Journal of Human Genetics (2012) 57, 326–334

www.nature.com/jhg

ORIGINAL ARTICLE

Genetic differences in the two main groups of the Japanese population based on autosomal SNPs and haplotypes

Yumi Yamaguchi-Kabata, Tatsuhiko Tsunoda, Natsuhiko Kumasaka, Atsushi Takahashi, Naoya Hosono, Michiaki Kubo, Yusuke Nakamura and Naoyuki Kamatani

Although the Japanese population has a rather low genetic diversity, we recently confirmed the presence of two main clusters (the Hondo and Ryukyu clusters) through principal component analysis of genome-wide single-nucleotide polymorphism (SNP) genotypes. Understanding the genetic differences between the two main clusters requires further genome-wide analyses based on a dense SNP set and comparison of haplotype frequencies. In the present study, we determined haplotypes for the Hondo cluster of the Japanese population by detecting SNP homozygotes with 388 591 autosomal SNPs from 18 379 individuals and estimated the haplotype frequencies. Haplotypes for the Ryukyu cluster were inferred by a statistical approach using the genotype data from 504 individuals. We then compared the haplotype frequencies between the Hondo and Ryukyu clusters. In most genomic regions, the haplotype frequencies in the Hondo and Ryukyu clusters were very similar. However, in addition to the human leukocyte antigen region on chromosome 6, other genomic regions (chromosomes 3, 4, 5, 7, 10 and 12) showed dissimilarities in haplotype frequency. These regions were enriched for genes involved in the immune system, cell–cell adhesion and the intracellular signaling cascade. These differentiated genomic regions between the Hondo and Ryukyu clusters are of interest because they (1) should be examined carefully in association studies and (2) likely contain genes responsible for morphological or physiological differences between the two groups.
Journal of Human Genetics (2012) **57,** 326–334; doi:10.1038/jhg.2012.26; published online 29 March 2012

Keywords: genetic differentiation; haplotype; population structure; single-nucleotide polymorphisms

INTRODUCTION

The Japanese population has a relatively low genetic diversity,[1] which was one of the reasons for the early success of genome-wide association studies (GWASs).[2,3] However, the results of several other studies on genetic variations in the Japanese population, which examined mitochondrial DNA-sequence variation,[4,5] polymorphic markers on the Y chromosome[5] or some polymorphic loci in autosomes,[6,7] support the hypothesis that the Japanese population has a 'dual structure' and that immigrants came to Japan in at least two major migrations.[8] These studies revealed that the Japanese population has three main genetic groups, the Hondo Japanese, who live on the main islands of Japan, the Ryukyu Japanese, who live on the Ryukyu Islands, and the Ainu Japanese, who are the indigenous people of Hokkaido, although their current population in Hokkaido is small ($\sim$0.5%). Furthermore, geographic clines of haplotype frequencies were found in the human leukocyte antigen (HLA) region and several loci for blood types.[8]

In a principal component analysis of genome-wide single-nucleotide polymorphism (SNP) genotypes in the Japanese population, we showed that most Japanese individuals fell into two main clusters (the Hondo and Ryukyu clusters).[9] Furthermore, genetic differentiation was observed among different regions in the Hondo people. Although the SNPs that are most differentiated between the Hondo and Ryukyu peoples have been identified,[9,10] a more detailed study of genetic differentiation between the two main clusters is desired for three reasons. First, understanding population structure is essential for the design of GWASs,[11–13] which are powerful tools for identifying disease-causing genes. To conduct more accurate GWASs of the Japanese population, it is important to know whether the population has a dual structure[10] and that the genetic backgrounds for the case and control samples are not biased. In addition, differentiated SNPs can be used as ancestry-informative markers to determine to which subpopulation each individual belongs.[14] Second, differentiated genomic regions are the genomic regions where spurious associations are likely to occur. Therefore, knowledge of the differentiated regions would help to make GWASs more robust. Third, differentiated genes are more likely to be involved in phenotypic variations[15] because some of them rapidly change in

Laboratory for Medical Informatics, Center for Genomic Medicine, The Institute of Physical and Chemical Research (RIKEN), Yokohama, Japan
Correspondence: Dr Y Yamaguchi-Kabata, Laboratory for Medical Informatics, Center for Genomic Medicine, The Institute of Physical and Chemical Research (RIKEN), 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045 Japan.
E-mail: yyamaguc@src.riken.jp
Received 18 September 2011; revised 30 January 2012; accepted 15 February 2012; published online 29 March 2012

allele frequency by adaptive evolution. Therefore, highly differentiated genomic regions could be good places to look for phenotype-associated genes. The Hondo and Okinawa peoples have some phenotypic differences, such as in ear wax type and hair thickness.[8] We previously showed that a SNP in *ABCC11*[16] and another SNP in *EDAR*[17] were the most differentiated nonsynonymous SNPs between the Hondo and Ryukyu clusters known so far.[9] Although these results may depend on the set of SNPs selected for genotyping, they suggest that highly differentiated genes are likely to be involved in phenotypic differences. In fact, the *EDAR* gene was recently shown to be also involved in the morphology of front teeth.[18]

Another advantage of genome-wide SNP genotype data is that they can be used for haplotype inference.[19] Understanding haplotype structure and frequency is important for associating genetic polymorphisms with a given trait and for inferring the genetic genealogy of alleles in a population.[20,21] If a haplotype catalog can be created from genome-wide SNP genotypes, it would be useful for looking at haplotypes at the genomic regions of interest. Furthermore, genome-wide haplotypes would be useful for evaluating genomic diversity of the population and differences between subpopulations. Use of haplotypes, as well as SNP genotypes, may be well suited for identifying genetic differences between closely related subpopulations because a recent recombination may have created new haplotypes that may result in a genetic difference between the two subpopulations. In contrast, analyses of common SNPs are based on only two alleles whose origins are relatively old. Therefore, comparison of haplotype frequencies can be used in addition to comparison of SNP allele frequencies to find genetic differentiation.

We previously showed that haplotype structure and frequency can be estimated from SNP homozygotes by the use of genotype data from 3397 individuals from the Japanese population.[22] In the present study, we applied this approach to genotype data of autosomal SNPs from 18379 individuals from the Hondo cluster of the Japanese population, determined haplotypes and estimated haplotype frequencies. The haplotypes of the Ryukyu cluster were analyzed separately with genotype data from 504 individuals. Our analysis revealed genomic regions with dissimilar haplotype frequencies. In addition to the HLA region in chromosome 6, many other genomic regions showed genetic differentiations between the two clusters. These differentiated regions between the two clusters would be good candidate regions to look for genes that are involved in phenotypic differences between the Hondo and Ryukyu populations.

MATERIALS AND METHODS

Subjects and genotype data

In this study, we used the same 19170 Japanese subjects that were analyzed in our previous study.[10] These individuals consisted of healthy controls from the Midosuji Rotary Club and case individuals from the BioBank Japan Project.[23] All the DNA samples were genotyped for 529 412 SNPs with Illumina 550K or 610K arrays (Illumina, Sandiego, CA, USA).

Selection of individuals for the two main clusters

Principal component analysis of the 19 170 Japanese individuals[10] generated two relatively distinct clusters for the Hondo and Ryukyu populations (Supplementary Figure S1). Using the eigenvalues for principal component 1, we selected 18 379 individuals for the Hondo cluster (principal component 1: −0.012 to 0.012) and 504 individuals for the Ryukyu cluster (principal component 1: −0.046 to −0.023).

Use of genotype data for haplotype analysis

Genotyped SNPs in autosomes (chromosomes 1–22) were selected for haplotype analyses if they satisfied the following three criteria: (1) the call rate was at least 99%, (2) genotype frequencies did not drastically depart from the Hardy–Weinberg equilibrium ($P \geqslant 10^{-7}$) and (3) the minor allele frequency was at least 0.05. After this filtering, the genotype data for 388 591 SNPs were selected and alleles in the SNP genotype data were converted into the corresponding alleles in the top strand with the genomic coordinate for each chromosome.

The genomic regions for all the autosomes were divided into non-overlapping bins having a fixed number of SNPs (4, 6 or 10 in this study). It should be noted that there are a small fraction of regions >1 Mbp where SNPs are very sparse that were excluded from the haplotype analysis, because haplotype inference would be inaccurate for these regions.

Haplotype analysis

We previously examined the efficiency of haplotype determination and frequency estimation based on SNP homozygotes,[22] and applied this approach to the genome-wide SNP genotype data from 18 379 individuals from the Hondo cluster. We evaluated the efficiency of the haplotype analysis for the Hondo cluster, because the reliability of the haplotype analysis based on SNP homozygotes depends on several factors (for example, the length of the region and the level of linkage disequilibrium).[22] Therefore, we examined the fraction and number of homozygotes to see whether they were enough for detection of haplotypes and estimation of haplotype frequencies. In addition, total frequencies of the haplotypes were examined to see whether undetected haplotypes were negligible in terms of frequency and whether the estimated haplotype frequencies were reliable. Haplotypes for the Ryukyu cluster were inferred and their frequencies were estimated using the computer program SNPHAP (www.gene.cimr.cam.ac.uk/clayton/software/) because the small sample size (504) may result in an inaccurate estimation of haplotype frequency based on SNP homozygotes.

To examine genetic differentiation between the Hondo and Ryukyu clusters, the F_{ST} value, as originally defined by Wright,[24] between the Hondo and Ryukyu clusters was calculated from the normalized haplotype frequencies. The haplotype frequencies estimated by counting SNP homozygotes were normalized so that the sum of frequencies was 1.0.

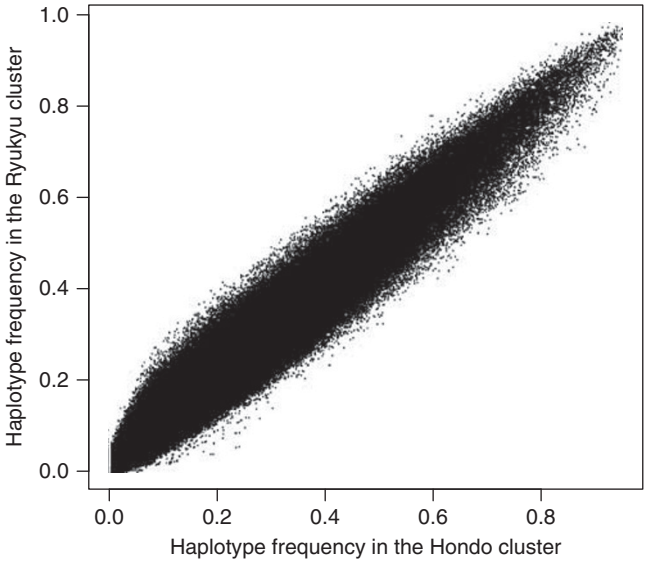

Figure 1 Comparison of haplotype frequencies between the Hondo and Ryukyu clusters. The frequencies of haplotypes for the Hondo cluster (X-axis) and the Ryukyu cluster (Y-axis) are shown in a scatter plot. All the genomic regions in autosomes were divided into non-overlapping bins having four SNPs, and 97 119 regions were analyzed. Haplotypes of the Hondo cluster were determined by detection of SNP homozygotes and the normalized haplotype frequencies (sum = 1.0) were used. Haplotypes of the Ryukyu cluster were inferred by use of the SNPHAP program. The correlation coefficient of haplotype frequency between the two clusters was 0.983.

Comparison of SNP allele frequencies between the Hondo and Ryukyu clusters

Genotyped autosomal SNPs were selected for comparison of allele frequencies in the two clusters if they satisfied the following three criteria: (1) the SNPs were polymorphic in the Japanese sample, (2) the genotype frequency did not drastically depart from the Hardy–Weinberg equilibrium ($P \geqslant 10^{-6}$) and (3) the call rate was at least 0.99. We selected 437 697 SNPs (discarding 65 202 SNPs) for comparison of allele frequencies. For each SNP site, we calculated F_{ST}, as originally defined,[24] between the Hondo and Ryukyu clusters.

RESULTS

Determination of haplotypes of the Hondo and Ryukyu clusters

Haplotypes for the Hondo cluster were determined by detecting SNP homozygotes.[22] To find the appropriate condition for haplotype analysis from SNP homozygotes, we conducted a genome-wide haplotype analysis with different numbers of SNPs (4, 6 and 10) and inspected the results by (1) fraction and number of homozygotes and (2) total frequencies of haplotypes (Supplementary Table S1). The fraction and number of homozygotes depended on the haplotype lengths.[22] When the genomic regions were divided into regions

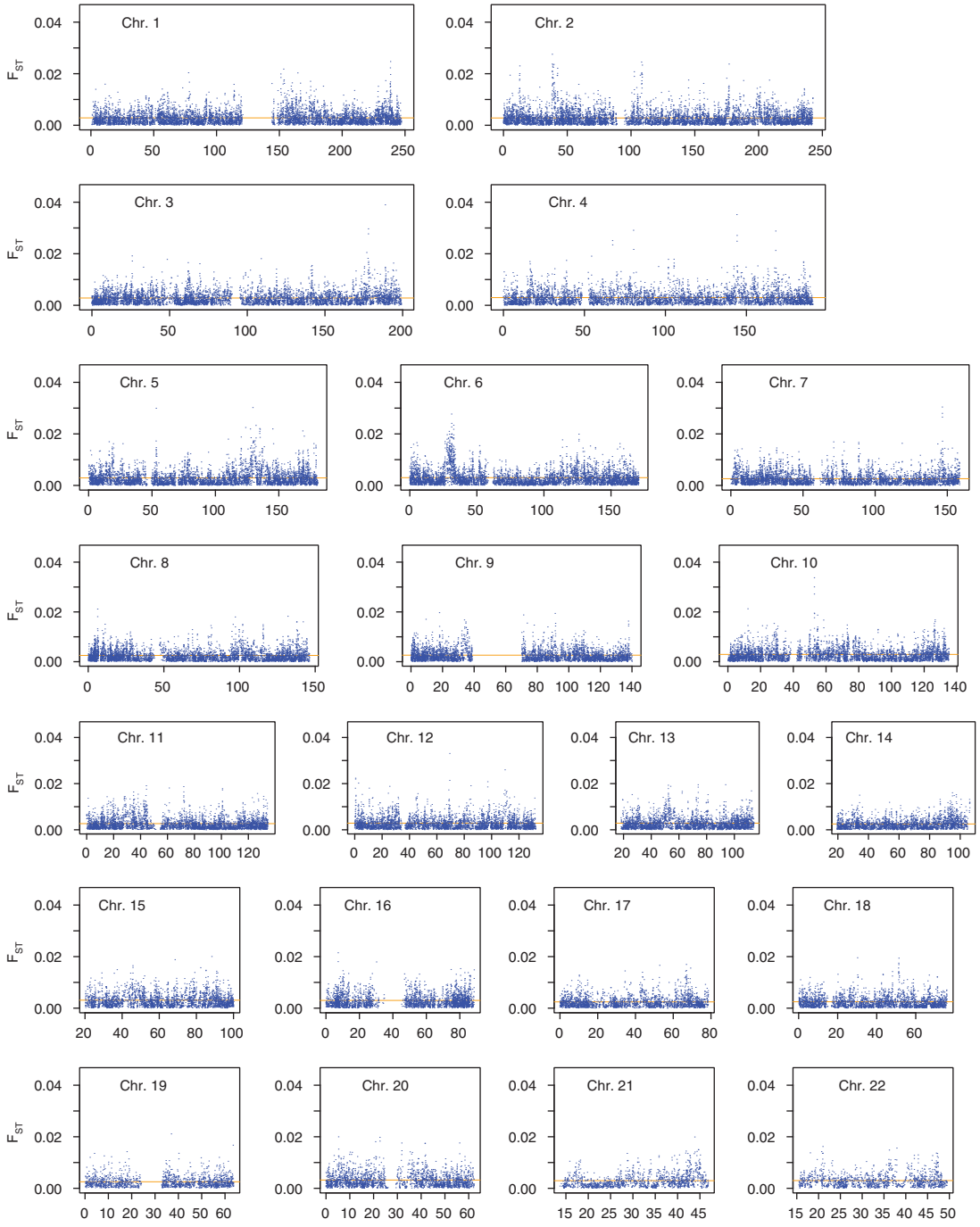

Figure 2 Chromosomal distribution of haplotype F_{ST} between the Hondo and Ryukyu clusters. F_{ST} values calculated with haplotype frequencies are shown along the genomic coordinate (Mbp) for each chromosome. Horizontal orange lines show the average value of haplotype F_{ST} for each chromosome.

having four SNPs, the proportion of SNP homozygotes were about 35% for the Hondo cluster (Supplementary Table S1). As the haplotype becomes longer, the fraction and number of homozygotes tended to decrease. When the genomic regions were divided into regions having four SNPs, the total frequency of haplotypes was 1.011, which was closer to 1.0 than the total frequencies obtained with other conditions. Based on these results, we decided to use the haplotypes with four SNPs. After discarding long haplotypes that may contain large gaps (19 regions, see Materials and methods), 97 119 genomic regions having four SNPs were used for further analysis, and each analyzed region had 6.28 haplotypes on average. Haplotypes for the Ryukyu cluster were inferred and haplotype frequency was estimated for the 97 119 regions by using the SNPHAP program. On average, 7.62 haplotypes were detected from the 504 Ryukyu individuals.

Differentiation between the Hondo and Ryukyu clusters by haplotype frequency

To evaluate genetic difference between the Hondo and Ryukyu clusters at the haplotype level, we compared haplotype frequencies between the two clusters. Generally, the haplotype frequencies in the two clusters were highly correlated (correlation coefficient was 0.983, Figure 1). The level of genetic differentiation between the two clusters was evaluated by F_{ST} with haplotype frequencies. The value of haplotype F_{ST} ranged from 0.0 to 0.039 among the 97 119 genomic regions covering all the autosomes and the average value of F_{ST} was 0.0028 (the empirical distribution of F_{ST} for all the analyzed regions is shown in Supplementary Figure S2). In spite of the low level of differentiation between the two clusters, a substantial proportion of SNPs were located in the tails of the distribution: 2030 of 97 119 regions have $F_{ST} \geqslant 0.01$. Therefore, we searched for genomic regions that showed relatively higher differentiation by the F_{ST} values.

To see whether some specific regions show relatively higher genetic differentiation, we examined chromosomal distribution of haplotype F_{ST} (Figure 2). These plots show that each chromosome has substantial variations in F_{ST} values.[25] Some local genomic regions show high F_{ST} values. In particular, the short arm of chromosome 6 had a long stretch of high F_{ST} haplotypes in the HLA region (approximate genomic positions 28 500 000–33 000 000). High F_{ST} regions were also found on the other genomic regions, as seen on the short arm of chromosome 9 (genomic position: around 35 081 154, proximal to the centromere) where the *PIGO* gene is located and on the long arm of chromosome 7 (genomic position: around 146 600 000) where the *CNTNAP2* gene (contactin-associated protein-like 2) is located.

By comparing haplotype frequencies, we detected the genomic regions that differed most in haplotype frequency between the two clusters (Table 1). A genomic region in chromosome 3 (genomic position: 188 873 942–188 884 675) showed the highest value of haplotype F_{ST} (0.039) in all the autosomal regions (Table 1). Although this region does not contain any protein-coding gene, *SST* and *RTP2* were located nearby. Chromosome 7 had three genomic regions (genomic position: around 146 600 000) adjacent to each other, which showed high values of F_{ST}. These regions contain *CNTNAP2*, whose polymorphism is associated with autism[26] and Pitt–Hopkins-like syndrome 1.[27]

Then we looked at haplotype frequencies at the most differentiated regions to see whether any haplotypes show marked differences in frequency between the two subpopulations (Table 2). We detected a few haplotypes whose difference in haplotype frequency is >0.1 at many of the most differentiated regions. The most differentiated region in chromosome 3 (approximate genomic position: 188 880 000) had a few haplotypes whose frequency differences were remarkable. The major haplotype in the Hondo cluster was CTGT (0.882), whereas its frequency was only slightly lower in the Ryukyu cluster (0.708). However, the haplotype TCAT is present at a frequency of 0.062 in the Hondo cluster, whereas its frequency was much higher in the Ryukyu cluster (0.214).

Table 1 Genomic regions showing the highest differentiation between the Hondo and Ryukyu clusters based on haplotype frequency

Chr	*Region*[a]	*Gene*	*Number of haplotypes*	*Haplotype* F_{ST}
2	38 340 074–38 344 371	—	7	0.0276
3	177 987 891–178 007 464	—	4	0.0277
3	178 013 068–178 039 364	—	9	0.0297
3	188 873 942–188 884 675	—	5	0.0390
4	67 416 498–67 441 195	—	6	0.0251
4	80 340 918–80 365 714	—	3	0.0291
4	144 156 353–144 192 195	*LOC729675*	3	0.0352
4	144 279 914–144 305 695	*LOC729675*	4	0.0271
4	168 169 585–168 208 659	*SPOCK3*	6	0.0288
5	53 395 816–53 399 911		2	0.0299
5	129 619 195–129 644 412		3	0.0302
6	31 240 064–31 244 432	*POU5F1*	5	0.0277
7	146 583 014–146 586 621	*CNTNAP2*	6	0.0304
7	146 597 809–146 600 980	*CNTNAP2*	7	0.0280
7	146 611 489–146 629 226	*CNTNAP2*	6	0.0265
10	52 904 934–52 907 653	*PRKG1*	7	0.0300
10	52 910 028–52 912 289	*PRKG1*	7	0.0272
10	52 912 465–52 914 651	*PRKG1*	4	0.0338
12	69 483 515–69 506 243	*PTPRR*	4	0.0331
12	109 877 902–109 894 920	—	3	0.0260

F_{ST} between the Hondo and Ryukyu clusters was calculated with haplotype frequencies. In all, 20 genomic regions showing the highest values are shown.
[a]Positions for the first and fourth SNPs.

To identify any functional bias in genes located at highly differentiated genomic regions, we examined overrepresented biological functions in these genes. The top 1% of highly differentiated genomic regions (971) were selected by the F_{ST} value and found to contain 379 genes. We divided the highly differentiated genes into two groups: genes in the HLA region (54) and genes in the non-HLA region (325), and conducted a gene-set enrichment analysis of each group. The HLA region was analyzed separately as it is known to be highly differentiated among populations, which may bias or obscure differences in other regions. We examined overrepresented biological functions in the differentiated regions using the PANTHER Classification System (http://www.pantherdb.org/). For the 54 differentiated genes in the HLA region, the molecular functions that are most overrepresented included antigen processing and presentation (Table 3a). On the other hand, for 325 differentiated genes in non-HLA regions, the molecular functions that are most overrepresented included cell–cell adhesion and intracellular-signaling cascade functions (Table 3b).

Differentiation by haplotype frequencies and allele frequencies

To determine to what extent differences in haplotype frequencies is correlated with differences in allele frequencies at single SNP sites, we calculated F_{ST} at all the SNP sites and examined the relationship between haplotype F_{ST} and F_{ST} at SNP sites. We used 437 697 autosomal SNPs to calculate F_{ST} by allele frequencies between the

Table 2 Haplotype frequencies in the most differentiated genomic regions

			Haplotype frequency		
Chr	*Region*[a]	*Haplotype*	*Hondo*	*Ryukyu*	*Difference*[b]
2	38 340 074–38 344 371	CCAC	0.0000	0.0010	0.0010
		CCAT	0.6309	0.4385	−0.1924
		CCGC	0.0000	0.0020	0.0020
		CTGC	0.0000	0.0020	0.0020
		TCAT	0.0147	0.0089	−0.0058
		TTGC	0.2960	0.4514	0.1554
		TTGT	0.0584	0.0962	0.0379
3	177 987 891–178 007 464	AACC	0.0735	0.0903	0.0168
		GACC	0.0102	0.0000	−0.0102
		GACT	0.7040	0.5308	−0.1732
		GGTT	0.2123	0.3790	0.1667
3	178 013 068–178 039 364	CCGA	0.7015	0.5186	−0.1829
		CCGG	0.0073	0.0017	−0.0057
		CTGA	0.0000	0.0057	0.0057
		CTTA	0.0000	0.0020	0.0020
		CTTG	0.0146	0.0147	0.0001
		TCGA	0.1280	0.3117	0.1837
		TCGG	0.0319	0.0252	−0.0067
		TTGA	0.0698	0.0866	0.0168
		TTTG	0.0469	0.0339	−0.0130
3	188 873 942–188 884 675	CTAT	0.0000	0.0010	0.0010
		CTGC	0.0328	0.0624	0.0296
		CTGT	0.8816	0.7084	−0.1732
		TCAC	0.0232	0.0140	−0.0092
		TCAT	0.0623	0.2142	0.1519
4	67 416 498–67 441 195	GAAA	0.8239	0.7065	−0.1174
		GAAC	0.0000	0.0029	0.0029
		GACA	0.0000	0.0010	0.0010
		TACA	0.0795	0.2320	0.1525
		TACC	0.0000	0.0001	0.0001
		TGAC	0.0966	0.0575	−0.0391
4	80 340 918–80 365 714	CACG	0.1507	0.2917	0.1410
		CGCG	0.0000	0.0020	0.0020
		TGTA	0.8493	0.7063	−0.1430
4	144 156 353–144 192 195	AACC	0.0074	0.0030	−0.0044
		AACT	0.7459	0.5694	−0.1765
		GGTC	0.2467	0.4276	0.1809
4	144 279 914–144 305 695	AGAG	0.0073	0.0050	−0.0024
		AGGG	0.2024	0.3522	0.1498
		GTAA	0.7903	0.6419	−0.1485
		GTGA	0.0000	0.0010	0.0010
4	168 169 585–168 208 659	ACCC	0.7475	0.5893	−0.1583
		ACTT	0.0128	0.0050	−0.0078
		ATCC	0.0000	0.0010	0.0010
		ATTT	0.0128	0.0079	−0.0048
		GCCC	0.0000	0.0040	0.0040
		GTTT	0.2269	0.3929	0.1659
5	53 395 816–53 399 911	ATCT	0.7719	0.6121	−0.1597
		GGTC	0.2281	0.3879	0.1597
5	129 619 195–129 644 412	AGTT	0.7071	0.5496	−0.1575
		ATGT	0.0952	0.0625	−0.0327
		GTGG	0.1977	0.3879	0.1902

Table 2 (Continued)

			Haplotype frequency		
Chr	*Region*[a]	*Haplotype*	*Hondo*	*Ryukyu*	*Difference*[b]
6	31 240 064–31 244 432	AGAA	0.0579	0.0188	−0.0391
		GGAA	0.3502	0.2113	−0.1389
		GGGG	0.0416	0.0357	−0.0059
		GTAG	0.0074	0.0020	−0.0054
		GTGG	0.5429	0.7321	0.1893
7	146 583 014–146 586 621	ACAC	0.0073	0.0109	0.0037
		ATGC	0.2022	0.3779	0.1757
		GCAC	0.0073	0.0079	0.0006
		GCAT	0.7508	0.5923	−0.1585
		GCGT	0.0000	0.0010	0.0010
		GTGC	0.0325	0.0100	−0.0225
7	146 597 809–146 600 980	CCGA	0.0000	0.0030	0.0030
		CCGG	0.7349	0.5823	−0.1526
		CTGG	0.0102	0.0080	−0.0022
		TCGG	0.0324	0.0100	−0.0224
		TTAA	0.2050	0.3780	0.1730
		TTAG	0.0102	0.0060	−0.0043
		TTGG	0.0072	0.0128	0.0056
7	146 611 489–146 629 226	CGAC	0.2194	0.3916	0.1722
		CGAT	0.0073	0.0032	−0.0040
		TAGC	0.0812	0.0737	−0.0075
		TAGT	0.6849	0.5196	−0.1653
		TGGC	0.0000	0.0000	0.0000
		TGGT	0.0073	0.0119	0.0046
10	52 904 934–52 907 653	CAAA	0.0889	0.2718	0.1829
		CAAG	0.0073	0.0000	−0.0073
		CAGG	0.0000	0.0010	0.0010
		CGAG	0.0550	0.0336	−0.0214
		CGGG	0.0073	0.0021	−0.0052
		TAAG	0.0651	0.0646	−0.0006
		TAGG	0.7764	0.6269	−0.1495
10	52 910 028–52 912 289	ATGC	0.8054	0.6391	−0.1663
		ATGT	0.0679	0.2101	0.1422
		GCAC	0.0412	0.0241	−0.0171
		GCAT	0.0646	0.0681	0.0036
		GTAC	0.0000	0.0069	0.0069
		GTAT	0.0000	0.0001	0.0001
		GTGT	0.0209	0.0516	0.0306
10	52 912 465–52 914 651	CTCA	0.7985	0.6379	−0.1607
		CTTC	0.0420	0.0327	−0.0092
		TCTC	0.1522	0.3294	0.1772
		TTCA	0.0073	0.0000	−0.0073
12	69 483 515–69 506 243	CCCC	0.0074	0.0208	0.0135
		CTCT	0.8704	0.7143	−0.1561
		TCCT	0.0000	0.0020	0.0020
		TCTC	0.1223	0.2629	0.1406
12	109 877 902–109 894 920	GTTC	0.4596	0.6181	0.1584
		TCGC	0.0000	0.0050	0.0050
		TCGT	0.5404	0.3770	−0.1634

[a]Positions for the first and fourth SNPs.
[b]Ryukyu frequency −Hondo frequency.

Table 3a Overrepresented functions of highly differentiated genes in the HLA region

Biological process	*All human genes*[a]	*Differentiated genes*[b]	*Expected*	P-*value*
Antigen processing and presentation	78	8	0.21	4.92E-11
Unclassified	6681	41	18.12	2.32E-10
Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	42	5	0.11	1.21E-07
Establishment or maintenance of chromatin architecture	302	9	0.82	1.22E-07
Organelle organization	328	9	0.89	2.43E-07
Cellular defense response	457	9	1.24	3.69E-06
Response to toxin	97	4	0.26	1.47E-04
Response to stimulus	1798	14	4.88	2.38E-04
Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	3825	21	10.37	6.26E-04
Immune system process	2628	15	7.13	3.45E-03
Cellular component organization	1443	10	3.91	4.99E-03

Enrichment of each biological function in the genes located in the differentiated regions was examined by using the PANTHER classification system.
[a]19911 genes (Reflist) as a catalog of all human genes.
[b]54 genes located at the differentiated genomic regions in the HLA region on chromosome 6.

Table 3b Overrepresented functions of highly differentiated genes in the non-HLA region

Biological Process	*All human genes*[a]	*Differentiated genes*[b]	*Expected*	P-*value*
Cell–cell adhesion	799	29	13.04	6.29E-05
Intracellular signaling cascade	1568	46	25.59	8.71E-05
Cytokinesis	238	13	3.88	1.84E-04
Nervous system development	1258	36	20.53	8.40E-04
Mitosis	635	22	10.36	8.88E-04
Cell adhesion	1333	37	21.76	1.22E-03
Cell motion	964	29	15.74	1.30E-03
Cell cycle	1840	46	30.03	2.57E-03
Signal transduction	4191	90	68.41	2.65E-03
Cell surface receptor linked signal transduction	2235	53	36.48	3.64E-03
Cell communication	4365	92	71.25	4.08E-03
Cellular glucose homeostasis	72	5	1.18	6.99E-03
Cellular process	6258	123	102.15	8.20E-03
Cell–matrix adhesion	173	8	2.82	8.22E-03
Homeostatic process	142	7	2.32	9.46E-03
System development	2031	47	33.15	9.53E-03

Enrichment of each biological function in the genes located in the differentiated regions was examined by using the PANTHER classification system.
[a]19911 genes (Reflist) as a catalog of all human genes.
[b]325 genes located at the differentiated genomic regions in the non-HLA regions.

Hondo and Ryukyu clusters. The most differentiated autosomal SNPs were found in the *MOG* gene of the HLA region in chromosome 6 (Supplementary Table S2), in agreement with the results of a previous study.[9] Examination of the most differentiated SNPs in gene regions (Supplementary Table S3) detected differentiated nonsynonymous SNPs (Supplementary Table S4). The most differentiated SNPs were found in the following annotated genes *FMN2*, *FBXL21, GEIN6* and *ZNF96*. Our previous study identified the most differentiated non-synonymous SNPs in *EDAR* and *ABCC11*. This discrepancy on the most differentiated nonsynonymous SNPs may be due to the differences in the SNPs that were selected for genotyping in the two studies.

Next, to examine the relationship between haplotype F_{ST} and F_{ST} at single SNP sites, we compared F_{ST} in two ways. First, we calculated the average value of F_{ST} at SNP sites within each region and examined the relationship with the haplotype F_{ST}. Second, we chose the largest F_{ST} for any SNP in each region and examined the relationship with the value of haplotype F_{ST}. We found that the haplotype F_{ST} was significantly correlated with these values of F_{ST} for each region (correlation coefficient was 0.837 for Figure 3a and 0.811 for Figure 3b).

However, the correlation between haplotype F_{ST} and the largest F_{ST} based on allele frequency was not very strong. To check the dissimilarity in two measures, we selected 971 genomic regions showing the highest values of haplotype F_{ST} (top 1%) and examined how many of them had the highest F_{ST} values at single SNP sites. By comparing with the top 1% genomic regions (971) having the highest F_{ST} at single SNP sites, we found that only 392 of the 971 genomic regions had the largest differences in both haplotype and allele frequencies. These results show that Hondo and Ryukyu clusters have genomic regions that are highly differentiated in haplotype frequency without a drastic difference in allele frequency at single SNP sites. Conversely, some genomic regions are highly differentiated in allele frequency but did not show drastic differences in haplotype frequency as single SNP sites. We considered that the former cases are to be investigated rather than the latter cases by two reasons. First, the latter cases may be explained by a weaker linkage of polymorphisms between SNP sites. Second, the merits of haplotype analysis may be found in the former cases. One example of a genomic region where haplotype frequencies showed drastic differences but did not have any highly differentiated SNP is in chromosome 1 (genomic position: 235 499 862–235 513 179) where the differentiation in haplotype frequency was 0.0130 (in the top 1%). This region contains the *RYR2* (ryanodine receptor 2) gene, whose mutations are associated with ventricular tachycardia and arrhythmogenic right-ventricular dysplasia. In this region, the frequency of haplotype TATC was 0.070 for the Hondo cluster and 0.190 for the Ryukyu cluster. However, no strongly differentiated SNP was observed in this region, the largest SNP F_{ST} being 0.0072. Another example is a genomic region in chromosome 4 (genomic position: 463 935–487 138) that contains *ZNF721* and *PIGG*. The haplotype F_{ST} was 0.0130 for this region, whereas the largest SNP F_{ST} value based on allele frequency was 0.0091. So far *PIGG* has not been related to any phenotype or disease, whereas other genes involved in phosphatidylinositol glycan anchor biosynthesis are known to be related to various kinds of diseases (for example, *PIGA* is known to be involved in paroxysmal nocturnal hemoglobinuria).

DISCUSSION

The present study examined the genetic differentiation between the Hondo and Ryukyu clusters in the Japanese population with SNP

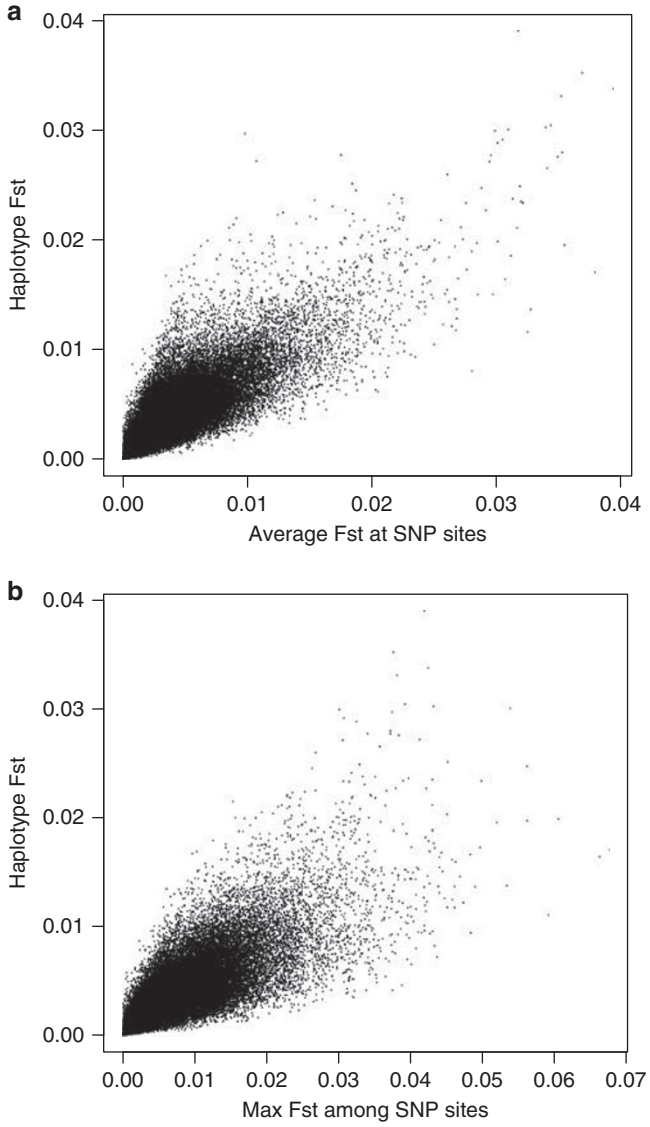

Figure 3 Relationships between differentiation in haplotype frequency and differentiation at SNP sites. F_{ST} at each SNP site between the Hondo and Ryukyu clusters was calculated with allele frequencies. As each region had four analyzed SNPs, the average or maximum value of F_{ST} in each region was used for comparison. (**a**) Haplotype F_{ST} and average values of F_{ST} at SNP sites within the region are shown in scatter plot. The correlation coefficient was 0.837 (0.835–0.839). (**b**) Haplotype F_{ST} and the maximum value of F_{ST} at SNP sites within the region are shown in a scatter plot. The correlation coefficient was 0.811 (0.809–0.814).

genotype data from about 400 000 autosomal SNP sites. Population differentiation between the two clusters was examined at both the allele frequency and haplotype levels. This is the first time that differentiation between these clusters was examined by genome-wide haplotypes. We identified many non-HLA regions with haplotype frequencies that were dissimilar between the Hondo and Ryukyu clusters. Previous studies have shown regional differences in haplotype frequency in the Japanese population mainly in the HLA region.[28,29] The present results suggest that geographic clines of haplotype frequencies exist in genomic regions other than the HLA region.

Jomon and Yayoi peoples differ in skeletal and cranial morphology. According to the dual structure model of the Japanese population,[8] regional differences of phenotypic variations of the Japanese may be explained by the varying fates of intermixture of the peoples in the second migration from Northeast Asia. In fact, there are morphological differences among different geographical regions in modern Japanese on skeletal, teeth, cranial and facial morphologies. However, the genetic determinants of these morphological differences have not been fully elucidated. Differentiated genomic regions found in the present study may be good candidates to search for the genetic determinants of the phenotypic differences between these peoples with the caveat that variations in the X and Y chromosomes and mitochondrial DNA were not investigated in this work. The differentiated genomic regions found in the present study may be good candidates to search for the genetic determinants of the phenotypic differences between two people.

Understanding the differentiation between subpopulations, in addition to being useful for avoiding false positive results in association studies, is also important for medical population genomics when disease prevalence varies among the populations. For example, the prevalence of closed angle glaucoma is higher in Okinawa than the main islands of Japan.[30] The Hondo and Okinawa peoples slightly differ in morphology, and some genetic factors may contribute to the phenotypic differences between them. Environmental factors may also affect the higher prevalence of glaucoma in Okinawa. Further studies are needed to clarify as to which and to what extent genetic factors contribute to the higher prevalence of glaucoma in Okinawa.

Differentiated genomic regions should be examined carefully in GWASs because spurious associations are likely to occur. On the other hand, spurious associations are less likely to occur in most other regions with little differentiation. In addition, some of the differentiated SNPs identified in this study can be used as ancestry-informative markers). A set of SNPs as ancestry-informative markers would be useful for identifying the subpopulation to which each individual belongs. The catalog of real haplotypes with their estimated frequencies, as we created in this study, will be useful for identifying causative polymorphisms for a trait, which are linked to the most associated SNPs in a GWAS. In particular, the haplotypes for the Hondo cluster were determined by SNP homozygotes without ambiguity, and the estimated haplotype frequencies were very similar to the frequencies by the SNPHAP program (correlation coefficient was 0.9995). The genome-wide haplotype catalog created in this study could be improved by investigating the haplotype block structure, which varies between genomic regions, because the strength of linkage of polymorphisms between SNP sites is different by regions. A genome-wide haplotype analysis is one of the ways to uncover a rough sketch of genome-sequence variations with a large number of samples, although sequencing individual genomes is becoming more convenient and less expensive. Through an appropriately designed haplotype analysis of many individuals, we may be able to identify the most variable, conserved or differentiated regions in human populations of interest.

ACKNOWLEDGEMENTS

We thank Drs Kazuharu Misawa, Yukinori Okada, Akihiro Fujimoto, Fuyuki Miya and Todd Johnson for helpful discussions and comments on this study, Keith Anthony Boroevich for helpful suggestions on the manuscript and Yoshiyuki Yukawa for technical assistance. We also thank all the members in the BioBank Japan Project for their effort in organizing the project and collecting samples. This study was supported by the Ministry of Education, Culture, Sports, Science and Technology.

1 Haga, H., Yamada, R., Ohnishi, Y., Nakamura, Y. & Tanaka, T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. *J. Hum. Genet.* **47,** 605–610 (2002).
2 Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R. & Tsunoda, T. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32,** 650–654 (2002).
3 Suzuki, A., Yamada, R., Chang, X., Tokuhiro, S., Sawada, T. & Suzuki, M. *et al.* Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* **34,** 395–402 (2003).
4 Tanaka, M., Cabrera, V. M., Gonzalez, A. M., Larruga, J. M., Takeyasu, T. & Fuku, N. *et al.* Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res.* **14,** 1832–1850 (2004).
5 Horai, S., Murayama, K., Hayasaka, K., Matsubayashi, S., Hattori, Y. & Fucharoen, G. *et al.* mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. *Am. J. Hum. Genet.* **59,** 579–590 (1996).
6 Omoto, K. & Saitou, N. Genetic origins of the Japanese: a partial support for the dual structure hypothesis. *Am. J. Hum. Genet.* **102,** 437–446 (1997).
7 Hatta, Y., Ohashi, J., Imanishi, T., Kamiyama, H., Iha, M. & Simabukuro, T. *et al.* HLA genes and haplotypes in Ryukyuans suggest recent gene flow to the Okinawa Islands. *Hum. Biol.* **71,** 353–365 (1999).
8 Hanihara, K. Dual structure model for the population history of the Japanese. *Japan Rev.* **2,** 1–33 (1991).
9 Yamaguchi-Kabata, Y., Nakazono, K., Takahashi, A., Saito, S., Hosono, N. & Kubo, M. *et al.* Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am. J. Hum. Genet.* **83,** 445–456 (2008).
10 Kumasaka, N., Yamaguchi-Kabata, Y., Takahashi, A., Kubo, M., Nakamura, Y. & Kamatani, N. Establishment of a standardized system to perform population structure analyses with limited sample size or with different sets of SNP genotypes. *J. Hum. Genet.* **55,** 525–533 (2010).
11 Tian, C., Gregersen, P. K. & Seldin, M. F. Accounting for ancestry: population substructure and genome-wide association studies. *Hum. Mol. Genet.* **17,** R143–R150 (2008).
12 Rosenberg, N. A. & Nordborg, M. A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. *Genetics* **173,** 1665–1678 (2006).
13 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
14 Bauchet, M., McEvoy, B., Pearson, L. N., Quillen, E. E., Sarkisian, T. & Hovhannesyan, K. *et al.* Measuring European population stratification with microarray genotype data. *Am. J. Hum. Genet.* **80,** 948–956 (2007).
15 Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12,** 1805–1814 (2002).
16 Yoshiura, K., Kinoshita, A., Ishida, T., Ninokata, A., Ishikawa, T. & Kaname, T. *et al.* A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat. Genet.* **38,** 324–330 (2006).
17 Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R. & Batubara, L. *et al.* A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum. Mol. Genet.* **17,** 835–843 (2008).
18 Kimura, R., Yamaguchi, T., Takeda, M., Kondo, O., Toma, T. & Haneji, K. *et al.* A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am. J. Hum. Genet.* **85,** 528–535 (2009).
19 Clark, A.G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7,** 111–122 (1990).
20 Tsunoda, T., Lathrop, G. M., Sekine, A., Yamada, R., Takahashi, A. & Ohnishi, Y. *et al.* Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome. *Hum. Mol. Genet.* **13,** 1623–1632 (2004).
21 The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
22 Yamaguchi-Kabata, Y., Tsunoda, T., Takahashi, A., Hosono, N., Kubo, M. & Nakamura, Y. *et al.* Making a haplotype catalog with estimated frequencies based on SNP homozygotes. *J. Hum. Genet.* **55,** 500–506 (2010).
23 Nakamura, Y. The BioBank Japan Project. *Clin. Adv. Hematol. Oncol.* **5,** 696–697 (2007).
24 Wright, S. The genetical structure of populations. *Ann. Eugenics* **15,** 323–354 (1951).
25 Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M. & Hill, W. G. Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* **15,** 1468–1476 (2005).
26 Bakkaloglu, B., O'Roak, B. J., Louvi, A., Gupta, A. R., Abelson, J. F. & Morgan, T. M. *et al.* Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. *Am. J. Hum. Genet.* **82,** 165–173 (2008).
27 Orrico, A., Galli, L., Zappella, M., Lam, C. W., Bonifacio, S. & Torricelli, F. *et al.* Possible case of Pitt-Hopkins syndrome in sibs. *Am. J. Med. Genet.* **103,** 157–159 (2001).
28 Tokunaga, K., Imanishi, T., Takahashi, K. & Juli, T. in *Prehistoric Mongoloid Dispersals* (eds Akazawa, T. & Szathary, E. J.) 187–197 (Oxford University Press, Oxford, 1996).
29 Tokunaga, K., Bannai, M., Imanishi, T. & Juli, T. in *The Origins and Past of Modern Humans: Towards Reconcilation* (eds Omoto, K. & Tobias, P. V.) 74–87 (World Scientific Publishing, Singapore, 1998).
30 Nakamura, Y., Ishikawa, S., Nakamura, Y., Hayakawa, K. & Sawaguchi, S. Incidence of acute angle-closure glaucoma in Okinawa (Japanese). *Atarashii Ganka* **17,** 683–686 (2000).

Supplementary Informationaccompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)