Table S13 Primers used for DNA re-sequencing.
(DOC)

Table S14 Primers used for construction of expression vectors.
(DOC)

Table S15 Oligonucleotides used for EMSAs and Luciferase assays.
(DOC)

Acknowledgments

We thank K. Kobayashi, M. Kitazato, K. Shimane, and all other members of the Laboratory for Autoimmune Diseases, CGM, RIKEN, for their advice and technical assistance. We also thank the members of BioBank Japan, the Rotary Club of Osaka-Midosuji District 2660 Rotary International, and Dr. Miyatake for supporting sample collection. The replication study of RA was performed under the support of the Genetics and Allied research in Rheumatic diseases Networking (GARNET) consortium.

Author Contributions

Conceived and designed the experiments: K Myouzen, Y Kochi, Y Okada, C Terao, K Ikari, K Ohmura, R Yamada, K Yamamoto. Performed the experiments: K Myouzen, Y Kochi, C Terao, A Suzuki, K Ikari, K Ohmura. Analyzed the data: K Myouzen, Y Kochi, Y Okada, C Terao, T Tsunoda, A Takahashi, R Yamada. Contributed reagents/materials/analysis tools: M Kubo, A Taniguchi, F Matsuda, K Ohmura, S Momohara, T Mimori, H Yamanaka, N Kamatani, Y Nakamura. Wrote the paper: K Myouzen, Y Kochi, Y Okada, C Terao, K Yamamoto.

References

1. Gabriel SE (2001) The epidemiology of rheumatoid arthritis. Rheum Dis Clin North Am 27: 269–281
2. Suzuki A, Yamada R, Chang X, Tokuhiro S, Sawada T, et al. (2003) Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet 34: 395–402
3. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, et al. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study. N Engl J Med 357: 1199–1209
4. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678
5. Gregersen PK, Amos CI, Lee AT, Lu Y, Remmers EF, et al. (2009) REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nat Genet 41: 820–823
6. Kochi Y, Okada Y, Suzuki A, Ikari K, Terao C, et al. (2010) A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. Nat Genet 42: 515–519
7. Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, et al. (2004) A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am J Hum Genet 75: 330–337
8. Adrianto I, Wen F, Templeton A, Wiley G, King JB, et al. (2011) Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nat Genet 43: 253–258
9. Thomas PD, Kejariwal A (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A 101: 15398–15403
10. Okada Y, Shimane K, Kochi Y, Tahira T, Suzuki A, et al. (2012) A Genome-Wide Association Study Identified AFF1 as a Susceptibility Locus for Systemic Lupus Eyrthematosus in Japanese. PLoS Genet 8: e1002455. doi:10.1371/journal.pgen.1002455
11. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42: 295–302
12. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073
13. Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PI, et al. (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet 39: 1477–1482
14. Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, et al. (2007) STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N Engl J Med 357: 977–986
15. Okada Y, Terao C, Ikari K, Kochi Y, Ohmura K, et al. (2012) Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat Genet 45: 511–516
16. Li Z, Nabel GJ (1997) A new member of the I kappaB protein family, I kappaB epsilon, inhibits RelA (p65)-mediated NF-kappaB transcription. Mol Cell Biol 17: 6184–6190
17. Whiteside ST, Epinat JC, Rice NR, Israel A (1997) I kappa B epsilon, a novel member of the I kappa B family, controls RelA and cRel NF-kappa B activity. Embo J 16: 1413–1426
18. Collier FM, Gregorio-King CC, Gough TJ, Talbot CD, Walder K, et al. (2004) Identification and characterization of a lymphocytic Rho-GTPase effector: rhotekin-2. Biochem Biophys Res Commun 324: 1360–1369
19. Collier FM, Loving A, Baker A, J., McLeod J, Walder K, et al. (2009) RTKN2 Induces NF-kappaB Dependent Resistance to Intrinsic Apoptosis in HEK cells and Regulates BCL-2 Gene in Human CD4+ Lymphocytes. J Cell Death 2: 9–23
20. Makarov SS (2001) NF-kappa B in rheumatoid arthritis: a pivotal regulator of inflammation, hyperplasia, and tissue destruction. Arthritis Res 3: 200–206
21. Kolbe D, Taylor J, Elnitski L, Eswara P, Li J, et al. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. Genome Res 14: 700–707
22. Taylor J, Tyekucheva S, King DC, Hardison RC, Miller W, et al. (2006) ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. Genome Res 16: 1596–1604
23. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497–1502
24. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5: 829–834
25. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560
26. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43–49
27. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, et al. (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods 3: 511–518
28. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325: 1246–1250
29. Yang TP, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, et al. (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics 26: 2474–2476
30. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853
31. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–1224
32. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 42: 508–514
33. Chu X, Pan CM, Zhao SX, Liang J, Gao GQ, et al. (2011) A genome-wide association study identifies two new risk loci for Graves' disease. Nat Genet 43: 897–901
34. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43: 1193–1201
35. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR (2011) Low-coverage sequencing: implications for design of complex trait association studies. Genome Res 21: 940–951
36. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482: 390–394
37. Musone SL, Taylor KE, Lu TT, Nititham J, Ferreira RC, et al. (2008) Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. Nat Genet 40: 1062–1064
38. Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, et al. (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet 40: 1216–1223
39. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, et al. (1988) The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 31: 315–324
40. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575
41. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10: 387–406

42. Akamatsu S, Takata R, Ashikawa K, Hosono N, Kamatani N, et al. (2010) A functional variant in NKX3.1 associated with prostate cancer susceptibility down-regulates NKX3.1 expression. Hum Mol Genet 19: 4265–4272

43. Andrews NC, Faller DV (1991) A rapid micropreparation technique for extraction of DNA-binding proteins from limiting numbers of mammalian cells. Nucleic Acids Res 19: 2499

Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population

Yukinori Okada[1–3,40], Chikashi Terao[4,5,40], Katsunori Ikari[6,40], Yuta Kochi[1,2,40], Koichiro Ohmura[5], Akari Suzuki[1], Takahisa Kawaguchi[4], Eli A Stahl[7,8], Fina A S Kurreeman[7–9], Nao Nishida[10], Hiroko Ohmiya[3], Keiko Myouzen[1], Meiko Takahashi[4], Tetsuji Sawada[11], Yuichi Nishioka[12], Masao Yukioka[13], Tsukasa Matsubara[14], Shigeyuki Wakitani[15], Ryota Teshima[16], Shigeto Tohma[17], Kiyoshi Takasugi[18], Kota Shimada[17], Akira Murasawa[19], Shigeru Honjo[20], Keitaro Matsuo[21], Hideo Tanaka[21], Kazuo Tajima[22], Taku Suzuki[6,23], Takuji Iwamoto[6,23], Yoshiya Kawamura[24], Hisashi Tanii[25], Yuji Okazaki[26], Tsukasa Sasaki[27], Peter K Gregersen[28], Leonid Padyukov[29], Jane Worthington[30], Katherine A Siminovitch[31], Mark Lathrop[32,33], Atsuo Taniguchi[6], Atsushi Takahashi[3], Katsushi Tokunaga[10], Michiaki Kubo[34], Yusuke Nakamura[35], Naoyuki Kamatani[36], Tsuneyo Mimori[5], Robert M Plenge[7,8], Hisashi Yamanaka[6], Shigeki Momohara[6,41], Ryo Yamada[37,41], Fumihiko Matsuda[4,38,39,41] & Kazuhiko Yamamoto[1,2,41]

Rheumatoid arthritis is a common autoimmune disease characterized by chronic inflammation. We report a meta-analysis of genome-wide association studies (GWAS) in a Japanese population including 4,074 individuals with rheumatoid arthritis (cases) and 16,891 controls, followed by a replication in 5,277 rheumatoid arthritis cases and 21,684 controls. Our study identified nine loci newly associated with rheumatoid arthritis at a threshold of $P < 5.0 \times 10^{-8}$, including *B3GNT2, ANXA3, CSF2, CD83, NFKBIE, ARID5B, PDE2A-ARAP1, PLD4* and *PTPN2*. *ANXA3* was also associated with susceptibility to systemic lupus erythematosus ($P = 0.0040$), and *B3GNT2* and *ARID5B* were associated with Graves' disease ($P = 3.5 \times 10^{-4}$ and 2.9×10^{-4}, respectively). We conducted a multi-ancestry comparative analysis with a previous meta-analysis in individuals of European descent (5,539 rheumatoid arthritis cases and 20,169 controls). This provided evidence of shared genetic risks of rheumatoid arthritis between the populations.

Rheumatoid arthritis is a complex autoimmune disease characterized by inflammation and the destruction of synovial joints and affects up to 1% of the population worldwide. To date, more than 35 rheumatoid arthritis susceptibility loci, including *HLA-DRB1*, *PTPN22*, *PADI4*, *STAT4*, *TNFAIP3* and *CCR6*, among others, have been identified by GWAS in multiple populations[1–12] and by several meta-analyses of the original GWAS[13–16]. In particular, each meta-analysis of these GWAS uncovered a number of loci that were not identified in the single GWAS, leading to recognition of the enormous power of the meta-analysis approach for detecting causal genes in disease. However, these previous meta-analyses have been performed solely in European populations[13–16] and not in Asian ones. As multi-ancestry studies on validated rheumatoid arthritis susceptibility loci showed the existence of both population-specific and shared genetic components of rheumatoid arthritis[10,17], additional studies in Asian populations might provide useful insight into the underlying genetic architecture of rheumatoid arthritis, which would otherwise be difficult to capture using the studies in a single population. Here, we report a meta-analysis of GWAS and a replication study for rheumatoid arthritis in a Japanese population that was conducted by the Genetics and Allied research in Rheumatic diseases NETworking (GARNET) consortium[10,12]. We subsequently performed a multi-ancestry comparative analysis that incorporated results from a previously conducted meta-analysis of individuals of European ancestry[15].

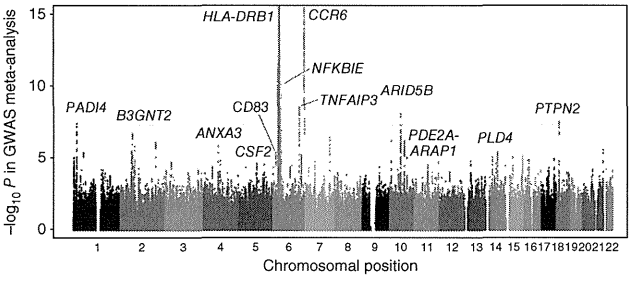

Figure 1 Manhattan plots of the GWAS meta-analysis for rheumatoid arthritis in the Japanese population. The genetic loci that satisfied the genome-wide significance threshold of $P < 5.0 \times 10^{-8}$ (gray line) in the meta-analysis or in the combined study of the meta-analysis and the replication study are presented. The *y* axis shows the $-\log_{10} P$ values of the SNPs in the meta-analysis. The SNPs for which the *P* values were smaller than 1.0×10^{-15} are indicated at the upper limit of the plot.

A full list of author affiliations appears at the end of the paper.

Received 24 October 2011; accepted 1 March 2012; published online 25 March 2012; doi:10.1038/ng.2231

Table 1 Results of the GWAS meta-analysis and the replication studies for rheumatoid arthritis

						Associations in Japanese							Associations in Europeans[c]			
						GWAS meta-analysis			Replication study		Combined study		GWAS meta-analysis			
							Allele 1 freq.						Allele 1 Freq.			
rsID[a]	Chr.	Position (bp)	Cytoband	Gene(s)	Allele 1/2	OR (95% CI)[b]	RA	Control	*P*	OR (95% CI)[b]	*P*	OR (95% CI)[b]	*P*	RA	Control	OR (95% CI)[b]	*P*
SNPs with significant associations ($P < 5.0 \times 10^{-8}$ in the combined study)																	
rs11900673	2	62306165	2p15	*B3GNT2*	T/C	1.15 (1.08–1.21)	0.31	0.28	3.5×10^{-6}	1.09 (1.04–1.14)	6.0×10^{-4}	1.11 (1.07–1.15)	1.1×10^{-8}	0.13	0.13	1.05 (0.98–1.13)	0.17
rs2867461	4	79732239	4q21	*ANXA3*	A/G	1.13 (1.08–1.19)	0.46	0.44	4.7×10^{-6}	1.12 (1.08–1.17)	1.2×10^{-7}	1.13 (1.09–1.17)	1.2×10^{-12}	0.37	0.37	0.98 (0.92–1.04)	0.52
rs657075	5	131458017	5q31	*CSF2*	A/G	1.12 (1.06–1.18)	0.38	0.36	3.2×10^{-5}	1.11 (1.06–1.16)	3.8×10^{-6}	1.12 (1.08–1.15)	2.8×10^{-10}	0.10	0.10	1.04 (0.95–1.13)	0.37
rs12529514	6	14204637	6p23	*CD83*	C/T	1.19 (1.10–1.27)	0.16	0.14	6.8×10^{-6}	1.11 (1.05–1.18)	6.0×10^{-4}	1.14 (1.09–1.19)	2.0×10^{-8}	0.055	0.053	1.11 (0.99–1.24)	0.074
rs2233434	6	44340898	6p21.1	*NFKBIE*	G/A	1.23 (1.16–1.31)	0.24	0.21	9.2×10^{-11}	1.17 (1.11–1.23)	2.2×10^{-9}	1.19 (1.15–1.24)	5.8×10^{-19}	0.059	0.040	1.57 (1.11–2.21)	0.0099
rs10821944	10	63455095	10q21	*ARID5B*	G/T	1.17 (1.11–1.23)	0.39	0.36	1.0×10^{-8}	1.15 (1.10–1.20)	3.0×10^{-10}	1.16 (1.12–1.20)	5.5×10^{-18}	0.29	0.26	1.11 (1.05–1.17)	1.9×10^{-4}
rs3781913	11	72051144	11q13	*PDE2A-ARAP1*	T/G	1.11 (1.05–1.17)	0.71	0.69	3.2×10^{-4}	1.13 (1.08–1.18)	6.7×10^{-7}	1.12 (1.08–1.16)	5.8×10^{-10}	0.45	0.43	1.04 (0.99–1.09)	0.13
rs2841277	14	104462050	14q32	*PLD4*	T/C	1.11 (1.05–1.18)	0.72	0.69	2.8×10^{-4}	1.18 (1.13–1.24)	7.0×10^{-12}	1.15 (1.11–1.19)	1.9×10^{-14}	0.47	0.46	1.02 (0.96–1.09)	0.54
rs2847297	18	12787694	18p11	*PTPN2*	G/A	1.16 (1.11–1.23)	0.37	0.33	3.5×10^{-8}	1.06 (1.01–1.11)	0.013	1.10 (1.07–1.14)	2.2×10^{-8}	0.36	0.34	1.10 (1.05–1.15)	9.2×10^{-5}
SNPs with suggestive associations ($5.0 \times 10^{-8} \leq P < 5.0 \times 10^{-6}$ in the combined study)																	
rs4937362	11	127997949	11q24	*ETS1-FLI1*	T/C	1.13 (1.07–1.19)	0.71	0.68	2.0×10^{-5}	1.07 (1.02–1.12)	0.0061	1.09 (1.06–1.13)	7.5×10^{-7}	0.46	0.44	1.06 (1.01–1.11)	0.015
rs3783637	14	54417868	14q22	*GCH1*	C/T	1.13 (1.07–1.20)	0.76	0.74	6.5×10^{-5}	1.07 (1.02–1.13)	0.0062	1.10 (1.06–1.14)	2.0×10^{-6}	0.88	0.88	0.99 (0.88–1.11)	0.87
rs1957895	14	60978085	14q23	*PRKCH*	G/T	1.12 (1.06–1.18)	0.40	0.39	4.1×10^{-5}	1.07 (1.02–1.12)	0.0022	1.09 (1.05–1.13)	3.6×10^{-7}	0.093	0.089	1.01 (0.95–1.07)	0.73
rs6496667	15	88694672	15q26	*ZNF774*	A/C	1.13 (1.07–1.19)	0.38	0.35	4.7×10^{-5}	1.07 (1.02–1.11)	0.0050	1.09 (1.05–1.13)	1.4×10^{-6}	0.21	0.20	1.07 (1.01–1.13)	0.031
rs7404928	16	23796341	16p12	*PRKCB1*	T/C	1.13 (1.07–1.19)	0.65	0.62	1.5×10^{-5}	1.05 (1.01–1.10)	0.026	1.08 (1.05–1.12)	4.0×10^{-6}	0.75	0.75	1.01 (0.94–1.09)	0.79
rs2280381	16	84576134	16q24	*IRF8*	T/C	1.16 (1.08–1.25)	0.86	0.84	1.0×10^{-4}	1.09 (1.03–1.15)	0.0049	1.12 (1.07–1.17)	2.4×10^{-6}	0.62	0.60	1.05 (0.99–1.11)	0.081
SNPs in previously reported rheumatoid arthritis susceptibility loci ($P < 5.0 \times 10^{-8}$ in the GWAS)																	
rs766449	1	17547439	1p36	*PADI4*	T/C	1.17 (1.11–1.24)	0.44	0.40	4.6×10^{-8}	–	–	–	–	0.38	0.37	1.09 (1.03–1.05)	0.0022
rs2157337	6	32609122	6p21.3	*HLA-DRB1*	C/T	1.99 (1.88–2.11)	0.59	0.44	2.6×10^{-118}	–	–	–	–	0.69	0.46	2.50 (2.39–2.62)	$< 1.0 \times 10^{-300}$
rs6932056	6	138284130	6q23	*TNFAIP3*	C/T	1.35 (1.23–1.49)	0.092	0.073	3.2×10^{-9}	–	–	–	–	0.044	0.034	1.41 (1.24–1.60)	1.3×10^{-7}
rs1571878	6	167460832	6q27	*CCR6*	C/T	1.31 (1.24–1.39)	0.54	0.48	3.2×10^{-19}	–	–	–	–	0.47	0.43	1.13 (1.08–1.19)	5.9×10^{-7}

Chr., chromosome; Freq., frequency; RA, rheumatoid arthritis; OR, odds ratio; CI, confidence interval.
[a]SNPs with $P < 5.0 \times 10^{-6}$ in the combined study of the GWAS meta-analysis and the replication study or SNPs with $P < 5.0 \times 10^{-8}$ in the GWAS meta-analysis are annotated according to forward strand and NCBI Build 36.3. Full results of the replication study are provided in **Supplementary Table 3**. [b]Odds ratio of allele 1. [c]Associations in the previous meta-analysis in European populations[15].

The meta-analysis included 4,074 rheumatoid arthritis cases (with 81.4% and 80.4% of the subjects being positive for antibody to cyclic citrullinated peptide (anti-CCP) and rheumatoid factor, respectively) and 16,891 controls from three GWAS of Japanese subjects (from the BioBank Japan Project[10,18], Kyoto University[12] and the Institute of Rheumatology Rheumatoid Arthritis (IORRA)[19]; **Supplementary Table 1**). After the application of stringent quality control criteria, including principal-component analysis (PCA; **Supplementary Fig. 1**) for each GWAS, the meta-analysis was conducted by evaluating ~2.0 million autosomal SNPs with minor allele frequencies (MAFs) ≥0.01, which were obtained through whole-genome imputation of genotypes on the basis of the HapMap Phase 2 East Asian panels (Japanese in Tokyo (JPT) and Han Chinese in Beijing (CHB)). The inflation factor of the test statistics in the meta-analysis λ_{GC} was as low as 1.036, suggesting no substantial effects of population structure (**Supplementary Table 2**). The quantile-quantile plot of *P* values showed a marked discrepancy in the values in its tail from those anticipated under the null hypothesis that there is no association—even after removal of the SNPs located in the human leukocyte antigen (HLA) region, the major rheumatoid arthritis susceptibility locus—thereby showing the presence of significant associations in the meta-analysis (**Supplementary Fig. 2**).

We identified seven loci in the current meta-analysis that satisfied the genome-wide significance threshold of $P < 5.0 \times 10^{-8}$. These included previously known rheumatoid arthritis susceptibility loci, such as *PADI4* at 1p36, *HLA-DRB1* at 6p21.3, *TNFAIP3* at 6q23 and *CCR6* at 6q27 (refs. 1,3,6,10,15) (the smallest $P = 2.6 \times 10^{-118}$ was found at the *HLA-DRB1* locus; **Fig. 1** and **Table 1**). To our knowledge, the other three loci identified, *NFKBIE* at 6p21.1, *ARID5B* at 10q21 and *PTPN2* at 18p11, are newly associated ($P = 9.2 \times 10^{-11}$, 1.0×10^{-8} and 3.5×10^{-8}, respectively).

To validate the associations identified in the meta-analysis, we conducted a replication study of two independent Japanese rheumatoid arthritis case-control cohorts (cohort 1: 3,830 rheumatoid arthritis cases and 17,920 controls, cohort 2: 1,447 rheumatoid arthritis cases and 3,764 controls; **Supplementary Table 1**). To increase the number of subjects and enhance statistical power, genotype data obtained from other GWAS projects conducted for non-autoimmune diseases in Japanese using Illumina platforms were used for the replication control panels. For each of the 46 loci that exhibited $P < 5.0 \times 10^{-4}$ in

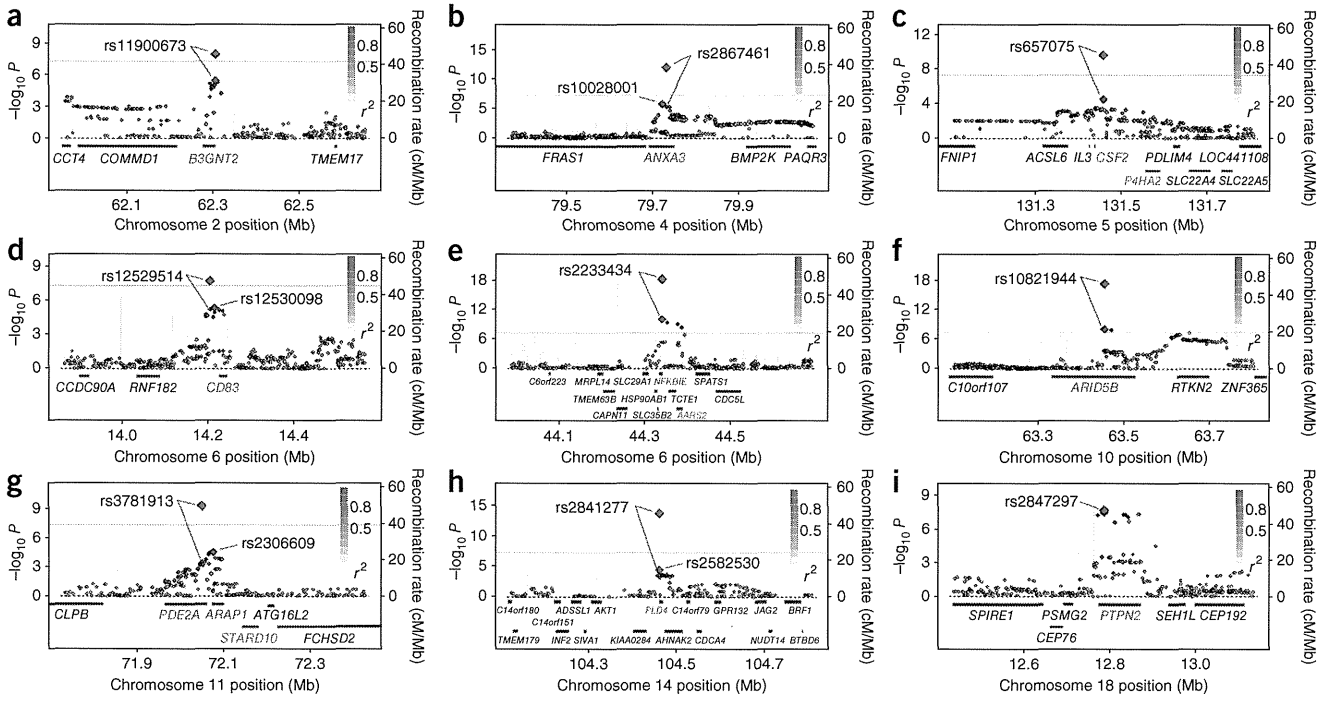

Figure 2 Regional plots of the loci newly associated with rheumatoid arthritis at the genome-wide significance threshold of $P < 5.0 \times 10^{-8}$ in the combined study of the meta-analysis and the replication study. (**a–i**) Regional plots are shown at *B3GNT2* (**a**), *ANXA3* (**b**), *CSF2* (**c**), *CD83* (**d**), *NFKBIE* (**e**), *ARID5B* (**f**), *PDE2A-ARAP1* (**g**), *PLD4* (**h**) and *PTPN2* (**i**). Diamonds represent the $-\log_{10} P$ values of the SNPs, and the red diamonds represent the $-\log_{10} P$ values of the SNPs in the meta-analysis. Red color for the smaller circles represents the r^2 value with the most significantly associated SNP (larger red circle). The purple circle represents the *P* value in the combined study. The blue line shows the recombination rates given by the HapMap Phase 2 east Asian populations (release 22). RefSeq genes at the loci are indicated below. Genes nearest to the marker SNPs at the loci are colored blue (**Supplementary Note**), and genes implicated in eQTL analysis are colored red (**Supplementary Table 4**). At 11q13, two genes (*PDE2A* and *ARAP1*) that are nearest to the SNP selected for the replication study and the most significant SNP in the meta-analysis are highlighted. The plots were drawn using SNP Annotation and Proxy Search (SNAP) version 2.2.

the meta-analysis and had not been reported as rheumatoid arthritis susceptibility loci[1–16], we selected a marker SNP for the replication study (Online Methods and **Supplementary Table 3**).

In the combined analyses of the meta-analysis and the replication study, including a total of 9,351 rheumatoid arthritis cases and 38,575 controls, we identified six newly associated loci, in addition to the *NFKBIE*, *ARID5B* and *PTPN2* loci, that satisfied the significance threshold of $P < 5.0 \times 10^{-8}$, including *B3GNT2* at 2p15, *ANXA3* at 4q21, *CSF2* at 5q31, *CD83* at 6p23, *PDE2A-ARAP1* at 11q13 and *PLD4* at 14q32 (**Figs. 1** and **2** and **Table 1**). Of these loci, *NFKBIE* had the smallest *P* value (5.8×10^{-19}). Although association with rheumatoid arthritis has been described for the *CSF2* and *PTPN2* loci[11,15,16,20,21], ours is the first report to our knowledge validating these associations with a threshold of $P < 5.0 \times 10^{-8}$. Suggestive associations were also observed in *ETS1-FLI1* at 11q24, *GCH1* at 14q22, *PRKCH* at 14q23, *ZNF774* at 15q26, *PRKCB1* at 16p12 and *IRF8* at 16q24 ($5.0 \times 10^{-8} \leq P < 5.0 \times 10^{-6}$). A summary of the genes in the newly associated loci and the results of *cis* expression quantitative trait locus (*cis* eQTL) analysis of the marker SNPs are provided (**Supplementary Table 4** and **Supplementary Note**).

Previous studies have reported associations of rheumatoid arthritis susceptibility loci with other autoimmune diseases[4,10,15,16]. Therefore, we assessed the association of these newly identified susceptibility loci with systemic lupus erythematosus (SLE) by examining the results of an SLE GWAS in the Japanese population (891 cases and 3,384 controls)[22] and in Graves' disease by genotyping 1,783 cases[10] (the controls from the SLE analysis were used for testing for Graves' disease). We observed significant associations of the *ANXA3* locus with SLE and of the *B3GNT2* and *ARID5B* loci with Graves' disease, which showed the same directional effects of the alleles as in rheumatoid arthritis ($P < 0.05/9 = 0.0056$, Bonferroni correction of the number of loci; **Supplementary Table 5**). It should be noted that relatively small sample sizes in the SLE and Graves' disease cohorts might yield limited statistical power, and further evaluations enrolling larger numbers of subjects would be desirable.

To highlight genetic backgrounds of rheumatoid arthritis that are common and divergent in different ancestry groups, we conducted a multi-ancestry comparative analysis of the present study in Japanese and a previous GWAS meta-analysis in Europeans that included 5,539 rheumatoid arthritis cases and 20,169 controls[15] (**Fig. 3a–c**). First, we compared associations in the reported[1–16] or newly identified rheumatoid arthritis susceptibility loci (**Fig. 3a** and **Supplementary Table 6**). Of the 46 rheumatoid arthritis risk variants evaluated, 6 were monomorphic in Japanese, and all were polymorphic in Europeans. We observed significant associations at 22 loci in Japanese and at 36 loci in Europeans (false discovery rate (FDR) < 0.05, $P < 0.0030$), with 14 loci being shared between the populations. Of the newly associated rheumatoid arthritis susceptibility loci identified in our Japanese meta-analysis, significant associations were also observed in the European meta-analysis at the *ARID5B* and *PTPN2* loci ($P = 1.9 \times 10^{-4}$ and 9.2×10^{-5}, respectively; **Table 1**). Significant positive correlation of odds ratios was observed between the studies ($r = 0.822$, $P = 8.1 \times 10^{-11}$; **Fig. 3b**), suggesting that a substantial proportion of genetic factors are shared between

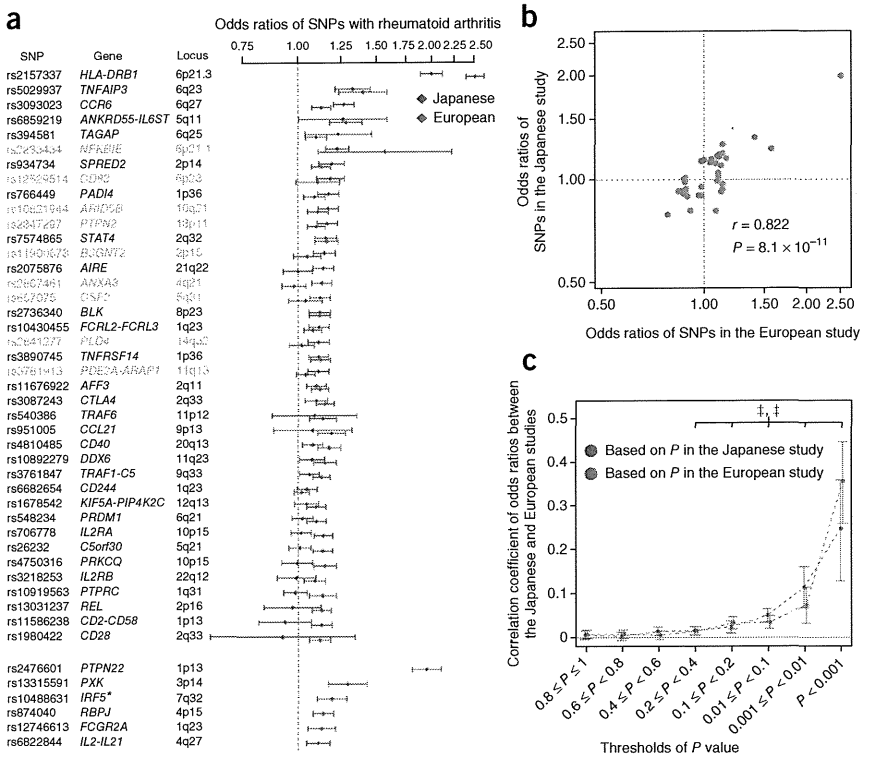

Figure 3 Overlap of the associations with rheumatoid arthritis between Japanese and European populations. (**a**) Forest plots of SNPs in the rheumatoid arthritis susceptibility loci (**Supplementary Table 6**). We selected the genetic loci that have been validated to be associated with rheumatoid arthritis susceptibility by showing associations in the reports of multiple cohorts or satisfying the genome-wide significant threshold ($P < 5.0 \times 10^{-8}$) in previous studies, including in the meta-analysis and replication phases[1–16]. For each of the loci, the most significant SNP among those reported in the previous or present study were selected[1–16]. SNPs in the newly identified rheumatoid arthritis susceptibility loci are colored green. Odds ratios and 95% confidence interval (CI) values are based on rheumatoid arthritis risk alleles, and the SNPs are ordered according to the odds ratios in the Japanese study. Several SNPs were monomorphic in the Japanese population. The odds ratios of these SNPs in the European study are presented below. The asterisk indicates that an association of another variant at the *IRF5* locus was reported in the Japanese population[24]. (**b**) Correlation of the odds ratios of the SNPs in the validated rheumatoid arthritis susceptibility loci between the two populations. SNPs that were polymorphic in both populations were used; odds ratios were based on the minor allele in the Japanese population. (**c**) Correlation of the odds ratios of the genome-wide SNPs, excluding the rheumatoid arthritis susceptibility loci. Correlations were evaluated for sets of SNPs stratified by the thresholds based on the meta-analysis *P* values in each population after pruning of the SNPs by LD ($r^2 < 0.3$). Correlation coefficient and 95% CI are indicated on the *y* axis. Significant correlation of the odds ratios was observed (‡, $P < 0.005$), even for the SNPs that showed moderate associations with rheumatoid arthritis (meta-analysis $P < 0.4$ in each population).

the two ancestry groups[17]. When the rheumatoid arthritis cases of the Japanese GWAS meta-analysis were stratified into anti-CCP–positive or rheumatoid factor–positive cases ($n = 3,209$) and controls ($n = 16,891$), similar results were observed (data not shown). Nevertheless, most of the SNPs assessed here are not necessarily causal variants, and further fine mapping of the loci is warranted to precisely evaluate the shared genetic predisposition between the populations.

Next, we compared regional associations within each of the loci and identified unique patterns in the *ARID5B* locus at 10q21 (**Supplementary Fig. 3**). In Japanese, three peaks of association were observed ($P = 1.0 \times 10^{-8}$ at rs10821944, $P = 5.7 \times 10^{-8}$ at rs10740069 and $P = 8.5 \times 10^{-6}$ at rs224311). These three variants were in weak linkage disequilibrium (LD) in Japanese ($r^2 < 0.10$), indicating independent associations with each of the other SNPs that satisfied a region-wide significance threshold of $P < 3.5 \times 10^{-5}$ (conditional $P = 4.3 \times 10^{-6}$, 1.7×10^{-5} and 1.8×10^{-5}, respectively) (**Supplementary Fig. 3**). In contrast, there was only one peak of association in Europeans ($P = 1.2 \times 10^{-6}$ at rs12764378; $r^2 = 0.59$ with rs10821944 in Europeans), and no additional association was observed in conditional analysis with rs12764378 (the smallest conditional $P = 2.2 \times 10^{-4}$), suggesting that the number of independent associations may be different at this locus in the two populations.

Finally, we conducted polygenic assessment for common variants showing modest associations to rheumatoid arthritis (those not meeting the genome-wide association threshold). This approach has been recognized to be a means to explain a substantial proportion of genetic risk[23]. For the SNPs that were shared between the two meta-analyses but not included in the validated rheumatoid arthritis susceptibility loci, we adopted LD pruning of the SNPs ($r^2 < 0.3$). We then evaluated the correlation of odds ratios of the SNPs between the two meta-analyses and observed a significant positive correlation ($r = 0.023$, $P < 1.0 \times 10^{-300}$). When the SNPs were stratified according to the *P* values in each meta-analysis, significant positive correlations of odds ratios were observed for the SNPs, even for those showing modest association ($P < 0.4$ in the meta-analysis of Japanese or Europeans; $r = 0.014$–0.36 for each *P* value range, $P < 0.005$ for each correlation test) (**Fig. 3c**). Correlations (r) of odds ratios observed herein suggest substantial overlap of the genetic risk of rheumatoid arthritis between the two populations, not only in the validated rheumatoid arthritis susceptibility loci but also at the loci showing nonsignificant associations. This suggests the usefulness of a meta-analysis approach involving multiple ancestry groups in identifying additional susceptibility loci.

In summary, we identified multiple new loci associated with rheumatoid arthritis through a large-scale meta-analysis of GWAS in Japanese. Multi-ancestry comparative analysis provided evidence of significant overlap in the genetic risks of rheumatoid arthritis between Japanese and Europeans. Thus, findings from the present study should contribute to the further understanding of the etiology of rheumatoid arthritis.

URLs. GARNET consortium, http://www.twmu.ac.jp/IOR/garnet/home.html; The BioBank Japan Project (in Japanese), http://biobankjp.org/; International HapMap Project, http://www.hapmap.org/; PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/; EIGENSTRAT, http://genepath.med.harvard.edu/~reich/Software.htm; MACH and mach2dat, http://www.sph.umich.edu/csg/abecasis/MACH/index.

html; R statistical software, http://cran.r-project.org/; SNAP, http://www.broadinstitute.org/mpg/snap/index.php; NCBI GEO database, http://www.ncbi.nlm.nih.gov/geo/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors acknowledge the essential role of the GARNET consortium in developing the study. In this study, the following GARNET members are included: CGM of RIKEN, University of Tokyo, the BioBank Japan Project, Kyoto University and IORRA. We would like to thank all the doctors and staff who participated in sample collection for the RIKEN cohort and the BioBank Japan Project. We thank K. Kobayashi and M. Kitazato for their technical assistance. We thank T. Raj for calculation of composite of multiple signals (CMS). We thank M. Kokubo for DNA extraction, GWAS genotyping and secretarial assistance. We would also like to thank H. Yoshifuji, N. Yukawa, D. Kawabata, T. Nojima, T. Usui and T. Fujii for collecting DNA samples. We thank Y. Katagiri for her technical efforts. We also appreciate the contribution of E. Inoue and other members of the Institute of Rheumatology, Tokyo Women's Medical University, for their efforts on the IORRA cohort. This study was supported in part by grants-in-aid from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan, the Ministry of Health, Labour and Welfare (MHLW) in Japan, the Japan Society for the Promotion of Science (JSPS), Core Research for Evolutional Science and Technology (CREST), Solution-Oriented Research for Science and Technology (SORST), INSERM and the Okawa Foundation for Information and Telecommunications.

AUTHOR CONTRIBUTIONS

Y. Okada, C.T., K.I., Y. Kochi and K.O. designed the study and drafted the manuscript. Y. Okada, C.T., K.I., T.K., H.O., N.N., M.T., M.L., K. Tokunaga and M.K. managed genotyping and manipulation of GWAS data. Y. Okada, Y. Kochi, C.T. and K.I. managed genotyping of replication cohorts. Y. Okada, T.K., H.O., E.A.S., A. Takahashi and R.Y. performed statistical analysis. Y. Kochi, A.S., K. Myouzen, T. Sawada, Y. Nishoka, M.Y., T. Matsubara, S.W., R.T. and S.T. collected samples and managed phenotype data for the rheumatoid arthritis cohorts from the BioBank Japan Project and CGM, RIKEN. C.T., K.O., T.K., M.T., K. Takasugi, K.S., A.M., S.H., K. Matsuo, H. Tanaka, K. Tajima and M.L. collected samples and managed phenotype data for the rheumatoid arthritis cohorts from Kyoto University. K.I., T. Suzuki, T.I., Y. Kawamura, H. Tanii, Y. Okazaki and T. Sakaki collected samples and managed phenotype data for the rheumatoid arthritis cohorts from IORRA. Y. Kochi managed the data for the SLE and Graves' disease cohorts. A.S., C.T. and K.I. analyzed the sera of subjects with rheumatoid arthritis. E.A.S., F.A.S.K., P.K.G., J.W., K.A.S., L.P. and R.M.P. managed the data for the rheumatoid arthritis cohorts in European populations. A. Taniguchi, A. Takahashi, K. Tokunaga, M.K., Y. Nakamura, N.K., T. Minori, R.M.P., H.Y., S.M., R.Y., F.M. and K.Y. supervised the overall study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Suzuki, A. *et al.* Functional haplotypes of *PADI4*, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* **34**, 395–402 (2003).
2. Kochi, Y. *et al.* A functional variant in *FCRL3*, encoding Fc receptor–like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nat. Genet.* **37**, 478–485 (2005).
3. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
4. Remmers, E.F. *et al. STAT4* and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N. Engl. J. Med.* **357**, 977–986 (2007).
5. Plenge, R.M. *et al. TRAF1–C5* as a risk locus for rheumatoid arthritis—a genomewide study. *N. Engl. J. Med.* **357**, 1199–1209 (2007).
6. Plenge, R.M. *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* **39**, 1477–1482 (2007).
7. Barton, A. *et al.* Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat. Genet.* **40**, 1156–1159 (2008).
8. Suzuki, A. *et al.* Functional SNPs in *CD244* increase the risk of rheumatoid arthritis in a Japanese population. *Nat. Genet.* **40**, 1224–1229 (2008).
9. Gregersen, P.K. *et al. REL*, encoding a member of the NF-κB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat. Genet.* **41**, 820–823 (2009).
10. Kochi, Y. *et al.* A regulatory variant in *CCR6* is associated with rheumatoid arthritis susceptibility. *Nat. Genet.* **42**, 515–519 (2010).
11. Freudenberg, J. *et al.* Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. *Arthritis Rheum.* **63**, 884–893 (2011).
12. Terao, C. *et al.* The human *AIRE* gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. *Hum. Mol. Genet.* **20**, 2680–2685 (2011).
13. Raychaudhuri, S. *et al.* Common variants at *CD40* and other loci confer risk of rheumatoid arthritis. *Nat. Genet.* **40**, 1216–1223 (2008).
14. Raychaudhuri, S. *et al.* Genetic variants at *CD28*, *PRDM1* and *CD2/CD58* are associated with rheumatoid arthritis risk. *Nat. Genet.* **41**, 1313–1318 (2009).
15. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
16. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).
17. Kurreeman, F. *et al.* Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am. J. Hum. Genet.* **88**, 57–69 (2011).
18. Nakamura, Y. The BioBank Japan Project. *Clin. Adv. Hematol. Oncol.* **5**, 696–697 (2007).
19. Yamanaka, H. *et al.* Influence of methotrexate dose on its efficacy and safety in rheumatoid arthritis patients: evidence based on the variety of prescribing approaches among practicing Japanese rheumatologists in a single institute-based large observational cohort (IORRA). *Mod. Rheumatol.* **17**, 98–105 (2007).
20. Yamada, R. *et al.* Association between a single-nucleotide polymorphism in the promoter of the human interleukin-3 gene and rheumatoid arthritis in Japanese patients, and maximum-likelihood estimation of combinatorial effect that two genetic loci have on susceptibility to the disease. *Am. J. Hum. Genet.* **68**, 674–685 (2001).
21. Tokuhiro, S. *et al.* An intronic SNP in a RUNX1 binding site of *SLC22A4*, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat. Genet.* **35**, 341–348 (2003).
22. Okada, Y. *et al.* A genome-wide association study identified *AFF1* as a susceptibility locus for systemic lupus eyrthematosus in Japanese. *PLoS Genet.* **8**, e1002455 (2012).
23. Stranger, B.E., Stahl, E.A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**, 367–383 (2011).
24. Shimane, K. *et al.* A single nucleotide polymorphism in the *IRF5* promoter region is associated with susceptibility to rheumatoid arthritis in the Japanese patients. *Ann. Rheum. Dis.* **68**, 377–383 (2009).

[1]Laboratory for Autoimmune Diseases, Center for Genomic Medicine (CGM), RIKEN, Yokohama, Japan. [2]Department of Allergy and Rheumatology, Graduate School of Medicine, University of Tokyo, Tokyo, Japan. [3]Laboratory for Statistical Analysis, CGM, RIKEN, Yokohama, Japan. [4]Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan. [5]Department of Rheumatology and Clinical Immunology, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [6]Institute of Rheumatology, Tokyo Women's Medical University, Tokyo, Japan. [7]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [8]Broad Institute, Cambridge, Massachusetts, USA. [9]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands. [10]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan. [11]Department of Rheumatology, Tokyo Medical University Hospital, Tokyo, Japan. [12]Yamanashi Prefectural Central Hospital, Yamanashi, Japan. [13]Department of Orthopaedic Surgery, Yukioka Hospital, Osaka, Japan. [14]Matsubara Mayflower Hospital, Hyogo, Japan. [15]Osaka Minami National Hospital, Osaka, Japan. [16]Department of Orthopedic Surgery, Tottori University, Tottori, Japan. [17]Department of Rheumatology, National Hospital Organization, Sagamihara Hospital, Kanagawa, Japan. [18]Center for Rheumatic Diseases, Dohgo Spa Hospital, Ehime, Japan. [19]Department of Rheumatology, Niigata Rheumatic Center, Niigata, Japan. [20]Saiseikai Takaoka Hospital, Toyama, Japan. [21]Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Aichi, Japan. [22]Aichi Cancer Center Hospital and Research Institute, Aichi, Japan. [23]Department of Orthopaedic Surgery, Keio University, Tokyo, Japan. [24]Yokohama Clinic, Warakukai Medical Corporation, Yokohama, Japan. [25]Department of Psychiatry, Mie University School of Medicine, Mie, Japan. [26]Metropolitan Matsuzawa Hospital, Tokyo, Japan. [27]Graduate School of Education, University of Tokyo, Tokyo, Japan. [28]The Feinstein Institute for Medical Research, North Shore–Long Island Jewish Health System, Manhasset, New York, USA. [29]Rheumatology Unit,

Department of Medicine in Solna, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden. [30]Arthritis Research Campaign–Epidemiology Unit, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK. [31]Division of Medicine, University of Toronto, Mount Sinai Hospital and University Health Network, Toronto, Ontario, Canada. [32]Commisariat a l'Energie Atomique (CEA), Institut Genomique, Centre National de Genotypage, Evry, France. [33]Fondation Jean Dausset, Centre d'Etude du Polymorphisme Humain, Paris, France. [34]Laboratory for Genotyping Development, CGM, RIKEN, Yokohama, Japan. [35]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan. [36]Laboratory for International Alliance, CGM, RIKEN, Yokohama, Japan. [37]Unit of Statistical Genetics, Center for Genomic Medicine Graduate School of Medicine Kyoto University, Kyoto, Japan. [38]Core Research for Evolutional Science and Technology (CREST) Program, Japan Science and Technology Agency, Kawaguchi, Japan. [39]Institut National de la Santé et de la Recherche Médicale (INSERM), Unité U852, Kyoto University Graduate School of Medicine, Kyoto, Japan. [40]These authors contributed equally to this work. [41]These authors jointly directed this work. Correspondence should be addressed to Y.K. (ykochi@src.riken.jp) or K.O. (ohmurako@kuhp.kyoto-u.ac.jp).

ONLINE METHODS

Subjects. The Japanese participants in the meta-analysis (4,074 rheumatoid arthritis cases and 16,891 controls) and the replication study (5,277 rheumatoid arthritis cases and 21,684 controls) were obtained through the collaborations of the GARNET consortium (**Supplementary Table 1**)[10,12]. The meta-analysis was conducted on three independent GWAS (from the BioBank Japan Project[18] with 2,414 rheumatoid arthritis cases and 14,245 controls[10], Kyoto University with 1,237 rheumatoid arthritis cases and 2,087 controls[12] and IORRA[19] with 423 rheumatoid arthritis cases and 559 controls). The replication study consisted of two independent cohorts (cohort 1 included 3,830 rheumatoid arthritis cases and 17,920 controls, and cohort 2 included 1,447 rheumatoid arthritis cases and 3,764 controls). We employed a case-control cohort of SLE (891 cases and 3,384 controls)[22] and 1,783 cases with Graves' disease[10]. Details of 5,539 rheumatoid arthritis cases and 20,169 controls included in the meta-analysis in European populations were described elsewhere[15]. All participants provided written informed consent for participation in the study, as approved by the ethical committees of the institutional review boards. Detailed descriptions of the participating subjects are provided (**Supplementary Note**).

Genotyping and quality control in the GWAS. Genotyping platforms and quality control criteria for the GWAS, including cutoff values for sample call rates, SNP call rates, MAF and Hardy-Weinberg *P* values, are given (**Supplementary Table 2**). For the subjects enrolled in each of three GWAS, we excluded closely related subjects with first- or second-degree kinship, which was estimated using PLINK version 1.06 (see URLs). We also excluded the subjects determined to be ancestry outliers from East Asian populations using PCA performed by EIGENSTRAT version 2.0 (see URLs) along with HapMap Phase 2 panels (release 24; **Supplementary Fig. 1**). Genotype imputation was performed on the basis of the HapMap Phase 2 East Asian populations, using MACH version 1.0.16 (see URLs) in a two-step procedure as described elsewhere[25]. We excluded imputed SNPs with MAF < 0.01 or *Rsq* < 0.5 from each of the GWAS. Associations of the SNPs with rheumatoid arthritis were assessed by logistic regression models assuming additive effects of the allele dosages of the SNPs using mach2dat software (see URLs).

Meta-analysis. We included 1,948,139 autosomal SNPs that satisfied quality control criteria in all three GWAS (**Supplementary Table 2**). SNP information was based on a forward strand of the NCBI build 36.3 reference sequence. The meta-analysis was performed using an inverse variance method assuming a fixed-effects model from the study-specific effect sizes (logarithm of odds ratio) and the standard errors of the coded alleles of the SNPs determined with the Java source code implemented by the authors[25]. Genomic control corrections[26] were carried out on test statistics of the GWAS using the study-specific inflation factor (λ_{GC}) and was applied or reapplied to the results of our current meta-analysis (**Supplementary Fig. 2**).

Replication study. We selected a SNP for the replication study from each of the loci that exhibited $P < 5.0 \times 10^{-4}$ in the meta-analysis that had not previously been reported as rheumatoid arthritis susceptibility loci[1–16] (**Supplementary Table 3**). For control subjects, we used genotype data obtained from additional GWAS for non-autoimmune diseases or healthy controls, genotyped using Illumina HumanHap550 BeadChips or HumanHap610-Quad BeadChips, and the cases for rheumatoid arthritis and Graves' disease were genotyped with the TaqMan genotyping system (Applied Biosystems; **Supplementary Table 1**). Selection of the SNP was conducted according to the following criteria: if the SNP with the most significant association in the locus was genotyped in the replication control panel, then that SNP was selected; otherwise, a tag SNP in the replication control panel with the strongest LD was selected (mean $r^2 = 0.89$). For the three SNPs that yielded low call rates (<90%), we alternatively selected proxy SNPs with the second strongest LD. As a result, average genotyping call rates of the SNPs were 99.9% and 99.0% for the controls and cases, respectively. We then evaluated concordance rates between the assayed genotypes by applying these two different methods to samples from 376 subjects who were randomly selected. This procedure yielded high concordance rates of ≥99.9%. Associations of the SNPs were evaluated using logistic regression assuming an additive-effects model of genotypes in R statistical software version 2.11.0 (see URLs). The combined study of the meta-analysis and replication study was performed using an inverse variance method assuming a fixed-effects model[25].

***Cis* eQTL analysis.** For each marker SNP of the newly identified rheumatoid arthritis susceptibility locus, correlations between SNP genotypes and expression levels of genes located 300 kb upstream or downstream of the SNP measured in B-lymphoblastoid cell lines (GSE6536) were evaluated using data from the HapMap Phase 2 east Asian populations[27].

Multi-ancestry analysis of the meta-analyses in Japanese and Europeans. We evaluated the associations of the variants in the validated rheumatoid arthritis susceptibility loci by comparing the results from the current meta-analysis in Japanese with those from a previous meta-analysis in Europeans[15]. We assessed two variants in the *IRF5* locus, where different causal variants were identified in the two populations[24]. For the conditional analysis of the regional associations in the *ARID5B* locus (**Supplementary Fig. 3**), we repeated the meta-analysis at that locus by incorporating genotypes of the referenced SNP(s) as additional covariate(s). For comparison of the odds ratios of the SNPs, we first selected SNPs that were shared between the meta-analyses in Japanese and Europeans. Next, we removed the SNPs located more than 1 Mb away from each of the marker SNPs in the validated rheumatoid arthritis susceptibility loci, except for in the HLA region, where we removed the SNPs located between 24,000,000 bp to 36,000,000 bp on chromosome 6 because of the existence of long-range haplotypes with rheumatoid arthritis susceptibility in this region[28]. LD pruning of the SNPs was conducted for the SNP pairs that were in LD ($r^2 \geq 0.3$) in both HapMap Phase 2 East Asian and Utah residents of Northern and Western European ancestry (CEU) populations (release 24). Correlations of the odds ratios were evaluated using R statistical software version 2.11.0.

25. Okada, Y. *et al.* Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. *PLoS Genet.* **7**, e1002067 (2011).
26. de Bakker, P.I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
27. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
28. Okada, Y. *et al.* Contribution of a haplotype in the HLA region to anti–cyclic citrullinated peptide antibody positivity in rheumatoid arthritis, independently of HLA-DRB1. *Arthritis Rheum.* **60**, 3582–3590 (2009).

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 1, 2012

Pp. 55–67
DOI: 10.1089/cmb.2010.0227

Population Model–Based Inter-Diplotype Similarity Measure for Accurate Diplotype Clustering

RITSUKO ONUKI,[1] RYO YAMADA,[2] RUI YAMAGUCHI,[3]
MINORU KANEHISA,[1] and TETSUO SHIBUYA[3]

ABSTRACT

Classification of the individuals' genotype data is important in various kinds of biomedical research. There are many sophisticated clustering algorithms, but most of them require some appropriate similarity measure between objects to be clustered. Hence, accurate inter-diplotype similarity measures are always required for classification of diplotypes. In this article, we propose a new accurate inter-diplotype similarity measure that we call the population model-based distance (PMD), so that we can cluster individuals with diplotype SNPs data (i.e., unphased-diplotypes) with higher accuracies. For unphased-diplotypes, the allele sharing distance (ASD) has been the standard to measure the genetic distance between the diplotypes of individuals. To achieve higher clustering accuracies, our new measure PMD makes good use of a given appropriate population model which has never been utilized in the ASD. As the population model, we propose to use an hidden Markov model (HMM)–based model. We call the PMD based on the model the HHD (HIT HMM–based Distance). We demonstrate the impact of the HHD on the diplotype classification through comprehensive large-scale experiments over the genome-wide 8930 data sets derived from the HapMap SNPs database. The experiments revealed that the HHD enables significantly more accurate clustering than the ASD.

Key words: algorithms, statistics, strings, suffix trees.

1. INTRODUCTION

SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) are the most fundamental genetic polymorphisms in human genomes (Kim and Misra, 2007), and classification of individuals with the individual SNPs data is very useful in various kinds of biomedical research, especially in population genetics and genetic epidemiology (Conrad et al., 2006; Jakobsson et al., 2008). Accurate classification of individual SNPs data will help study of genotype variations, especially when different genotypes prevail in different populations or subgroups.

There are various sophisticated clustering methods for general data (not limited for clustering SNPs data), many of which (e.g., Ward's method [Team RDC, 2007; Ward, 1963; Ward and Hook, 1963],

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto Japan.
[2]Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan.
[3]Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan.

k-Medoid [Kaufman and Rousseuw, 1990], DBSCAN [Ester et al., 1996], and most of the phylogenetic clustering algorithms such as the famous neighbor joining method [Saitou and Nei, 1987]) require appropriate similarity measures between target objects. Designing accurate similarity measure for the objects to be clustered is essential for these similarity-based clustering algorithms.

For SNPs data, there have been proposed various clustering algorithms for clustering haplotypes (i.e., haplotype-alleles, not diplotypes),[1] and various types of similarity measures have been proposed for haplotype data (Jin et al., 2010; Li and Jiang, 2005; Li et al., 2006).[2] But the human genome is diallelic, and in many cases we observe only the unordered (i.e., unphased) pair of alleles at each locus, instead of ordered (i.e., phased) allele data, due to the high costs required for deciphering unphased allele data to accurate phased ones. In this article, we call a phased pair of haplotypes a "haplotype-diplotype," and we call an unphased pair of haplotypes a "unphased-diplotype."

Much work has been done on clustering the unphased-diplotype data. They can be categorized into two types: distance-based methods (Bowcock et al., 1994; Gao and Starmer, 2007) and statistics-based methods (Falush et al., 2003; Pritchard et al., 2000). The distance-based methods utilize a distance measure between two objects, while statistics-based methods are based on the statistical behavior of objects. In this article, we focus on the distance-based clustering methods for unphased-diplotype data. Most previous distance-based methods utilize a similarity measure called the allele sharing distance (ASD) (Gao and Martin, 2009; Jakobsson et al., 2008; Mao et al., 2007; Witherspoon et al., 2007) (see Section 2.1.1). The ASD is a simple and straightforward extension of the Hamming distance, and is the most standard and frequently used similarity measure between a pair of unphased-diplotypes.

In genetic analysis, it is very important to consider properties of populations that are different among genetically distinct populations (Beaty et al., 2005; Fallin et al., 2001; Witherspoon et al., 2007). It should also be true with designing similarity measures for unphased-diplotypes. But the measure ASD does not utilize any population information in obtaining the similarity values. Thus, in this article, we will first propose a new similarity measure called the population model-based distance (PMD) for unphased-diplotypes, which incorporates the population information from an appropriate population model. As the model, we will propose to use an hidden Markov model (HMM)–based model predicted by a standard HMM-based phasing software called HIT (Rastas et al., 2005). We call the PMD based on the model the HHD (the HIT HMM-based distance). We will show the superiority of our new measure HHD over the previous standard ASD through comprehensive experiments over the genome-wide HapMap data (International HapMap Consortium, 2005).

The organization of this article is as follows. In Section 2, we describe previous work on which our method is based. In Section 3, we describe our new measure. In Section 4, we compare the ASD and the HHD through comprehensive experiments over large-scale HapMap data sets to evaluate the impact of the HHD. In Section 5, we conclude.

1.1. Notations and definitions

We assume all SNPs are diallelic. We consider n diplotypes over m SNP loci from the same chromosome. These loci are numbered $1, 2, \cdots, m$ in the physical order. A SNP-allele for a SNP locus is an element in set $\mathcal{S} = \{1, 0\}$ where 1 and 0 denote the major and minor SNP-alleles, respectively. A haplotype-allele is a sequence of SNP-alleles and is represented by a sequence in $\mathcal{S}^m$ (e.g., $10101 \in \mathcal{S}^5$). A SNP-diplotype for a SNP locus is an unordered pair of SNP-allele in $\mathcal{D} = \mathcal{S} \times \mathcal{S}$(e.g., $\{0, 1\} \in \mathcal{D}$). An unphased-diplotype is a sequence of SNP-diplotype and is represented by a sequence in $\mathcal{D}^m$(e.g., $\{1, 0\} - \{0, 0\} - \{1, 0\} - \{1, 1\} - \{1, 0\} \in \mathcal{D}^5$). Given unphased-diplotypes, the phasing problem is to find the most probable corresponding haplotype-allele pairs that could have generated the unphased-diplotypes. A phased haplotype-allele pair is called a haplotype-diplotype (e.g., {10010, 00111}).

[1]There are also many algorithms proposed for clustering SNP loci (Yang and Tabus, 2007), instead of individuals, but we do not deal with these problems in this article.

[2]Various inter-population distances have also been proposed (Cornuet et al., 1999), but we will not deal with these in this article.

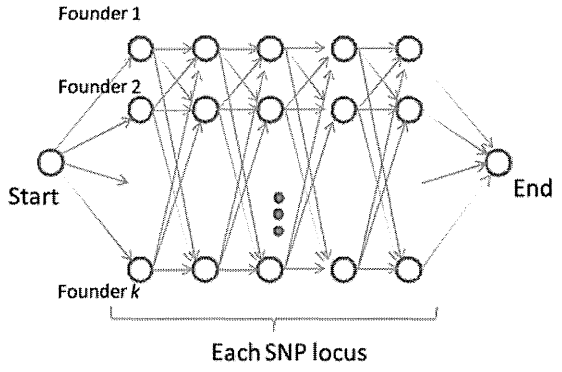

FIG. 1. The HMM model of the HIT. In the HMM, a set of nodes in a row corresponds to states of one founder (i.e., ancestor) haplotype-allele. A set of nodes in a column corresponds to states of one locus. Each node (except for the start and end nodes) emits 1 or 0 with some estimated probabilities, which correspond to the major and minor alleles respectively. A path from the start node to the end node corresponds to a haplotype-allele. The HMM emits a haplotype-diplotype as an unordered pair of two paths from the start node to the end node, randomly based on the probabilities estimated for edges. The observers can only see the unphased-diplotype that corresponds to the emitted haplotype-diplotype.

2. PREVIOUS WORK

In this section, we describe previous work on which our work is based. In Section 2.1, we describe the definitions of measures in previous work (e.g., the ASD). In Section 2.2, we describe the HIT algorithm on which our new distance measure is based. In Section 2.3, we describe a clustering algorithm and an evaluation method for clustering that we will use in the experiments in Section 4.

2.1. *Previous measures for inter-individual genetic distances*

2.1.1. Allele sharing distance. The most standard inter-diplotype distance is the ASD (Gao and Martin, 2009; Jakobsson et al., 2008; Mao et al., 2007; Witherspoon et al., 2007), defined as follows. For two unphased-diplotypes $\mathbf{g}, \mathbf{g}' \in \mathcal{D}^m$ (i.e., m is the number of SNP loci), the ASD between the diplotypes $\mathbf{g}$ and $\mathbf{g}'$ is defined as follows:

$$D(\mathbf{g}, \mathbf{g}') = \frac{1}{2m} \sum_{\ell=1}^{m} d(\mathbf{g}[\ell], \mathbf{g}'[\ell]), \tag{1}$$

where $\mathbf{g}[\ell]$ denotes the ℓ-th SNP-diplotype of unphased-diplotype $\mathbf{g}$, and $d(\mathbf{g}[\ell],\mathbf{g}'[\ell])$ is the number of SNP-alleles which are not shared between $\mathbf{g}$ and $\mathbf{g}'$ at the ℓ-th locus.

2.1.2. Haplotype similarity measure. The most common and simplest measurement for the similarity between DNA sequences, including the haplotype-allele data, is the hamming distance (Cover and Thomas, 1991; Isaev, 2004; Lesk, 2005; Li and Jiang, 2005; Tzeng et al., 2003). For a haplotype-allele $\mathbf{h} \in \mathcal{S}^m$ (where m is the length of $\mathbf{h}$), let $\mathbf{h}[k]$ denote the SNP-allele at the k-th locus of $\mathbf{h}$. The hamming distance between two haplotype-alleles $\mathbf{h}$ and $\mathbf{h}'$ is defined as

$$s(h, h') = \sum_{k=1}^{m} I(\mathbf{h}[k], \mathbf{h}'[k]), \tag{2}$$

where $I(a, b) = 0$ if $a = b$ and $I(a, b) = 1$ otherwise. As the hamming distance is length-dependent, we define the following $A(\mathbf{h}, \mathbf{h}')$ as a length-independent distance between haplotype-alleles $\mathbf{h}$ and $\mathbf{h}'$:

$$A(\mathbf{h}, \mathbf{h}') = \frac{s(\mathbf{h}, \mathbf{h}')}{m}. \tag{3}$$

2.2. *HIT algorithm*

The Haplotype Inference Technique (HIT) algorithm (Rastas et al., 2005) is an HMM-based algorithm for phasing unphased-diplotypes. The algorithm utilizes the HMM (Rabiner and Juang, 1986). The HMM of the HIT is designed to simulate multiple set of ancestors (i.e., founders).[3] The HMM is trained from a set

[3]According to Rastas et al. (2005), the optimal number of ancestors is around 7 for most cases. Thus, we also use the HMM model with 7 ancestors in the experiments in Section 4.

of unphased-diplotypes in an unsupervised way with the EM algorithm (Durbin et al., 1998). Figure 1 shows the HMM model used in the HIT. The HIT algorithm phases an unphased haplotype-diplotype by heuristically finding the haplotype-diplotype with the highest emission probability from the HMM.

2.3. *Clustering methods*

In this section, we describe the clustering method and the method for evaluating the results, which we will use in Section 4.

2.3.1. Ward's method. We use Ward's minimum variance algorithm (Team RDC, 2007; Ward, 1963; Ward and Hook, 1963), which is a widely used hierarchical clustering method, to infer clusters based on the ASD or the HHD in Section 4.[4] Given n items $I_1, I_2, \cdots, I_n$, a distance matrix $\{w_{ij}\}$ where w_{ij} denotes the distance between I_i and I_j, and some fixed positive integer k ($k < n$), the Ward's method clusters the n items into k clusters by the following $n - k - 1$ steps.[5] At first the algorithm considers n clusters each of which contains only 1 item, i.e., $\mathcal{C}_1 = \{\{I_1\}, \{I_2\}, \cdots, \{I_n\}\}$. Then the algorithm reduces the number of clusters one by one in each step as follows. In the m-th step of the algorithm, two clusters are merged into a cluster to minimize $\sum_{C \in \mathcal{C}_{m+1}} \sum_{I_i, I_j \in C} w_{ij}^2 / |C|$, where $\mathcal{C}_i$ denotes the set of clusters before the i-th step of the algorithm. This bottom-up approach is repeated until $|\mathcal{C}_m| = k$.

2.3.2. How to evaluate the clustering results. To evaluate the clustering results, we use the classification error rate (CER) (Gao and Starmer, 2007). The CER is the rate of elements that are assigned to incorrect clusters in clustering results. To know the assignment is correct or not, we need to know the labels of each cluster, but Ward's algorithm does not assign any labels onto the output clusters. In the experiment, we use the minimum CER among all the possible assignments of the population labels, to evaluate the clustering results.

3. NEW UNPHASED-DIPLOTYPE DISTANCE MEASURES

In this section, we first propose in Section 3.1 a new measure for the distance between two unphased-diplotypes, the PMD. The PMD is a general concept of distance measures, and we will give an example of the PMD which we call the HHD in Section 3.2. In Section 3.3, we discuss the properties of the proposed measures.

3.1. *Population model–based distance*

Before defining our new measure called the PMD, we first extend the haplotype similarity measure described in Section 2.1.2 so that we can deal with the distances between two haplotype-diplotypes instead of haplotype-alleles, as follows. Let $a = \{\mathbf{h}_1, \mathbf{h}_2\}$ and $a' = \{\mathbf{h}'_1, \mathbf{h}'_2\}$ be haplotype-diplotypes to be compared, where $\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}'_1, \mathbf{h}'_2 \in \mathcal{S}^m$. We define the distance between haplotype-diplotypes a and a' as

$$H(a, a') = \min\left\{\frac{A(\mathbf{h}_1, \mathbf{h}'_1) + A(\mathbf{h}_2, \mathbf{h}'_2)}{2}, \frac{A(\mathbf{h}_1, \mathbf{h}'_2) + A(\mathbf{h}_2, \mathbf{h}'_1)}{2}\right\}, \tag{4}$$

where A is the haplotype similarity measure defined in Section 2.1.2. But we cannot compute this value for unphased-diplotypes, as we cannot know the actual haplotype-diplotypes. To enable it, we extend the above haplotype-diplotype distance H for unphased-diplotypes by utilizing some given population model $\mathcal{M}$ as follows.

For any unphased-diplotype, we can enumerate corresponding haplotype-diplotype candidates.[6] For example, there are four haplotype-diplotype candidates for unphased-diplotype $\{1, 0\} - \{1, 0\} - \{1, 0\}$, i.e., $\{111, 000\}$, $\{110, 001\}$, $\{101, 010\}$, and $\{011, 011\}$. For unphased-diplotypes $\mathbf{g}, \mathbf{g}' \in \mathcal{D}^m$, let $c_i = \{\mathbf{h}_{i1}, \mathbf{h}_{i2}\}$ $(1 \le i \le M)$ and $c'_j = \{\mathbf{h}'_{\mathbf{j1}}, \mathbf{h}'_{\mathbf{j2}}\}$ $(1 \le j \le M')$ be the i-th and the j-th candidate haplotype-diplotypes for

[4]We used the statistical software, R, to implement this algorithm.
[5]The ASD or the HHD values will be used as w_{ij} in Section 4.
[6]*Phasing* is the process of finding the most probable haplotype-diplotype, utilizing some population information.

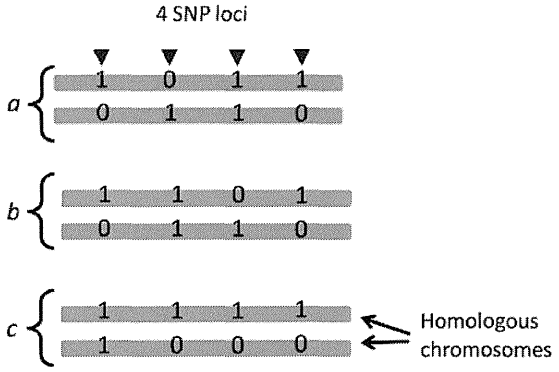

FIG. 2. Haplotype-diplotype examples on which we can observe difference between the ASD and the PMD.

$\mathbf{g}$ and $\mathbf{g}'$, respectively. M and M' are the numbers of haplotype-diplotype candidates for $\mathbf{g}$ and $\mathbf{g}'$, respectively.

If we were given a population model $\mathcal{M}$, we can compute the probability $Prob(c|\mathbf{g}, \mathcal{M})$ that a haplotype-diplotype candidate c is correct for the unphased-diplotype data $\mathbf{g}$. Let $p_i = Prob(c_i|\mathbf{g}, \mathcal{M})$ and $p'_j = Prob(c'_j|\mathbf{g}', \mathcal{M})$ be the conditional probabilities of the candidate haplotype-diplotypes c_i and c'_j under the model $\mathcal{M}$. Then the $PMD_{\mathcal{M}}$ between two haplotype-diplotypes $\mathbf{g}$ and $\mathbf{g}'$ is defined as follows:

$$PMD_{\mathcal{M}}(\mathbf{g}, \mathbf{g}') = \sum_{i=1}^{M} \sum_{j=1}^{M'} H(c_i, c'_j) \cdot q_i \cdot q'_j, \tag{5}$$

where $q_i = p_i / (\sum_{k=1}^{M} p_k)$ and $q'_j = p'_j / (\sum_{k=1}^{M'} p'_k)$. q_i and q'_j are the normalized predicted conditional probabilities of the candidate haplotype-diplotypes c_i and c'_j, respectively.[7] Note that the PMD is the expected value of the distance between candidate haplotype-diplotypes, $H(c_i, c'_j)$, under the population model $\mathcal{M}$.

3.2. HIT HMM-based Distance

To compute the PMD in Section 3.1, we need an appropriate model for the population. In the following, we propose an example of the PMD that we call the HHD.[8] To define the HHD, we propose to use the HMM model used in the HIT algorithm (Rastas et al., 2005) (described in Section 2.2) as the population model for the PMD as follows.

The HMM defined in the HIT algorithm can be considered as a predicted population model. Thus, we first train the HMM from all the unphased-diplotype data that are in our hand, and then we define the HHD as follows. Let $\mathcal{M}^*$ denote the HMM model obtained with the HIT. Then we define the HHD as

$$HHD(\mathbf{g}, \mathbf{g}') = PMD_{\mathcal{M}^*}(\mathbf{g}, \mathbf{g}'). \tag{6}$$

Note that the probability of each haplotype-diplotype candidate is computed as the conditional emission probability of the candidate from the HMM, which can be computed by the forward algorithm (Durbin et al., 1998) for the HMM.

3.3. Discussions on the PMD

3.3.1. The PMD and the multiple founder hypothesis. In many regions (especially in important regions) of the human genome, the haplotype-alleles of the majority in populations can be categorized into a small number of types (Bhatia et al., 2010; Cirulli and Goldstein, 2010), which suggest that only a small number of founder (or ancestral) haplotype-alleles spread over the population on those regions. This

[7]Note that $\sum_{k=1}^{M} p_k = \sum_{k=1}^{M} p'_k = 1$ and there is no need to normalize the probabilities if we enumerate all the candidates. But we need to normalize them in case we ignore the candidates with very small probabilities. When we compute the HHD (which will be introduced in Section 3.2), we ignore candidates with very small probabilities.

[8]We also introduce other simpler examples of the PMD in Section 3.3.1.

TABLE 1. DISTANCES BETWEEN THE INDIVIDUALS IN FIGURE 2

	(1) ASD				(2) $H = PMD_{\mathcal{M}_1}$				(3) $PMD_{\mathcal{M}_2}$		
	a	*b*	*c*		*a*	*b*	*c*		*a*	*b*	*c*
a	0	0.25	0.25	*a*	0	0.25	0.5	*a*	0	0.301	0.450
b	—	0	0.25	*b*	—	0	0.5	*b*	—	0	0.500
c	—	—	0	*c*	—	—	0	*c*	—	—	0

hypothesis of the existence of (a few but) multiple founder haplotype-alleles is very important and effective for various kinds of research, for example, the design of the experiments of linkage disequilibrium mapping (Chung et al., 2008; Gonzalez et al., 1999; Haiman et al., 2003) and the evolutionary history analysis of populations (Ahmad et al., 2002; Gaudieri et al., 1997).

The PMD well reflects the existence of the founder haplotype-alleles. In the example given in Figure 2, there are three individuals with haplotype-diplotypes $a = \{1011, 0110\}$, $b = \{1101, 0110\}$, and $c = \{1111, 1000\}$, but we assume that we know only the unphased-haplotypes, i.e., $\{1, 0\} - \{1, 0\} - \{1, 1\} - \{1, 0\}$, $\{1, 0\} - \{1, 1\}$-$\{1, 0\} - \{1, 0\}$ and $\{1, 1\} - \{1, 0\} - \{1, 0\} - \{1, 0\}$, respectively. We can easily see that the ASD between any two of these three individuals is 0.25 (Table 1*(1)*), and therefore we cannot cluster these three individuals based on the ASD.

The distance between two sequences are often measured by the number of point mutations between them (i.e., we consider two sequences to be very distant to each other if there are many mutations between them). We can define the number of mutations under the assumption of existence of multiple founder haplotype-alleles (for details, see the Appendix). Table 2 shows the number under the assumption that there are two founder haplotype-alleles. According to the table, the clustering result of the three individuals should be the one in Figure 3, which cannot be obtained with the ASD. Note that the clustered individuals *a* and *b* share the same haplotype-allele, i.e., 0110, which also supports the validity of the clustering result.

Unlike the ASD, the haplotype-diplotype distance H reflects the numbers in Table 2 very well. The H value between individuals *a* and *b* is 0.25, which is the same value as the ASD, but H between *a* and *c* and H between *b* and *c* are 0.5 (Table 1*(2)*), which enable us to cluster the individuals as in Figure 3. It means the H values are more appropriate than the ASD values under the existence of the founder haplotype-alleles, at least in this case.

But we cannot compute the real H values unless we know the real haplotype-diplotypes. Instead, we can estimate them by computing the PMD if we are given some population model. Consider the two population models given in Table 3, where haplotype frequencies in the population are given.[9] Under the model $\mathcal{M}_1$, we can phase any of the three individuals' unphased-haplotypes correctly with 100% confidence, and the resulting $PMD_{\mathcal{M}_1}$ values are the same as the H values (Table 1*(2)*). But we cannot predict unphased-haplotypes with such high confidence in many cases, as in the case of the population model $\mathcal{M}_2$ where we have multiple haplotype-diplotype candidates for each unphased diplotype (see Table 4 and Table 1*(3)*).

If we cluster the three individuals based on the $H = PMD_{\mathcal{M}_1}$ values, we can obtain the same clusters as in Figure 3. Furthermore, we can still get the same clusters even if we use the $PMD_{\mathcal{M}_2}$ values instead. Thus, we assume that the PMD is more suitable than the ASD under the multiple founder hypothesis, if we are given an appropriate population model.

3.3.2. Influences of the linkage equilibrium. It is easy to imagine that the linkage equilibrium (LE) and the linkage disequilibrium (LD) should affect the similarity measures. In fact, the variance of the distribution of the ASD values among the individuals should converges to some value in $\Theta(1/m)$ where m is the number of the SNP loci in the region according to the central limit theorem, if the loci are independent to each other. It means that the variance of the ASD values should be smaller on the regions of LE. The PMD and its example HHD should also be influenced by the LE/LD. We compared the influences of the LE/LD to the ASD and the HHD by checking distances on the LE/LD regions obtained from the HapMap database (release 24) (International HapMap Consortium, 2005) as follows.

[9]The population models could be represented by many other methods. For example, we consider HMM-based models in Section 3.2.

TABLE 2. NUMBER OF MUATIONS BETWEEN EACH INDIVIDUAL UNDER THE ASSUMPTION THAT THERE ARE TWO FOUNDERS

	a	*b*	*c*
a	0	2	4
b	—	0	4
c	—	—	0

See Appendix how we obtain the number of mutaions for each pair of individuals.

We can determine whether a region is near to LE or to LD by counting the number of haplotype tagging SNPs (htSNPs) (Carlson et al., 2004; Johnson et al., 2001; Ke and Cardon, 2003; Meng et al., 2003; Rinaldo et al., 2005). The htSNPs are selected so that each SNP in the given region has a correlation larger than a threshold with at least one of the htSNPs. Thus, the regions with many htSNPs can be considered to be near the LE, and regions with few htSNPs can be considered to be near the LD.

We divided the set of SNPs in chromosome 1 into 658 blocks, each of which consists of 100 consecutive SNPs. For each block B, we counted the number h_B of htSNPs obtained by the software Tagger (de Bakker et al., 2005) with the default settings. We selected 100 blocks with the 100 smallest h_B values as the LD regions and also selected 100 blocks with the 100 largest h_B values as the LE regions.

For each of all these regions, we computed the ASD and the HHD measures among the 270 individuals in HapMap (which are the same as the 270 individuals used in Section 4), and computed the variances among the obtained $270 \times 269/2 = 36315$ distances of the ASD and of the HHD. Table 5 shows the difference between the variances of the ASD and the HHD measures. According to the P-values in the table, the HHD reflects the LD/LE effects more than the ASD.

4. APPLICATION TO HAPMAP DATA SETS

4.1. Data sets

In the experiments in Section 4.2, we will use the unphased-diplotype data sets of 22 autosomal chromosomes and X chromosome derived from HapMap release 24 (International HapMap Consortium, 2005). The data sets consist of unphased-diplotypes of 270 individuals: 90 Yoruba in Ibadan, Nigeria (YRI); 90 Utah residents with ancestry from northern and western Europe (CEU, from the CEPH diversity panel); and 90 Japanese in Tokyo, Japan, and Han Chinese in Beijing, China (CHB + JPT). There are 894,398 SNPs that are genotyped for all the above 270 individuals, which we used for our experiments. We divided the SNP set into 8,930 blocks, each of which consists of consecutive 100 SNPs, and we will perform comprehensive experiments against each of these blocks in Section 4.2.

4.2. Experimental results

In this section, we demonstrate the impact of incorporating the population information, by comparing the clustering accuracies by the ASD and that by the HHD on the HapMap data described in Section 4.1.

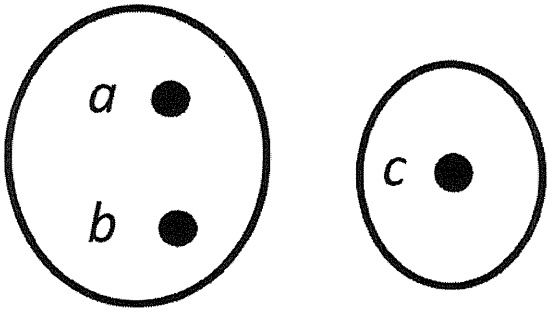

FIG. 3. Clustering results for individuals in Figure 2 based on the numbers of mutations (Table 2), $H = PMD_{\mathcal{M}_1}$ distances (Table 1*(2)*), or $PMD_{\mathcal{M}_2}$ distances (Table 1*(3)*). On the other hand, the ASD distances (Table 1*(1)*) cannot deduce this result.

TABLE 3. POPULATION MODEL EXAMPLES GIVEN AS HAPLOTYPE-ALLELE FREQUENCIES

	Frequency in population	
Haplotype-allele	*(i)* $\mathcal{M}_1$	*(ii)* $\mathcal{M}_2$
1111	0.40	0.20
1110	0.00	0.07
1101	0.20	0.08
1011	0.25	0.10
0011	0.00	0.05
0110	0.10	0.30
0101	0.00	0.05
1100	0.00	0.05
1000	0.05	0.10
Others	0.00	0.00

Against each of the 8,930 blocks, we performed Ward's clustering algorithm (see Section 2.3.1) based on the ASD and also did the same based on the HHD, and compared the CERs (see Section 2.3.2) of their results (Table 6). The difference of the results in relation to the number of htSNPs, i.e., h_B (see Section 3.3.2), is also shown.

The mean of CERs based on the HHD (i.e., 0.3557) is better than that for the ASD (i.e., 0.3611). The P-value of the t-test to see the difference between them is 0.004177, which means the CERs of the HHD is significantly better than that of the ASD. The number of data sets where the HHD (or the ASD) shows better performance than the ASD (or the HHD) are checked with the sign test. Among all the data sets, the HHD is superior to the ASD on 4366 data sets and inferior to the ASD in 3696 data sets. The results of two measures were the same in the other 868 data sets. The P-value of the sign test of all of these results is $8.98 \cdot 10^{-14}$, which means that the HHD is significantly superior to the ASD.

The CERs decrease with increasing h_B for both the ASD and the HHD, but the differences of CERs between the ASD and the HHD also increases as h_B increase (Fig. 4). We call the result HDD's success if the HHD's CER is lower than that of the ASD, and vice versa. The ratio of the HHD's success increases with increasing h_B. The ratio of ASD's success also increases with increasing h_B. The difference of ratios of success between the ASD and the HHD is getting larger as h_B increases. The ratio of the case when the ASD and the HHD have the same results are getting lower as h_B increases (Fig. 5).

The HHD is superior to ASD especially when $80 \leq h_B < 90$. It is a reasonable result as we should be able to better cluster individuals if we have more information (i.e., LE). The difference of ratios of success

TABLE 4. CONDITIONAL PROBABILITIES OF CANDIDATE HAPLOTYPE-DIPLOTYPES FOR INDIVIDUALS IN FIGURE 2 BASED ON THE POPULATION MODELS IN TABLE 3

			Conditional probability	
Individual	*Unphased-diplotype*	*Candidate haplotype-diplotype*	*(i)* $\mathcal{M}_1$	*(ii)* $\mathcal{M}_2$
	{1,0}-{1,0}-{1,1}-{1,0}	{1011, 0110}	1.0000	0.8955
a		{1110, 0011}	0.0000	0.1045
		Others	0.0000	0.0000
	{1,0}-{1,1}-{1,0}-{1,0}	{1101, 0110}	1.0000	0.8727
b		{1110, 0101}	0.0000	0.1273
		Others	0.0000	0.0000
	{1,1}-{1,0}-{1,0}-{1,0}	{1111, 1000}	1.0000	0.8000
c		{1011, 1100}	0.0000	0.2000
		Others	0.0000	0.0000

TABLE 5. MEANS OF VARIANCES OF ASD/HHD MEASURES ON THE REGIONS WHERE THE SNPS ARE WEAKLY CORRELATED AND HIGHLY CORRELATED IN CHROMOSOME 1

	Mean of variances		
	LE	*LD*	P-*value*
ASD	0.00267	0.00546	$2.066 \cdot 10^{-16}$
HHD	0.00248	0.00539	$1.637 \cdot 10^{-17}$

The LE and LD columns show the means of variances on the LE regions (i.e., regions with many htSNPs) and those on the LD regions (i.e., regions with a few htSNPs), respectively. The difference of the variances between weakly and highly correlated regions are tested by t-test for each of the measures. The P-value column shows the P-value of the t-test.

between the ASD and the HHD also becomes largest when $80 < h_B < 90$. In this case, the HHD is superior on 13 data sets, while the ASD is superior only on six data sets among the remaining 18 data sets.

5. CONCLUSION

We proposed a new inter-diplotype similarity measure that we call the PMD. The PMD improves the previous ASD measure by utilizing a population model. As one of such population models, we propose to use the HMM population model used in the phasing algorithm HIT. We call the PMD based on the HIT's HMM the HHD. The HHD utilizes the predicted conditional probabilities of haplotype-diplotypes of unphased-diplotype emitted from the HIT's HMM. Based on comprehensive experiments over 8930 genome-wide data sets of HapMap, we showed that the HHD significantly outperforms the ASD. We also discussed the relationships between the clustering accuracies and the LD.

There are many future tasks to do related to this work. The HHD requires much larger computation time than the ASD, and one future task should be to improve the computation speed of the HHD. There are still data sets for which the HHD is not superior to the ASD. It would be very interesting if we can predict the regions where the HHD is inferior to the ASD, before computing these measures. Another future task is to improve the population model, as it should directly improve the performance of the PMD. From the biological viewpoint, it would also be very interesting if we can utilize our clustering algorithms to identify

TABLE 6. THE EXPERIMENTAL RESULTS AND THEIR RELATIONSHIPS TO THE h_B VALUES

		Mean of CERs		*Comparison of CERs*			
h_B	♯*blocks*	*ASD*	*HHD*	$CER_{ASD} < CER_{HHD}$	$CER_{HHD} < CER_{ASD}$	$CER_{ASD} = CER_{HHD}$	P-*value of sign test*
$0 \sim 10$	1	0.5630	0.5630	0 (0.0)	0 (0.0)	1 (1.0)	
$10 \sim 20$	44	0.4733	0.4678	9 (0.2045)	13 (0.2955)	22 (0.5)	0.5235
$20 \sim 30$	223	0.4363	0.4305	62 (0.2780)	82 (0.3677)	79 (0.3543)	0.1130
$30 \sim 40$	993	0.4240	0.4207	380 (0.3827)	418 (0.4209)	195 (0.1964)	0.1902
$40 \sim 50$	2364	0.3929	0.3877	975 (0.4124)	1131 (0.4784)	258 (0.1091)	$7.276 \cdot 10^{-4}$*
$50 \sim 60$	3063	0.3567	0.3514	1327 (0.4332)	1528 (0.4989)	208 (0.06793)	$1.808 \cdot 10^{-4}$*
$60 \sim 70$	1822	0.3052	0.2997	772 (0.4237)	970 (0.5324)	80 (0.04391)	$2.303 \cdot 10^{-6}$*
$70 \sim 80$	399	0.2584	0.2465	165 (0.4135)	211 (0.5288)	23 (0.05764)	0.02018*
$80 \sim 90$	21	0.2178	0.1944	6 (0.2857)	13 (0.6190)	2 (0.09524)	0.1671
$90 \sim 100$	0	—	—	—	—	—	—
Total	8930	0.3611	0.3557	3696 (0.4139)	4366 (0.4889)	868 (0.09720)	$8.98 \cdot 10^{-14}$*

The ♯blocks column shows the numbers of blocks with the specified h_B values. In the Comparison of CERs columns, the $CER_{ASD} < CER_{HHD}$/$CER_{ASD} > CER_{HHD}$/$CER_{ASD} = CER_{HHD}$ columns show the numbers (and the ratios) of data (with the specified h_B values) where the ASD performed better/the HHD performed better/the performance of the two measures are exactly the same, respectively. $x \sim y$ indicates that $x \leq h_B < y$, and * means the result of the sign test is significant (i.e., ≤ 0.05).

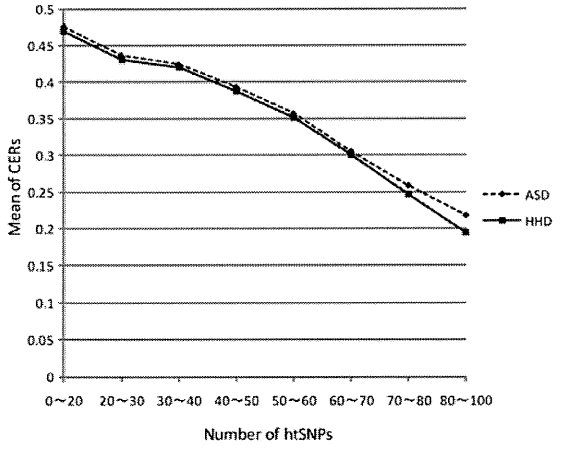

FIG. 4. The plot of h_B values and the means of CERs for both the ASD and the HHD. $x \sim y$ indicates that $x \leq h_B < y$. The HHD is superior to the ASD in all the cases.

gene functions of the target genome regions, especially the regions that affect the disease prevalence and drug responses (Bamshad et al., 2004; Wiencke, 2004; Wilson et al., 2001).

6. APPENDIX

Counting number of mutations under founder hypothesis

Suppose that founder haplotype-alleles $\mathbf{f}_1, \ldots, \mathbf{f}_m$ has been evolved into the present-day haplotpye-alleles of individuals p and q, without any recombinations. Let p_1 and p_2 be the haplotype-alleles of p and $\mathbf{q}_1$ and $\mathbf{q}_2$ be the haplotype-alleles of q. We can consider that the number of mutations between p and q under the assumption of founders $\mathbf{f}_1, \ldots, \mathbf{f}_m$ as

$$S_{\mathbf{f}_1,\ldots,\mathbf{f}_m}(p,q) = \min\left\{ \sum_{i=1}^{2} \min_{j=1}^{m}\{dist(\mathbf{p}_i, \mathbf{f}_j) + dist(\mathbf{q}_i, \mathbf{f}_j)\}, \right.$$
$$\left. \sum_{i=1}^{2} \min_{j=1}^{m}\{dist(\mathbf{p}_i, \mathbf{f}_j) + dist(\mathbf{q}_{2-i}, \mathbf{f}_j)\} \right\}, \tag{7}$$

where $dist()$ denotes the ordinary number of mutations between the two sequences.

But we cannot know the appropriate set of founder haplotype-alleles. Instead, we can define the number of mutations between two individuals under the assumption that there are m founders as

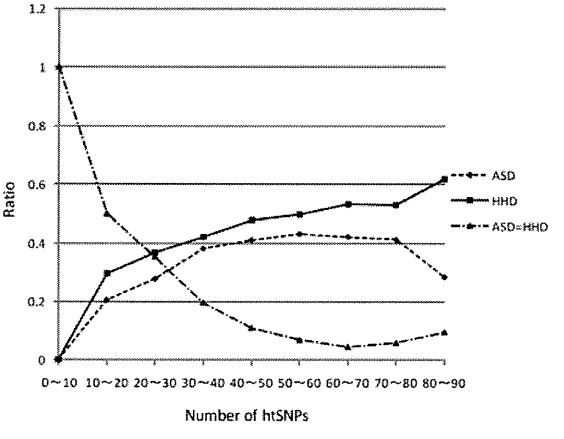

FIG. 5. The plot of h_B values and the ratios of success for both the ASD and the HHD. The line ASD = HHD indicates the results in which the performance of the two measures are the exactly the same. $x \sim y$ indicates that $x \leq h_B < y$. The HHD is superior to the ASD in all the cases.

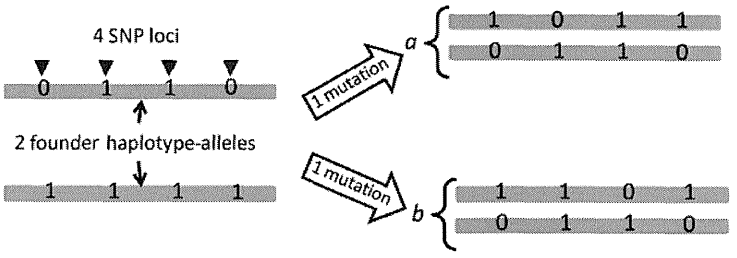

FIG. 6. The optimal founder haplotype-allele pair (when $m = 2$) for the individuals a and b in Figure 2.

$$S_m^*(p, q) = \min_{\mathbf{f}_1, \ldots, \mathbf{f}_m} S_{\mathbf{f}_1, \ldots, \mathbf{f}_m}(p, q). \tag{8}$$

Table 2 shows all the $S_2^*()$ values for all the pairs among individuals a, b, and c in Figure 2. Figure 6 shows the founder pair $\mathbf{f_1}$, $\mathbf{f_2}$ that minimizes the $S_{\mathbf{f}_1, \mathbf{f}_2}(a, b)$ value.

ACKNOWLEDGMENTS

The experiments in this work were done on the Super Computer System of the Human Genome Center, the Institute of Medical Science, the University of Tokyo.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Ahmad, T., Neville, M., Marshall, S.E., et al. 2002. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum. Mol. Genet.* 12, 647–656.

Bamshad, M., Wooding, S., Salisbury, B.A., et al. 2004. Deconstructing the relationship between genetics and race. *Nat. Rev. Genet.* 5, 598–609.

Beaty, T.H., Fallin, M.D., Hetmanski, J.B., et al. 2005. Haplotype diversity in 11 candidate genes across four populations. *Genetics* 171, 259–267.

Bhatia, G., Bansal, V., Harismendy, O., et al. 2010. A covering method for detecting genetic associations between rare variants and common phenotypes. *Plos Comput. Biol.* 6, 1–12.

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., et al. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368, 455–457.

Carlson, C.S., Eberle, M.A., Rieder, M.J., et al. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* 74, 106–120.

Chung, P.Y.J., Beyens, G., Guanabens, N., et al. 2008. Founder effect in different European countries for the recurrent P392L SQSTM1 mutation in Paget's disease of bone. *Calcif. Tissue. Int.* 83, 34–42.

Cirulli, E.T., and Goldstein, D.B. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* 11, 415–425.

Conrad, D.F., Jakobsson, M., Coop, G., et al. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* 38, 1251–1260.

Cornuet, J.M., Sylvain, P., Luikart, G., et al. 1999. New methods employing multilocus genotypes to select or exlcude populations as origins of individuals. *Genetics* 153, 1989–2000.

Cover, T.M., and Thomas, J.A. 1991. *Elements of Information Theory*, John Wiley & Sons, New York.

de Bakker, P.I.W., Yelensky, R., Pe'er, I., et al. 2005. Efficiency and power in genetic association studies. *Nat. Genet.* 37, 1217–1223.

Durbin, R., Eddy, S., Krogh, A., et al. 1998. *Biological Sequence Analysis*. Cambridge Press, New York.

Ester, M., Kriegel, H.P., Sander, J., et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining* 226–231.