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 $c'_j = \{\mathbf{h}'_{j1}, \mathbf{h}'_{j2}\}$ $(1 \le j \le M')$ 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.

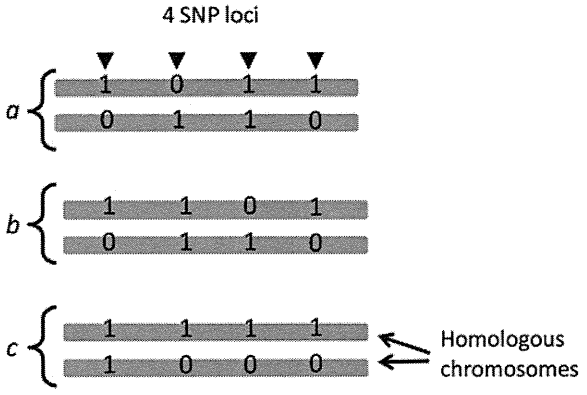

FIG. 2. Haplotype-diplotype examples on which we can observe difference between the ASD and the PMD.

$\mathbf{g}$ 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.

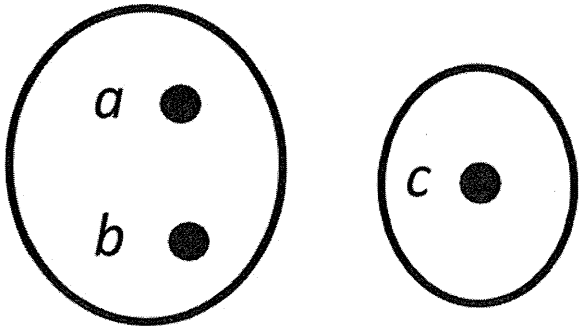

FIG. 3. Clustering results for individuals in Figure 2 based on the numbers of mutations (Table 2), $H = PMD_{\mathcal{M}_1}$ distances (Table 1*(2)*), or $PMD_{\mathcal{M}_2}$ distances (Table 1*(3)*). On the other hand, the ASD distances (Table 1*(1)*) cannot deduce this result.

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).

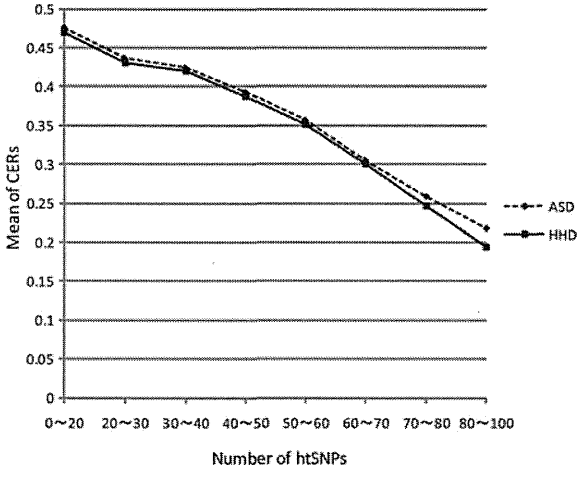

FIG. 4. The plot of h_B values and the means of CERs for both the ASD and the HHD. $x \sim y$ indicates that $x \leq h_B < y$. The HHD is superior to the ASD in all the cases.

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

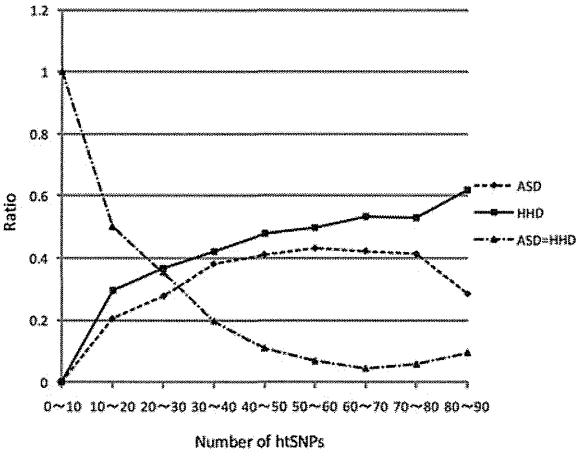

FIG. 5. The plot of h_B values and the ratios of success for both the ASD and the HHD. The line ASD = HHD indicates the results in which the performance of the two measures are the exactly the same. $x \sim y$ indicates that $x \leq h_B < y$. The HHD is superior to the ASD in all the cases.

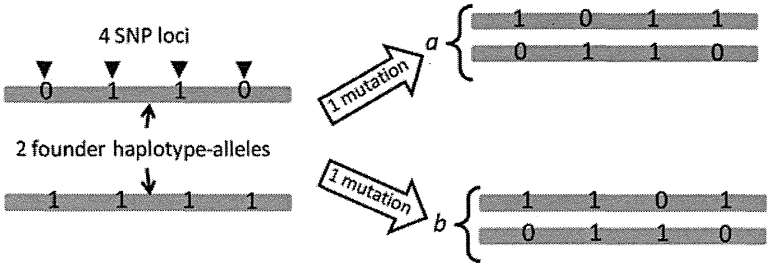

FIG. 6. The optimal founder haplotype-allele pair (when $m = 2$) for the individuals a and b in Figure 2.

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

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.

Fallin, D., Cohen, A., Essioux, L., et al. 2001. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res.* 11, 143–151.

Falush, D., Stephens, M., and Pritchard, J.K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587.

Gao, X., and Martin, E.R. 2009. Using allele sharing distance for detecting human population stratification. *Hum. Hered.* 68, 3.

Gao, X., and Starmer, J. 2007. Human population structure detection via multilocus genotype clustering. *BMC Genet.* 8, 34.

Gaudieri, S., Leelayuwat, C., Tay, G.K., et al. 1997. The Major Histocompatability Complex (MHC) contains conserved polymorphic genomic sequences that are shuffled by recombination to form ethnic-specific haplotypes. *J. Mol. Evol.* 45, 17–23.

Gonzalez, E., Bamshad, M., Sato, N., et al. 1999. Race-specific HIV-1 disease-modifying effects associated with CCR5 haplotypes. *Proc. Natl. Acad. Sci. USA* 96, 12004–12009.

Haiman, C.A., Stram, D.O., Pike, M.C., et al. 2003. A comprehensive haplotype analysis of CYP19 and breast cancer risk: The Multiethnic Cohort. *Hum. Mol. Genet.* 12, 2679–2692.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437, 1299–1320. Available at www.hapmap.org. Accessed November 1, 2011.

Isaev, A. 2004. *Introduction to mathematical methods to bioinformatics*. Springer, New York.

Jakobsson, M., Scolz, S.W., Scheet, P., et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451, 998–1003.

Jin, L., Zhu, W., and Guo, J. 2010. Genome-wide association studies using haplotype clustering with a new haplotype similarity. *Genet. Epidemiol.* 34, 633–641.

Johnson, G.C.L., Esposito, L., Barratt, B.J., et al. 2001. Haplotype tagging for the identification of common disease genes. *Nat. Genet.* 29, 233–237.

Kaufman, L, and Rousseeuw, P. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.

Ke, X., and Cardon, L.R. 2003. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* 19, 287–288.

Kim, S., and Misra, A. 2007. SNP genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.* 9, 289–320.

Lesk, A.M. 2005. Introduction to Bioinformatics, 2nd ed. *Oxford*, New York.

Li, J., and Jiang, T. 2005. Haplotype-based linkage disequilibrium mapping via direct data mining. *Bioinformatics* 21, 4384–4393.

Li, J., Zhou, Y., and Elston, R.C. 2006. Haplotype-based quantitative trait mapping using a clustering algorithm. *BMC Bioinform.* 7, 258.

Mao, X., Bigham, A.W., Mei, R., et al. 2007. A genomewide admixture mapping panel for Hispanic/Latino populations. *Am. J. Hum. Genet.* 80, 1171–1178.

Meng, Z., Zaykin, D.V., Xu, C., et al. 2003. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am. J. Hum. Genet.* 73, 115–130.

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Rabiner, L.R., and Juang, B.H. 1986.An introduction to hidden Markov models. *IEEE ASSP Mag.* 3, 4–16.

Rastas, P., Koivisto, P.M., Mannila, H., et al. 2005. A hidden Markov technique for haplotype reconstruction. *Lect. Notes Bioinform.* 3692, 140–151.

Rinaldo, A., Bacanu, S., Devlin, B., et al. 2005. Characterization of multilocus linkage disequilibrium. *Genet. Epidemiol.* 28, 193–206.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Small, K.M., Mialet-Perez, J., Seman, C.A., et al. 2004. Polymorphisms of cardiac presynaptic α_{2C} adrenergic receptors: diverse intragenic variability with haplotype-specific functional effects. *Proc. Natl. Acad. Sci. USA* 101, 13020–13025.

Team RDC. 2007. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.*

Tzeng, J.Y., Devlin, B., Wasserman, L., et al. 2003. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.* 72, 891–902.

Ward, J.H. 1963. Hierarchical grouping procedure to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244.

Ward, J.H., and Hook, M.E. 1963. Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educ. Psychol. Measure.* 23, 69–81.

Wiencke, J.K. 2004. Impact of race/ethnicity on molecular pathways in human cancer. *Nat. Rev. Cancer* 4, 79–84.

Wilson, F.W., Weale, M.E., Smith, A.C., et al. 2001. Population genetic structure of variable drug response. *Nat. Genet.* 29, 265–269.

Witherspoon, D.J., Wooding, S., Rogers, A.R., et al. 2007. Genetic similarities within and between human populations. *Genetics* 176, 351–359.

Yang, Y., and Tabus, I. 2007. Haplotype block partitioning using a normalized maximum likelihood model. *Proc. IEEE Genomic Signal Process. Stat.* 1–4.

Address correspondence to:
Ms. Ritsuko Onuki
Bioinformatics Center
Institute for Chemical Research
Kyoto University
Gokasho, Uji
Kyoto 611-0011, Japan

E-mail: onuki@hgc.jp

Review Article

Recent advances on the genetics of rheumatoid arthritis: current topics and the future

Koichiro Ohmura[1,*], Chikashi Terao[2] and Tsuneyo Mimori[1]

[1]Department of Rheumatology and Clinical Immunology, Kyoto University Graduate School of Medicine, Kyoto, Japan
[2]Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan

Rheumatoid arthritis (RA) is a chronic autoimmune disease that causes severe joint pain and eventually joint deformity. Recent large cohort studies and the rapid progression of genotyping platforms have enabled identification of more than 30 susceptibility genes for RA. *HLA* is the major genetic determinant for RA for which a shared epitope hypothesis (70th-74th amino acids of HLA-DR β chain determine susceptibility) has been accepted. However, recent detailed single nucleotide polymorphism (SNP) typing of the *HLA* region and imputation method revealed that the most important amino acid positions of the HLA-DR β chain are the 11th in addition to the 71st and the 74th. HLA-B (at position 9) and HLA-DPB1 (at position 9) are also important determinants. This revised shared epitope hypothesis will form a new theory for *HLA* association. Another topic is that anti-citrullinated protein antibody (ACPA)-negative RA has been shown to be genetically different from ACPA-positive RA. Many susceptibility genes including *HLA* were not associated with ACPA-negative RA; however, we have shown that some *HLA* alleles are associated with ACPA-negative RA. In this review, we present some new findings regarding *HLA* as well as some recently discovered susceptibility genes for RA.

Rec.4/6/2012, Acc.5/14/2012, pp90-98

*Correspondence should be addressed to:
Koichiro Ohmura, Department of Rheumatology and Clinical Immunology, Kyoto University Graduate School of Medicine, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan. Phone: +81-75-751-4380, Fax: +81-75-751-4338, Email: ohmurako@kuhp.kyoto-u.ac.jp

Key words genetics, GWAS, HLA, rheumatoid arthritis, SNP

Introduction

Since 2003, when sequencing of the human genome was completed, there has been a burst of identification of new susceptibility genes for RA. In the last several years in particular, more than 30 genes or loci have been identified as RA-related genes[1]. This activity was supported by the development of SNP genotyping platform, which enables us to type hundreds of millions of SNPs in a few weeks, even in a relatively small lab. In addition, a growing number of large cohorts were formed to tackle the elucidation of RA pathogenesis, which provided substantial power to detect genes of significance[2].

ACPA is a specific autoantibody of RA, and its target antigens are citrullinated vimentin, filaggrin, α-enolase, and others[3]. It is a useful marker not only for diagnosis of RA, but also for predicting disease course[4]. ACPA-positive RA

is clinically severer than ACPA-negative RA. Moreover, it has been suggested that ACPA-positive RA is genetically distinct from ACPA-negative RA[5,6].

Here we present the recent advances in RA genetics and also discuss the genetic differences between ACPA-positive and ACPA-negative RA.

Human leukocyte antigen (HLA)

Genetic predisposition to RA has been investigated intensively. *HLA* is a major determinant of RA susceptibility and *HLA-DRB1**01:01, *01:02, *04:01, *04:04, *04:05, *04:08, *04:10, *04:13, *04:16, *10:01, *14:02 and *14:06 were reported to be associated with RA development. Among these *HLA-DRB1* alleles, there are common amino acid sequences at the 70th-74th residues of the HLA-DRβ chain (QKRAA, QRRAA or RRRAA), which is called a 'shared epitope' (SE)[7]. The association of *HLA-DRB1* SE with RA has been replicated in many ethnic groups[8]. However, recently the important role of Leucine at 67th position (Leu67)[9,10] and Valine at 11th position (Val11)[10] for RA development and resistant effect on RA development by Aspatic acid at 70th position (Asp70)[11] were also reported. In addition, Raychaudhuri et al. used existing genome-wide SNP data of >5,000 ACPA-positive RA cases and ~15,000 controls and imputed (expected SNP genotypes in silico from adjacent SNP gentotypes and linkage disequilibrium information) the gap SNP genotypes of HLA locus and reported the following findings. They showed that three amino acid positions (11, 71 and 74) of HLA-DRβ chain as well as single-amino acid positions in HLA-B (at position 9) and HLA-DPβ chain (at position 9) explain most of the MHC association with RA[12]. All these positions are located in peptide-binding grooves, as shown in Fig.1. Among these positions, position 11 of HLA-DRβ chain showed the strongest association with RA development ($p<10^{-581}$ for position 11). As shown in Table 1, Val11 and Leu11 are the key amino acids for susceptibility and

Table 1 Effect estimates of the 3 amino acids associated with risk of RA

HLA-DRβ1 amino acid at position			multivariate OR	95%CI	*HLA-DRB1* alleles
11	71	74			
Val	Lys	Ala	4.44	4.02-4.91	*04:01
Val	Arg	Ala	4.22	3.75-4.75	*04:08, *04:05, *04:04, *10:01
Leu	Arg	Ala	2.17	1.94-2.42	*01:02, *01:01
Pro	Arg	Ala	2.04	1.59-2.62	*16:01
Val	Arg	Glu	1.65	1.24-2.19	*04:03, *04:07
Asp	Arg	Glu	1.65	1.29-2.10	*09:01
Val	Glu	Ala	1.43	1.04-1.96	*04:02
Pro	Ala	Ala	1.00	Reference	*15:01, *15:02
Ser	Arg	Ala	0.88	0.77-1.00	*11:01, *11:04, *12:01
Ser	Arg	Leu	0.71	0.57-0.89	*08:01, *08:04
Ser	Lys	Arg	0.63	0.54-0.73	*03:01
Ser	Glu	Ala	0.59	0.51-0.68	*11:02, *11:03, *13:01, *13:02

Estimate effects for haplotypes of *HLA-DRB1*. For each haplotype, the multivariate effect is given as an odds ratio (OR), taking the most frequent haplotype (Pro-Ala-Ala) in the control samples as the reference (that is, given that the haplotype has an OR of 1). Classical shared epitope alleles are shown in bold. This table is modified from a previous report[12].

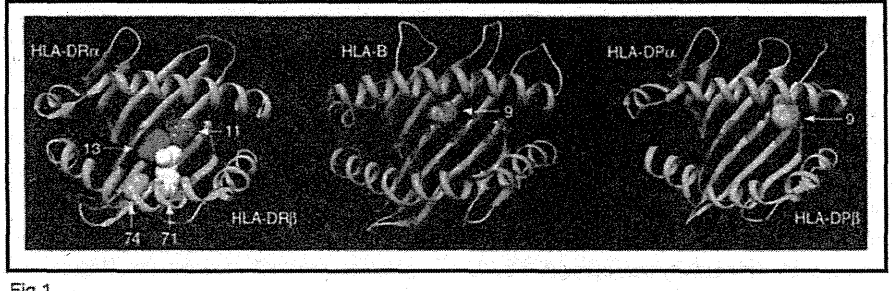

Fig.1
Three-dimensional ribbon models for the HLA-DR, HLA-B and HLA-DP proteins. Key amino acid positions identified by the association analysis are highlighted. This figure is taken from a previous report[12].

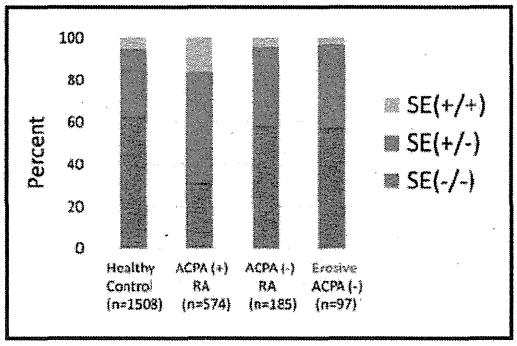

Fig.2
Prevalence of individuals carrying double SE, single SE or no SE is shown in healthy control, ACPA-positive RA, ACPA-negative RA and ACPA-negative RA with typical bone erosion as determined by X-ray. This clearly shows that ACPA-negative RA is distinct from ACPA-positive RA. This figure is illustrated based on our previous report[13].

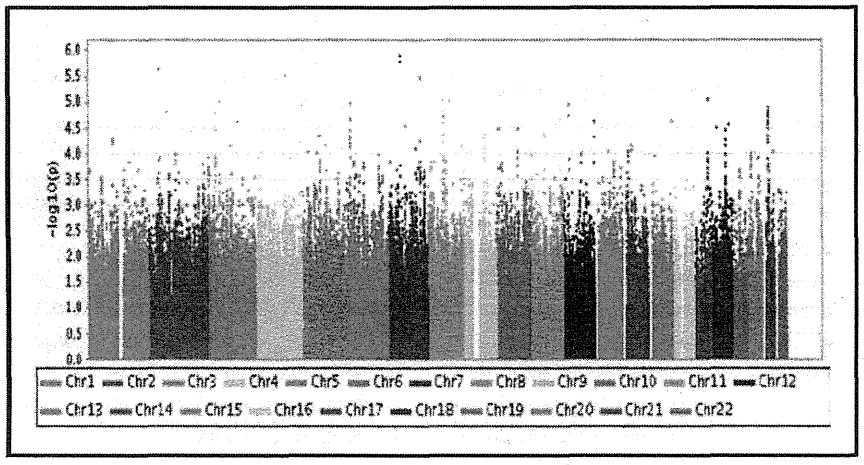

Fig 3
Probability plot for association with ACPA-negative RA (n=774) versus healthy controls (n=1079). This figure is taken from a previous report[6].

Ser11 is protective, for example, even though positions 71 and 74 are the SE types, Ser11 offsets such effects. Since most of the SE alleles have Valine or Leucine at position 11, Leucine at position 67, and do not have Serine at position 11 nor Aspatic acid at position 70, the results of previous studies using SE would not have been affected by the recent findings. Thus, key amino acid positions of HLA-DRβ chain for RA development seem to be 11th, 70th, 71st, and 74th positions and there still are some debates which positions have the primary effect. Anyway, these positions seem to be important for citrullinated peptide presentation.

HLA association with ACPA-negative RA

In 2005, a Dutch group reported that the association of SE was only exhibited with ACPA-positive RA and no association was seen with the ACPA-negative RA patients[10]. We have replicated the results in the Japanese population, and also showed that similar results were obtained even when we selected only bone-erosive ACPA-negative RA[13], which strongly suggests that this observation is not due to the contamination of non-RA arthritic diseases in ACPA-negative RA subset (Fig.2).

First of all, is there a genetic predisposition for ACPA-negative RA? From a twin study, heritability of ACPA-negative RA has been estimated and is thought to be as high as that of ACPA-positive RA[14].

Next, is *HLA* associated with ACPA-negative RA? A genome-wide association study (GWAS) meta-analysis of ACPA-negative RA showed that *HLA-DR* locus in chromosome 6 had no peak of association (see Fig.3)[6], suggest-

ing that the impact of *HLA* for development of ACPA-negative RA is not as large as that of ACPA-positive RA. In the study, the p-value of the *HLA* locus for ACPA-positive RA reached the order of 10^{-60}; in contrast, that for ACPA-negative RA reached the order of 10^{-4}. However, this does not mean that *HLA* is not associated with ACPA-negative RA, but probably means that ACPA-positive RA is a rather homogeneous subset in terms of *HLA* usage compared with ACPA-negative RA. ACPA-negative RA might have more variations of autoantigen (probably not citrullinated). In ACPA-positive RA, *HLA* usage is rather homogeneous, probably because citrullinated proteins or peptides are the common autoantigens among such patients that have SE-carrying *HLA*.

What *HLA* alleles are associated with ACPA-negative RA? In Caucasians, *HLA-DR3* and *DR13* have been reported to be associated with ACPA-negative RA[15-17]. As *HLA-DR3* association was seen in 3 independent European cohorts, it is probably true in Caucasians. In Japanese, we found that multiple *HLA-DRB1* alleles, including *12:01, *14:03 and *04:05, were associated with ACPA-negative RA susceptibility in the Japanese population[18]. *HLA-DR3* alleles were not shown because they are very rare in Japanese. We also found that *HLA-DRB1**15:02 and *13:02 were protective against ACPA-negative RA development. It is noteworthy that one of the SE alleles, *HLA-DRB1**04:05, was associated with ACPA-negative RA. Other SE alleles were not associated with ACPA-negative RA. This implies that the association of *04:05 with ACPA-negative RA is not due to the common amino acid sequence of SE because SE-carrying alleles other than *04:05 are not associated. Therefore, other mechanisms are suggested.

It seems there are two subsets in ACPA-negative RA based on RF positivity. Mackie *et al.* recently reported that *HLA-DRB1* SE is associated with ACPA(-)RF(+) RA but not with ACPA(-) RF(-) RA[19]. We have similar data for the Japanese population and showed that there are some specific *HLA-DRB1* alleles associated with ACPA(-) RF(+) RA or ACPA(-)RF(-) RA (Fig.4). For example, *04:05 and *09:01 were specifically associated with ACPA(-)RF(+) subset, and DR8/DR8 homozygote and DR14 were specifically associated with ACPA(-)RF(-) subset, whereas *12:01 was associated with both subsets. In contrast, ACPA(+)RA could not be separated by *HLA-DR* allelic usage.

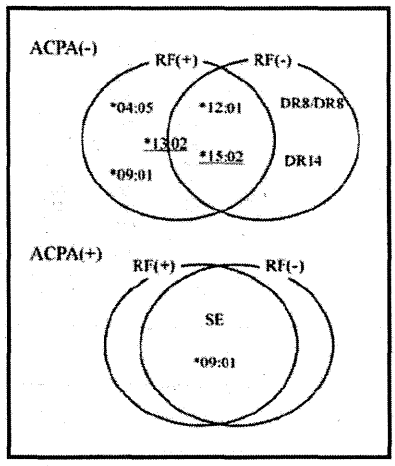

Fig.4
Scheme of HLA-DRB1 allele association with RF(+) or RF(-) subset of ACPA(-) RA or ACPA(+)RA in Japanese. Underline represents the protective allele. This figure is taken from our unpublished results.

Non-HLA genes associated with RA

A lot of genetic polymorphisms of candidate genes were tested for association with RA and reported to be associated with it, but most of them were not replicated. Perhaps the positive results are due to publication bias and relatively small sample sizes. Since 2003[20], genome-wide association studies (GWAS) have been applied to RA[21-26] and recently several meta-analyses of GWAS were performed[27-29]. Sample sizes also jumped from several hundred to tens of thousands. As a result, 30-40 genes or loci were detected to be significantly ($p<5\times10^{-8}$) associated with RA[1]. Many of these SNPs are located not in the genes (exons and introns), but near the genes, while some of the SNPs are located in exons and cause amino acid substitution (e.g. *PTPN22*). In many cases, the real causative SNPs or variants are still unknown. The list of SNPs in Table 2 shows the most strongly associated SNPs in the studies, but the real causative variants may exist somewhere else. The associated genes shown in Table 2 are classified by their main function. These genetic variants satisfied the genome-wide significance ($p<5\times10^{-8}$) or region-wide significance after Bonferroni's correction with multiple replication. Some of them are specific to Caucasians, mainly due to the absence of polymorphisms such as *PTPN22* and *RBPJ*, while some are specific to the Japanese or

Table 2 Candidate genes with confirmed association with rheumatoid arthritis

Gene	Best p-value	OR	Association† in		landmark SNP	SNP position	reference
			Caucasians	Japanese*			
(1)Intracellular signaling molecules and receptors							
PTPN22	9.1×10^{-74}	1.94	++	NA	rs2476601	exon	27
TRAF1-C5	4.0×10^{-14}	1.32	++	-	rs3761847	near	22
MBP	2.7×10^{-8}	1.23	-	++	rs2000811	intron	26
TNFAIP3	8.9×10^{-13}	1.22	++	++	rs6920220	near	27
BLK	5.7×10^{-9}	1.19	++	+	rs2736340	near	24
SPRED2	5.3×10^{-10}	1.13	++	+	rs934734	intron	27
TAGAP	3.8×10^{-7}	0.91	+	-	rs394581	near	44
TRAF6	3.9×10^{-8}	0.89	+	-	rs540386	intron	44
PTPRC	6.7×10^{-7}	0.88	+	-	rs10919563	intron	44
PRKCQ	4.4×10^{-8}	0.88	+	-	rs4750316	near	45
(2)Transcription factor							
REL	3.1×10^{-14}	1.25	++	-	rs13031237	intron	24
IRF5	4.2×10^{-11}	1.25	++	+	rs10488631/ rs13225818	near/near	27
STAT4	1.7×10^{-11}	1.24	++	++	rs7574865	intron	46
RBPJ	1.0×10^{-10}	1.18	++	NA	rs874040	near	27
AIRE	3.6×10^{-9}	1.18	-	++	rs2075876	intron	33
AFF3	1.0×10^{-14}	1.15	++	+	rs11676922	near	27
PRDM1	2.1×10^{-8}	1.11	++	-	rs6822844	near	44
(3)Cytokines and cytokine receptors							
CCR6	7.7×10^{-19}	1.19	++	++	rs3093024	near	25
IL2RB	4.6×10^{-8}	1.13	++	-	rs3218253	intron	47
IL2RA	1.4×10^{-11}	1.11	++	-	rs706778	intron	27
TNFRSF14	1.1×10^{-7}	0.92	+	+	rs3890745	near	45
CCL21	3.9×10^{-10}	0.87	++	-	rs951005	near	27
ANKRD55-IL6ST	9.6×10^{-12}	0.85	++	-	rs6859219	near	27
IL2-IL21	5.6×10^{-8}	0.78	+	NA	rs6822844	near	46
(4)Membrane receptors and costimulatory molecules							
HLA-DRB1	$<10^{-299}$	2.88	++	++	rs6910071	exon	27
FCRL3	8.5×10^{-7}	2.15	+	+	rs10430455	near	48
CD244	7.0×10^{-8}	1.31	-	+	rs6682654	intron	49
CD2-CD58	1.0×10^{-9}	1.13	++	-	rs11586238	near	44
CD28	1.3×10^{-9}	1.13	++	-	rs1980422	near	44
FCGR2A	1.5×10^{-5}	1.12	+	NA	rs12746613	near	44
CTLA4	6.3×10^{-9}	0.86	++	+	rs231735	near	27
CD40	2.8×10^{-9}	0.85	++	-	rs4810485	intron	27
(5)Enzymes							
PADI4	4.6×10^{-8}	1.97	+	++	rs766449	intron	20
PXK	3.1×10^{-14}	1.13	++	NA	rs13315591	near	27
DDX6	1.1×10^{-8}	0.87	++	-	rs10892279	near	28
(6)Unknown							
KIF5A-PIP4K2C	8.8×10^{-8}	0.89	+	-	rs1678542	near	45
C5orf30	4.1×10^{-8}	0.93	++	-	rs26232	intron	27

NA: not applicable due to the lack of polymorphism in Japanese

*Associations in Japanese are mainly based on our recent reports[29].

†, ++: $p<5\times10^{-8}$, +: $1\times10^{-4}<p<5\times10^{-8}$ with confirmation in other studies, -: no association

Asians, such as *AIRE*, although the reasons for this are unknown.

It is noteworthy that the list of genes includes many T cell receptor (TCR) and costimulatory signal molecules, many NFκB signal molecules and some B-cell-activation molecules, clearly indicating the importance of T and B cells and inflammatory response, especially the NFκB signal pathway. Interestingly, many molecules such as PTPN22, TNFAIP3, CTLA4 and FCRL3 are negative regulators of receptor signaling.

Here we introduce some recently discovered RA-associated genetic polymorphisms.

1)CCR6

CCR6 encodes chemokine receptor 6, which is a surface marker of Th17, a subset of T helper cells producing IL-17. We identified that genetic variation of *CCR6* is associated with RA (p=7.7x10^{-19}, OR=1.19) in Japanese by the combination of GWAS and replication studies[25]. *CCR6* genetic polymorphism is also associated with RA in Caucasians (p=1.5x10^{-11}, OR=1.11)[27]. It is interesting that not only the identified marker SNP (rs3093024) but also the functional dinucleotide polymorphism (rs968334 and the adjacent new SNP: CA, CG and TG variants, TA was not detected) was found to be associated with CCR6 expression (CA< CG<TG) and serum IL-17 level. This is quite an important finding in that Th17 involvement in the RA pathogenesis was supported genetically because there are some arguments that Th17 is not as important in human RA as in the mouse arthritis models[30,31]. *CCR6* variant is more strongly associated with ACPA (+) RA and is also associated with Graves' disease and Crohn's disease.

2)AIRE

AIRE is a key regulatory molecule of self-antigen presentation in medullary thymic epithelial cells (mTEC). *AIRE* knockout mice lack expression of organ-specific peripheral antigens (e.g. insulin, salivary protein 1, type II collagen) in the mTEC of thymus, which leads to the development of organ-specific autoimmune diseases[32]. Combination of GWAS and replication studies in Japan revealed that genetic polymorphisms of the *AIRE* gene are associated with RA[33]. There were two SNPs with genome-wide significance, one of which is located in an intron and correlated with the decreased expression of *AIRE* gene. This is in concordance with *AIRE* knockout mice developing more rapid and severe collagen-induced arthritis[34]. The other SNP is located in exon 7, which introduces amino acid alteration (S278R) at the SAND domain, and these two SNPs are in strong linkage disequilibrium. Such altered AIRE molecule may have reduced AIRE function.

3)MBP

MBP encodes myelin basic protein, which is a constituent of the myelin sheath of peripheral nerves. We conducted GWAS and replication studies with 2 different cohorts and identified *MBP* as a susceptibility gene for RA[26]. We also found that ~70% of RA patients have anti-MBP antibody in the serum. This was surprising because MBP is an autoantigen for multiple sclerosis (MS) and RA patients do not show such neurological symptoms as MS patients do. However, soon we found that this is not so surprising. First, MBP has several isoforms and the long isoform of MBP is called Golli-MBP[35,36]. Identified SNP is located in the intron of *Golli-MBP*. *Golli-MBP* is expressed in the hematopoietic cells and was shown to function as a negative regulator of TCR signaling through PKC[37]. *Golli-MBP* knockout T cells showed stronger reaction than the wild-type T cells[26]. Moreover, we found that anti-MBP antibody in the sera of RA recognized citrullinated MBP, but not non-citrullinated MBP. Since MBP is a well-known antigen that is physiologically citrullinated and a number of citrullinated proteins are the targets of RA autoantibodies[39], it is not surprising that MBP becomes one of the targets of RA autoimmunity. However, it has not been well studied how the *MBP* polymorphism is linked to the pathogenesis of RA. The *MBP* polymorphism is not associated with RA in Caucasians.

4)TNFAIP3

The *TNFAIP3* gene encodes a cytoplasmic zinc finger protein that possesses both ubiquitination and deubiquitination properties and is a major negative regulator of TNF-induced NF-κB signaling pathways. *TNFAIP3* polymorphism showed relatively high odds ratio for RA in both Caucasians and Japanese (odds ratios of 1.22 and 1.35, respectively). Several different polymorphisms have been associated with autoimmunity, including a nonsynonymous coding SNP (Phe127Cys), with some evidence of reduced negative regulatory ability for TNF-induced NF-κB signaling[40]. In addition to *TNFAIP3*, a number of genes related to NF-κB signaling (e.g. *TRAF1*, *CD40*, *Rel* and

NFKBIE) were reported to be associated with RA, clearly indicating the importance of NF-κB signaling in the pathogenesis of RA.

In the near future: rare variants

The genetic influence of each polymorphism is very modest (OR mostly ranging from 1.1 to 1.5). Therefore, there is no obvious clinical utility to predict the development of RA with such polymorphisms. This may change as the obtained knowledge becomes more complete, but currently all the known genetic variants can explain only ~15% of the genetic component[41]. This will not change very much even though we have found >100 associated genes with common variants (SNPs). Since most of the GWASs adopt common SNPs with a population prevalence of >3-5%, there may be some rare genetic variants with high genetic impacts. Sialic acid acetylesterase (*SIAE*) is an enzyme that negatively regulates B lymphocyte antigen receptor signaling and is required for the maintenance of immunological tolerance. By sequencing the *SIAE* exons, various defective variants were found in various autoimmune diseases including RA[42]. Defective variants were found in only 2 out of 648 (0.3%) healthy European subjects, whereas 24 out of 923 (2.6%) autoimmune disease patients had defective variants (OR=8.62). The odds ratio for RA was 8.31. Although this result was not successfully replicated in a larger study[43], some unknown rare variants may have strong impacts on the development of RA.

Now that the sequencing technology has developed markedly and is becoming less expensive, finding rare genetic variants associated with RA by whole-genome sequencing is realistic. As a first step, researchers started sequencing only exons of the whole genome, which is called the exome sequence, because it is much more economical than whole-genome sequencing. However, in the very near future, it is announced that the whole-genome sequence of one person can be read for $1,000 in a day. From this point onwards, it will be more realistic to understand completely the impact of genetic variants on the development of RA.

Acknowledgements

We would like to appreciate Prof. Matsuda of Dept. of Genomic Medicine in Kyoto University for all the supports on the research of our genetic studies. We would also like to appreciate all the members of GARNET consortium for the meta-analysis of GWAS and the HLA data on the Japanese RA. This work was supported by Grants-in-aid from the Ministry of Health, Labor and Welfare of Japan and from the Ministry of Education, Culture, Sports, Sciences and Technology of Japan. We have no conflict of interests to be declared.

References

1) Bax M, van Heemst J, Huizinga TW, Toes RE: Genetics of rheumatoid arthritis: what have we learned? Immunogenetics. 2011; 63: 459-466.
2) Gregersen PK, Olsson LM: Recent advances in the genetics of autoimmune disease. Annu Rev Immunol. 2009; 27: 363-391.
3) van Venrooij WJ, van Beers JJ, Pruijn GJ: Anti-CCP antibodies: the past, the present and the future. Nat Rev Rheumatol. 2011; 7: 391-398.
4) Kroot EJ, de Jong BA, van Leeuwen MA, et al: The prognostic value of anti-cyclic citrullinated peptide antibody in patients with recent-onset rheumatoid arthritis. Arthritis Rheum. 2000; 43: 1831-1835.
5) Kallberg H, Padyukov L, Plenge RM, et al: Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. Am J Hum Genet. 2007; 80: 867-875.
6) Padyukov L, Seielstad M, Ong RT, et al: A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. Ann Rheum Dis. 2011; 70: 259-265.
7) Gregersen PK, Silver J, Winchester RJ: The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum. 1987; 30: 1205-1213.
8) Deighton CM, Walker DJ, Griffiths ID, Roberts DF: The contribution of HLA to rheumatoid arthritis. Clin Genet. 1989; 36: 178-182.
9) de Vries N, Tijssen H, van Riel PL, van de Putte LB: Reshaping the shared epitope hypothesis: HLA-associated risk for rheumatoid arthritis is encoded by amino acid substitutions at positions 67-74 of the HLA-DRB1 molecule. Arthritis Rheum. 2002; 46: 921-928.
10) Freed BM, Schuyler RP, Aubrey MT: Association of the HLA-DRB1 epitope LA(67, 74) with rheumatoid arthritis and citrullinated vimentin binding. Arthritis Rheum. 2011; 63: 3733-3739.
11) Mattey DL, Dawes PT, Gonzalez-Gay MA, et al: HLA-DRB1 alleles encoding an aspartic acid at position 70 protect against development of rheumatoid arthritis. J Rheumatol. 2001; 28: 232-239.

12) Raychaudhuri S, Sandor C, Stahl EA, et al: Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet. 2012; 44: 291-296.
13) Ohmura K, Terao C, Maruya E, et al: Anti-citrullinated peptide antibody-negative RA is a genetically distinct subset: a definitive study using only bone-erosive ACPA-negative rheumatoid arthritis. Rheumatology (Oxford). 2010; 49: 2298-2304.
14) Ding B, Padyukov L, Lundstrom E, et al: Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in the extended major histocompatibility complex region. Arthritis Rheum. 2009; 60: 30-38.
15) Verpoort KN, van Gaalen FA, van der Helm-van Mil AH, et al: Association of HLA-DR3 with anti-cyclic citrullinated peptide antibody-negative rheumatoid arthritis. Arthritis Rheum. 2005; 52: 3058-3062.
16) Vignal C, Bansal AT, Balding DJ, et al: Genetic association of the major histocompatibility complex with rheumatoid arthritis implicates two non-DRB1 loci. Arthritis Rheum. 2009; 60: 53-62.
17) Lundstrom E, Kallberg H, Smolnikova M, et al: Opposing effects of HLA-DRB1*13 alleles on the risk of developing anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. Arthritis Rheum. 2009; 60: 924-930.
18) Terao C, Ohmura K, Kochi Y, et al: A large-scale association study identified multiple HLA-DRB1 alleles associated with ACPA-negative rheumatoid arthritis in Japanese subjects. Ann Rheum Dis. 2011; 70: 2134-2139.
19) Mackie SL, Taylor JC, Martin SG, et al: A spectrum of susceptibility to rheumatoid arthritis within HLA-DRB1: stratification by autoantibody status in a large UK population. Genes Immun. 2012; 13: 120-128.
20) Suzuki A, Yamada R, Chang X, et al: Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet. 2003; 34: 395-402.
21) Consortium TWTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447: 661-678.
22) Plenge RM, Seielstad M, Padyukov L, et al: TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. N Engl J Med. 2007; 357: 1199-1209.
23) Plenge RM, Cotsapas C, Davies L, et al: Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet. 2007; 39: 1477-1482.
24) Gregersen PK, Amos CI, Lee AT, et al: REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nat Genet. 2009; 41: 820-823.
25) Kochi Y, Okada Y, Suzuki A, et al: A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. Nat Genet. 2010; 42: 515-519.
26) Terao C, Ohmura K, Katayama M, et al: Myelin basic protein as a novel genetic risk factor in rheumatoid arthritis--a genome-wide study combined with immunological analyses. PLoS One. 2011; 6: e20457.
27) Stahl EA, Raychaudhuri S, Remmers EF, et al: Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010; 42: 508-514.
28) Zhernakova A, Stahl EA, Trynka G, et al: Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet. 2011; 7: e1002004.
29) Okada Y, Terao C, Ikari K, et al: Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat Genet. 2012; 44: 511-516.
30) Kato H, Fox DA: Are Th17 cells an appropriate new target in the treatment of rheumatoid arthritis? Clin Transl Sci. 2010; 3: 319-326.
31) Genovese MC, Durez P, Richards HB, et al: One year efficacy and safety results of a phase II trial of secukinumab in patients with rheumatoid arthritis. Arthritis Rheum. 2011; 63: S149-S150.
32) Anderson MS, Venanzi ES, Klein L, et al: Projection of an immunological self shadow within the thymus by the aire protein. Science. 2002; 298: 1395-1401.
33) Terao C, Yamada R, Ohmura K, 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. 2011; 20: 2680-2685.
34) Campbell IK, Kinkel SA, Drake SF, et al: Autoimmune regulator controls T cell help for pathogenetic autoantibody production in collagen-induced arthritis. Arthritis Rheum. 2009; 60: 1683-1693.
35) Pribyl TM, Campagnoni CW, Kampf K, et al: The hu-

man myelin basic protein gene is included within a 179-kilobase transcription unit: expression in the immune and central nervous systems. Proc Natl Acad Sci USA. 1993; 90: 10695-10699.

36) Feng JM: Minireview: expression and function of golli protein in immune system. Neurochem Res. 2007; 32: 273-278.

37) Feng JM, Fernandes AO, Campagnoni CW, Hu YH, Campagnoni AT: The golli-myelin basic protein negatively regulates signal transduction in T lymphocytes. J Neuroimmunol. 2004; 152: 57-66.

38) Feng JM, Hu YK, Xie LH, et al: Golli protein negatively regulates store depletion-induced calcium influx in T cells. Immunity. 2006; 24: 717-727.

39) Conrad K, Roggenbuck D, Reinhold D, Dorner T. Profiling of rheumatoid arthritis associated autoantibodies. Autoimmun Rev. 2010; 9: 431-435.

40) Musone SL, Taylor KE, Lu TT, et al: Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. Nat Genet. 2008; 40: 1062-1064.

41) Raychaudhuri S: Recent advances in the genetics of rheumatoid arthritis. Curr Opin Rheumatol. 2010; 22: 109-118.

42) Surolia I, Pirnie SP, Chellappa V, et al: Functionally defective germline variants of sialic acid acetylesterase in autoimmunity. Nature. 2010; 466: 243-247.

43) Hunt KA, Smyth DJ, Balschun T, et al: Rare and functional SIAE variants are not associated with autoimmune disease risk in up to 66,924 individuals of European ancestry. Nat Genet. 2012; 44: 3-5.

44) Raychaudhuri S, Thomson BP, Remmers EF, et al: Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat Genet. 2009; 41: 1313-1318.

45) Raychaudhuri S, Remmers EF, Lee AT, et al: Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet. 2008; 40: 1216-1223.

46) Daha NA, Kurreeman FA, Marques RB, et al: Confirmation of STAT4, IL2/IL21, and CTLA4 polymorphisms in rheumatoid arthritis. Arthritis Rheum. 2009; 60: 1255-1260.

47) Barton A, Thomson W, Ke X, et al: Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. Nat Genet. 2008; 40: 1156-1159.

48) Kochi Y, Yamada R, Suzuki A, et al: A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. Nat Genet. 2005; 37: 478-485.

49) Suzuki A, Yamada R, Kochi Y, et al: Functional SNPs in CD244 increase the risk of rheumatoid arthritis in a Japanese population. Nat Genet. 2008; 40: 1224-1229.

ACPA-Negative RA Consists of Two Genetically Distinct Subsets Based on RF Positivity in Japanese

Chikashi Terao[1,2]*, Koichiro Ohmura[1]*, Katsunori Ikari[3], Yuta Kochi[4], Etsuko Maruya[5], Masaki Katayama[1], Kimiko Yurugi[6], Kota Shimada[7], Akira Murasawa[8], Shigeru Honjo[9], Kiyoshi Takasugi[10], Keitaro Matsuo[11], Kazuo Tajima[11], Akari Suzuki[4], Kazuhiko Yamamoto[12], Shigeki Momohara[3], Hisashi Yamanaka[3], Ryo Yamada[2], Hiroo Saji[5], Fumihiko Matsuda[2,13,14], Tsuneyo Mimori[1]

1 Department of Rheumatology and Clinical Immunology, Kyoto University Graduate School of Medicine, Kyoto, Japan, **2** Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan, **3** Institute of Rheumatology, Tokyo Women's Medical University, Tokyo, Japan, **4** Laboratory for Autoimmune Diseases, Center for Genomic Medicine, RIKEN, Yokohama, Japan, **5** HLA Laboratory, Kyoto, Japan, **6** Department of Transfusion Medicine and Cell Therapy, Kyoto University Hospital, Kyoto, Japan, **7** Department of Rheumatology, Sagamihara National Hospital, National Hospital Organization, Sagamihara, Japan, **8** Department of Rheumatology, Niigata Rheumatic Center, Niigata, Japan, **9** Rheumatoid Arthritis Center, Saiseikai Takaoka Hospital, Toyama, Japan, **10** Department of Internal Medicine, Center for Rheumatic Diseases, Dohgo Spa Hospital, Matsuyama, Japan, **11** Aichi Cancer Center Hospital and Research Institute, Nagoya, Japan, **12** Department of Allergy and Rheumatology, Graduate School of Medicine, University of Tokyo, Tokyo, Japan, **13** CREST program, Japan Science and Technology Agency, Kawaguchi, Saitama, Japan, **14** Institut National de la Sante et de la Recherche Medicale (INSERM) Unite U852, Kyoto University Graduate School of Medicine, Kyoto, Japan

Abstract

HLA-DRB1, especially the shared epitope (SE), is strongly associated with rheumatoid arthritis (RA). However, recent studies have shown that SE is at most weakly associated with RA without anti-citrullinated peptide/protein antibody (ACPA). We have recently reported that ACPA-negative RA is associated with specific HLA-DRB1 alleles and diplotypes. Here, we attempted to detect genetically different subsets of ACPA-negative RA by classifying ACPA-negative RA patients into two groups based on their positivity for rheumatoid factor (RF). HLA-DRB1 genotyping data for totally 954 ACPA-negative RA patients and 2,008 healthy individuals in two independent sets were used. HLA-DRB1 allele and diplotype frequencies were compared among the ACPA-negative RF-positive RA patients, ACPA-negative RF-negative RA patients, and controls in each set. Combined results were also analyzed. A similar analysis was performed in 685 ACPA-positive RA patients classified according to their RF positivity. As a result, HLA-DRB1*04:05 and *09:01 showed strong associations with ACPA-negative RF-positive RA in the combined analysis ($p = 8.8 \times 10^{-6}$ and 0.0011, OR: 1.57 (1.28–1.91) and 1.37 (1.13–1.65), respectively). We also found that HLA-DR14 and the HLA-DR8 homozygote were associated with ACPA-negative RF-negative RA ($p = 0.00022$ and 0.00013, OR: 1.52 (1.21–1.89) and 3.08 (1.68–5.64), respectively). These association tendencies were found in each set. On the contrary, we could not detect any significant differences between ACPA-positive RA subsets. As a conclusion, ACPA-negative RA includes two genetically distinct subsets according to RF positivity in Japan, which display different associations with HLA-DRB1. ACPA-negative RF-positive RA is strongly associated with HLA-DRB1*04:05 and *09:01. ACPA-negative RF-negative RA is associated with DR14 and the HLA-DR8 homozygote.

Citation: Terao C, Ohmura K, Ikari K, Kochi Y, Maruya E, et al. (2012) ACPA-Negative RA Consists of Two Genetically Distinct Subsets Based on RF Positivity in Japanese. PLoS ONE 7(7): e40067. doi:10.1371/journal.pone.0040067

Editor: Pierre Bobé, Institut Jacques Monod, France

Received March 10, 2012; **Accepted** May 31, 2012; **Published** July 6, 2012

Funding: This study was supported by Grants-in-aid from the Ministry of Health, Labor, and Welfare of Japan and from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, as well as by research grants from the Japan Rheumatism Foundation, the Waksman Foundation, and the Mitsubishi Pharma Research Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding received for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: a0001101@kuhp.kyoto-u.ac.jp (CT); ohmurako@kuhp.kyoto-u.ac.jp (KO)

Introduction

Rheumatoid arthritis (RA) is the most common cause of chronic arthritis worldwide and results in severe joint destruction [1]. Genetic and environmental factors have been shown to be associated with its onset [2–3]. Among the susceptibility genes to RA, HLA-DRB1 has been shown to be the strongest genetic determinant of RA susceptibility, and its association with RA susceptibility has been repeatedly shown to be independent of ethnicity [4–5]. A common amino acid sequence extending from the 70th to 74th in the HLA-DRβ chain, which is known as the "shared epitope (SE)", is considered to be the reason for the association between HLA-DRB1 and RA, and the association between the SE and RA has been reported to be ethnicity-independent [6–8]. However, recent studies have shown that the SE is strongly associated with RA patients who have anti-citrullinated peptide/protein antibodies (ACPA), which is a highly specific marker of RA [9], but that it is not or only weakly associated with RA without ACPA [7,10–11]. Among the various HLA-DRB1 alleles, HLA-DR3 [12] and HLA-DR13 [13] were reported to be associated with ACPA-negative RA in populations of European descent, but these results were not confirmed in a