Supporting Information

Xu et al. 10.1073/pnas.1118892109

SI Materials and Methods

Populations and Samples. The Pan-Asian dataset consists of one Papuan population from Papua New Guinea and one Melanesian population from Bougainville obtained from the database of the Human Genome Diversity Project–Centre d'Etude du Polymorphisme Humain and 288 unrelated individuals representing 13 Indonesian populations obtained from the Pan-Asia SNP Project (1) (Table 1). The Affymetrix 6.0 dataset consists of 61 individuals from seven populations from Indonesia and one Papuan population from the southern highlands of Papua New Guinea (2, 3). Detailed information on these populations is presented in Table 2.

Markers and Their Positions. In the Pan-Asian dataset, genotype data of 13 Indonesian populations generated using Affymetrix Genechip Human Mapping 50K Xba array were obtained from the Pan-Asia SNP Project (1). Detailed information about data generation, filtration, and data quality control was described elsewhere (1). Genotype data of one Papuan population and one Melanesian population were generated using Illumina Genechip Human Mapping 650K Xba array, and details were described elsewhere (4). All of the analyses in this study used the markers that genotyped in both Pan-Asia and HGDP-CEPH samples. With data integration, we obtained 19,934 SNPs shared by 15 population samples. The physical positions of SNPs were based on the *Homo sapiens* Genome Build 37. The average spacing between adjacent markers was 50.5 kb, with a minimum of 11 bp and a maximum of 26.8 Mb; the median between marker distances was 18.3 kb. In the Affymetrix 6.0 dataset, all samples were genotyped on the Affymetrix 6.0 platform as described previously (2, 3). After data cleaning and integration, there were 685,582 SNPs for analysis.

Statistical Analysis. ***Analysis of the Pan-Asian dataset.*** Principle component analysis was performed at the individual level using EIGENSOFT version 3.0 (5). Unbiased estimates of F_{ST} were calculated using the work by Weir and Hill (6) with PEAS V1.0 (7). Great circle distance calculations followed the approach in the work by Ramachandran et al. (8). The tree of populations was reconstructed based on the F_{ST} distances and the neighbor-joining algorithm (9) implemented in the Molecular Evolutionary Genetics Analysis software package (MEGA version 4.0) (10).

Given the large number of markers in our dataset, genetic analyses can be performed at the level of individual, making no presumption of group membership. We applied a Bayesian cluster analysis as implemented in the STRUCTURE program (11) and a maximum likelihood method as implemented in the *frappe* program (12) to infer the genetic ancestry of individuals. Our approach is solely based on genotype without incorporating any information on sampling location or population affiliation of each individual. We ran STRUCTURE from $K = 1$ to $K = 15$, and 10 repeats were done for each K values. All STRUCTURE runs used 10,000 iterations after a burn-in of length 20,000 with the admixture model and assuming that allele frequencies were correlated (11). The *frappe* program was run for 100,000 iterations from $K = 2$–15 and repeated 10 times for each single K value. According to the distribution of the posterior probability as provided by STRUCTURE and *frappe* analyses, the most probable and appropriate number of clusters should be three in our dataset.

In estimating the admixture time of East Indonesian populations, we selected a panel of 2,807 ancestry informative markers with large allele frequency differences ($F_{ST} > 0.3$) between ID-MT and Papuan and ran STRUCTURE with the linkage model to estimate recombination rates in seven Eastern Indonesian populations. In this model, STRUCTURE reports not only the overall ancestry for each individual but also the probability of origin of each allele. The break points were inferred according to the estimated origin of each allele. The program STRUCTURE was run with 100,000 iterations, 200,000 burn-ins, and 10,000 admixture burn-ins.

Analysis of the Affymetrix 6.0 dataset. Principle component analysis, admixture proportions, and time of admixture estimation were performed using the StepPCO software (13). Individual ancestry components were inferred using a maximum likelihood method as implemented in the *frappe* program (12). We ran analyses for $K = 2$ and $K = 3$ and performed three independent runs for each K value. The analysis of admixture rates on the autosomes vs. X chromosome was based on 36,415 X-linked SNPs. For the time of admixture estimation, Borneo and New Guinea were used as parental groups. To make the sample sizes equal for the two parental groups, 16 individuals were selected at random from the NGH population. To limit the analysis only to variation defined by the parental groups and exclude any signal in the admixed groups that comes from genetic drift or other sources of admixture, the first principle axis was calculated only between the parental groups, and the admixed group was then projected onto this axis (14, 15). The admixture signal along each chromosome was obtained, and the width of the ancestry blocks was estimated as described previously (13). The method is sensitive to small sample sizes, and the admixed Indonesian groups were, therefore, combined into the Nusa Tenggaras group (10 individuals from the islands of Alor, Timor, Roti, and Flores) and the Moluccas (10 individuals from the islands of Hiri and Ternate). The time estimate is based on comparison with the data from 100 forward simulations with a 40% migration rate (13).

1. Abdulla MA, et al. (2009) Mapping human genetic diversity in Asia. *Science* 326: 1541–1545.
2. Reich D, et al. (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89:516–528.
3. Wollstein A, et al. (2010) Demographic history of Oceania inferred from genome-wide data. *Curr Biol* 20:1983–1992.
4. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
5. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190.
6. Weir BS, Hill WG (2002) Estimating F-statistics. *Annu Rev Genet* 36:721–750.
7. Xu S, Gupta S, Jin L (2010) PEAS V1.0: A package for elementary analysis of SNP data. *Mol Ecol Resour* 10:1085–1088.
8. Ramachandran S, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* 102:15942–15947.
9. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.
10. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599.
11. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164: 1567–1587.
12. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol* 28:289–301.
13. Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M (2011) Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol* 12:R19.
14. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* 461:489–494.
15. McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genet* 5:e1000686.

SI Text

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

Mahmood Ameen Abdulla[a], Ikhlak Ahmed[b], Anunchai Assawamakin[c,d], Jong Bhak[e], Samir K. Brahmachari[b], Gayvelline C. Calacal[f], Amit Chaurasia[b], Chien-Hsiun Chen[g], Jieming Chen[h], Yuan-Tsong Chen[g], Jiayou Chu[i], Eva Maria C. Cutiongco-de la Paz[j], Maria Corazon A. De Ungria[f], Frederick C. Delfin[f], Juli Edo[a], Suthat Fucharoen[c],Ho Ghang[e], Takashi Gojobori[k,l], Junsong Han[m], Sheng-Feng Ho[g], Boon Peng Hoh[n], Wei Huang[o], Hidetoshi Inoko[p], Pankaj Jha[b], Timothy A. Jinam[1], Li Jin[q,r], Jongsun Jung[s], Daoroong Kangwanpong[t], Jatupol Kampuansai[t], Giulia C. Kennedy[u,v], Preeti Khurana[w], Hyung-Lae Kim[s], Kwangjoong Kim[s], Sangsoo Kim[x], Woo-Yeon Kim[e], Kuchan Kimm[y], Ryosuke Kimura[z], Tomohiro Koike[k], Supasak Kulawonganunchai[d], Vikrant Kumar[h], Poh San Lai[aa,bb], Jong-Young Lee[s], Sunghoon Lee[e], Edison T. Liu[h], Partha P. Majumder[cc], Kiran Kumar Mandapati[w], Sangkot Marzuki[dd], Wayne Mitchell[ee,ff], Mitali Mukerji[b], Kenji Naritomi[gg], Chumpol Ngamphiw[d], Norio Niikawa[hh], Nao Nishida[z], Bermseok Oh[s], Sangho Oh[e], Jun Ohashi[z], Akira Oka[p], Rick Ong[h], Carmencita D. Padilla[j], Prasit Palittapongarnpim[ii], Henry B. Perdigon[f], Maude Elvira Phipps[a,jj], Eileen Png[h], Yoshiyuki Sakaki[kk], Jazelyn M. Salvador[f], Yuliana Sandraling[dd], Vinod Scaria[b], Mark Seielstad[h], Mohd Ros Sidek[n], Amit Sinha[b], Metawee Srikummool[t], Herawati Sudoyo[dd], Sumio Sugano[ll], Helena Suryadi[dd], Yoshiyuki Suzuki[k], Kristina A. Tabbada[f], Adrian Tan[h], Katsushi Tokunaga[z], Sissades Tongsima[d], Lilian P. Villamor[f], Eric Wang[u,v], Ying Wang[o], Haifeng Wang[o], Jer-Yuarn Wu[g], Huasheng Xiao[m], Shuhua Xu[r], Jin Ok Yang[e], Yin Yao Shugart[mm], Hyang-Sook Yoo[e], Wentao Yuan[o], Guoping Zhao[o], Bin Alwi Zilfalil[n], and Indian Genome Variation Consortium[b]

[a]Department of Molecular Medicine, Faculty of Medicine, and Department of Anthropology, Faculty of Arts and Social Sciences, University of Malaya, Kuala Lumpur, 50603, Malaysia; [b]Council for Scientific and Industrial Research, Institute of Genomics and Integrative Biology, Delhi 110007, India; [c]Mahidol University, Salaya Campus, Puttamonthon, Nakornpathom 73170, Thailand; [d]Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, Pathumtani 12120, Thailand; [e]Korean BioInformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), Yuseong-gu, Deajeon 305–806, Korea; [f]DNA Analysis Laboratory, Natural Sciences Research Institute, University of the Philippines, Diliman, Quezon City 1101, Philippines; [g]Institute of Biomedical Sciences, Academia Sinica, Nangang, Taipei City 115, Taiwan; [h]Genome Institute of Singapore, 138672, Singapore; [i]Institute of Medical Biology, Chinese Academy of Medical Science, Kunming 650118, China; [j]Institute of Human Genetics, National Institutes of Health, University of the Philippines Manila, Ermita Manila 1000, Philippines; [k]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Mishima, Shizuoka 411–8540, Japan; [l]Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo 135–0064, Japan; [m]National Engineering Center for Biochip at Shanghai, Shanghai 201203, China; [n]Human Genome Center, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia; [o]Ministry of Science and Technology-Shanghai Laboratory of Disease and Health Genomics, Chinese National Human Genome Center Shanghai, Shanghai 201203, China; [p]Division of Molecular Medical Science and Molecular Medicine, Department of Molecular Life Science, Tokai University School of Medicine, Isehara-A Kanagawa-Pref A259-1193, Japan; [q]State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China; [r]Chinese Academy of Sciences-Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; [s]Korea National Institute of Health, Eunpyung-Gu, Seoul, 122–701, Korea; [t]Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai 50202, Thailand; [u]Genomics Collaborations, Affymetrix, Santa Clara, CA 95051; [v]Veracyte, South San Francisco, CA 94080; [w]The Centre for Genomic Applications (an Institute of Genomics and Integrative Biology-Institute of Molecular Medicine Collaboration), New Delhi 110020, India; [x]Soongsil University, Dongjak-gu, Seoul 156–743, Korea; [y]Eulji University College of Medicine, Dae-jeon City 301–832, Korea; [z]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Bunkyo-ku, Tokyo 113–0033, Japan; [aa]Department of Pediatrics, Yong Loo Lin School of Medicine, National University of Singapore, National University Hospital, 119074, Singapore; [bb]Population Genetics Laboratory, Defense Medical and Environmental Research Institute, Defence Science Organisation National Laboratories, 117510, Singapore; [cc]Indian Statistical Institute (Kolkata), Kolkata 700108, India; [dd]Eijkman Institute for Molecular Biology, Jakarta 10430, Indonesia; [ee]Informatics Experimental Therapeutic Centre, 03–01 Nanos, 138669, Singapore; [ff]Division of Information Sciences, School of Computer Engineering, Nanyang Technological University, 639798, Singapore; [gg]Department of Medical Genetics, Faculty of Medicine, University of the Ryukyus, Nishihara, Okinawa 903–0215, Japan; [hh]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Tobetsu 061–0293, Japan; [ii]National Science and Technology Development Agency, Pathumtani 12120, Thailand; [jj]Monash University (Sunway Campus), 46150 Bandar Sunway, Selangor, Malaysia; [kk]RIKEN Genomic Sciences Center, Tsurumi-ku, Yokohama 230–0045, Japan; [ll]Laboratory of Functional Genomics, Department of Medical Genome Sciences Graduate School of Frontier Sciences, Shirokanedai Laboratory, University of Tokyo, Minato-ku, Tokyo 108–8639, Japan; and [mm]Genomic Research Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892

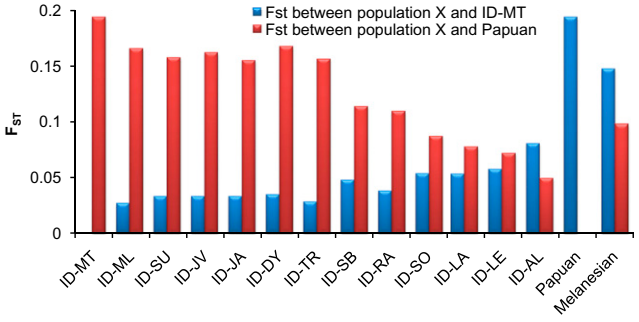

Fig. S1. F_{ST} clines for 15 population samples. Pairwise F_{ST} values of all populations in the Pan-Asian SNP dataset were compared with the easternmost (Papuan) and westernmost (ID-MT) populations, respectively.

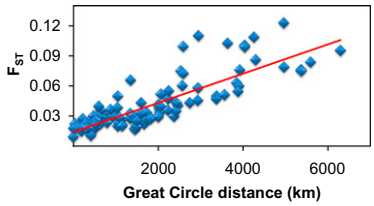

Fig. S2. The relationship between genetic distance (F_{ST}) and geographic distance computed using great circle distances. Blue diamonds represent pairwise population comparisons. Red line is the regression line fitted to the data [$F_{ST} = 1.4 \times 10^{-2} + (1.5 \times 10^{-5}) \times$ (geographic distance in kilometers)]. R^2 for the linear regression of genetic distance on geographic distance is 0.69 (Mantel test, $P < 0.00005$).

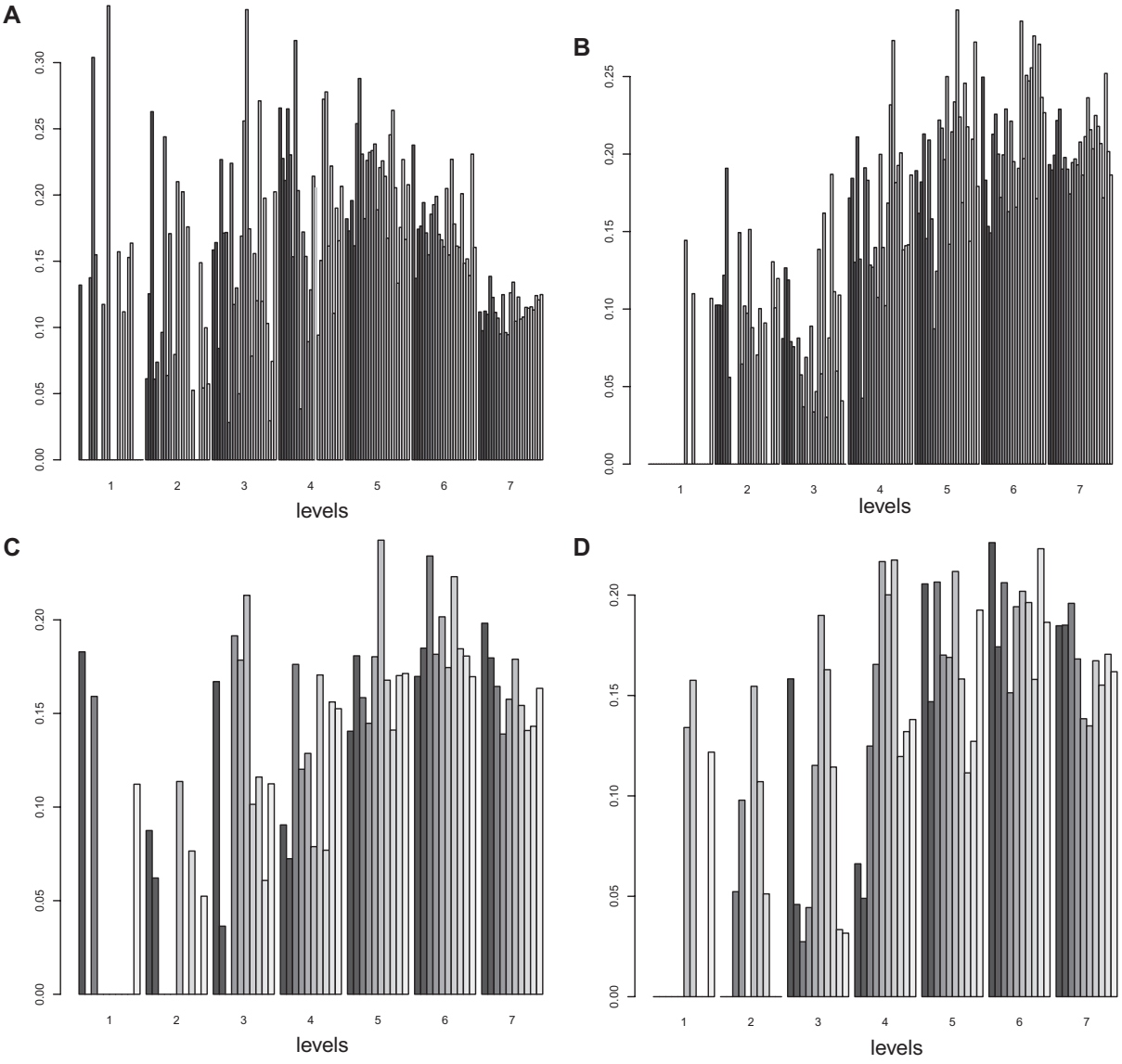

Fig. S3. Distributions of the WT levels illustrating how the wavelet transform spectrum is different for populations with single vs. multiple admixture events. WT levels from four populations are plotted (each bar represents one individual, and each individual has a measurement for each level). The height of the columns indicates the abundance of wavelets of a particular frequency present in the signal, starting with the lowest wave frequencies (widest recombination blocks, representing recent admixture) on the left and progressing to the highest wave frequencies (narrowest recombination blocks, representing older admixture) on the right. (*A*) Distribution of the WT levels for Fiji, a population previously inferred to have experienced multiple admixture events (1, 2). (*B*) Distribution of the WT levels for Polynesia, a population previously inferred to have experienced a single admixture event. (*C*) Distribution of the WT levels for the Nusa Tenggaras. (*D*) Distribution of the WT levels for the Moluccas.

1. Wollstein A, et al. (2010) Demographic history of Oceania inferred from genome-wide data. *Curr Biol* 20:1983–1992.
2. Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M (2011) Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol* 12:R19.

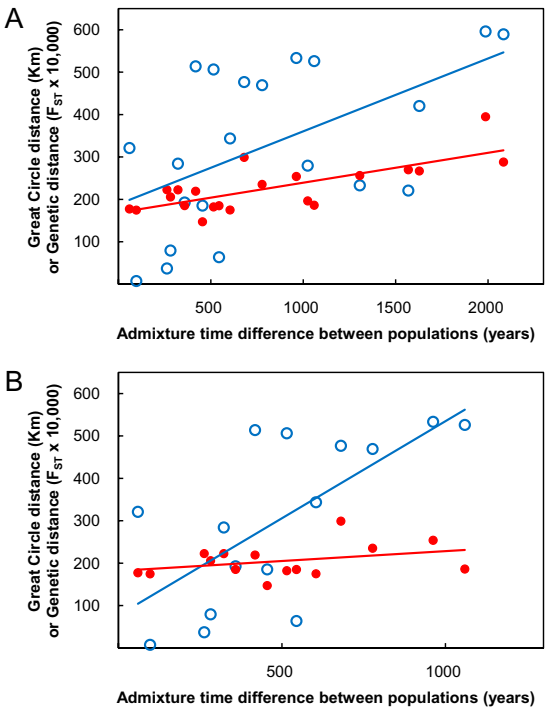

Fig. S4. The relationship between admixture time difference (ATD), geographic distance computed using great circle distance (GCD), and genetic distance (F_{ST}) in the Pan-Asian SNP dataset. (*A*) Data for all seven populations (ID-TR, ID-SB, ID-RA, ID-SO, ID-LA, ID-LE, and ID-AL). Blue circles represent pairwise population comparisons of GCD. Blue line is the regression line fitted to the data of GCD on ATD (GCD = 188.5 + 0.17 × ATD). R^2 for the linear regression of GCD on ATD is 0.30 (P = 0.01). Red dots represent pairwise population comparisons of F_{ST}. Red line is the regression line fitted to the data of F_{ST} on ATD (F_{ST} = 168.9 + 0.07 × ATD). R^2 for the linear regression of F_{ST} on ATD is 0.55 ($P = 1.2 \times 10^{-4}$). (*B*) Data for six populations with ID-TR removed, because ID-TR showed low levels of admixture and wide confidence intervals in admixture time estimation (Fig. 1). Blue circles represent pairwise population comparisons of GCD. Blue line is the regression line fitted to the data of GCD on ATD (GCD = 76.9 + 0.46 × ATD). R^2 for the linear regression of GCD on ATD is 0.45 (P = 0.006). Red dots represent pairwise population comparisons of F_{ST}. Red line is the regression line fitted to the data of F_{ST} on ATD (F_{ST} = 181.8 + 0.05 × ATD). R^2 for the linear regression of F_{ST} on ATD is 0.12 (P = 0.2).

Table S1. Pairwise F_{ST} between populations

	ID-MT	ID-ML	ID-SU	ID-JA	ID-JV	ID-DY	ID-TR	ID-SB	ID-SO	ID-RA	ID-LA	ID-LE	ID-AL	Papuan
ID-ML	0.027													
ID-SU	0.033	0.009												
ID-JA	0.033	0.010	0.001											
ID-JV	0.033	0.010	0.000	0.000										
ID-DY	0.035	0.012	0.011	0.012	0.012									
ID-TR	0.028	0.010	0.012	0.013	0.014	0.016								
ID-SB	0.038	0.018	0.018	0.018	0.019	0.022	0.014							
ID-SO	0.054	0.031	0.028	0.029	0.029	0.035	0.028	0.011						
ID-RA	0.048	0.024	0.022	0.023	0.023	0.027	0.023	0.013	0.016					
ID-LA	0.053	0.033	0.033	0.033	0.034	0.038	0.028	0.010	0.009	0.017				
ID-LE	0.058	0.036	0.036	0.037	0.037	0.040	0.031	0.011	0.010	0.019	0.003			
ID-AL	0.081	0.057	0.056	0.057	0.058	0.062	0.052	0.024	0.017	0.031	0.011	0.008		
Papuan	0.193	0.165	0.157	0.155	0.161	0.167	0.156	0.109	0.087	0.113	0.078	0.072	0.049	
Melanesian	0.147	0.127	0.119	0.119	0.124	0.131	0.119	0.090	0.080	0.094	0.074	0.071	0.066	0.098

Table S2. Admixture proportion of populations (%) estimated from 19,934 SNPs

	Cluster 1	Cluster 2	Cluster 3
ID-MT	99.9	0.0	0.0
ID-ML	98.5	1.0	0.5
ID-SU	98.9	0.7	0.4
ID-JV	99.2	0.6	0.2
ID-JA	99.4	0.4	0.2
ID-DY	99.2	0.5	0.4
ID-TR	94.7	5.1	0.2
ID-SB	76.2	23.7	0.1
ID-RA	75.6	24.1	0.3
ID-SO	64.9	34.9	0.3
ID-LA	60.7	39.0	0.2
ID-LE	57.7	41.9	0.5
ID-AL	43.8	55.4	0.9
Papuan	1.2	98.4	0.4
Melanesian	6.7	47.3	46.0

Cluster 1, Asian; Cluster 2, Papuan; Cluster 3, Melanesian. Note that admixture proportions in the table are the results averaged from 10 independent structure runs; the variation of the estimations from different runs is very small, and SDs are less than 1%.