Table 2 The association of single nucleotide polymorphism with allopurinol-related Japanese patients with Stevens–Johnson syndrome or toxic epidermal necrolysis

							Dominant genotype mode		Allelic frequency mode	
Order	*SNP*	*Chromosome*	*Closest gene*	*Distance to gene (bp)*	*Case*[a]	*Control*[a]	P	*Odds ratio (95% CI)*	P	*MAF (%)*
1	rs2734583	6p21.3	*BAT1*	0	0/6/8	0/11/980	2.44×10^{-8}	66.8 (19.8–225.0)	4.62×10^{-8}	0.55
1	rs3094011	6p21.3	*HCP5*	6553	0/6/8	0/11/980	2.44×10^{-8}	66.8 (19.8–225.0)	4.62×10^{-8}	0.55
1	GA005234	6p22.1	*MICC*	0	0/6/8	0/11/980	2.44×10^{-8}	66.8 (19.8–225.0)	4.62×10^{-8}	0.55
4	rs3099844	6p21.3	*HCP5*	3693	1/5/8	0/11/978	2.47×10^{-8}	66.7 (19.8–224.5)	1.33×10^{-9}	0.56
5	rs9267445	6p21.1	*PPIAP9*	3776	0/6/8	0/11/971	2.58×10^{-8}	66.2 (19.7–222.9)	4.87×10^{-8}	0.56
6	rs17190526	6p21.3	*PSORS1C1*	−446	0/6/8	0/12/979	3.64×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs9263726	6p21.3	*PSORS1C1*	0	0/6/8	0/12/979	3.64×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs2233945	6p21.3	*PSORS1C1*	0	0/6/8	0/12/979	3.64×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs9263733	6p21.3	*POLR2LP*	139	0/6/8	0/12/979	3.64×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs9263745	6p21.3	*CCHCR1*	0	0/6/8	0/12/979	3.64×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs130077	6p21.3	*CCHCR1*	0	0/6/8	0/12/979	2.44×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs9263781	6p21.3	*CCHCR1*	0	0/6/8	0/12/979	2.44×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs9263785	6p21.3	*CCHCR1*	0	0/6/8	0/12/979	2.44×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs9263794	6p21.3	*TCF19*	0	0/6/8	0/12/979	2.47×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs1044870	6p21.3	*TCF19*	0	0/6/8	0/12/979	2.58×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs9263796	6p21.3	*POU5F1*	0	0/6/8	0/12/979	3.64×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs9263800	6p21.3	*POU5F1*	0	0/6/8	0/12/979	3.64×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
6	rs4084090	6p21.3	*HLA-C*	17691	0/6/8	0/12/979	3.64×10^{-8}	61.2 (18.4–203.5)	6.87×10^{-8}	0.61
19	rs3131643	6p21.3	*HCP5*	0	1/5/8	0/12/977	3.68×10^{-8}	61.1 (18.4–203.1)	2.08×10^{-9}	0.61
20	rs9263827	6p21.3	*PSORS1C3*	−3369	0/6/8	0/12/974	3.75×10^{-8}	60.9 (18.3–202.5)	7.07×10^{-8}	0.61
20	rs1634776	6p21.3	*HLA-B*	12661	0/6/8	0/12/974	3.75×10^{-8}	60.9 (18.3–202.5)	7.07×10^{-8}	0.61

Abbreviations: CI, confidence interval; MAF; minor allelic frequency; SNP, ingle nucleotide polymorphism.

[a] Number of subjects in minor homo/hetero/major homo.

Table 3 *HLA* types and representative genotypes in 6p21 of allopurinol-related Japanese patients with Stevens–Johnson syndrome or toxic epidermal necrolysis

ID	HLA-A		HLA-B		HLA-Cw		*rs2734583*	*rs3099844*	*rs9267445*	*rs9263726*	*rs3131643*	*rs1634776*
1	2402	3303	4002	**5801**	0302	0304	T/**C**	C/**A**	G/**C**	G/**A**	C/**T**	G/**A**
2	2402	3101	1501	5601	0303	0401	T/T	C/C	G/G	G/G	C/C	G/G
3	2402	3101	5201	**5801**	0302	1202	T/**C**	C/**A**	G/**C**	G/**A**	C/**T**	G/**A**
4	1101	1101	4801	**5801**	0302	0803	T/**C**	**A/A**	G/**C**	G/**A**	**T/T**	G/**A**
5	2402	2602	4006	5101	0801	1402	T/T	C/C	G/G	G/G	C/C	G/G
6	0201	1101	1518	3501	0401	0801	T/T	C/C	G/G	G/G	C/C	G/G
7	2402	3303	5201	**5801**	0302	1202	T/**C**	C/**A**	G/**C**	G/**A**	C/**T**	G/**A**
8	0201	2402	1527	4003	0304	0401	T/T	C/C	G/G	G/G	C/C	G/G
9	2402	2402	3501	5201	0303	1202	T/T	C/C	G/G	G/G	C/C	G/G
10	0210	1101	4002	4006	0401	0801	T/T	C/C	G/G	G/G	C/C	G/G
11	0207	2402	4601	5101	0102	1402	T/T	C/C	G/G	G/G	C/C	G/G
12	2402	3101	3901	4001	0304	0702	T/T	C/C	G/G	G/G	C/C	G/G
13	0207	3303	4601	**5801**	0102	0302	T/**C**	C/**A**	G/**C**	G/**A**	C/**T**	G/**A**
14	3101	3303	3901	**5801**	0302	0702	T/**C**	C/**A**	G/**C**	G/**A**	C/**T**	G/**A**
15	2402	3303	5101	**5801**	0302	1402	T/**C**	C/**A**	NA	G/**A**	**T/T**	NA
16	0201	3303	3802	**5801**	0302	0702	T/**C**	C/**A**	NA	G/**A**	**T/T**	NA
17	2402	3303	0702	**5801**	0302	0702	T/**C**	C/**A**	NA	G/**A**	C/**T**	NA
18	2402	3303	5101	**5801**	0302	0304	T/**C**	C/**A**	NA	G/**A**	**T/T**	NA

Abbreviations: HLA, human leukocyte antigen; NA, not available.

Single nucleotide polymorphisms data of rs2734583, rs3099844, rs9263726 and rs3131643 are from BeadChip analysis and TaqMan genotyping analysis. Single nucleotide polymorphisms data of rs9267445 and rs1634776 are from BeadChip analysis.

Underlines of HLA types mean that these types are in linkage disequilibrium. HLA-B*5801s are expressed by bold types.

Bold types of the nucleotide mean the variant allele.

Table 4a Association between *HLA-A* alleles and allopurinol-induced Stevens–Johnson syndrome or toxic epidermal necrolysis

	Number of alleles detected (allele frequency)			
HLA-A allele	*Case,* n = *36 (%)*	*General population control* (n = *986)*[a] *(%)*	P	*Odds ratio (95% CI)*
0201	3 (8.3)	10.9	0.7895	
0206	0 (0)	10.4	0.0426	
0207	2 (5.6)	3.4	0.3650	
0210	1 (2.8)	0.1	0.0692	
1101	4 (11.1)	8.1	0.5299	
2402	13 (36.1)	35.6	1.000	1.02 (0.51–2.04)
2601	0 (0)	9.8	0.0417	
2602	1 (2.8)	2.2	0.5657	
3101	4 (11.1)	7.7	0.5195	
3303	8 (22.2)	7.9	0.0077	3.32 (1.46–7.54)

Abbreviations: CI, confidence interval; HLA, human leukocyte antigen.
We listed the *HLA-A* types of which the allele frequencies in the Japanese population are more than 9% or which were detected in this study.
[a]General population control data are cited from Tanaka *et al.*[40]

Table 4b Association between *HLA-B* alleles and allopurinol-induced Stevens–Johnson syndrome or toxic epidermal necrolysis

	Number of alleles detected (allele frequency)			
HLA-B allele	*Case,* n = *36 (%)*	*General population control* (n = *986)*[a] *(%)*	P	*Odds ratio (95% CI)*
0702	1 (2.8)	5.2	1.000	
1501	1 (2.8)	7.2	0.5076	
1518	1 (2.8)	0.9	0.3025	
1527	1 (2.8)	0	0.0352	
3501	2 (5.6)	8.6	0.7621	
3802	1 (2.8)	0.3	0.1338	
3901	2 (5.6)	4.0	0.6520	
4001	1 (2.8)	5.1	1.0000	
4002	2 (5.6)	8.2	0.7620	
4003	1 (2.8)	1.1	0.3512	
4006	2 (5.6)	5.3	0.7150	
4403	0 (0)	6.9	0.1648	
4601	2 (5.6)	3.8	0.6441	
4801	1 (2.8)	2.7	1.0000	
5101	4 (11.1)	7.9	0.5244	
5201	3 (8.3)	13.7	0.4624	
5401	0 (0)	6.5	0.1620	
5601	1 (2.8)	1.0	0.3273	
5801	10 (27.8)	0.6	5.388×10^{-12}	62.8 (21.2–185.8)

Abbreviations: CI, confidence interval; HLA, human leukocyte antigen.
We listed the *HLA-B* types of which the allele frequencies in the Japanese population are more than 6.5% or which were detected in this study.
[a]General population control data are cited from Tanaka *et al.*[40]

*LD of HLA-B*5801 with SNPs on chromosome 6*

We compared the genotypic distributions of six SNPs, which were significantly associated with SJS/TEN (Table 2), with *HLA* types because these SNPs are located near the *HLA-B* gene. These 6 SNPs listed in Table 3 represent 21 SNPs in Table 2 because the other 15 SNPs are in absolute LD with 1 of the 6 SNPs. Representative six variants of the significant SNPs on chromosome 6 were found in all of the SJS/TEN patients who carried the *HLA-B*5801* (10 patients) (Table 3). Therefore, in order to evaluate LD in the Japanese

Table 4c Association between *HLA-Cw* alleles and allopurinol-induced Stevens–Johnson syndrome or toxic epidermal necrolysis

	Number of alleles detected (allele frequency)			
HLA-Cw allele	*Case,* n = *36 (%)*	*General population control (*n = *234)*[a] *(%)*	P	*Odds ratio (95% CI)*
0102	2 (5.6)	17.0	0.0859	
0302	10 (27.8)	0	5.303×10^{-10}	
0303	2 (5.6)	7.8	1.000	
0304	4 (11.1)	11.3	1.000	
0401	4 (11.1)	6.5	0.2961	
0702	4 (11.1)	11.3	1.000	
0801	3 (8.3)	10.9	0.7777	
0803	1 (2.8)	2.6	1.000	
1202	3 (8.3)	10.4	1.000	
1402	3 (8.3)	5.7	0.4559	
1403	0 (0)	12.2	0.0192	

Abbreviations: CI, confidence interval; HLA, human leukocyte antigen.
We listed the *HLA-Cw* types of which the allele frequencies in the Japanese population are more than 10% or which were detected in this study.
[a]General population control data are cited from Tokunaga *et al.*[41]

Table 5 The linkage disequilibrium between *HLA* types and representative single nucleotide polymorphisms on 6p21 of 206 Japanese individuals

HLA	*rs3099844*	*rs3131643*	*rs2734583*	*rs9267445*	*rs9263726*	*rs1634776*
A	0.821	0.621	0.835	0.798	0.847	0.803
B	0.973	0.873	1.000	1.000	1.000	0.996
Cw	0.984	0.773	1.000	1.000	1.000	0.909

Abbreviation: HLA, human leukocyte antigen.
Data are expressed in *D'*.

Table 6 The linkage disequilibrium between representative single nucleotide polymorphisms on 6p21 and *HLA-B*5801* of 206 Japanese individuals

SNP	*D'*	r^2
rs3099844	0.930	0.866
rs3131643	0.929	0.674
rs2734583	1.000	0.931
rs9267445	1.000	0.896
rs9263726	1.000	1.000
rs1634776	1.000	0.905

Abbreviation: SNP, single nucleotide polymorphism.

population, LD coefficients (*D'*) were calculated between classical class 1 *HLA* types and six representative SNPs at 6p21, using the *HLA*-type and SNPs genotype data of 206 Japanese individuals, including 141 SJS/TEN cases and an additional 65 non-SJS/TEN Japanese subjects. As shown in Tables 5 and 6 representative SNPs on chromosome 6 showed LD for the *HLAs*. In particular, three SNPs (rs2734583, rs9267445 and rs9263726) showed a strong linkage with *HLA-B* and *Cw* alleles (Table 5). LD between six representative SNPs in 6p21 and *HLA-B*5801* are shown in Table 6. A novel observation was the absolute LD ($D'=1$, $r^2=1$) between rs9263726 in *PSORS1C1* and the *HLA-B*5801* allele.

Discussion

In order to explore new genetic biomarkers associated with the occurrence of allopurinol-related SJS/TEN Japanese patients, we conducted a GWAS using 890321 SNPs from patients with allopurinol-related SJS/TEN and an ethnically matched control group. The GWAS data indicated that most SNPs significantly associated with allopurinol-related SJS/TEN are located on or close to genes that overlap the 6p21 region, especially the genes neighboring *HLA-B*. There was no significantly associated SNP in any other region of the genome (Figures 1 and 2 and Table 2), indicating that the 6p21 region has the most important role in the progress of allopurinol-related SJS/TEN. We expected to find SJS/TEN-associated SNPs, which are unrelated to *HLA-B*5801* from this GWAS study because the association of *HLA-B*5801* with SJS/TEN is incomplete (10/18) in Japanese patients in contrast to Han Chinese[7] and Thai patients.[8] However, most of significant SNPs were closely linked with *HLA-B*5801* (Table 6). Previous studies have indicated that a SNP (rs2395029) in the *HCP5*, which is on 6p21.3, is strongly associated with human immunodeficiency virus-1 set points,[28–30] abacavir-induced hypersensitivity[24–26] and flucloxacillin-induced liver injury.[31] This SNP is in strong LD with *HLA-B*5701* in Caucasians.[25] Another SNP in 6p21 in *PSORS1C1*, a psoriasis-susceptibility candidate gene, was related with psoriasis in Swedish and Canadian populations[17,18] and exhibits LD with *HLA-Cw*0602* in Canadian populations.[18] These reports suggest that SNPs located in 6p21 link with a specific type of classical class I *HLA* that could be an alternative biomarker for the physiological phenomenon. Therefore, we examined the LD between these SNPs, shown in Table 2, and *HLA-B*5801*, which has been regarded as a genetic biomarker of SJS/TEN not only in Han Chinese,[7] but also in Caucasians[9] and Japanese.[10] We found that all of the Japanese patients with the allopurinol-related SJS/TEN who had the *HLA-B*5801* (10 patients) also had variant SNPs of genes that are located in 6p21, including *BAT1*, *HCP5*, *PPIAP9*, *PSORS1C1* and *HLA-B* (Table 3). The analysis of the LD coefficients between SNPs located in 6p21 and *HLA* types in the Japanese population indicated that these SNPs are in strong LD with *HLA* types (Table 5), and an absolute LD between rs9263726 in *PSORS1C1* and *HLA-B*5801* was observed in the Japanese population (Table 6). These results mean that all subjects (14 individuals including 10 with allopurinol-related SJS/TEN) who carry *HLA-B*5801* are in complete accord with all subjects with minor A allele of rs9263726 in the Japanese population. Therefore, rs9263726 in *PSORS1C1* is an alternative biomarker for *HLA-B*5801* in the Japanese population. Conventional genotyping of rs9263726 based on allelic discrimination offers several advantages over *HLA-B* typing, which is determined by genotyping of several SNPs forming the *HLA-B*5801* haplotype. Various broadly used technologies (for example, TaqMan genotyping) allow the standardized identification of two distinct alleles in one reaction tube, limiting the risk of contamination and allowing high-throughput genotyping with high sensitivity and specificity. In addition, the test is largely independent of both the performance of and interpretation by laboratory personnel. SNP genotyping is also less time consuming and cheaper than sequence-based *HLA* typing, and it does not require specialized laboratories. Therefore, the easy detection of these SNPs has a practical and economical advantage in clinical application for predicting the onset of allopurinol-related SJS/TEN. Although the previous report revealed that three SNPs in *HLA* region strongly associated with allopurinol-related SCAR in Han Chinese,[7] the two SNPs analyzed by the Illumina Human 1M-DUO BeadChip showed only weak association in the Japanese. This ethnic difference might be due to the difference of LD.

The functional analysis of genes that carry these SNPs—including *HCP5*, *BAT1*, *PSORS1C1*, *CCHCR1*, *TCF19* and *POU5F1*—in the pathogenesis of allopurinol-related SJS/TEN might be useful for determining their relevance. *CCHCR1* is a regulator of keratinocyte proliferation or differentiation and is overexpressed in keratinocytes in psoriatic lesions.[20–23] *TCF19* is a potential trans-activating factor that could play an important role in the transcription of genes required for the later stages of cell cycle progression.[27] Possible psoriasis candidate genes near *HLA-B* include *PSORS1C1*,[17–19] *CCHCR1*,[22,23] and *POU5F1*.[32,33] Mutations in *BAT1* may be associated with rheumatoid arthritis.[34–36] *HCP5* encodes an endogenous retroviral element mainly that is expressed in immune cells and there is evidence that the SNP in this gene is protective against human immunodeficiency virus-1 infection.[37–39] The functions and relevance of these genes suggest that the pathogenesis of allopurinol-related SJS/TEN might involve not only an immune system disorder, but also processes of cell proliferation and differentiation.

In conclusion, the results of this GWAS of allopurinol-related SJS/TEN in Japanese patients show that SNPs in genes located in 6p21, which are in LD with *HLA-B*5801*, are strongly associated with the cutaneous adverse reaction. Therefore, these SNPs, especially rs9263726, prove to be predictors for allopurinol-related SJS/TEN in Japanese, and their genes might be involved in the pathogenesis of allopurinol-related SJS/TEN. The OR of rs9263726 is extremely high from this case–control study and the typing cost of SNP is much cheaper than that of *HLA* typing. Moreover, the SJS/TEN has a very severe adverse reaction of allopurinol, which is high mortality. Therefore, we believe that the screening of rs9263726 genotype before allopurinol administration is necessary to prevent SJS/TEN in allopurinol-treated Japanese patients, although its allele frequency is very low in the Japanese. Association analyses of other ethnic populations are needed for confirming and comparing the results obtained in this study. *In vitro* functional studies of these genes are also necessary for identification of the physiological and molecular pathways leading to allopurinol-related SJS/TEN.

Conflict of interest

The authors declare no conflict of interest except one member of JPDSC, Mitsubishi Tanabe Pharma, which is a distributor of allopurinol in Japan.

Acknowledgments

We thank all the patients and physicians for their cooperation with this study. We also thank Riken Genesis for the Illumina BeadChip analysis and StaGen for the statistical analysis. This study was supported in part by a Health and Labor Science Research Grant (Research on Advanced Medical Technology) from the Ministry of Health, Labor and Welfare of Japan.

References

1 Wortmann RL. Gout and hyperuricemia. *Curr Opin Rheumatol* 2002; **14**: 281–286.

2 Chung WH, Hung SI, Chen YT. Human leukocyte antigens and drug hypersensitivity. *Curr Opin Allergy Clin Immunol* 2007; **7**: 317–323.

3 Tohkin M, Ishiguro A, Kaniwa N, Saito Y, Kurose K, Hasegawa R. Prediction of severe adverse drug reactions using pharmacogenetic biomarkers. *Drug Metab Pharmacokinet* 2010; **25**: 122–133.

4 Bastuji-Garin S, Rzany B, Stern RS, Shear NH, Naldi L, Roujeau JC. Clinical classification of cases of toxic epidermal necrolysis, Stevens-Johnson syndrome, and erythema multiforme. *Arch Dermatol* 1993; **129**: 92–96.

5 French LE. Toxic epidermal necrolysis and Stevens Johnson syndrome: our current understanding. *Allergol Int* 2006; **55**: 9–16.

6 Bowman C, Delrieu O. Immunogenetics of drug-induced skin blistering disorders. Part I: perspective. *Pharmacogenomics* 2009; **10**: 601–621.

7 Hung SI, Chung WH, Liou LB, Chu CC, Lin M, Huang HP *et al.* HLA-B*5801 allele as a genetic marker for severe cutaneous adverse reactions caused by allopurinol. *Proc Natl Acad Sci USA* 2005; **102**: 4134–4139.

8 Tassaneeyakul W, Jantararoungtong T, Chen P, Lin PY, Tiamkao S, Khunarkornsiri U *et al.* Strong association between HLA-B*5801 and allopurinol-induced Stevens-Johnson syndrome and toxic epidermal necrolysis in a Thai population. *Pharmacogenet Genomics* 2009; **19**: 704–709.

9 Lonjou C, Borot N, Sekula P, Ledger N, Thomas L, Halevy S *et al.* A European study of HLA-B in Stevens-Johnson syndrome and toxic epidermal necrolysis related to five high-risk drugs. *Pharmacogenet Genomics* 2008; **18**: 99–107.

10 Kaniwa N, Saito Y, Aihara M, Matsunaga K, Tohkin M, Kurose K *et al.* HLA-B locus in Japanese patients with anti-epileptics and allopurinol-related Stevens-Johnson syndrome and toxic epidermal necrolysis. *Pharmacogenomics* 2008; **9**: 1617–1622.

11 Wilke RA, Lin DW, Roden DM, Watkins PB, Flockhart D, Zineh I *et al.* Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. *Nat Rev Drug Discov* 2007; **6**: 904–916.

12 Nakamura Y. Pharmacogenomics and drug toxicity. *N Engl J Med* 2008; **359**: 856–858.

13 Daly AK, Day CP. Genetic association studies in drug-induced liver injury. *Semin Liver Dis* 2009; **29**: 400–411.

14 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.

15 Zhao JH, Curtis D, Sham PC. Model-free analysis and permutation tests for allelic associations. *Hum Hered* 2000; **50**: 133–139.

16 Zhao JH. 2LD, GENECOUNTING and HAP: computer programs for linkage disequilibrium analysis. *Bioinformatics* 2004; **20**: 1325–1326.

17 Holm SJ, Carlen LM, Mallbris L, Stahle-Backdahl M, O'Brien KP. Polymorphisms in the SEEK1 and SPR1 genes on 6p21.3 associate with psoriasis in the Swedish population. *Exp Dermatol* 2003; **12**: 435–444.

18 Rahman P, Butt C, Siannis F, Farewell VT, Peddle L, Pellett FJ *et al.* Association of SEEK1 and psoriatic arthritis in two distinct Canadian populations. *Ann Rheum Dis* 2005; **64**: 1370–1372.

19 Zhang XJ, He PP, Wang ZX, Zhang J, Li YB, Wang HY *et al.* Evidence for a major psoriasis susceptibility locus at 6p21(PSORS1) and a novel candidate region at 4q31 by genome-wide scan in Chinese hans. *J Invest Dermatol* 2002; **119**: 1361–1366.

20 Suomela S, Elomaa O, Skoog T, Ala-aho R, Jeskanen L, Parssinen J *et al.* CCHCR1 is up-regulated in skin cancer and associated with EGFR expression. *PLoS One* 2009; **4**: e6030.

21 Tiala I, Wakkinen J, Suomela S, Puolakkainen P, Tammi R, Forsberg S *et al.* The PSORS1 locus gene CCHCR1 affects keratinocyte proliferation in transgenic mice. *Hum Mol Genet* 2008; **17**: 1043–1051.

22 Suomela S, Kainu K, Onkamo P, Tiala I, Himberg J, Koskinen L *et al.* Clinical associations of the risk alleles of HLA-Cw6 and CCHCR1*WWCC in psoriasis. *Acta Derm Venereol* 2007; **87**: 127–134.

23 Tiala I, Suomela S, Huuhtanen J, Wakkinen J, Holtta-Vuori M, Kainu K *et al.* The CCHCR1 (HCR) gene is relevant for skin steroidogenesis and downregulated in cultured psoriatic keratinocytes. *J Mol Med* 2007; **85**: 589–601.

24 Hughes AR, Mosteller M, Bansal AT, Davies K, Haneline SA, Lai EH *et al.* Association of genetic variations in HLA-B region with hypersensitivity to abacavir in some, but not all, populations. *Pharmacogenomics* 2004; **5**: 203–211.

25 Colombo S, Rauch A, Rotger M, Fellay J, Martinez R, Fux C *et al.* The HCP5 single-nucleotide polymorphism: a simple screening tool for prediction of hypersensitivity reaction to abacavir. *J Infect Dis* 2008; **198**: 864–867.

26 Mallal S, Phillips E, Carosi G, Molina JM, Workman C, Tomazic J *et al.* HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med* 2008; **358**: 568–579.

27 Teraoka Y, Naruse TK, Oka A, Matsuzawa Y, Shiina T, Iizuka M *et al.* Genetic polymorphisms in the cell growth regulated gene, SC1 telomeric of the HLA-C gene and lack of association of psoriasis vulgaris. *Tissue Antigens* 2000; **55**: 206–211.

28 Migueles SA, Sabbaghian MS, Shupert WL, Bettinotti MP, Marincola FM, Martino L *et al.* HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc Natl Acad Sci USA* 2000; **97**: 2709–2714.

29 Altfeld M, Addo MM, Rosenberg ES, Hecht FM, Lee PK, Vogel M *et al.* Influence of HLA-B57 on clinical presentation and viral control during acute HIV-1 infection. *AIDS* 2003; **17**: 2581–2591.

30 Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science* 2007; **317**: 944–947.

31 Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A *et al.* HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat Genet* 2009; **41**: 816–819.

32 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* 1999; **8**: 2165–2170.

33 Chang YT, Hsu CY, Chou CT, Lin MW, Shiao YM, Tsai CY *et al.* The genetic polymorphisms of POU5F1 gene are associated with psoriasis vulgaris in Chinese. *J Dermatol Sci* 2007; **46**: 153–156.

34 Okamoto K, Makino S, Yoshikawa Y, Takaki A, Nagatsuka Y, Ota M *et al.* Identification of I kappa BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. *Am J Hum Genet* 2003; **72**: 303–312.

35 Kilding R, Iles MM, Timms JM, Worthington J, Wilson AG. Additional genetic susceptibility for rheumatoid arthritis telomeric of the DRB1 locus. *Arthritis Rheum* 2004; **50**: 763–769.

36 Quinones-Lombrana A, Lopez-Soto A, Ballina-Garcia FJ, Alperi-Lopez M, Queiro-Silva R, Lopez-Vazquez A *et al.* BAT1 promoter polymorphism is associated with rheumatoid arthritis susceptibility. *J Rheumatol* 2008; **35**: 741–744.

37 van Manen D, Kootstra NA, Boeser-Nunnink B, Handulle MA, van't Wout AB, Schuitemaker H. Association of HLA-C and HCP5 gene regions with the clinical course of HIV-1 infection. *AIDS* 2009; **23**: 19–28.

38 Catano G, Kulkarni H, He W, Marconi VC, Agan BK, Landrum M *et al.* HIV-1 disease-influencing effects associated with ZNRD1, HCP5 and HLA-C alleles are attributable mainly to either HLA-A10 or HLA-B*57 alleles. *PLoS One* 2008; **3**: e3636.

39 Han Y, Lai J, Barditch-Crovo P, Gallant JE, Williams TM, Siliciano RF *et al.* The role of protective HCP5 and HLA-C associated polymorphisms in the control of HIV-1 replication in a subset of elite suppressors. *AIDS* 2008; **22**: 541–544.

40 Tanaka H, Akaza T, Juji T. Report of the Japanese Central Bone Marrow Data Center. *Clin Transpl* 1996; **9**: 139–144.

41 Tokunaga K, Ishikawa Y, Ogawa A, Wang H, Mitsunaga S, Moriyama S *et al.* Sequence-based association analysis of HLA class I and II alleles in Japanese supports conservation of common haplotypes. *Immunogenetics* 1997; **46**: 199–205.

Appendix

Japan Pharmacogenomics Data Science Consortium (JPDSC)

The Japan Pharmacogenomics Data Science Consortium is composed of Astellas Pharma, Otsuka Pharmaceutical, Daiichi Sankyo, Taisho Pharmaceutical, Takeda Pharmaceutical and Mitsubishi Tanabe Pharma, and is chaired by Ichiro Nakaoka (Takeda Pharmaceutical).

Gene 516 (2013) 33–38

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Identification of a novel gene by whole human genome tiling array

Hirokazu Ishida [a,b], Tomohito Yagi [a], Masami Tanaka [a], Yuichi Tokuda [a], Kazumi Kamoi [b], Fumiya Hongo [b], Akihiro Kawauchi [b], Masakazu Nakano [a], Tsuneharu Miki [b], Kei Tashiro [a,*]

[a] *Department of Genomic Medical Sciences, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kyoto, Japan*
[b] *Department of Urology, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kyoto, Japan*

ARTICLE INFO

Article history:
Accepted 21 November 2012
Available online 19 December 2012

Keywords:
Tiling array
Microarray
Novel gene
Prostate

ABSTRACT

When the whole human genome sequence was determined by the Human Genome Project, the number of identified genes was fewer than expected. However, recent studies suggest that undiscovered transcripts still exist in the human genome. Furthermore, a new technology, the DNA microarray, which can simultaneously characterize huge amounts of genome sequence data, has become a useful tool for analyzing genetic changes in various diseases. A version of this tool, the tiling DNA microarray, was designed to search all the transcripts of the entire human genome, and provides huge amounts of data, including both exon and intron sequences, by a simple process. Although some previous studies using tiling DNA microarray analysis have indicated that numerous novel transcripts can be found in the human genome, none of them has reported any novel full-length human genes. Here, to find novel genes, we analyzed all the transcripts expressed in normal human prostate cells using this microarray. Because the optimal analytical parameters for using tiling DNA microarray data for this purpose had not been established, we established parameters for extracting the most likely regions for novel transcripts. The three parameters we optimized were the threshold for positive signal intensity, the Max gap, and the Min run, which we set to detect all transcriptional regions that were above the average length of known exons and had a signal intensity in the top 5%. We succeeded in obtaining the full-length sequence of one novel gene, located on chromosome 12q24.13. We named the novel gene "POTAGE". Its 5841-bp mRNA consists of 26 exons. We detected part of exon 2 in the tiling data analysis. The full-length sequence was then obtained by RT-PCR and RACE. Although the function of POTAGE is unclear, its sequence showed high homology with genes in other species, suggesting it might have an important or essential function. This study demonstrates that the tiling DNA microarray can be useful for identifying novel human genes.

1. Introduction

DNA microarray analysis has been established as one of the most useful technologies for investigating the underlying pathogenesis of various diseases (Castellano et al., 2009; Heintzman et al., 2009; Hussain et al., 2009; Takata et al., 2010; Xu et al., 2005; Yeager et al., 2007; Yeager et al., 2009). Annotated gene expression levels and single nucleotide polymorphisms can be conveniently evaluated using this technique. The tiling DNA microarray is a variation that was developed for investigating all the transcripts of the whole genome, including those of undiscovered genes (Bertone et al., 2004; Johnson et al., 2005; Kapranov et al., 2002; Mockler et al., 2005; Royce et al., 2005; Schadt et al., 2004; Shoemaker et al., 2001). This innovation also allows us to investigate the pathogenesis of various diseases.

The Human Genome Project reported the first complete sequence of the human genome in 2003. This project found 30,000 fewer expressed genes than had been expected (International Human Genome Sequencing Consortium, 2004; Lander et al., 2001; Venter et al., 2001). However, even though the findings also suggested that more than 98% of all genomic sequences are not transcribed (Cheng et al., 2005), recent studies on these "non-coding" DNA regions have revealed that they have many functions. In addition, these regions contain computationally predicted genes that may encode functional DNA and/or proteins. These observations suggest that novel genes that are transcribed into RNA may be found in these regions.

Because the original DNA microarray technology, used for evaluating annotated gene expression levels, was designed with relatively few probes, usually covering only the 5′-ends of annotated genes, it is not very useful for finding undiscovered transcripts in unexplored genomic regions. Specifically, the number and location of the probes meant that transcribed regions that lay between the probes could

Abbreviations: mRNA, messenger RNA; RT-PCR, reverse transcription-polymerase chain reaction; RACE, rapid amplification of cDNA ends; kb, kilobases; GAPDH, glyceraldehyde-3-phosphate dehydrogenase.

[*] Corresponding author at: Department of Genomic Medical Sciences, Kyoto Prefectural University of Medicine, Kawaramachi-Hirokoji, Kamigyo-ku, Kyoto, 602-8566, Japan. Tel.: +81 75 251 5346; fax: +81 75 251 5346.

E-mail address: tashiro@koto.kpu-m.ac.jp (K. Tashiro).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.076

not be detected. In contrast, tiling DNA microarrays are useful for mapping novel transcripts, because the "tiling" feature consists of 25-mer oligonucleotide probes that are tiled at approximately 35-bp intervals, as measured from the central position of the adjacent probe. Therefore, a gap of only approximately 10 bp lies between probes, which cover the entire genome except the telomeres and centromeres (Sasaki et al., 2007). Tiling DNA microarray data have improved gene annotations and revealed the extensive transcriptions of non-coding RNAs. The closely spaced probes allow for the accurate measurement of small transcriptional features, such as single exons or small introns. This technology is now allowing us to investigate undiscovered transcripts as well as the expression of annotated genes. In this regard, the tiling DNA microarray is one of the most powerful and fruitful tools for evaluating both annotated genes and novel transcripts that have unclear functions. Previous reports using tiling DNA microarray have demonstrated novel transcripts in the human genome. However, full-length novel genes have not been reported (Kampa et al., 2004; Kapranov et al., 2005; Nelson et al., 2008; Weile et al., 2007).

In this study, we used tiling DNA microarray to seek undiscovered transcripts, and we demonstrated its usefulness for identifying a novel coding gene.

2. Materials and methods

2.1. Cell culture

Primary normal prostate epithelial cells (PrECs) were purchased from Lonza (Walkersville, MD) and maintained in prostate epithelial cell media (PrEGM Bullet Kit-Lonza) supplemented with a mixture of various growth factors (Single Quots-Lonza). Cells were seeded at recommended densities and cultured at 37 °C at 5% CO_2. Media were changed every 48 h.

2.2. RNA and DNA preparation

Total RNA was extracted from PrECs with the RNeasy Plus Mini Kit (Qiagen, Valencia, CA). RNA quality was evaluated by spectrophotometry with a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA) and by gel electrophoresis.

Genomic DNA was extracted from cancerous and normal frozen prostate tissue. The samples were minced and mixed well in lysis buffer with proteinase-K (to 0.2 mg/mL) and SDS (to 0.1%) at 55 °C. DNA was separated from the proteinaceous component by two extractions with an equal volume of phenol/chloroform isoamyl alcohol. The aqueous phase was mixed with 2.5 volumes of 100% ethanol and 0.1 volumes of 3 M sodium acetate and centrifuged at 12,000 ×g for 20 min at 4 °C. The DNA pellet was washed with cold 70% ethanol and allowed to air dry before resuspension in TE (10 mM Tris–HCl pH 7.5, 1 mM EDTA).

2.3. Affymetrix GeneChip® hybridization

Affymetrix Human Tiling 1.0R Array GeneChip® (Tiling array; Affymetrix, Santa Clara, CA) arrays were used for triplicate hybridizations. For the microarray hybridization, we followed the protocol described in the Affymetrix GeneChip® Whole Transcript Double-Stranded Target Assay Manual (Affymetrix, Santa Clara, CA). In brief, 6 μg of total RNA was purified by ribosomal RNA reduction using the RiboMinus Human/Mouse Transcriptome Isolation Kit (Invitrogen Co., Carlsbad, CA) and cleaned up. A single-stranded cDNA was synthesized using a T7-$(N)_6$ primer, and the cDNA was made double-stranded. The ds cDNA was amplified by in vitro transcription into complementary RNA (cRNA) and cleaned up. The second cycle ds cDNA was synthesized using the amplified cRNA as a template. The ds DNA was cleaned up, fragmented, and labeled with biotin. The fragmented ds DNA was used for hybridization to the microarrays at 45 °C for 16 h with a rotation rate of 60 rpm using a GeneChip® Hybridization Oven (Affymetrix, Santa Clara, CA). The microarrays were washed and stained using an Affymetrix GeneChip® Fluidics Station 450 and scanned by an Affymetrix GeneChip® Scanner 3000.

2.4. Tiling array data analysis

To handle the data generated by using probes that hybridize throughout the whole genome, we extracted the positive data as follows (Supplemental Fig. 1). As the distance used to locally group positional data for statistical analysis, we set the bandwidth at the maximum recommended level (73 bp). This setting increases the reliability of the signal intensity derived from a perfectly matched probe vs. a mismatched probe. After removing data that showed no signal intensity (43.9% of all probes in 14 arrays), the threshold for each array was set to filter out all but the top 5% of probe intensities. A positive probe was defined as one having a signal intensity greater than threshold (Supplemental Table 1, Supplemental Fig. 2) (Eisenberg and Levanon, 2003). To evaluate the DNA regions hybridizing with positive probes, we used a Max gap parameter (the maximum tolerated gap between positive positions in the derivation of detected regions) of 70 bp, to permit the hybridization of a negative probe between two positive probes. The Min run parameter (the minimum size of a detected region) was set at 140 bp, which is the approximate average length of all exons identified among the annotated genes of NCBI (http://www.ncbi.nlm.nih.gov/) Build 36.2 (Supplemental Table 1, Supplemental Fig. 3). The regions whose signals passed our three parameter settings were compared to those of annotated genes by the probe position, according to the information provided by NCBI Build 36 in the Affymetrix Integrated Genome Browser (IGB) (Nicol et al., 2009), to remove regions that overlapped with annotated genes. Next, the data were carefully divided into known or unknown transcripts by checking each sequence against the latest annotations. New transcripts that appeared within an annotated gene, even if not in the exonic sequences, were also considered to be gene-related transcripts, and we excluded them from further analysis. Finally, cases of two or more novel regions lying within 5-kbp on the genome were defined as "zones," and were investigated further.

2.5. RT-PCR and rapid amplification of cDNA ends (RACE)

Total RNA was extracted from the specimens using the RNeasy Plus Mini Kit (Qiagen, Valencia, CA). First-strand cDNA synthesis was performed using the SuperScript III First-Strand Synthesis System with an oligo $(dT)_{20}$ primer for RT-PCR (Invitrogen Co., Carlsbad CA), according to standard procedures.

Rapid amplification of cDNA ends (RACE) was performed with a GeneRacer kit (Invitrogen Co., Carlsbad, CA) and SMART RACE cDNA Amplification Kit (Clontech Laboratories, CA, USA), according to the manufacturer's instructions.

2.6. Sequencing analysis

Amplified RT-PCR and RACE products of target regions were sequenced with a BigDye terminator v1.1 or v3.1 Cycle Sequencing kit (Applied Biosystems, CA, USA) using Applied Biosystems 3130 Genetic Analyzers. The primers for the sequencing analysis were designed according to the results of each RACE analysis. The primer sequences are described in the supplementary information.

2.7. Quantitative RT-PCR assay in multiple human tissues

To evaluate the levels of POTAGE expression in human tissues, quantitative PCR (QPCR) was performed using the Stratagene Mx3005P real-time QPCR system with the Brilliant II Fast SYBR Green QPCR Master Mix (Agilent Technologies, CA, USA) and Human MTC Panels I and II, which include heart, brain, placenta, lung, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, testis, ovary, small intestine, colon, and peripheral leukocytes (Clontech Laboratories, CA, USA). These assays were performed at least four times in duplicate. To normalize the values in the quantitative assays, the level of *beta-Actin* was assessed as a control. The primer sequences for *beta-actin* were (forward) 5′-ATTGCCGACAGGATGCAGAA-3′ and (reverse) 5′-ACATCTGCTGGAAGGTGGACAG-3′. After the QRT-PCR assay, each sample was examined by agarose gel electrophoresis to evaluate the amplification of a single RT-PCR product.

2.8. Methylation assay with normal and cancerous human prostate tissue

DNA methylation is an important epigenetic mechanism of gene regulation. To investigate the methylation status of POTAGE in the human prostate, we performed methylation PCR using human prostate tissues. Normal and cancerous prostate samples were obtained from six patients, with their informed consent, during radical retropubic prostatectomy for clinically localized prostate adenocarcinoma, performed at the Kyoto Prefectural University of Medicine. All procedures were conducted in accordance with the Helsinki declaration. This study was approved by the Institutional Review Board of Kyoto Prefectural University of Medicine.

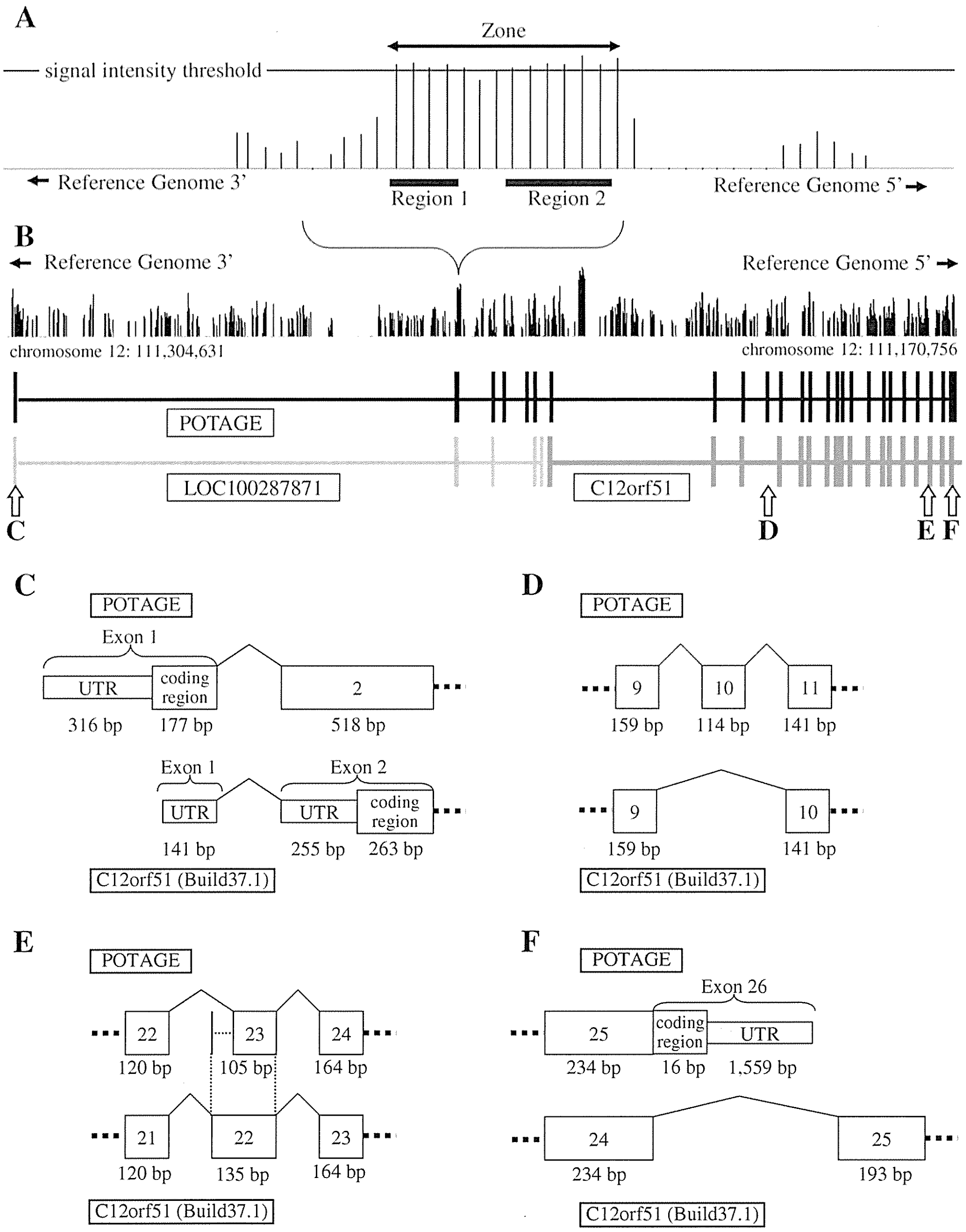


Fig. 1. (A) For this study, we defined two concepts: the region and the zone. A region was the genomic area that passed our settings for all three parameters and did not overlap with annotated genes. Zones were two or more regions that lay within 5-kbp of one another, on the chromosome. Zones were assumed to encode at least some portion of a novel gene. (B) Each novel zone was investigated in detail by RT-PCR and RACE. Shown is the region containing the novel protein, which overlapped with exon 2 of LOC100287871, a predicted gene in NCBI Build 36.3. We obtained the full-length mRNA sequence encoded by this region. (C, D, F) Although 22 of the 26 exons of the novel gene also served as exons of C12orf51, exons 1, 10, and 26 were new. (E) Exon 23 of the novel gene had a 30-bp deletion.

The genomic DNA samples from the six patients were treated with the MethylEasy Xceed Rapid DNA Bisulphite Modification Kit (Human Genetics Signatures Pty Ltd, Australia), according to the manufacturer's instructions. The methylation and unmethylation primers for POTAGE were designed using the CpG island searcher (http://cpgislands.usc.edu/) (Takai and Jones, 2003) and MethPrimer (http://www.urogene.org/methprimer/index1.html) web sites (Li and Dahiya, 2002). After the amplification, the PCR products were separated by electrophoresis on an agarose gel, and fragments in the expected range were excised and purified using the QIAquick Gel Extraction Kit (Qiagen, Valencia, CA). The purified PCR products were ligated using the pGEM-T Easy Vector System (Promega, WI, USA), and at least 20 independent clones were sequenced with the T7 (5′-TAATACGACTCACTATAGGG-3′) and SP6 (5′-ATTTAGGTGACACTATAGAA-3′) primers.

3. Results

3.1. Analysis of tiling array data

Our goal was to evaluate all the mRNAs expressed in human prostate cells, using the tiling array in triplicate. The signal intensity and P-value for each probe were determined by quantile normalization (Bolstad et al., 2003), after the raw intensity data from triplicate microarrays were transformed with the Affymetrix Tiling Analysis Software ver. 1.1. All of the extracted signal data were mapped to their genomic position and visualized in the IGB. Because the tiling array was designed based on information from NCBI Build 34, the results were translated to Build 36 automatically.

Because the thresholds determining positive signal intensity were determined on the basis of the signals from all the probes in each tiling array, the thresholds were slightly different for each array (see Materials and methods). The values for two other parameters (Max gap and Min run) were the same for all the arrays (see Materials and methods). The three parameter settings enabled us to predict the genomic locations likely to contain transcribed sequences. After comparing the sequences from our predicted regions with those of annotated genes (NCBI Build 37.1), we found 319 regions in the entire genomic sequence that encoded undiscovered transcripts. After the novel regions were obtained, the novel zones were defined by tiling data analysis. Finally, we defined 17 zones containing two or more regions within 5-kbp of each other (Fig. 1A, Supplemental Fig. 1 and Supplemental Table 1).

3.2. RT-PCR and RACE analysis of the novel region of human chromosome 12

We next designed primer sets for each region that were appropriate for performing RT-PCR analysis with the single-strand cDNA obtained from normal prostate cells. Each RT-PCR product was sequenced to confirm the amplification of the target sequences. Even when a positive tiling array signal was confirmed, no region was studied further without the successful amplification of the correct sequence. In addition, single regions that did not have any positive regions in the flanking regions were also excluded. After the RT-PCR analysis, primers for 5′- and 3′-RACE were designed on the basis of both the tiling array data and information from NCBI Build 37.1. All of the 5′- and 3′-RACE experiments were performed with single-stranded cDNA obtained from normal human prostate tissue.

Finally, we obtained the full-length sequence of POTAGE on chromosome 12q24.13 (Supplemental Table 2, primer Nos. 1–4, Supplemental Figs. 4, 5). However we succeeded to obtain the 17 zones by tiling array data, we failed to confirm 16 zones by RT-PCR and/or RACE.

The novel mRNA sequence we obtained consisted of 26 exons within an mRNA of 5841 bp. The gene was located on chromosome 12q24.13. Assuming that the tiling array data might indicate one of the exons in the novel transcripts, we performed RT-PCR and RACE of the regions in this zone. From these results, we found that 4 of the 6 5′-most exons of POTAGE belonged to the hypothetical protein LOC100287871 (http://www.ncbi.nlm.nih.gov/gene/?term=LOC100287871). Moreover, 18 of the 19 3′-most exons overlapped with part of the 5′-end of predicted gene C12orf51 in NCBI Build 37.1. There were three novel exons in POTAGE: exons 1, 10, and 26. Exon 23 of POTAGE contained a 30-bp deletion compared with the 5′-end of C12orf51 exon 22. The remaining 22 exons of POTAGE shared 100% identity with C12orf51 (Figs. 1B–F, Supplemental Table 3).

We also investigated the sequence flanking exons 25 and 26 of POTAGE in detail. We found at least two human isoforms of these exons. One isoform included exon 26 of POTAGE as its 3′-end; this isoform was equivalent to POTAGE. The other isoform had a different 3′-end; that is, some other exon followed exon 25. For example, C12orf51 was partially encoded by the other isoform.

To explore the possible function of POTAGE, motif and homology searches were performed using the MOTIF Search (http://motif.genome.jp/), Pfam (http://pfam.sanger.ac.uk/), and NCBI web sites. No major motif was found in the nucleotide acid sequence or the deduced protein sequence.

3.3. Comparison of POTAGE expression level in multiple human tissues

We evaluated the expression levels of POTAGE with region-specific primer pairs in multiple human tissues (Human MTC Panels I and II), using semi-quantitative real-time RT-PCR (Supplemental Table 2, primer No. 5). POTAGE was expressed in every human tissue examined. The relative expression levels were calculated as the levels normalized to the beta-actin expression in each sample. The highest expression level of POTAGE was observed in the testis (Fig. 2). We created primer pairs to evaluate the level of expression of the other isoform, and performed real time RT-PCR using the same conditions as for POTAGE. While the expression level of the other isoform also was higher in testis, it was different from that of POTAGE, in that the other isoform was also highly expressed in skeletal muscle (data not shown).

3.4. Methylation assay for the 5′-upstream CG-rich region of POTAGE

Because DNA methylation in the 5′-upstream CG-rich region of a gene is related to the repression of gene expression, we investigated the methylation status of POTAGE using a methylation-specific PCR assay, to discover if differences in methylation could explain the lower expression level of POTAGE in normal prostate tissue compared to other tissues (Jones and Baylin, 2007; Laird, 2003; Ting et al., 2006;

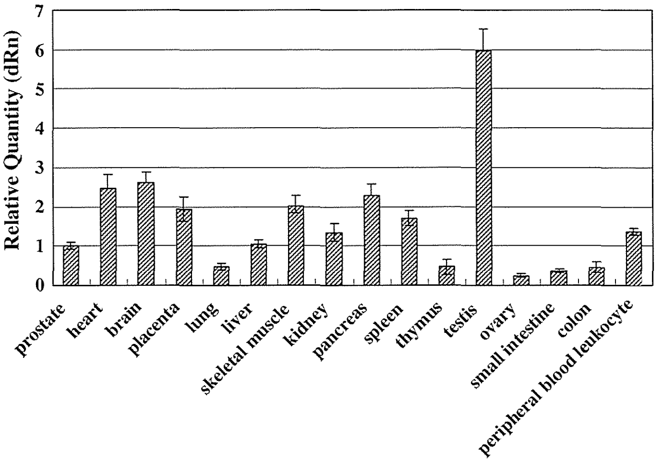


Fig. 2. The expression levels of the novel gene in multiple human tissues using semi-quantitative real-time RT-PCR. The novel gene was expressed in every human tissue examined in this study, and its level was highest in the testis.

Table 1
Homology among exons of the novel gene in human, mouse, rat, and fugu (pufferfish).

		cDNA (translated region)			
		Human	Mouse	Rat	Fugu
Amino acids	Human		89.18%	88.80%	70.60%
	Mouse	97.27%		95.14%	72.59%
	Rat	97.05%	98.19%		72.47%
	Fugu	81.15%	81.00%	80.92%	

The percentage of identical amino acids was essentially constant across species.

Vanaja et al., 2009). In addition, we examined the methylation condition in prostate cancer tissue, to assess any relationship between the level of expression and prostate oncogenesis. We examined the methylation of CpG 101 (UCSC (http://genome.ucsc.edu/) GRCh37/hg39), a CpG island located 557 bp upstream of POTAGE (Supplemental Table 2, Nos. 6–7, Supplemental Fig. 6). However, the CpG island was not methylated in normal or cancerous prostate tissues.

4. Discussion

New genomic technologies have yielded much useful information about the whole human genome, and both experimental and computational approaches have been developed to handle the accumulation of data. Our approach using the tiling array supported the importance of choosing the appropriate settings for the three parameters, threshold of signal intensity, Max gap, and Min run, when examining the tiling data to evaluate mRNA expression or discover novel genes. The settings of these parameters were critical to our finding the few pieces of relevant information among the enormous quantities of tiling array data. Because our data demonstrated that the signal patterns of many undiscovered regions were very short or very close to annotated genes, we excluded unknown regions with these patterns to obtain novel genes that were independent of the known genes. Therefore, the three parameters in our data were chosen to be stringent, to reduce the amount of data that would require further investigation.

First, our parameter settings allowed us to extract 17 zones containing novel regions from our entire set of tiling array data. All of the zones consisted of two or more novel regions within about 5-kbp of one another. We assumed that each novel region might represent one or more exons of a novel gene. In 16 of the 17 zones, each region was confirmed as encoding a transcript. However, no amplification product included neighboring sequences.

Finally, we determined the sequence of a full-length novel gene from the RT-PCR and RACE results of 1 of the 17 zones. The novel mRNA was 5841 bp in length and contained 26 exons, which were encoded by sequences spread over 133,876 bp in the genome, on chromosome 12q24.13. There were two predicted genes, hypothetical protein LOC100287871 and C12orf51, that were close to our discovered gene (NCBI Build 36.3). The information on hypothetical protein LOC100287871 was replaced with the predicted region of C12orf51 in NCBI Build 37.1. To confirm the sequence of POTAGE, we also referred to the sequence of the predicted transcript at NCBI Build 37.1.

Among the 26 exons of POTAGE, three were novel (exons 1, 10, and 26). POTAGE also shared 23 exons with C12orf51 (NCBI Build 37.1), and the 22nd exon of POTAGE (the 23rd exon of C12orf51) contained a 30-bp deletion.

The expression level of POTAGE was lower in the prostate than in most other tissues. To investigate whether POTAGE's expression was suppressed by methylation, we examined the genomic methylation in the 5′-upstream CpG island of POTAGE in normal prostate and prostate cancer, to look for possible associations between malignancies and expression of POTAGE. However, the region we assayed was not methylated in normal or cancerous prostate tissue. Therefore, the different expression levels in several tissues, including prostate cancer, are unlikely to be regulated by the methylation of the 5′-upstream CG rich region of POTAGE.

POTAGE has no major motif in the nucleotide acid sequence or the deduced protein sequence. Interestingly, the sequence of POTAGE had high homology to transcripts in other species, such as mouse and rat. The predicted protein had a 97% sequence identity between human and mouse, and almost the same homology between human and rat (Table 1). The alignment of the amino acid sequences, which was constructed using CLUSTAL W ver. 1.83 (http://clustalw.ddbj.nig.ac.jp/top-j.html) with Kimura's correction, between human and five other species, showed the closest matches among different species (Supplemental Fig. 7). Furthermore, the phylogenetic relationship based on the amino acid alignments of these six species also revealed high protein homology among human and other species (Fig. 3). On the other hand, the nucleotide sequence identities between human and mouse of two housekeeping proteins (beta-Actin and GAPDH) are 92% and 89%. The between-species percent identity of POTAGE was higher than that of these housekeeping genes. Therefore, although the function of POTAGE is currently unknown, its high sequence homology among different species suggests that it may have an important or essential biological function.

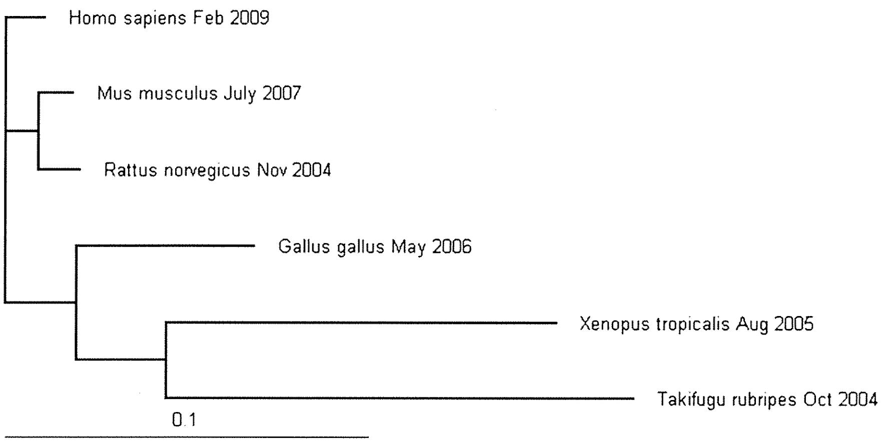

Homo sapiens Feb 2009
Mus musculus July 2007
Rattus norvegicus Nov 2004
Gallus gallus May 2006
Xenopus tropicalis Aug 2005
Takifugu rubripes Oct 2004
0.1

Fig. 3. Phylogram depicting the relationship between the deduced amino acid sequence encoded by the novel gene in humans and its homologues in five other species, using Tree View (ver. 1.6.6). The phylogram was based on the alignment of the amino acid sequences (see Supplemental Fig. 7).

In summary, we identified a novel gene in a search of the whole human genome using the powerful new tiling array tool. Although analyzing the tiling array data was no simple matter, it was still useful for detecting a novel gene.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2012.11.076.

Acknowledgments

T. Ichikawa is thanked for excellent secretarial assistance. R. Sato and F. Sato (SASA Plus Co., Ltd., Fukuoka, Japan) are thanked for the assistant of data management and the development of the specific tool for sequence analysis. This work was supported by grants from Grants-in-Aid for Scientific Research.

References

Bertone, P., et al., 2004. Global identification of human transcribed sequences with genome tiling arrays. Science 306, 2242–2246.

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185–193.

Castellano, L., et al., 2009. The estrogen receptor-alpha-induced microRNA signature regulates itself and its transcriptional response. Proc. Natl. Acad. Sci. U. S. A. 106, 15732–15737.

Cheng, J., et al., 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308, 1149–1154.

Eisenberg, E., Levanon, E.Y., 2003. Human housekeeping genes are compact. Trends Genet. 19, 362–365.

Heintzman, N.D., et al., 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459, 108–112.

Hussain, M., et al., 2009. Tobacco smoke induces polycomb-mediated repression of Dickkopf-1 in lung cancer cells. Cancer Res. 69, 3570–3578.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

Johnson, J.M., Edwards, S., Shoemaker, D., Schadt, E.E., 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet. 21, 93–102.

Jones, P.A., Baylin, S.B., 2007. The epigenomics of cancer. Cell 128, 683–692.

Kampa, D., et al., 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. 14, 331–342.

Kapranov, P., et al., 2002. Large-scale transcriptional activity in chromosomes 21 and 22. Science 296, 916–919.

Kapranov, P., et al., 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res. 15, 987–997.

Laird, P.W., 2003. The power and the promise of DNA methylation markers. Nat. Rev. Cancer 3, 253–266.

Lander, E.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, L.C., Dahiya, R., 2002. MethPrimer: designing primers for methylation PCRs. Bioinformatics 18, 1427–1431.

Mockler, T.C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S.E., Ecker, J.R., 2005. Applications of DNA tiling arrays for whole-genome analysis. Genomics 85, 1–15.

Nelson, C.M., et al., 2008. Whole genome transcription profiling of *Anaplasma phagocytophilum* in human and tick host cells by tiling array analysis. BMC Genomics 9, 364.

Nicol, J.W., Helt, G.A., Blanchard Jr., S.G., Raja, A., Loraine, A.E., 2009. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 25, 2730–2731.

Royce, T.E., et al., 2005. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. Trends Genet. 21, 466–475.

Sasaki, D., Kondo, S., Maeda, N., Gingeras, T.R., Hasegawa, Y., Hayashizaki, Y., 2007. Characteristics of oligonucleotide tiling arrays measured by hybridizing full-length cDNA clones: causes of signal variation and false positive signals. Genomics 89, 541–551.

Schadt, E.E., et al., 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol. 5, R73.

Shoemaker, D.D., et al., 2001. Experimental annotation of the human genome using microarray technology. Nature 409, 922–927.

Takai, D., Jones, P.A., 2003. The CpG island searcher: a new WWW resource. Silico Biol, 3, pp. 235–240.

Takata, R., Akamatsu, S., Kubo, M., Takahashi, A., Hosono, N., Kawaguchi, T., Tsunoda, T., Inazawa, J., Kamatani, N., Ogawa, O., Fujioka, T., Nakamura, Y., Nakagawa, H., 2010. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. Nat. Genet. 42, 751–754.

Ting, A.H., McGarvey, K.M., Baylin, S.B., 2006. The cancer epigenome—components and functional correlates. Genes Dev. 20, 3215–3231.

Vanaja, D.K., et al., 2009. Hypermethylation of genes for diagnosis and risk stratification of prostate cancer. Cancer Invest. 27, 549–560.

Venter, J.C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Weile, C., Gardner, P.P., Hedegaard, M.M., Vinther, J., 2007. Use of tiling array data and RNA secondary structure predictions to identify noncoding RNA genes. BMC Genomics 8, 244.

Xu, J., et al., 2005. A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. Am. J. Hum. Genet. 77, 219–229.

Yeager, M., et al., 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat. Genet. 39, 645–649.

Yeager, M., et al., 2009. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. Nat. Genet. 41, 1055–1057.

Proteomic Pattern Analysis Discriminates Among Multiple Sclerosis–Related Disorders

Mika Komori, MD,[1] Yumiko Matsuyama, PhD,[2] Takashi Nirasawa, PhD,[2] Herbert Thiele, PhD,[3] Michael Becker, PhD,[3] Theodore Alexandrov, PhD,[4] Takahiko Saida, MD, PhD,[5] Masami Tanaka, MD, PhD,[5] Hidenori Matsuo, MD, PhD,[6] Hidekazu Tomimoto, MD, PhD,[1] Ryosuke Takahashi, MD, PhD,[1] Kei Tashiro, MD, PhD,[7] Masaya Ikegawa, MD, PhD,[7] and Takayuki Kondo, MD, PhD[8]

Objective: To use a new, unbiased biomarker discovery strategy to obtain and assess proteomic data from cerebrospinal fluid (CSF) of patients with multiple sclerosis (MS)-related disorders.
Methods: CSF protein profiles were analyzed from 107 patients with either MS-related disorders (including relapsing remitting MS [RRMS], primary progressive MS [PPMS], anti-aquaporin4 antibody seropositive–neuromyelitis optica spectrum disorder [SP-NMOSD], and seronegative-NMOSD with long cord lesions on spinal magnetic resonance imaging [SN-NMOSD]), amyotrophic lateral sclerosis (ALS), or other inflammatory neurological diseases (used as controls). CSF peptides/proteins were purified with magnetic beads, and directly measured by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. The obtained spectra were analyzed with multivariate statistics and pattern matching algorithms. These analyses were replicated in an independent sample set of 84 patients composed of those with MS-related disorders or with other neurological diseases (the second cohort).
Results: MS-related disorders differed considerably in terms of CSF protein profiles. SP-NMOSD and SN-NMOSD, both of which fit within the NMO spectrum, were distinguishable from RRMS with high cross-validation accuracy on a support vector machine classifier, especially in relapse phases. Some peaks derived from samples of relapsed SP-NMOSD can discriminate RRMS with high area under curve scores (>0.95) and this was reproduced on the second cohort. The similarity of proteomic patterns between selected neurological diseases were demonstrated by pattern matching analysis. To our surprise, the spectral differences between RRMS and PPMS were much larger than those of PPMS and ALS.
Interpretation: Our findings suggest that CSF proteomic pattern analysis can increase the accuracy of disease diagnosis of MS-related disorders and will aid physicians in appropriate therapeutic decision-making.

ANN NEUROL 2012;71:614–623

Multiple sclerosis (MS)-related disorders are inflammatory diseases of the central nervous system (CNS), and are characterized by different degrees of autoimmune involvement and neurodegeneration.[1] Categories of MS-related disorders include relapsing-remitting MS (RRMS), secondary progressive MS (SPMS), primary progressive MS (PPMS), progressive relapsing MS (PRMS), Balo's concentric sclerosis, and neuromyelitis optica (NMO). It is crucial to differentially diagnose these disorders in order to select the appropriate treatment course that will benefit the patient. Since effective therapy has only been established for RRMS, it is

View this article online at wileyonlinelibrary.com. DOI: 10.1002/ana.22633

Received Oct 16, 2010, and in revised form Aug 29, 2011. Accepted for publication Sep 2, 2011.

Address correspondence to Dr Masaya Ikegawa, Department of Genomic Medical Sciences, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kajii-cho 465 Kawaramachi Hirokoji, Kamigyo-ku, Kyoto, Japan, 602-8566. E-mail: mikegawa@koto.kup-m.ac.jp

From the [1]Department of Neurology, Graduate School of Medicine, Kyoto University, Kyoto, Japan; [2]Bruker Daltonics K.K., Yokohama, Japan; [3]Bruker Daltonik GmbH, Bremen, Germany; [4]Center for Industrial Mathematics, University of Bremen, Bremen, Germany; [5]Department of Neurology, Utano National Hospital, Kyoto, Japan; [6]Division of Clinical Research, Nagasaki Medical Center of Neurology, Nagasaki, Japan; [7]Department of Genomic Medical Sciences, Graduate School of Medicine, Kyoto Prefectural University of Medicine, Kyoto, Japan; and [8]The Tazuke Kofukai Medical Research Institute, Kitano Hospital, Osaka, Japan.

Additional Supporting Information can be found in the online version of this article.

necessary that we gain a comprehensive understanding of how RRMS pathogenesis is similar to the other MS-related disorders so that similarly efficacious treatments may be developed. It is particularly important to differentiate RRMS and NMO, given their largely overlapping clinical characteristics, their particular prevalence in East Asia, and because the optimal treatments for the diseases differ considerably.[2,3] The current Mayo NMO diagnostic criteria[4] requires clinical episodes of both optic neuritis and myelitis to definitively identify NMO. Anti-aquaporin-4 (AQP4) antibody was discovered as a biomarker of NMO[5,6]; however, this designation has evoked some controversy. First, the clinical spectrum of disorders defined by the presence of anti-AQP4 antibody[2] encompasses recurrent optic neuritis or myelitis alone and Asian "optic-spinal MS" with long cord lesion (LCL) on spinal magnetic resonance imaging (MRI),[2,7] which are not included in the 2006 Mayo NMO criteria.[4] NMO-specific brain lesions have also been classified.[8] Second, although LCL is 1 of the most characteristic features of NMO, a considerable number of subjects with LCL present as seronegative for the anti-AQP4 antibody. Even for those patients fulfilling the 2006 Mayo NMO criteria, 24% to 67% have been reported as being seronegative for the anti-AQP4 antibody.[4,5,9,10] Although the term "NMO spectrum disorder (NMOSD)" has been used for these disorders, it is not clear whether these seronegative subjects are the result of inadequate clinical diagnostic criteria, suboptimal assay sensitivity, or different targeted antigens. Moreover, it remains unknown whether seropositive and seronegative NMO patients have shared or distinct pathogenesis.

PPMS carries a poor prognosis, and no successful therapeutic trials have been accomplished to date.[11] The notion that RRMS and PPMS can be treated as a single disorder is unproven, although sufficiently large and lengthy research studies are underway.[12] It is necessary to know whether the 2 forms of MS have distinct pathogeneses as they may require different therapeutic strategies. It also remains uncertain, although it has been speculated, that the progressive form of MS undergoes more significant neurodegeneration than does the relapsing form of MS.[13]

The cerebrospinal fluid (CSF) is considered by many to be a window into the brain through which one might identify promising new biomarkers of neurodegenerative disorders. Indeed, several biomarkers for MS-related disorders have been reported from studies on CSF.[14–22] Recently, CSF glial fibrillary acidic protein (CSF-GFAP) and S100B were proposed as promising biomarker candidates of NMO attacks.[23] CSF-GFAP was found to be elevated more than 1,000-fold in NMO patients, as compared to control cases, and the increased concentrations were determined to be the result of astrocyte destruction. However, complex traits of MS-related disorders defy strict association with a single biological process. CSF-GFAP elevation can result from multiple diseases in which astrocyte destruction is induced.[23]

The advent of mass spectrometry technologies have made it possible to uncover distinct molecular components associated with particular disease states.[20,24,25] However, as illustrated by the CSF-GFAP example above, similar changes of single components can be induced by multiple mechanisms. Hence, it may not be realistic to expect to find a single biomarker for complex disease processes that involve multiple underlying molecular mechanisms in their pathogenesis. Proteomic pattern analysis, a new method to search for biomarkers, is suitable for this purpose as it examines a panel of molecules; moreover, this approach can effectively distinguish seemingly closely-related diseases of a complex nature, such as MS-related disorders.

In this study, we analyzed CSF proteomic patterns from MS-related and non-MS control diseases by using magnetic bead-based enrichment of CSF peptides and proteins followed by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry.[26] The current study reveals distinct CSF proteomic patterns between MS and anti-AQP4 antibody-defined disorder, confirmed by the 2 separated analysis with the same protocol but for different patient sets (which we have designated the first cohort and the second cohort). While the disorder characterized by LCL was found to have a proteomic pattern similar to the anti-AQP4 antibody-defined disorder in the first cohort, the result of the second cohort was inconsistent, indicating that this disorder have more than 1 proteomic pattern, and that the proteomic pattern is useful in classifying the disorders with LCL. We also succeeded in evaluating and visualizing the similarity of the proteomic pattern between neurological diseases analyzed.

Patients and Methods

Study Design

Patients were enrolled on the basis of clinical and laboratory features consistent with MS-related disorders. RRMS and PPMS subjects were diagnosed based on the revised McDonald criteria from 2005 (Polman and colleagues[27]). Patients seropositive for anti-AQP4 antibody were defined as seropositive NMOSD (SP-NMOSD), regardless of the distribution of their lesion. Seronegative NMOSD (SN-NMOSD) was applied to those who were seronegative for anti-AQP4 antibody and fit the McDonald MS criteria,[27] but exhibited LCL on spinal MRI.

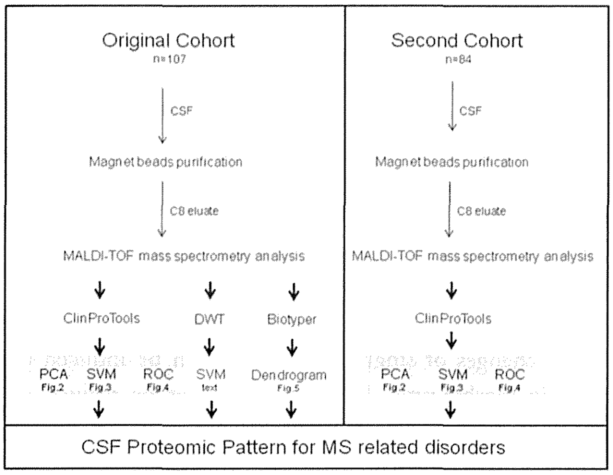

Original Cohort
n=107
CSF
Magnet beads purification
C8 eluate
MALDI-TOF mass spectrometry analysis
ClinProTools
DWT
Biotyper
PCA
Fig.2
SVM
Fig.3
ROC
Fig.4
SVM
text
Dendrogram
Fig.5
Second Cohort
n=84
CSF
Magnet beads purification
C8 eluate
MALDI-TOF mass spectrometry analysis
ClinProTools
PCA
Fig.2
SVM
Fig.3
ROC
Fig.4
CSF Proteomic Pattern for MS related disorders

FIGURE 1: General workflow. This study was composed of 2 cohorts: the original one (n = 107) and the second one (n = 84). C8 = reversed phase column; DWT = discrete wavelet transformation; MALDI-TOF = matrix assisted laser desorption ionization–time of flight; PCA = principal component analysis, ROC = receiver operating characteristic curve; SVM = support vector machine.

Anti-AQP4 antibody was assayed in all patients with MS-related disorders. The assay was performed in a blinded manner (no patient information) by the standard method.[28] Patients with RRMS, SP-NMOSD, and SN-NMOSD were further divided among 2 clinical phases: relapse and remission. CSF samples from patients in the relapse phase were obtained prior to any treatment being administered to counter acute worsening, such as high-dose intravenous corticosteroid injection. The amyotrophic lateral sclerosis (ALS) group was considered as the control group and represented other degenerative neurological disease. The other inflammatory neurological disease (OIND) group was composed of patients suffering from aseptic encephalomeningitis (AEM), Guillain-Barr syndrome (GBS), and chronic inflammatory demyelinating polyneuropathy (CIDP). The second cohort study was also performed with the same protocol and it included RRMS, SP-NMOSD, SN-NMOSD, and OIND composed of AEM, GBS and CIDP.

The study protocol was approved by our ethics committee prior to enrolling patients, and all subjects provided written informed consent. The general workflow is illustrated in Figure 1.

Preparation of CSF samples for ClinProt Analysis

CSF samples were sent for routine diagnostics, including quantification of total protein, calculation of immunoglobulin G (IgG) index based on serum albumin and IgG concentration, and presence of oligoclonal IgG bands (OCB) as detected by isoelectric focusing and immunofixation. No sample contained more than 500 erythrocytes per microliter. All samples were centrifuged for 10 minutes at 3,000 rpm to separate the cellular elements for removal before storage at −80°C. Samples were prepared for analysis immediately upon thawing. A 5μl aliquot of the CSF was purified using magnetic beads with functionalized surface (hydrophobic interaction C8, MB-HIC 8; Bruker Daltonik GmbH, Bremen, Germany), according to the manufacturer's protocol. For mass spectrometric analysis, 1μl of the bead elute was mixed with 10μl of matrix solution (0.6g/liter a-cyano-4-hydroxycinnamic acid in 2:1 ethanol/acetone); 1μl of the mixture was then spotted in quadruplicate on a MALDI target MTP AnchorChip 600/384 (Bruker Daltonik GmbH).

Mass Spectrometry

Samples applied to the chip were analyzed on an Autoflex II MALDI-TOF mass spectrometer, operating in positive-ion linear mode (Bruker Daltonik GmbH). To generate a spectrum, 1,000 laser shots were acquired from random positions for each matrix spot. Four independent spectra were acquired for each spot. Acquisition was controlled by flexControl 3.0 software (Bruker Daltonik GmbH), using the AutoXecute tool and fuzzy control of laser intensity. The mass range analyzed was 1,000 to 15,000 mass to charge ratio (m/z). Spectra were externally calibrated using a mixture of standardized protein/peptide calibrants (ClinProt Standard, Bruker Daltonik GmbH).

Analysis of Proteomic Profile Spectra

The resulting spectra were analyzed using ClinProTools 2.2 bioinformatic software (Bruker Daltonik GmbH); the process included intensity normalization and spectral alignment using prominent internal peaks. As ClinProTools allows for discovery of discriminative peaks of spectra and can estimate how discriminative they are, we used the program to generate estimates of the respective potential for accurate diagnosis for each peak. Concomitant measures of specificity and sensitivity were also calculated. Peaks of interest were selected from the total average spectra, using a signal-to-noise threshold of 5.0. Finally, the ClinProTools was used to carry out comparative analysis of peak intensities between groups/disease classes, and to calculate corresponding statistics. When comparing 2 groups, we used analysis of variance (ANOVA) or the Wilcoxon-Mann-Whitney test. A cross-validation was performed on the same data by randomly assigning a group number to each CSF sample and then repeating the Wilcoxon-Mann-Whitney test.

Multivariate statistical analysis techniques, including principal component analysis (PCA)[29] and the support vector machine (SVM) algorithm (from ClinProTools), were employed to extract, display, and rank the variance within each data set. Through the calculation process of principal components (PCs), different weightings were assigned to each variable based on their contribution to the explained variance of a PC; in this manner, the contribution of single peaks to the variance covered by the respective PC was determined. To confirm the accuracy of SVM analysis, discrete wavelet transformation combined with SVM (DWT)[30] was employed.

We performed leave-1-out cross-validation (LOOCV) experiments using the SVM algorithm. In these analyses, different combinations of peptides selected by the Mann-Whitney U-test at different adjusted p value cutoffs were used to build the models and find significant peaks. The best models, ie, the

TABLE 1: Clinical and Laboratory Values for the First Cohort

	RRMS		SP-NMOSD		SN-NMOSD		PPMS	ALS	OIND
Clinical phase	Relapse	Remission	Relapse	Remission	Relapse	Remission			
Total patients	12	17	11	11	6	6	12	17	15
Male/female	6/6	6/11	3/8	0/11	0/6	0/6	4/8	13/4	11/4
Mean age at sampling, yr (range)	30 (17–40)	37 (16–57)	54 (33–64)	48 (17–81)	41 (37–60)	50 (38–58)	42 (32–51)	67 (44–72)	46 (15–62)
Disease duration, yr (range)	1.3 (0.1–17.3)	4.0 (0.3–27.3)	4.7 (0.3–8.1)	2.3 (0.2–10.0)	7.1 (1.4–10.0)	4.1 (0.5–19.0)	6.0 (0.5–20.0)	1.3 (0.1–10.0)	0.1 (0.0–1.3)
Average Expanded Disability Status Scale of Kurtzke (range)	2.3 (0.0–6.0)	1.0 (0.0–6.5)	7.5 (3.0–9.0)	5.0 (2.0–8.5)	8.0 (6.0–9.0)	8.3 (1.0–9.0)	4.5 (2.5–9.0)	n.e.	n.e.
Optic neuritis/total patients	5/12	4/17	9/11	8/11	5/6	4/6	0/12	n.e.	n.e.
Spinal MRI evidence of long cord lesions/total patients	0/12	0/17	11/11	9/11	6/6	6/6	0/12	n.e.	n.e.
Serum anti-AQP4 antibody positive/total patients	0/12	0/17	11/11	11/11	0/6	0/6	0/12	n.e.	n.e.
Fulfilling NMO criteria/total patients	0/12	0/17	9/11	7/11	5/6	4/6	0/12	n.e.	n.e.
CSF protein concentration, mg/dl (range)	36.0 (21.0–45.0)	30.0 (16.0–47.8)	60.0 (33.0–170.0)	30.1 (16.0–61.0)	65.1 (45.0–85.0)	33.0 (21.0–65.0)	39.5 (27.0–63.0)	35.8 (28.1–68.1)	54.8 (20.0–534.0)
IgG index, mg/dl (range)	0.67 (0.44–1.51)	0.61 (0.44–1.03)	0.62 (0.43–0.82)	0.53 (0.42–0.69)	0.57 (0.45–0.77)	0.49 (0.44–0.59)	0.87 (0.44–2.49)	0.49 (0.40–0.65)	0.65 (0.41–0.78)
Oligoclonal IgG bands/total patients	7/11	8/15	0/9	0/6	0/3	1/5	7/10	n.e.	n.e.

ALS = amyotrophic lateral sclerosis; AQP4 = aquaporin-4; CSF = cerebrospinal fluid; IgG = immunoglobulin G; MRI = magnetic resonance imaging; MS = multiple sclerosis; n.e. =; NMO = neuromyelitis optica; NMOSD = neuromyelitis optica spectrum disorder; OIND = other inflammatory neurological disease; PPMS = primary progressive MS; RRMS = relapsing-remitting MS; SN-NMOSD = anti-aquaporin4 antibody seronegative NMOSDO; SP-NMOSD = anti-aquaporin4 antibody seropositive NMOSD.

TABLE 2: Clinical and Laboratory Values for the Second Cohort

	RRMS		SP-NMOSD		SN-NMOSD		PPMS	ALS	OIND
Clinical phase	Relapse	Remission	Relapse	Remission	Relapse	Remission			
Total patients	15	15	4	7	2	5	2	18	16
Male/female	10/5	5/10	0/4	2/5	0/2	2/3	2/0	10/8	12/4
Mean age at sampling, yr (range)	39 (19–54)	43 (29–68)	47 (44–52)	48 (30–62)	48 (35–60)	48 (22–66)	58 (51–64)	62 (42–78)	48 (22–74)
Disease duration, yr (range)	4.4 (0.1–20.0)	5.3 (0.2–24.0)	2.3 (0.1–7.0)	1.2 (0.2–6.3)	3.3 (0.3–6.3)	7.1 (2.0–18.2)	7.9 (3.8–12.0)	2.2 (0.3–6.0)	0.9 (0.0–6.0)
Average Expanded Disability Status Scale of Kurtzke (range)	3.3 (2.0–6.5)	4.3 (1.0–7.5)	7.7 (7.5–8.0)	4.1 (1.0–7.0)	7.8 (7.5–8.0)	6.3 (6.0–6.5)	6.3 (5.5–7.0)	n.e.	n.e.
Optic neuritis/total patients	3/15	3/15	2/4	5/7	1/2	2/5	1/2	n.e.	n.e.
Spinal MRI evidence of long cord lesions/total patients	0/15	0/15	4/4	4/7	2/2	5/5	0/2	n.e.	n.e.
Serum anti-AQP4 antibody positive/total patients	0/15	0/15	4/4	7/7	0/2	0/5	0/2	n.e.	n.e.
CSF protein concentration, mg/dl (range)	36.6 (23.0–76.1)	36.4 (22.0–51.0)	86.7 (52.0–126.0)	39.7 (27.8–72.0)	58.0 (49.0–67.0)	44.0 (25.0–52.0)	29.8 (28.5–31.0)	41.7 (26.0–90.5)	81.5 (20.0–347.0)
IgG index, mg/dl (range)	0.89 (0.43–2.12)	0.54 (0.37–0.94)	0.70 (0.43–0.71)	0.47 (0.42–0.51)	0.55 (0.46–0.62)	0.52 (0.47–0.62)	0.59 (0.40–0.79)	0.51 (0.35–1.01)	0.56 (0.46–0.64)
Oligoclonal IgG bands/total patients	8/14	4/12	0/4	0/7	0/2	0/4	1/2	n.e.	n.e.

ALS = amyotrophic lateral sclerosis; AQP4 = aquaporin-4; CSF = cerebrospinal fluid; IgG = immunoglobulin G; MRI = magnetic resonance imaging; MS = multiple sclerosis; n.e. =; NMOSD = neuromyelitis optica spectrum disorder; OIND = other inflammatory neurological disease; PPMS = primary progressive MS; RRMS = relapsing-remitting MS; SN-NMOSD = anti-aquaporin4 antibody seronegative NMOSDO; SP-NMOSD = anti-aquaporin4 antibody seropositive NMOSD.

ones giving the smaller classification error rate in the cross-validation of the first cohort, were tested. We then examined an independent collection of 84 CSF samples as a second cohort and replicated the above mentioned calculation as shown in Figure 1.

Pattern Matching of the CSF Proteomic Spectra

We applied the MALDI Biotyper algorithm (Bruker Daltonik GmbH)[31,32] to the spectra obtained from the CSF samples, according to the manufacturer's protocols. For phylogenetic analysis, we hierarchically clustered mass spectra corresponding to each disease group. For graphical correlations, an average statistical algorithm was implemented in the MALDI Biotyper software. Reference spectra were analyzed and compared for the nine disease stages (ALS, PPMS, RRMS remission, RRMS relapse, OIND, SP-NMOSD remission, SP-NMOSD relapse, SN-NMOSD remission, and SN-NMOSD relapse). Based on the distance values obtained, a list of mass signals and their intensities was taken into consideration, and a dendrogram was produced by a similar scoring method using a set of mass spectra to determine distance values between disease groups. According to previous analogous bacterial identification experiments for group-by-group comparisons,[31] distance levels <500 were considered to indicate "reliable similar classification." We applied this standard of distance values to our evaluation of the similarity of CSF protein patterns.

Results

Subject and CSF Characteristics of the First and the Second Cohort

We analyzed 107 CSF samples from patients diagnosed with RRMS (n = 29), SP-NMOSD (n = 22), SN-NMOSD (n = 12), PPMS (n = 12), ALS (n = 17), and OIND (n = 15). The OIND group was composed of patients suffering from AEM (n = 7), GBS (n = 4), and CIDP (n = 4). Clinical information, including routine CSF and MRI findings, are summarized in Table 1. Of the 22 SP-NMOSD patients, 16 (72.7%) fulfilled the 2006 Mayo NMO criteria,[4] as did 9 of the 12 (75.0%) SN-NMOSD patients. The remaining 6 individuals with SP-NMOSD and 3 with SN-NMOSD lacked detectable optic nerve lesions or spinal lesions.

We put 84 CSF samples from patients diagnosed with RRMS (n = 30), SP-NMOSD (n = 11), SN-NMOSD (n = 7), PPMS (n = 2), ALS (n = 18), and OIND (n = 16) as the second cohort. OIND includes 8 CIDP, 4 GBS, and 4 AEM. Clinical information of these patients is summarized in Table 2.

Discrimination Among MS-Related Disorders by PCA

In the mass range analyzed (m/z 1,000–15,000), an average of 108 peaks per subset spectrum were detected at a signal-to-noise threshold of 5.0 (Supplementary Figure 1). The loading coefficients were found to indicate that more than 50 peaks predominantly contributed to the separation of RRMS relapse and SP-NMOSD relapse groups (Supplementary Figure 2).

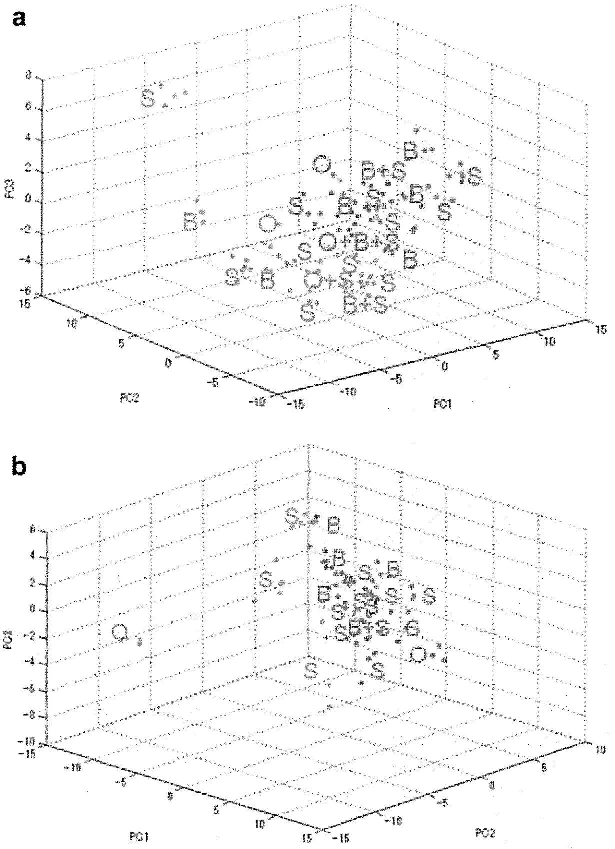
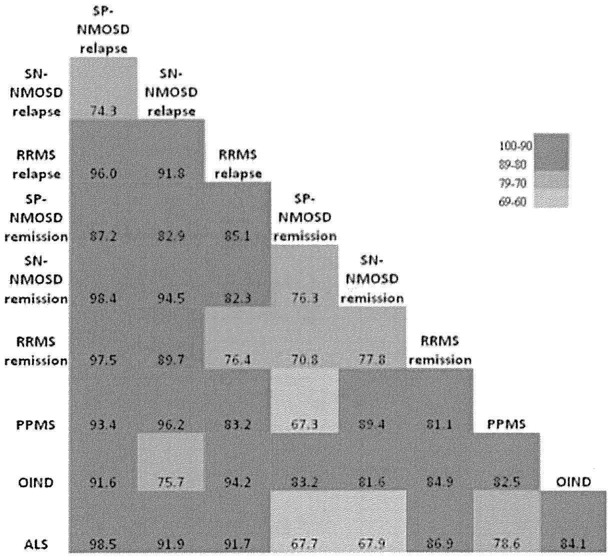


FIGURE 2: Representation of the principal components (PCs) generation from a data set. For each CSF sample, 4 measurements were automatically performed and represented a combination of clinically manifested CNS lesion sites. O = optic nerve; S = spinal cord; B = brain. (A) Principal component analysis (PCA) score plot indicating that RRMS relapse patients (*red*) clustered separately from SP-NMOSD relapse patients (*green*) based on the first cohort. (B) PCA score plot indicating that RRMS relapse patients (*red*) clustered separately from SP-NMOSD relapse patients (*green*) based on the second cohort.

Comparative analysis of spectra between disease groups revealed a number of peaks with significant differences in intensity. PCA showed that discrimination between each disease category was distinct. For example, RRMS relapse and SP-NMOSD relapse was clearly discriminated by PCA (Fig 2A). This result is also replicated by the second cohort (see Fig 2B). Furthermore, distribution of CNS lesions do not have significant power on proteomic pattern discrimination.

One can argue a possibility that difference in the sex ratio, age, disease severity, presence or absence of OCB, or distribution of CNS lesions have significant power on proteomic patterns. To address this question, the PCA analysis was carried. The results statistically and



	SP-NMOSD relapse	SN-NMOSD relapse	RRMS relapse	SP-NMOSD remission	SN-NMOSD remission	RRMS remission	PPMS	OIND
SN-NMOSD relapse	74.3							
RRMS relapse	96.0	91.8						
SP-NMOSD remission	87.2	82.9	85.1					
SN-NMOSD remission	98.4	94.5	82.3	76.3				
RRMS remission	97.5	89.7	76.4	70.8	77.8			
PPMS	93.4	96.2	83.2	67.3	89.4	81.1		
OIND	91.6	75.7	94.2	83.2	81.6	84.9	82.5	
ALS	98.5	91.9	91.7	67.7	67.9	86.9	78.6	84.1

FIGURE 3: Heat map of the SVM analysis to differentiate patients with SP-NMOSD, SN-NMOSD, RRMS, PPMS, OIND, and ALS. Cross-validation analysis provided an estimate of the success rate for the SVM model to separate user-defined groups of spectra. RRMS and SP-NMOSD were well-separated by SVM. The distinct proteomic profiles of RRMS and SP-NMOSD were confirmed in relapse phase and, to a lesser extent, in the remission phase. The variation of proteomic profiles between relapse and remission phases was robust for SP-NMOSD, but less so for RRMS. The *x*-axis and *y*-axis represent the samples, ordered by group. The accuracy is denoted by color: 90% to 100% (*red*); 80% to 89% (*pink*); 70% to 79% (*purple*); and 60% to 69% (*blue*) (see inset key).

reproducibly showed that each of the above differences had little impact on distinction of proteomic patterns of relevant diseases (Supplementary Figure 3A and B).

Discrimination Among MS-Related Disorders by SVM

To automatically detect differences of the obtained spectra at different disease stages, we applied a supervised model generation procedure in combination with SVM, an approach based on machine learning. Cross-validation analysis provided an estimate of the success rate for the SVM model to separate user-defined groups of spectra (Fig 3). For example, SVM analysis between RRMS remission and OIND resulted in cross-validation accuracy of 84.9%; between RRMS relapse and OIND, accuracy was determined to be 94.2%. The RRMS and SP-NMOSD groups were well separated by SVM (96.0%, RRMS relapse vs SP-NMOSD relapse; 70.8%, RRMS remission vs SP-NMOSD remission). Therefore, this method was able to confirm that patients with RRMS and SP-NMOSD exhibited distinct proteomic profiles in the relapse phase, and to a lesser extent in the remission phase. The change of proteomic profiles that occurred between relapse and remission phases was found to be more prominent in SP-NMOSD, and less so in RRMS. SVM cross-validation accuracies were 76.4% for RRMS and 87.2% for SP-NMOSD. SVM analysis between RRMS remission and PPMS yielded an accuracy rate of 81.1%; moreover, SVM between RRMS relapse and PPMS had an accuracy of 83.2%. RRMS and SN-NMOSD showed distinct features (91.8%, RRMS relapse vs SN-NMOSD relapse; 77.8%, RRMS remission vs SN-NMOSD remission). SVM analyses between SP-NMOSD and SN-NMOSD were found to have an

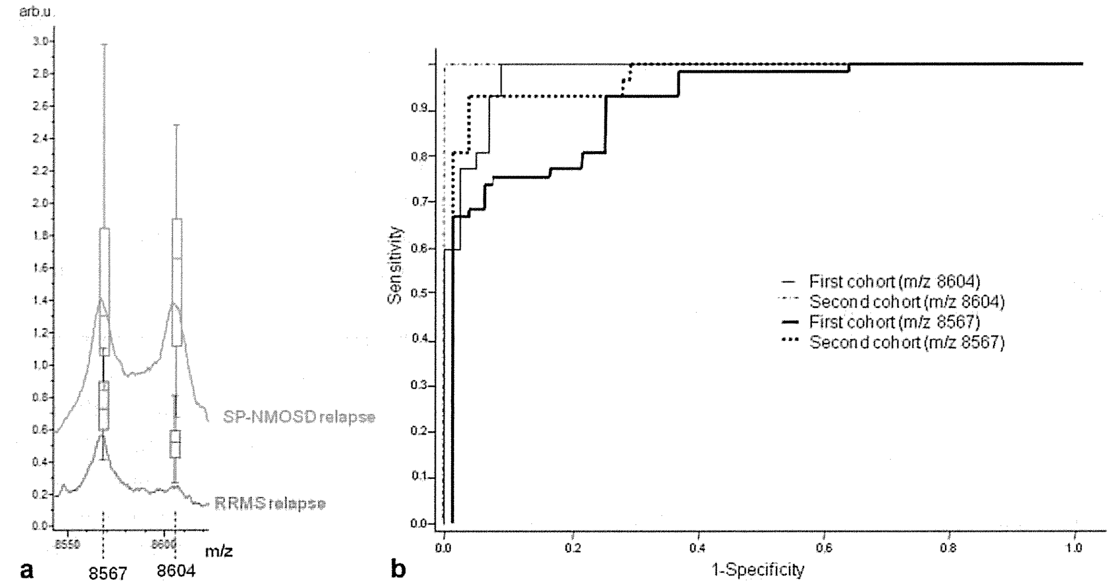


FIGURE 4: Analysis by receiver operating characteristic curves. (A) Average spectral features and box-and-whisker plots at m/z 8567 and 8604, representative markers for RRMS vs SP-NMOSD in relapse phase. SP-NMOSD relapse (*green*); RRMS relapse (*red*). The *x*-axis and *y*-axis represent the relative intensity and m/z, respectively. (B) Changes in key markers of RRMS relapse and SP-NMOSD relapse were validated by ROC curve of m/z 8567 and 8604.

TABLE 3: AUC Determined by ROC Analysis for Each Peak Used in the Cluster to Differentiate RRMS and SP-NMOSD/SN-NMOSD Relapse Phases

m/z	RRMS relapse vs SP-NMOSD relapse[a]	RRMS relapse vs SN-NMOSD relapse[a]
8604	0.996/1	0.991/0.810
6970	0.980/0.884	0.990/0719
4644	0.970/0.617	0.979/0.559
8567	0.957/0.979	0.976/0.734
7033	0.950/0.810	0.956/0.542

[a]Values are given as first cohort/second cohort.AUC = area under the ROC curve; m/z = mass to charge ratio; NMOSD = neuromyelitis optica spectrum disorder; ROC = receiver operating characteristic; RRMS = relapsing remitting multiple sclerosis; SN-NMOSD = anti-aquaporin4 antibody seronegative NMOSD; SP-NMOSD = anti-aquaporin4 antibody seropositive NMOSD.

accuracy of 74.3% in relapse and 76.3% in remission. DWT emphasized the distinction between RRMS and SP-NMOSD in relapse stages, with 86.5% double cross-validation accuracy. These observations are well validated by the second cohort as shown in Supplementary Figure 4.

Analysis by Receiver Operating Characteristics Curves

The sensitivity and specificity of each peak was calculated by the receiver operating characteristics (ROC) analysis (Fig 4A, B; Table 3). Each of the peaks showed high accuracy in discriminating RRMS from SP-NMOSD in relapse phases, by using the area under the ROC curve (AUC). The areas of these peaks in the spectra were found to be statistically different. Moreover, 5 of these peaks had AUC > 0.95 when RRMS relapse was compared with SP-NMOSD relapse. This analysis was even better replicated by the second cohort (see Fig 4B). In Table.3, while discrimination between RRMS relapse and the SP-NMOSD relapse phase was well validated by the first and second cohort, discrimination between RRMS relapse and SN-NMOSD was not replicated. In Supplementary Figure 5A and 5B, this situation is visualized by 2D distribution view of 2 selected peaks, x = 8604 and y = 8567 in the spectrum of loaded model generation classes.

"Pattern Matching" Spectral Differences Among Reference Disease Groups

Using the MALDI Biotyper 2.0 software, 9 reference disease groups were analyzed and their similarities were visualized in a dendrogram (Fig 5). The CSF proteomic patterns of disease were clearly divided into 3 main groups with distance values >500. One group was composed of RRMS, PPMS, and ALS, while a second group included SP-NMOSD, SN-NMOSD, and OIND. Except for the comparison between SP-NMOSD and SN-NMOSD, the distance values of proteomic patterns between relapse and remission phases for the same disorder were shorter than that in comparison to different disease groups. The distance value of proteomic patterns between the remission phases of SP-NMOSD and SN-NMOSD was 88, but the distance value increased to 446 for the relapse phases; this finding indicated that the

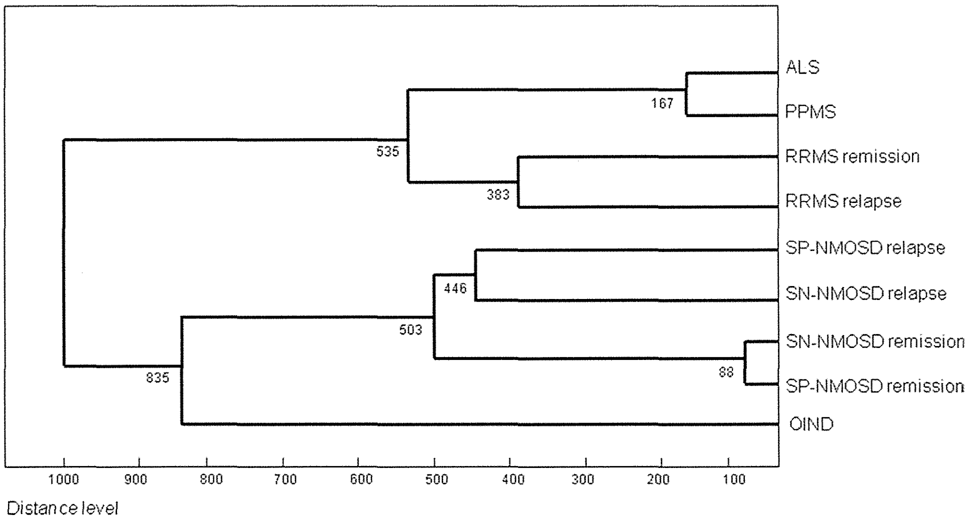


FIGURE 5: Score-oriented dendrogram of MALDI-TOF mass spectral profiles generated by the MALDI Biotyper. Reference spectra were generated for each of 9 different disease groups. Similarity was visualized by a rooted tree. The distance level is presented as percentage.

patterns still resembled each other at relapse, but the differences were clearer than in remission stages. Surprisingly, PPMS, RRMS, and ALS classified with the same group. The distance values of proteomic patterns between PPMS and ALS was 167, and distinctly shorter than that of PPMS and any of the other inflammatory diseases, including SP-NMOSD and SN-NMOSD.

Discussion

In this study, we demonstrated that CSF proteomic patterns can effectively discriminate MS from other MS-related disorders. Proteomic profiles of CSF from the relapse phase of SP-NMOSD and SN-NMOSD are distinct from those of RRMS. As there is a strong need to develop therapeutic guidelines specific to each of the MS-related disorders, detection of process-specific biomarkers represents an important new direction toward this end. As we have shown here, CSF proteomic pattern analysis could afford clinicians the possibility to make clear distinctions among MS-related disorders that are much less influenced by the size and distribution of disease lesions. CSF sampling without trypsin digestion presents an advantage in analyzing the native CSF proteome, thus potentially allowing for the direct measurement of enzymatically cleaved proteins of pathological relevance.[20]

SVM, a mathematical algorithm based on supervised learning methods, has proven to be a useful tool to detect differences between created models, especially when small datasets are applied. To confirm the accuracy of SVM in this study, in a completely different approach of spectral processing we employed DWT, which is a specific kind of Fourier transformation. This method is also considered to be superior in treating relatively small number of samples. In the relapse phase of RRMS and SP-NMOSD, discrimination of each disease by CSF proteomic profiling was much easier to accomplish than in the remission phase, indicating that dynamic autodestructive processes may be reflected in the CSF proteomic profiles.

SN-NMOSD and SP-NMOSD have only recently be recognized as components of the NMO-spectrum.[2] However, controversy exists as to where SN-NMOSD should be classified between RRMS and SP-NMOSD. Hence, the rational selection of therapy for SN-NMOSD remains unclear. In the first cohort, 5 common discriminative mass spectra peaks with high AUC scores (>0.95) between SN-NMOSD and SP-NMOSD were found to be discriminative between the 2 NMO-spectrum disorders and RRMS in relapse phases. We believe from this result that most of SN-NMOSD has similar or identical pathogenesis to SP-NMOSD. However, the 2 samples with SN-NMOSD in the second cohort have a different pattern (see Table 3; Supplementary Figure 5A and B). There is little possibility that the difference was due to a methodological error since the reproducibility was quite robust for the other disorders in analysis. It is more likely that SN-NMOSD have more than 1 population with distinct pathogenesis. Indeed, LCL findings in MRI can admit more than 1 condition. Adjacent spinal lesions in advanced stage of MS can be indistinguishable to a single continuous long lesion in the MRI study. The result may show that proteomic pattern can be a strong tool for clarifying multiple disorders not easy to separate by conventional methods.

The MALDI Biotyper was developed as a mass spectrometry-based platform for identification and classification of microorganisms.[31] The patterns of protein masses observed by MALDI-mass spectrometry have been successfully used for accurate classification and identification of bacteria. In this study, we have applied the MALDI Biotyper software solution for discriminating proteomic patterns in human neurological disease. To our surprise, the resulting dendrogram was composed of 3 major "islands." ALS and PPMS were classified as the disease entities nearest to RRMS, with relapse and remission stages composing the first island. These disease entities were clearly discriminated from SP-NMOSD and SN-NMOSD, with relapse and remission stages composing the second island on the dendrogram. OIND, which included disorders with a prominent CNS inflammatory feature, produced the third island, which was closer to the second island of SP-NMOSD and SN-NMOSD. The finding that PPMS was situated next to ALS may indicate that the CSF proteomic patterns for PPMS and ALS represent less inflammation or merely a small skew from the normal state. However, recent reports have suggested another possibility in that PPMS has more neurodegenerative features than other MS-related disorders.[13,33,34] Alternatively the observation here may support involvement of the immune system in ALS to mediate either neurotoxicity or neuroprotection events.[35] These are open questions that will likely be answered in future studies.

MRI evidence of LCL is currently considered the most characteristic feature of NMO,[4] and is indispensable for NMO diagnosis. However, early therapeutic intervention can prevent the extension of spinal cord lesions. In this situation, tracking of disease progress can be impeded by the relative stabilization of lesions. Our procedure is not influenced by size and distribution of lesions, and will provide more solid information to