nature genetics

# Meta-analysis of genome scans and replication identify *CD6*, *IRF8* and *TNFRSF1A* as new multiple sclerosis susceptibility loci

Philip L De Jager[1–3], Xiaoming Jia[4], Joanne Wang[5,6], Paul I W de Bakker[3,4], Linda Ottoboni[1–3], Neelum T Aggarwal[7], Laura Piccio[8], Soumya Raychaudhuri[3,9], Dong Tran[3], Cristin Aubin[3], Rebeccah Briskin[2], Susan Romano[1], International MS Genetics Consortium, Sergio E Baranzini[5], Jacob L McCauley[10], Margaret A Pericak-Vance[10], Jonathan L Haines[11], Rachel A Gibson[12], Yvonne Naeglin[13], Bernard Uitdehaag[14], Paul M Matthews[12], Ludwig Kappos[13], Chris Polman[14], Wendy L McArdle[15], David P Strachan[16], Denis Evans[7], Anne H Cross[8], Mark J Daly[3,17], Alastair Compston[18], Stephen J Sawcer[18], Howard L Weiner[1], Stephen L Hauser[5,6,19], David A Hafler[1,3,19] & Jorge R Oksenberg[5,6,19]



**We report the results of a meta-analysis of genome-wide association scans for multiple sclerosis (MS) susceptibility that includes 2,624 subjects with MS and 7,220 control subjects. Replication in an independent set of 2,215 subjects with MS and 2,116 control subjects validates new MS susceptibility loci at *TNFRSF1A* (combined $P = 1.59 \times 10^{-11}$), *IRF8* ($P = 3.73 \times 10^{-9}$) and *CD6* ($P = 3.79 \times 10^{-9}$). *TNFRSF1A* harbors two independent susceptibility alleles: rs1800693 is a common variant with modest effect (odds ratio = 1.2), whereas rs4149584 is a nonsynonymous coding polymorphism of low frequency but with stronger effect (allele frequency = 0.02; odds ratio = 1.6). We also report that the susceptibility allele near *IRF8*, which encodes a transcription factor known to function in type I interferon signaling, is associated with higher mRNA expression of interferon-response pathway genes in subjects with MS.**

Multiple sclerosis is thought to emerge when genetically susceptible individuals encounter environmental triggers and initiate an inflammatory reaction against self-antigens in the central nervous system (CNS); these events result in recurring episodes of inflammatory demyelination and, in many cases, a progressive neurodegenerative process[1]. The genetic architecture underlying susceptibility to MS is complex, and there are no known mendelian forms. As seen with many other inflammatory diseases, the major histocompatibility complex (MHC) has long been associated with MS, and both class I and class II susceptibility alleles exist[2,3]. However, a recent genome-wide association study (GWAS) revealed the existence of multiple non-MHC MS susceptibility loci of modest effect[4]. The role of three such loci—*CLEC16A*, *IL2RA* and *IL7R*—has now been well validated by other investigators and by our own replication efforts[5–7]. Given the success of the GWAS approach in MS, we extended earlier gene discovery efforts by pooling together data from three separate genome-wide studies exploring the genetic architecture of MS and report three newly identified susceptibility loci for MS.

## RESULTS

### GWAS meta-analysis and replication

We conducted a meta-analysis of genome-wide data from (i) 895 subjects with MS genotyped in the original scan by the International MS Genetic Consortium[4], (ii) 969 subjects with MS scanned by the

[1]Division of Molecular Immunology, Center for Neurologic Diseases, Department of Neurology, Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [2]Partners Center for Personalized Genetic Medicine, Boston, Massachusetts, USA. [3]Program in Medical & Population Genetics, Broad Institute of Harvard University and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [5]Department of Neurology and [6]Institute for Human Genetics, School of Medicine, University of California, San Francisco, San Francisco, California, USA. [7]Rush Alzheimer Disease Center & Department of Neurological Sciences, Rush University, Chicago, Illinois, USA. [8]Department of Neurology, Washington University, St. Louis, Missouri, USA. [9]Division of Immunology, Allergy and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [10]Miami Institute for Human Genomics, University of Miami, Miller School of Medicine, Miami, Florida, USA. [11]Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, Tennessee, USA. [12]GlaxoSmithKline Clinical Imaging Centre, Hammersmith Hospital and Department of Clinical Neurosciences, Imperial College, London. [13]Department of Neurology, University Hospital Basel, Basel, Switzerland. [14]Department of Neurology, Vrije Universiteit Medical Centre, Amsterdam, The Netherlands. [15]Department of Social Medicine, University of Bristol, Bristol, UK. [16]Division of Community Health sciences, St. George's, University of London, London, UK. [17]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [18]University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's Hospital, Cambridge, UK. [19]These authors contributed equally to this work. Correspondence should be addressed to P.L.D. (pdejager@rics.bwh.harvard.edu).

Received 8 December 2008; accepted 21 May 2009; published online 14 June 2009; doi:10.1038/ng.401

**Table 1 Sample sets used in the meta-analysis and replication studies**

| | Meta-analysis sample collections | | | | | | |
|---|---|---|---|---|---|---|---|
| Stratum | IMSGC UK | IMSGC US | BWH | Gene MSA CH | Gene MSA NL | Gene MSA US | Total |
| Number of cases | 453 | 342 | 860 | 230 | 253 | 486 | 2,624 |
| Number of controls | 2,950[a] | 1,679[b] | 1,720[c] | 232[d] | 208[d] | 431[d] | 7,220 |
| Genotyping platform | Affy 500K | Affy 500K | Affy 6.0 | Illumina 550 | Illumina 550 | Illumina 550 | |

| | Replication sample collections | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Stratum | US | | | | | UK | | |
| Collection | BWH | WU | ACP | UCSF | RUSH | UC | 1958 BC | Total |
| Number of cases | 228 | 152 | 597 | 407 | 0 | 831 | 0 | 2,215 |
| Number of controls | 407[e] | 13 | 35 | 142 | 489 | 0 | 1,030 | 2,116 |

In each pair of matched cases and controls, all subjects are genotyped using the same genome-wide platform.
[a]Wellcome Trust Case Control Consortium–healthy control subjects. [b]NIMH–healthy control subjects. [c]MIGen study–healthy control subjects and subjects with a history of early myocardial infarction & BWH healthy control subjects. [d]Gene MSA–healthy control subjects recruited at the respective subject recruitment sites for this MS study. [e]BWH controls–these subjects of European ancestry recruited in the Boston area include (1) unaffected spouses from our MS Genetics collections ($n = 14$), (2) the BWH PhenoGenetic Project subjects ($n = 292$), and healthy subjects from the HPCGG collection ($n = 101$) (see Online Methods for details). These subjects do not overlap with BWH control subjects used in the meta-analysis. 1958 BC, 1958 birth cohort; ACP, Accelerated Cure Project; BWH, Brigham & Women's Hospital; CH, Switzerland; IMSGC, International MS Genetics Consortium; NL, Netherlands; RUSH, RUSH University; UC, University of Cambridge, UK; UCSF, University California, San Francisco; UK, United Kingdom; US, United States; WU, Washington University, St. Louis.



GeneMSA consortium[8] and (iii) an unpublished set of data generated from 860 subjects with MS recruited at the Partners MS Center in Boston, Massachusetts. A detailed description of the component sample sets is presented in **Table 1**, and their clinical characteristics are outlined in **Supplementary Table 1a** online. As each of the studies used a different genotyping platform (**Table 1**), we used the phased chromosomes of HapMap samples of European ancestry (CEU)[9] and the MACH algorithm (see URLs section in Online Methods)[10] to impute missing autosomal SNPs with a minor allele frequency >0.01 in each of the three datasets. This effort produced a dataset containing a common panel of 2.56 million SNPs in 2,624 subjects with MS and 7,220 healthy control subjects. We then implemented a meta-analysis method that combines the association results from each of the six strata of subjects outlined in **Table 1**, taking into account the imputation uncertainty for each SNP (see Online Methods)[11]. Overall, the degree of statistical inflation was modest (genomic inflation factor $\lambda = 1.054$).

To organize the top results of the meta-analysis, we assembled SNPs into groups that were highly correlated with one another ($r^2 > 0.5$), selecting the SNP with the most extreme evidence of association (lowest $P$ value) to serve as the representative tagging marker for that group. We list the top 100 independent loci with the lowest $P$ values for association with susceptibility to MS, ranging from the *CD58* locus (rs12025416, $P = 4.74 \times 10^{-8}$) to the 14q31.3 locus (rs2022771, $P = 4.89 \times 10^{-5}$) in **Supplementary Table 2** online. Each locus is defined by a single tagging SNP and contains those polymorphisms that are in linkage disequilibrium (LD) with it. These 100 loci form the core of the SNP panel that was genotyped in the replication sample set (**Table 1** and **Supplementary Table 2**). Given the preponderance of women over men among individuals affected with MS, we also conducted a secondary genome-wide regression analysis that included a term for gender and a term for subject source to account for the structure of our subject samples (see Online Methods). From this analysis, 41 of the top 50 loci with a term for gender were not redundant with known loci or loci selected by the primary analysis; thus, these 41 SNPs were also included in the replication panel.

To supplement the 141 susceptibility loci selected in an unbiased manner, we selected an additional 47 SNPs for replication using one of the following strategies (**Supplementary Table 2**). First, we screened all SNPs with a $P < 10^{-3}$ in the meta-analysis and selected 32 loci that included candidate genes implicated in MS or pathologic inflammation according to a search of current literature. Second, we selected eight nonsynonymous coding SNPs (nscSNPs) with $P < 10^{-3}$ in the meta-analysis that also had a $P < 0.01$ in an independent screen for nscSNPs in MS[12]. Finally, for reference, we included seven SNPs previously associated with MS at or near a genome-wide level of significance. The putative association of the rs10492972 marker in the *KIF1B* locus with MS susceptibility[13] was not known at the time the replication panel was designed. As this locus did not offer evidence of association in our meta-analysis ($P = 0.72$), it was not included in the replication study. In all, we genotyped 188 SNPs in the replication samples from the UK and the US (**Table 1** and **Supplementary Table 1b**), of which 180 SNPs provided high-quality data for subsequent Cochran-Mantel-Haenszel analysis as well as a joint analysis of the replication and meta-analysis results (**Supplementary Table 3** online). The relative success of each of our SNP selection strategies is reported in **Supplementary Table 3**. We note that we do not have genome-wide estimates of ancestry for the subjects in the replication study, and therefore we cannot assess the level of population stratification that may exist within the separate UK and US strata of the replication samples.

Among the 180 SNPs that met quality-control criteria, we observed an excess of associations in the replication stage that was consistent with the direction of effect observed in the meta-analysis (**Fig. 1**). In **Table 2**, we present the top results of the replication analysis and the combined evidence for association of these loci. The known MHC class I and class II associations were detected in the replication samples: rs3135388 is the surrogate marker for the *HLA DRB1*1501* risk allele and rs2523393 is a surrogate marker for the *HLA B*4402* allele (**Table 2** and Online Methods)[14]. Consistent with previous findings, the associations of *HLA B*4402* and *HLA DRB1*1501* were independent (**Supplementary Table 4** online).

Outside the MHC, the previously validated associations with the *CLEC16A*, *IL2RA* and *IL7R* loci were observed (**Table 2**), and we now validate the *CD58* locus at a level of genome-wide significance ($P = 3.10 \times 10^{-10}$) (**Table 2**). The best *CD58* marker, rs2300747,

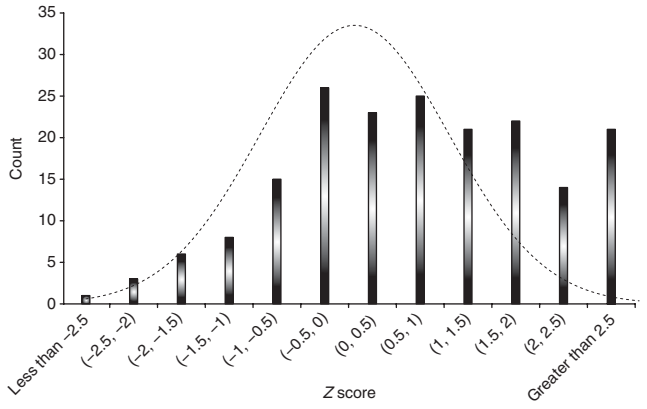


**Figure 1** Enrichment of associations in the replication stage that are consistent with the meta-analysis. Plot shows a histogram of the absolute values of the replication study *Z* scores from the 180 SNPs included in the replication analysis (see **Supplementary Table 3** for detailed results). In this case, the direction of the association reflects consistency with the results of the meta-analysis: a positive *Z* score is in the same direction in both the meta-analysis and the replication analysis while a negative *Z* score highlights the fact that the two analyses are discordant. An excess of concordant, highly significant associations is noted. A null distribution is plotted to highlight this enrichment.

was identified in earlier fine-mapping exercises in this locus[15] and is in strong linkage disequilibrium (LD) ($r^2 = 0.73$ in HapMap CEU samples) with the *CD58* marker selected from the meta-analysis, rs12025416 ($P = 1.16 \times 10^{-9}$) (**Supplementary Table 3**). Logistic regression revealed no evidence of an independent effect at rs12025416 in the replication samples (**Supplementary Table 4**). Thus, we see no evidence for allelic heterogeneity at the *CD58* locus in our data, and our previously identified *CD58* SNP (rs2300747) remains the best marker of a susceptibility allele within the *CD58* locus.

In the replication data, we found strong evidence for the presence of three previously unreported associations with genome-wide significance ($P < 5 \times 10^{-8}$) in the joint analysis: they are located in the *TNFRSF1A* (rs1800693, $P = 1.59 \times 10^{-11}$), *IRF8* (rs17445836, $P = 3.73 \times 10^{-9}$) and *CD6* (rs17824933, $P = 3.79 \times 10^{-9}$) loci. All three loci were selected for replication in an unbiased manner on the basis of the results of the meta-analysis (**Supplementary Table 2**). The rs1800693[T] allele associated with increased risk of MS (odds ratio (OR) = 1.20, 95% confidence interval (CI) = 1.10–1.31, in the replication stage) is found within the sixth intron of *TNFRSF1A*, and thus we refer to this region as the *TNFRSF1A* locus. Another gene, *PLEKHG6*, is found within the block of LD that contains rs1800693, so, formally, either or both genes could be associated with MS susceptibility (**Fig. 2a**). However, none of the SNPs that are in LD with rs1800693 are located within the *PLEKHG6* gene region (**Fig. 2a**); future fine-mapping efforts and follow-up studies will be needed to definitively resolve the functional basis of this association. Nonetheless, current literature supports a role for *TNFRSF1A*, as it has previously been implicated in tumor necrosis factor–associated periodic syndrome (TRAPS). This rare syndrome consists of recurrent episodes of systemic inflammation with variable symptoms including fever, abdominal pain, myalgia, arthralgia, exanthema and ocular involvement[16], and many affected individuals have been shown to have one of over 57 coding or noncoding mutations in *TNFRSF1A* (see URLs section in Online Methods). Most of these alleles are rare

**Table 2 Results of the replication and combined analysis**

| | | | | | Replication study | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | SNP | A1 | A2 | BP | $\chi^2$ | MAF_US | MAF_UK | OR | L95 | U95 | $P_{replication}$ | $P_{meta}$ | $P_{Joint}$ | Locus |
| | Previously published MS loci | | | | | | | | | | | | | |
| 6 | rs3135388 | A | G | 32521029 | 335.10 | 0.22 | 0.21 | 2.75 | 2.46 | 3.07 | $7.35 \times 10^{-75}$ | $7.44 \times 10^{-164}$ | $3.80 \times 10^{-225}$ | *HLA-DRB1* |
| 6 | rs2523393 | G | A | 29813638 | 30.55 | 0.41 | 0.41 | 0.78 | 0.72 | 0.85 | $3.26 \times 10^{-8}$ | $3.36 \times 10^{-11}$ | $1.04 \times 10^{-17}$ | *HLA-B* |
| 1 | rs2300747 | G | A | 116905738 | 14.31 | 0.12 | 0.11 | 0.77 | 0.68 | 0.88 | $1.55 \times 10^{-4}$ | $2.44 \times 10^{-7}$ | $3.10 \times 10^{-10}$ | *CD58* |
| 10 | rs2104286 | C | T | 6139051 | 7.87 | 0.24 | 0.25 | 0.87 | 0.79 | 0.96 | 0.0050 | $1.52 \times 10^{-6}$ | $9.33 \times 10^{-8}$ | *IL2RA* |
| 16 | rs11865121 | A | C | 11074189 | 8.42 | 0.31 | 0.29 | 0.87 | 0.80 | 0.96 | 0.0037 | $1.30 \times 10^{-5}$ | $1.77 \times 10^{-7}$ | *CLEC16A* |
| 5 | rs6897932 | T | C | 35910332 | 5.60 | 0.25 | 0.27 | 0.89 | 0.81 | 0.98 | 0.0180 | $7.71 \times 10^{-6}$ | $1.67 \times 10^{-6}$ | *IL7R* |
| | Newly identified loci with genome-wide level of evidence | | | | | | | | | | | | | |
| 12 | rs1800693 | C | T | 6310270 | 17.57 | 0.45 | 0.42 | 1.20 | 1.10 | 1.31 | $2.77 \times 10^{-5}$ | $7.52 \times 10^{-8}$ | $1.59 \times 10^{-11}$ | *TNFRSF1A* |
| 12 | rs4149584 | T | C | 6312904 | 8.20 | 0.022 | 0.018 | 1.58 | 1.15 | 2.17 | 0.0042 | 0.00035 | $5.25 \times 10^{-6}$ | *TNFRSF1A*[a] |
| 16 | rs17445836 | A | G | 84575164 | 17.40 | 0.19 | 0.22 | 0.80 | 0.72 | 0.89 | $3.03 \times 10^{-5}$ | $3.05 \times 10^{-5}$ | $3.73 \times 10^{-9}$ | *IRF8* |
| 11 | rs17824933 | G | C | 60517188 | 10.39 | 0.25 | 0.23 | 1.18 | 1.07 | 1.30 | 0.0013 | $2.32 \times 10^{-7}$ | $3.79 \times 10^{-9}$ | *CD6* |
| | Newly identified loci with suggestive evidence | | | | | | | | | | | | | |
| 2 | rs882300 | T | C | 136692725 | 15.75 | 0.39 | 0.46 | 0.84 | 0.77 | 0.92 | 0.000517 | $7.23 \times 10^{-5}$ | $1.37 \times 10^{-7}$ | *CXCR4* |
| 5 | rs6896969 | A | C | 40460183 | 4.58 | 0.38 | 0.40 | 0.91 | 0.83 | 0.99 | 0.0324 | $1.44 \times 10^{-7}$ | $2.40 \times 10^{-7}$ | *PTGER4*[b] |
| 12 | rs1790100 | G | T | 122222678 | 3.86 | 0.24 | 0.22 | 1.11 | 1.00 | 1.22 | 0.0495 | $2.74 \times 10^{-7}$ | $7.21 \times 10^{-7}$ | *MPHOSPH9* |
| 10 | rs1250540 | G | A | 80706013 | 5.91 | 0.35 | 0.40 | 1.12 | 1.02 | 1.22 | 0.0151 | $9.89 \times 10^{-6}$ | $1.59 \times 10^{-6}$ | *ZMIZ1* |
| 3 | rs4680534 | C | T | 161181639 | 6.18 | 0.37 | 0.36 | 1.12 | 1.02 | 1.22 | 0.0129 | $6.80 \times 10^{-6}$ | $5.58 \times 10^{-6}$ | *IL12A*[b] |
| 1 | rs2760524 | A | G | 190797171 | 5.88 | 0.16 | 0.17 | 0.87 | 0.77 | 0.97 | 0.0153 | $1.07 \times 10^{-4}$ | $9.77 \times 10^{-6}$ | *RGS1*[b] |
| 6 | rs9321619 | G | A | 137916101 | 7.78 | 0.47 | 0.46 | 0.89 | 0.81 | 0.96 | 0.0053 | $9.34 \times 10^{-4}$ | $1.71 \times 10^{-5}$ | *OLIG3-TNFAIP3*[b] |

[a]The rs4149584 nonsynonymous coding SNP was genotyped after the replication effort was concluded; given the result at rs1800693 and suggestive evidence of association at rs4149584 in the meta-analysis, we genotyped rs4149584 separately in the replication sample set. The two SNPs have independent effects on MS susceptibility (**Supplementary Table 4**). [b]These loci have previously validated associations with other inflammatory diseases at a genome-wide level of significance: *PTGER4*, Crohn's disease; *IL12A* and *RGS1*, celiac disease; *OLIG3-TNFAIP3*, psoriasis, rheumatoid arthritis and systemic lupus erythematosus. MS association results for the associated SNP in another disease are reported in **Supplementary Table 5**. A1, minor allele; A2, major allele; BP, physical location of the SNP in build 36; MAF, minor allele frequency in US and UK strata; L95/U95, lower and upper bounds of the 95% confidence interval for the OR. At each locus, the OR is stated relative to the minor allele. Here, we list all loci with evidence of genome-wide significance ($P < 5 \times 10^{-8}$) as well as loci with suggestive results, which are defined as either (i) a joint $P < 1 \times 10^{-6}$ or (ii) a joint $P < 1 \times 10^{-4}$ and evidence of association to another inflammatory disease.

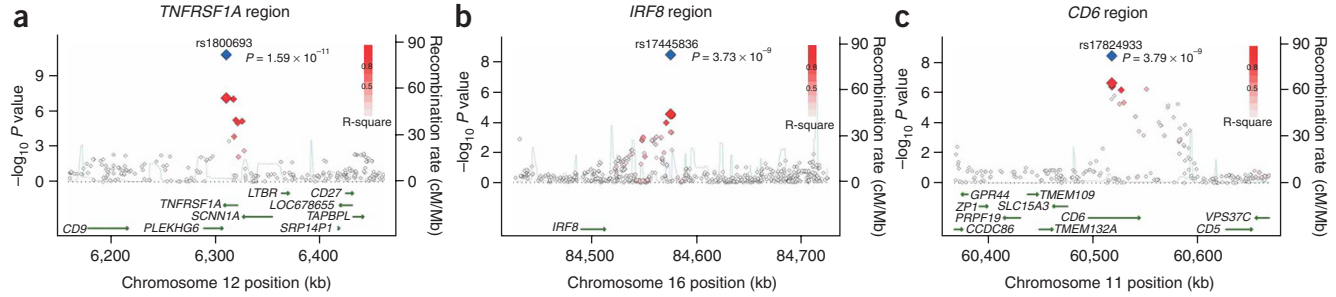


**Figure 2** Three previously unidentified loci, *TNFRSF1A*, *IRF8* and *CD6*, with genome-wide level of evidence of association to MS. (**a**) Illustration of the *TNFRSF1A* locus, with the local recombination rate plotted in light blue over this 200-kb chromosomal segment centered on rs1800693. Each diamond represents one SNP found in this locus, and the most associated SNP in the meta-analysis, rs1800693, is marked by a red diamond. A blue diamond is used to represent the level of evidence associated with rs1800693 in the joint analysis that includes the replication data. The color of each circle is defined by the extent of LD with rs1800693: red ($r^2 > 0.8$), orange ($0.8 > r^2 > 0.5$), gray ($0.5 > r^2 > 0.3$) and white ($r^2 < 0.3$). Physical positions are based on build 36 of the human genome. rs1800693 is located in an intron of *TNFRSF1A*. (**b**) Illustration of the *IRF8* locus, with the most associated SNP in this locus, rs17445836, highlighted by red (meta-analysis result) and blue (joint analysis result) diamonds. Here, we also present all SNPs found within a 200-kb window centered on rs17445836 and define SNP colors based on LD to rs17445836. (**c**) Illustration of the *CD6* locus, with the most associated SNP in this locus, rs17824933, highlighted by red (meta-analysis result) and blue (joint analysis result) diamonds. Here, we also present all SNPs found within a 200-kb window centered on rs17824933 and define SNP colors based on LD to rs17824933. In this case, *CD6* is the only gene found in the large block of LD that contains the association to MS susceptibility. Linkage disequilibrium maps are presented for all three loci in **Supplementary Figure 1a–c** online.



variants found in certain pedigrees, but a few less-penetrant alleles are segregating in European populations at <0.05 frequency[17,18]. Notably, a number of subjects with demyelinating or demyelinating-like diseases have been recently reported to harbor such variants (such as the R92Q substitution), but the slight excess in the proportion of these polymorphisms in MS subjects was not significant[16,19,20]. In our meta-analysis, only the R92Q polymorphism (rs4149584, labeled R121Q in dbSNP) has been analyzed. It has substantial evidence of association with MS susceptibility in the meta-analysis ($P = 0.0003$) as well as the replication effort ($P = 0.0042$), and it is not in LD with rs1800693 ($r^2 = 0.041$ in HapMap CEU samples), the common *TNFRSF1A* variant identified in our study (**Table 2**). Conditional analysis suggests that the two SNPs represent independent associations (**Supplementary Table 4**). Although a detailed investigation of common and rare variants in this locus is necessary to fully characterize MS-related effects, our study provides the first definitive link between the *TNFRSF1A* locus and susceptibility to demyelinating disease at both a high-frequency polymorphism of modest effect (rs1800693) and a low-frequency polymorphism of stronger effect (rs4149584). Altogether, rs1800693[T] and rs4149584[T] are excellent candidate risk alleles for other inflammatory diseases, particularly those with rheumatologic features.

The association of rs17445836[A] (OR = 0.80, 95% CI = 0.72–0.89) is also new and is found in a region of elevated recombination rate and lower LD (**Fig. 2b**). On the centromeric side, this SNP is located within 61 kb of *IRF8* (interferon response factor 8; also known as interferon consensus sequence binding protein 1, *ICSBP1*), and we therefore refer to this association as being in the *IRF8* locus because the closest telomeric gene, *FOXF1*, is 526 kb away. As its name implies, IRF8 is one of the several transcription factors that regulate responses to type I interferons (α and β interferons) by binding the interferon-stimulated response element (ISRE) (MIM601565). It has many roles that include involvement in B-cell germinal center development as well as macrophage cell function[21,22].

The third locus contains the rs17824933[G] susceptibility allele (OR = 1.18, 95% CI = 1.07–1.30) and is bounded by two peaks of recombination (**Fig. 2c**) between which only one gene, *CD6*, is found. The excess of extreme results with modest LD ($0.8 > r^2 > 0.5$) to rs17824933 at the telomeric end of the block of LD suggests that there may be an independent association within this locus. CD6, like CD58, is a molecule involved in T-cell costimulation and differentiation[23,24]; it may therefore have a role in modulating the activation and proliferation of T cells in the context of an inflammatory disease. In fact, these properties led to its targeting with a blocking monoclonal antibody in a clinical trial treating individuals with MS[25]. Finally, the soluble form of CD6 may also function as a pattern recognition receptor and affects the serum level of TNFα in this context in mice[26]. Thus, the rs17824933[G] susceptibility allele found in the first intron of *CD6* may have functional repercussions that interact with those of the *TNFRSF1A* locus.

Four other loci with $P < 10^{-4}$ in our joint analysis (**Table 2**) have previously validated associations with other inflammatory diseases

**Table 3 Gene set enrichment analysis pathways with coordinated RNA expression relative to the *IRF8* rs17445836[G] susceptibility allele**

| | Untreated | | | IFNβ-treated | | |
|---|---|---|---|---|---|---|
| Gene set | Size | NES | FDR *q* | Size | NES | FDR *q* |
| TAKEDA_NUP8_HOXA9_10D_UP | 27 | 2.8 | $<10^{-4}$ | 20 | 3.0 | $<10^{-4}$ |
| TAKEDA_NUP8_HOXA9_16D_UP | 19 | 2.7 | $<10^{-4}$ | 15 | 2.8 | $<10^{-4}$ |
| TAKEDA_NUP8_HOXA9_3D_UP | 29 | 2.7 | $<10^{-4}$ | 30 | 3.0 | $<10^{-4}$ |
| TAKEDA_NUP8_HOXA9_8D_UP | 24 | 2.7 | $<10^{-4}$ | 18 | 2.8 | $<10^{-4}$ |
| IFNA_HCMV_6HRS_UP | 21 | 2.3 | $<10^{-4}$ | 15 | 2.5 | $<10^{-4}$ |
| RADAEVA_IFNA_UP | 16 | 2.2 | $<10^{-4}$ | 17 | 2.8 | $<10^{-4}$ |
| DER_IFNB_UP | 17 | 2.2 | $5 \times 10^{-4}$ | 20 | 2.4 | $1 \times 10^{-4}$ |
| REOVIRUS_HEK293_UP | 23 | 2.1 | $7 \times 10^{-4}$ | 29 | 2.7 | $<10^{-4}$ |

This table presents only those gene sets with an FDR *q* value <0.001 in both the analysis of untreated subjects and that of IFNβ-treated subjects. When the analysis was repeated after excluding the ten subjects who had a diagnosis of CIS at the time of sampling, the same results are returned. "Size" reports the number of genes that are coordinately regulated within each gene set; "NES" is the "normalized enrichment score," an outcome parameter of the gene set enrichment method; "FDR q-val" reports a measure of significance corrected for the number of gene sets tested. Of note, the gene sets with a "TAKEDA" prefix are defined as sets of genes upregulated after transformation of $CD34^+$ hematopoietic cells with the NUP98/HOXA9 fusion protein that is found in certain translocations associated with myelodysplastic syndromes and acute myeloid leukemia. This fusion protein is known to strongly upregulate interferon β expression and, hence, interferon response pathways. **Supplementary Tables 6a** and **6b** present the detailed results of our analyses.

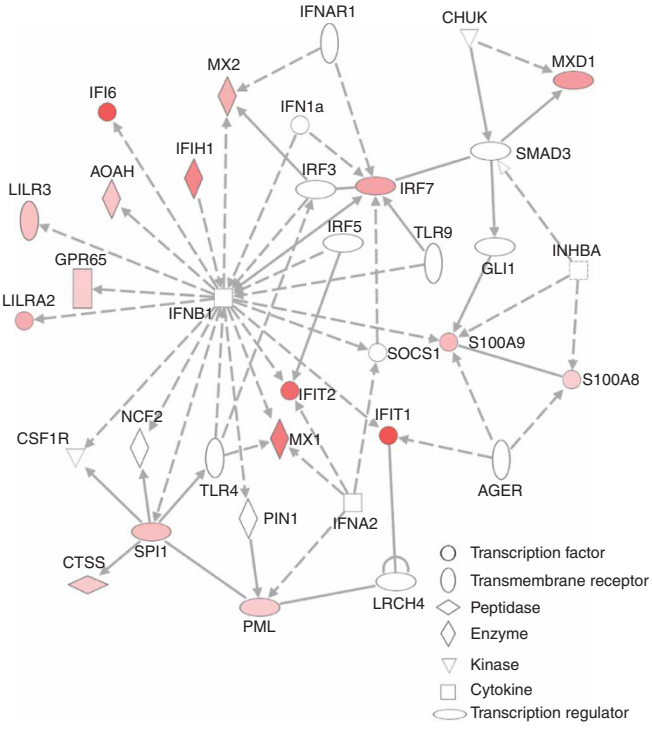


**Figure 3** Interferon response genes are coordinately upregulated relative to the rs17445836[G] allele of *IRF8*. Using the Ingenuity Pathways Analysis software suite, we illustrate the network of genes in the "interferon signaling pathway" whose expression is found to be correlated with the rs17445836[G] allele in untreated MS subjects. The diagram highlights those genes in this pathway whose expression are upregulated (red) in the presence of the rs17445836[G] allele. The magnitude of each gene's association is reflected by the intensity of the color, with brighter red indicating a stronger correlation between rs17445836[G] and RNA expression. A white color denotes a gene found in the associated pathway but which failed to meet a nominal $P < 0.05$ threshold for association of its RNA expression with rs17445836[G]. The shapes of each gene symbol denote the class of that gene as defined by the Ingenuity Pathways application: horizontal oval, transcription regulator; vertical oval, transmembrane receptor; rectangle, G protein–coupled receptor; horizontal diamond, peptidase; vertical diamond, enzyme; triangle, kinase; square, cytokine; circle, other.

and are therefore likely to be true MS susceptibility loci: *IL12A*, *OLIG3-TNFAIP3*, *PTGER4* and *RGS1*. The putative *IL12A* and *RGS1* MS susceptibility alleles are in strong LD with known celiac disease susceptibility alleles[27], as is the *PTGER4* MS allele and the known Crohn's disease allele[28] in this locus (**Supplementary Table 5a** online). On the other hand, the signal of association within the *OLIG3-TNFAIP3* locus seems to be distinct from known associations to psoriasis, rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE) (**Supplementary Table 5a**)[29–32]. Given the discovery of these strong candidate MS loci, we extended this comparative analysis to a larger number of loci by comparing our list of 100 top MS loci selected for replication (**Supplementary Table 2**) to the list of 76 Crohn's disease loci in which replication was attempted[28]. We found seven loci with substantial evidence of association in both diseases: these include not only loci with a validated role in one disease (*IL12B* and *PTGER4* in Crohn's disease as well as *IRF8* in MS) but also loci with suggested roles in both diseases (*BCL2*, *NEDD4L*, *PPA2* and *STAT3*; **Supplementary Table 5b**).

## Expression studies

Given the newly discovered association to the *IRF8* locus that contains an important transcription factor involved in responses to type I interferons, we explored its possible functional consequences by investigating a set of RNA data (Affymetrix U133 2.0 array) captured from the peripheral blood mononuclear cells (PBMCs) of 240 subjects of European ancestry with either remitting-relapsing MS (RRMS, $n = 230$) or a clinically isolated demyelinating syndrome (CIS, $n = 10$), many of which go on to develop MS (**Supplementary Table 1c**). These subjects can be classified into three categories: untreated subjects ($n = 82$), interferon β (IFNβ)-treated subjects ($n = 94$) and glatiramer acetate (GA)-treated subjects ($n = 64$). We used an unbiased approach to assess these data for the hypothesis that the *IRF8* locus may have a modest but broad effect on RNA expression from genes involved in interferon response. Specifically, we applied a gene set enrichment analysis (GSEA) methodology[33] to explore the results of a quantitative trait (eQTL) analysis correlating rs17445836 with our genome-wide RNA expression data from subjects with MS and CIS. As *IRF8* is known to be an interferon-response gene, its function could be affected either by IFNβ treatment or by GA treatment, which is reported to suppress IFNβ expression (S. Zamvil, University of California, San Francisco, personal communication). Thus, we pursued this GSEA screen of RNA data separately in each of our three subject subsets (untreated, IFNβ-treated and GA-treated). Sixteen gene sets that meet our threshold of significance (an FDR $q$ value $<0.05$) have genes that are coordinately upregulated in the presence of the rs17445836[G] allele in both the untreated and the IFNβ-treated subject subsets. Specifically, each of the 16 shared gene sets contain genes whose expression is coordinately enhanced under an additive model for rs17445836[G] association. In **Table 3**, we present the most associated of these 16 gene sets, that is, those gene sets that have an FDR $q$ value $<0.001$ in both sets of subjects. All eight of these most associated gene sets are primarily defined as being interferon-responding or are known to contain responses to type I interferons. Detailed results of each analysis are presented in **Supplementary Table 6a,b** online. Upregulation of interferon pathway genes in peripheral blood has previously been noted in ~50% of untreated subjects with MS[34,35], so the overlap between the untreated and IFNβ-treated subsets suggests that these results are consistent with our current knowledge of pathophysiology in MS. The lack of replication of the results of the untreated group in the GA-treated group is intriguing; it could be due to the smaller size of this subject subset ($n = 64$) and/or the suppression of IFNβ expression by GA. Further validation experiments in these subject subsets are needed to confirm our observations and explore the interactions of these MS treatments with the effect of the rs17445836[G] allele.

To control for potential bias in our analysis method, we repeated this investigation of the quantitative trait analysis results using the Ingenuity Pathways Analysis software suite (see URLs section in Online Methods). Here, using the same set of quantitative trait analysis results, we find significant co-regulation, relative to the rs17445836[G] allele, of genes within Ingenuity Systems' predefined "canonical interferon signaling pathway" among both untreated subjects ($P = 0.001$) and IFNβ-treated subjects ($P = 0.01$) (**Fig. 3**). The GA-treated subjects do not have a significant co-regulation of genes in this pathway. We have also repeated the GSEA and Ingenuity analyses using the best markers for MS susceptibility in the *CD6*, *CD58* and *TNFRSF1A* loci that were validated in this meta-analysis; none of these three loci show significant co-regulation within the interferon pathway (data not shown) relative to the best susceptibility marker in each locus. In addition, we examined a publicly available dataset generated

from a different cell type (EBV-transformed B cells) for the effect of the rs17445836[G] susceptibility allele on interferon response but did not observe this association in the small sample of HapMap cell lines of European ancestry (data not shown)[36]. Our data therefore suggest that both at baseline and during chronic exposure to exogenous INFβ the rs17445836[G] susceptibility allele may have a widespread but specific effect on gene expression in PBMC from subjects with MS, particularly within the interferon response pathway in which IRF8 is known to function (**Fig. 3**).

Only one probe in our RNA dataset provided information on the *IRF8* gene itself, and this probe shows no evidence of correlation between rs17445836[G] and *IRF8* expression. Thus, the mechanism by which rs17445836[G] influences gene expression remains unknown at this time, and more comprehensive studies of the expression of *IRF8* and its RNA isoforms in specific cell populations are needed to address this question.

## DISCUSSION

Our current data suggest that dysregulation of interferon responses may be one of the early events that contribute to the onset of MS. Upregulation of interferon responses has been noted not only in a subset of MS subjects[34,35], but also in subjects with other inflammatory diseases (dermatomyositis[37], rheumatoid arthritis and SLE[38,39]), and may reflect a shared feature of autoimmunity. However, the role of interferons in the onset of MS remains to be better defined. In addition, other pathways may also be affected by the *IRF8* variant, such as a gene set defined in response to TNFα stimulation that is coordinately upregulated in untreated MS subjects with the rs17445836[G] allele of *IRF8* (FDR $q$ value $<10^{-4}$, **Supplementary Table 6a**). This observation suggests a link between the functional consequences of the *IRF8* locus and those of the newly identified *TNFRSF1A* locus.

The possible role of the *TNFRSF1A* alleles in multiple sclerosis is informed by functional data from human studies. The TNFα pathway is implicated in MS susceptibility as a result of observations from human clinical data: treatment with monoclonal antibodies to TNFα may trigger acute episodes of CNS inflammation in subjects with MS[40]. A phase II clinical trial with a TNFRSF1A:IgG1 fusion protein (lenercept) also reported increased clinical attacks, although these occurred in the absence of enhanced disease activity on magnetic-resonance imaging and disability[41]. Furthermore, demyelinating lesions are a possible adverse event in subjects with Crohn's disease or rheumatoid arthritis treated with monoclonal antibodies to TNFα[42]. Thus, genetic and functional data now merge and suggest that dysregulation of the TNFα pathway has a role in the onset of MS, with diminished TNFα activity being associated with onset of CNS inflammatory lesions in clinical data. The suggested association of the *OLIG3-TNFAIP3* locus fits within this theme, as may the association of *CD6*: soluble CD6 may function as a pattern recognition receptor and influence circulating levels of TNFα[26]. With its observations relating to responses to class I interferons and TNFα, this study focuses our attention on dysregulation within the innate immune system in MS susceptibility. Dysfunction of the innate immune system, an important first line of defense against pathogens, has long been noted in immunopathology studies of MS. This history has contributed to the longstanding hypothesis of a viral or microbial trigger for MS, with the best evidence residing with the Epstein Barr Virus (EBV), and we must now consider our new associations in the context of such environmental risk factors[43].

Given the nearly 3:1 preponderance of women in all of our cohorts, we also need to better understand the impact of gender on MS susceptibility. Our secondary analysis that includes a term for gender was conducted for this reason and was successful in highlighting the role of the *CXCR4* locus ($P = 1.37 \times 10^{-7}$) (**Supplementary Table 3**). Although this locus was selected for replication on the basis of the secondary gender analysis, it showed strong evidence of replication in our primary replication analysis without a term for gender. The results of the secondary analysis of the replication data that includes gender as a covariate are shown in **Supplementary Table 7** online. The top results of this analysis generally mirror those of the primary replication analysis.

Overall, most validated and strongly suggested MS susceptibility loci (*CD6*, *CD58*, *CLEC16A*, *HLA-B*, *HLA-DRB1*, *IRF8*, *IL2RA*, *IL7R*, *IL12A*, *OLIG3-TNFAIP3*, *PTGER4*, *RGS1* and *TNFRSF1A*) have well-known and primarily immunologic functions. This is particularly true for our newly validated loci (*CD6*, *IRF8* and *TNFRSF1A*) that were selected for replication in an unbiased manner. In addition, many of these MS susceptibility loci have validated roles in other inflammatory diseases. Thus, we inform the ongoing debate of the relative roles of neurodegeneration and inflammation in the onset of MS by reporting a preponderance of current genetic evidence in favor of early immune dysregulation that may trigger secondary neurodegenerative processes. A definitive evaluation of this question awaits a more complete map of genetic susceptibility factors and a more comprehensive understanding of the functions of the associated genes in different cell types. This search for further susceptibility loci is also guided by the important observation that a less common variant (frequency of ~2% in European populations) of stronger effect has now been associated with susceptibility to MS in the *TNFRSF1A* locus. This suggests that future investigations of complex traits in MS will have to target this class of low-frequency alleles, which are typically only poorly interrogated by the current set of genome-wide SNP genotyping platforms.

## 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

P.L.D. is a Harry Weaver Neuroscience Scholar Award of the National MS Society (NMSS); he is also a William C. Fowler Scholar in Multiple Sclerosis Research and is supported by a National Institute of Neurological Disorders and Stroke (NINDS) K08 grant, NS46341. D.A.H. is a Jacob Javits Scholar of the US National Institutes of Health; he is also supported by NINDS P01 AI039671, R01 NS049477, R01NS046630, NMSS Collaborative MS Research Award and NMSS RG3567A. The International MS Genetics Consortium is supported by R01NS049477. L.P. is supported by an NMSS fellowship grant (FG1665-A-1). The genome-wide data on the BWH subjects and the RNA data on MS and CIS subjects from the CLIMB study were generated as part of a collaboration with Affymetrix, Inc. We thank the Myocardial Infarction Genetics Consortium (MIGen) study for the use of their genotype data as control data in our study. The MIGen study was funded by the US National Institutes of Health and National Heart, Lung, and Blood Institute's STAMPEED genomics research program and a grant from the National Center for Research Resources. We acknowledge use of genotype data from the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. We thank R. Lincoln and R. Gomez for expert specimen management at UCSF as well as A. Santaniello for database management. We thank the Accelerated Cure Project for its work in collecting samples from subjects with MS and for making these samples available to MS investigators. We also thank the following clinicians for contributing to sample collection efforts: Accelerated Cure project, E. Frohman, B. Greenberg, P. Riskind, S. Sadiq, B. Thrower and T. Vollmer; Washington University, B.J. Parks and R.T. Naismith. Finally, we thank the Brigham & Women's Hospital PhenoGenetic Project for providing DNA samples from healthy subjects that were used in the replication effort of this study.

### AUTHOR CONTRIBUTIONS

P.L.D., D.A.H., S.L.H., P.M.M. and J.R.O. designed the study. P.L.D. and J.R.O. wrote the manuscript. P.I.W.d.B., P.L.D., S.R., M.J.D., D.T., J.W., S.E.B. and X.J.

performed analytical work. P.I.W.d.B., X.J. and M.J.D. developed the meta-analysis method while S.R. developed the subject matching algorithm. L.O. and P.L.D. performed the quality control analysis and quantitative trait analysis of the RNA from MS PBMC samples. C.A. generated and processed genotype data for analysis. P.L.D., N.T.A., L.P., R.B., R.A.G., P.M.M., Y.N., L.K., B.U., C.P., W.L.M., D.P.S., D.E., A.H.C., A.C., S.J.S., H.L.W., S.L.H., J.R.O. and D.A.H. contributed to DNA sample collection and genetic data. J.L.M., M.A.P.-V. and J.L.H. contributed to the interpretation of the results. All authors have read and contributed to the manuscript.

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



1. Hauser, S.L. & Oksenberg, J.R. The neurobiology of multiple sclerosis: genes, inflammation, and neurodegeneration. *Neuron* **52**, 61–76 (2006).
2. Barcellos, L.F. *et al.* Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. *Hum. Mol. Genet.* **15**, 2813–2824 (2006).
3. Yeo, T.W. *et al.* A second major histocompatibility complex susceptibility locus for multiple sclerosis. *Ann. Neurol.* **61**, 228–236 (2007).
4. Hafler, D.A. *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **357**, 851–862 (2007).
5. 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).
6. International Multiple Sclerosis Genetics Consortium. Refining genetic associations in multiple sclerosis. *Lancet Neurol.* **7**, 567–569 (2008).
7. Ramagopalan, S.V., Anderson, C., Sadovnick, A.D. & Ebers, G.C. Genomewide study of multiple sclerosis. *N. Engl. J. Med.* **357**, 2199–2200 (2007).
8. Baranzini, S.E. *et al.* Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum. Mol. Genet.* **18**, 767–778 (2009).
9. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
10. Li, Y. & Abecasis, G.R. Rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.* **S79**, 2290 (2006).
11. de Bakker, P.I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
12. Burton, P.R. *et al.* Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* **39**, 1329–1337 (2007).
13. Aulchenko, Y.S. *et al.* Genetic variation in the *KIF1B* locus influences susceptibility to multiple sclerosis. *Nat. Genet.* **40**, 1402–1403 (2008).
14. de Bakker, P.I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).
15. De Jager, P.L. *et al.* The role of the *CD58* locus in multiple sclerosis. *Proc. Natl. Acad. Sci. USA* **106**, 5264–5269 (2009).
16. Hoffmann, L.A. *et al.* TNFRSF1A R92Q mutation in association with a multiple sclerosis-like demyelinating syndrome. *Neurology* **70**, 1155–1156 (2008).
17. Kumpfel, T. *et al.* Late-onset tumor necrosis factor receptor-associated periodic syndrome in multiple sclerosis patients carrying the TNFRSF1A R92Q mutation. *Arthritis Rheum.* **56**, 2774–2783 (2007).
18. Aksentijevich, I. *et al.* The tumor-necrosis-factor receptor-associated periodic syndrome: new mutations in *TNFRSF1A*, ancestral origins, genotype-phenotype studies, and evidence for further genetic heterogeneity of periodic fevers. *Am. J. Hum. Genet.* **69**, 301–314 (2001).
19. Wildemann, B. *et al.* The tumor-necrosis-factor-associated periodic syndrome, the brain, and tumor-necrosis-factor-alpha antagonists. *Neurology* **68**, 1742–1744 (2007).
20. Jenne, D.E. *et al.* The low-penetrance R92Q mutation of the tumour necrosis factor superfamily 1A gene is neither a major risk factor for Wegener's granulomatosis nor multiple sclerosis. *Ann. Rheum. Dis.* **66**, 1266–1267 (2007).
21. Pedchenko, T.V., Park, G.Y., Joo, M., Blackwell, T.S. & Christman, J.W. Inducible binding of PU.1 and interacting proteins to the Toll-like receptor 4 promoter during endotoxemia. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **289**, L429–L437 (2005).
22. Lee, C.H. *et al.* Regulation of the germinal center gene program by interferon (IFN) regulatory factor 8/IFN consensus sequence-binding protein. *J. Exp. Med.* **203**, 63–72 (2006).
23. Hassan, N.J. *et al.* CD6 regulates T-cell responses through activation-dependent recruitment of the positive regulator SLP-76. *Mol. Cell. Biol.* **26**, 6727–6738 (2006).
24. Castro, M.A. *et al.* Extracellular isoforms of CD6 generated by alternative splicing regulate targeting of CD6 to the immunological synapse. *J. Immunol.* **178**, 4351–4361 (2007).
25. Hafler, D.A. *et al.* Immunologic responses of progressive multiple sclerosis patients treated with an anti-T-cell monoclonal antibody, anti-T12. *Neurology* **36**, 777–784 (1986).
26. Sarrias, M.R. *et al.* CD6 binds to pathogen-associated molecular patterns and protects from LPS-induced septic shock. *Proc. Natl. Acad. Sci. USA* **104**, 11724–11729 (2007).
27. Hunt, K.A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* **40**, 395–402 (2008).
28. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
29. Graham, R.R. *et al.* Genetic variants near *TNFAIP3* on 6q23 are associated with systemic lupus erythematosus. *Nat. Genet.* **40**, 1059–1061 (2008).
30. Raychaudhuri, S. *et al.* Common variants at *CD40* and other loci confer risk of rheumatoid arthritis. *Nat. Genet.* **40**, 1216–1223 (2008).
31. Musone, S.L. *et al.* Multiple polymorphisms in the *TNFAIP3* region are independently associated with systemic lupus erythematosus. *Nat. Genet.* **40**, 1062–1064 (2008).
32. Nair, R.P. *et al.* Genome-wide scan reveals association of psoriasis with IL-23 and NF-κB pathways. *Nat. Genet.* **41**, 199–204 (2009).
33. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
34. van Baarsen, L.G. *et al.* A subtype of multiple sclerosis defined by an activated immune defense program. *Genes Immun.* **7**, 522–531 (2006).
35. Degre, M., Dahl, H. & Vandvik, B. Interferon in the serum and cerebrospinal fluid in patients with multiple sclerosis and other neurological disorders. *Acta Neurol. Scand.* **53**, 152–160 (1976).
36. Stranger, B.E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).
37. Greenberg, S.A. *et al.* Interferon-α/β-mediated innate immune mechanisms in dermatomyositis. *Ann. Neurol.* **57**, 664–678 (2005).
38. van der Pouw Kraan, T.C. *et al.* Rheumatoid arthritis subtypes identified by genomic profiling of peripheral blood cells: assignment of a type I interferon signature in a subpopulation of patients. *Ann. Rheum. Dis.* **66**, 1008–1014 (2007).
39. Baechler, E.C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl. Acad. Sci. USA* **100**, 2610–2615 (2003).
40. van Oosten, B.W. *et al.* Increased MRI activity and immune activation in two multiple sclerosis patients treated with the monoclonal anti-tumor necrosis factor antibody cA2. *Neurology* **47**, 1531–1534 (1996).
41. The Lenercept Multiple Sclerosis Study Group and The University of British Columbia MS/MRI Analysis Group. TNF neutralization in MS: results of a randomized, placebo-controlled multicenter study. *Neurology* **53**, 457–465 (1999).
42. Siddiqui, M.A. & Scott, L.J. Spotlight on infliximab in Crohn disease and rheumatoid arthritis. *BioDrugs* **20**, 67–70 (2006).
43. De Jager, P.L. *et al.* Integrating risk factors: HLA-DRB1*1501 and Epstein-Barr virus in multiple sclerosis. *Neurology* **70**, 1113–1118 (2008).

## ONLINE METHODS

**Study design.** Briefly, we conducted a fixed-effects meta-analysis of three whole-genome association scans for MS susceptibility (**Table 1**) based on the observed (imputed) and expected allele dosage of each SNP, taking into account the empirically observed variance of the allele dosage. As each dataset was generated on a different SNP array, we used the MACH algorithm[10] to impute genotypes on a single panel of 2.56 million SNPs; the meta-analysis was executed over this entire collection of SNPs. They were ranked based on the meta-analysis *P* values, and 186 SNPs outside of the MHC as well as two known MHC SNPs were selected for replication in an independent set of subjects (**Table 1**). We used the Cochran-Mantel-Haenszel method implemented in PLINK[44] to analyze the replication data, and we conducted a joint analysis by combining the results of the meta-analysis and replication studies using a sample size–weighted *Z* score (**Table 2**). Then, RNA data from PBMCs of subjects with MS and CIS (**Supplementary Table 1c**) were explored for evidence that the *IRF8* risk allele (rs17445836[G]) influences expression of genes in the interferon response pathway. We therefore explored broader effects on transcription in our PBMC RNA dataset (Affymetrix U133 2.0 plus) by analyzing the results of a comprehensive quantitative trait analysis across all RNA probes in relationship to rs17445836[G] using a gene set enrichment methodology[33].

**Human subjects and genome-wide data in the meta-analysis.** Summary information on the MS subjects and healthy control subjects are shown in **Table 1**. Below, we offer additional details on each sample set. All subjects with MS meet the McDonald criteria for a diagnosis of MS[45]. The demographic profiles of the sample collections are presented in **Supplementary Table 1a**.

In the initial MS genome-wide association scan by the IMSGC[4], subjects with MS were genotyped on the Affymetrix platform using the GeneChip Human Mapping 500K Array set. All healthy control subjects also have data generated on the Affymetrix platform using the same array. The MS subjects from the US were matched to healthy control subject data generated by the National Institute of Mental Health, and the MS subjects from the UK were matched to healthy control subject data generated by the Wellcome Trust. All details related to the quality parameters of these data (SNPs and subjects) are presented elsewhere[4]. Here, we use the same dataset that was analyzed in our earlier study, after removing subjects that were also genotyped in one of the other two studies: 54 subjects were genotyped in the original IMSGC study and the BWH study, and 82 subjects were genotyped in the IMSGC study and in the Gene MSA study. All of these duplicates were US subjects, and they were removed from the US component of the IMSGC dataset.

The BWH dataset is new. Its subjects with MS ($n = 860$) and healthy control ($n = 270$) subjects were genotyped on the Affymetrix Genome-wide Human SNP Array 6.0 (Genechip 6.0) at the Broad Institute's Center for Genotyping and processed for quality control (QC) using the PLINK software suite. We applied its standard quality control pipeline for subjects (genotype success rate >95%, sex concordant, excess inter/intra-heterozygosity) and for SNPs (Hardy-Weinberg equilibrium $P > 1 \times 10^{-6}$; MAF > 0.01, genotype call rate >0.95; misshap test $>1 \times 10^{-9}$) to these data. In a second step, EIGENSTRAT[46] was used to identify population stratification outliers. Given the limited number of healthy control subjects that we had genotyped, we selected additional control subjects from an existing dataset of 2,681 subjects with genotypes generated by the MIGen consortium using the Affymetrix Genechip 6.0 platform at the Broad Institute's Center for Genotyping in a scan for loci associated with susceptibility to early myocardial infarction (MI). As there is no known association between MS and early MI, we selected control subjects for our analysis from a combined pool of (i) healthy control subjects from BWH that were recruited for MS studies (spouses and friends of MS subjects), (ii) healthy control subjects from the MIGen study, and (iii) early MI cases from MIGen. The SNP content was reduced to the 709,690 SNPs that had passed quality control in both studies. Each subject with MS in the BWH dataset was then matched to two subjects drawn from the combined control subject pool (MIGen and BWH subjects) using the first principal component distance calculated by EIGENSTRAT as described below. As recommended by the authors of EIGENSTRAT, we excluded X-chromosome SNPs from the calculation of eigenvectors.

The Gene MSA study consists of three sets of samples: (i) GeneMSA-NL: 253 subjects with MS and 208 healthy control subjects were collected at the Vrije Universiteit Medical Centre in Amsterdam, Netherlands, (ii) GeneMSA-CH: 230 subjects with MS and 232 healthy control subjects were collected at the University Hospital Basel, Basel, Switzerland, and (iii) GeneMSA-US: 486 subjects with MS and 431 healthy control subjects were collected at the University of California, San Francisco. All subjects were genotyped genome-wide in a single batch using the HumanHap550 Beadchip produced by Illumina. Details of the quality control pipeline for these data are described in detail in a prior publication[8].

**Matching case and control subjects in the BWH discovery sample.** As described above, we matched BWH MS cases to a pool of control samples consisting of subjects from the MIGen study as well as healthy control subjects from BWH. All of these subjects had Affymetrix 6.0 genome-wide SNP data, genotyped at the Broad Institute. We selected a subset of ancestry-informative markers (709,690 SNPs that passed stringent quality controls in both the MIGen study and our own study). We excluded X-chromosome SNPs and then used EIGENSTRAT[46] to define genetic eigenvectors. We then matched cases and controls using the following strategy. First, we randomly selected a subject with MS (case). Second, for each case we selected the most genetically similar control from the pool of available unmatched control subjects. We defined similarity with a Euclidean distance metric based on the top eigenvector:

$$d_{i,j} = \sqrt{(a_i - a_j)^2}$$

where $d_{i,j}$ is the distance between case *i* and control *j*, and *a* is the value of the case or control in the first eigenvector. Steps 1 and 2 are repeated until a total of two controls are selected for each case.

This resulted in a matched collection of 860 cases and 1,720 controls. Of note, this method was intended to match cases and controls on the basis of the top two eigenvectors, but an error in the code prevented the use of the second eigenvector in the calculation. The error was discovered in the final stage of manuscript preparation. The gain in matching from using a second eigenvector was tested and is marginal in this dataset.

**Genome-wide genotype imputation and meta-analysis method.** A meta-analysis was conducted in 9,844 unique individuals (2,624 cases and 7,220 controls) to identify genetic loci associated with multiple sclerosis. These samples came from three separate genome-wide association studies (six separate cohorts or strata) that are described in **Table 1**. To maintain the existing case-control relationships, our analysis approach involved combining the results of independent analyses performed in the six strata that are outlined in **Table 1**. Specifically, to control for population stratification, all individuals were stratified into 268 clusters on the basis of pairwise identity by state (IBS) between individuals within each of the six strata using PLINK: this process yielded 83 clusters in BWH, 73 clusters in IMSGC/UK, 46 clusters in IMSGC/US, 33 clusters in GeneMSA-US, 15 clusters in GeneMSA-NL, and 18 clusters in GeneMSA-CH. These clusters were defined using only those genotyped SNP data that were available in each cohort. In parallel, we used MACH version 1.0.5 (ref. 10) to impute the genotypes of 2,557,248 SNPs across the genome based on phased chromosomes (haplotypes) of the CEU population in HapMap release 21 (NCBI build 35). Imputation was conducted on all samples, ignoring case-control status, to avoid introducing artifacts between cases and controls. We elected to use probabilistic dosages in our statistical analysis rather than hard genotype calls to account for the uncertainty of imputation at each locus. Standard quality metrics were applied to the imputed data: we considered only those SNPs with <5% genotype missing rate, minor allele frequency >0.01, and Hardy-Weinberg $P > 10^{-6}$.

We conducted a fixed-effects meta-analysis across all clusters based on the observed and expected allele dosage, taking into account the empirically observed variance of the allele dosage:

$$z_{meta} = \frac{\sum (p_o - p_e)}{\sqrt{\sum \mathrm{var}(p)}}$$

where $p_o$ and $p_e$ are the cumulative observed and expected allele dosage in cases per cluster, respectively. We take into account imputation uncertainty by taking the empirically observed variance of the allele dosage (computed per cohort) for var(*p*) if the average maximal posterior probability of an imputed SNP

$<0.99$ (that is, for poorly imputed SNPs). Otherwise, if the average maximal posterior probability of an imputed SNP $>0.99$ (that is, for well imputed SNPs), we take for var($p$) the binomial variance of the allele dosage (which is equal to $p(1-p)$).

For the 2,557,248 SNPs examined in 9,813 individuals, the genomic inflation factor was $\lambda = 1.054$. Given the unique role of the major histocompatibility complex (located on chromosome 6) in MS, we also computed the genomic inflation factor after excluding SNPs found on chromosome 6: $\lambda = 1.048$. There were 2,142 SNPs that reached genome-wide significance ($P < 5 \times 10^{-8}$) in the meta-analysis. Of these, 2,141 SNPs were on chromosome 6, and one SNP (rs12025416, $P = 4.7 \times 10^{-8}$) was found within the *CD58* locus on chromosome 1. The most significant SNP (rs6931337, $P = 2.8 \times 10^{-167}$) was on chromosome 6. A SNP within HLA-DRA (rs3135388, $P = 7.4 \times 10^{-164}$) previously identified to be associated with MS is in high LD with rs6931337. As rs3135388 has been used previously as a surrogate for the HLA-DRB1*1501 susceptibility allele, we used it again for this purpose in the replication analysis.

**Identifying independent blocks of association for replication.** To organize our top results of our meta-analysis and select loci for the replication study, we used the "clump" routine from PLINK[44]. We applied an iterative approach where the marker with the most extreme evidence of association was used as the starting point and all other SNPs with an $r^2 > 0.5$ with best marker were grouped into one locus. The process was then repeated until 100 independent loci were defined. Known susceptibility alleles were included, based on previous work (ref. 4 and IMAGEN consortium, unpublished data). For HLA-B*4402, which emerged from a parallel set of analyses by the IMAGEN consortium (J. Oksenberg, unpublished data), the best surrogate marker (rs2743951) could not be designed as a SNP assay, and therefore rs2523393 was selected based on its strong LD ($r^2 = 0.92$) with rs2743951 in HapMap CEU samples (J. Oksenberg, unpublished data).

**Subjects used in the replication analysis.** There are two strata in our panel of replication samples (**Table 1**). As we do not have genome-wide data on these individuals, we matched them by country of origin and limited the analysis to subjects of self-reported European ancestry. **Supplementary Table 1b** contains the pertinent demographic details of these subjects. The UK component consists of an additional 831 subjects with MS collected at the University of Cambridge. These cases are matched to 1,030 subjects from the 1958 birth cohort who had not been genotyped genome-wide as part of the Wellcome Trust project and therefore represent an independent set of UK control samples from those used in the meta-analysis. The US stratum of the replication panel consists of subjects with MS from four different collections (demographic details are provided in **Supplementary Table 1b**): (i) BWH, 228 cases and 14 healthy controls; (ii) Accelerated Cure Project, 597 cases and 35 healthy controls; (iii) Washington University. 152 cases and 13 healthy controls; and (iv) UCSF, 407 cases and 142 healthy controls. To supplement the pool of healthy control subjects, we also included subjects of self-declared European ancestry from (i) the PhenoGenetic project, 292 healthy control subjects from a tissue bank of samples from subjects recruited in the Greater Boston metropolitan area to provide fresh blood for immunogenetic and other analyses; (ii) the healthy control collection at the Harvard/Partners Center for Genetics & Genomics, 101 healthy control subjects recruited from the Greater Boston metropolitan area (see URLS section below), and (iii) 489 healthy control subjects from the Chicago Health and Aging Project, a population-based study of healthy, nondemented, aging subjects centered in a suburb of Chicago[47].

**Genotyping platforms.** The platforms used to generate genome-wide data in each component of our meta-analysis are listed above in the description of these components. The Sequenom MASS Array platform in its iPLEX format was used to genotype the panel of SNPs selected for replication in the replication samples. Of the 188 SNPs selected for replication, 180 SNPs had data that met our quality control parameters: HWE $P > 1 \times 10^{-6}$; MAF $> 0.01$, genotype call rate $>0.95$.

**Replication analysis and joint analysis.** The replication analysis was conducted using a Cochran-Mantel Haenszel (CMH) approach as implemented in PLINK[44]. As genome-wide data were not available, we divided subjects (cases and controls) into two strata based on the country in which the sample was collected (US or UK); these two strata were used in the CMH analysis that is reported in detail in **Supplementary Table 3**. To perform a joint analysis of the meta-analysis and replication data, we calculated $Z$ scores for each of the two components of the analysis that were then added to calculate a joint $Z$ score (based on an estimated effective sample size (cases + controls) of 4,000 subjects for the meta-analysis and 4,500 subjects for the replication stage), from which the final $P$ values are determined. This approach is described in detail in a prior publication[11].

**Logistic regression for assessing independence of loci.** To assess whether SNPs in the same locus may have distinct effects on susceptibility to MS, we implemented a logistic regression analysis using stepwise selection, with the rank 1 SNP (SNP with most extreme $P$ value) being forced into the model first. We then calculated the residual effect of each of the other SNPs after accounting for the effect of the SNP with the most extreme evidence of association. The covariate in the model is the country of origin (US/UK) to account for possible population heterogeneity between the US and UK samples.

**Secondary analyses with a covariate for gender.** In the secondary analysis of the data included in our meta-analysis, we implemented a logistic regression analysis, with case-control as the outcome variable, and cohort of origin (six cohorts outlined in **Table 1**) and gender as two covariates in the model to account for possible heterogeneity between the cohorts and different sex ratio between case and control groups of the six cohorts. We selected 41 of the top SNPs that were not redundant with the results of the meta-analysis for inclusion in the replication effort. We used the same method to perform a secondary analysis of the replication data (see **Supplementary Table 7**); in that case, a covariate for the country of origin (UK or US) was included to minimize the possible effect of population stratification.

**RNA data and analysis.** Between July 2002 and October 2007, PBMC samples were collected from relapsing-remitting MS subjects and CIS subjects as part of the Comprehensive Longitudinal Investigation of MS at the Brigham & Women's Hospital (**Supplementary Table 1c**)[48]. CIS subjects differ from MS subjects by having had only one clinical episode of demyelination; MS subjects, by definition, must have at least two such events or one event and evidence of disease activity in a paraclinical measure such as MRI[45]. Nonethless, the pathophyisology is shared between these two sets of subjects, and they are treated in the same manner in a clinical environment[49]. PBMCs were isolated from heparinized blood by centrifugation on a Ficoll-Hypaque (Amersham Biosciences) gradient, and immediately frozen in 90% FBS and 10% DMSO. All blood samples were processed within 3 h of phlebotomy. Total RNA from frozen samples was isolated using a homogenization shredding system in a micro-centrifuge spin-column format (QIAshredder, Qiagen), followed by total RNA purification using selective binding columns (RNeasy Mini Kit, Qiagen), according to the manufacturer's protocol.

RNA concentration was determined using the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies). RNA quality was assessed on Agilent Bioanalyzer 2100 using the Agilent RNA 6000 Nano Assay kit (Agilent Technologies). The overall total RNA quality was evaluated by A260/A280 ratio (ratio $>$ 1.8) and electropherogram (score $>7$). Two micrograms of extracted RNA were reversed transcribed *in vitro* (Two-Cycle cDNA Synthesis Kit, Affymetrix), labeled (IVT Labeling Kit, Affymetrix) and hybridized on Affymetrix gene chip U133 2.0 plus. The GeneChip arrays were scanned on a GeneChip Scanner 3000.

Once generated, the RNA data underwent a rigorous quality control (QC) analysis using the recommended pipeline available in the R package simpleaffy and affyPLM (Bioconductor). The quality parameters that we monitored included (i) background noise, (ii) percentage of present called probe sets, (iii) scaling factor, (iv) information about exogenous control transcripts from the Affymetrix Poly-A control kit, and (v) the ratio of intensities of 3′/5′ probes for the housekeeping genes *GAPDH* and *β-actin*. We then normalized the data using GCRMA.

From our collection, 240 RNA profiles met our QC criteria and had genotypes for rs1800693 (*TNFRSF1A*), rs17445836 (*IRF8*) and rs17824933 (*CD6*). Next, we analyzed the correlation of rs17445836 (*ICSBP1*) with probes from its three target genes: *AICD*, *BCL6* and *TLR4*. For these analyses, all

subjects were considered together. We also performed secondary analyses in each treatment group (untreated, $n = 82$; IFNβ-treated, $n = 94$; GA-treated, $n = 64$), and these individual results mirrored the results obtained in the pooled analysis in each case (data not shown).

**Gene set enrichment analysis of quantitative trait analysis results** We implemented a quantitative trait analysis, using the Wald test as implemented in PLINK, of the rs17445836 polymorphism with all 54,676 Affymetrix U133 2.0 Plus probes; in all, 22,757 genes are sampled. Each treatment category (subjects are either untreated, IFNβ-treated and GA-treated) was analyzed separately using the normalized expression data described in the previous section. Only probes with an expression value $>3$ in each of the 240 subjects were considered for the next phase of the analysis. We use the Wald test as implemented in PLINK[44] for our quantitative trait analysis. The β value from the Wald test is used as the input variable of each probe in downstream analyses. If two or more probes mapped to the same gene, they were collapsed into one mean β value for that gene.

Gene set enrichment analysis (GSEA) version 2.0.1 (ref. 33) and the manually curated section C2 of MSigDB database were used in our subsequent analyses. Gene sets were preprocessed to exclude gene sets which contained $<15$ or $>200$ genes from our collection of 828 probes from untreated subjects that exceeded our threshold of $P < 0.05$ in the quantitative trait analysis. 93 out of a possible 1,892 gene sets met these criteria and were tested in our analysis. We performed 1,000 permutations of the analysis using the weighted enrichment statistic to estimate the statistical significance of our results. The process was repeated for the IFNβ-treated subjects (1,413 probes and 147 gene sets met our criteria and were tested) and the GA-treated subjects (3,191 probes and 223 gene sets). We consider replicated those results that were associated in both the untreated subjects and at least one of the treatment categories at an FDR $q$ value $<0.05$ (with the same direction of effect). Lack of replication between the two treated categories is difficult to interpret given the different mechanisms of action for GA and IFNβ. Sixteen genes tested met this criterion of replication (**Table 3**). Detailed results are presented in **Supplementary Table 6a**. The GA-treated group (the smallest subject group) does not have a significant overlap with either of the other two groups of subjects.

The GSEA report for each gene set includes (i) the number of genes used to evaluate a particular gene set, (ii) a normalized enrichment score (NES) which accounts for differences in gene set size and number of permutation performed, and (iii) a false discovery rate (FDR) $q$ value, a measure of statistical significance that accounts for the number of hypotheses tested.

**Ingenuity analysis.** The Ingenuity Pathway Analysis (IPA) tool (IPA Tool; Ingenuity Systems) was used to test whether its pre-defined "canonical interferon response pathway" was enriched in genes whose expression is correlated with the rs17445836[G] allele in our dataset generated from the PBMCs of subjects with MS. The data file we uploaded into this analysis tool is the same that was explored using the Gene Set Enrichment method described above. In short, it consists of all Affymetrix probesets who meet a $P < 0.05$ threshold in our quantitative trait analysis testing an additive model of association with rs17445836[G] (see previous section). For each probeset, its Affymetrix probeset ID and its β value from the quantitative trait analysis are loaded into the analysis tool. Each probeset ID is mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base.

The interferon signaling pathway that we have tested is the one found in the Ingenuity Pathways Analysis library of canonical pathways. This pathway contains 29 genes. Since we are testing a single hypothesis (association of higher interferon pathway gene expression with to rs17445836[G]) in this analysis, we use Fisher's exact test to derive a $P$ value that estimates the significance of the enrichment of interferon pathway genes in the list of genes that we have found to be associated with rs17445836[G].

**URLs.** MACH algorithm, http://www.sph.umich.edu/csg/abecasis/MACH/download/; Infevers, http://fmf.igh.cnrs.fr/ISSAID/infevers; Ingenuity Systems, http://www.ingenuity.com; http://www.hpcgg.org/BiosampleServices/overview.jsp.


44. 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).
45. 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).
46. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
47. Bienias, J.L., Beckett, L.A., Bennett, D.A., Wilson, R.S. & Evans, D.A. Design of the Chicago Health and Aging Project (CHAP). *J. Alzheimers Dis.* **5**, 349–355 (2003).
48. Gauthier, S.A., Glanz, B.I., Mandel, M. & Weiner, H.L. A model for the comprehensive investigation of a chronic autoimmune disease: the multiple sclerosis CLIMB study. *Autoimmun. Rev.* **5**, 532–536 (2006).
49. Miller, D., Barkhof, F., Montalban, X., Thompson, A. & Filippi, M. Clinically isolated syndromes suggestive of multiple sclerosis, part 2: non-conventional MRI, recovery processes, and management. *Lancet Neurol.* **4**, 341–348 (2005).