our selection criteria: statistical significance ($P<0.1$) and the same direction of allelic effect on AN susceptibility between the first and the second screening results. In the third screening, among the 158 markers tested, 16 satisfied our selection criteria: statistical significance ($P<0.1$) and the same direction of allelic effect throughout the MS screening stages. We determined a significance threshold to control false-positive rates (nominal $\alpha=0.05$) in the first stage of MS screening. In the second and third stages, considering that our sample sizes of cases and controls are not large, we set the significance thresholds (nominal $\alpha=0.1$) to maintain the overall statistical power of the screening.

To determine the definite allele frequencies of the selected 16 MS markers, we performed individual genotyping on all the AN cases ($n=320$) and controls ($n=341$) used in the first to third screenings. Of the 16 markers, 10 showed a statistically significant difference by Fisher's exact test in the comparison between controls and the AN cohort (Table 1). After correction of multiple tests with the number of alleles, 7 of 10 markers remained statistically significant ($Pc<0.05$).

SNP association analysis to narrow down the regions responsible for AN susceptibility

From the 10 MS markers that were found to be associated with AN, we selected 7 (shown in bold in Table 1), on the basis of the gene content around each marker, as targets for SNP association analysis to narrow down disease susceptibility loci. We primarily selected a collection of evenly spaced SNPs (11.1-kb interval on average) within a several 100-kb region surrounding each candidate MS marker, although it should be noted that intragenic SNPs were preferentially selected from the loci of 1q41 (spermatogenesis associated 17 (*SPATA17*)), 5q15 (*CDH18*) and 18q22 (*NETO1*). The number of SNPs subjected to association analysis and the size and nucleotide positions of the corresponding genomic interval are listed for each locus in Table 2.

In total, we performed genotyping for 333 SNPs on 331 AN cases and 872 controls. Among the 251 SNPs that satisfied thresholds for the Hardy–Weinberg equilibrium (exact test $P>0.01$) and minor allele frequency (>5%), 24 showed a statistically significant association (nominal $P<0.05$) with the AN cohort (Table 2). For each of the seven loci analyzed, nominal *P*-values were corrected with 10 000 iterated permutations using Haploview 4.0. In all, 3 SNPs, all of which are located on 1q41, out of the 24 SNPs remained statistically significant ($Pc<0.05$) (underlined in Table 2).

Subsequently, using Haploview 4.0, we inferred LD block structures for each candidate chromosomal region, and performed a haplotype association analysis (100 000 iterated permutations) for the constructed LD blocks. Significant association ($Pc<0.05$) with the AN cohort was detected in three of the six SNP haplotype blocks defined in the 1q41 locus, and in one of the eight blocks defined in the 11q22 locus (Figure 1 and Table 3).

1q41

A total of 38 SNPs were selected for genotyping within a 337.3-kb interval, including the AN-associated MS marker D1S0562i. Among 30 SNPs subjected to association analysis, 7 showed a statistically significant association ($P<0.05$) with the AN cohort (Table 2). All the seven SNPs were located at 3′-downstream of the *SPATA17* gene (Figure 1, left). SNP rs2048332 showed the most significant association (allelic $P=0.00023$) and was further analyzed under different genetic models. Association analysis under a recessive model for rs2048332 showed the lowest *P*-value of 0.00015 with the CC genotype, indicating that the CC genotype of rs2048332 has a susceptible effect on the AN phenotype in the Japanese (odds ratio=1.73, confidence interval, 1.30–2.31). Among the three AN-associated haplotype blocks (1q41-#4, #5 and #6 in Figure 1, left and Table 3), 1q41-#5 that comprised two SNPs, namely rs1397178 and rs2048332, spanning a 10.2-kb interval, was found to be most significantly associated ($Pc=0.0039$).

The AN-associated MS marker D1S0562i was located in block 1q41-#6, which comprised five SNPs spanning a 38.1-kb interval, and was also associated with AN ($Pc=0.038$). Four of the five SNPs binned to haplotype block 1q41-#6, rs17691163, rs34418611, rs1934216 and rs1538555, were in a relatively strong pairwise LD ($D'=0.72$–0.75) with D1S0562i, whereas the most significant SNP, rs2048332, in 1q41-#5 block was in modest LD ($D'=0.46$) with it. The four SNPs and D1S0562i (rs17691163–D1S0562i–rs34418611–rs1934216–rs1538555) were selected as tags captured through LD in block 1q41-#6. These haplotype tags were subjected to an MS-SNP haplotype-based association analysis: one haplotype (G-2-A-T-G), tagged by an AN-associated risk allele of D1S0562i (Supplementary Table 2), was significantly associated with AN ($Pc=0.0065$) (Table 4).

Table 1 Ten microsatellite markers showing statistically significant differences in the individual genotyping

				Allele frequencies		*2×2*		*2×m*		
MS Marker	*Cytoband*	*No. of alleles*	*Positive alleles*[a]	*Control (N=341)*	*AN (N=320)*	P	Pc	P	*Odds ratio*	*95% CI*
D1S0016i	**1p36**	12	2	13.0%	20.3%	0.00047	0.0056	0.014	1.70	1.26–2.28
D1S0562i	**1q41**	9	2	12.4%	18.1%	0.0043	0.039	0.054	1.57	1.16–2.12
D5S0853i	**5q15**	17	1	12.5%	7.7%	0.0047	0.080	0.29	0.59	0.41–0.85
D11S0389i	**11q13**	16	1	6.5%	10.8%	0.038	0.61	0.85	1.75	1.18–2.60
D11S0268i	**11q22**	16	2	17.8%	26.0%	0.00039	0.0062	0.0057	1.62	1.24–2.11
D12S0245i	12q14.1	3	2	50.0%	57.5%	0.0067	0.020	0.024	1.36	1.09–1.69
D12S0848i	12q23.2	11	1	14.6%	21.1%	0.0029	0.032	0.040	1.56	1.17–2.08
G09961	**16q12**	11	1	17.6%	11.1%	0.00075	0.008	0.0043	0.58	0.43–0.80
D18S0019i	**18q22**	13	3	36.8%	26.5%	0.000070	0.00091	0.0071	0.61	0.48–0.77
D19S0081i	19p13.3	13	1	14.2%	10.4%	0.044	0.57	0.017	0.70	0.50–0.98

Abbreviations: *P*, Fisher's exact test *P*-value; *Pc*, *P*-value corrected for the number of alleles.
P- and *Pc*-values smaller than 0.05 are underlined.
[a]Number of the alleles showing significant differences ($P<0.05$). When more than one positive allele were detected, the allele frequencies and the *P and Pc*-values of the positive allele showing the most significant difference are listed.

Table 2 SNP allelic association with AN

Cytoband, # of SNPs subjected to association analysis, interval size, and nucleotide positions (NCBI Build 36.1)	*SNP ID*	*Location (NCBI Build 36.1)*	*Gene symbol*	*SNP Type*	*Allele*	*Frequency*		*HWE*			*Odds*	
						Control (n=872)	*Case (n=331)*	P-value	P-value	Pc	*ratio*	*95% CI*
1p36	rs3753499	chr.1 7834024	*UTS2*	Intron	T	0.868	0.903	0.59	0.019	0.375	1.42	1.06~1.90
54 SNPs, 332.8 kb	rs106968	chr.1 7897131		Intergenic	G	0.397	0.447	0.017	0.027	0.480	1.23	1.02~1.47
chr1: 7695967–8028788												
1q41	rs11117962	chr.1 216116886		Intergenic	T	0.424	0.500	0.41	0.00088	0.106	1.36	1.13~1.62
30 SNPs, 337.3 kb	rs907057	chr.1 216125221		Intergenic	C	0.354	0.400	0.29	0.039	0.710	1.21	1.01~1.46
chr1: 215887318–216224643	rs7538903	chr.1 216135516		Intergenic	C	0.453	0.530	0.60	0.00074	0.022	1.36	1.14~1.63
	rs1397178	chr.1 216166200		Intergenic	T	0.283	0.326	0.16	0.039	0.643	1.23	1.01~1.49
	rs2048332	chr.1 216176389		Intergenic	C	0.446	0.530	0.61	0.00023	0.009	1.40	1.17~1.68
	rs12125437	chr.1 216186564		Intergenic	T	0.311	0.382	0.75	0.00092	0.062	1.37	1.14~1.65
	rs1538555	chr.1 216224643		Intergenic	G	0.311	0.388	0.71	0.00036	0.039	1.40	1.16~1.69
11q13	rs11227668	chr.11 66568914	*SYT12*	Intron	A	0.214	0.258	0.39	0.019	0.243	1.28	1.04~1.58
20 SNPs, 441.3 kb	rs3741190	chr.11 66573083	*SYT12*	UTR 3	T	0.207	0.248	0.82	0.029	0.330	1.26	1.02~1.56
chr11: 66365567–66806868												
11q22	rs12574821	chr.11 99333169	*CNTN5*	Intron	G	0.207	0.264	0.74	0.0027	0.100	1.38	1.12~1.69
56 SNPs, 699.8 kb	rs1349782	chr.11 99335431	*CNTN5*	Intron	G	0.395	0.443	0.49	0.034	0.678	1.22	1.01~1.46
chr11: 98931758–99631604	rs6590474	chr.11 99358383	*CNTN5*	Intron	A	0.164	0.214	0.68	0.004	0.137	1.39	1.11~1.75
16q12	rs2270355	chr.16 50958219		Intergenic	G	0.086	0.115	0.29	0.031	0.512	1.38	1.03~1.84
34 SNPs, 324.9 kb	rs1111482	chr.16 51021032		Intergenic	T	0.858	0.902	0.23	0.0040	0.094	1.52	1.14~2.04
chr16: 50819155–51144082	rs11647880	chr.16 51022612		Intergenic	A	0.858	0.902	0.49	0.0045	0.102	1.52	1.14~2.03
	rs8062936	chr.16 51033470	*TOX3*	Intron	A	0.863	0.903	0.30	0.0077	0.171	1.48	1.11~1.99
	rs2052287	chr.16 51059669	*TOX3*	Intron	G	0.755	0.801	0.73	0.019	0.347	1.30	1.04~1.62
	rs2287144	chr.16 51059819	*TOX3*	Intron	G	0.756	0.800	0.72	0.022	0.397	1.29	1.04~1.61
	rs3095620	chr.16 51068809	*TOX3*	Intron	T	0.880	0.913	0.016	0.020	0.376	1.44	1.06~1.96
18q22	rs2000728	chr.18 68647991	*NETO1*	Intron	G	0.554	0.599	0.04	0.047	0.667	1.20	1.00~1.44
27 SNPs, 1026.4 kb	rs1477490	chr.18 68910075		Intergenic	T	0.385	0.447	0.92	0.0061	0.137	1.29	1.07~1.55
chr18: 68441726–69468090	rs1477491	chr.18 68910515		Intergenic	A	0.383	0.447	0.84	0.0044	0.103	1.30	1.09~1.56

Abbreviations: *P*, chi-square test *P*-value; *Pc*, 10000 iterated permutation *P*-value.
Pc-values smaller than 0.05 are underlined.
For the 5q15 locus, 30 SNPs within a 521.8 kb interval (chr5: 19501188-20341733) were subjected to association analysis, and none of them showed a stastitically significant association ($P<0.05$) with AN.

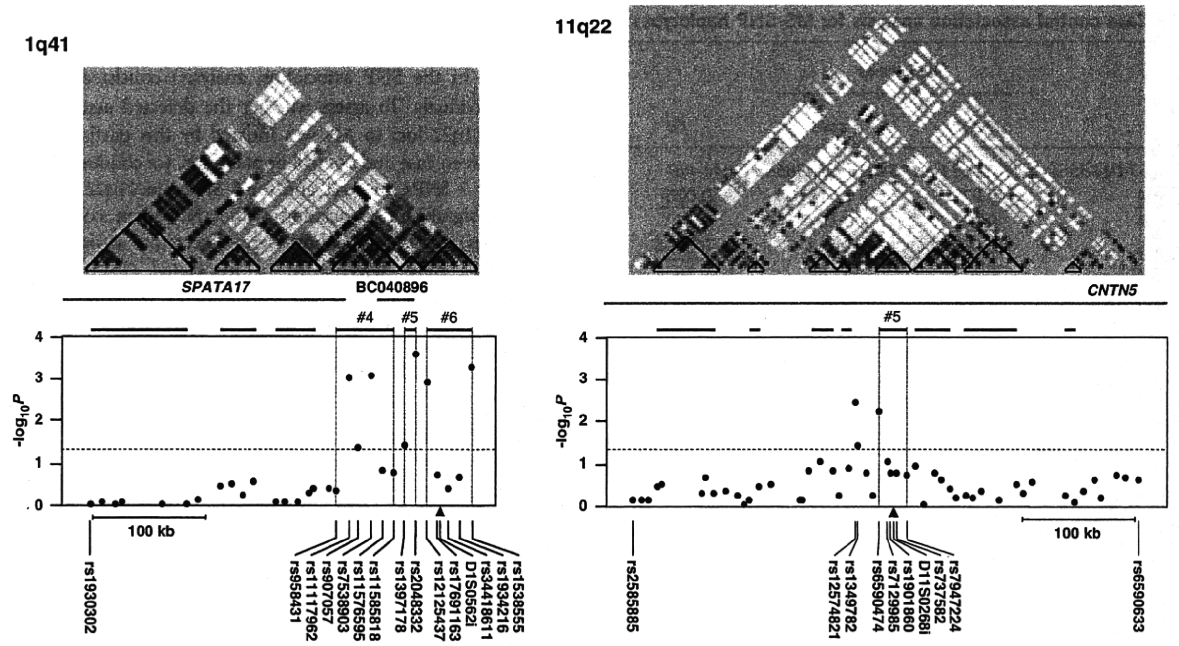

Figure 1 Single-nucleotide polymorphism (SNP) and haplotype association analyses for the 1q41 (left) and the 11q22 (right) loci. For each locus, the linkage disequilibrium (LD) plot (top), resided gene(s) (middle) and the *P*-value plot (bottom) are shown. In LD plots, the extent of LD between two SNPs is shown by the standard color scheme (*D*′/LOD) of Haploview 4.0. In *P*-value plots, closed dots show the minus log *P*-value (y axis) and the physical location (x axis) of SNPs. Minus log *P*-values were calculated by χ^2 tests for the genotyping data of anorexia nervosa (AN) cases (n=331) and controls (n=872). The horizontal dashed line corresponds to the *P*-value of 0.05. Black and red horizontal bars above the *P*-value plots correspond to the LD blocks defined by the confidence intervals method (Haploview 4.0). LD blocks showing statistical significance in the haplotype association analysis (*Pc*<0.05 in Table 3) are shown by red bars. The rs numbers of the SNPs showing statistical significance (*P*<0.05 in Table 1), the SNPs binned to the AN-associated LD blocks and the SNPs at the ends of the genomic interval are shown underneath the *P*-value plot. The positions of AN-associated MS markers (D1S0562i and D11S0268i) are shown by blue rectangles. The *SPATA17* gene and an uncharacterized mRNA sequence, BC040896, which are transcribed from left to right orientation, are mapped in the 337.3-kb interval between SNPs rs1930302 and rs1538555 on 1q41. For the 11q22 locus, the 474.6-kb region between SNPs rs2585885 and rs6590633, which are located in intron 2 and intron 16 of the *CNTN5* gene (NM_014361), respectively, is shown.

Table 3 LD blocks in 1q41 and 11q22 loci and haplotype association analysis

Locus & block#	*Size (kb)*	*Start*	*End*	*Gene*	*# of SNPs*	*# of haplotypes*	*# of positive haplotypes (P<0.05)*	*Frequency (controls)*	*Frequency (AN)*	P	*Pc (permutation P-value)*
1q41											
#1	85.1	rs1930302	rs11578064	*SPATA17*	6	3	0	0.676	0.676	0.99	1
#2	29.8	rs6604558	rs10495075	*SPATA17*	4	3	0	0.535	0.559	0.29	1
#3	32.9	rs11578620	rs1510262	*SPATA17*	5	4	0	0.586	0.570	0.48	1
#4	49.9	rs958431	rs11585818	*SPATA17/BC040896*	6	7	2	0.520	0.447	0.0014	0.031
#5	10.2	rs1397178	rs2048332	BC040896	2	3	2	0.548	0.461	0.00010	0.0039
#6	38.1	rs12125437	rs1538555		5	4	2	0.139	0.191	0.0019	0.038
11q22											
#1	48.0	rs2585885	rs1145408	*CNTN5*	5	5	0	0.487	0.508	0.37	1
#2	8.6	rs4754649	rs11221713	*CNTN5*	2	3	1	0.037	0.020	0.032	0.56
#3	22.7	rs3824932	rs10894179	*CNTN5*	3	3	0	0.654	0.623	0.16	0.99
#4	6.5	rs11221996	rs1530997	*CNTN5*	2	3	0	0.445	0.431	0.55	1
#5	20.2	rs6590474	rs7947224	*CNTN5*	5	3	2	0.730	0.655	0.00030	0.0078
#6	32.0	rs770569	rs4754665	*CNTN5*	5	5	0	0.482	0.447	0.13	1
#7	45.9	rs10750469	rs12806530	*CNTN5*	5	6	0	0.409	0.397	0.61	1
#8	7.3	rs7115626	rs6590633	*CNTN5*	2	3	0	0.798	0.793	0.81	1

Abbreviations: *P*, chi-square test *P*-value; *Pc*, 100 000 iterated permutation *P*-value.
P- and *Pc*-values smaller than 0.05 are underlined.
When no haplotype shows statistically significant association among multiple haplotypes inferred, the frequencies, *P* value, and *Pc* value of the major haplotype are shown.
When more than one haplotype show statistically significant association, the frequencies, *P* value, and *Pc* value of the most significantly associated haplotype are shown.

Table 4 Case-control association analysis for MS-SNP haplotypes

Haplotype	Frequencies		Pc
	control	case	
rs17691163-D1S0562i-rs34418611-rs1934216-rs1538555 (1q41-#6)			
G-6-A-T-A	0.279	0.218	0.025
G-7-A-T-A	0.141	0.127	0.51
G-2-A-T-G	0.098	0.153	0.0065
G-4-G-A-G	0.111	0.105	0.76
G-5-A-T-A	0.102	0.080	0.20
G-4-A-T-A	0.066	0.073	0.68
A-6-A-T-A	0.051	0.052	0.89
rs6590474-D11S0268i-rs737582-rs7947224 (11q22-#5)			
C-2-G-T	0.479	0.408	0.023
C-3-G-T	0.288	0.242	0.105
A-4-G-T	0.124	0.212	0.00003
C-5-A-C	0.050	0.062	0.377

Abbreviation: *Pc* (permutation *P*-value, *n*=100 000).
Haplotypes whose control frequency is >=0.05 are shown.
Pc-values smaller than 0.05 and the corresponding haplotypes are underlined.

This association was comparable with those observed in the SNP-haplotype analysis in two haplotype blocks 1q41-#5 (*Pc*=0.0039) and 1q41-#6 (*Pc*=0.038) in terms of statistical significance.

11q22

A total of 66 SNPs were selected from a 699.8-kb interval surrounding the AN-associated MS marker D11S0268i. Among the 56 SNPs subjected to association analysis, 3 (rs12574821, rs1349782 and rs6590474) showed a statistically significant association ($P<0.05$) with the AN cohort (Table 2). These associated SNPs were found to be located in the eighth intron of the *CNTN5* gene (GenBank accession no. NM_014361). Although these three SNPs did not hold statistical significance after multiple-testing correction by permutation tests, a haplotype composed of five SNPs (rs6590474, rs7129985, rs1901860, rs737582 and rs7947224) spanning a 20.2-kb interval showed a statistically significant association with AN (*Pc*=0.0082) in the haplotype association analysis (11q22-#5 in Figure 1, right and Table 3). Exon 9 of *CNTN5* was included in the 20.2-kb interval.

The AN-associated MS marker, D11S0268i, was located in the AN-associated 11q22-#5 block. In this block, D11S0268i was in a strong pairwise LD (D'=0.81–0.90) with each of the five SNPs binned to this block. As three SNPs (rs6590474, rs737582 and rs7947224) and D11S0268i were selected as tags captured through LD in the 11q22-#5 block, we further conducted an MS-SNP haplotype-based association study within the block using these four markers (rs6590474, D11S0268i, rs737582 and rs7947224). As shown in Table 4, the A-4-G-T haplotype was overrepresented in AN cases with the greatest statistical significance (*Pc*=0.00003). This MS-SNP haplotype contained both of the significantly associated risk alleles (A allele and 4) at SNP rs6590474 (Table 2) and D11S0268i (Supplementary Table 3).

Assessment of possible gender effects in the detected association of 1q41 and 11q22 with AN

Owing to the limited number of female individuals whose age matches with the average age of the AN cases in our control samples, we adopted a population-based control group to search for AN susceptibility loci. Therefore, although 180 female controls (control group 1, average age: 34.5 years) were genotyped in the first and second stages of MS screening, an additional 692 control individuals enrolled in the later stages (161 individuals in the third stage of MS screening and all of 692 individuals in the SNP association analysis) consisted of male and female individuals. To assess whether the detected association of the 1q41 and 11q22 loci to AN was inflated by this partial mismatch in gender between case and control populations, we conducted a stratified analysis for 7 SNPs on 1q41 and for 3 SNPs on 11q22 (Table 2) that showed a significant association with AN. When control group 1 (180 females) and the AN cohort were subjected to association analysis, all 10 SNPs were detected to be associated ($P<0.05$) with AN (Supplementary Table 4). These results assure that the association of the 1q41 and 11q22 loci with AN detected in this study is not because of inflation caused by the inclusion of male individuals in the control population.

Association analysis for a BN cohort

To assess whether the genomic intervals identified to be associated with AN in this study are also involved in the genetic etiology of BN, we conducted an SNP association analysis for BN cases and controls. The 7 SNPs from the 1q41 locus and the 3 SNPs from the 11q22 locus, which showed a statistically significant association ($P<0.05$) with AN before multiple-testing correction, were subjected to SNP genotyping on the cohort of 125 BN cases. None of the 10 SNPs showed a statistically significant association with BN (data not shown).

DISCUSSION

We have completed a genome-wide association analysis for AN using 23 465 MS markers. To our knowledge, this is the first GWAS performed for EDs. Among the 10 candidate loci we identified, 9 are reported to be associated with AN for the first time in this study. Only one locus, D1S0016i on 1p36, overlaps with the chromosomal region of 1p33-p36 that has already been reported to show significant linkage to AN.[12]

Through an SNP association analysis for the seven selected candidate regions to narrow down genomic intervals involved with AN susceptibility, we tentatively identified a 10.2-kb genomic interval (the haplotype block 1q41-#5) located at 3′-downstream of the *SPATA17* gene as a region associated with AN. The MS-SNP haplotype-based association analysis also indicated the association of haplotype block 1q41-#6 with AN with a similar statistical significance. *SPATA17* encodes a 361 amino-acid protein that contains three highly conserved IQ motifs and is strongly expressed in the testis.[30] It is unknown whether the SPATA17 protein has any physiological roles in neuronal tissues. It should be noted that the 10.2-kb critical interval coincides with the exon–intron structure of the uncharacterized mRNA sequence BC040896, which is derived from a cDNA library made from brain (adult medulla) RNA.

Another genomic region identified to be associated with AN in this study is a 20.2-kb interval (the haplotype block 11q22-#5), spanning from the eighth to the ninth intron of the *CNTN5* gene on 11q22. Furthermore, we found that one MS-SNP haplotype (A-4-G-T), which includes two AN-associated risk alleles at SNP rs6590474 and MS D11S0268i, was significantly overrepresented in AN cases. *CNTN5* encodes a member of the contactin family known to function during the formation of neuronal interactions. It is reported that the mouse line deficient of *Cntn1*, another member of the contactin family, exhibits an ataxic and anorectic phenotype.[31,32] In human adult tissues examined using northern blot analysis, *CNTN5* has been shown to be predominantly expressed in the brain and thyroid.[33] In various regions of an adult brain examined, the gene was found to be expressed with highest levels in the occipital lobe and amygdala,

followed by the cerebral cortex, frontal lobe, thalamus and temporal lobe.[33] Although neuronal activity in the auditory system is reported to be impaired in the mouse line deficient for *Cntn5*, no anorexic phenotype has been described.[34]

Although causative SNPs are not yet determined, we have successfully mapped genetic association with AN to at least two genomic regions on 1q41 and 11q22 and narrowed down an AN-associated genomic interval for each locus by haplotype association analysis. Further replication analysis using independent patient/control populations for AN-associated SNPs and functional analyses for the genes or for particular genomic regions in these loci will better clarify the impact of these SNPs/genes in the genetic etiology of AN. It should also be noted that additional common variants are likely to have roles in the development of AN because this study was not well powered to detect susceptible loci with relatively small genetic risks. Additional gender/age-matched cohorts consisting of much larger numbers of cases and controls need to be used to improve statistical power in MS-based genome-wide association analysis.

ACKNOWLEDGEMENTS

This study was performed under the management of the Japan Biological Informatics Consortium (JBIC) and was supported by grants from the New Energy and Industrial Technology Development Organization (NEDO). This study was also supported by Grant-in-Aid for Scientific Research on Priority Areas from MEXT, and by Grant-in-Aid for Young Scientists (B) from JSPS. We thank Karin Ohki for her technical assistance.

1 Tozzi, F., Thornton, L. M., Klump, K. L., Fichter, M. M., Halmi, K. A., Kaplan, A. S. *et al.* Symptom fluctuation in eating disorders: correlates of diagnostic crossover. *Am. J. Psychiatry* **162,** 732–740 (2005).
2 Nishimura, H., Komaki, G., Ando, T., Nakahara, T., Oka, T., Kawai, K. *et al.* Japanese Genetic Research Group for Eating Disorders. Psychological and weight-related characteristics of patients with anorexia nervosa-restricting type who later develop bulimia nervosa. *Biopsychosoc. Med.* **2,** 5 (2008).
3 Fairburn, C. G., Harrison, P. J. Eating disorders. *Lancet* **361,** 407–416 (2003).
4 Birmingham, C. L., Su, J., Hlynsky, J. A., Goldner, E. M., Gao, M. The mortality rate from anorexia nervosa. *Int. J. Eat. Disord.* **38,** 143–146 (2005).
5 Keel, P. K., Mitchell, J. E. Outcome in bulimia nervosa. *Am. J. Psychiatry* **154,** 313–321 (1997).
6 Lilenfeld, L. R., Kaye, W. H., Greeno, C. G., Merikangas, K. R., Plotnicov, K., Pollice, C. *et al.* A controlled family study of anorexia nervosa and bulimia nervosa: psychiatric disorders in first-degree relatives and effects of proband comorbidity. *Arch. Gen. Psychiatry* **55,** 603–610 (1998).
7 Strober, M., Freeman, R., Lampert, C., Diamond, J., Kaye, W. Controlled family study of anorexia nervosa and bulimia nervosa: evidence of shared liability and transmission of partial syndromes. *Am. J. Psychiatry* **157,** 393–401 (2000).
8 Bulik, C. M., Sullivan, P. F., Wade, T. D., Kendler, K. S. Twin studies of eating disorders: a review. *Int. J. Eat. Disord.* **27,** 1–20 (2000).
9 Wade, T. D., Bulik, C. M., Neale, M., Kendler, K. S. Anorexia nervosa and major depression: shared genetic and environmental risk factors. *Am. J. Psychiatry* **157,** 469–471 (2000).
10 Bulik, C. M., Slof-Op't Landt, M. C., van Furth, E. F., Sullivan, P. F. The genetics of anorexia nervosa. *Annu. Rev. Nutr.* **27,** 263–275 (2007).
11 Pinheiro, A. P., Sullivan, P. F., Bacaltchuck, J., Prado-Lima, P. A., Bulik, C. M. Genetics in eating disorders: extending the boundaries of research. *Rev. Bras. Psiquiatr.* **28,** 218–225 (2006).
12 Grice, D. E., Halmi, K. A., Fichter, M. M., Strober, M., Woodside, D. B., Treasure, J. T. *et al.* Evidence for a susceptibility gene for anorexia nervosa on chromosome 1. *Am. J. Hum. Genet.* **70,** 787–792 (2002).
13 Devlin, B., Bacanu, S. A., Klump, K. L., Bulik, C. M., Fichter, M. M., Halmi, K. A. *et al.* Linkage analysis of anorexia nervosa incorporating behavioral covariates. *Hum. Mol. Genet.* **11,** 689–696 (2002).
14 Bacanu, S. A., Bulik, C. M., Klump, K. L., Fichter, M. M., Halmi, K. A., Keel, P. *et al.* Linkage analysis of anorexia and bulimia nervosa cohorts using selected behavioral phenotypes as quantitative traits or covariates. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **139,** 61–68 (2005).
15 Bergen, A. W., van den Bree, M. B., Yeager, M., Welch, R., Ganjei, J. K., Haque, K. *et al.* Candidate genes for anorexia nervosa in the 1p33-36 linkage region: serotonin 1D and delta opioid receptor loci exhibit significant association to anorexia nervosa. *Mol. Psychiatry* **8,** 397–406 (2003).
16 Bulik, C. M., Devlin, B., Bacanu, S. A., Thornton, L., Klump, K. L., Fichter, M. M. *et al.* Significant linkage on chromosome 10p in families with bulimia nervosa. *Am. J. Hum. Genet.* **72,** 200–207 (2003).
17 Klump, K. L., Gobrogge, K. L. A review and primer of molecular genetic studies of anorexia nervosa. *Int. J. Eat. Disord.* **37,** S43–S48 (2005).
18 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
19 International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
20 Oka, A., Tamiya, G., Tomizawa, M., Ota, M., Katsuyama, Y., Makino, S. *et al.* Association analysis using refined microsatellite markers localizes a susceptibility locus for psoriasis vulgaris within a 111 kb segment telomeric to the HLA-C gene. *Hum. Mol. Genet.* **8,** 2165–2170 (1999).
21 Oka, A., Hayashi, H., Tomizawa, M., Okamoto, K., Suyun, L., Hui, J. *et al.* 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 (2003).
22 Tamiya, G., Shinya, M., Imanishi, T., Ikuta, T., Makino, S., Okamoto, K. *et al.* Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. *Hum. Mol. Genet.* **14,** 2305–2321 (2005).
23 Kimura, T., Kobayashi, T., Munkhbat, B., Oyungerel, G., Bilegtsaikhan, T., Anar, D. *et al.* Genome-wide association analysis with selective genotyping identifies candidate loci for adult height at 8q21.13 and 15q22.33-q23 in Mongolians. *Hum. Genet.* **123,** 655–660 (2008).
24 Yatsu, K., Mizuki, N., Hirawa, N., Oka, A., Itoh, N., Yamane, T. *et al.* High-resolution mapping for essential hypertension using microsatellite markers. *Hypertension* **49,** 446–452 (2007).
25 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders,* 4th edn. (American Psychiatric Association: Washington, DC, 1994).
26 Collins, H. E., Li, H., Inda, S. E., Anderson, J., Laiho, K., Tuomilehto, J. *et al.* A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. *Hum. Genet.* **106,** 218–226 (2000).
27 Purcell, S., Cherny, S. S., Sham, P. C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19,** 149–150 (2003).
28 Barrett, J. C., Fry, B., Maller, J., Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).
29 Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).
30 Deng, Y., Hu, LS., Lu, GX. Expression and identification of a novel apoptosis gene Spata17 (MSRG-11) in mouse spermatogenic cells. *Acta. Biochim. Biophys. Sin. (Shanghai)* **38,** 37–45 (2006).
31 Johansen, J. E., Fetissov, S. O., Bergström, U., Nilsson, I., Faÿ, C., Ranscht, B. *et al.* Evidence for hypothalamic dysregulation in mouse models of anorexia as well as in humans. *Physiol. Behav.* **92,** 278–282 (2007).
32 Fetissov, S. O., Bergström, U., Johansen, J. E., Hökfelt, T., Schalling, M., Ranscht, B. Alterations of arcuate nucleus neuropeptidergic development in contactin-deficient mice: comparison with anorexia and food-deprived mice. *Eur. J. Neurosci.* **22,** 3217–3228 (2005).
33 Kamei, Y., Takeda, Y., Teramoto, K., Tsutsumi, O., Taketani, Y., Watanabe, K. Human NB-2 of the contactin subgroup molecules: chromosomal localization of the gene (CNTN5) and distinct expression pattern from other subgroup members. *Genomics* **69,** 113–119 (2000).
34 Li, H., Takeda, Y., Niki, H., Ogawa, J., Kobayashi, S., Kai, N. *et al.* Aberrant responses to acoustic stimuli in mice deficient for neural recognition molecule NB-2. *Eur. J. Neurosci.* **17,** 929–936 (2003).

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

Transgenic Res
DOI 10.1007/s11248-011-9505-y

TECHNICAL UPDATE

Targeted transgenesis through pronuclear injection of improved vectors into in vitro fertilized eggs

Masato Ohtsuka · Hiromi Miura · Hirofumi Nakaoka · Minoru Kimura · Masahiro Sato · Hidetoshi Inoko

Received: 16 November 2010 / Accepted: 13 March 2011

We recently described a new transgenic technique termed pronuclear injection-based targeted transgenesis (PITT) (Ohtsuka et al. 2010), which enables targeted integration of transgenes into predetermined loci (e.g. *Rosa26*) by Cre-*loxP*-based recombination in fertilized eggs, leading to stable and reproducible transgene expression.

In this study, two steps of PITT were modified. The donor vector was improved by introducing an FLPe expression cassette (Fig. 1a). Therefore, self-removal of the extra sequence flanked by FRT sequences occurs without additional FLPe administration. In addition, the plasmid backbone of the donor vector was modified from pUC119-to pBR322-based to reduce plasmid instability, which is sometimes associated with genomic DNA fragment cloning into a high-copy plasmid. Moreover, in vitro fertilized eggs were used in this study; in the previous study, fertilized eggs obtained by natural mating were used. In this study, fertilized eggs (>100) were obtained simultaneously using epididymal spermatozoa isolated from a single male carrying the targeted allele.

Results are listed in Fig. 1b. Three independent donor vectors, pAWK, pAWV and pAXV, were microinjected with a Cre expression vector into in vitro fertilized eggs. The resulting PITT efficiency was 5/93 (5.4%; 95% CI, 1.8–12.1%), which is similar to that reported previously (4.3%; 95% CI, 2.7–6.5%), indicating the feasibility of this modified method and reproducibility of PITT. Regarding removal of the extra sequence, the recombinase-mediated cassette exchange $(RMCE)^{\Delta ex}$ allele was successfully created in 60% of founder mice (3/5; indicated by *). Although the other two founder mice did not possess the $RMCE^{\Delta ex}$ allele, the extra sequences of these founders can be successfully removed by crossing them with an FLPe deleter mouse (Ohtsuka et al. 2010). Notably, all tested lines (5/5) exhibited successful germline transmission

Electronic supplementary material The online version of this article (doi:10.1007/s11248-011-9505-y) contains supplementary material, which is available to authorized users.

M. Ohtsuka (✉) · H. Miura · M. Kimura · H. Inoko
Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan
e-mail: masato@is.icc.u-tokai.ac.jp

H. Nakaoka
Division of Human Genetics, Department of Integrated Genetics, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan

M. Sato
Section of Gene Expression Regulation, Frontier Science Research Center, Kagoshima University, 1-21-20 Korimoto, Kagoshima, Kagoshima 890-0065, Japan

Published online: 25 March 2011

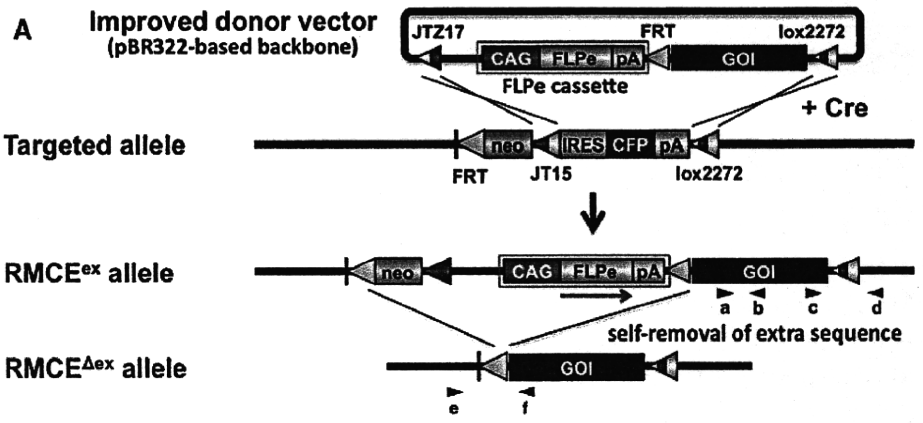

Fig. 1 **a** PITT strategy using the improved vector and **b** PITT efficiency. Also see supplementary material

B

Efficiency of PITT using new donor vectors and *in vitro* fertilized eggs

Plasmids injected	Eggs injected	Embryos transferred	Pups obtained[a]	Random integration	PITT[b] (%; b/a*100 [95%CI])	*RMCE$^{\Delta ex}$ allele
pAWK	304	170	9	0	1 (11.1 [0.3-48.2])	0
pAWV	290	172	30	1	3 (10.0 [2.1-26.5])	2
pAXV	305	186	54	N.D.	1 (1.9 [0.0-9.9])	1
total	899	528	93	-	5 (5.4 [1.8-12.1])	3

(data not shown). Detailed methods and comments are provided in the supplementary material.

Acknowledgments Funded by the Grant-in-Aid for Young Scientists (B) (20700368) to M.O. from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

Reference

Ohtsuka M, Ogiwara S, Miura H, Mizutani A, Warita T, Sato M, Imai K, Hozumi K, Sato T, Tanaka M, Kimura M, Inoko H (2010) Pronuclear injection-based mouse targeted transgenesis for reproducible and highly efficient transgene expression. Nucleic Acids Res 38:e198

Journal of Human Genetics (2010) 55, 656–661

www.nature.com/jhg

ORIGINAL ARTICLE

Genome-wide association study to identify genetic variants present in Japanese patients harboring intracranial aneurysms

Koichi Akiyama[1], Akira Narita[1], Hirofumi Nakaoka[1], Tailin Cui[1], Tomoko Takahashi[1], Katsuhito Yasuno[1], Atsushi Tajima[1], Boris Krischek[2], Ken Yamamoto[3], Hidetoshi Kasuya[4], Akira Hata[5] and Ituro Inoue[1]

An intracranial aneurysm (IA), which results in a subarachnoid hemorrhage with a high mortality on rupture, is a major public health concern. To identify genetic susceptibility loci for IA, we carried out a multistage association study using genome-wide single nucleotide polymorphisms (SNPs) in Japanese case–control subjects. In this study, we assessed evidence for association in standard approaches, and additional tests with adjusting sex effects that act between genetic effect and disease. Consequently, five SNPs ($P=1.31\times10^{-5}$ for rs1930095 of intergenic region; $P=1.32\times10^{-5}$ for rs4628172 of *TMEM195*; $P=2.78\times10^{-5}$ for rs7781293 of *TMEM195*; $P=4.93\times10^{-5}$ for rs7550260 of *ARHGEF11*; and $P=3.63\times10^{-5}$ for rs9864101 of *IQSEC1*) with probabilities of being false positives <0.5 were associated with IA in Japanese population, and the susceptibility genes could have a role in actin remodeling in the *ELN/LIMK* pathway. This study indicates the presence of several susceptibility loci that deserve further investigation in the Japanese population.
Journal of Human Genetics (2010) **55,** 656–661; doi:10.1038/jhg.2010.82; published online 8 July 2010

Keywords: cerebrovascular disease; genetics; genome-wide association study; intracranial aneurysm; sex effect; single nucleotide polymorphism; subarachnoid hemorrhage

INTRODUCTION

Intracranial aneurysm (IA) (MIM105800) is a cerebrovascular disease and has a reported prevalence of 3–6%. As the rupture of an IA leads to subarachnoid hemorrhage (SAH), which often results in death or severe physical disability, the disease is considered to be a major public health concern.[1–3] Much of the etiology of IA still remains unknown, but subjects with a family history of IA have a higher risk of being affected by IA, suggesting that some genetic components contribute to predisposition to developing an IA. Previously, we conducted a genome-wide linkage study in affected Japanese sib pairs[4] and showed through subsequent association studies that the single nucleotide polymorphisms (SNPs) of *ELN* and *LIMK1* on chromosome 7 was significantly associated with IA.[5] We also discovered one SNP of *LOXL2* on chromosome 8 associated with IA, and a possible gene–gene interaction of *LOXL2* with *ELN/LIMK1*.[6] Other loci of the linkage analyses in the Japanese population include 14q23, replicated by an association study.[7] In other ethnic groups, the perlecan gene (HSPG2) at 1p36.1–36.4,[8] the versican gene (CSPG2) at 5q14.3[9] and Kallikrein at 19q13[10] have been proposed as susceptibility genes for IA by case–control association study, among others.

Genome-wide association studies (GWASs) on complex diseases using high-density SNP arrays have recently been exemplified by a number of groups.[11,12] In a previous study, a multicenter collaboration regarding IA genetics led to complete a multistage GWAS that identified common variants that contribute to IA formation in three large cohorts: a Finnish cohort of 920 cases and 985 controls, a Dutch cohort of 781 cases and 6424 controls and a Japanese cohort of 495 cases and 676 controls.[11] Consequently, common SNPs with odds ratios (ORs) of 1.22–1.36 on chromosomes 2q, 8q and 9p were identified and showed significant association with IA. Because the three loci account only for a small fraction of the genetic variance underlying IA, it remains a challenge to identify another genetic variant such as common variants with population-specific impacts on the risk of IA.

Intracranial aneurysm is most likely to have a multifactorial etiology, involving complex interactions of genetic and environmental risk factors. Smoking habits, hypertension, frequent alcohol intake and female gender are known risk factors for aneurysm formation and subsequent SAH.[13–16] There is a possibility that these nongenetic risk factors might hinder the identification of genetic effects. Therefore, in the case of evaluating genuine genetic effects of SNPs contributing to

[1]Department of Molecular Life Sciences, Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Kanagawa, Japan; [2]Department of Neurosurgery, University of Tübingen, Tübingen, Germany; [3]Division of Molecular Population Genetics, Department of Molecular Genetics, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan; [4]Department of Neurosurgery, Neurological Institute, Tokyo Women's Medical University, Tokyo, Japan and [5]Department of Public Health, Graduate School of Medicine, Chiba University, Chiba, Japan
Correspondence: Dr I Inoue, Department of Molecular Life Sciences, Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan.
E-mail: ituro@is.icc.u-tokai.ac.jp
Received 23 February 2010; revised 30 May 2010; accepted 5 June 2010; published online 8 July 2010

IA, we first dealt with the interaction of SNPs with the invariable risk factor, female gender, at a genome-wide level, rather than those with other modifiable lifestyle factors such as smoking habits.

To gain a more comprehensive understanding of IA pathogenesis, we carried out a multistage association study using genome-wide SNPs in Japanese case–control data sets. For this purpose, we conducted a genome-wide assessment of SNP–gender interactions by statistically examining whether gender could affect an association of each SNP with IA. After controlling for the gender-effect, we identified five promising candidate loci associated with IA, by classifying and applying the best-suited tests for SNPs.

MATERIALS AND METHODS

Study design

We first performed a two-stage GWAS in Japanese IA case–control cohorts. To perform the most powerful and cost-effective study, the top-ranked 1% of the genotyped SNPs of each GWA stage were used for the following stage (1%×1%=0.01% of tagged autosomal SNPs). The statistical power, $1-\beta_{GWAS}$, is estimated to be more than 85% for the two-stage association study, assuming that α_{GWAS} is 2×10^{-7}, the prevalence of IA is 5% in the Japanese population, genotypic relative risk is 1.6 and allele frequency of the polymorphism is 0.45, calculated by the CaTS Power Calculator for two-stage association studies, available on the web site of Center of Statistical Genetics, University of Michigan (Supplementary Figure 1).[17] Additional statistical tests to examine possible confounding between SNP and gender, and effect modification of the genetic effect by gender were performed (the details are shown below). For the analysis of SNPs modified by gender, we also calculated statistical powers specific for male and female cohorts in the GWAS. These powers ($1-\beta_{male}$ and $1-\beta_{female}$) are estimated to be more than 18 and 22%, respectively, under the same conditions in which $1-\beta_{GWAS}$ was estimated.

In the final intensified study, we estimated false-positive report probabilities (FPRPs) of associations of the SNPs selected carefully from the GWAS stage by using all available case–control samples including GWAS cohorts to examine which SNPs would be noteworthy to be reported.

Subjects

The Ethics Committees of Tokai University, Tokyo Women's Medical University and Chiba University approved the study protocols, and all the participants gave written informed consent. We recruited a total of 1069 IA patients and 904 controls from the three university hospitals described above and their affiliated hospitals. The 495 cases (48.9%) and 676 controls (74.8%) have been also used in the previous GWAS by Bilguvar *et al.*[11] Of the cases, 27.7% had familial history of IA, and SAH was observed in 727 cases (68.0%). The female to male ratio of cases and controls were 1.56 and 0.62, respectively. The presence of IA was confirmed by conventional angiography, three-dimensional computed tomography angiography, magnetic resonance angiography or surgical findings, when applicative. Controls were recruited at the time of periodic health examination or from outpatients with conditions other than IAs and/or SAHs, such as idiopathic headache. About 92% of them were examined by magnetic resonance excluding the presence of IA. All the remaining control subjects had no medical and familial history of SAH. Adding them to the control group did not influence genetic frequency in this study (data not shown). The information for lifestyles and hypertension was obtained by a standard questionnaire, and each question was asked and confirmed by a skilled research coordinator and medical doctor (summarized in Table 1). Genomic DNA was extracted from peripheral blood or saliva according to the standard protocol.

Exclusion criteria for subjects

In the first stage of the GWAS, 499 subjects (300 patients/199 controls) were genotyped, 920 (460/460) were added to the second stage and the remaining 534 (310/224) were additionally genotyped in the intensified study. We excluded subjects in whom the genotyping call rate was <0.95 and 0.7 in first and second stages of GWAS, respectively.

In the first stage of GWAS, we also excluded subjects who genetically deviated from most other subjects on the basis of the result of multidimensional

Table 1 The detailed numbers of the subjects used in the study

	Non-IA	*IA*	*Total*
Starting cohort size			
Total	904	1069	1973
Male	551	408	959
Female	353	661	1014
Partitioning of samples			
GWAS (first stage)			
Total	199	300	499
Male	115	112	227
Female	84	188	272
GWAS (newly added in second stage)			
Total	460	460	920
Male	280	177	457
Female	180	283	463
GWAS (second stage)			
Total	659	760	1419
Male	395	289	684
Female	264	471	735
Intensive study			
Total	882	1069	1951
Male	537	408	945
Female	345	661	1006
Final cohort (quality controlled)			
GWAS (first stage)			
Total	194	288	482
Male	111	108	219
Female	83	180	263
GWAS (newly added in second stage)			
Total	455	452	907
Male	276	171	507
Female	179	281	399
GWAS (second stage)			
Total	649	740	1389
Male	387	279	726
Female	262	461	662
Intensified study			
Total	853	1027	1880
Male	515	391	906
Female	338	636	974
Mean age (±s.d.)	63.2 (±14.0)	55.1 (±11.3)	59.1 (±13.3)
SAH+ (%)	—	67.8	—
Hypertension[a,b]			
Yes	346	505	851
No	355	342	697
Smoking habit[a]			
Ever	316	431	747
Never	370	401	771
Alcohol drinking habit[a]			
Ever	326	304	630
Never	217	309	526

Abbreviations: GWAS, genome-wide association study; IA, intracranial aneurysm; SAH, subarachnoid hemorrhage.
[a]Differences between total and analyzed sample size were caused by incomplete clinical information.
[b]Hypertension was defined as systolic blood pressure >140 mm Hg and/or diastolic blood pressure >90 mm Hg.

scaling analysis, which was performed to identify genetic outliers. Furthermore, to detect cryptic relatedness (undiscovered kinship among subjects), probabilities of identity-by-state were estimated for each pair of subjects, and the data sets were reduced so that there was no substantial genetic kinship within four degrees among the remaining subjects. The calculation of identity-by-state probabilities and multidimensional scaling analysis were carried out using the software package PLINK.[18]

In the intensified study, we also excluded subjects with a genotyping call rate <0.8.

Genotyping and exclusion criteria for SNPs

In the first stage of the GWAS, genotyping was carried out using the HumanHap300 or HumanHap300-Duo Genotyping BeadChips (Illumina, San Diego, CA, USA), and a total of 312 712 SNPs were tested in both BeadChips. The SNPs were then filtered on the basis of genotyping call rate ($>95.0\%$ in either two types of BeadChips), minor allele frequency (>0.02) and deviation from Hardy–Weinberg equilibrium ($P<1.0\times10^{-5}$). The X-linked SNPs were also excluded.

In the second stage, the top-ranked 1% of all the SNPs analyzed in the first stage and the newly added 920 subjects were genotyped using the GoldenGate genotyping assay (Illumina). These 2304 SNPs were also quality controlled on the basis of genotyping call rate ($>90.0\%$), minor allele frequency (>0.01) and the Hardy–Weinberg equilibrium ($P<1.0\times10^{-4}$). To confirm the concordance of the genotype data obtained on different platforms, 20 of the subjects (10 patients/10 controls) included in the first stage were regenotyped as well.

In the intensified study, the remaining 532 subjects and all the other subjects studied in the previous two stages were genotyped for the 22 SNPs whose *P*-values ranked 1% of all those analyzed in the second stage using the TaqMan SNP Genotyping Assays (Applied Biosystems, Foster City, CA, USA). We confirmed the reliability of the genotype data by comparing the genotypes of the subjects studied in the previous two stages with those obtained by the HumanHap300/HumanHap300-Duo Genotyping BeadChips or the GoldenGate genotyping assay. The genotype concordance rate of the TaqMan data in the intensified study with the HumanHap data in the first GWAS stage was $99.4\% \pm 0.5$ (mean ± s.d.), and that with the GoldenGate data in the second GWAS stage was $98.9\% \pm 1.1$.

Statistical analysis

To test for association of each SNP with IA, we assumed an additive (in log-odds scale) model. The sex-specific effect modification can be characterized by the large difference of ORs between males and females. We estimated sex-specific OR by using a logistic regression model. Under the null hypothesis of no effect modification by sex, the statistic ψ, the logarithm of the ratio of these sex-specific ORs, follows asymptotically normal distribution with variance σ^2, the sum of the variances of sex-specific ORs.[19]

$$\psi = \log \mathrm{OR}_{\mathrm{male}}/\mathrm{OR}_{\mathrm{female}} \approx N(0, \sigma^2) \text{ (average } \sigma^2 = 0.06 \text{ at the GWAS stage)}$$

If the test *P*-value was <0.2, (corresponding to $\exp(\psi)=\mathrm{OR}_{\mathrm{male}}/\mathrm{OR}_{\mathrm{female}}>1.37$ or <0.73 under the above average σ^2), which might be large or small enough to consider that the SNP-effect differed between males and females, we stratified our data by sex, tested using each stratum, and defined the SNPs as 'class M'.

We supposed that gender was a confounding factor for association between the SNP and IA if the absolute value of the log-ratio of OR adjusted by gender to crude OR is >0.1 (roughly $>10\%$ difference of ORs), and defined as 'class C'. The adjusted and crude ORs were calculated using the logistic regression models with and without gender as a covariate, respectively.

For SNPs that were neither modified nor confounded by gender, defined as 'class N', the Cochran–Armitage (CA) trend test (one degree of freedom) and general genotypic test (two degrees of freedom) were performed. In contrast to the SNPs that were modified by gender, those tests were separately performed in male and female subjects (class M). For the SNPs that were confounded by gender, the Mantel extension test (one degree of freedom), and the generalized Cochran–Mantel–Haenszel test (two degrees of freedom) were performed (class C).

We also examined the validity of the assumption of additivity (in log-odds scale) in the association tests by comparing the logistic regression models with and without dominance effect.

To explore the likelihood that the results represent true associations, we computed the FPRP under different assumptions.[20] The previous probability of a true association was set between 0.01 and 0.0001, as it is regarded as an adequate level for genome-wide scan.[21] We considered that the SNP might be noteworthy when the FPRP was below 0.5 for suggestive association or 0.2 for significant association.[20,22]

RESULTS

Quality control

At the first stage of the GWAS, genome-wide genotyping in 499 Japanese subjects was performed, and we extracted 250 507 of 312 712 SNPs through the rigorous filtering described above. According to multidimensional scaling analysis based on identity-by-state distance (Supplementary Figure 2), 12 genetic outliers and 3 related individuals were excluded from the association analysis. Two subjects with low genotyping call rate were further excluded, and samples obtained from a total of 482 subjects were finally used in the first stage study. The detailed numbers of the subjects after filtering through the GWAS stage and the intensified study were shown in Table 1.

Genome-wide case–control association study

In the first stage of the GWAS, 250 507 filtered autosomal SNPs were analyzed in 482 subjects. In the following stage, 1833 filtered SNPs were analyzed in a cohort of 1389 subjects, which included 482 subjects studied in the first stage and 907 who were newly added. The summary of statistical analysis in the first stage is available on the genome-wide association database (https://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_study.cgi?id=cerebral).

We investigated genetic effects of the filtered 250 507 autosomal SNPs in the first stage, of which 50 740 (20.2%) were modified by gender (class M). We then applied the CA trend test and the 2 d.f. genetic test to each of the sex-stratified data sets (Table 2). Meanwhile, 32 413 SNPs (13.0%) were confounded by gender (class C). We estimated the association by Mantel extension test and generalized Cochran–Mantel–Haenszel test. For the remaining 167 354 (66.8%), the CA trend test and the 2 d.f. genetic test were applied to the crude data set (class N). We arranged the *P*-values from all tests in ascending order and recruited the top 2304 SNPs, corresponding to 0.92% of the filtered SNPs, to the second stage. Of the 2304 SNPs, 1043 were from class M, 205 from class C and 1056 from class N. By this screening, we could select out SNPs whose *P*-values were <0.008.

In the second stage, 920 samples (460 controls/460 cases) were genotyped by GoldenGate assay for SNPs that had passed the first

Table 2 SNP classification according to possible effect modification and confounding by sex

	Sex effect modification (class M)	*Sex confounding (class C)*	*Others (class N)*	*Total*
Passed SNPs in first stage	50 740	32 413	167 354	250 507
		↓ 0.92% ($P<0.008$)[a]		
Genotyped in second stage	1043	205	1056	2304
		↓ SNP quality control		
Passed SNPs in second stage	838	162	833	1833
		↓ 0.95% ($P<2\times10^{-4}$)[a]		
Intensified study	7	4	11	22

Abbreviation: SNP, single nucleotide polymorphism.
[a]The *P*-values were assessed by proper tests in each category.

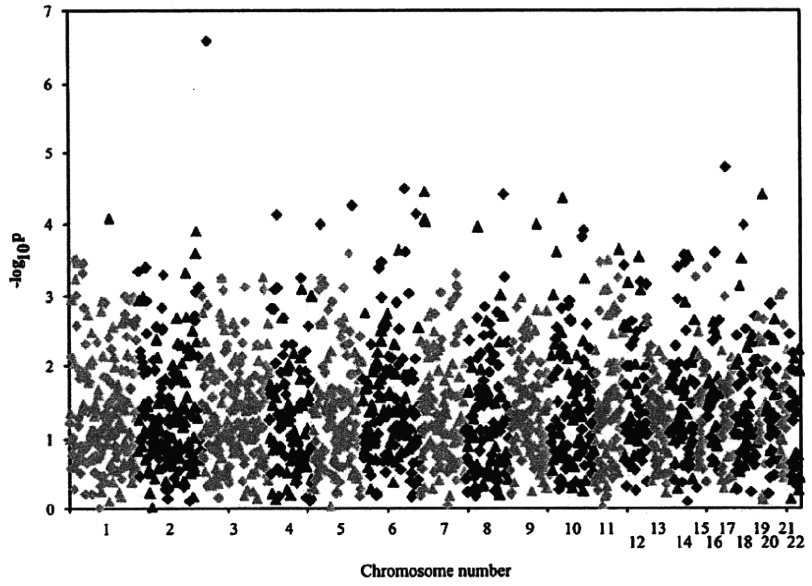

Figure 1 The scatter plot of $-\log_{10}P$ against all the analysis. The four data sets were plotted in $-\log_{10}P$ of GWAS. Triangles, diamonds and squares indicate SNPs in class N, M and C, respectively.

stage. Following the quality control described in Materials and methods for samples and SNPs, we combined the data set of both stages. On the basis of the same criteria of classification as the first stage, association tests were performed for the 1803 SNPs that passed quality control (Figure 1 and Supplementary Table 1). Although none of the SNPs reached genome-wide significance level (α=0.05/250 507$\approx 2\times10^{-7}$) after Bonferroni's correction, moderate associations were observed. We recruited the top 22 from 1833 SNPs, corresponding to 1.2% of filtered SNPs. The *P*-values of the 22 SNPs were <0.0002 (Table 2). Of the 22 SNPs, 7 (32%) were selected from class M, 4 (18%) from class C and 11 (50%) from class N.

Intensified study

We performed an intensified study for the 22 SNPs using larger case–control samples comprised of the GWAS cohort and additional samples to determine whether these SNPs would be noteworthy to be reported (Table 3). Specifically, they were genotyped in 1973 samples (904 controls/1069 cases). Following the quality control, we performed a statistical test for each SNP according to the classification at the GWA stages, and additionally calculated FPRPs. We finally identified five SNPs (rs7550260, rs7781293, rs4628172, rs1930095 and rs9864101) whose FPRP values with most stringent previous probability (1 in 10 000) were <0.5. Of them, only rs1930095, on chromosomes 9 had FPRP under significance level (FPRP=0.2) ($P=1.31\times10^{-5}$, FPRP=0.17). It is an intergenic SNP and about 250 kb upstream from the nearest gene, actin-like 7B. Of the four SNPs with FPRPs under suggestive level (FPRP=0.5), rs4628172 and rs7781293 ($P=1.32\times10^{-5}$ and 2.78×10^{-5}, respectively) were in physical proximity to each other, and located on the transmembrane protein 195 gene (*TMEM195*) on chromosome 7. The SNP rs7550260 with a *P*-value of 4.93×10^{-5} were located on Rho guanine nucleotide-exchange factor 11 (*ARHGEF11*) on chromosome 1. The last one, rs9864101 existed within the intron of the IQ motif and the Sec7 domain 1 gene (*IQSEC1*), annotated by expressed sequence tag, on chromosome 3. Per-allele ORs of the T allele of rs9864101 in male and female cohorts were 1.49 (95% confidence interval (CI), 1.24–1.80; $P=3.63\times10^{-5}$) and 1.07 (95% CI, 0.89–1.29; P=0.46), respectively. The effect modification by gender was significant (ψ-statistic=0.33; σ^2=0.018; P=0.014), and thus the association of the SNP with IA was observed only in the male subjects. The *P*-values of the five SNPs described above were led by CA trend test, and the ORs increased linearly with the number of risk allele. Furthermore, for these five SNPs, we investigated whether the association results were affected by potential confounding factors, gender and hypertension. When performing a multivariate logistic regression with gender and hypertension as covariates, we still observed significant associations of the SNPs (data not shown).

In contrast, SNPs in class C and associated only in female subjects had high FPRP values and were weakly or not associated with IA.

DISCUSSION

In this study, we performed a two-stage genome-wide scan in Japanese cohorts, followed by an intensified study, through which we identified five promising SNPs that may be responsible for predisposition to IA. The two SNPs with the second and forth smallest *P*-value, rs4628172 ($P=1.32\times10^{-5}$) and rs7781293 ($P=2.78\times10^{-5}$) were located on *TMEM195* on chromosome 7. The r^2 linkage disequilibrium between the SNPs was 0.78 in our data set, and their physical distance was 4.1 kb in length. *TMEM195* is a member of the fatty-acid hydroxylase family, which has transmembrane domains, although the function is not fully understood. The SNP rs7550260 was located on *ARHGEF11* on chromosome 1. The encoded protein may form a complex with G proteins and stimulate Rho-dependent signals. Intriguingly, in rat, overexpression of the ortholog induces the reorganization of the actin cytoskeleton and the formation of membrane ruffling.[23] The SNP rs9864101, whose effect on IA was modified by gender, is located in the gene *IQSEC1*. ARF-GEP100 protein (ADP-ribosylation factor–guanine nucleotide-exchange protein—100 kDa) encoded by this gene activates ADP-ribosylation factor protein, ARF6, and regulates cell adhesion through recycling E-cadherin and remodeling actin cytoskeleton.[24] We speculate that these three products might act in the uniform pathway of actin remodeling with *ELN/LIMK*[25] in smooth

Table 3 SNPs associated suggestively with IA in the GWAS

								Odds ratio[a] (95% CI)			FPRP[b]			
Sex effect	*SNP*	*Chromosome*	*Position*	*Non-risk allele*	*Risk allele*	*Risk allele frequency (control/case)*	P[c]	*Per allele*[d]	*Heterozygote*[e]	*Homozygote*[f]	*0.01*	*0.001*	*0.0001*	*Gene*
Confounding														
(class C)	rs7702812	5	154 485 773	A	G	0.44/0.48	1.44×10^{-2}	1.18 (1.03–1.35)	1.14 (0.92–1.42)	1.40 (1.07–1.83)	0.589	0.935	0.993	None
	rs10872573	6	100 917 343	T	C	0.30/0.35	5.02×10^{-4}	1.28 (1.11–1.47)	1.29 (1.05–1.57)	1.62 (1.19–2.21)	0.049	0.343	0.839	TBX18
	rs7923449	10	116 313 710	G	A	0.70/0.75	5.25×10^{-4}	1.30 (1.12–1.50)	1.26 (0.87–1.81)	1.65 (1.16–2.36)	0.052	0.357	0.847	None
	rs1978503	18	51 815 280	G	A	0.84/0.88	8.89×10^{-4}	1.38 (1.14–1.67)	1.29 (0.65–2.56)	1.81 (0.93–3.51)	0.103	0.537	0.921	None
Modification														
Male	**rs9864101**	**3**	**13 165 384**	**C**	**T**	**0.46/0.56**	**3.63×10^{-5}**	**1.49 (1.23–1.80)**	**1.64 (1.17–2.29)**	**2.23 (1.52–3.27)**	**0.006**	**0.060**	**0.391**	**IQSEC1**
(class M)	rs2703888	4	38 246 901	T	C	0.62/0.70	2.25×10^{-4}	1.45 (1.19–1.77)	1.51 (0.97–2.35)	2.14 (1.38–3.34)	0.037	0.277	0.793	None
	rs7209819	17	29 068 391	A	C	0.15/0.22	1.00×10^{-4}	1.61 (1.26–2.05)	1.54 (1.15–2.07)	3.01 (1.43–6.36)	0.039	0.290	0.804	ACCN1
Female	rs1565873	5	26 410 688	G	A	0.21/0.26	1.39×10^{-2}	1.31 (1.06–1.63)	1.29 (0.97–1.71)	1.79 (1.02–3.13)	0.613	0.941	0.994	None
(class M)	rs6570836	6	148 662 711	A	C	0.43/0.51	2.24×10^{-3}	1.33 (1.11–1.61)	1.47 (1.08–2.00)	1.75 (1.21–2.54)	0.204	0.721	0.963	None
	rs6470572	8	128 837 336	C	A	0.72/0.79	4.90×10^{-4}	1.44 (1.15–1.79)	0.78 (0.43–1.43)	1.36 (0.75–2.46)	0.155	0.649	0.949	None
	rs945335	10	112 877 539	T	C	0.39/0.46	1.25×10^{-3}	1.37 (1.13–1.66)	1.13 (0.84–1.52)	2.04 (1.36–3.05)	0.134	0.611	0.940	ABLIM1
Others														
(class N)	**rs7550260**	**1**	**155 268 285**	**C**	**A**	**0.34/0.41**	**4.93×10^{-5}**	**1.32 (1.15–1.50)**	**1.39 (1.13–1.70)**	**1.67 (1.26–2.22)**	**0.005**	**0.050**	**0.347**	**ARHGEF11**
	rs1367878	2	235 797 955	G	A	0.13/0.16	1.55×10^{-3}	1.34 (1.12–1.61)	1.31 (1.05–1.62)	2.05 (1.08–3.88)	0.153	0.646	0.948	None
	rs215939	6	85 523 005	A	C	0.57/0.61	8.39×10^{-3}	1.19 (1.05–1.35)	1.27 (0.98–1.64)	1.44 (1.10–1.88)	0.456	0.894	0.988	None
	rs7781293	**7**	**15 457 508**	**A**	**C**	**0.52/0.59**	**2.78×10^{-5}**	**1.32 (1.16–1.50)**	**1.59 (1.25–2.03)**	**1.81 (1.39–2.35)**	**0.003**	**0.029**	**0.232**	**TMEM195**
	rs4628172	**7**	**15 461 675**	**G**	**T**	**0.52/0.59**	**1.32×10^{-5}**	**1.30 (1.14–1.48)**	**1.69 (1.32–2.16)**	**1.78 (1.37–2.31)**	**0.006**	**0.057**	**0.376**	**TMEM195**
	rs2389409	7	15 473 314	G	A	0.58/0.63	5.66×10^{-4}	1.25 (1.09–1.42)	1.56 (1.20–2.03)	1.66 (1.27–2.17)	0.080	0.467	0.898	TMEM195
	rs3176292	8	21 959 262	A	G	0.68/0.74	1.76×10^{-4}	1.31 (1.14–1.51)	1.12 (0.80–1.57)	1.56 (1.12–2.18)	0.018	0.159	0.654	FGF17
	rs1930095	**9**	**110 394 739**	**T**	**C**	**0.16/0.21**	**1.31×10^{-5}**	**1.44 (1.22–1.71)**	**1.54 (1.25–1.89)**	**1.74 (1.08–2.79)**	**0.002**	**0.021**	**0.179**	**None**
	rs268300	10	43 900 845	G	A	0.15/0.20	1.01×10^{-4}	1.40 (1.18–1.67)	1.33 (1.08–1.63)	2.45 (1.39–4.33)	0.014	0.122	0.582	None
	rs1150229	11	113 338 362	G	A	0.45/0.51	1.09×10^{-4}	1.28 (1.13–1.46)	1.13 (0.91–1.41)	1.67 (1.29–2.16)	0.011	0.103	0.534	None
	rs873286	19	51 855 614	G	A	0.25/0.31	5.33×10^{-4}	1.28 (1.11–1.48)	1.22 (1.01–1.49)	1.75 (1.25–2.45)	0.052	0.358	0.848	DACT3

Abbreviations: CI, confidence interval; FPRP, false-positive report probability; GWAS, genome-wide association study; IA, intracranial aneurysm; SNP, single nucleotide polymorphism.
[a]Odds ratios (OR) were estimated by using a logistic.
[b]False-positive report probability under different prior probability. The SNPs with FPRP below 0.5 in 0.0001 prior probability are indicated in bold.
[c]The *P*-values were calculated by trend test or Mantel extension test, according to the class of each SNP.
[d]OR and 95% CI per copy of risk allele.
[e]OR for heterozygotes compared with nonrisk allele homozygotes.
[f]OR for risk allele homozygotes compared with nonrisk allele homozygotes.

muscle cells in blood vessels, and that disturbance of these gene functions might cause IA formation through altering plasticity of the arterial wall.

Among the SNPs to which gender acts as an effect modifier or confounder, rs9864101 was the only associated SNP (with FPRP < 0.5) whose genetic effect on IA was exerted in males only. Possibly, there was no association between IA and the SNPs in class M and C, except for rs9864101, but their risk allele frequencies or ORs could also have been too low to detect their genetic effects in this study design. In particular, for the SNPs in class M, the data set must be stratified by gender and then the sample size decreased by half. Thereby the statistical power is decreased.

We only focused on gender as a potential confounder or effect modifier of genetic association with IA in this study, because female gender is an established risk factor of IA. The gender is a distinct parameter, thus, it is filled almost entirely and accurately. On the other hand, the lifestyles, that is, smoking habit and frequent alcohol intake, of each individual are unstable over time and not completely described. We did not perform the GWA analysis after controlling for these important risk factors. Therefore, we may overlook SNPs modified and/or confounded by nongenetic factors other than gender. We also recognize the importance of the analysis taking lifestyle risk factors into account at a discovery stage, and thus the analysis is a challenge for the future.

Because Bilguvar *et al.*[11] have replicated three SNPs significantly associated with IA by using a subset of our Japanese cohort (rs1429412 on chromosome 2q, rs10958409 on chromosome 8q and rs1333040 on chromosome 9p in 495 IA cases and 676 non-IA controls), the SNPs are likely to be true positives. In our study, we tested for association between the three SNPs and IA at the first stage of the current GWAS (194 IA cases and 288 controls) and obtained ORs and 95% CIs for the SNPs as follows: rs1429412, OR=1.13 (95% CI, 0.79–1.33); rs10958409, 1.20 (0.78–1.53); and rs1333040, 1.29 (0.98–1.72). The ORs were not so different from those in the previous study,[11] but the effect sizes showed no statistical significance. This GWAS has limited power to detect loci with small effect sizes (for example, <10% power to detect common variants with genotypic relative risk of 1.3; also see Supplementary Figure 1). We therefore think that this is the primary reason we could not detect the previously identified loci in this study. On the other hand, none of the SNPs identified in this study were associated with IA in Bilguvar *et al.*' s study (data not shown, personal communication from Professor M. Gunel, Yale University). Although we cannot rule out the possibility of the false-positive SNPs, our findings are suggested to be Japanese population-specific susceptibility variants, and the genes and genetic regions would be powerful candidates for IA-predisposing factors in the Japanese population. To prove this, it will be necessary to carry out more detailed investigations using larger sample sizes of independent sets of Japanese and/or Asian cohorts. We further provided the information of allele frequencies of 1833 SNPs filtered out through the current GWAS (Supplementary Table 1). We expect that this is helpful for future studies in search of Japanese- and/or Asian-specific susceptibility loci for IA.

ACKNOWLEDGEMENTS

We are grateful to the DNA donors and the supporting staffs for making this study possible. We thank Dr Hiroyuki Akagawa for his assistance with DNA specimen management. We also thank Makiko Funamizu, Midori Yamamoto, Kayako Fukuyama, Kozue Otaka, Eriko Tokubo, Hiromi Kamura and Miho Takabe for their technical help. This work was supported in part by CREST (Core Research for Evolutionary Science and Technology) of Japan Science and Technology (II) and a Grant-in-Aid for scientific research on Priority Area 'Applied Genomics' from the Japanese Ministry of Education, Science, Sports and Culture (II).

1 Inagawa, T., Tokuda, Y., Ohbayashi, N., Takaya, M. & Moritake, K. Study of aneurysmal subarachnoid hemorrhage in Izumo City, Japan. *Stroke* **26**, 761–766 (1995).
2 Longstreth, W. T. Jr., Nelson, L. M., Koepsell, T. D. & van Belle, G. Clinical course of spontaneous subarachnoid hemorrhage: a population-based study in King County, Washington. *Neurology* **43**, 712–718 (1993).
3 Schievink, W. I., Wijdicks, E. F., Parisi, J. E., Piepgras, D. G. & Whisnant, J. P. Sudden death from aneurysmal subarachnoid hemorrhage. *Neurology* **45**, 871–874 (1995).
4 Onda, H., Kasuya, H., Yoneyama, T., Takakura, K., Hori, T., Takeda, J. *et al.* Genome-wide-linkage and haplotype-association studies map intracranial aneurysm to chromosome 7q11. *Am. J. Hum. Genet.* **69**, 804–819 (2001).
5 Akagawa, H., Tajima, A., Sakamoto, Y., Krischek, B., Yoneyama, T., Kasuya, H. *et al.* A haplotype spanning two genes, ELN and LIMK1, decreases their transcripts and confers susceptibility to intracranial aneurysms. *Hum. Mol. Genet.* **15**, 1722–1734 (2006).
6 Akagawa, H., Narita, A., Yamada, H., Tajima, A., Krischek, B., Kasuya, H. *et al.* Systematic screening of lysyl oxidase-like (LOXL) family genes demonstrates that LOXL2 is a susceptibility gene to intracranial aneurysms. *Hum. Genet.* **121**, 377–387 (2007).
7 Mineharu, Y., Inoue, K., Inoue, S., Kikuchi, K., Ohishi, H., Nozaki, K. *et al.* Association analyses confirming a susceptibility locus for intracranial aneurysm at chromosome 14q23. *J. Hum. Genet.* **53**, 325–332 (2008).
8 Ruigrok, Y. M., Rinkel, G. J., van't Slot, R., Wolfs, M., Tang, S. & Wijmenga, C. Evidence in favor of the contribution of genes involved in the maintenance of the extracellular matrix of the arterial wall to the development of intracranial aneurysms. *Hum. Mol. Genet.* **15**, 3361–3368 (2006).
9 Ruigrok, Y. M., Rinkel, G. J. & Wijmenga, C. The versican gene and the risk of intracranial aneurysms. *Stroke* **37**, 2372–2374 (2006).
10 Weinsheimer, S., Goddard, K. A., Parrado, A. R., Lu, Q., Sinha, M., Lebedeva, E. R. *et al.* Association of kallikrein gene polymorphisms with intracranial aneurysms. *Stroke* **38**, 2670–2676 (2007).
11 Bilguvar, K., Yasuno, K., Niemela, M., Ruigrok, Y. M., von Und Zu Fraunberg, M., van Duijn, C. M. *et al.* Susceptibility loci for intracranial aneurysm in European and Japanese populations. *Nat. Genet.* **40**, 1472–1477 (2008).
12 WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
13 Broderick, J. P., Viscoli, C. M., Brott, T., Kernan, W. N., Brass, L. M., Feldmann, E. *et al.* Major risk factors for aneurysmal subarachnoid hemorrhage in the young are modifiable. *Stroke* **34**, 1375–1381 (2003).
14 Krischek, B. & Inoue, I. The genetics of intracranial aneurysms. *J. Hum. Genet.* **51**, 587–594 (2006).
15 Sankai, T., Iso, H., Shimamoto, T., Kitamura, A., Naito, Y., Sato, S. *et al.* Prospective study on alcohol intake and risk of subarachnoid hemorrhage among Japanese men and women. *Alcohol. Clin. Exp. Res.* **24**, 386–389 (2000).
16 Yamada, S., Koizumi, A., Iso, H., Wada, Y., Watanabe, Y., Date, C. *et al.* Risk factors for fatal subarachnoid hemorrhage: the Japan Collaborative Cohort Study. *Stroke* **34**, 2781–2787 (2003).
17 Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
18 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
19 Thomas, D. C. *Statistical Methods in Genetic Epidemiology* (Oxford University, New York, 2004).
20 Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl Cancer Inst.* **96**, 434–442 (2004).
21 Chanock, S. Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. *Dis. Markers* **17**, 89–98 (2001).
22 Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B. *et al.* Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* **357**, 443–453 (2007).
23 Jackson, M., Song, W., Liu, M. Y., Jin, L., Dykes-Hoberg, M., Lin, C. I. *et al.* Modulation of the neuronal glutamate transporter EAAT4 by two interacting proteins. *Nature* **410**, 89–93 (2001).
24 Hiroi, T., Someya, A., Thompson, W., Moss, J. & Vaughan, M. GEP100/BRAG2: activator of ADP-ribosylation factor 6 for regulation of cell adhesion and actin cytoskeleton via E-cadherin and alpha-catenin. *Proc. Natl Acad. Sci. USA* **103**, 10672–10677 (2006).
25 Maekawa, M., Ishizaki, T., Boku, S., Watanabe, N., Fujita, A., Iwamatsu, A. *et al.* Signaling from Rho to the actin cytoskeleton through protein kinases ROCK and LIM-kinase. *Science* **285**, 895–898 (1999).

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

Online journal ISSN 1676-5680
GMR Genetics and Molecular Research

Positional effects of polymorphisms in probe-target sequences on genoplot images of oligonucleotide microarrays

T.L. Cui[1], H. Nakaoka[1], K. Akiyama[1], H. Kamura[1], K. Hosomichi[1], J. Bae[2], H. Cheong[2], H. Shin[2,3], T. Yada[4] and I. Inoue[1]

[1]Division of Molecular Life Science, School of Medicine, Tokai University, Kanagawa, Japan
[2]Department of Life Science, Sogang University, Seoul, Korea
[3]Department of Genetic Epidemiology, SNP Genetics, Inc., Complex B, Seoul, Korea
[4]Graduate School of informatics, Kyoto University, Kyoto, Japan

Corresponding author: T.L. Cui
E-mail: tailinc@is.icc.u-tokai.ac.jp

Genet. Mol. Res. 9 (1): 524-531 (2010)
Received November 20, 2009
Accepted January 4, 2010
Published March 23, 2010

ABSTRACT. Single nucleotide polymorphisms (SNPs) present in probe-target sequences (SPTS) have been shown to be associated with abnormal genoplot images. We explored the effects of SPTS positions on genoplot images using a data set from a genome-wide association study typed on an Illumina Human Hap300 platform. We screened the physical genomic positions of 308,330 autosomal probes to identify SPTS candidates deposited in dbSNP. The genoplot images across 293 individuals were inspected further in SNPs bearing an SPTS candidate. We identified 35,185 SNPs bearing a single SPTS candidate, including 264 SNPs showing abnormal genoplot images. The frequencies of SPTS at distances within 10 bases from the target SNP were significantly higher in the 264 SNPs showing abnormal genoplot images, than in the remaining 34,921 SNPs (49.62 *vs* 12.87%; Fisher exact test; $P = 2.2 \times 10^{-16}$). Of these 264 SNPs, we randomly selected 20 SNPs and resequenced them in 97 individuals. An SPTS within 10 bases of the target SNP was confirmed in all 20 SNPs, except for one SNP with a small deletion (7 bases) in the probe-

target sequence. Taken together, these results suggest an association of a proximal SPTS with an abnormal genoplot image, which could result in spurious genotype detections, highlighting the importance of minimizing systematic errors in microarray experiments.

Key words: Probe-target sequence; Genoplot image; Positional effects; Oligonucleotide microarray

INTRODUCTION

A typical genoplot image of a genotyping microarray shows three discrete groups corresponding to homozygous and heterozygous genotypes (Gunderson et al., 2005). However, in some situations, the genoplot represents an abnormal cluster pattern (abnormal genoplot image), which interferes with genotype detections (Franke et al., 2008).

An abnormal genoplot image may represent a biologically meaningful failure, for instance, the presence of untyped third alleles, such as a deletion copy number variation (CNV) mapped within a single nucleotide polymorphism (SNP) site, or a polymorphism present in probe-target sequences (SPTS) (McCarroll et al., 2006; Franke et al., 2008). In case of a deletion CNV, the abnormal genoplot image is shown to correspond to individuals with hemizygous and homozygous deletions. Consequently, individual genotypes with a hemizygous deletion are miscalled as homozygotes, while individual genotypes with homozygous deletions are called as missing (McCarroll et al., 2006).

Hybridization-based technologies, such as microarray, rely on the precise probe interaction between a microarray probe and its target sequence to ensure specific and accurate signal intensity measurements (Fan et al., 2007; Sliwerska et al., 2007; Zhang et al., 2007; Bemmo et al., 2008; Wang et al., 2008). The presence of SNP present in an SPTS may affect hybridization affinities due to single nucleotide mismatch of SPTS with a microarray probe, consequently leading to an abnormal genoplot image and false-positive results (Franke et al., 2008; Benovoy et al., 2008).

With the increasing volume of SNPs deposited in public databases, such as NCBI dbSNP, it is possible that there is an SNP within the immediately adjacent locus that is complementary to the 50-base probe (used in the Illumina Infinium chemistry). This raises questions of which positions of SPTS are associated with the abnormal genoplot image.

In the present study, we explored the association of SPTS positions with the genoplot image, using a data set from a genome-wide association study typed on Illumina Human Hap300 platform.

MATERIAL AND METHODS

Data sets for SPTS discovery

A data set from our unpublished genome-wide association study (GWAS) of 293 unrelated individuals that passed quality control and that had been typed on the Illumina Infinium II Human Hap300 BeadChip platform for 317,503 SNPs was used in this analysis. These included 97 controls and 196 patients with intracranial aneurysm in a Japanese population. The

Ethics Committees of Tokyo Women's Medical University approved the study protocols, and all participants gave written informed consent.

Identification of SNPs bearing SPTS candidate deposited in dbSNP

To identify these SPTS candidates, we screened the physical genomic positions in which the 50-base probes annealed and determined whether more SNPs were known in these loci in dbSNP (Build 127) using an automated in-house developed algorithm. According to this algorithm, a BLAST hit is considered to be an SPTS candidate if an SNP of dbSNP is located within a probe-targeted genomic region. All analyses were performed on the NCBI Build 36 Genomic Assembly. We identified 35,185 SNPs bearing a single SPTS candidate (fraction D in Figure 1) and 5317 SNPs bearing multi-SPTS (at least 2 SPTS candidates) (fraction C in Figure 1). Sets of 5317 SNPs and 15,139 SNPs that mapped within a common CNV (individual frequency above 5%) deposited in Database of Genomic Variants (http://projects.tcag.ca/variation) were removed from our assay (fraction B in Figure 1).

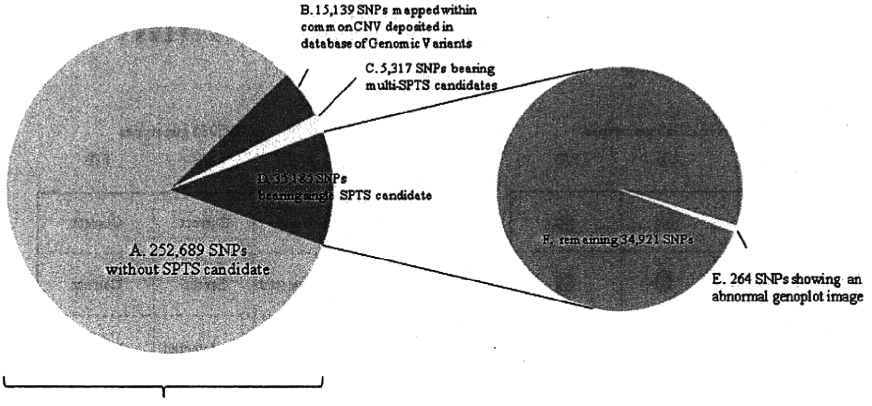

Figure 1. Identification of single nucleotide polymorphism (SNP) candidates with the presense of probe-target sequence (SPTS) of Illumina Human Hap300 BeadChip. **A.** 252,689 SNPs without SPTS candidate. **B.** 15,139 SNPs mapped within common copy number variation (CNV) deposited in Database of Genomic Variants. **C.** 5,317 SNPs bearing multiple SPTS candidates. **D.** 35,185 SNPs bearing single SPTS candidate. **E.** 264 SNPs showing an abnormal genoplot image. **F.** Remaining 34,921 SNPs bearing single SPTS candidate.

Visual inspection of genoplot

Typically, the genoplot image shows three typical clusters (A/A, A/B and B/B) corresponding to homozygous and heterozygous genotypes, when the minor allele frequency is sufficiently high (Sapolsky et al., 1999; Chen and Kwok, 1999; van Heel et al., 2007). However, the individual genotypes were grouped into additional two or three extra-clusters that fall below expectations in the presence of an SPTS (Franke et al., 2008). We refer to this

as an abnormal genoplot image. For instance, individual genotypes of rs1325065 (T/C) with an SPTS and rs9404149 (T/G) positioned at 2 bases from rs1325065 showed an abnormal genoplot image consisting of three typical clusters (green, yellow and blue circles) and three extra-clusters (pink, violet and black circles) (Figure 2A).

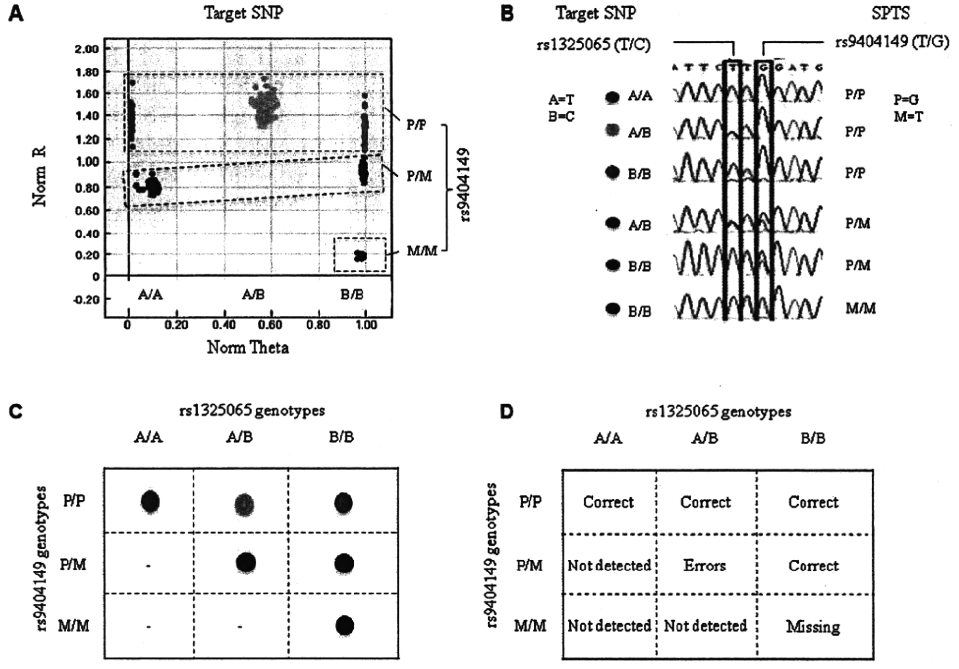

Figure 2. Abnormal genoplot image corresponding to the presence of single nucleotide polymorphism (SNP) present in a probe-target sequence (SPTS). **A.** Genoplot of rs1325065 (T/C) bearing an SPTS, rs9404149 (T/G) positioned at distances 2 bases from rs1325065. Individual genotypes grouped into six clusters comprising three typical clusters (A/A: green, A/B: yellow, and B/B: blue circles) and three extra-clusters (A/B: pink, B/B: violet, and B/B: black circles). P/P: perfect match allelic homozygote of rs9404149, P/M: heterozygote of rs9404149, and M/M: mismatch allelic homozygote of rs9404149. The clusters included in the dashed frames are shown to correspond to the genotypes P/P, P/M and M/M of rs9404149, respectively. **B.** Resequencing for region flanking rs1325065 and rs9404149. The color-coded circles represent the genotypes of both rs1325065 and rs9404149, corresponding to the clusters in *Panel A*. **C.** Genotype combination between rs1325065 (target SNP) and rs9404149 (SPTS) for each cluster. **D.** Genotype combinations between rs1325065 (target SNP) and rs9404149 (SPTS) for missing and erroneous genotypes.

To identify SNPs showing an abnormal genoplot image, we carried out visual inspection of the genoplot image in these 35,185 SNPs bearing a single SPTS candidate using the Beadstudio software (Illumina® Beadstudio 3.0), and observed 264 SNPs showing an abnormal genoplot image (group E in Figure 1).

Resequencing

Of these 264 SNPs showing an abnormal genoplot image, we randomly selected 20 SNP candidates with an SPTS, and carried out direct sequencing in 97 control individuals using the ABI 3130XL Genetic Analyzer sequencer (Applied Biosystems).

RESULTS AND DISCUSSION

In this study, we identified 35,185 SNPs bearing a single SPTS candidate (fraction D in Figure 1), of which 264 SNPs showed an abnormal genoplot image (fraction E in Figure 1). We found that the frequencies of SPTS positioned at distances ranging from 1 to 50 bases (probe size) gradually decrease with proximity to the target SNP in 35,185 SNPs (Figure 3A), which is apparently different from expected uniform distributions (Zhao and Boerwinkle, 2002; Madsen et al., 2007). However, their frequency at distance 1 base is over-represented, because of the high CpG mutation rate (Hwang and Green, 2004).

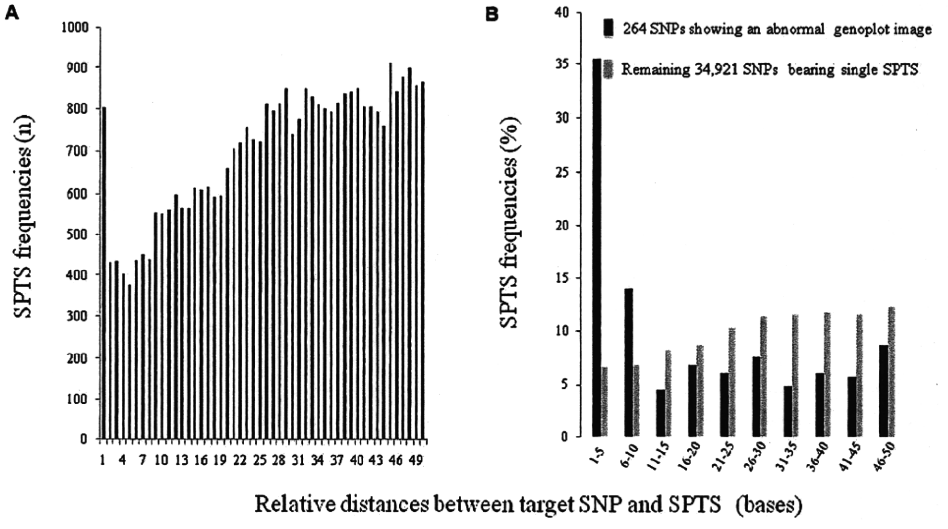

Figure 3. Distributions of probe-target sequence (SPTS) frequencies at different distances ranging from 1 to 50 bases relative to target single nucleotide polymorphism (SNP). **A.** SPTS frequencies in 35,185 SNPs bearing single SPTS candidate. **B.** SPTS frequencies in 264 SNPs showing an abnormal genoplot image and remaining 34,921 SNPs.

The frequencies of SPTS positioned at distances within 10 bases (defined as proximal SPTS) from the target SNP are significantly higher in these 264 SNPs showing an abnormal genoplot image, than in the remaining 34,921 SNPs without an abnormal genoplot image (49.62 *vs* 12.87%; Fisher exact test; $P = 2.2 \times 10^{-16}$) (Figure 3B). In contrast, the frequencies of SPTS positioned at distances 11 to 50 bases (defined as distal SPTS) are significantly higher

in 34,921 SNPs than the 264 SNPs showing an abnormal cluster pattern. This different distribution of SPTS frequencies between SNPs with and without an abnormal genoplot image suggests the association of proximal SPTS with the abnormal genoplot image.

Of these 264 SNPs, we randomly selected 20 SNPs and resequenced them in 97 individuals to confirm the presence of a proximal SPTS, which correlated with the abnormal genoplot image identified. Interestingly, a proximal SPTS was confirmed in 19 of 20 SNPs (Tables 1 and 2). Of these 19 proximal SPTS, 9 proximal SPTS are novel SNPs and are not deposited in the dbSNP. Small deletion (7 bases) present in the probe-target sequence was confirmed in one SNP, rs1014824 (Table 1).

Table 1. Confirmation of a proximal probe-target sequence (SPTS) in 10 single nucleotide polymorphisms (SNPs) showing abnormal genoplot image.

Target SNP	Chr	SPTS	Relative distances (bp)	
			dbSNP[a]	Resequencing[b]
rs6431746	2	Novel	23	3 and 23
rs7677996	4	Novel	37	4 and 37
rs2245050	5	Novel	17	1 and 17
rs4896566	6	Novel	25	6 and 25
rs2543046	8	Novel	48	4 and 48
rs2441706	8	Novel	15	1 and 15
rs1962249	10	Novel	32	3 and 32
rs3818246	14	Novel	49	2 and 49
rs4786015	16	Novel	49	1 and 49
rs1014824	18		45	7-bp deletion

[a]Relative distances between targeted SNP and an SPTS deposited in NCBI dbSNP. [b]Relative distances between target SNP and an SPTS that was confirmed by direct sequencing.

Table 2. Effects of probe-target sequence (SPTS) on genotype detections in 10 single nucleotide polymorphisms (SNPs) showing an abnormal genoplot image.

Target SNP	Chr	SPTS	Relative distance	D'	n	Genotype combination between target SNP and SPTS {Error (%) /Missing (%)}		
						Ht *vs* P/M[a]	Ho *vs* M/M[b]	Other genotype combinations (Ho or Ht/P/P, Ho/Ht)
rs7424350	2	rs34721305	7	1	97	100/0	0/100	0/0
rs4373124	4	rs12645377	1	1	97	100/0	0/100	0/0
rs1325065	6	rs9404149	2	1	97	100/0	0/100	0/0
rs6954269	7	rs6953985	1	1	97	100/0	0/100	0/0
rs2543046	8	rs35615372	4	1	97	100/0	0/100	0/0
rs7079697	10	rs11200310	1	1	97	100/0	0/100	0/0
rs7103853	11	rs4756745	2	1	97	100/0	0/100	0/0
rs9538229	13	rs7326315	1	1	97	100/0	0/100	0/0
rs845567	20	rs6109557	2	1	97	100/0	0/100	0/0
rs396999	21	rs35907272	2	1	97	100/0	0/100	0/0

Ht and Ho = heterozygote and homozygote of target SNP, respectively. P/M, M/M and P/P = mismatch allelic heterozygote, mismatch allelic homozygote and perfect match allelic homozygote of SPTS, respectively. [a]Genotype combination of Ht *vs* P/M, and Ht are misclassified as Ho, at an error rate = 100%. [b]Genotype combination of Ho *vs* M/M, and Ho are called as missing, at a missing rate = 100%.

The Illumina Haman Hap300 platform uses relatively long probes (50 bases in length), which are less sensitive to single nucleotide mismatch, and hybridize immediately adjacent to the targeted SNP sites, followed by extension of a single nucleotide (Gunderson et al., 2005,

2006). Thus, a distal SPTS positioned at distances over 10 bases may allow efficient primer extension. However, a proximal SPTS is expected to affect hybridization affinities, leading to reduced allele-specific signal intensity measurements and abnormal genoplot image.

We noticed that the abnormal genoplot image is dependent on not only target SNP genotypes but also SPTS genotype (Figure 2A). The typical clusters were shown to correspond to an SPTS with perfect match allelic homozygote (P/P), while the extra-clusters were shown to correspond to an SPTS with mismatch allelic genotypes (P/M and M/M) (Figure 2A-C). The levels of signal intensity measurements (y-axis) somewhat reflected the genotypes, P/P, P/M and M/M, respectively (Figure 3A). Therefore, it is possible to deduce a set of genotypic imputation rules for SPTS through linkage disequilibrium pattern and cluster pattern (Franke et al., 2008). This direct detection of SPTS genotypes sometimes resulted in the identification of a novel SPTS, in perfect linkage disequilibrium with the target SNP (Table 1). This will be of particular interest in studies of exonic polymorphism, where rare SPTS could translate into phenotypic changes.

As with most genotyping platforms, it is difficult to distinguish heterozygous genotypes from homozygous genotypes in the case of a proximal SPTS (McCarroll et al., 2006). We found that the missing and erroneous genotypes occurred in fixed genotype combinations between target SNP and SPTS (Figure 2A-D). Following resequencing for 10 SNPs in 97 individuals, we found that individual genotypes with heterozygote (Ht) are misclassified as homozygote (Ho) in the presence of a proximal SPTS with heterozygote (P/M), at an error rate = 100%, while individual genotypes with homozygote (Ho) are called as missing in the presence of a proximal SPTS with mismatch allelic homozygote (M/M), at a missing rate = 100% (Table 2). No missing and erroneous genotypes were confirmed in the presence of a proximal SPTS with perfect match allelic homozygote (P/P) (data not shown).

In view of the perfect linkage disequilibrium (D' = 1) between proximal SPTS and target SNP (Table 2), it is possible to "correct" the genotypes of a target SNP from homozygote to heterozygote so that false-positive and false-negative associations can be prevented.

CONCLUSIONS

We demonstrated that proximal but not distal SPTS are associated with an abnormal genoplot image, which could result in spurious genotype detections. Undetected proximal SPTS are therefore likely to have an effect on the validity of SNP genotyping platforms, highlighting the importance of minimizing systematic errors in microarray experiments.

ACKNOWLEDGMENTS

We gratefully acknowledge the clinicians of Tokyo Women's Medical University who collected samples and obtained consent from sample donors. We express our appreciation to all study participants for donating their blood and time. We thank Dr. Christian Harkensee (University of Newcastle upon Tyne) for editing a number of drafts of this manuscript.

REFERENCES

Bemmo A, Benovoy D, Kwan T, Gaffney DJ, et al. (2008). Gene expression and isoform variation analysis using Affymetrix Exon Arrays. *BMC Genomics* 9: 529.