ARTICLE

Genomics & Informatics
Vol. 9(4) 181-188, December 2011

http://dx.doi.org/10.5808/GI.2011.9.4.181
pISSN 1598-866X eISSN 2234-0742

How Many SNPs Should Be Used for the Human Phylogeny of Highly Related Ethnicities? A Case of Pan Asian 63 Ethnicities

Hoyoung Ghang[1], Youngjoo Han[1], Sangjin Jeong[1,2], Jong Bhak[3], Sunghoon Lee[3], Tae-Hyung Kim[3], Chulhong Kim[3], Sangsoo Kim[4], Fahd Al-Mulla[5], Chan-Hyun Youn[1]*, Hyang-Sook Yoo[6]* and The HUGO Pan-Asian SNP Consortium

[1]Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, [2]Electronics and Telecommunications Research Institute of Korea, [3]Theragen Bio Institute, Theragen Etex Co. Ltd., Suwon 443-270, [4]Department of Bioinformatics & Life Sciences, Soongsil University, Seoul 156-743, [5]Department of Pathology, University of Kuwait, 13110, Kuwait, [6]Korea Research Institute of Bioscience and Biotechnology (KRIBB), Deajeon 305-806, Korea

Abstract

In planning a model-based phylogenic study for highly related ethnic data, the SNP marker number is an important factor to determine for relationship inferences. Genotype frequency data, utilizing a sub sampling method, from 63 Pan Asian ethnic groups was used for determining the minimum SNP number required to establish such relationships. Bootstrap random sub-samplings were done from 5.6K PASNPi SNP data. DA distance was calculated and neighbour-joining trees were drawn with every re-sampling data set. Consensus trees were made with the same 100 sub-samples and bootstrap proportions were calculated. The tree consistency to the one obtained from the whole marker set, improved with increasing marker numbers. The bootstrap proportions became reliable when more than 7,000 SNPs were used at a time. Within highly related ethnic groups, the minimum SNPs number for a robust neighbor-joining tree inference was about 7,000 for a 95% bootstrap support.

Keywords: neighbour-joining, phylogeny, minimum SNP, ethnic group

Introduction

Autosomal Single Nucleotide Polymorphisms (SNPs) are now widely used for human linkage analyses and demographic history inferences (Abdulla *et al*., 2009; Cavalli-Sforza and Feldman, 2003; Collins *et al*., 1997; Li *et al*., 2008; Wang *et al*., 1998). Other markers such as Short Tandem Repeat (STR), Y chromosomal variation, and mitochondrial variations have been used for the same purpose widely (Agrawal and Khan, 2005; Karafet *et al*., 2008; Mountain and Cavalli-Sforza, 1997; Torroni *et al*., 2006). SNPs have some advantages as (a) they are highly abundant in whole human genome compared to STR, (b) they can be detected with high efficiency by genotyping, and (c) they are preserved over generations compared to the Y chromosome and the mitochondrial genome.

In phylogeny, the search for a minimum marker number has a long history (Felsenstein, 1988; Lecointre *et al*., 1994; Liu and Muse, 2005; Zharkikh and Li, 1992a; Zharkikh and Li, 1992b). Earlier studies have mainly used sequence itself over species. When phylogeny was used as a tool in human evolution and relationship study between ethnicities, so many issues surfaced: genetic distance calculation method, tree drawing method, and marker type disputes (Glover *et al*., 2010; Lin and Nei, 1991; Nei, 1978a; Nei, 1978b; Nei and Roychoudhury, 1974; Tateno *et al*., 1994). The cause of these arguments were mainly that human phylogeny is a study of micro-evolution. Nowadays, SNPs are being utilized for relationship inferences (Hinch *et al*., 2011; Li *et al*., 2010; Travis, 2009). However, the elementary question, regarding the minimum marker number one can use to establish phylogeny-related ethnicity in humans remains unsolved.

In the phylogenic analysis, more SNPs give more information contents, and, hence, a more accurate phylogenic tree. However, when sub-sets of SNPs should be tested for a simulation or a hypothesis, there should be a criteria for a minimum number of markers. Furthermore, for cases using highly related sample groups (ethnicities) or numerous sample groups at a time, the need for the minimum marker number becomes larger. Here, we report the minimum number of SNP that can be used for a robust phylogenic analysis with a highly

*Corresponding authors: E-mail chyoun@kaist.ac.kr
Tel +82-42-350-6126, Fax +82-42-350-7260
E-mail yoohyang@kribb.re.kr
Tel +82-42-860-4170, Fax +82-42-879-8119

Received 14 November 2011, Revised 29 November 2011, Accepted 1 December 2011

related 63 Pan Asian ethnic groups.

Methods

Sample data: PASNPi genotyping data

The 63 ethnic group samples were selected from PASNPi genotyping data (Table 1). PASNPi data were obtained from 72 PASNP ethnic and four HapMap groups. Samples that represent 72 ethnic groups in Pan Asia were obtained from ten Asian countries and Affymetrix data from USA. 1,833 distinct non-duplicated individuals were genotyped with the Affymetrix 50K Xba chip probing 58,960 SNPs. The data included HapMap data, which consists of 209 individuals representing four populations (http://www.hapmap.org/). Common markers between PASNPi and HapMap data were 56,025. To use the 56,025 common marker based NJ tree as a comparison reference tree in the simulation, the tree robustness was tested 1,000 times with 72 ethnic groups. During the test, the bootstrap proportions (BPs) of 13 ethnic groups were not robust (lowest BP of the excluded groups was 44). Because those 13 groups can adversely influence the simulation, they were filtered out to get more accurate result. Removed ethnicities were ID-KR, ID-TB, MY-MN, SG-ML, MY-KN, ID-TR, MY-JH, MY-TM, MY-KS, ID-SU, ID-JA, ID-JV, and MY-BD.

Phylogenic analysis and simulation test

Sub-marker sets were created from 1,000 to 30,000 SNP markers increasing 500 markers at each time. Each

Table 1. Abbreviations of 76 ethnic groups

Ethnic group code	Ethnicity	Ethnic group code	Ethnicity	Ethnic group code	Ethnicity
AX-AI	Karitiana, Maya, Quechua, Auca, Pima	ID-SU	Sunda	PI-MA	Minanubu
AX-AM	Ami	ID-TB	Batak Toba	PI-MW	Mamanwa
AX-AT	Atayal	ID-TR	Toraja	PI-UB	Filipino
AX-ME	Melanesians	IN-DR	Proto-Austroloids	PI-UI	Filipino
CEU-NA	European	IN-EL	Caucasoids (may have admixture with Mongoloids)	PI-UN	Filipino
CHB-NA	Han	IN-IL	Caucasoids	SG-CH	Chinese
CN-CC	Zhuang	IN-NI	Mongoloid features	SG-ID	Indian
CN-GA	Han	IN-NL	Caucasoids	SG-ML	Malay
CN-HM	Hmong	IN-SP	Caucasoids	TH-HM	Hmong (Miao)
CN-JI	Jiamao	IN-TB	Mongoloid features	TH-KA	Karen
CN-JN	Jinuo	IN-WI	Caucasoids	TH-LW	Lawa
CN-SH	Han	IN-WL	Caucasoids	TH-MA	Mlabri
CN-UG	Uyghur	JP-ML	Japanese	TH-MO	Mon
CN-WA	Wa	JP-RK	Ryukyuan	TH-PL	Paluang
ID-AL	Alorese	JPT-NA	Japanese	TH-PP	Plang
ID-DY	Dayak	KR-KR	Korean	TH-TK	Tai Khuen
ID-JA	Javanese	MY-BD	Bidayuh	TH-TL	Tai Lue
ID-JV	Javanese	MY-JH	Negrito	TH-TN	H'tin
ID-KR	Batak Karo	MY-KN	Malay	TH-TU	Tai Yuan
ID-LA	Lamaholot	MY-KS	Negrito	TH-TY	Tai Yong
ID-LE	Lembata	MY-MN	Malay	TH-YA	Yao
ID-ML	Malay	MY-TM	Proto-Malay	TW-HA	Chinese
ID-MT	Mentawai	PI-AE	Ayta	TW-HB	Chinese
ID-RA	Manggarai	PI-AG	Agta	YRI-NA	Yoruban
ID-SB	Kambera	PI-AT	Ati		
ID-SO	Manggarai	PI-IR	Iraya		

The ethnic group codes consist of a two-letter country code followed by another two-letter ethnicity code. Country codes are as follows: [CN: China], [ID: Indonesia], [IN: India], [JP: Japan], [KR: Korea], [MY: Malaysia], [PI: Philippine], [SG: Singapore], [TH: Thailand], [TW: Taiwan], [AX: Affymetrix (not a country)]. Four HapMap samples are as follows: [CHB: Han Chinese in Beijing, China], [CEU: Americans with northern and western European ancestry in Utah, USA], [JPT: Japanese in Tokyo, Japan], and [YRI: Yoruba in Ibadan, Nigeria]. The sampling map of ethnicities was given the earlier PASNPi paper (Abdulla *et al.*, 2009).

sub-marker set was sampled 100 times randomly from whole 56,025 SNPs and was bootstrapped 100 times. Bootstrapping was restricted 100 times because of the computational load. Phylogenic trees were drawn with a neighbor-joining method (Saitou and Nei, 1987) and a consensus tree method, Consense, in the PHYLIP package (Felsenstein, 1989). Genetic distance based on allele frequencies of SNPs was measured with Nei's D_A distance (Nei *et al*., 1983). Takezaki and Nei showed that Nei's D_A and Cavalli-sforza and Edwards's chord distance were more appropriate to get a good quality of tree topologies (Takezaki and Nei, 1996). We used the D_A distance in the phylogenic analysis. Nei's D_A distance between population X and population Y was de-

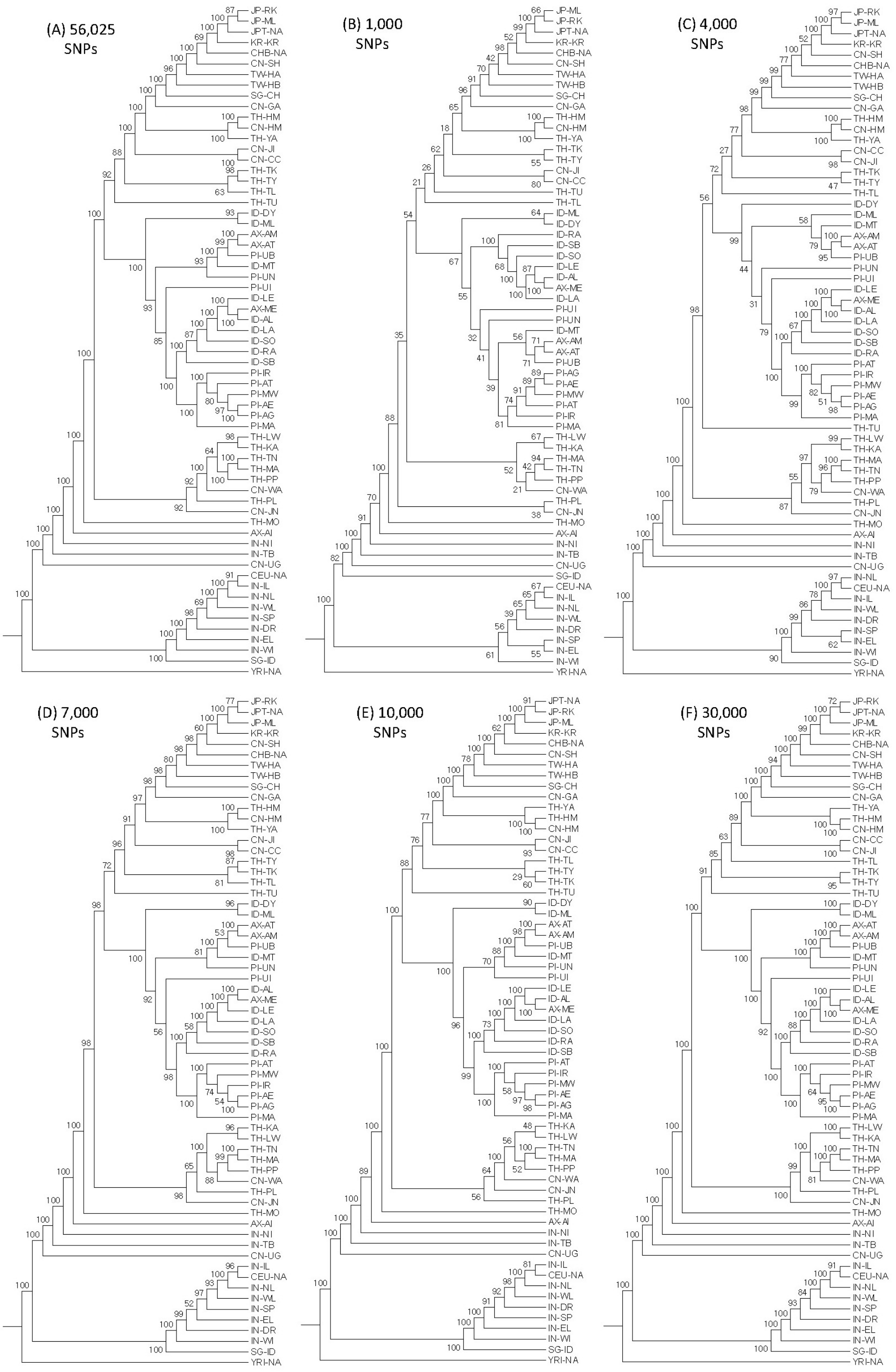

Fig. 1. Representative phylogenic trees. Whole SNP based tree and representative bootstrap trees of each sub-sample are selected. Each bootstrap tree use different number of markers: (A) Whole 56,025 SNPs, (B) 1,000 SNPs, (C) 5,000 SNPs, (D) 7,000 SNPs, (E) 10,000 SNPs, (F) 30,000 SNPs. The index of ethnic group ID is on the Table 1.

fined by

$$D_A = 1 - \frac{1}{r}\sum_{j}^{r}\sum_{i}^{m_j}\sqrt{x_{ij}y_{ij}}$$

where x_{ij} and y_{ij} are the frequencies of the *i*-th allele at the *j*-th locus in populations X and Y, respectively, m_j is the number of alleles at the *j*-th locus, and *r* is the number of loci examined.

In the simulation test for the minimum SNP number required for a robust tree, jackknife and bootstrap re-sampling methods were executed alternatively (Lecointre *et al.*, 1994). As a similarity measure of tree robustness, bootstrap support (BS) was defined by

$$BS = \frac{\textit{Mean BP of test trees of each sub sample set}}{\textit{Mean BP of the reference tree}} \times 100$$

Because of a hard computing task for sub-sampling and bootstrapping, the simulation was performed on a Workflow-based Genomic Cyber Computing (GCC) system that is the main computing platform controlled by computationally intensive workflows in the high performance computing environment (Youn *et al.*, 2011).

Results

Tree topology accuracy and robustness

Five representative bootstrap trees from each 100 sub-sample set and the one of whole 56,025 SNPs were expressed in Fig. 1. The whole SNP based-phylogenic tree of 63 Pan Asian ethnic groups had a stable topology when it was bootstrapped 100 times. The lowest BP was 63 in the group of three Thailand ethnic groups (TH-TK, TH-TY, and TH-TL) and most of the nodes had 100% BPs.

Compared to the one of whole SNPs, tree topologies from 1,000 SNPs to 3,000 SNPs were not consistent within the 100 sub-sets of each random picking number. Furthermore, the BPs were low (the lowest one was 18% in Fig. 1B, a tree of 1,000 SNPs) and, therefore, the robustness of each tree was not supported. When 4,000 SNPs were used in the analysis (Fig. 1C), the BPs were more stable than earlier ones (the lowest one was 27% in the joint node of North East Asians). However, they were not comparable with the one using whole SNP. Additionally, some miss-groupings were observed at the same time (ID-DY and ID-ML, PI-UI and PI-UN, IN-SP and IN-EL, and etc). In the 7,000 SNP based tree, the topology difference to the whole SNP- based tree was very low, and just a few end node joint problems were observed (three Japanese ethnic groups (JP), the location of PI-IR and CN-WA). BPs were high and the tree robustness was acceptable. The lowest BP was 52% in the joint node of Indian ethnicities (INs) and that location had the same problem in the whole SNP based tree (BP was 69%).

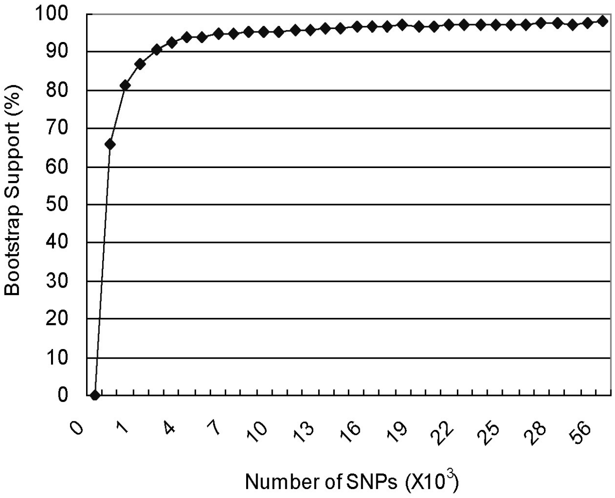

Fig. 2. Bootstrap support (BS) relative to each SNP number. BSs calculated with each 100 sub-sample set are on the Y axis. Sub-samplings of SNPs were done randomly from 1,000 SNPs to 30,000 SNP.

BP increased according to SNP number

Within the 100 random picking sub-samples of the same SNP number, there was some difference in values. Especially, the small number of SNPs (<3,000 SNPs) resulted in higher variability of BPs, and hence, low robustness. With the increase of an analysed SNP number, the tree robustness was improved (Fig. 2). The mean of BP within 100 sub-sample sets were charted in

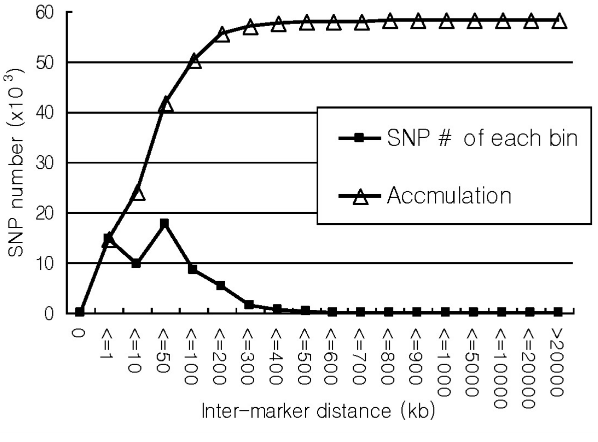

Fig. 3. Inter-marker distance of genotyping data. The marker numbers of each distance (X axis) are on Y axis.

the Fig. 2. The BPs with more than 7,000 SNPs were similar to the ones of whole SNP based tree (BS was more than 95%).

Discussion

If SNPs in a linkage disequilibrium (LD) bin or a same haplotype block are used in a phylogenic analysis, the genetic distance and the resulting tree can be biased to the related SNPs. Thus, the tree could be representing some partial markers that do not reflect the genome-wide relationship pattern. In a 2002 study of Gabriel and his colleagues, it was known that about 90% of LD bins span within 100kb in Asian human genome (Gabriel *et al.*, 2002). About 50,000 SNPs in the Affymetrix 50K chip had the proximity problem (Fig. 3). However, since the strategy of Affymetrix marker selection reflects tag SNPs, most of the genotyped markers were not in one bin or block (Matsuzaki *et al.*, 2004; Nicolae *et al.*, 2006). Furthermore, there were 63 ethnic groups involved in the analysis and they had somewhat different genomic structures within them that are not known yet. Thus, concrete bins or blocks common within 63 ethnicities were not identifiable.

Most of the current human relationship studies use a genome-wide SNP chips (Hinch *et al.*, 2011; Li *et al.*, 2010; Travis, 2009). The small number of markers has worked well within highly different ethnic groups (Agrawal and Khan, 2005; Cavalli-Sforza and Feldman, 2003; Mountain and Cavalli-Sforza, 1997). When highly related ethnicities or a number of ethnicities are considered in a study, a larger marker numbers will be a good strategy. However, when a bulk of markers was used, there would be an inevitable problem of proximity between markers, which can cause a bias to some specific haplotype blocks or LD bins. As an alternative method, based on informative marker sets were studied (Jung *et al.*, 2010; Liu and Muse, 2005). However, those informative marker sets could be less useful when they are used with another third ethnic group, which was not considered during the marker design itself. Thus, random marker could be more informative within numerous ethnicities.

Acknowledgements

The authors would like to thank all colleagues who contributed this study. We are grateful to Mr. Daeui Park, Byung-Chul Ghim, and Ms, JM Kim for their support in the data analysis and the analysis system. This research was supported by a grant from Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2011-0020522) and by a grant from the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MOST), a KOSEF grant (No. R11-2008-044-03004-0, S.M.A.).

References

Abdulla, M.A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S.K., Calacal, G.C., Chaurasia, A., Chen, C.H., Chen, J., Chen, Y.T., Chu, J., Cutiongco-de la Paz, E.M., De Ungria, M.C., Delfin, F.C., Edo, J., Fuchareon, S., Ghang, H., Gojobori, T., Han, J., Ho, S.F., Hoh, B.P., Huang, W., Inoko, H., Jha, P., Jinam, T.A., Jin, L., Jung, J., Kangwanpong, D., Kampuansai, J., Kennedy, G.C., Khurana, P., Kim, H.L., Kim, K., Kim, S., Kim, W.Y., Kimm, K., Kimura, R., Koike, T., Kulawonganunchai, S., Kumar, V., Lai, P.S., Lee, J.Y., Lee, S., Liu, E.T., Majumder, P.P., Mandapati, K.K., Marzuki, S., Mitchell, W., Mukerji, M., Naritomi, K., Ngamphiw, C., Niikawa, N., Nishida, N., Oh, B., Oh, S., Ohashi, J., Oka, A., Ong, R., Padilla, C.D., Palittapongarnpim, P., Perdigon, H.B., Phipps, M.E., Png, E., Sakaki, Y., Salvador, J.M., Sandraling, Y., Scaria, V., Seielstad, M., Sidek, M.R., Sinha, A., Srikummool, M., Sudoyo, H., Sugano, S., Suryadi, H., Suzuki, Y., Tabbada, K.A., Tan, A., Tokunaga, K., Tongsima, S., Villamor, L.P., Wang, E., Wang, Y., Wang, H., Wu, J.Y., Xiao, H., Xu, S., Yang, J.O., Shugart, Y.Y., Yoo, H.S., Yuan, W., Zhao, G., and Zilfalil, B.A. (2009). Mapping human genetic diversity in Asia. *Science* 326, 1541-1545.

Agrawal, S. and Khan, F. (2005). Reconstructing recent human phylogenies with forensic STR loci: a statistical approach. *BMC Genet.* 6, 47.

Cavalli-Sforza, L.L. and Feldman, M.W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.* 33 Suppl, 266-275.

Collins, F.S., Guyer, M.S., and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science* 278, 1580-1581.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22, 521-565.

Felsenstein, J. (1989). PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics* 5, 164-166.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* 296, 2225-2229.

Glover, K.A., Hansen, M.M., Lien, S., Als, T.D., Hoyheim, B., and Skaala, O. (2010). A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genet.* 11, 2.

Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., Aldrich, M.C., Ambrosone, C.B., Amos, C., Bandera, E.V., Berndt, S.I., Bernstein, L., Blot, W.J., Bock, C.H., Boerwinkle, E., Cai, Q., Caporaso, N., Casey, G., Cupples, L.A., Deming, S.L., Diver, W.R., Divers, J., Fornage, M., Gillanders, E.M., Glessner, J., Harris, C.C., Hu, J.J., Ingles, S.A., Isaacs, W., John, E.M., Kao, W.H., Keating, B., Kittles, R.A., Kolonel, L.N., Larkin, E., Le Marchand, L., McNeill, L.H., Millikan, R.C., Murphy, A., Musani, S., Neslund-Dudas, C., Nyante, S., Papanicolaou, G.J., Press, M.F., Psaty, B.M., Reiner, A.P., Rich, S.S., Rodriguez-Gil, J.L., Rotter, J.I., Rybicki, B.A., Schwartz, A.G., Signorello, L.B., Spitz, M., Strom, S.S., Thun, M.J., Tucker, M.A., Wang, Z., Wiencke, J.K., Witte, J.S., Wrensch, M., Wu, X., Yamamura, Y., Zanetti, K.A., Zheng, W., Ziegler, R.G., Zhu, X., Redline, S., Hirschhorn, J.N., Henderson, B.E., Taylor, H.A., Jr., Price, A.L., Hakonarson, H., Chanock, S.J., Haiman, C.A., Wilson, J.G., Reich, D., and Myers, S.R. (2011). The landscape of recombination in African Americans. *Nature* 476, 170-175.

Jung, J., Kang, H., Cho, Y.S., Oh, J.H., Ryu, M.H., Chung, H.W., Seo, J.S., Lee, J.E., Oh, B., Bhak, J., and Kim, H.L. (2010). Gene Flow between the Korean Peninsula and Its Neighboring Countries. *PLoS One* 5, e11855.

Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L., and Hammer, M.F. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 18, 830-838.

Lecointre, G., Philippe, H., Van Le, H.L., and Le Guyader, H. (1994). How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. *Mol. Phylogenet. Evol.* 3, 292-309.

Li, D., Sun, Y., Lu, Y., Mustavich, L.F., Ou, C., Zhou, Z., Li, S., Jin, L., and Li, H. (2010). Genetic origin of Kadai-speaking Gelong people on Hainan island viewed from Y chromosomes. *J. Hum. Genet.* 55, 462-468.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100-1104.

Lin, J. and Nei, M. (1991). Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data. *Mol. Biol. Evol.* 8, 356-365.

Liu, K. and Muse, S.V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21, 2128-2129.

Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H., Yang, G., Kennedy, G.C., Webster, T.A., Cawley, S., Walsh, P.S., Jones, K.W., Fodor, S.P., and Mei, R. (2004). Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* 1, 109-111.

Mountain, J.L. and Cavalli-Sforza, L.L. (1997). Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am. J. Hum. Genet.* 61, 705-718.

Nei, M. (1978a). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89, 583-590.

Nei, M. (1978b). The theory of genetic distance and evolution of human races. *Jinrui Idengaku Zasshi.* 23, 341-369.

Nei, M. and Roychoudhury, A.K. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics* 76, 379-390.

Nei, M., Tajima, F. and Tateno, Y. (1983). Accuracy of Estimated Phylogenetic Trees from Molecular-Data.2. Gene-Frequency Data. *J. Mol. Evol.* 19, 153-170.

Nicolae, D.L., Wen, X., Voight, B.F. and Cox, N.J. (2006). Coverage and characteristics of the Affymetrix GeneChip Human Mapping 100K SNP set. *PLoS Genet.* 2, e67.

Saitou, N. and Nei, M. (1987). The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* 4, 406-425.

Takezaki, N. and Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* 144, 389-399.

Tateno, Y., Takezaki, N. and Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* 11, 261-277.

Torroni, A., Achilli, A., Macaulay, V., Richards, M. and Bandelt, H.J. (2006). Harvesting the fruit of the human mtDNA tree. *Trends Genet.* 22, 339-345.

Travis, J. (2009). Forensic science. Scientists decry isotope, DNA testing of 'nationality'. *Science* 326, 30-31.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M. and Lander, E.S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077-1082.

Youn, C.H., Shim, E.B., Lim, S., Cho, Y.M., Hong, H.K., Choi, Y.S., Park, H.D. and Lee, H.K. (2011). A cooperative metabolic syndrome estimation with high precision sensing unit. *IEEE Trans. Biomed. Eng.* 58, 809-813.

Zharkikh, A. and Li, W.H. (1992a). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Mol. Biol. Evol.* 9, 1119-1147.

Zharkikh, A. and Li, W.H. (1992b). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: II. Four taxa without a molecular clock. *J. Mol. Evol.* 35, 356-366.

The participants of the HUGO Pan-Asian SNP Consortium are arranged by surname alphabetically in the following

Mahmood Ameen Abdulla,[1] Ikhlak Ahmed,[2] Anunchai Assawamakin,[3,4] Jong Bhak,[5] Samir K. Brahmachari,[2] Gayvelline C. Calacal,[6] Amit Chaurasia,[2] Chien-Hsiun

Chen,[7] Jieming Chen,[8] Yuan-Tsong Chen,[7] Jiayou Chu,[9] Eva Maria C. Cutiongco-de la Paz,[10] Maria Corazon A. De Ungria,[6] Frederick C. Delfin,[6] Juli Edo,[1] Suthat Fuchareon,[3] Ho Ghang,[5] Takashi Gojobori,[11,12] Junsong Han,[13] Sheng-Feng Ho,[7] Boon Peng Hoh,[14] Wei Huang,[15] Hidetoshi Inoko,[16] Pankaj Jha,[2] Timothy A. Jinam,[1] Li Jin,[17,37] Jongsun Jung,[18] Daoroong Kangwanpong,[19] Jatupol Kampuansai,[19] Giulia C. Kennedy,[20,21] Preeti Khurana,[22] Hyung-Lae Kim,[18] Kwangjoong Kim,[18] Sangsoo Kim,[23] Woo-Yeon Kim,[5] Kuchan Kimm,[24] Ryosuke Kimura,[25] Tomohiro Koike,[11] Supasak Kulawonganunchai,[4] Vikrant Kumar,[8] Poh San Lai,[26,27] Jong-Young Lee,[18] Sunghoon Lee,[5] Edison T. Liu,[8] Partha P. Majumder,[28] Kiran Kumar Mandapati,[22] Sangkot Marzuki,[29] Wayne Mitchell,[30,31] Mitali Mukerji,[2] Kenji Naritomi,[32] Chumpol Ngamphiw,[4] Norio Niikawa,[39] Nao Nishida,[25] Bermseok Oh,[18] Sangho Oh,[5] Jun Ohashi,[25] Akira Oka,[16] Rick Ong,[8] Carmencita D. Padilla,[10] Prasit Palittapongarnpim,[33] Henry B. Perdigon,[6] Maude Elvira Phipps,[1,34] Eileen Png,[8] Yoshiyuki Sakaki,[35] Jazelyn M. Salvador,[6] Yuliana Sandraling,[29] Vinod Scaria,[2] Mark Seielstad,[8] Mohd Ros Sidek,[14] Amit Sinha,[2] Metawee Srikummool,[19] Herawati Sudoyo,[29] Sumio Sugano,[36] Helena Suryadi,[29] Yoshiyuki Suzuki,[11] Kristina A. Tabbada,[6] Adrian Tan,[8] Katsushi Tokunaga,[25] Sissades Tongsima,[4] Lilian P. Villamor,[6] Eric Wang,[20,21] Ying Wang,[15] Haifeng Wang,[15] Jer-Yuarn Wu,[7] Huasheng Xiao,[13] Shuhua Xu,[37] Jin Ok Yang,[5] Yin Yao Shugart,[38] Hyang-Sook Yoo,[5] Wentao Yuan,[15] Guoping Zhao,[15] Bin Alwi Zilfalil,[14] Indian Genome Variation Consortium[2]

[1]Department of Molecular Medicine, Faculty of Medicine, and the Department of Anthropology, Faculty of Arts and Social Sciences, University of Malaya, Kuala Lumpur, 50603, Malaysia. [2]Institute of Genomics and Integrative Biology, Council for Scientific and Industrial Research, Mall Road, Delhi 110007, India. [3]Mahidol University, Salaya Campus, 25/25 M. 3, Puttamonthon 4 Road, Puttamonthon, Nakornpathom 73170, Thailand. [4]Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, Thailand Science Park, Pathumtani 12120, Thailand. [5]Korean BioInformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), 111 Gwahangno, Yuseong-gu, Deajeon 305-806, Korea. [6]DNA Analysis Laboratory, Natural Sciences Research Institute, University of the Philippines, Diliman, Quezon City 1101, Philippines. [7]Institute of Biomedical Sciences, Academia Sinica, 128 Sec 2 Academia Road Nangang, Taipei City 115, Taiwan. [8]Genome Institute of Singapore, 60 Biopolis Street 02-01, 138672, Singapore. [9]Institute of Medical Biology, Chinese Academy of Medical Science, Kunming, China. [10]Institute of Human Genetics, National Institutes of Health, University of the Philippines Manila, 625 Pedro Gil Street, Ermita Manila 1000, Philippines. [11]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411-8540, Japan. [12]Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan. [13]National Engineering Center for Biochip at Shanghai, 151 Li Bing Road, Shanghai 201203, China. [14]Human Genome Center, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia. [15]MOST-Shanghai Laboratory of Disease and Health Genomics, Chinese National Human Genome Center Shanghai, 250 Bi Bo Road, Shanghai 201203, China. [16]Department of Molecular Life Science Division of Molecular Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara-A Kanagawa-Pref A259-1193, Japan. [17]State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, 220 Handan Road, Shanghai 200433, China. [18]Korea National Institute of Health, 194, Tongil-Lo, Eunpyung-Gu, Seoul, 122-701, Korea. [19]Department of Biology, Faculty of Science, Chiang Mai University, 239 Huay Kaew Road, Chiang Mai 50202, Thailand. [20]Genomics Collaborations, Affymetrix, 3420 Central Expressway, Santa Clara, CA 95051, USA. [21]Veracyte, 7000 Shoreline Court, Suite 250, South San Francisco, CA 94080, USA. [22]The Centre for Genomic Applications (an IGIB-IMM Collaboration), 254 Ground Floor, Phase III Okhla Industrial Estate, New Delhi 110020, India. [23]Soongsil University, Sangdo-5-dong 1-1, Dongjak-gu, Seoul 156-743, Korea. [24]Eulji University College of Medicine, 143-5 Yong-du-dong Jung-gu, Dae-jeon City 301-832, Korea. [25]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. [26]Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, National University Hospital, 5 Lower Kent Ridge Road, 119074, Singapore. [27]Population Genetics Lab, Defence Medical and Environmental Research Institute, DSO National Laboratories, 27 Medical Drive, 117510, Singapore. [28]Indian Statistical Institute (Kolkata) 203 Barrackpore Trunk Road, Kolkata 700108, India. [29]Eijkman Institute for Molecular Biology, Jl. Diponegoro 69, Jakarta 10430, Indonesia. [30]Informatics Experimental Therapeutic Centre, 31 Biopolis Way, 03-01 Nanos, 138669, Singapore. [31]Division of Information Sciences, School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore. [32]Department of Medical Genetics, University of the

Ryukyus Faculty of Medicine, Nishihara, 207 Uehara, Okinawa 903-0215, Japan. [33]National Science and Technology Development Agency, 111 Thailand Science Park, Pathumtani 12120, Thailand. [34]Monash University (Sunway Campus), Jalan Lagoon Selatan, 46150 Bandar Sunway, Selangor, Malaysia. [35]RIKEN Genomic Sciences Center, W502, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan. [36]Laboratory of Functional Genomics, Department of Medical Genome Sciences Graduate School of Frontier Sciences, University of Tokyo (Shirokanedai Laboratory), 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. [37]Chinese Academy of Sciences-Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Rd., Shanghai 200031, China. [38]Genomic Research Branch, National Institute of Mental Health, National Institutes of Health, 6001 Executive Boulevard, Bethesda, MD 20892 USA. [39]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Tobetsu 061-0293, Japan.