Human Molecular Genetics, 2005, Vol. 14, No. 16 ***2305–2321***
doi:10.1093/hmg/ddi234
Advance Access published on July 6, 2005

Whole genome association study of rheumatoid arthritis using 27 039 microsatellites

Gen Tamiya[1,2], Minori Shinya[1,2], Tadashi Imanishi[3], Tomoki Ikuta[1,2], Satoshi Makino[1], Koichi Okamoto[1,2,4], Koh Furugaki[1,2,4], Toshiko Matsumoto[5], Shuhei Mano[1], Satoshi Ando[1], Yasuyuki Nozaki[5], Wataru Yukawa[2,5], Ryo Nakashige[5], Daisuke Yamaguchi[5], Hideo Ishibashi[2,6], Manabu Yonekura[2,7], Yuu Nakami[2,5], Seiken Takayama[7], Takaho Endo[1], Takuya Saruwatari[2,8], Masaru Yagura[1], Yoko Yoshikawa[9], Kei Fujimoto[1], Akira Oka[1], Suenori Chiku[10], Samuel E.V. Linsen[11], Marius J. Giphart[11], Jerzy K Kulski[1,12], Toru Fukazawa[13], Hiroshi Hashimoto[13], Minoru Kimura[1], Yuuichi Hoshina[14], Yasuo Suzuki[14], Tomomitsu Hotta[14], Joji Mochida[15], Takatoshi Minezaki[15], Koichiro Komai[16], Shunichi Shiozawa[16], Atsuo Taniguchi[17], Hisashi Yamanaka[17], Naoyuki Kamatani[2,17], Takashi Gojobori[2,18], Seiamak Bahram[19] and Hidetoshi Inoko[1,2,3,*]

[1]Department of Molecular Life Science, Course of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Bohseidai, Isehara, Kanagawa 259-1193, Japan, [2]Japan Biological Information Research Center, Japan Biological Informatics Consortium, Tokyo 135-0064, Japan, [3]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan, [4]Chugai Pharmaceutical Corporation Ltd., Gotemba, Shizuoka 412-8513, Japan, [5]Hitachi Software Engineering Corporation, Ltd., Tokyo 140-002, Japan, [6]Applied Biosystems Japan Ltd., Tokyo 104-0032, Japan, [7]Mitsui Knowledge Industry Corporation Ltd., Tokyo 164-8555, Japan, [8]NTT DATA Corporation Ltd., Tokyo 135-6033, Japan, [9]Nisshinbo Industries Inc., Chiba, Chiba 267-0056, Japan, [10]Fuji Research Institute Corporation Ltd., Tokyo 101-0054, Japan, [11]Department of ImmunoHaematology and Blood Transfusion, Leiden University Medical Center, 2300RC Leiden, The Netherlands, [12]Centre for Bioinformatics and Biological Computing, School of Information Technology, Murdoch University, Murdoch, Weatern Australia 6150, Australia, [13]Department of Rheumatology and Internal Medicine, Juntendo University, Tokyo 113-8421, Japan, [14]Department of Hematology, Rheumatology and Endocrinology, Course of Medical Science, and [15]Department of Orthopedics Surgery, Course of Surgical Science, Tokai University School of Medicine, Isehara, Kanagawa 259-1193, Japan, [16]Department of Rheumatology, Faculty of Health Science, School of Medicine, Kobe University, Kobe, Hyougo 654-0142, Japan, [17]Institute of Rheumatology, Tokyo Women's Medical University, Tokyo 162-8666, Japan, [18]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan and [19]INSERM-CReS, Immunogénétique Moléculaire Humaine, Centre de Recherche d'Immunologie et d'Hématologie, 67085 Strasbourg, France

Received April 29, 2005; Revised June 15, 2005; Accepted June 28, 2005

A major goal of current human genome-wide studies is to identify the genetic basis of complex disorders. However, the availability of an unbiased, reliable, cost efficient and comprehensive methodology to analyze the entire genome for complex disease association is still largely lacking or problematic. Therefore, we have developed a practical and efficient strategy for whole genome association studies of complex diseases by charting the human genome at 100 kb intervals using a collection of 27 039 microsatellites and the DNA pool-

*To whom correspondence should be addressed. Tel: +81 463-93-1121, ext. 2312; Fax: +81 463-94-8884; E.mail: hinoko@is.icc.u-tokai.ac.jp

ing method in three successive genomic screens of independent case–control populations. The final step in our methodology consists of fine mapping of the candidate susceptible DNA regions by single nucleotide polymorphisms (SNPs) analysis. This approach was validated upon application to rheumatoid arthritis, a destructive joint disease affecting up to 1% of the population. A total of 47 candidate regions were identified. The top seven loci, withstanding the most stringent statistical tests, were dissected down to individual genes and/or SNPs on four chromosomes, including the previously known 6p21.3-encoded Major Histocompatibility Complex gene, *HLA-DRB1*. Hence, microsatellite-based genome-wide association analysis complemented by end stage SNP typing provides a new tool for genetic dissection of multifactorial pathologies including common diseases.

INTRODUCTION

With the ongoing success at unraveling the molecular basis of Mendelian disorders, the genomic community is now poised to tackle the genetics of the inherently more sophisticated 'complex disorders', so-called because they are the fruit of numerous interactions between the individual's complex genetic background (a few or multiple alleles at multiple gene loci) and the environment (1). Mendelian disorders are comparatively rare, whereas complex diseases affect larger sections of the population and have become, especially in industrialized nations, of major public health concern. In contrast to Mendelian diseases, where the path to successful identification of the causative mutation—autosomal dominant, autosomal recessive or X-linked—is well defined with many examples, the search for underlying mutations/polymorphisms in complex diseases have enjoyed only a few clear-cut successes. In fact, linkage studies in multiplex families followed by positional cloning, which is the standard procedure for tracking monogenic diseases, have proved to be inadequate once applied to complex diseases. Among the many broadly defined segments of the human genome that have been linked to such diseases as rheumatoid arthritis (RA), asthma, schizophrenia and so on, only a few have been replicated in independent studies (2). Association studies on the other hand do not require a large number of sib-pairs for analysis, as they are based on genotyping specific markers or sets of anonymous markers in independent cohorts of affected and healthy individuals, although the main bottleneck is a combination of marker density and sample size.

Genetic association studies can be performed in two ways. The candidate gene approach is hypothesis-driven and directly bound by the systemic knowledge of a biological process, whereas whole genome association studies can theoretically tackle the entire genome at once in an unbiased fashion. The main bottleneck for the feasibility of the latter approach is the scarcity of dense polymorphic markers across the whole genome. In principle, two types of markers are at hand for disease association studies, the microsatellites and single nucleotide polymorphisms (SNPs), with each type of markers presenting advantages as well as inconveniences. In comparison with microsatellites, SNPs are thought to be genetically more stable, due to a lower mutation rate, they are bi-allelic in nature and show a rather low degree of heterozygosity (on average, ~20%) as well as a comparatively much shorter range of linkage disequilibrium (LD). This means that for an efficient pan-genome analysis, millions of SNP may need to be simultaneously analyzed. However, completion of the human 'Haplotype Map' (HapMap) project will bring down this number of testable SNPs to several hundred thousands so-called haplotype tag SNPs (3). Further, recent developments, such as DNA chip-based technology, have attained high-throughput and cost-effective SNP typing.

Microsatellites, if carefully chosen, are highly polymorphic, show a high degree of heterozygosity (on average, ~70%) and LD lengths in the 100 kb range (4–12) when compared with the shorter, ~30 kb, range for SNPs, probably due to their older age (8,13–17). Therefore, the advantage of microsatellite is that a collection of a relatively small number of polymorphic markers (e.g. tens of thousands of microsatellite markers versus hundreds of thousands or millions of SNPs) could make whole genome association analyses an immediate reality (18). Namely, a genome scan is first performed using microsatellite markers at orders of magnitude fewer than SNP markers for identification of the incriminated region(s) within the 100 kb range, followed by high-density SNP typing (kilobase range) in order to ultimately find the responsible base mutation(s) or polymorphism(s). We have previously tested this combined approach to narrow down disease critical regions to 100 kb by microsatellite typing (4–7,9–11) and then identify susceptible loci by SNP typing (19) within 100 kb segments of the 3.6 Mb Major Histocompatibility Complex (MHC, also called the HLA) region notorious for its strong association with a large number of so-called 'HLA-associated diseases'. The pertinence of this extension from the HLA region to the entire genome was corroborated by the recent finding that LD and variation in the HLA region were essentially not different from those in the rest of the genome (20).

Here, we report on our use of 27 037 microsatellites in the first human whole genome case–control association study of RA, a chronic multifactorial debilitating systemic inflammatory disease presumably of autoimmune etiology. Our methodology relies on four main components: (i) the identification of enough microsatellites in order to chart the genome at 100 kb intervals, (ii) a three-phased genomic screen, i.e. replication of the data in three independent case–control populations, in order to reduce the type I error rate (21,22), (iii) the confirmation of the 'pool association' by separate (unpooled) genotyping of individual DNAs for the positive microsatellite markers (23) and (iv) fine dissection to susceptible gene regions, again by genotyping individual DNAs with a set of SNPs surrounding the target area.

RESULTS

Charting the genome with a high-density set of microsatellites

Extending our investigation from the HLA region of chromosome 6 to encompass the entire genome required a level of resolution that was unavailable at the onset of our enterprise to undertake whole genome-wide association analysis. On the basis of the knowledge accumulated from a large number of recent data that the average length of LD between disease susceptible SNPs and nearby microsatellite alleles is $\geq$100 kb (4–12), a microsatellite-based map of the genome at a 100 kb density will therefore make a whole genome association analysis a reality. Although the LD pattern is variable between different regions of the human genome depending on several factors such as allele frequency, mutation and recombination, the use of average spacing of genetic markers across the entire genome is a practical solution in genome-wide association analysis prior to the availability of a genome-wide LD map. Therefore, our first step for genome-wide analysis was to collect enough microsatellite markers ($>$27 000 microsatellites, one microsatellite for every 100 kb) to cover the euchromatic area ($\sim$90%) of the human genome (3 Gb) (3×10^9 kb $\times$ 0.90 $\div$ 100 kb = 27 000). The remaining part of the genome was mostly heterochromatin restricted mainly to centromeres and telomeres, rich in repetitive sequences and believed to lack any expressed genes. This 100 kb spacing would enable us to doubly screen a 100 kb genomic interval for the presence of a disease susceptible loci by two neighboring microsatellites across the whole genome. Indeed, we believe that the screening and detection of two neighboring microsatellites on both sides of the susceptible locus, when the intervening genomic sequence of $\geq$100 kb is in LD, is the most logical, reliable and practical step to whole genome-wide analysis. When LD happens to be $<$100 kb, but $>$50 kb around the susceptible locus, a microsatellite on either side can detect it in association mapping. This means that the maximum length of LD by which microsatellite markers should detect susceptible locus in this method can be 50 kb.

Microsatellite sequences were computationally detected from all the chromosomes except for the Y chromosome which is known to contain few expressed genes in the human genome sequence (NCBI build 35), and polymerase chain reaction (PCR) primers were designed for the uniform amplification of selected repeats. Among the 66 089 microsatellites investigated, 27 158 polymorphic markers that corresponded to our selection criteria were selected (see Materials and Methods) and localized on the human genome draft sequence (Fig. 1). The great majority of microsatellite markers, 20 755 are reported here for the first time, whereas 6403 were previously known (the CEPH Genotype database: http://www.cephb.fr/cephdb and the CHLC Genetic Mapping database: http://lpgws.nci.nih.gov/html-chlc/ChlcMarkers.html). We eliminated 119 from our total list of markers because we found them to be located on the Y chromosome rather than on autosomes as initially reported. The remaining 27 039 microsatellites that we finally accepted for our association studies had an average heterozygosity of 0.67 $\pm$ 0.16, an average of 6.4 $\pm$ 3.1 alleles and an average spacing of 108.1 kb (SD = 64.5 kb; max = 930.1 kb) (Supplementary Material, Table S1). Among these 27 039 microsatellites, 77 markers have intervals of over 400 kb mainly due to the absence of any identifiable polymorphic markers with shorter intervals (see Supplementary Material, Table S1 and data not shown). However, only $\sim$5% of the entire human genome region (150 Mb) was limited to a resolution of $>$200 kb in LD due to the distance intervals between the polymorphic microsatellite markers of $>$200 kb (interval genomic segments of $>$200 kb between two neighboring microsatellites where microsatellites on both sides cannot detect the presence of a disease susceptible locus in the middle part away from both ends because of the 100 kb length of LD). If susceptibility genes are located in these intervals, we may have therefore momentarily lost the opportunity to find them.

Phased genomic screens using DNA pools

In order to bring down substantially the cost and the technical burden linked to genotyping thousands of microsatellites without losing any significant amounts of data, the DNA pooling method was implemented. Because the absolute equality of individual DNA quantities is the key factor in this methodology, we employed a highly accurate quantitative procedure to construct a pooled DNA template for PCR amplifications (11,24). This pool was composed of strictly measured DNA concentrations, extracted from 125 Japanese individuals. Given that multiple peak patterns representing marker polymorphisms (21) showed distinct patterns in our Japanese pool when compared with that in two European pools (due to different allelic distributions between two populations) (data not shown), we concluded that the multiple peaks reflected original length polymorphisms and not experimental artifacts. Therefore, the DNA pooling method enabled us to obtain the allele frequencies of microsatellites in 125 individuals at once just by the measurement of the heights of multiple peaks, an approach directly applicable to association analysis. The accuracy of the pooling method was confirmed by the absence of any significant difference ($P > 0.05$) in the allele frequencies obtained by pooling against individual typing (21) (Supplementary Material, Fig. S1). Because the measurement error of our pooling method is $<$2% (11) (Supplementary Material, Fig. S1), we calculated that the expected difference in the allele frequency between cases and controls of $<$4%, which corresponds to 1.5-fold, 1.2-fold and 1.1-fold genotype relative risk if the allele frequency is 10, 30 and 60%, respectively, may be missed in a genome-wide scan (23).

An initial set of 375 RA patients and an identical number of control samples, all of Japanese descent, were equally divided into three pairs of 125 cases and 125 controls each, in order to initiate the three-step genomic screen. In the first screening, 125 cases and 125 controls were subjected to association analysis using all of the 27 039 microsatellites. Among them, microsatellites showing statistical significance of $P < 0.05$ were subjected to a second screening phase with a separate 125 cases and 125 controls. The microsatellites showing statistical significance of $P < 0.05$ in the second screening were then subjected to a third and final screening step with another distinct 125 cases and 125 controls. The power estimates of association testing that we calculated for

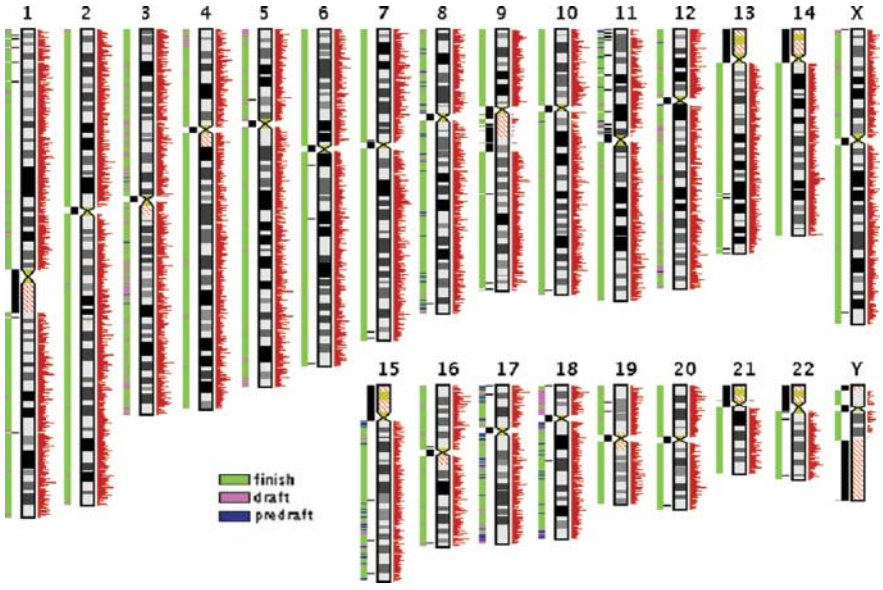

Figure 1. Distribution of 27 039 polymorphic microsatellite markers across the human genome. First column indicates status of draft sequence (green: finished, pink: draft and dark blue: predraft); black bars in second columns indicate sequence gaps, whereas red bars on right side of each chromosome represent the number of polymorphic marker per every 500 kb. These markers were mapped *in silico* on the NCBI build 35.

each of the three screenings were ~0.9 and 0.5 to detect a genotype relative risk of 1.8 and 1.5, respectively, when D' (degree of LD between the marker and disease-responsible allele) is 0.8 and the frequency of the microsatellite marker and disease-responsible allele is 0.25 (18,22). This means that, in three successive screens, the powers to detect a genotype relative risk of 1.8 and 1.5 are 0.73 (0.9^3) and 0.13 (0.5^3), respectively. Therefore, if a genotype relative risk is lower than 1.5 and/or frequency of disease-associated allele is much lower than 0.25, a considerable number of disease-associated microsatellites may be missed in this screening strategy.

Microsatellites that had remained statistically significant in all three screening steps were ultimately confirmed by individual genotyping using the same set of 375 patients and 375 controls. Such phased screens intended to sequentially replicate the results in the three independent sample populations are an essential step to eliminate many of the pseudo-positives resulting from type I errors (21,22). To calculate *P*-values, two types of the Fisher's exact test for the 2×2 contingency tables for each individual allele and the $2 \times m$ contingency tables for each locus were used, where m refers to the number of marker alleles observed in a population. The number of multiple comparisons in this mapping is $n + 1$, where n refers to the number of multiple comparisons in the 2×2 Fisher's exact test. Supposing that microsatellites are only bi-allelic, the first screening phase would theoretically include more than 1352 significantly associated pseudo-positive microsatellites ($27\,039 \times \alpha' = 2636$; $\alpha' = 1 - (1 - 0.05)^{n+1}$, $P < 0.05$ in the Fisher's exact test). Further, this number will drop to 257 ($2636 \times \alpha' = 257$) and 25 ($257 \times \alpha' = 25$) microsatellites in the second and third screenings, respectively. Because this microsatellite-based association analysis was followed by SNP association analysis using a DNA sample set including a further 565 cases and 565 controls (Table 2), the final list of markers would be expected to be free of the pseudo-positives. However, as the average allele number of 27 039 microsatellites used here is 6.4, multiple comparisons for each of the microsatellite alleles should be made in the *P*-value tests for each of the three screenings. The detection of more pseudo-positive markers might be expected with increased allelic numbers, although multiple comparisons for microsatellite alleles are not completely independent of each other when evaluating statistical independence. This issue will be discussed later ($n = 1.4$ for microsatellites used in this study, see Discussion). Prior to embarking on these screens, however, we verified through the Pritchard's method (25), using 69 randomly selected microsatellites from each of one ~22 and X chromosomes (enough to successfully perform such analysis), the absence of any significant stratification in either case or control populations (Supplementary Material, Table S2) (discussed subsequently). The accomplishment of this test is important in order to prevent the so-called 'spurious associations' generated by population stratifications, especially for late-onset diseases such as RA (26), where it is rather difficult to collect adequate internal controls.

In the first screen using all 27 039 microsatellites, we found significant association ($P < 0.05$) for 2847 markers as assessed by the Fisher's exact test, for the either 2×2 or $2 \times m$ contingency tables. In the second screen, 372 of these 2847 markers continued to show significant association, whereas after the third screen, the significant association was reduced to the 133 positive markers (Supplementary Material, Table S3 and Fig. S2) (Fig. 2). The relatively higher number of positive markers compared with what would be statistically expected might be partly due to the experimental artifacts inherent to the DNA pooling method, as this has been previously reported in analyses other than multiple testing (23,27). As the final step, we aimed to confirm each single positive marker by individually genotyping each and every

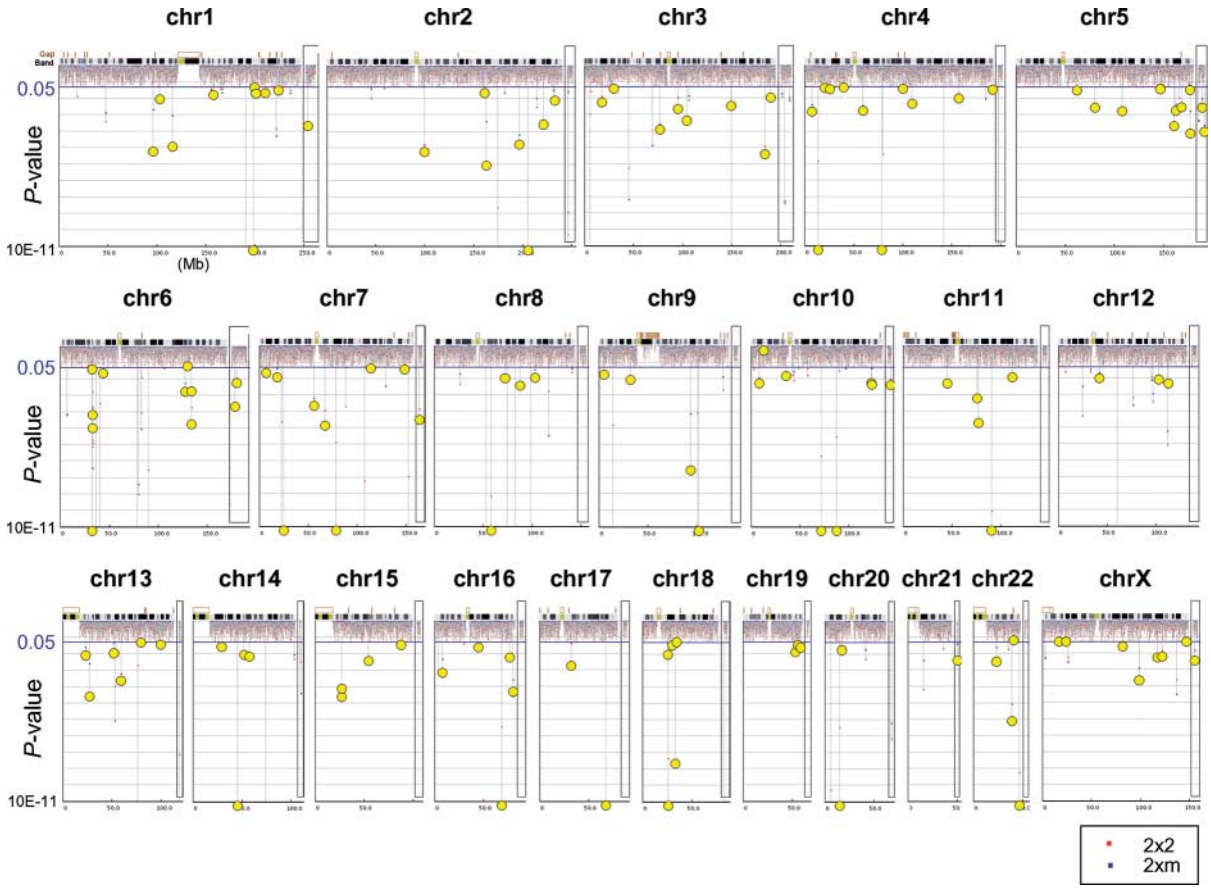

Figure 2. Genome-wide mapping of rheumatoid arthritis (RA) susceptibility loci using 27 039 microsatellites. One hundred and thirty three microsatellites showing significant association ($P < 0.05$) in the first, second and third screens by the DNA pooling method are indicated by yellow circles. Small dots (red and blue) pinpoint microsatellites failing to show significant association in the first, second or third screen with *P*-values. Red and blue confer to *P*-values obtained by the Fisher's exact test for the 2×2 and $2 \times m$ contingency tables, respectively. Dots in parenthesis on right side of each chromosome indicate microsatellites which could not be precisely mapped for their location on chromosomes.

DNA sample within the screened populations. Only 47 markers passed this ultimate test. Most of the other 86 failed markers possibly represented experimental artifacts associated with the DNA pooling method, such as PCR run to run variations in PCR conditions and in peak height in electrophoresis, PCR ghost peaks due to dissociation of labeled fluorescence reagent from a primer oligonucleotide, complications resulting from stutter and A addition bands which might be inherent to particular microsatellites and the effect of sample size in the association test (125 versus 375). Overall, the error or unsuccessful rate in our DNA pooling method was calculated as less than 3.2×10^{-3} (87/27 039). As a result, as many as 26 992 microsatellites ($27\,039 - 47 = 26\,992$) used as genetic markers showed no significant difference in allele frequencies between the patient and control groups. Among the 47 markers which survived three screenings by the DNA pooling technique followed by individual genotyping, the genomic segments around 11 markers were found to coincide with the ones previously suggested to be the RA gene candidate regions by genome-wide mapping based on linkage analysis, whereas the remaining 36 markers represented new RA-candidate regions defined for the first time by microsatellite-based association mapping in this study (data not shown).

As disease-responsible alleles with low frequency are supposed to explain a smaller fraction of the genetic susceptibility to disease when compared with those with high frequency, 24 of 47 positive markers were reserved for future analysis because of their low (<0.05) frequency, therefore leaving 23 positive markers for subsequent analysis (Table 1). Among the latter group, the seven microsatellites that revealed seven distinct genomic regions with the highest significance in association with RA were then subjected to fine mapping using SNPs.

Fine mapping by SNP and haplotype analysis

Among the seven most significant markers, four—the first (D6S0588i), the second (D6S0483i), the third (D6S1061) and the fifth (D6S0025i)—were located in the HLA region on chromosome 6p21.3, whereas the fourth (D11S0497i), the sixth (D10S0168i) and the seventh (D14S0452i) were located on chromosomes 11q13.4, 10p13 and 14q23.1, respectively (cytobands refer to the NCBI build 35) (Table 1 and Fig. 3). In order to further dissect each genomic region, we selected a collection of evenly spaced SNPs (coding and non-coding) within a several hundred kilobase perimeter

Table 1. Twenty-three positive microsatellite markers from individual genotyping

Markers	Cytobands	Number of allele	Positive allele	Allele frequencies		Fisher's exact *P*-values				Odds ratio	95% Cl
				Control	Case	2 × 2	P_c	2 × 2	P_c		
D6S0588i	6p21.3	10	5	0.430	0.572	0.000000055	0.000014	0	0	1.78	1.45–2.18
D6S0483i	6p21.3	18	7	0.089	0.176	0.00000092	0.00024	0	0	2.18	1.59–2.98
D6S1061	6p21.3	24	16	0.095	0.183	0.000001	0.00026	0	0	2.14	1.57–2.90
D11S0497i	11q13.4	5	2	0.513	0.613	0.000031	0.008	0.00052	0.012	1.55	1.26–1.91
D6S0025i	6p21.3	6	2	0.125	0.185	0.002	0.51	0.0005	0.012	1.59	1.20–2.11
D10S0168i	10p13	4	2	0.408	0.499	0.0005	0.13	0.001	0.024	1.44	1.18–1.77
D14S0452i	14q23.1	9	4	0.370	0.452	0.001	0.26	0.0006	0.014	1.40	1.14–1.72
D8S0127i	8q13.3	16	3	0.116	0.069	0.002	1	0.009	0.25	0.57	0.40–0.81
D7S0086i	7p21.1	11	4	0.095	0.053	0.002	1	0.03	0.75	0.54	0.36–0.80
D10S0607i	10q26.13	5	1	0.827	0.882	0.003	1	0.02	0.5	1.59	1.19–2.14
D13S0561i	13q31.1	10	8	0.130	0.183	0.005	1	0.16	1	1.50	1.13–2.00
G08462	5q14.1	9	4	0.190	0.136	0.005	1	0.09	1	0.67	0.51–0.89
D16S0496i	16q12.2	10	7	0.204	0.267	0.005	1	0.07	1	1.41	1.11–1.79
D5S0228i	5q12.1	11	7	0.305	0.371	0.004	1	0.02	0.5	1.35	1.09–1.67
D5S400	5q34	18	2	0.063	0.101	0.008	1	0.03	0.75	1.69	1.15–2.46
D6S0811i	6q22.33	6	3	0.445	0.515	0.008	1	0.01	0.25	1.31	1.07–1.61
D20S910	20p12.1	14	7	0.301	0.365	0.009	1	0.18	1	1.34	1.08–1.66
D4S0017i	4q25	22	5	0.071	0.111	0.009	1	0.12	1	1.64	1.14–2.35
D16S0232i	16q24.1	4	2	0.444	0.380	0.01	1	0.06	1	0.77	0.63–0.95
D3S1500i	3p24.3	4	1	0.781	0.725	0.01	1	0.005	0.13	0.74	0.58–0.94
D20S470	20p12.1	14	7	0.111	0.073	0.02	1	0.59	1	0.64	0.45–0.91
DXS0486i	Xq25	8	1	0.118	0.090	0.09	1	0.19	1	0.68	0.51–1.04
D18S0090i	18q12.1	20	13	0.193	0.153	0.05	1	0.54	1	0.76	0.58–0.99

P_c means corrected *P*-values by Bonferroni's correction. The Fisher's exact test was carried out in the case and control populations ($n = 375$ each). This means allele frequency of which has the lowest *P*-value in the locus.

surrounding each candidate region from the dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) and JSNP (http://snp.ims.u-tokyo.ac.jp) databases. In the HLA region, we selected additional SNPs from the *IkBL* to *C4B* genes to confirm previously reported associations around the centromeric end of the HLA class III region. These SNPs were selected from the Applied Biosystems SNP database (http://www.appliedbiosystems.com/).

We genotyped 165 SNPs after expanding our sample size (the combined population consisting of the previously tested 375 cases and 375 controls, and an additional population set composed of 565 cases and an equal number of controls, i.e. 940 patients and 940 healthy individuals) (Supplementary Material, Table S4). Among these 165 SNPs, 45 displayed a statistically significant ($P < 0.05$) association in the combined population (Table 2 and Fig. 3). Essentially the same results were obtained when statistical significance was assessed using only the newly recruited 565 cases and 565 controls (data not shown), indicating that these SNPs represented real-positive markers and not pseudo-positive ones. Of these positive SNPs, 25 remained significant ($P_c < 0.05$) even after Bonferroni's correction (Table 2). We then inferred the LD block structures for these 165 SNPs within the population using the EM algorithm (28) (Fig. 3) and carried out the case–control association study using the newly constructed haplotypes in each block (Table 3).

6p21.3

In the 6p21.3-located HLA region, 29 SNPs were statistically significant ($P < 0.05$) after the SNP association test of the combined ($n = 2 \times 940$) population (Table 2). Two previously known MHC associations were replicated. In the first instance, genotyping *HLA-DRB1* unveiled the *HLA-DRB1*0405* allele as the most significant RA-associated locus ($P = 9.7 \times 10^{-20}$; $P_c = 5.1 \times 10^{-18}$) in the combined ($n = 2 \times 940$) population (Table 2) as this allele is widely known to be associated with RA Japanese as well as other populations (29,30). Thus, the first and the fifth most strongly associated microsatellites, *D6S0588i* and *D6S0025i*, respectively, were in strong LD with the *HLA-DRB1* gene. In addition to *HLA-DRB1*, we were also able to confirm the association of the *IkBL* (MIM*601022) promoter SNP, rs3219185, with the disease ($P = 3.8 \times 10^{-6}$; $P_c = 2.0 \times 10^{-4}$) in the combined population (19), although the frequency of the minor allele in this SNP is relatively low. The association between *IkBL* and RA was not detected by microsatellites in our screening steps because no microsatellite near the *IkBL* gene was included in our marker set. Importantly, we also detected by SNP analysis two new candidate loci within the HLA region that were strongly associated with RA. These were SNPs around the *NOTCH4* (MIM*164951) and *Tenascin-XB (TNXB)* (MIM*600985) genes, which are located ~250 and 300 kb from *HLA-DRB1*, respectively (Fig. 3).

NOTCH4, a member of NOTCH transmembrane receptors family, is a proto-oncogene which contains epidermal growth factor (EGF) repeats. It is believed to be involved in signal transduction in a host of basic biological processes such as cell proliferation, cell differentiation as well as angiogenesis (31). Within *NOTCH4*, nine SNPs, among which two were non-synonymous, were significantly associated with RA. The strongest association ($P = 1.1 \times 10^{-11}$;

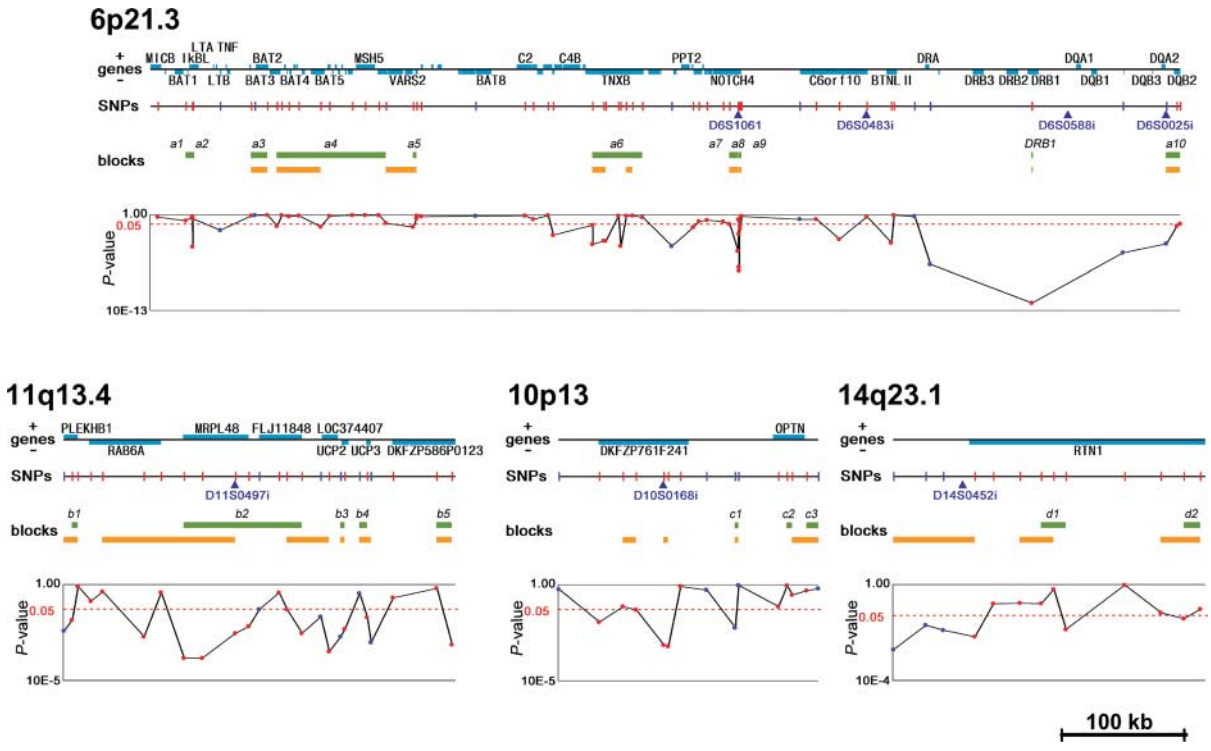

Figure 3. Single nucleotide polymorphisms (SNP) allelic association within the candidate regions. On each chromosome region, the first line refers to gene map whereas the second represents the SNP distribution. Below the second line, small blue arrowheads indicate positive microsatellite markers. In a long interval between the DRA and DQB1 genes, there were no suitable SNPs. Third line has linkage disequilibrium (LD) blocks inferred by the EM (green bar) and Clark (orange bar) algorithms. Graph represents SNP allelic associations. The plot shows log *P* versus the physical location in kb. Blue plot refers to intergenic SNPs, whereas red plot to genic SNPs [i.e. 5′-untranslated (UT), coding—synonymous and non-synonymous—3′-UT as well as intronic]. The red dashed line indicates significance threshold ($P = 0.05$).

$P_c = 5.8 \times 10^{-10}$) in the combined ($n = 2 \times 940$) population was observed for rs2071282, located within the fourth exon (encoding the fourth extracellular EGF repeat) and having a Leu203Pro substitution. On the other hand, rs915894 in exon 3 (third EGF repeat), although modestly significant ($P = 0.001$; $P_c = 0.052$), has a Lys116Gln amino acid exchange (Table 2).

The *TNXB* gene encodes a large extracellular matrix protein harboring 34 fibronectin type III-like (FNIII) and 18 EGF repeats. It is involved in a yet unknown manner in collagen fibril deposition in connective tissues (32). Within *TNXB*, five SNPs were significantly associated with RA, four of which were non-synonymous variations. Among these, rs185819 in exon 10 showed the strongest association ($P = 3.7 \times 10^{-5}$; $P_c = 1.9 \times 10^{-3}$) in the combined ($n = 2 \times 940$) population. It encodes a His1248Arg exchange in the seventh FNIII repeat. Other SNPs, rs2075563, Glu3260Lys, in exon 29 (26th FNIII repeat), rs2269428, His2363Pro, in exon 21 (18th FNIII repeat) and rs3749960, Phe2300Tyr, in exon 20 (17th FNIII repeat), were also significantly associated with RA (Table 2).

Haplotype analysis replicated these results for *IkBL, NOTCH4* and *TNXB* (Table 3). In order to estimate the influence of the *HLA-DRB1* on the associations of *IkBL, TNXB* and *NOTCH4* with RA, we carried out multiple logistic regression and Mantel-Haenszel tests for the SNPs in *IkBL, TNXB* and *NOTCH4* with those in *HLA-DRB1*. Multiple logistic regression showed that three genes were significantly RA-associated ($P < 0.05$) under a partially recessive model, *DRB1*0405* (odds ratios, ORs = 2.29–8.84), rs3219185 in *IkBL* (ORs = 1.16–2.66) and rs185819 in *TNXB* (ORs = 1.00–1.62) (Supplementary Material, Table S5). Under a partially dominant model, two intragenic SNPs, *DRB1* (ORs = 2.16–4.69) and *TNXB* (ORs = 1.02–2.01) were significant. In comparison, when focusing on the shared epitope (SE) of DRB1 (33) only under the partially recessive model SE (ORs = 1.79–3.87), *IkBL* (ORs = 1.11–2.54) and rs2071282 in *NOTCH4* (ORs = 1.13–7.14) were significant. These results may suggest that the four HLA candidate loci, *DRB1, IkBL, TNXB* and *NOTCH4*, can independently contribute to RA under a particular inheritance mode. The Mantel-Haenszel weighted ORs also supported the independent contribution of these candidate loci to the genetic susceptibility of RA (data not shown). In addition, when the stratification test of the association with RA in the subgroups without *DRB1*0405* or SE was carried out, the results indicated that the three *DRB1*-independent susceptible loci, *IkBL, TNXB* and *NOTCH4* in the HLA region can contribute to the development or maintenance of RA (Supplementary Material, Table S6).

11q13.4

The target region on 11q13.4 contains eight possible candidate genes *MRPL48, UCP2, UCP3, RAB6A, FLJ11848,*

Table 2. SNP allelic association

Cytobands	SNPs	Genes				Case : control = 940 : 940			P_c	Odds ratio	95% CI
					Allele	Frequencies		*P*-value*[2]			
		Name	Portion	Amino acid		Control	Case				
6p21.3	rs3219185	IkBL	promoter		G	0.929	0.964	0.0000038	0.00020	2.01	1.49–2.71
	rs769178	–			A	0.186	0.227	0.002	0.10	1.29	1.10–1.51
	rs2242656	BAT3	intron8		A	0.866	0.886	0.07	1	1.20	0.99–1.46
	rs805273	BAT5	intron4		C	0.866	0.886	0.07	1	1.20	0.99–1.46
	rs2242668	LSM2	intron2		A	0.882	0.901	0.08	1	1.21	0.99–1.49
	rs74534	DOM3Z	intron5		T	0.912	0.935	0.008	0.42	1.39	1.09–1.78
	rs2242569	TNXB	exon29		G	0.073	0.093	0.03	1	1.30	1.03–1.64
	rs2075563	TNXB	exon29^{+}*1	Glu3260Lys	G	0.106	0.162	0.00000076	0.00004	1.62	1.34–1.96
	rs2269428	TNXB	exon21^{+}	His2363Pro	A	0.107	0.159	0.000003	0.00016	1.58	1.30–1.91
	rs3749960	TNXB	exon20^{+}	Phe2300Tyr	T	0.107	0.160	0.0000024	0.00013	1.59	1.31–1.92
	rs185819	TNXB	exon10^{+}	His1248Arg	A	0.647	0.711	0.000037	0.0019	1.34	1.17–1.53
	rs204999	–			G	0.936	0.965	0.000042	0.0022	1.90	1.40–2.59
	rs2071289	EGFL8	exon6^{+}	Glu204Ala	A	0.019	0.036	0.002	0.10	1.92	1.28–2.89
	rs2849012	NOTCH4	intron7		G	0.692	0.762	0.0000016	0.000082	1.43	1.23–1.65
	rs520688	NOTCH4	exon5		G	0.326	0.408	0.00000022	0.000011	1.42	1.25–1.63
	rs2071284	NOTCH4	intron4		A	0.113	0.189	0.000000000080	0.0000000042	1.83	1.52–2.20
	rs2071283	NOTCH4	exon4		A	0.112	0.189	0.000000000057	0.000000003	1.84	1.53–2.21
	rs2071282	NOTCH4	exon4^{+}	Leu203Pro	T	0.113	0.193	0.000000000011	0.00000000058	1.87	1.56–2.25
	rs2071281	NOTCH4	exon4		T	0.113	0.189	0.00000000011	0.0000000058	1.82	1.52–2.19
	rs415929	NOTCH4	exon4		G	0.329	0.408	0.00000055	0.000029	1.41	1.23–1.61
	rs915894	NOTCH4	exon3^{+}	Lys116Gln	A	0.503	0.556	0.001	0.052	1.24	1.09–1.41
	rs443198	NOTCH4	exon3		T	0.504	0.569	0.000076	0.0039	1.30	1.14–1.47
	rs2273019	C6orf10	intron11		A	0.406	0.470	0.000091	0.0047	1.30	1.14–1.48
	rs2294878	BTNL2	intron2		C	0.644	0.733	0.0000000039	0.0000002	1.52	1.3–21.75
	rs2227139	–			A	0.607	0.718	0.00000000000087	0.000000000045	1.65	1.44–1.89
		HLA-DRB1			*0405	0.147	0.267	0.000000000000000000097	0.00000000000000000051	2.11	1.79–2.49
	rs2647012	–			A	0.827	0.887	0.00000013	0.0000067	1.65	1.37–1.98
	rs2071798	–			T	0.713	0.768	0.00015	0.0076	1.33	1.15–1.54
	rs1049110	HLA-DQB2	exon5^{+}	Gln 161Arg	A	0.778	0.802	0.07	1	1.16	0.99–1.36
11q13.4	rs3781909	–			C	0.428	0.480	0.001	0.052	1.24	1.09–1.41
	rs2008734	PLEKHB1	intron5		T	0.425	0.476	0.002	0.10	1.22	1.08–1.39
	rs2140893	RAB6A	intron1		C	0.509	0.564	0.00076	0.039	1.25	1.10–1.42
	rs17922174	MPRL48	5′-UTR		A	0.522	0.580	0.00045	0.023	1.26	1.11–1.44
	rs1792160	MPRL48	intron3		A	0.522	0.580	0.00035	0.018	1.27	1.11–1.44
	rs1792193	MPRL48	intron5		T	0.551	0.606	0.00075	0.039	1.25	1.10–1.43
	rs1061090	MPRL48	3′-UTR		C	0.971	0.976	0.4	1	1.21	0.81–1.80
	rs3741138	FLJ11848	exon7^{+}	Ala209Gly	C	0.833	0.862	0.01	0.73	1.25	1.05–1.50
	rs935985	FLJ11848	intron11		C	0.804	0.838	0.007	0.36	1.26	1.07–1.49
	rs637028	–			T	0.831	0.867	0.003	0.16	1.32	1.10–1.58
	rs653263	LOC374407	exon3		A	0.428	0.471	0.008	0.42	1.19	1.05–1.35
	rs655717	–			T	0.483	0.526	0.009	0.47	1.19	1.05–1.35
	rs660339	UCP2	exon4^{+}	Ala55Val	G	0.487	0.528	0.01	0.68	1.18	1.04–1.34

	rs2075577	UCP3	exon5	G	0.445	0.477	0.05	1	1.14	1.00–1.29
	rs1800849	–		G	0.667	0.711	0.004	0.21	1.23	1.07–1.41
	rs1527302	DKFZP586P0123	intron2	T	0.657	0.703	0.003	0.16	1.24	1.08–1.42
10p13	rs2280076	DKFZP761F241	3′-UTR	A	0.768	0.768	1	1	1.00	0.86–1.16
	rs2668907	DKFZP761F241	intron2	A	0.438	0.453	0.4	1	1.06	0.93–1.21
	rs662141	DKFZP761F241	intron2	T	0.481	0.499	0.3	1	1.08	0.95–1.22
	rs1347979	–		G	0.828	0.858	0.01	0.73	1.25	1.05–1.50
14q23.1	rs725951	–		T	0.787	0.823	0.006	0.31	1.26	1.07–1.48
	rs2073318	–		G	0.784	0.823	0.002	0.10	1.29	1.10–1.51
	rs1980579	–		T	0.782	0.822	0.002	0.10	1.29	1.10–1.51
	rs1950789	RTN1	intron8	C	0.810	0.848	0.002	0.10	1.31	1.11–1.56
	rs182138	RTN1	intron3	C	0.788	0.835	0.0002	0.012	1.36	1.16–1.61
	rs927326	RTN1	intron1	A	0.462	0.502	0.02	0.83	1.17	1.03–1.33

*1 Nonsynonymous SNPs.
*2 Fisher's exact test *P*-value in 2 × 2table of alleles.

Table 3. LD blocks and haplotype association with RA

Cytobands	Block*		SNPS		Included genes	Number of SNPs	Number of haplotype	Positive haplotype	Haplotype frequencies				Fisher's exact *P*-values		Odd ratio	95% CI
	Name	Size (kb)	End	Start	Name				Control	95% CI	Case	95% CI	2 × 2	P_c		
6p21.3	*a1*	8.26	rs2071595	rs2071592	BAT1-IkBL	5	5	4	0.074	0.055–0.093	0.028	0.018–0.040	0.000033	0.0037	0.35	0.21–0.59
	a2	0.04	rs2239708	rs2071591	IkBL	2	3	1	0.451	0.415–0.485	0.495	0.459–0.0531	0.10	1	1.19	0.97–1.46
	a3	19.03	rs2269475	rs1046089	BAT2	3	4	4	0.008	0.0003–0.015	0.012	0.005–0.020	0.45	1	1.51	0.53–4.26
	a4	127.97	rs2242656	rs707929	BAT3-C6orf27	10	11	4	0.072	0.056–0.090	0.030	0.019–0.041	0.00014	0.016	0.39	0.23–0.64
	a5	4.13	rs2242668	rs2075800	LSM2-HSPA1L	2	3	3	0.126	0.102–0.150	0.089	0.070–0.109	0.024	1	0.68	0.49–0.94
	a6	58.45	rs2242569	rs429150	TNXB	9	9	2	0.362	0.326–0.399	0.264	0.232–0.296	0.000046	0.0051	0.63	0.50–0.78
	a7	9.75	rs206018	rs2849012	NOTCH4	2	3	1	0.673	0.640–0.706	0.774	0.745–0.805	0.000015	0.0017	1.66	1.32–2.09
	a8	0.65	rs422951	rs415929	NOTCH4	8	3	3	0.103	0.083–0.126	0.203	0.175–0.231	0.00000010	0.000011	2.20	1.64–2.95
	a9	0.02	rs915894	rs443198	NOTCH4	2	4	2	0.493	0.460–0.531	0.432	0.400–0.465	0.020	1	0.78	0.64–0.96
	DRB1	0.26	rs2308754	rs1141742	DRB1	64	29	*405	0.129		0.276		0.0000000000013	0.00000000014	2.67	2.04–3.49
	a10	16.35	rs2071798	rs2071550	DQB2	3	4	4	0.069	0.052–0.089	0.025	0.014–0.037	0.000077	0.0086	0.35	0.20–0.60
11q13.4	*b1*	6.86	rs2008734	rs6590	PLEKHB1	2	3	2	0.412	0.376–0.447	0.479	0.445–0.513	0.011	1	1.31	0.07–1.61
	b2	139.37	rs1792174	rs93.5985	MRPL48-FLJ11848	8	6	1	0.500	0.461–0.533	0.595	0.565–0.629	0.00019	0.021	1.48	1.21–1.82
	b3	4.89	rs655717	rs660339	UCP2	2	2	2	0.454	0.417–0.489	0.535	0.501–0.569	0.0027	0.31	1.38	1.12–1.69
	b4	9.02	rs668514	rs2075577	UCP3	2	3	2	0.352	0.320–0.387	0.273	0.240–0.306	0.0010	0.11	0.69	0.55–0.86
	b5	17.81	rs866650	rs1527302	DKFZP586P0123	2	3	2	0.353	0.320–0.386	0.267	0.238–0.297	0.00044	0.049	0.67	0.54–0.84
10p13	*c1*	3.89	rs1347979	rs571066		2	4	3	0.175	0.150–0.203	0.120	0.097–0.145	0.0035	0.39	0.64	0.48–0.86
	c2	6.07	rs2244380	rs765884	OPTN	2	3	3	0.101	0.079–0.123	0.121	0.100–0.145	0.25	1	1.22	0.89–1.69
	c3	13.64	rs999999	rs1324252	OPTN	2	3	3	0.039	0.027–0.052	0.047	0.032–0.063	0.52	1	1.22	0.74–2.01
14q23.1	*d1*	28.86	rs1952043	rs2182138	RTN1	3	5	3	0.207	0.180–0.238	0.158	0.133–0.185	0.014	1	0.72	0.55–0.93
	d2	19.22	rs927326	rs2064992	RTN1	2	4	2	0.457	0.423–0.493	0.509	0.472–0.545	0.050	1	1.23	1.01–1.51

*LD blocks were inferred by the EM algorithm.

LOC374407, DKFZP586P0123 and *PLEKHB1*. Three of these genes, *MRPL48, UCP2* and *UCP3*, are mitochondrial-related genes (Fig. 3). *MRPL48* was recently identified on the basis of homology to mammalian mitochondrial ribosomal proteins (MRPs) (34). *UCP2* (MIM*601693) and *UCP3* (MIM*602044) encode transporter proteins on the inner mitochondrial membrane, which are related to energy expenditure. *UCP2* has been implicated in the genetics of obesity and diabetes (35). The *RAB6A* gene encodes a RAS-associated protein (MIM*179513) and it is centromerically located with respect to *MRPL48*. Of the other genes, *FLJ11848* has WD40 repeats that are related to a wide variety of functions including cell–cell interactions (36). *LOC374407* appears to have protein homology to heat shock protein 40 homolog (*HSP40* homolog) and a structural similarity to spermatogenesis apoptosis-related protein. *DKFZP586P0123* has one protein kinase C conserved region. Finally, *PLEKHB1* (MIM*607651) encodes a conserved protein (94% mouse–human amino acid identity) containing a pleckstrin homology domain as well as several casein kinase II (see MIM*115440) phosphorylation sites and a potential protein kinase C phosphorylation site.

Within this group of possible candidate genes, 15 of 25 polymorphic SNPs were statistically significant after the SNP association test in the combined population. Although the positive SNPs were scattered throughout the candidate region, the first and second most significant associations were observed for two SNPs, rs1792174 ($P = 0.00045$) in 5′-UTR and rs1792160 ($P = 0.00035$) in intron 3 of the *MRPL48* gene. *MRPL48* also had two other positive SNPs, rs1792193 ($P = 0.0076$) in intron 5 and rs1051090 ($P = 0.0075$) in the 3′-UTR region. Positive SNPs were also observed in *UCP2, UCP3, RAB38* and *FLJ11848*. However, only one common haplotype in the block *b2*, including *MRPL48* and *FLJ11848*, showed a significant association that was as strong as the single SNP in *MRPL48* (Fig. 3). The positive SNPs in *MRPL48* were also confirmed after Bonferroni's correction in the combined population (Table 2), indicating that *MRPL48* is the strongest candidate locus for RA in this region.

10q13, 14q23.1 and *PADI4*

The candidate region at chromosomal position 10p13 contains two genes of interest, *DKFZP761F241* of yet unknown function and *optineurin* (OPTN) which encodes an optic neuropathy inducing protein involved in the development of primary open-angle glaucoma (37) (Fig. 3). Only one SNP (rs1347979) in *DKFZP761F241* remained significant in the combined population. Although this SNP association was not confirmed after correction in the combined population, *DKFZP761F241* still remains the RA-susceptibility locus in the region most likely via a yet-to-be identified SNP(s). The candidate region on 14q23.1 contained a single locus, *reticulon 1* (MIM*600865), a member of a group of neuroendocrine-specific proteins. Even after Bonferroni's correction in the combined samples, rs2182138 in the third intron of *reticulon 1* remained statistically significant ($P = 0.0002$).

Finally, LD-mapping using SNPs in a previously reported linkage group for RA identified peptidylarginine deiminases 4 *(PADI4)* as an RA-susceptibility locus (38). We replicated four positive SNPs, padi89 ($P = 0.002$), padi90 ($P = 0.004$), rs874881 ($P = 0.002$) and rs2240340 ($P = 0.002$) in our population study of *PADI4*. Moreover, we confirmed that the microsatellite, *D1S1144i*, in intron 6 of the *PADI4* gene was included in the 47 microsatellite set that passed as a positive marker with a marginally significant RA-association ($P = 0.008$) in the three-phased pooling DNA screenings and in the individual genotyping test, although the associated allele frequency was low (3.7%) in the control population (data not shown).

Expression analysis

We performed a comprehensive human expression tissue-scan on ten of the identified RA-associated genes using quantitative reverse transcription–PCR (RT–PCR). The target tissues included synovial cell lines obtained from patients with RA and osteoarthritis (OA). Those from OA patients were employed as a control. We observed consistently high expression of *NOTCH4* in the lung and of *TNXB* in a number of tissues including the adrenal gland (Fig. 4). Our results also showed that all genes were expressed in the RA synovial cells, with *TNXB* and *NOTCH4* showing the highest level of expression in contrast to *RTN1* which displayed the lowest level. We also compared expression levels of these genes between RA and OA synovial cells where the latter was employed as a control (Table 4). By the Student's *t*-test, expression levels of *MRPL48* ($P = 0.049$) and *DKFZP761F241* ($P = 0.027$) genes showed a relatively significant difference between RA and OA synovial tissues, which may support our preliminary association data that *MRPL48* and *DKFZP761F241* are involved in the pathogenesis of RA. The expression of *MRPL48* in the RA synovial tissues was about twice the levels of the OA tissues. Three-quarters of the RA tissue donors were homozygous for a positive haplotype in the block *b2* of the *MRPL48* locus.

DISCUSSION

The strategy we report here for genome-wide association mapping of complex disease relies on a sequential case–control analysis starting with three separate screens of DNA pools and ending with individual genotyping of positive markers that had successfully passed a stringent selection process (23). Incriminated regions were then genetically dissected down to candidate susceptible loci using SNP and haplotype analysis. More than 99% of the 27 039 microsatellites collected as polymorphic markers in the Japanese population in this study were found to be also polymorphic in the Caucasian population and so probably can be applied to genome-wide association mapping of complex disease in most of the world-wide populations. In order to improve the power of the DNA pooling method, multiple measurements of pools and multiple moderately sized pools such as 50 may be incorporated in future studies to take advantage of the allele-frequency estimates from them (23).

The numbers of cases and controls ($N = 125$) employed in our screening system were possibly not high enough to

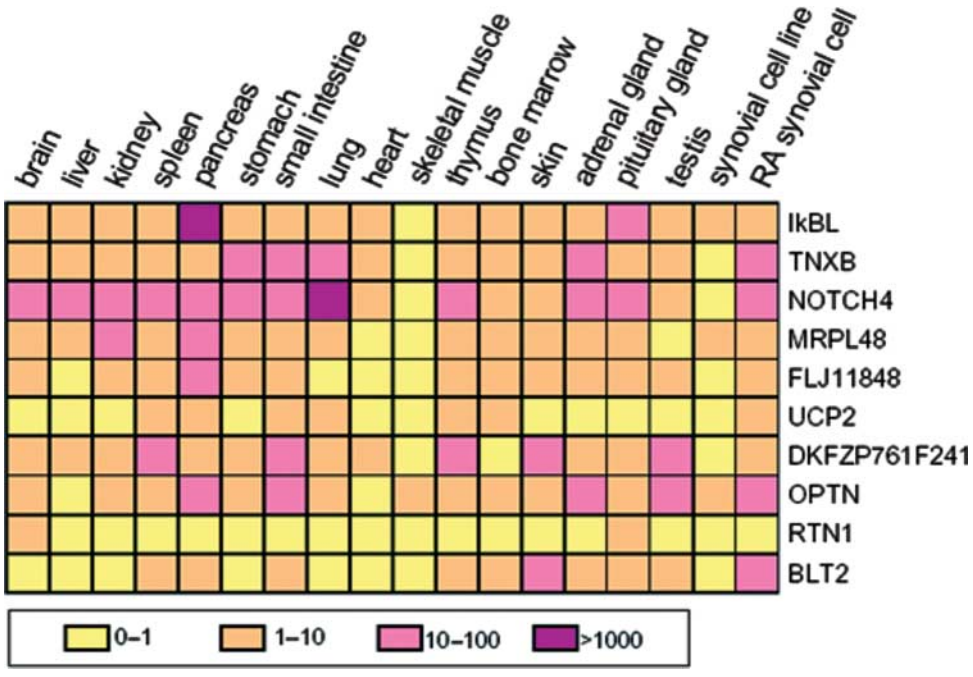

Figure 4. Gene expression analyses by quantitative reverse transcription–polymerase chain reaction (RT–PCR). The *BLT2* (leukotriene B4 receptor subtype 2) gene was employed as a positive control, which has been known to have strong expression in the RA synovial tissues (65). Each cell indicates average expression level of the standardized quantity data after rejection by the Smirnov's test.

Table 4. Expression levels of RA candidate genes in OA and RA synovial cells

	OA synovial cell		RA synovial cell	
Gene	Average	SD	Average	SD
IkBL	3.1	2.4	1.0	0.8
TNXB	339.3	349.7	80.7	35.1
NOTCH4	36.7	0.4	39.7	30.8
MRPL48	2.5*	0.2	4.4	1.6
FLJ11848	0.9	0	1.3	0.7
UCP2	1.5	0.9	2.8	2.1
DKFZP761 F241	17.5*	8.1	7.1	0.8
OPTN	7.4	2.4	12.7	3.4
RTN1	0.8	0.1	0.9	0.6
BLT2**	51.2	70.7	11.1	6.9

*Expression levels of *MRPL48* ($P = 0.049$) and *DKFZP761F241* ($P = 0.027$) genes showed relatively significant difference between RA and OA synovial tissues.
**The *BLT2* (leukotriene B4 receptor subtype 2) gene was employed as a positive control, which has been known to have strong expression in the RA synovial tissues (65).

detect disease-responsible loci with low genetic contribution to the disease susceptibility because of the limited number of samples used for each of the three independent association screenings, as described in Results section. A considerable number of other disease-associated microsatellites, for which a genotype relative risk was lower than 1.5 and/or frequency of disease-associated alleles was low, may have been missed in our present screening system. Although we confirmed the *HLA-DRB1* and *PADI4* genes as RA-susceptible loci in this study, the recently identified RA loci, *RUNX1* (39), *SLC22A4* (39) and *PTPN22* (40), were not included in the 47 positive microsatellite RA-candidate regions that we identified by the three-phased screening method. Consequently, more cases and controls may need to be employed for each of the three screenings in order to increase the statistical power in future association analyses.

Employment of 300 cases and 300 patients in each of three screenings is considered to be large enough to detect susceptible loci with a genotype relative risk of 1.5 with 95% probability and with a power of association testing at more than 0.9 in a genome-wide LD-mapping system (18). The large difference between the predicted and observed positives at each of the first, second and third screenings was detected by assuming that the microsatellites are only bi-allelic ($n = 1$; n: number of times of multiple comparisons). This difference might be explained by independent multiple comparisons of more than two alleles of microsatellites in P-value tests because the average allele number of 27 039 microsatellites used here was 6.4. Of 47 microsatellites that survived all three screenings by the DNA pooling technique and individual genotyping, only seven most statistically significant markers were used to finally identify and examine the seven candidate susceptible loci by SNP analysis. Therefore, on the basis of differential selective screening, we reached a statistical confidence level where we believed that the top seven markers represented the real-positive markers in our association mapping study of RA. We also applied the 27 039 microsatellites to genome-wide association mapping of several other complex diseases such as psoriasis vulgaris, hypertension, diabetes mellitus, Parkinson disease, etc. and found that a similar number of microsatellites were significant in each of the first, second and third screenings, and individual typing as in the case of the RA study. On the basis of these observations and also

on the assumption that all of the 47 microsatellites remaining after a three-phased screen represent pseudo-positive markers in association mapping, then '*n*' (*n* refers to the number of multiple comparisons in the 2×2 Fisher's exact test) in our microsatellite mapping system was calculated to be 1.4. This is very low, when compared with the average number of alleles of 27 039 microsatellites used in this study, 6.4. Even if dozens of microsatellites still survive as pseudo-positive markers after a three-phased screen, most of them can be excluded theoretically in the following SNP association analysis by incorporating additional cases and controls. That means that SNPs identified by a three-phased screen followed by an association SNP test represent real-positive markers for diseases. More strictly, however, a second SNP association test employing another set of cases and controls might be preferred to completely eliminate pseudo-positive markers, although the power of detection may be decreased by an increase in the number of association tests.

The large difference between the predicted and observed positives at each of the first, second and third screenings might also be explained in part by experimental artifacts linked to the DNA pooling method (23,27). Microsatellites vulnerable to artifact formation by DNA pooling need to be identified and removed from those used in genome-wide association mapping. In addition, a more accurate and reliable DNA pooling typing technique may need to be developed to carry out microsatellite-based genome-wide association mapping of complex diseases more efficiently.

The technical aspects of our association study began with the analysis of microsatellite markers that were previously used in linkage analysis and in the development of a series of initial human genetic maps (41–43). However, these microsatellite markers were not previously applied consistently to whole genome association analyses possibly because their overall number was insufficient for efficient whole genome analysis and there was a lack of an adequate support technology. On this basis, our first step for genome-wide analysis was to collect 27 037 polymorphic microsatellite markers estimated to be of sufficient number to cover the euchromatic area (∼90%) of the human genome (3 Gb) at ∼100 kb intervals. Therefore, we developed and tested 20 755 new polymorphic microsatellite markers that represented the great majority used in our study of RA and are reported here for the first time. The new microsatellite markers, presented here, should be useful for future whole genome association studies of the many other human chronic and infectious diseases that have yet to be investigated systematically. As the LD pattern is variable between different regions of the human genome and so the LD length is <50 kb in some regions of the human genome, it may be better to collect more polymorphic microsatellite markers on the basis of LD map, namely genetic distance but not physical distance like this work, on the whole human genome. The number of polymorphic microsatellites on the human genome that can be used as genetic markers in association mapping was estimated to be approximately 200 000 according to our recent *in silico* screening of the entire human genome sequences. Therefore, it will be possible to collect much more polymorphic microsatellites and employ them in our genome-wide association method with a higher density in order to increase the statistical power of the detection of susceptible loci.

Advantages inherent in the use of microsatellites over dimorphic SNP markers in genetic association studies have been repeatedly emphasized (14,15,44–46). Disadvantage of microsatellite may be their high mutation rates, 10^{-3} to 10^{-4}/site/generation (one mutation/20 000 ∼200 000 years) when compared with the lower mutation rates, 10^{-8}/site/generation, of SNPs (47). Mutation of microsatellites that were associated originally with a disease susceptibility allele(s) may subsequently result in a loss of disease association between the mutated microsatellites and such susceptibility allele(s) in association analysis. Thus, if a susceptible SNP is contained on a haplotype which has multiple microsatellite alleles, no single microsatellite allele may have strong correlation with the susceptible SNP, especially with low allelic frequency (<10%) (18). In that case, it will be necessary to assess associations of microsatellite with respect to every possible combination of each allele of microsatellite to detect such a susceptible SNP in association mapping. However, a new mutated allele in the population might serve as a new informative genetic marker showing a longer LD due to its young age (15,18). This issue can only be satisfactorily addressed through more extensive microsatellite-based association mapping of a significant number of diseases. Nevertheless, it should be noted that the presence of a strong LD between *HLA-DRB1* and its nearby microsatellite alleles reported in this study as well as between *HLA-DQB1, HLA-A, HLA-B* or *HLA-C* and their nearby microsatellite alleles (48) were consistently recognized despite the fact that the age of HLA alleles are considered to be ancient (although species-specific), having been traced back to around the emergence of *Homo Sapiens* 200 000 ∼1 000 000 years ago (49). These facts suggest that microsatellites are genetically stable enough to be applied to large-scale association mapping if a high-density marker set is used, and even if mutated, the newly generated allele might serve as a new genetic marker.

We successfully detected and confirmed the already well-known susceptibility gene for RA, *HLA-DRB1*. In addition to this gene, through a combination of microsatellite and SNP analysis, we identified two new candidate RA loci, *TNXB* and *NOTCH4*, in the HLA candidate region marked by the microsatellite marker located 250 kb away from *HLA-DRB1*. These findings are not only consistent with the previous data that suggested the existence of distinct LD blocks containing these loci (20,50) but also hint to the existence of an additional MHC-linked loci for RA (51). Importantly, analyses using the multiple logistic regression and Mantel-Haenszel tests showed that the positive SNPs in *TNXB* and *NOTCH4* were independent of *HLA-DRB1*0405* or SE, under both partially dominant as well as partially recessive models (Supplementary Material, Tables S5 and S6). Mutations in *TNXB* have been identified in a number of patients suffering from the Ehlers–Danlos syndrome (MIM*600985), a disorder of connective tissue due to defects in fibrillar collagen structure, deposition and/or metabolism, where *TNXB* has been shown to be involved (33). Amino acid exchanges of the *TNXB* product may therefore functionally predispose to RA through a yet-to-be identified pathway in collagen metabolism. Given that the murine type II collagen-induced arthritis mimics human RA (52), a

similar involvement of collagen might be in play in human. Further, it must be noted that expression changes in TNXB were observed in synovial samples from RA (53). *NOTCH4* was the other MHC-linked RA-associated candidate gene identified in our study. As invasive hyperplastic synoviocytes are often linked with cartilage and joint destruction, it is not hard to envision how *NOTCH4* might be involved in this pathological activity. Accordingly, it was recently reported that tumor necrosis factor (TNF), a pivotal cytokine in RA pathogenesis, upregulates *NOTCH4* in rheumatoid synovial fibroblasts in contrast to normal synovial fibroblasts. Hence, it was suggested that the *NOTCH4* product might be involved in hyperproliferation of RA synovial cells, a prerequisite to joint destruction (54). We also found candidate RA loci on 11q13.4, 10q13 and 14q23.1. On 11q13.4 and 10q13, although *MRPL48* and *DKFZP761F241* functions are still unknown, significant differences in their mRNA expression levels between affected and unaffected synovial tissues tend to corroborate their involvement in RA pathogenesis. How the 14q23.1-located *reticulon 1* gene which encodes a member of a group of neuroendocrine-specific proteins is involved in the development of RA remains to be investigated. Nevertheless, to prove that the new candidate RA genes identified in this association study are true susceptible loci, additional evidence will depend on the identification of all SNPs in the candidate regions followed by additional SNP association studies and supporting data from functional analysis by directed assays in cell or animal model systems.

What is the ultimate extent of the genetic contribution to RA pathogenesis? It is noteworthy that the seven loci dissected by genetic mapping here might represent only a minor fraction of the total genetic component to RA as we have detected up to 47 markers that were significantly linked to RA. It will be, therefore, important to identify the other RA-susceptibility loci that are hidden in the remaining 40 uncharacterized candidate regions, although some of them might represent pseudo-positive markers. Because the strength of statistical significance depends largely on the genetic distance between microsatellite and disease susceptibility locus, the comparatively lesser significance for these 40 candidate regions, when compared with the top seven microsatellites investigated here, does not necessarily mean lesser contribution to disease pathogenesis *per se*.

Finally and in regard to the detailed analysis of the top seven microsatellite markers, it is clear that they were positioned on particular LD blocks (Fig. 3), related to the 'Clark blocks' rather than the 'EM blocks'. In many cases, positive microsatellite alleles were also apparently associated with positive SNP haplotypes in these blocks. Such association between them is likely because the combination between them roughly depends on each frequency in the random mating population. This observed relationship between microsatellite alleles and SNP haplotypes indicates a possible compatibility between our microsatellite set and the HapMap consortium data (55), as recently suggested (56,57). In this regard, our whole genome case–control association study supports the practical synergism between these two seemingly distinct approaches to not only use successfully in complex diseases gene mapping but also in human evolutionary investigations.

In conclusion, we have performed the first genome-wide association analysis of a complex human disease using the densest set of polymorphic microsatellite markers available to date with an original and multi-step methodology. The outcome is a rapid and efficient path for the detection of susceptibility genes for complex disorders. The successful accomplishment of our analysis for RA-susceptibility genes opens the door for parallel investigations into a host of other multifactorial disorders including the relatively frequent common diseases such as asthma, type II diabetes, obesity, atherosclerosis, schizophrenia and psoriasis. The whole genome association study of common diseases using our approach as outlined here may ultimately lead to the identification of hitherto untapped biological pathways, multiplying the number of molecular targets for the development of specific therapeutic agents.

MATERIALS AND METHODS

Subjects

A total of 940 individuals affected with RA and an equal number of healthy unrelated individuals of Japanese origin participated in this study. Upon the approval of our experimental procedures by the relevant ethical committee in each participating center, we obtained informed consent from all affected and healthy individuals whose DNA samples were used in the analyses. The RA diagnosis was made according to the American College of Rheumatology diagnostic criteria (58). All personal identifiers associated with medical information and blood samples were carefully eliminated and replaced with anonymous identities in each recruiting institution. The average age of onset for the disease was 47.7 ± 13.1 years in the case population and the gender ratio was 1:4 (male:female). The healthy control population was also optimally matched for age and gender. For all involved samples, we reconfirmed gender assignment by amelogenin genotyping (59). *HLA-DRB1* was genotyped by direct PCR sequencing according to previous protocols (60). Finally, the quality of DNA from each individual was PCR-checked prior to further analysis.

Microsatellite detection and PCR primer design

A number of high-speed programs (*Apollo*: detection of microsatellites from genomic sequence; *Discovery*: design of PCR primers by batch treatment; *gPCR*: localization of PCR primer on the human genome sequence; *MSMK*: microsatellites database in human; *MICOS*: viewer of microsatellite map on the human chromosomes) were developed in our laboratory in order to efficiently process the massive genomic sequences contained within each human chromosome. Microsatellite sequences displaying two to six repetitive units were detected using the Apollo program, which is also compatible with Sputnik, in four versions of the human genome draft sequence: Golden Path June 2004 to the NCBI build 35. Microsatellites were investigated for repeat polymorphisms using 200 healthy Japanese with the DNA pooling method (discussed subsequently). Our criteria for the selection of

microsatellites were (i) di-nucleotide repeats with more than ten times repeat and tri-, tetra- and penta-nucleotide repeats with more than five times repeat and (ii) polymorphic microsatellites with heterozygosity of >30%, but not those with heterozygosity of >85% to eliminate unstable and highly mutated microsatellites. The discovery program, compatible with PrimerExpress, automatically designed the PCR primers for a uniform reaction condition. We also chose PCR primers which contained no SNPs in the sequences in order to prevent differential amplification (23). Finally, using the gPCR program, compatible with e-PCR, we certified each primer set to amplify a single copy on the NCBI build 35. Detailed information on 27 039 microsatellites is available at the JBIRC (Japan Biological Information Research Center) homepage (http://www.jbirc.aist.go.jp/gdbs/).

DNA pool construction and typing

The DNA pooling method for microsatellite typing was performed according to the protocol of Collins *et al.* (24) after slight modification (11). DNA was extracted using the QIAamp DNA blood kit (QIAGEN) under standardized conditions to prevent variation of DNA quality. This was followed by a 0.8% agarose gel electrophoresis in order to check for DNA degradation and/or RNA contamination. Following measurement of optical density in order to check for protein contamination, the DNA concentration was determined through three successive measurements using the PicoGreen fluorescence assay (Molecular Probes) as previously described (24). The standardized pipetting and aliquoting of the DNA samples were robotically performed using Biomek 2000 and Multimek 96 (Beckman). The pooled DNA template for $2 \times 27\,039$ microsatellite typing was prepared immediately after the DNA quantification. The quality of the pooled DNA was ascertained by comparing allelic distributions between individual and pooled typing results using 96 microsatellite markers (Supplementary Material, Fig. S1). Measurement error of our pooling methods is <2% (11). Stuttering of peak heights did not introduce appreciable artifacts in allele frequency estimations. After the initial tests, the 27 039 PCR reaction mixtures containing all components except primers were prepared and then aliquotted into the 96-well reaction plates and stored until use. The microsatellite pooled typing and individual genotyping procedures after the PCR reaction were carried out according to standard protocols using the ABI3700 DNA analyzer (Applied Biosystems). The standardized preparations allowed reproducibility and accuracy to be maintained for the pooled DNA typing throughout the experiment. Various kinds of information such as peak positions and heights were automatically extracted by the PickPeak and MultiPeaks programs, developed by Applied Biosystems Japan, from the multipeak pattern in the chromatogram ABI fas files. Because peaks including those with stutter and shadow were automatically extracted and compared for their height between cases and controls by these programs in association studies, most of the positive markers which remained statistically significant could be confirmed by individual genotyping using the same set of patients and controls.

SNP genotyping

SNPs were selected around candidate regions from the dbSNP at the NCBI homepage (http://www.ncbi.nlm.nih.gov/projects/SNP/) from the JSNP database at the homepage of the Institute of Medical Science of Tokyo University (http://snp.ims.u-tokyo.ac.jp) as well as from the SNP database of Applied Biosystems (http://www.appliedbiosystems.com/). The SNPs were genotyped using the TaqMan assays or direct sequencing and all the information regarding these SNPs is listed in Supplementary Material, Table S4. The TaqMan assays were carried out using standard protocols for the ABI PRISM 7900HT Sequence Detection System using a 384-well block module and automation accessory (Applied Biosystems). The direct sequencing of PCR products was carried out according to standard procedures using the ABI3700 DNA analyzer (Applied Biosystems).

Statistical analyses

The Pritchard's method (25) was employed for the detection of stratification in case and control populations. To calculate *P*-values, we used two types of the Fisher's exact test for the 2×2 contingency tables for each individual allele and the $2 \times m$ contingency tables for each locus, where m refers to the number of marker alleles observed in a population. The Markov chain/Monte Carlo simulation method was employed to execute the Fisher's exact test for the $2 \times m$ contingency table. The simple 'allelic' but not 'genotypic' association was presented for the 2×2 contingency tables for microsatellites, SNPs and haplotypes. We corrected these *P*-values (P_c) by the Bonferroni's correction where the coefficient was the total number of the contingency tables tested. These analyses were carried out using the software package, MCFishman. Other basic statistical analyses, including the multiple logistic regression and Mantel-Haenszel tests, were carried out using the SPSS program package as well as Microsoft Excel. We inferred LD block structures for these SNPs using confidence intervals of the D' value as a LD measure (61,62). We also estimated haplotypes in each block and their frequencies by both the EM (28) and Clark (63) algorithms. Finally, to assess reliability of each block haplotype, we calculated the 95% confidence interval from each haplotype frequency distribution given by bootstrap resampling of up to 2000 times based on the estimated haplotype frequencies, which is implemented in the Right program (64).

Expression analysis

Total RNA was isolated by ISOGEN (Nippon Gene) from surgically obtained synovial membranes from eight RA and four OA patients, and from a synovial cell line (SW982) obtained from the American Type Culture Collection. RNAs from various other tissues were obtained commercially from Clontech, Invitrogen, Origene and Stratagene. The quality and quantity of these RNAs were assessed using Agilent 2100 Bioanalyzer (Agilent) and their quantity confirmed by the RiboGreen RNA fluorescence assay (Molecular Probes). Using these total RNAs as templates, complimentary DNAs were synthesized using random hexamers and TaqMan

Reverse Transcription Reagents kit (Applied Biosystems). cDNA specific primers and probes were obtained by the 'Assay-by-Design (AbD)' for the ten genes tested and by the 'Assay-on-Demand (AoD)' for *GAPD* (glyceraldehyde-3-phosphate dehydrogenase), used as a housekeeping control gene, all provided by Applied Biosystems. After preliminary experiments, we used a final concentration of 210 nM of probe, 756 nM of primers and 0.48 ng/μl cDNA in 50 μl reaction volume in 96-well reaction plates on ABI PRISM 7900, all according to standard procedures recommended by Applied Biosystems. We processed each plate three times and calculated the average and SD (standard deviation) for each sample. Quantity estimates were calculated each time using a standard curve in each well. All standardized quantity data were adjusted to *GAPD* and tested by the Smirnov's test with a 5% significance level. The *BLT2* (leukotriene B4 receptor subtype 2) gene was employed as a positive control, which has been known to have strong expression in RA synovial tissues (65). After the reciprocal transformation for all standardized quantity data, we carried out the Student's *t*-test for the expression levels between averages of the RA and OA synovial tissues.

SUPPLEMENTARY MATERIALS

Supplementary material is available at HMG Online.

ELECTRONIC DATABASE INFORMATION

Information on the physical map of microsatellites and primer sequences used for their amplification in this study have been deposited in Genbank and given the accession numbers provided in Supplementary Material Table 7 (Fig. 1) and also available at the JBIRC (Japan Biological Information Research Center) homepage (http://www.jbirc.aist.go.jp/gdbs/).

ACKNOWLEDGEMENTS

We would like to thank M. Tomizawa, E. Tokubo, A. Takaki, H. Ando, S. Adachi, K. Yoshida, Y. Makino, K. Kobayashi, T. Shinomiya, S. Harada, M. Matsuzawa and S. Yamamoto for technical assistance, T. Ichihara (Nisshinbo Research and Development Center), N. Yasuda and T. Tamura (JBIRC), S. Hashimoto and H. Sano (JBiC), Y. Eguchi (MKI), M. Morikawa (GenoDive Pharm) for suggestions or help in this work and finally J.-L. Mandel, M. Koenig (both at IGBMC) and J. Sibilia (Strasbourg University Hospital) for critical reading of the manuscript. This work was performed under the management of Japan Biological Informatics Consortium (JBIC) and supported by grants from the New Energy and Industrial Technology Development Organization (NEDO). This research was also supported by 'Special Coordination Funds for Promoting Science and Technology' from the Japan Science and Technology Agency and 'Research for the Future Program' from the Japan Society for the Promotion Science. S.B. and H.I. wish to thank an INSERM-JSPS collaborative grant. Funding to pay the Open Access publication charges for this article was provided by the grant from NEDO.

Conflict of Interest Statement. None declared.

REFERENCES

1. Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease.*Trends Genet.*, **17**, 502–510.
2. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
3. The International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
4. Oka, A., Tamiya, G., Tomizawa, M., Ota, M., Katsuyama, Y., Makino, S., Shiina, T., Yoshitome, M., Iizuka, M., Sasao, Y. *et al.* (1999) Association analysis using refined microsatellite markers localizes a susceptible locus for psoriasis vulgaris within a 111 kb segment telomeric of the *HLA-C* gene. *Hum. Mol. Genet.*, **8**, 2165–2170.
5. Ota, M., Mizuki N., Katsuyama, Y., Tamiya, G., Shiina, T., Oka, A., Ando, H., Kimura, M., Goto, K., Ohno, S. *et al.* (1999) The critical region for Behcet's disease in the human major histocompatibility complex is reduced to a 46 kb segment centromeric of HLA-B by association analysis using refined microsatellite mapping. *Am. J. Hum. Genet.*, **64**, 1406–1410.
6. Keicho, N., Ohashi, J., Tamiya, G., Nakata, K., Taguchi, Y., Azuma, A., Ohishi, N., Emi, M., Park, M.H., Inoko, H. *et al.* (2000) Fine localization of a major disease-susceptibility locus for diffuse panbronchiolitis. *Am. J. Hum. Genet.*, **66**, 501–507.
7. Mizuki, N., Ota, M., Yabuki, K., Katsuyama, Y., Ando, H., Palimeris, G.D., Kaklamani, E., Accorinti, M., Pivetti-Pezzi, P., Ohno, S. *et al.* (2000) Localization of the pathogenic gene of Behcet's disease by microsatellite analysis of three different populations. *Invest. Ophthalmol. Vis. Sci.*, **41**, 3702–3708.
8. Abecasis, D.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., Leaves, N.I., Anderson, G.G., Zhang, Y., Lench, N.J., Carey, A. *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.*, **68**, 191–197.
9. Ota, M., Katsuyama, Y., Kimura, A., Tsuchiya, K., Kondo, M., Naruse, T., Mizuki, N., Itoh, K., Sasazuki, T. and Inoko, H. (2001) A second susceptibility gene for developing rheumatoid arthritis in the human MHC is localized within a 70 kb interval telomeric of the *TNF* genes in the HLA class III region. *Genomics*, **71**, 263–270.
10. Matsuzaka, Y., Makino, S., Okamoto, K., Oka, A., Tsujimura, A., Matsumiya, K., Takahara, S., Okuyama, A., Sada, M., Gotoh, R. *et al.* (2002) Susceptibility locus for non-obstructive azoospermia is localized within the HLA-DR/DQ subregion: primary role of DQB1*0604. *Tissue Antigens*, **60**, 53–63.
11. Oka, A., Hayashi, H., Tomizawa, M., Okamoto, K., Suyun, L., Hui, J., Kulski, J.K., Beilby, J., Tamiya, G. and Inoko, H. (2003) Localization of a non-melanoma skin cancer susceptibility region within the major histocompatibility complex by association analysis using microsatellite markers. *Tissue Antigens*, **61**, 203–210.
12. Zhang, Y., Leaves, N.I., Anderson, G.G., Ponting, C.P., Broxholme, J., Holt, R., Edser, P., Bhattacharyya, S., Dunham, A., Adcock, I.M. *et al.* (2003) Positional cloning of a quantitative trait locus on chromosome 13q14 that influences immunoglobulin E levels and asthma. *Nat. Genet.*, **34**, 181–186.
13. Koch, H.G., McClay, J., Loh, E.W., Higuchi, S., Zhao., J.H., Sham, P., Ball, D. and Craig, I.W. (2000) Allele association studies with SSR and SNP markers at known physical distances within a 1 Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb. *Hum. Mol. Genet.*, **9**, 2993–2999.
14. Mohlke, K.L., Lange, E.M., Valle, T.T., Ghosh, S., Magnuson, V.L., Silander, K., Watanabe, R.M., Chines, P.S., Bergman, R.N., Tuomilehto, J. *et al.* (2001) Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. *Genome Res.*, **11**, 1221–1226.
15. Terwilliger, J.D., Haghighi, F., Hiekkalinna, T.S. and Goring, H.H. (2002) A biased assessment of the use of SNPs in human complex traits. *Curr. Opin. Genet. Dev.*, **12**, 726–734.
16. Shifman, S., Kuypers, J., Kokoris, M., Yakir, B. and Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.*, **12**, 771–776.
17. Varilo, T., Paunio, T., Parker, A., Perola, M., Meyer, J., Terwilliger. J.D. and Peltonen, L. (2003) The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Hum. Mol. Genet.*, **12**, 51–59.

18. Ohashi, J. and Tokunaga, K. (2003) Power of genome-wide linkage disequilibrium testing by using microsatellite markers. *J. Hum. Genet.*, **48**, 487–491.
19. Okamoto, K., Makino, S., Yoshikawa, Y., Takaki, A., Nagatsuka, Y., Ota, M., Tamiya, G., Kimura, A., Bahram, S. and Inoko, H. (2003) Identification of I kappa BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. *Am. J. Hum. Genet.*, **72**, 303–312.
20. Walsh, E.C., Mather, K.A., Schaffner, S.F., Farwell, L., Daly, M.J., Patterson, N., Cullen, M., Carrington, M., Bugawan, T.L., Erlich, H. *et al.* (2003) An integrated haplotype map of the human major histocompatibility complex. *Am. J. Hum. Genet.*, **73**, 580–590.
21. Barcellos, L.F., Klitz, W., Field, L.L., Tobias, R., Bowcock, A.M., Wilson, R., Nelson, M.P., Nagatomi, J. and Thomson, G. (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.*, **61**, 734–747.
22. Saito, A. and Kamatani, N. (2002) Strategies for genome-wide association studies: optimization of study designs by the stepwise focusing method. *J. Hum. Genet.*, **47**, 360–365.
23. Sham, P.C., Bader, J.S., Craig, I., O'Donovan, M. and Owen, M. (2002) DNA pooling: a tool for large-scale association studies. *Nat. Rev. Genet.*, **3**, 862–871.
24. Collins, H.E., Li, H., Inda, S.E., Anderson, J., Laiho, K., Tuomilehto, J. and Seldin, M.F. (2000) A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. *Hum. Genet.*, **106**, 218–226.
25. Pritchard, J.K. and Rosenberg, N.A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, **65**, 220–228.
26. Risch, N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
27. Shaw, S.H., Carrasquillo, M.M., Kashuk, C., Puffenberger, E.G. and Chakravarti, A. (1998) Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.*, **8**, 111–123.
28. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.
29. Wakitani, S., Murata, N., Toda, Y., Ogawa, R., Kaneshige, T., Nishimura, Y. and Ochi, T. (1997) The relationship between HLA-DRB1 alleles and disease subsets of rheumatoid arthritis in Japanese. *Br. J. Rheumatol.*, **36**, 630–636.
30. Shibue, T., Tsuchiya, N., Komata, T., Matsushita, M., Shiota, M., Ohashi, J., Wakui, M., Matsuta, K. and Tokunaga, K. (2000) Tumor necrosis factor alpha 5′-flanking region, tumor necrosis factor receptor II, and HLA-DRB1 polymorphisms in Japanese patients with rheumatoid arthritis. *Arthritis Rheum.*, **43**, 753–757.
31. Yung, Yu.C., Yang, Z., Blanchong, C.A. and Miller, W. (2000) The human and mouse MHC class III region: a parade of 21 genes at the centromeric segment. *Immunol. Today*, **21**, 320–328.
32. Mao, J.R., Taylor, G., Dean, W.B., Wagner, D.R., Afzal, V., Lotz, J.C., Rubin, E.M. and Bristow, J. (2002) Tenascin-X deficiency mimics Ehlers–Danlos syndrome in mice through alteration of collagen deposition. *Nat. Genet.*, **30**, 421–425.
33. Gonzalez-Gay, M.A., Garcia-Porrua, C. and Hajeer, A.H. (2002) Influence of human leukocyte antigen-DRB1 on the susceptibility and severity of rheumatoid arthritis. *Semin. Arthritis Rheum.*, **31**, 355–360.
34. Zhang, Z. and Gerstein, M. (2003) Identification and characterization of over 100 mitochondrial ribosomal protein pseudogenes in the human genome. *Genomics*, **81**, 468–480.
35. Hesselink, M.K., Mensink, M. and Schrauwen, P. (2003) Human uncoupling protein-3 and obesity: an update. *Obes. Res.*, **11**, 1429–1443.
36. Smith, T.F, Gaitatzes, C., Saxena, K. and Neer, E.J. (1999) The WD repeat: a common architecture f or diverse functions. *Trends Biochem. Sci.*, **24**, 181–185.
37. Rezaie, T., Child, A., Hitchings, R., Brice, G., Miller, L., Coca-Prados, M., Heon, E., Krupin, T., Ritch, R., Kreutzer, D. *et al.* (2002) Adult-onset primary open-angle glaucoma caused by mutations in optineurin. *Science*, **295**, 1077–1079.
38. Suzuki, A., Yamada, R., Chang, X., Tokuhiro, S., Sawada, T., Suzuki, M., Nagasaki, M, Nakayama-Hamada, M., Kawaida, R., Ono, M. *et al.* (2003) Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.*, **34**, 395–402.
39. Tokuhiro, S., Yamada, R., Chang, X., Suzuki, A., Kochi, Y., Sawada, T., Suzuki, M., Nagasaki, M., Ohtsuki, M., Ono, M. *et al.* (2003) An intronic SNP in RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat. Genet.*, **35**, 341–348.
40. Begovich, A.B., Carlton, V.E., Honigberg, L.A., Schrodi, S.J., Chokkalingam, A.P., Alexander, H.C., Ardlie, K.G., Huang, Q., Smith, A.M., Spoerke, J.M. *et al.* (2004) A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.*, **75**, 330–337.
41. Murray, J.C., Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V.C., Sunden, S., Duyk, G.M. *et al.* (1994) A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science*, **265**, 2049–2054.
42. Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., *et al.* (1996) A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature*, **380**, 152–154.
43. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–247.
44. Ott, J. and Rabinowitz, D. (1997) The effect of marker heterozygosity on the power to detect linkage disequilibrium. *Genetics*, **147**, 927–930.
45. Chapman, N.H. and Wijsman, E.M. (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am. J. Hum. Genet.*, **63**, 1872–1885.
46. Sham, P.C., Zhao, J.H. and Curtis, D. (2000) The effect of marker characteristics on the power to detect linkage disequilibrium due to single or multiple ancestral mutations. *Ann. Hum. Genet.*, **64**, 161–169.
47. Gray, I.C., Campbell, D.A. and Spurr, N.K. (2000) Single nucleotide polymorphisms as tools in human genetics. *Hum. Mol. Genet.*, **9**, 2403–2408.
48. Li, S., Kawata, H., Katsuyama, Y., Ota, M., Morishima, Y., Mano, S., Kulski, J.K., Naruse, T. and Inoko, H. (2004) Association of polymorphic MHC microsatellites with GDVH, survival and leukemia relapse in translated hematopoietic stem cell transplant donor/recipient matched at 5 HLA loci. *Tissue Antigens*, **63**, 362–368.
49. Bergstrom, T.F., Josefsson, A., Erlich, H.A. and Gyllensten, U. (1998) Recent origin of HLA-DRB1 alleles and implications for human evolution. *Nat. Genet.*, **18**, 237–242.
50. Cullen, M., Perfetto, S.P., Klitz, W., Nelson, G. and Carrington, M. (2002) High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am. J. Hum.Genet.*, **71**, 759–776.
51. Jawaheer, D., Li, W., Graham, R.R., Chen, W., Damle, A., Xiao, X., Monteiro, J., Khalili, H., Lee, A., Lundsten, R., *et al.* (2002) Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. *Am. J. Hum. Genet.*, **71**, 585–594.
52. Moore, A.R. (2003) Collagen-induced arthritis. *Methods Mol. Biol.*, **225**, 175–179 (2003).
53. Li, T., Warris, V., Ma, J., Lassus, J., Yoshida, T., Santavirta, S., Virtanen, I. and Konttinen, Y.T. (2000) Distribution of tenascin-X in different synovial samples and synovial membrane-like interface tissue from aseptic loosening of total hip replacement. *Rheumatol. Int.*, **19**, 177–183.
54. Ando, K., Kanazawa, S., Tetsuka, T., Ohta, S., Jiang, X., Tada, T., Kobayashi, M., Matsui, N. and Okamoto, T. (2003) Induction of Notch signaling by tumor necrosis factor in rheumatoid synovial fibroblasts. *Oncogene*, **22**, 7796–7803.
55. Cardon, L.R. and Abecasis, G.R. (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet.*, **19**, 135–140.
56. Mateu, E., Perez-Lezaun, A., Martinez-Arias, R., Andres, A., Valles, M., Bertranpetit, J. and Calafell, F. (2002) PKLR—GBA region shows almost complete linkage disequilibrium over 70 kb in a set of worldwide populations. *Hum. Genet.* **110**, 532–544.
57. Gianfrancesco, F., Esposito, T., Ombra, M.N., Forabosco, P., Maninchedda, G., Fattorini, M., Casula, S., Vaccargiu, S., Casu, G., Cardia, F., *et al.* (2003) Identification of a novel gene and a common variant associated with uric acid nephrolithiasis in a Sardinian genetic isolate. *Am. J. Hum. Genet.*, **72**, 1479–1491.

58. Smith, C.A. and Arnett, F.C. Jr. (1991) Diagnosing rheumatoid arthritis: current criteria. *Am. Fam. Physician*, **44**, 863–870.
59. Akane, A., Shiono, H., Matsubara, K., Nakahori, Y., Seki, S., Nagafuchi, S., Yamada, M. and Nakagome, Y. (2001) Sex identification of forensic specimens by polymerase chain reaction (PCR): two alternative methods. *Forensic Sci. Int.*, **49**, 81–88.
60. Voorter, C.E., Rozemuller, E.H., de Bruyn-Geraets, D., van der Zwan, A.W., Tilanus, M.G. and van den Berg-Loonen, E.M. (1997) Comparison of DRB sequence-based typing using different strategies. *Tissue Antigens*, **49**, 471–476.
61. Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S. *et al.* (2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, **418**, 544–548.
62. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
63. Clark, A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7**, 111–122.
64. Mano, S., Yasuda, N., Katoh, T., Tounai, K., Inoko, H., Imanishi, T., Tamiya, G. and Gojobori, T. (2004) Notes on the maximum likelihood estimation of haplotype frequencies. *Ann. Hum. Genet.*, **68**, 257–264.
65. Hashimoto, A., Endo, H., Hayashi, I., Murakami, Y., Kitasato, H., Kono, S., Matsui, T., Tanaka, S., Nishimura, A., Urabe, K. *et al.* (2003) Differential expression of leukotriene B4 receptor subtypes (BLT1 and BLT2) in human synovial tissues and synovial fluid leukocytes of patients with rheumatoid arthritis. *J. Rheumatol.*, **30**, 1712–1718.