nature genetics

Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20

The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene)[1]

To identify multiple sclerosis (MS) susceptibility loci, we conducted a genome-wide association study (GWAS) in 1,618 cases and used shared data for 3,413 controls. We performed replication in an independent set of 2,256 cases and 2,310 controls, for a total of 3,874 cases and 5,723 controls. We identified risk-associated SNPs on chromosome 12q13–14 (rs703842, $P = 5.4 \times 10^{-11}$; rs10876994, $P = 2.7 \times 10^{-10}$; rs12368653, $P = 1.0 \times 10^{-7}$) and upstream of *CD40* on chromosome 20q13 (rs6074022, $P = 1.3 \times 10^{-7}$; rs1569723, $P = 2.9 \times 10^{-7}$). Both loci are also associated with other autoimmune diseases[1–5]. We also replicated several known MS associations (HLA-DR15, $P = 7.0 \times 10^{-184}$; *CD58*, $P = 9.6 \times 10^{-8}$; *EVI5-RPL5*, $P = 2.5 \times 10^{-6}$; *IL2RA*, $P = 7.4 \times 10^{-6}$; *CLEC16A*, $P = 1.1 \times 10^{-4}$; *IL7R*, $P = 1.3 \times 10^{-3}$; *TYK2*, $P = 3.5 \times 10^{-3}$) and observed a statistical interaction between SNPs in *EVI5-RPL5* and HLA-DR15 ($P = 0.001$).

Multiple sclerosis (MS) is a potentially devastating disease of the central nervous system that mainly affects young adults of European ancestry (MIM126200). Predisposition to MS is influenced by genetic and environmental risk factors, and the HLA-DRB1*1501-DQB1*602 (HLA-DR15) haplotype in the major histocompatibility complex (MHC) is the predominant genetic risk factor in Northern Europeans. A genome-wide association study (GWAS) conducted by the International MS Genetics Consortium (IMSGC) in 931 MS trio families identified additional genetic associations with *IL7R* and *IL2RA*[6] and these are now considered unequivocal MS susceptibility genes, as is *CLEC16A* (*KIAA0350*)[7,8]. More recently, *CD226* (ref. 8), *KIF1B* (ref. 9) and *TYK2* (ref. 10) have been identified as genes likely to be involved in MS susceptibility.

To identify additional MS susceptibility loci, we conducted a GWAS in European-ancestry MS cases from Australia and New Zealand (ANZ) (**Supplementary Table 1** online) and used control data from the United Kingdom (UK) and from individuals of European ancestry from the United States (US) (**Table 1**). After application of strict quality control filters, the effective sample size was 1,618 MS cases and 3,413 controls (Online Methods). In total, 310,504 SNPs were common to the case and control datasets and 303,431 SNPs were successfully genotyped in more than 95% of samples. We removed SNPs with a minor allele frequency (MAF) <1% and those in significant Hardy-Weinberg disequilibrium ($P < 1 \times 10^{-7}$) in any one of three datasets (ANZ, UK, US), leaving 302,098 SNPs. Of the remaining autosomal SNPs, 46,827 were identified as having significantly different call rates ($P < 1 \times 10^{-4}$) between the datasets. Missing genotypes were imputed for these SNPs and also for the remaining genotyped SNPs in the GWAS (Online Methods).

We tested each SNP for association with disease susceptibility using the Cochran-Armitage test for trend. After correction for population stratification, there was still a modest overall inflation of test statistics, with a genomic inflation factor (λ) of 1.10 after excluding SNPs in the MHC (**Supplementary Fig. 1a** online). The excess of small P values was largely removed after exclusion of these SNPs. The only region of the genome showing association at the genome-wide level in the GWAS was the MHC on chromosome 6p21.3 (**Supplementary Fig. 1b**). The most strongly associated SNP in the MHC was rs9271366 ($P = 6.9 \times 10^{-89}$), which tags the HLA-DR15 haplotype (correlation = 0.98; see Online Methods). This SNP was later genotyped in the replication phase to enable stratification of subjects on HLA-DR15.

SNPs were prioritized for the replication phase if they ranked in the 500 most associated ($P < 9.2 \times 10^{-4}$) SNPs in the GWAS or met other biologically plausible criteria (Online Methods). We also imputed genotypes probabilistically at all Phase 2 HapMap SNPs, and selected some of these for genotyping in the replication phase. The replication sample consisted of 2,256 cases and 2,310 controls from ANZ (**Table 1**), producing a combined total sample size for the discovery and replication cohorts of 3,874 cases and 5,723 controls. Replication genotyping was conducted for 100 SNPs, and 91 SNPs passed quality control filters (Online Methods). Twenty-one SNPs achieved suggestive evidence of association ($P \leq 0.05$). Of these, six were associated with two previously unidentified MS susceptibility loci on chromosomes 12q13–14 and 20q13 (**Table 2**), nine were located in other suggestively associated loci (**Supplementary Table 2** online) and another six were located in previously identified MS susceptibility loci (**Table 3**). We did not find convincing evidence that risk-associated SNPs influence the clinical course of MS (**Supplementary Note** online).

At chromosome 12q13–14, 15 genotyped SNPs in a 288-kb genomic segment (chromosome 12: 56,255,005–56,542,981) were suggestively associated ($P < 0.01$) in the GWAS ($P = 3.8 \times 10^{-6}$ to $P = 7.4 \times 10^{-3}$), defining a candidate region that encompassed

[1]A complete list of authors and affiliations is provided at the end of the article. Correspondence should be directed to J.P.R. (justin.rubio@florey.edu.au).

Received 23 February; accepted 29 April; published online 14 June 2009; doi:10.1038/ng.396

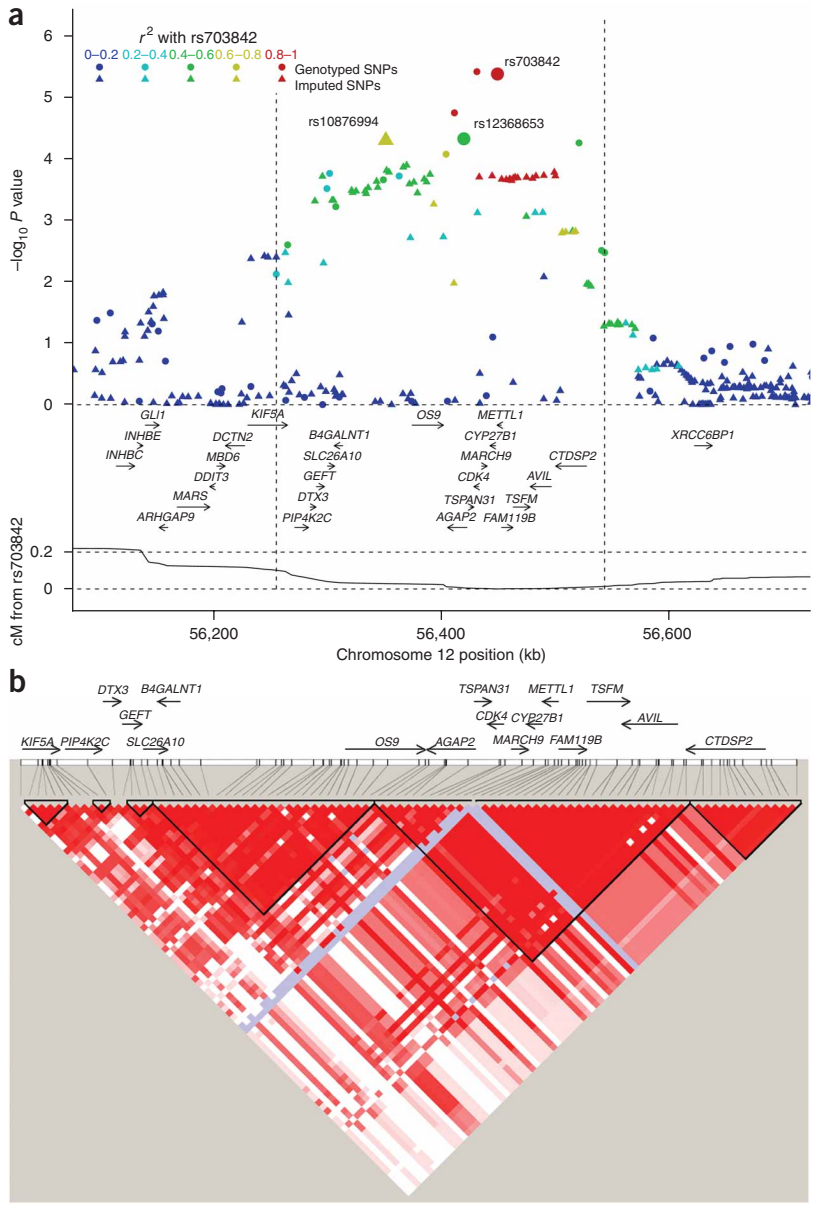

Figure 1 Chromosome 12q13–14 region and association with MS. (**a**) SNP association plot for a 600-kb region (56,100 kb to 56,700 kb) of chromosome 12 (NCBI dbSNP genome build 128 (October 2007)). *P* value ($-\log_{10}$) is plotted for SNPs on the *y* axis in relation to their scaled genomic position on the *x* axis. Filled circles represent genotyped SNPs and filled triangles imputed SNPs from the HapMap database. Larger shapes indicate SNPs (rs703842, rs10876994, rs12368653) that were genotyped in the replication phase. LD is shown in a range of colors relative to rs703842. Arrows identify genes, their approximate genomic size and direction of transcription. Genetic distance is indicated at the bottom of the plot in centimorgans (cM) in relation to rs703842. Vertical dashed lines encompass a putative candidate region that includes 17 genes. (**b**) LD structure at the chromosome 12 locus for the region including the 17 candidate genes (chromosome 12: 56,255,005–56,542,981) in CEU HapMap samples was visualized using default settings in Haploview version 4.1 (ref. 28). Inter-SNP LD is represented by shaded squares with red shading indicating strongest LD ($D' = 1$), and white indicating little or no LD between markers. Black triangles indicate the location of haplotype blocks.

17 known genes (**Fig. 1a**). Three SNPs genotyped in the replication phase were associated at the genome-wide level in the combined cohort (rs703842, $P = 5.4 \times 10^{-11}$; rs10876994, $P = 2.7 \times 10^{-10}$; rs12368653, $P = 1.0 \times 10^{-7}$). The seventh most associated non-MHC SNP in the GWAS, rs703842 ($P = 4.1 \times 10^{-6}$), is located in the 3′ UTR of the *METTL1* (methyltransferase-like protein 1) gene, 1.76 kb upstream of the *CYP27B1* (cytochrome P450 family 27 subfamily B) gene. There are moderate levels of linkage disequilbrium (LD) between rs703842, rs10876994 and rs12368653 ($r^2 = 0.28$–0.63; **Fig. 1b**). Fitting a logistic regression model with both rs703842 and rs10876994, we found that the association at rs10876994 is strongest ($P = 0.004$ versus $P = 0.086$), suggesting that rs10876994 may be in stronger LD with the causal disease allele. However, although these findings are suggestive of a single causal disease allele, they do not rule out the possibility of allelic heterogeneity or multiple disease loci. A SNP in *KIF5A* (rs1678542), located at the proximal boundary of the putative candidate region, that is associated with both rheumatoid arthritis[2] ($P = 8.8 \times 10^{-8}$) and type 1 diabetes (T1D)[4] ($P = 3.0 \times 10^{-4}$) showed modest evidence of association with MS risk ($P = 3 \times 10^{-3}$). For MS, at least, the peak of association seems to be skewed distal of *KIF5A*, and thus other genes in the region seem more likely candidates for the causal disease locus. Assuming a common causal autoimmune susceptibility gene underlies these associations, the strongest candidate given available genetic, immunological and epidemiological evidence is *CYP27B1*.

CYP27B1 encodes the enzyme 25-hydroxyvitamin D-1 alpha hydroxylase, which hydroxylates 25-hydroxyvitamin D into the bioactive form (1,25-dihydroxyvitamin D ($1,25(OH)_2D$)). The $1,25(OH)_2D$ form regulates calcium metabolism through the vitamin D receptor (VDR) and has important immune functions, modulating innate immunity and regulating adaptive immunity and tolerance[11] and B-cell homeostasis[12]. It can also direct activated T cells toward a T helper type 2 anti-inflammatory phenotype[13] and induce dendritic cells with tolerogenic properties[14].

In some animal models of autoimmunity, $1,25(OH)_2D$ and the vitamin D endocrine system are important in the prevention of disease onset and progression[15,16], and epidemiological data suggest a link between vitamin D deficiency and increased incidence of MS and other autoimmune diseases[17–19]. Mutations in *CYP27B1* cause vitamin D–dependent rickets type I (VDDR1, MIM264700), but common variants are associated with risk of T1D[3]. The T1D risk-associated variants in *CYP27B1* (−1260 C>A (rs10877012) and +2838T>C (rs4646536)) are in complete LD with each other and with rs703842, the most strongly associated SNP in this study. We found the same allelic association with MS as has been observed for T1D, providing further evidence for the existence of a common causal disease allele.

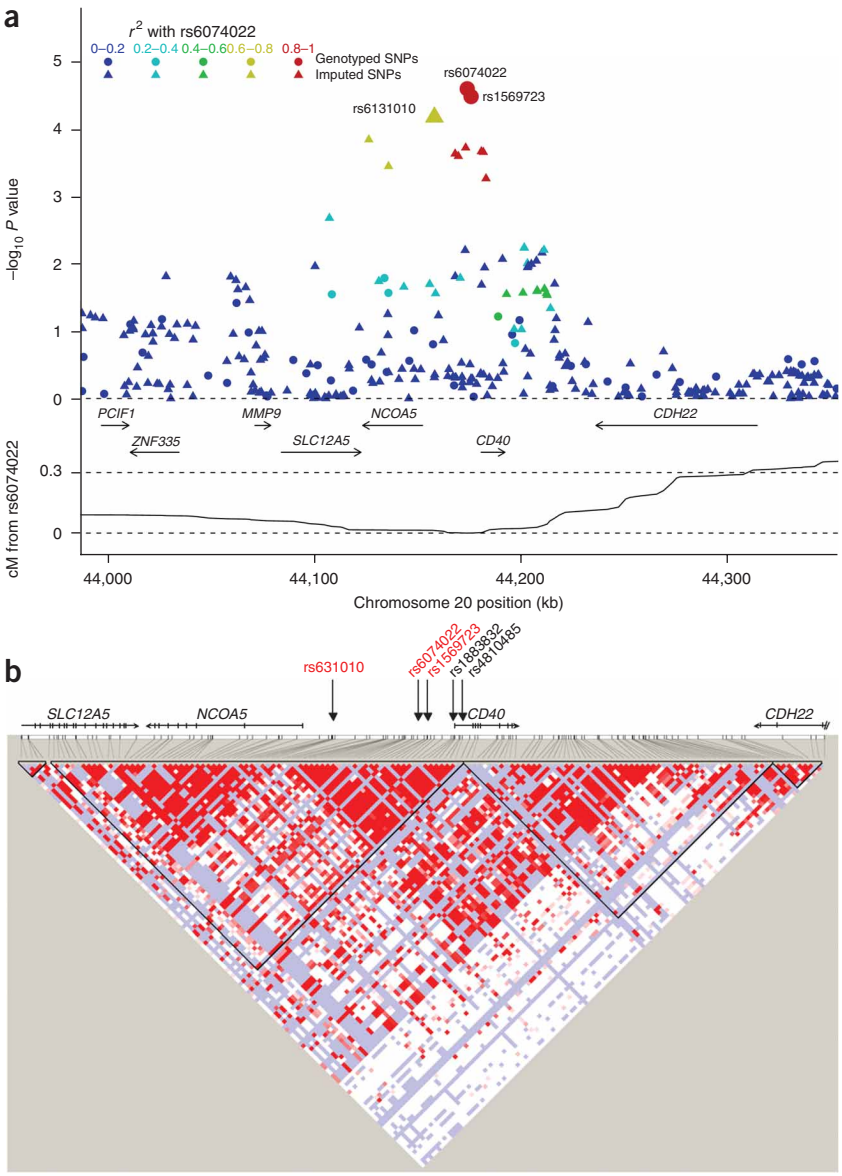

Figure 2 Chromosome 20q13 region and association with MS. (**a**) SNP association plot for a 340-kb region (40,000 kb to 44,340 kb) of chromosome 20. (**b**) LD structure at the chromosome 20 locus (44,100 kb to 44,250 kb). Red SNP markers were genotyped in the replication phase of this study (**Table 3**). rs4810485 is associated with susceptibility to rheumatoid arthritis[2] and rs1883832 is associated with Graves' disease[1]. $D' = 1$ for all SNP combinations involving rs607422, rs1569723, rs1883832 and rs4810485. Vertical lines in genes represent the approximate locations of exons.

Another plausible autoimmune candidate gene on chromosome 12q13–14, *CDK4* (cyclin-dependent kinase 4), can be modulated by CD40 signaling[20] and promote autoreactivity in the non-obese diabetic (NOD) mouse[21]. CDK4 inhibitors can ameliorate mouse collagen-induced arthritis (CIA)[22]. *CDK4* is downregulated in T cells from Japanese MS cases compared to controls[23].

Eleven SNPs in a 56.65-kb region on chromosome 20q13.12 (chromosome 20: 44,126,005–44,182,658) were associated ($P < 0.001$) with susceptibility to MS in the GWAS (**Fig. 2a**). Two SNPs were located in the *NCOA5* (nuclear receptor coactivator 5) gene (rs13037326, $P = 1.4 \times 10^{-4}$; rs2425752, $P = 3.6 \times 10^{-4}$) and three SNPs were located in *CD40* (rs1883832, $P = 2.2 \times 10^{-4}$; rs4810485, $P = 2.2 \times 10^{-4}$; rs4239702, $P = 5.4 \times 10^{-4}$). A further six SNPs were located in the intergenic region between *NCOA5* and *CD40*, including the three most associated markers in this region (rs6074022, rs1569723, rs6131010), which were genotyped in the replication phase (**Table 2** and **Fig. 2a**). In the combined sample, rs6074022 ($P = 1.3 \times 10^{-7}$), which was ranked nineteenth most associated non-MHC SNP in the GWAS, and rs1569723 ($P = 2.9 \times 10^{-7}$) achieved a significance level of $P < 5 \times 10^{-7}$. In HapMap samples, rs6074022 and rs1569723 are in complete LD ($r^2 = 1$) and both are in LD with rs6131010 ($r^2 = 0.7$) (**Fig. 2b**). These variants are located 5′ of *CD40*, and their risk-associated alleles are encompassed by the same extended haplotype and haplotype block that includes a portion of the *CD40* exonic region (**Supplementary Fig. 2** online).

On chromosome 20q13, *CD40* is clearly a strong candidate MS susceptibility gene because of its association with rheumatoid arthritis[2] and Graves' disease[1]. CD40 (TNFRSF5) is a member of the tumor necrosis factor receptor superfamily and is expressed as a costimulatory molecule on B cells, dendritic cells, macrophages and microglia. Together with its ligand, CD40L (CD154), which is expressed on activated $CD4^+$ T-helper cells, CD40 is recognized as an important regulator of both humoral and cellular immunity[24].

Table 1 Genotyping summary

	GWAS cases (Illumina Hap370CNV)		GWAS controls (Illumina Infinium)			Replication cases (Sequenom iPlex)		Replication controls (Sequenom iPlex)	
Origin	AUS	NZ	AUS[a]	UK[b]	US[c]	AUS	NZ	AUS[d]	NZ[e]
Samples (*n*)	1,226	392	41	1,425	1,947	1,382	874	1,840	470
All samples (*n*)	1,618			3,413		2,256		2,310	

AUS, Australia; NZ, New Zealand; UK, United Kingdom; US, United States of America. Genotyping platforms used for the GWAS and replication study are indicated in parentheses. [a]AUS controls involved in a concurrent gene expression study (unpublished data). [b]Controls from the 1958 British (UK) Birth Cohort. [c]US controls from the Illumina iControlDB database. [d]AUS controls recruited through the Australian Bone Marrow Donor Registry. [e]NZ controls recruited through The University of Otago.

Table 2 SNP associations at previously undescribed MS susceptibility loci on chromosomes 12 and 20

						GWAS					Replication				Combined			
SNP	Chr	Position[a]	Selection criteria for replication	Minor allele	Major allele	Genotyped/ imputed	MAF cases	MAF controls	*P* value	OR	MAF cases	MAF controls	*P* value	OR	MAF cases	MAF controls	*P* value	OR
rs703842	12	56,449,006	ANZ, IMSGC	G	A	Gen	0.281	0.326	4.1×10^{-6}	0.81	0.293	0.339	1.4×10^{-6}	0.81	0.288	0.331	5.4×10^{-11}	0.81
rs10876994	12	56,351,004	ANZ, IMSGC	C	A	Imp	0.219	0.256	4.9×10^{-5}	0.82	0.228	0.276	6.1×10^{-8}	0.78	0.224	0.264	2.7×10^{-10}	0.80
rs12368653	12	56,419,523	ANZ, IMSGC	A	G	Gen	0.521	0.477	4.6×10^{-5}	1.19	0.510	0.474	2.6×10^{-4}	1.15	0.515	0.476	1.0×10^{-7}	1.17
rs6074022	20	44,173,603	ANZ, cand	G	A	Gen	0.287	0.247	2.5×10^{-5}	1.22	0.275	0.245	4.6×10^{-4}	1.17	0.280	0.247	1.3×10^{-7}	1.20
rs1569723	20	44,175,471	ANZ, cand	C	A	Gen	0.287	0.248	3.3×10^{-5}	1.22	0.274	0.245	7.4×10^{-4}	1.16	0.279	0.247	2.9×10^{-7}	1.19
rs6131010	20	44,157,712	ANZ, cand	A	G	Imp	0.291	0.253	6.7×10^{-5}	1.21	0.287	0.262	3.0×10^{-3}	1.14	0.289	0.257	8.5×10^{-7}	1.18

[a]SNP positions are derived from NCBI dbSNP genome build 128 (October 2007). Criteria for selection of SNPs for replication genotyping are as follows: 'ANZ' denotes SNPs ranked in the top 50 most associated non-MHC SNPs that were genotyped or the top 100 imputed SNPs in the current study (Online Methods). 'IMSGC' indicates SNPs located in a genomic region showing evidence of association ($P < 0.05$) in a GWAS performed by the International MS Genetics Consortium (IMSGC)[6]. 'Cand' refers to SNPs that are located either within or close to a gene that was considered a biologically plausible candidate.

In animal models of autoimmunity, disruption of the CD40–CD40L interaction with antibodies to CD40L prevents the onset of disease[25,26].

SNPs associated with MS risk upstream of *CD40* (rs6074022 and rs1569723) are in complete LD ($r^2 = 1$) with rs4810485G>T, a SNP in *CD40* that is associated with susceptibility to rheumatoid arthritis[2], and with rs1883832C>T, a SNP that is associated with Graves' disease[1]. The rs1883832C>T SNP juxtaposes (−1) the ATG translation initiation codon, disrupting the Kozak consensus sequence, and the common C allele has been shown to enhance the efficiency of CD40 translation[27]. rs1883832C has a predisposing effect on both Graves' disease and rheumatoid arthritis, but seems to be protective in MS.

In the combined sample, the HLA-DR15 tag SNP, rs9271366, was unequivocally associated with MS risk ($P = 7.0 \times 10^{-184}$, OR = 2.78), and rs1335532 ($P = 9.6 \times 10^{-8}$) in *CD58* was associated at $P < 5 \times 10^{-7}$ (**Table 3**). We also found suggestive evidence of association with other MS susceptibility genes including *EVI5-RPL5* ($P = 2.5 \times 10^{-6}$) and *IL2RA* ($P = 7.4 \times 10^{-6}$), and more modest evidence for *IL7R* ($P = 1.3 \times 10^{-3}$) and *TYK2* ($P = 3.5 \times 10^{-3}$). In *CLEC16A* (*KIAA0350*), rs7404554 showed suggestive association in the GWAS ($P = 1.1 \times 10^{-4}$) but failed quality control in the replication phase. We did not confirm previously reported associations with *KIF1B* (rs10492972; $P = 1.0$) and *CD226* (rs2051322; $P = 0.65$).

We observed a modest statistical interaction ($P = 0.001$, or $P = 0.038$ after correction for multiple testing) between the *EVI5-RPL5* locus and HLA-DR15 (rs9271366) in the combined samples manifesting as an association in HLA-DR15 non-carriers (OR = 1.29, 95% CI = 1.18–1.41) but not in HLA-DR15 carriers (OR = 1.03, 95% CI = 0.93–1.14). No other interactions were detected between any other associated SNPs and HLA-DR15, and there was no evidence of dominant effects at any associated SNPs (**Supplementary Table 3** online). A modest statistical interaction was identified between associated SNPs on chromosomes 12 (rs703842) and 20 (rs6074022) ($P = 0.03$), with the effects of risk alleles at one locus elevated in the presence of risk alleles at the other, but this did not withstand Bonferroni correction ($P = 0.71$).

In conclusion, we have identified two new MS susceptibility loci in chromosomal regions associated with susceptibility to other autoimmune diseases. Further studies will be required to replicate these associations and to map the causal variants. Confirmation of *CYP27B1* as the causal gene on chromosome 12q13–14 would pave the way for more extensive clinical trials to determine the benefits of vitamin D supplementation or VDR agonists for the prevention and treatment of MS.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank individuals with MS in Australia and New Zealand for supporting this research. We are grateful to M. Tanner for database management and J. Wright and C. Remediakis from Multiple Sclerosis Research Australia (MSRA) for expediting this research. J.P.R. and M.B. are supported by Career Development Awards from The National Health and Medical Research Council of Australia (NHMRC). M.A.B. is an NHMRC Principal Research Fellow. H.B. is an NHMRC Peter Doherty Post-doctoral Fellow. J.F. is an MSRA Post-doctoral Fellow. M.B.C. is supported by a grant from the John Hunter Hospital Charitable Trust Fund and a special grant from Macquarie Bank. Replication genotyping was conducted at the Murdoch

Table 3 SNP associations in known and putative MS susceptibility genes

						GWAS					Replication				Combined			
SNP	Chr	Position	Locus	Minor allele	Major allele	Genotyped/ imputed	MAF cases	MAF controls	*P* value	OR	MAF cases	MAF controls	*P* value	OR	MAF cases	MAF controls	*P* value	OR
rs9271366	6	32,694,832	*HLA-DRB1*	G	A	Gen	0.329	0.155	6.9×10^{-89}	2.68	0.339	0.151	8.7×10^{-100}	2.88	0.335	0.153	7.0×10^{-184}	2.78
rs1335532	1	116,902,480	*CD58*	G	A	Gen	0.101	0.124	1.1×10^{-3}	0.80	0.106	0.134	1.1×10^{-5}	0.76	0.104	0.128	9.6×10^{-8}	0.78
rs6604026	1	93,076,191	*EVI5-RPL5*	G	A	Gen	0.310	0.273	1.3×10^{-4}	1.20	0.305	0.278	2.1×10^{-3}	1.14	0.307	0.275	2.5×10^{-6}	1.17
rs2104286	10	6,139,051	*IL2RA*	G	A	Imp	0.234	0.263	2.1×10^{-3}	0.86	0.241	0.272	2.6×10^{-4}	0.85	0.238	0.267	7.4×10^{-6}	0.86
rs11808092	1	92,845,816	*EVI5-RPL5*	A	C	Gen	0.292	0.256	9.6×10^{-5}	1.20	0.281	0.260	1.3×10^{-2}	1.11	0.286	0.258	1.5×10^{-5}	1.16
rs7404554	16	10,960,425	*CLEC16A*	G	A	Gen	0.434	0.394	1.1×10^{-4}	1.18	Assay failed	–	–	–	–			
rs6897932	5	35,910,332	*IL7R*	A	G	Gen	0.253	0.267	1.3×10^{-1}	0.93	0.231	0.258	1.3×10^{-3}	0.86	0.240	0.264	1.3×10^{-3}	0.91
rs8118449	19	10,381,064	*TYK2*	A	G	Gen	0.293	0.328	5.2×10^{-4}	0.85	0.321	0.327	2.6×10^{-1}	0.97	0.309	0.327	3.5×10^{-3}	0.91
rs2051322	18	65,701,566	*CD226*	A	G	Gen	0.102	0.116	3.8×10^{-2}	0.87	0.121	0.112	9.1×10^{-1}	1.09	0.113	0.114	6.5×10^{-1}	0.98
rs10492972	1	10,275,699	*KIF1B*	C	T	Imp	0.323	0.315	0.44	1.04	0.316	0.324	7.9×10^{-1}	0.96	0.319	0.319	1.0	1.00

Children's Research Institute Sequenom Platform Facility. This work was supported by MSRA, John T. Reid Charitable Trusts, Trish MS Research Foundation and the Australian Research Council, under the Linkage Projects Scheme (LP0776744).

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Jacobson, E.M. *et al.* A CD40 Kozak sequence polymorphism and susceptibility to antibody-mediated autoimmune conditions: the role of CD40 tissue-specific expression. *Genes Immun.* **8**, 205–214 (2007).
2. Raychaudhuri, S. *et al.* Common variants at *CD40* and other loci confer risk of rheumatoid arthritis. *Nat. Genet.* **40**, 1216–1223 (2008).
3. Bailey, R. *et al.* Association of the vitamin D metabolism gene *CYP27B1* with type 1 diabetes. *Diabetes* **56**, 2616–2621 (2007).
4. Fung, E. *et al.* Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/*TNFAIP3* as a susceptibility locus. *Genes Immun.* **10**, 188–191 (2008).
5. Barton, A. *et al.* Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat. Genet.* **40**, 1156–1159 (2008).
6. Hafler, D.A. *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **357**, 851–862 (2007).
7. Rubio, J.P. *et al.* Replication of *KIAA0350, IL2RA, RPL5* and *CD58* as multiple sclerosis susceptibility genes in Australians. *Genes Immun.* **9**, 624–630 (2008).
8. International Multiple Sclerosis Genetics Consortium. The expanding genetic overlap between multiple sclerosis and type I diabetes. *Genes Immun.* **10**, 11–14 (2009).
9. Aulchenko, Y.S. *et al.* Genetic variation in the *KIF1B* locus influences susceptibility to multiple sclerosis. *Nat. Genet.* **40**, 1402–1403 (2008).
10. Ban, M. *et al.* Replication analysis identifies *TYK2* as a multiple sclerosis susceptibility factor. *Eur. J. Hum. Genet.* advance online publication, doi: 10.1038/ejhg.2009.41 (18 March 2009).
11. Adorini, L. & Penna, G. Control of autoimmune diseases by the vitamin D endocrine system. *Nat. Clin. Pract. Rheumatol.* **4**, 404–412 (2008).
12. Chen, S. *et al.* Modulatory effects of 1,25-dihydroxyvitamin D3 on human B cell differentiation. *J. Immunol.* **179**, 1634–1647 (2007).
13. Boonstra, A. *et al.* 1alpha,25-Dihydroxyvitamin d3 has a direct effect on naive CD4(+) T cells to enhance the development of Th2 cells. *J. Immunol.* **167**, 4974–4980 (2001).
14. Penna, G. & Adorini, L. 1 Alpha,25-dihydroxyvitamin D3 inhibits differentiation, maturation, activation, and survival of dendritic cells leading to impaired alloreactive T cell activation. *J. Immunol.* **164**, 2405–2411 (2000).
15. Lemire, J.M. & Archer, D.C. 1,25-dihydroxyvitamin D3 prevents the *in vivo* induction of murine experimental autoimmune encephalomyelitis. *J. Clin. Invest.* **87**, 1103–1107 (1991).
16. Cantorna, M.T., Hayes, C.E. & DeLuca, H.F. 1,25-Dihydroxycholecalciferol inhibits the progression of arthritis in murine models of human arthritis. *J. Nutr.* **128**, 68–72 (1998).
17. Munger, K.L., Levin, L.I., Hollis, B.W., Howard, N.S. & Ascherio, A. Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. *J. Am. Med. Assoc.* **296**, 2832–2838 (2006).
18. Holick, M.F. Vitamin D deficiency. *N. Engl. J. Med.* **357**, 266–281 (2007).
19. van der Mei, I.A. *et al.* Vitamin D levels in people with multiple sclerosis and community controls in Tasmania, Australia. *J. Neurol.* **254**, 581–590 (2007).
20. Ishida, T. *et al.* CD40 signaling-mediated induction of Bcl-XL, Cdk4, and Cdk6. Implication of their cooperation in selective B cell growth. *J. Immunol.* **155**, 5527–5535 (1995).
21. Marzo, N. *et al.* Cyclin-dependent kinase 4 hyperactivity promotes autoreactivity in the immune system but protects pancreatic cell mass from autoimmune destruction in the nonobese diabetic mouse model. *J. Immunol.* **180**, 1189–1198 (2008).
22. Sekine, C. *et al.* Successful treatment of animal models of rheumatoid arthritis with small-molecule cyclin-dependent kinase inhibitors. *J. Immunol.* **180**, 1954–1961 (2008).
23. Satoh, J. *et al.* T cell gene expression profiling identifies distinct subgroups of Japanese multiple sclerosis patients. *J. Neuroimmunol.* **174**, 108–118 (2006).
24. Durie, F.H., Foy, T.M., Masters, S.R., Laman, J.D. & Noelle, R.J. The role of CD40 in the regulation of humoral and cell-mediated immunity. *Immunol. Today* **15**, 406–411 (1994).
25. Gerritse, K. *et al.* CD40–CD40 ligand interactions in experimental allergic encephalomyelitis and multiple sclerosis. *Proc. Natl. Acad. Sci. USA* **93**, 2499–2504 (1996).
26. Durie, F.H. *et al.* Prevention of collagen-induced arthritis with an antibody to gp39, the ligand for CD40. *Science* **261**, 1328–1330 (1993).
27. Jacobson, E.M., Concepcion, E., Oashi, T. & Tomer, Y.A. Graves' disease-associated Kozak sequence single-nucleotide polymorphism enhances the efficiency of CD40 gene translation: a case for translational pathophysiology. *Endocrinology* **146**, 2684–2691 (2005).
28. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).

The full list of authors and affiliations is as follows:

The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene)

Study design and management committee: Melanie Bahlo[2], David R Booth[3], Simon A Broadley[4,5], Matthew A Brown[6,7], Simon J Foote[8], Lyn R Griffiths[9], Trevor J Kilpatrick[10–12], Jeanette Lechner-Scott[13,14], Pablo Moscato[14,15], Victoria M Perreau[10], Justin P Rubio[11], Rodney J Scott[13–15], Jim Stankovich[8], Graeme J Stewart[3], Bruce V Taylor[8], James Wiley[16] (Chair)

Sample processing, data handling and genotyping: Matthew A Brown[6,7], David R Booth[3], Glynnis Clarke[17], Mathew B Cox[14,15], Peter A Csurhes[18], Patrick Danoy[6], Karen Drysdale[8], Judith Field[11], Simon J Foote[8], Judith M Greer[18], Lyn R Griffiths[9], Preethi Guru[8], Johanna Hadler[6], Brendan J McMorran[8], Cathy J Jensen[11], Laura J Johnson[11], Ruth McCallum[19], Marilyn Merriman[19], Tony Merriman[19], Karen Pryce[6], Rodney J Scott[13–15], Graeme J Stewart[3], Lotfi Tajouri[9], Ella J Wilkins[11], Justin P Rubio[11] (Chair)

Data analysis: Melanie Bahlo[2], Matthew A Brown[6,7], Brian L Browning[20], Sharon R Browning[20], Devindri Perera[8], Justin P Rubio[11], Jim Stankovich[8] (Chief analyst)

Phenotyping: Simon Broadley[4,5], Helmut Butzkueven[11,21], William M Carroll[22,23], Caron Chapman[24], Allan G Kermode[22,23], Mark Marriott[12], Deborah Mason[25], Robert N Heard[3], Michael P Pender[26,27], Mark Slee[28], Niall Tubridy[29], Jeanette Lechner-Scott[13,14], Bruce V Taylor[8], Ernest Willoughby[30], Trevor J Kilpatrick[10–12] (Chair)

[2]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. [3]The Westmead Millenium Institute, Westmead, New South Wales, Australia. [4]School of Medicine, Griffith University, Queensland, Australia. [5]Department of Neurology, Gold Coast Hospital, Queensland, Australia. [6]Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland, Australia. [7]Botnar Research Centre, Nuffield Department of Orthopaedic Surgery, University of Oxford, Oxford, UK. [8]Menzies Research Institute, University of Tasmania, Hobart, Tasmania. [9]Genomics Research Centre, Griffith University, Queensland, Australia. [10]Centre for Neuroscience, University of Melbourne, Victoria, Australia. [11]The Howard Florey Institute, University of Melbourne, Victoria, Australia. [12]Royal Melbourne Hospital, Parkville, Victoria, Australia. [13]John Hunter Hospital, Hunter New England Health Service, Newcastle, New South Wales, Australia. [14]Hunter Medical Research Institute, Newcastle, New South Wales, Australia. [15]Centre for Bioinformatics, Biomarker Discovery and Information-based Medicine, University of Newcastle, New South Wales, Australia. [16]Department of Medicine, Nepean Hospital, Penrith, New South Wales, Australia. [17]Christchurch School of Medicine & Health Sciences, University of Otago, New Zealand. [18]UQ Centre for Clinical Research, University of Queensland, Queensland, Australia. [19]Department of Biochemistry, University of Otago, Dunedin, New Zealand. [20]Department of Statistics, The University of Auckland, Auckland, New Zealand. [21]Department of Neurology, Box Hill Hospital, Victoria, Australia. [22]Sir Charles Gairdner Hospital, Nedlands, Western Australia, Australia. [23]Australian Neuromuscular Research Institute, Nedlands, West Australia, Australia. [24]Barwon Health, Geelong, Victoria, Australia. [25]Canterbury District Health Board, Christchurch, New Zealand. [26]School of Medicine, University of Queensland, Queensland, Australia. [27]Department of Neurology, Royal Brisbane and Women's Hospital, Queensland, Australia. [28]School of Medicine, Department of Neurology, Flinders University, Bedford Park, Adelaide, South Australia, Australia. [29]Department of Neurology, St. Vincent's University Hospital, Dublin, Republic of Ireland. [30]Auckland District Health board, Auckland, New Zealand.

ONLINE METHODS

Study subjects. A total of 4,336 MS cases from Australia and New Zealand (ANZ) were involved in this study. Australian MS cases were self-identified volunteers recruited at centers located in Adelaide, Brisbane, Gold Coast, Hobart, Melbourne, Newcastle, Perth and Sydney. MS cases from New Zealand (NZ) were recruited across the country as part of a recent national prevalence survey. Control genotypes for the GWAS were provided by the Sanger Institute (Cambridge, UK) for the 1958 British Birth Cohort (n = 1,438), and for healthy US controls of Caucasian descent from the Illumina iControlDB (n = 2,083). Control subjects genotyped in the replication phase were de-identified donors from the Australian Bone Marrow Donor Registry (ABMDR) (n = 2,060) and population controls recruited through The University of Otago in NZ (n = 493). All cases and controls involved in the GWAS and the replication phase were included on the basis of reported European Caucasian ancestry. All cases and controls were recruited with written and informed consent according to the guidelines of local institutional ethics committees.

Phenotyping. Phenotypic assessments of MS cases were conducted according to established clinical[29,30] brain imaging and laboratory[31] criteria. All cases had either definite MS, clinically definite MS or laboratory-supported definite MS according to the McDonald and Poser criteria, respectively. **Supplementary Table 1** contains demographic information for MS cases involved in this study, after data cleaning.

DNA samples. Genomic DNA (gDNA) from Australian MS cases and controls was extracted from whole blood using a variety of standard laboratory approaches. gDNA from NZ MS cases was isolated from saliva self-collected into Oragene DNA tubes according to the manufacturer's instructions (DNA-genotek). DNA concentrations were assessed using pico green fluorescence, ultraviolet (OD_{260nm}) spectrophotometry or on an ethidium bromide–stained low-percentage agarose gel compared to a high-molecular-weight standard. Because of possible bacterial gDNA contamination and difficulty in obtaining reliable pico green and spectrophotometry measurements, all saliva DNA samples were estimated and assessed for their integrity by agarose gel electrophoresis and using at least one other method.

GWAS genotyping and quality control. Overall, 2,000 DNA samples from ANZ MS cases (n = 1,952) and ANZ controls (n = 48) were genotyped on the Illumina Infinium Hap370CNV array. Genotype clustering was conducted with the Illumina BeadStudio program, with cases from each recruiting center and DNA extracted from saliva and peripheral blood clustered separately to control for minor differences in clustering properties between centers. Samples from 1,925 individuals (96.3%) passed initial quality control filtering, with SNP genotype call rates $\geq$98%. Cryptic relationships between genotyped individuals were identified using pairwise identity-by-descent (IBD) estimation using the program PLINK[32], and 89 individuals were removed. Using the program EIGENSTRAT[33], we used principal components analysis (PCA) to investigate hidden population structure in the combined remaining case and control data. PCA was conducted on a subset of 77,704 SNPs not in LD, which were common to all sample sets. Through successive iterations, 178 outliers (35 ANZ cases and 1 control; 8 UK controls; 134 US controls) were excluded because they were located more than six s.d. from the mean along the first 10 principal components. Cases and controls differed significantly in their mean positions along eigenvectors 1, 2 and 3, and positions along these eigenvectors were recorded for subsequent analyses.

Eigenvectors 1 and 2 were both correlated with SNPs near the lactase gene (*LCT*) on chromosome 2q21.3. MS cases self-reporting southern or central European ancestry were often positioned at one end of the distribution along these eigenvectors (**Supplementary Fig. 3** online). Three case samples were removed because of discrepancies between reported sex and sex inferred from genotypes, and seven control samples were removed because it was not possible to infer sex from genotypes using PLINK. Data for 128 cases in two genotyping batches with below-average call rates were removed after inspection of genotype cluster plots. This left a final sample size of 1,618 MS cases and 3,413 controls.

Genotype imputation. Multilocus imputation analyses were run on the cleaned set of case and control genotypes for autosomes only. Beagle 3.0 (ref. 34) was used with default settings for these analyses. Two analyses were run to (i) impute the most likely values of missing genotypes[35], and (ii) to calculate posterior genotype probabilities at all HapMap SNPs. To impute genotypes at all HapMap SNPs, unphased HapMap phase II CEU trios (release 24, build 36, forward strand) were used as a reference panel. These samples were simultaneously phased and used for imputation of genotypes in the case and control samples. Individuals more than three s.d. from the mean along the first three eigenvectors in the EIGENSTRAT analysis (n = 209) were excluded from this imputation procedure, leaving 1,500 cases and 3,322 controls.

Care was taken to merge the cleaned genotype data with the HapMap data. In order to speed up computation while not reducing the accuracy of the imputation[34], the sample was split up into 50 batches. The three genotyping groups (ANZ cases and controls, UK 1958 BC and US controls) were spread evenly through sample batches to avoid potential bias.

Genome-wide association testing. For SNPs passing quality control, we tested for association using the Cochran-Armitage test for trend implemented in PLINK[32], with missing genotypes on autosomes imputed as described above. The genomic inflation factor, λ, was estimated from the median χ^2 statistic after excluding SNPs in the MHC. To assess the validity of the imputation procedure and the likelihood of false positives due to population stratification, we conducted several subsidiary tests of association: (i) without using imputed values for missing genotypes (this test was not run on the 46,827 SNPs showing significantly different call rates between the datasets, and it was the main test for SNPs on the X chromosome); (ii) discarding samples more than 3 s.d. from the mean along the first three principal components in the EIGENSTRAT analysis, leaving 1,500 cases and 3,322 controls; and (iii) discarding samples as in (ii) and adjusting for positions along the first three principal components by treating the principal components as continuous or categorical covariates in a logistic regression analysis. Results for these tests are shown for SNPs at the chromosome 12 and 20 loci in **Supplementary Table 4** online.

Genotype intensity cluster plots were examined for the top 500 ranked non-MHC SNPs. For the imputed HapMap SNPs, posterior genotype probabilities were used to perform the association analysis. For each individual at each marker, we calculated a posterior allele dosage, which is twice the posterior probability of the AA genotype plus the posterior probability of the AB genotype (where A represents one of the alleles for the marker and B the other allele). We used a two sample t-test to determine whether the mean of the distribution of posterior allele dosages for the cases differed from that of the controls. As a check, we also conducted the same test to compare the two control groups (UK versus US).

We carried out several quality control checks on genotyped and imputed SNPs before including them in the replication phase: (i) examination of genotype cluster plots for the ANZ samples and UK controls combined (genotype intensity information was not available for the US controls); (ii) removal of SNPs with significant differences in allele frequencies between the UK and US controls; (iii) interrogation of P values at nearby genotyped SNPs, particularly for imputed SNPs, to provide increased evidence of 'true' association at the test SNP.

The replication phase. We calculated that our sample of 1,618 cases and 3,413 controls would provide 7% power to identify genotyped loci conferring an allelic odds ratio (OR) of 1.2 (multiplicative effect and risk allele frequency of 0.2) in the GWAS at a significance level of $P = 5 \times 10^{-7}$, or would provide 59% power to identify such a locus at the threshold used here ($P = 9.2 \times 10^{-4}$) to prioritize SNPs for the replication phase. SNPs passing quality control were considered for genotyping in the replication phase if they ranked in the top 500 most associated non-MHC SNPs ($P < 9.2 \times 10^{-4}$), either using the primary test or one of the three subsidiary tests (above), or if they ranked in the top 500 most associated imputed SNPs ($P < 1.0 \times 10^{-4}$). All SNPs selected for the replication phase fulfilled at least one of the following additional statistical and biological criteria: (i) ranked in the top 50 most associated genotyped SNPs ($P < 7.2 \times 10^{-5}$) or the top 100 imputed SNPs ($P < 1.7 \times 10^{-5}$) not in the MHC, or were located within a small (<200 kb) distance of such a SNP; (ii) were in, or close to, genes showing differential gene expression between cases and controls in a concurrent microarray study (unpublished data); (iii) were located within or in close proximity to associated genes involved in MS, inflammation, myelination, T-cell function or autoimmune disease;

(iv) overlapped with regions or genes associated ($P < 0.05$) in the IMSGC GWAS[6]. A SNP, rs9271366, that tagged the HLA-DR15 (DRB1*1501-DQB1*0602) haplotype in 732 of 740 DR15-positive subjects (prediction accuracy 98.9%, correlation 0.98) was genotyped to enable interaction analyses.

A total of 100 SNPs were genotyped in the replication phase using the Sequenom MassARRAY system and iPlex Gold chemistry. Individuals with successful genotype calls at less than 90% of SNPs were discarded (128 cases and 243 controls), leaving 2,256 cases and 2,310 controls. One assay, rs7404554 (in *CLEC16A*), failed completely. A further eight SNPs were excluded from further consideration: four with a call rate less than 95%, two with significantly different call rates in cases and controls ($P < 10^{-4}$) and two in Hardy-Weinberg disequilibrium ($P < 10^{-7}$) in cases and controls. Of the 91 SNPs that passed these filters, 10 were in previously known and putative MS susceptibility genes, 53 were selected on the basis of statistical criteria ('i' and 'iv', above), 14 were selected because of biological plausibility ('ii' and 'iii', above), and 14 were selected on both statistical and biological criteria (**Supplementary Table 5** online). The Cochran-Armitage test for trend was again used to test for association, and one-sided *P* values were calculated by comparing the direction of association in the replication sample with the direction of association in the GWAS.

Combined, two-sided *P* values and odds ratios were calculated using a 2 × 2 × 2 Cochran-Mantel-Haenszel test in PLINK. For SNPs genotyped in the replication phase that were imputed from HapMap data in the GWAS (26 of 91), combined *P* values were calculated using the same *t*-test on posterior allele dosages as described above. For individuals in the replication sample, where SNPs were genotyped, posterior genotype probabilities were set at 1 and 0. Logistic regression was used to test for interactions between associated SNPs and HLA-DR15, and between SNPs on chromosomes 12 and 20, and to test for significant dominance effects. A Bonferroni correction was applied to correct for performing these 38 tests.

29. Poser, C.M. *et al.* New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann. Neurol.* **13**, 227–231 (1983).
30. McDonald, W.I. *et al.* Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann. Neurol.* **50**, 121–127 (2001).
31. Paty, D.W. *et al.* MRI in the diagnosis of MS: a prospective study with comparison of clinical evaluation, evoked potentials, oligoclonal banding, and CT. *Neurology* **38**, 180–185 (1988).
32. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
33. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
34. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
35. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).