoRS), existing both Affymetrix and Illumina microarray chip. Finally, we obtained information about the effects of these SNP markers on ADME (15). ADME describes the disposition of a pharmaceutical compound within an organism and influence the drug levels and kinetics of drug exposure to the tissues. We downloaded protein information of ADME functions from the Pharmainformatics database (http://bidd.nus.edu.sg/group/admeap/admeap.asp) and obtained SNP markers with effects on ADME by mapping into dbSNP (16) loci using in-house Python scripts.

Allele shared distance (ASD)

To measure kinship coefficients using SNP genotyping data, we considered all possible pairs of individuals in each family; married couples were excluded because they are genetically un-related to each other. We classified these pairs as 1st to 6th relatives according to their degree of kinship. Then, the genetic distance (D) between the two individuals in each pair was calculated. The distance between individual *i* and *j* was defined as

$$D_k = \frac{1}{2U} \sum S_{ij} \qquad \text{(eq. 1)}$$

where S_{ij} is the number of alleles shared between individuals *i* and *j*, U is the number of total loci evaluated, and *k* is degree of kinship. D_k means a genetic distance of k^{th} degree of kinship in a family. From this distance, we obtained ASD values for the kth degree of kinship using a simple function, defined as $ASD_k = 1\text{-}D_k$. If two individuals have the same alleles at nearly all loci, the ASD value will be close to 0; if individuals have no alleles in common, the ASD will be close to 1. We calculated the ASD values for all the degrees of kinship from the two families.

Measurement of uncertainty of genetic distance

To obtain a confidence interval (CI) of ASD for each degree of kinship, we calculated the sample mean and standard error of the mean of the k^{th} degree. We calculated a CI with a 95% confidence level. These confidence intervals are defined as the uncertainty of the measurement, the dispersion of the values that could reasonably be attributed to the measured variable.

Acknowedgements

This study was supported by the KRIBB Research Initiative Program and by the Korean Ministry of Education, Science and Technology (MEST) under grant numbers (20120006002 and 20120006011). This study was also supported by funding of the National Center for Standard Reference Data at the Korea Research Institute of Standards and Science.

REFERENCES

1. Budowle, B. and van Daal, A. (2008) Forensically relevant SNP classes. *Biotechniques* **44**, 603-608, 610.
2. Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X. Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A. and Rothberg, J. M. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876.
3. Ginsburg, G. S. and McCarthy, J. J. (2001) Personalized medicine: revolutionizing drug discovery and patient care. *Trends. Biotechnol.* **19**, 491-496.
4. Kim, S. and Misra, A. (2007) SNP genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.* **9**, 289-320.
5. Barrett, J. C. and Cardon, L. R. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, 659-662.
6. Sahana, G., Guldbrandtsen, B., Janss, L. and Lund, M. S. (2010) Comparison of association mapping methods in a complex pedigreed population. *Genet. Epidemiol.* **34**, 455-462.
7. Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33(Suppl)**, 228-237.
8. Kirkpatrick, B., Halperin, E. and Karp, R. M. (2010) Haplotype inference in complex pedigrees. *J. Comput. Biol.* **17**, 269-280.
9. Lin, S., Ding, J., Dong, C., Liu, Z., Ma, Z. J., Wan, S. and Xu, Y. (2005) Comparisons of methods for linkage analysis and haplotype reconstruction using extended pedigree data. *BMC. Genet.* **6(Suppl 1)**, S76.
10. Takahata, N., Satta, Y. and Klein, J. (1992) Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* **130**, 925-938.
11. Dupanloup, I. and Bertorelle, G. (2001) Inferring admixture proportions from molecular data: extension to any number

of parental populations. *Mol. Biol. Evol.* **18**, 672-675.

12. Paik, S. H., Kim, H. J., Lee, S., Im, S. W., Ju, Y. S., Yeon, J. H., Jo, S. J., Eun, H. C., Seo, J. S., Kim, J. I. and Kwon, O. S. (2011) Linkage and association scan for tanning ability in an isolated Mongolian population. *BMB Rep.* **44**, 741-746.
13. Gao, X. and Starmer, J. (2007) Human population structure detection via multilocus genotype clustering. *BMC Genet.* **8**, 34.
14. Yoon, D., Ban, H. J., Kim, Y. J., Kim, E. J., Kim, H. C., Han, B. G., Park, J. W., Hong, S. J., Cho, S. H., Park, K. and Lee, J. S. (2012) Replication of genome-wide association studies on asthma and allergic diseases in Korean adult population. *BMB Rep.* **45**, 305-310.
15. Balani, S. K., Miwa, G. T., Gan, L. S., Wu, J. T. and Lee, F. W. (2005) Strategy of utilizing in vitro and in vivo ADME tools for lead optimization and drug candidate selection. *Curr. Top. Med. Chem.* **5**, 1033-1038.
16. Day, I. N. (2010) dbSNP in the detail and copy number complexities. *Hum. Mutat.* **31**, 2-4.