Evidence for Polygenic Susceptibility to Multiple Sclerosis—The Shape of Things to Come

The International Multiple Sclerosis Genetics Consortium (IMSGC)[1,*]

It is well established that the risk of developing multiple sclerosis is substantially increased in the relatives of affected individuals and that most of this increase is genetically determined. The observed pattern of familial recurrence risk has long suggested that multiple variants are involved, but it has proven difficult to identify individual risk variants and little has been established about the genetic architecture underlying susceptibility. By using data from two independent genome-wide association studies (GWAS), we demonstrate that a substantial proportion of the thousands of variants that individually fail to show statistically significant evidence of association have allele frequencies in cases that are skewed away from the null distribution through the effects of multiple as-yet-unidentified risk loci. The collective effect of 12,627 SNPs with Cochran-Mantel-Haenszel test ($p < 0.2$) in our discovery GWAS set optimally explains ~3% of the variance in MS risk in our independent target GWAS set, estimated by Nagelkerke's pseudo-R^2. This model has a highly significant fit ($p = 9.90E-19$). These results statistically demonstrate a polygenic component to MS susceptibility and suggest that the risk alleles identified to date represent just the tip of an iceberg of risk variants likely to include hundreds of modest effects and possibly thousands of very small effects.

Multiple sclerosis (MS [MIM 126200]) is a complex demyelinating disorder of the central nervous system. The etiology of the disease is not well understood but epidemiological studies provide undeniable evidence that genetic factors are involved.[1–4] Despite these extensive data, there was little progress identifying relevant genes prior to the emergence of genome-wide association studies (GWAS). For more than 30 years the only known MS genetic risk locus was the major histocompatibility complex (MHC) on chromosome 6p21.[5,6] All this changed in 2007 when the International Multiple Sclerosis Genetics Consortium (IMSGC) completed the first MS GWAS[7] and identified one new susceptibility locus, *IL2RA* (MIM 147730), and simultaneously with two other reports confirmed another susceptibility locus, *IL7R* (MIM 146661).[8,9] Since then progress has been rapid and the list of established MS risk loci has grown considerably (see Table 1). To date all of the newly identified MS risk alleles are common, exert only modest individual effects on risk (odds ratios 1.1–1.3), and ostensibly act independently. Even in combination with the much more substantial effects attributable to the MHC, these known MS risk loci account for less than half of the familial clustering observed epidemiologically. It is inescapable that other genetic risk factors exist and that much remains to be discovered about the genetics of MS. To aid in this discovery process, it would be helpful to estimate the form of this as-yet-undiscovered genetically determined risk.

GWAS, which by design examine common variation, have naturally driven the identification of common risk alleles for MS rather than rare variants. At the same time, the virtual absence of large extended MS families[10] coupled with the total absence of any consistent evidence for linkage outside the MHC region[11] indicates that rare alleles exerting very large effects are unlikely to play a prominent role in the genetics of MS. The relationship between familial recurrence risk and the degree of relatedness provides some guidance to the underlying genetic architecture of MS[12–14] and several studies have demonstrated that this relationship is distinctly nonlinear, suggesting that susceptibility is probably determined by multiple risk alleles, each exerting modest individual effects.[1–4,15,16] Unfortunately, this type of segregation analysis is insensitive to the number of involved loci and is largely unable to distinguish between dozens of modest effects, hundreds of weak effects, or thousands of very weak effects.[15] In line with the polygenic model proposed by Fisher,[17] it is logical to anticipate that the genetic architecture underlying susceptibility to MS will involve a wide spectrum of risk allele frequencies and effect sizes.[13] Under such a polygenic model, we would expect that many of the SNPs tested in a GWAS will be genuinely associated with disease, although most to only a minor degree. In this context we should see a systematic inflation of association scores across the genome, the pattern of which would be a reflection of the linkage disequilibrium (LD) between tested variants and the underlying risk alleles. The genome-wide inflation of association scores, beyond that which would be expected through sampling variance alone, is a well-recognized aspect of GWAS; however, most of this phenomenon is known to stem from subtle experimental biases, such as hidden population stratification and differential missingness.[13] In a recently published ground-breaking study, the International Schizophrenia Consortium (ISC)[18] confirmed the existence of polygenic influences in schizophrenia (MIM 181500) by showing that a small fraction of the variance in disease status (~3%) was significantly associated with a score based on multiple SNPs, none of

Table 1. Established Non-MHC MS Risk Alleles

Susceptibility Gene	RS Number	OR Estimate	Allele	Reference
IL7RA	rs6897932	1.18	C	Gregory et al., 2007;[8] IMSGC, 2007;[7] Lundmark et al., 2007[9]
IL2RA	rs12722489	1.25	C	IMSGC, 2007[7]
IL2RA	rs2104286	1.19	T	IMSGC, 2008[25]
CLEC16A	rs6498169	1.14	G	IMSGC, 2007[7]
CLEC16A	rs12708716	1.19	A	IMSGC, 2009[26]
CD58	rs12044852	1.24	C	IMSGC, 2007[7]
CD58	rs2300747	0.77	G	De Jager et al., 2008[19]
CD226	rs763361	1.12	A	IMSGC, 2009[26]
TYK2	rs34536443	1.32	G	Ban et al., 2009[27]
CD6	rs17824933	1.18	G	De Jager et al., 2008[19]
IRF8	rs17445836	0.8	A	De Jager et al., 2008[19]
TNFRSF1A	rs1800693	1.2	C	De Jager et al., 2008[19]
TNFRSF1A	rs4149584	1.58	T	De Jager et al., 2008[19]
METTL1	rs703842	0.81	G	ANZgene, 2009[28]
METTL1	rs10876994	0.82	C	ANZgene, 2009[28]
METTL1	rs12368653	1.15	A	ANZgene, 2009[28]
CD40	rs6074022	1.22	G	ANZgene, 2009[28]
CD40	rs1569723	1.22	C	ANZgene, 2009[28]

which were significantly associated with the disease at a genome-wide level in their own right. In short, they showed that despite sampling variance and the confounding influences of population stratification and differential missingness, the systematic influences of polygenic effects can be demonstrated. The ISC further demonstrated that the extent of variance explained increased as they incorporated additional SNPs with less significant p values into their scoring. With simulations they showed that this pattern most probably indicated that polygenic risk in schizophrenia is determined by modest numbers of larger effects and increasing numbers of smaller effects.[18]

Two previously published MS GWAS data sets were employed in this analysis. Data from the International Multiple Sclerosis Genetics Consortium (IMSGC) GWAS[7] was used in the discovery phase, while data from the GWAS performed by the Partners MS Center at the Brigham and Women's Hospital (BWH) and published as part of a recent meta-analysis[19] was used in the target phase. In both of these GWAS, MS was diagnosed according to standardized clinical criteria[20] and stringent quality control (QC) measures were applied to the data prior to analysis (see original publications for details[7,19]).

The IMSGC GWAS[7] was based on the Affymetrix Gene-Chip Human Mapping 500K Array Set and consisted of reliable data from 334,923 SNPs in 931 affected individuals and 2,431 controls: 1,475 from the Wellcome Trust Case

Control Consortium (randomly selected from the 1958 British Birth Cohort and the UK National Blood Transfusion Service) and 956 from the National Institute of Mental Health. Of the 931 cases, 453 were ascertained from sites across the UK, and 478 were ascertained from the Partners Healthcare MS Center in Boston and the UCSF MS Center. The BWH GWAS[19] was based on the Affymetrix Genome-wide Human SNP Array 6.0 (GeneChip 6.0) which, after excluding 54 individuals included in both GWAS, provided reliable data from 713,683 SNPs in 806 affected individuals ascertained from the Partners Healthcare MS Center in Boston and 1,720 controls ascertained by the MIGen consortium[21] (394 from Spain, 63 from Sweden, 12 from Finland, 746 from Seattle) and an additional 357 controls ascertained in Boston.

To avoid the issue of scoring hemizygous individuals, we based our analysis solely on autosomal SNPs and excluded X and Y linked variants. SNPs with minor allele frequency <0.05 were also excluded because in the sample sizes available to us these SNPs have little or no power. To avoid confounding influences from established MS susceptibility variants (Table 1), we excluded these along with all SNPs within 1 Mb of each established risk variant. Because of the more extensive LD in the MHC region and greater influence of this region, exclusion was extended to the entire MHC (chromosome 6, 29700KB–33300KB). After applying these exclusions, we were left with 311,315 SNPs with high-quality genotypes common to both studies. By using the genotypes from both discovery and target sets, we further pruned this consensus SNP panel with the PLINK software suite to remove SNPs in strong linkage disequilibrium with each other and thereby generate a set of 59,470 essentially independent SNPs. We used an r^2 threshold of 0.2 within a 500 SNP window, sliding 5 SNPs at a time.

We applied established methods for computing individual risk score profiles[22,23] to assess the polygenic effect of multiple sclerosis. With the PLINK software suite, Cochran-Mantel-Haenszel test statistics were computed for all SNPs in the discovery set, stratifying by geographic origin (UK versus US samples). Based on these CMH association tests, overlapping sets of scoring SNPs were generated based on test significance threshold (p_T): $p_T < 0.01$, $p_T < 0.05$, and p_T deciles between 0.1 and 1 (Table 2).

By using the odds ratio point estimates from the discovery set, we calculated an aggregate score for each individual in the target set equal to the number of score alleles weighted by the log of the odds ratio. Missing genotypes in target individuals were given the mean score for that locus as seen across the individuals not missing a genotype at that locus. All score profiles were calculated via PLINK's score function.

To allow correction for any residual signal attributable to established risk alleles that had not been removed by our regional exclusion of SNPs flanking established risk alleles, we generated a score profile based on the known susceptibility variants listed in Table 1 (with rs3135388 as the

**Table 2. Target GWAS Statistics for Each Scoring SNP Set**

p Value Threshold Range	Number of SNPs	p Value	Nagelkerke's Pseudo-R^2
0–0.01	746	3.14E-09	0.0126
0–0.05	3,272	1.39E-16	0.0243
0–0.1	6,426	2.45E-18	0.0271
0–0.2	12,627	2.62E-19	0.0287
0–0.3	18,564	7.53E-19	0.0279
0–0.4	24,565	1.29E-19	0.0292
0–0.5	30,365	8.93E-20	0.0294
0–0.6	36,160	3.35E-20	0.0301
0–0.7	42,075	1.22E-20	0.0308
0–0.8	48,012	1.73E-20	0.0305
0–0.9	53,748	1.05E-20	0.0309
0–1	59,470	6.12E-21	0.0313

The number of scoring SNPs falling into each p value threshold defined set is shown (note these numbers are of course overlapping). Both the p value and the pseudo-R^2 are reported as a difference between a full model containing score and covariates and a reduced model containing covariates only, evaluated by likelihood ratio test.

Figure 1. Nagelkerke's Pseudo-R^2 for Each Scoring SNP Set
Variance explained in the target set, measured by Nagelkerke's pseudo-R^2, is shown for each of the seven scoring SNP sets as defined by arbitrary discovery GWAS p value thresholds.

surrogate for the MHC effect). Where susceptibility SNPs were not typed in the target set, we selected proxies with the highest r^2 to the susceptibility SNP, with ties broken by lowest physical distance (Table S1 available online).

With the aggregate scores generated in the target set, we conducted logistic regression analysis to test the relationship between the computed scores and disease status. The score based on known-susceptibility variants and the total number of nonmissing alleles over all SNPs were included as covariates in the analysis. Logistic regression analysis was conducted with STATA 10.1. The variance in case/control status explained by the score statistic was estimated as the difference in variance (with the Nagelkerke pseudo-R^2) between a model including score and the covariates (full model) versus the covariates alone (reduced model). p values were computed as 1-df likelihood ratio tests between the full and reduced models.

To test for polygenic influences in multiple sclerosis, we considered data from two MS GWAS[7,19] and explored how the risk of MS was related to the effects of multiple loci considered collectively ("en masse").[18] After excluding SNPs in LD with known MS susceptibility variants (see Table 1) and filtering SNPs on the basis of quality control measures, allele frequency, and mutual LD, we established a set of 59,470 essentially independent autosomal SNPs common to both MS GWAS. These scoring SNPs were then considered in seven sets defined on the basis of nominal significance thresholds observed in the first GWAS[7] (Table 2 indicates the number of scoring SNPs in each significance threshold defined set). The first GWAS[7] (the discovery data set) was used to establish a score for each allele at each of the 59,470 SNPs. Although these
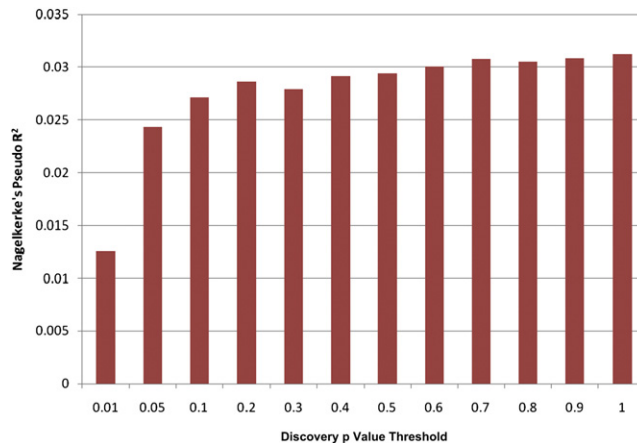
odds ratio point estimates have wide confidence intervals, they are inevitably correlated with the true risk for the assessed alleles. In the target data set (the second MS GWAS[19]), we then calculated the aggregate score for each sample by summing the number of risk alleles present in the individual weighted by the score for that allele as calculated in the discovery set. These aggregate scores were then subject to logistic regression analysis to determine whether they account for significant variance in disease status within the target data set. This process was repeated for each of the sets of scoring SNPs.

Each of the scoring SNP sets showed highly significant evidence for association to disease status (p = 3.14E-9–p = 6.12E-21), confirming that polygenic inheritance determined by as-yet-unidentified genetic variation is relevant in multiple sclerosis (see Figure 1 and Table 2).

Our analysis provides highly significant evidence for a polygenic component to the genetics of MS. As with other complex diseases, it is becoming increasingly clear that single genes explaining a large proportion of heritability do not exist.[24] The results of this study indicate that MS risk is governed in part by the cumulative effect of multiple genetic variants scattered across the genome, each contributing only a modest individual effect.

In contrast to the results of the schizophrenia study, the variance explained in MS (as measured by the Nagelkerke's pseudo-R^2 value) tended to stabilize once the SNP inclusion cut-off reached an initial p value of 0.2. On the surface, this might suggest that SNPs with somewhat larger effects (e.g., odds ratios of 1.1–1.3) are more prevalent in MS than in schizophrenia. However, the fact that the level of pseudo-R^2 does not fall in scoring sets based on less significant thresholds suggests that this flattening may be due to limited power more than an absence of genuinely associated variants among those SNPs with p values of 0.2–0.5.

Some portion of the identified polygenic effect could be due to cryptic population stratification remaining in the samples of European descent. However, common methods

for correcting for such stratification rely on the assumption that there is no widely dispersed genomic effect on the trait. This is exactly what we are testing, so applying such correction will probably eliminate a true polygenic effect. As pointed out in the original paper from the ISC,[18] en masse effects like those we have demonstrated are unlikely to result from stratification because the same structure would need to be present in both data sets to produce correlation between studies, a highly unlikely scenario.

Although this analysis provides convincing evidence for the existence of a polygenic influence on susceptibility, it provides little if any guidance as to the location of the relevant variants. Mapping these additional effects can be achieved only through larger GWAS. Most of the common variants with odds ratio > 1.2 should emerge from a GWAS involving ~10,000 cases and ~10,000 controls, but even larger studies will be required to map weaker effects.

Supplemental Data

Supplemental Data include one table and can be found with this article online at http://www.cell.com/AJHG.

Web Resources

The URL for data presented herein is as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/

References

1. Kenealy, S.J., Pericak-Vance, M.A., and Haines, J.L. (2003). The genetic epidemiology of multiple sclerosis. J. Neuroimmunol. 143, 7–12.

2. Mumford, C.J., Wood, N.W., Kellar-Wood, H., Thorpe, J.W., Miller, D.H., and Compston, D.A. (1994). The British Isles survey of multiple sclerosis in twins. Neurology 44, 11–15.

3. Robertson, N.P., Fraser, M., Deans, J., Clayton, D., Walker, N., and Compston, D.A. (1996). Age-adjusted recurrence risks for relatives of patients with multiple sclerosis. Brain 119, 449–455.

4. Sadovnick, A.D., and Ebers, G.C. (1995). Genetics of multiple sclerosis. Neurol. Clin. 13, 99–118.

5. Barcellos, L.F., Sawcer, S., Ramsay, P.P., Baranzini, S.E., Thomson, G., Briggs, F., Cree, B.C., Begovich, A.B., Villoslada, P., Montalban, X., et al. International Multiple Sclerosis Genetics Consortium. (2006). Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. Hum. Mol. Genet. 15, 2813–2824.

6. Yeo, T.W., De Jager, P.L., Gregory, S.G., Barcellos, L.F., Walton, A., Goris, A., Fenoglio, C., Ban, M., Taylor, C.J., Goodman, R.S., et al. International Multiple Sclerosis Genetics Consortium. (2007). A second major histocompatibility complex susceptibility locus for multiple sclerosis. Ann. Neurol. 61, 228–236.

7. Hafler, D.A., Compston, A., Sawcer, S., Lander, E.S., Daly, M.J., De Jager, P.L., de Bakker, P.I., Gabriel, S.B., Mirel, D.B., Ivinson, A.J., et al. International Multiple Sclerosis Genetics Consortium. (2007). Risk alleles for multiple sclerosis identified by a genomewide study. N. Engl. J. Med. 357, 851–862.

8. Gregory, S.G., Schmidt, S., Seth, P., Oksenberg, J.R., Hart, J., Prokop, A., Caillier, S.J., Ban, M., Goris, A., Barcellos, L.F., et al. Multiple Sclerosis Genetics Group. (2007). Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. Nat. Genet. 39, 1083–1091.

9. Lundmark, F., Duvefelt, K., Iacobaeus, E., Kockum, I., Wallström, E., Khademi, M., Oturai, A., Ryder, L.P., Saarela, J., Harbo, H.F., et al. (2007). Variation in interleukin 7 receptor alpha chain (IL7R) influences risk of multiple sclerosis. Nat. Genet. 39, 1108–1113.

10. Willer, C.J., Dyment, D.A., Cherny, S., Ramagopalan, S.V., Herrera, B.M., Morrison, K.M., Sadovnick, A.D., Risch, N.J., and Ebers, G.C. (2007). A genome-wide scan in forty large pedigrees with multiple sclerosis. J. Hum. Genet. 52, 955–962.

11. Sawcer, S., Ban, M., Maranian, M., Yeo, T.W., Compston, A., Kirby, A., Daly, M.J., De Jager, P.L., Walsh, E., Lander, E.S., et al. International Multiple Sclerosis Genetics Consortium. (2005). A high-density screen for linkage in multiple sclerosis. Am. J. Hum. Genet. 77, 454–467.

12. Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. Am. J. Hum. Genet. 46, 222–228.

13. Wang, W.Y., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: Theoretical and practical concerns. Nat. Rev. Genet. 6, 109–118.

14. Yang, Q., Khoury, M.J., Friedman, J., Little, J., and Flanders, W.D. (2005). How many genes underlie the occurrence of common complex diseases in the population? Int. J. Epidemiol. 34, 1129–1137.

15. Lindsey, J.W. (2005). Familial recurrence rates and genetic models of multiple sclerosis. Am. J. Med. Genet. A. 135, 53–58.

16. Sadovnick, A.D., Dyment, D., and Ebers, G.C. (1997). Genetic epidemiology of multiple sclerosis. Epidemiol. Rev. 19, 99–106.

17. Fisher, R.A. (2009). The correlations between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. *52*, 399–433.

18. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P., and International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

19. De Jager, P.L., Jia, X., Wang, J., de Bakker, P.I., Ottoboni, L., Aggarwal, N.T., Piccio, L., Raychaudhuri, S., Tran, D., Aubin, C., et al. International MS Genetics Consortium. (2009). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat. Genet. *41*, 776–782.

20. McDonald, W.I., Compston, A., Edan, G., Goodkin, D., Hartung, H.P., Lublin, F.D., McFarland, H.F., Paty, D.W., Polman, C.H., Reingold, S.C., et al. (2001). Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann. Neurol. *50*, 121–127.

21. Kathiresan, S., Voight, B.F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P.M., Anand, S., Engert, J.C., Samani, N.J., Schunkert, H., et al. Myocardial Infarction Genetics Consortium, Wellcome Trust Case Control Consortium. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat. Genet. *41*, 334–341.

22. Wray, N.R., Goddard, M.E., and Visscher, P.M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. *17*, 1520–1528.

23. Evans, D.M., Visscher, P.M., and Wray, N.R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum. Mol. Genet. *18*, 3525–3531.

24. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

25. International Multiple Sclerosis Genetics Consortium (IMSGC). (2008). Refining genetic associations in multiple sclerosis. Lancet Neurol. *7*, 567–569.

26. International Multiple Sclerosis Genetics Consortium (IMSGC). (2009). The expanding genetic overlap between multiple sclerosis and type I diabetes. Genes Immun. *10*, 11–14.

27. Ban, M., Goris, A., Lorentzen, A.R., Baker, A., Mihalova, T., Ingram, G., Booth, D.R., Heard, R.N., Stewart, G.J., Bogaert, E., et al. Wellcome Trust Case-Control Consortium (WTCCC). (2009). Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. Eur. J. Hum. Genet. *17*, 1309–1313.

28. Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). (2009). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat. Genet. *41*, 824–828.